TESIS DOCTORAL

2025-2026

FROM DETECTION TO CHARACTERIZATION: A HOLISTIC AND PERSPECTIVIST APPROACH TO MEDIA BIAS ANALYSIS

Francisco Javier Rodrigo Ginés

PROGRAMA DE DOCTORADO EN SISTEMAS INTELIGENTES

Prof.ª Dra. Laura Plaza
Prof. Dr. Jorge Carrillo de Albornoz

This page intentionally left blank.

From detection to characterization: a holistic and perspectivist approach to media bias analysis

Dissertation Thesis

Francisco-Javier Rodrigo-Ginés

June 3, 2026

Submitted in partial fulfillment of the requirements
for the degree of *Doctor en Sistemas Inteligentes*

to the

Escuela Internacional de Doctorado
at Universidad Nacional de Educación a Distancia (UNED)

Supervised by:
Prof. Dr. Jorge Carrillo-de-Albornoz
Prof. Dra. Laura Plaza

COLOPHON

This thesis was typeset using LaTeX and the `memoir` documentclass. It is based on Aaron Turon's thesis *Understanding and expressing scalable concurrency*, itself a mixture of `classic thesis` by André Miede and `tufte-latex`, based on Edward Tufte's *Beautiful Evidence*.

The bibliography was processed by Biblatex. All graphics and plots are made with PGF/Ti*k*Z, Seaborn, and Inkscape.

The body text is set 10/13pt on a 26pc measure. The margin text is set in `scriptsize`. Matthew Carter's Charter acts as both the text and display typeface. Monospaced text uses Jim Lyles's `Bitstream Vera Mono` ("Bera Mono").

Throughout the writing process, the OpenAI ChatGPT-4o LLM model was employed as a grammar checker and for enhancing the clarity of the text. It was not, however, used for generating original ideas or content. All creative and intellectual work presented here remains entirely the author's own.

Given that this thesis explores bias in the media, some examples may contain strong or politically charged content. Every effort has been made to approach these cases with neutrality, focusing on their analytical and methodological relevance rather than endorsing any particular perspective.

"*Not all those who wander are lost*"
—J.R.R. Tolkien

This page intentionally left blank.

Abstract

Media bias, understood as the systematic tendency of news outlets to present information in ways that favour particular perspectives through linguistic, rhetorical, or editorial strategies, has become a central concern for democratic accountability and media literacy. Over the past decade, a growing body of research has addressed the automatic detection of media bias using Natural Language Processing and Machine Learning techniques. However, existing approaches predominantly rely on document-level binary classification, shallow lexical features, and single-dataset evaluation, which limits both their analytical depth and their practical applicability.

This dissertation addresses the limitations of current automatic media bias detection resources and proposes a holistic, perspectivist, and multi-level framework to overcome them. Through extensive cross-dataset experiments, we demonstrate that existing corpora fail to generalize bias detection beyond their original domains. This lack of transferability stems from insufficient linguistic, cultural, and topical representativeness, as well as from oversimplified modeling assumptions that reduce media bias to a single binary or categorical label per article.

To tackle this issue, we propose that media bias should be analyzed as a context-sensitive and cognitively mediated phenomenon, requiring more fine-grained, socially aware, and flexible modeling. This involves shifting the unit of analysis from entire articles or media outlets to individual sentences, explicitly capturing different manifestations of bias such as sensationalism, mind reading, biased word choice, or subjective qualifiers. It also demands embracing linguistic and sociocultural variation and treating bias perception as inherently perspectivist and contested.

Based on these premises, we first develop a comprehensive taxonomy of media bias manifestations and categories, grounded in linguistic theory and communication studies. We then introduce MBBMD (Media Bias Bias-Mitigated Dataset), a novel Spanish-language resource annotated at three hierarchical levels: sentence, document, and bias manifestation category. Crucially, the dataset adopts a perspectivist approach that preserves annotator disagreement as a meaningful signal rather than noise, capturing how bias perception varies across ideological perspectives.

Building upon MBBMD, we design a hierarchical system for media bias analysis that integrates three types of models: (1) linguistic-heuristic algorithms that detect specific bias manifestations at the sentence level; (2) transformer-based classifiers fine-tuned for bias detection across ideological perspectives; and (3) ensemble metamodels that aggregate multi-level outputs to generate final predictions.

Our experiments show that this hierarchical and perspectivist approach yields substantial improvements. Specifically, document-level performance increases from 47% F1-score (flat document-level baseline) to 77% F1-score (∼64% relative gain), reaching a level of generalization and robustness that rivals models trained and tested on the same dataset. Moreover, by targeting concrete manifestations of media bias and preserving intersubjective disagreement, the system supports higher explainability, making it more suitable for real-world applications and human-in-the-loop evaluation.

Universidad Nacional de Educación a Distancia - PhD in Intelligent Systems

From detection to characterization: a holistic and perspectivist approach to media bias analysis

Modeling media bias: challenges of definition, annotation, and generalization

Most datasets for media bias detection rely on majority-vote labels, overlooking the diversity of interpretations bias can evoke. Models trained on such data often fail to generalize across contexts, exposing deeper conceptual and methodological gaps:

- No clear, shared definition of what constitutes media bias
- Poor generalization of models across datasets and domains
- Lack of attention to the reader's perspective in annotation and evaluation

2D t-SNE visualization highlights output separation among five media bias detection models

Research hypotheses

A holistic, taxonomy-based and perspectivist approach outperforms traditional models in media bias detection.

H1: Fine-grained linguistic instantiation

≡ Media bias is instantiated at the sentence level through linguistic and rhetorical mechanisms.

✓ Validated via sentence-level classifiers + ablation.

H2: Perspectival perception

≡ Media perception depends on ideological and cognitive priors.

✓ Validated via Perspectivist annotations + ideology-aware models.

H3: Multi-level integration

≡ Media bias emerges from the integration of cues across representational levels..

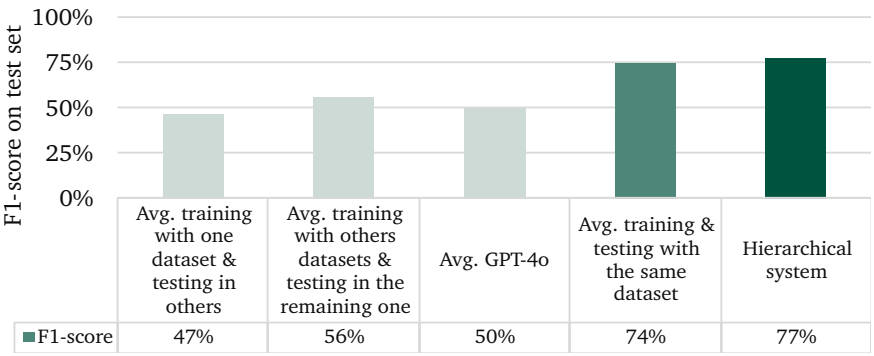
✓ Validated via Hierarchical modelling + ensemble integration.

Foundation and architecture of the holistic approach

To operationalize media bias from a multidimensional perspective, we developed a threefold foundation:

- A two-tier **taxonomy of media bias** manifestations, distinguishing rhetorical strategies by context (e.g., statement, coverage, gatekeeping) and by authorial intent (ideological bias vs. spin). It defines 17 manifestations grounded in discourse theory.
- A novel **Media Bias Bias-Mitigated Dataset** (MBBMD) structured along this taxonomy. It includes annotations at:
  - Sentence level (2,000 labeled bias spans)
  - Document level (ideologically disaggregated judgments)
  - Category level (taxonomy-aligned categories)
- A **perspectivist hierarchical system** integrating:
  - Sentence-level bias cues
  - Ideology-aware document classifiers
  - Continuous media bias intensity modeling

This modular architecture, powered by ensemble learning, validates our hypotheses and achieves robust generalization across datasets.



Main contributions: a media bias taxonomy, a perspectivist dataset structured around it, and a holistic model spanning from fine-grained to document and category-level bias detection.

Resumen

El sesgo mediático, entendido como la tendencia sistemática de los medios de comunicación a presentar la información de manera que favorezca determinadas perspectivas a través de estrategias lingüísticas, retóricas o editoriales, se ha convertido en una preocupación central para la rendición de cuentas democrática y la alfabetización mediática. Durante la última década, un creciente cuerpo de investigación ha abordado la detección automática del sesgo mediático mediante técnicas de Procesamiento del Lenguaje Natural y Aprendizaje Automático. Sin embargo, los enfoques existentes se basan predominantemente en clasificación binaria a nivel de documento, características léxicas superficiales y evaluación sobre un único corpus, lo que limita tanto su profundidad analítica como su aplicabilidad práctica.

Esta tesis aborda las limitaciones de los recursos actuales para la detección automática de sesgo mediático y propone un enfoque holístico, perspectivista y multinivel para superarlas. A través de experimentos de generalización entre conjuntos de datos, se demuestra que los corpus existentes no permiten transferir la detección de sesgo más allá de sus propios dominios. Esta falta de transferibilidad se debe a una escasa representatividad lingüística, cultural y temática, así como a supuestos metodológicos reduccionistas que simplifican el sesgo mediático como una etiqueta binaria o categórica por artículo.

Frente a ello, defendemos que el sesgo mediático debe analizarse como un fenómeno sensible al contexto y mediado cognitivamente, lo que requiere modelos más granulares, conscientes del entorno sociolingüístico, y capaces de representar la subjetividad inherente a su percepción. Esto implica cambiar la unidad de análisis desde el artículo completo hacia la oración, capturando manifestaciones concretas del sesgo como el sensacionalismo, el "mind reading", la elección léxica sesgada, o el uso de calificativos subjetivos. Asimismo, abogamos por incorporar la variación cultural y tratar el sesgo como un fenómeno perspectivista y contestado.

A partir de esta base teórica, se desarrolla una taxonomía de manifestaciones y categorías del sesgo mediático fundamentada en la teoría lingüística y los estudios de comunicación. Sobre esta taxonomía se construye MBBMD (Media Bias Bias-Mitigated Dataset), un nuevo recurso en español anotado jerárquicamente en tres niveles: oración, documento y categoría de sesgo. El proceso de anotación sigue un protocolo perspectivista que preserva el desacuerdo entre anotadores como una señal significativa, capturando cómo la percepción del sesgo varía según la perspectiva ideológica.

Sobre el dataset MBBMD se construye un sistema jerárquico de análisis del sesgo mediático que integra tres tipos de modelos: (1) heurísticas lingüísticas para detectar manifestaciones específicas de sesgo a nivel de oración; (2) clasificadores basados en transformers adaptados al sesgo y a las distintas perspectivas ideológicas; y (3) metamodelos de ensemble que agregan las salidas de los modelos anteriores para producir una predicción final.

Los resultados experimentales muestran mejoras sustanciales: la puntuación F1 a nivel de documento aumenta del 47% (baseline plano a nivel de documento) al 77% (una ganancia relativa de ∼64%), alcanzando un nivel de generalización que antes sólo se lograba entrenando y evaluando sobre el mismo corpus. Además, al centrarse en manifestaciones concretas del sesgo y preservar el desacuerdo intersubjetivo, el sistema ofrece mayor explicabilidad, lo que lo hace más adecuado para su uso en entornos reales y evaluación human-in-the-loop.

x

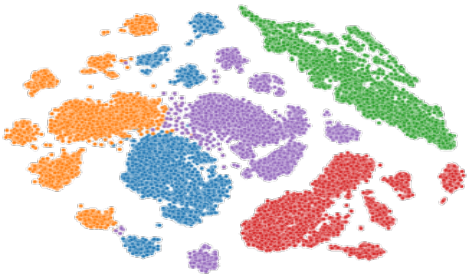
Universidad Nacional de Educación a Distancia – Doctorado en Sistemas Inteligentes

From detection to characterization :
a holistic and perspectivist approach to media bias analysis

Modelar el sesgo mediático: desafíos en la definición, anotación y generalización

La mayoría de los conjuntos de datos para la detección de sesgo mediático ignoran la diversidad de interpretaciones que el sesgo puede suscitar. Los modelos entrenados con este tipo de datos no generalizan bien entre contextos, lo que pone de manifiesto carencias conceptuales y metodológicas más profundas:

- No existe una definición clara y compartida de qué constituye el sesgo mediático
- Los modelos generalizan mal entre distintos datasets.
- No se tiene en cuenta la perspectiva del lector durante la anotación ni la evaluación



Visualización 2D t-SNE que muestra separación de salidas entre cinco modelos de detección de sesgo

Hipótesis de la tesis

Un enfoque holístico y perspectivista, supera a los modelos tradicionales en la detección de sesgo mediático.

H1: Detección fina de sesgo
≡ El sesgo mediático se instancia a nivel de frase mediante mecanismos lingüísticos y retóricos.
✓ Validado con clasificadores a nivel de frase + ablación.

H2: Percepción perspectivista
≡ La percepción del sesgo está mediada por disposiciones ideológicas y cognitivas.
✓ Validado con anotación perspectivista + modelos sensibles a la ideología..

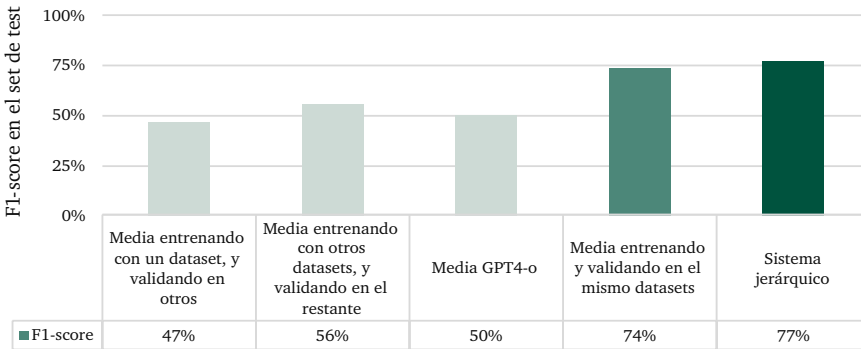
H3: Integración multinivel
≡ El sesgo mediático emerge de la integración de señales en distintos niveles representacionales.
✓ Validado mediante modelado jerárquico + ensembles.

Fundamentos y arquitectura del enfoque holístico

Para operacionalizar el sesgo mediático desde una perspectiva multidimensional, desarrollamos una base tripartita:

- Una **taxonomía de sesgo mediático** de dos niveles, que distingue estrategias retóricas por contexto (declaración, cobertura, y selección temática) y por intención del autor (sesgo ideológico vs. spin). Define 17 manifestaciones mapeadas desde la teoría del discurso.
- Un novedoso **conjunto de datos de sesgo mediático** (MBBMD) estructurado en torno a esta taxonomía. Incluye anotaciones a:
  - Nivel de frase (más de 2.000 sesgos anotados)
  - Nivel documental (juicios ideológicos desagregados)
  - Nivel de categoría (clases alineadas con la taxonomía)
- Un **sistema jerárquico perspectivista** que integra:
  - Pistas lingüísticas a nivel de frase
  - Clasificadores documentales ideológicamente informados
  - Modelado continuo de la intensidad del sesgo

Esta arquitectura modular, valida nuestras hipótesis y logra una generalización robusta entre dominios.



Contribuciones principales: una taxonomía de sesgo mediático, un conjunto de datos perspectivistas estructurado en torno a ella y un modelo holístico que abarca desde la detección de sesgo a nivel de granularidad fina hasta la detección de sesgo a nivel de documento y categoría.

Acknowledgments

Writing these acknowledgements after having revised this thesis so many times feels, at the very least, weird. Perhaps because bringing a work like this to an end is not only about finishing a text, but about coming to terms with all the time, decisions, and renunciations that have made it possible.

This thesis is also the result of a social and institutional context that should not be taken for granted. Being able to pursue a doctoral degree, together with everything it entails before, during, and after, cannot be explained solely by individual effort. It is deeply tied to the existence of a welfare state that supports public education. Not so long ago, a path like this would have been simply unthinkable. That it is possible today does not make it permanent: it is a collective achievement that, as citizens, we have a responsibility to preserve and protect.

I am grateful to the University as a universal institution: a shared space for knowledge, debate, and transmission. Within that journey, the University of Jaén holds a fundamental place. Many of the questions that eventually shaped this thesis were first formed there, and I keep academic and personal memories from that period. I would like to explicitly acknowledge Víctor Manuel Ribas, whose letter of recommendation made access to the doctoral program possible; without that gesture, this thesis would simply not exist. I also thank the Universidad Nacional de Educación a Distancia, for its open and inclusive spirit, and the NLP&IR research group in particular, for their support and welcome. I consider both institutions, UJAEN and UNED, to be my *alma mater*, each of them shaping my academic trajectory in distinct yet complementary ways.

Throughout this process, I have also relied on colleagues from other universities and whose support proved especially valuable during the doctorate. Flor Plaza and Jorge Chamorro have been a constant reference, offering alternative perspectives on academic life, thoughtful conversations, and help that was both practical and intellectual.

To all those who, over the past six years, have repeatedly asked me "how is the thesis going?". You probably do not know how complicated, and at times genuinely painful, that question could be to answer. Even so, that persistent interest has often served, perhaps unknowingly, as a reminder that this work was still there, unfinished, and that it had to be brought to an end.

To my supervisors, Jorge Carrillo de Albornoz and Laura Plaza, I owe a particular debt of gratitude. Their guidance, judgment, and support have been a constant point of reference, essential in turning scattered ideas into a coherent and defensible piece of work, and in learning how to conduct research rigorously without losing autonomy.

I would also like to thank my parents, for something far more basic and, for that very reason, more important: instilling in me the value of studying, sustained effort, and intellectual curiosity. That foundation long predates this thesis and runs through it entirely.

Last but not least, Celia deserves an acknowledgement of her own. This doctorate has unfolded alongside many other aspects of life, and she has been present through all of them. She has shared impossible deadlines, moments of blockage, stolen weekends, and a thesis that at times seemed endless. She has offered calm when it was needed, perspective when it was lost, and steady support when the process became particularly demanding. Completing this thesis is as much her achievement as it is mine, not as a figure of speech, but as a fact. I am also sincerely grateful to her family for their support and closeness, and for helping to sustain this journey over the years.

This work opens with a phrase that has accompanied this entire trajectory: *not all those who wander are lost*. Pursuing a part-time PhD while working in industry, along a path that is not always linear, can appear erratic from the outside. Yet even in moments of dispersion, there has always been a sense of direction. This thesis stands as the confirmation that the path, with all its turns, was neither accidental nor in vain.

This page intentionally left blank.

Contents

This page intentionally left blank.

List of abbreviations

Core concepts

AI Artificial Intelligence

ML Machine Learning

DL Deep Learning

NLP Natural Language Processing

Learning paradigms and methods

LR Logistic Regression

SVM Support Vector Machine

RF Random Forest

NB Naive Bayes

MTL Multi-Task Learning

PCA Principal Component Analysis

DBSCAN Density-Based Spatial Clustering of Applications with Noise

LDA Latent Dirichlet Allocation

Neural networks and language models

LLM Large Language Model

RNN Recurrent Neural Network

LSTM Long Short-Term Memory

Bi-LSTM Bidirectional Long Short-Term Memory

BERT Bidirectional Encoder Representations from Transformers

BART Bidirectional and Auto-Regressive Transformers

ConvBERT Convolutional BERT

RoBERTa Robustly Optimized BERT Pretraining Approach

DistilBERT Distilled BERT

DA-BERT Domain-Adaptive BERT

DA-RoBERTa Domain-Adaptive RoBERTa

DA-BART Domain-Adaptive BART

XLM-RoBERTa Cross-Lingual RoBERTa

GPT Generative Pre-trained Transformer

LLaMA Large Language Model Meta AI

CLIP Contrastive Language-Image Pretraining

RAG Retrieval-Augmented Generation

Linguistic and textual representations

PoS Part-of-Speech

NER Named Entity Recognition

LIWC Linguistic Inquiry and Word Count

TF-IDF Term Frequency-Inverse Document Frequency

PMI Pointwise Mutual Information

USE Universal Sentence Encoder

HITS Hyperlink-Induced Topic Search

PMF Probability Mass Function

▸ **Explainability and interpretability techniques**

LIME Local Interpretable Model-agnostic Explanations

SHAP SHapley Additive exPlanations

▸ **Datasets and frameworks for media bias detection**

MBFC Media Bias/Fact Check

MBIC Media Bias Identification Corpus

BASIL Bias Annotation in Sentence-level Interpretations of Language

BABE Bias Annotation for Biased Examples

FIGNEWS Framing and Information Gathering in News

NELA News Landscape Dataset

MBBMD Media Bias Bias-Mitigated Dataset

▸ **Bias mitigation strategies: data-centric and annotation-centric**

CDA Counterfactual Data Augmentation

DACE Data-Augmented Contextual Enrichment

RACE Retrieval-Augmented Contextual Enrichment

LeWiDi Learning with Disagreements

▸ **Principles and reporting standards**

FAIR Findable, Accessible, Interoperable, Reusable

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

▸ **Evaluation metrics**

MSE Mean Squared Error

RMSE Root Mean Squared Error

MAE Mean Absolute Error

R^2 Coefficient of Determination (R-squared)

ROC-AUC Receiver Operating Characteristic - Area Under the Curve

BLEU Bilingual Evaluation Understudy

ROUGE Recall-Oriented Understudy for Gisting Evaluation

PERMANOVA Permutational Multivariate Analysis of Variance

List of Figures

List of Tables

This page intentionally left blank.

List of Algorithms

This page intentionally left blank.

1

Introduction

"The beginning of knowledge is the discovery of something we do not understand."
—Frank Herbert

In an era where information flows more freely and rapidly than ever before, the rise of digital media has fundamentally reshaped the landscape of communication and knowledge dissemination. While these technological advancements have democratized access to information, they have also introduced significant challenges, particularly the prevalent problem of disinformation. Disinformation, defined as the intentional spread of false or misleading information with the aim of deceiving, has emerged as a severe threat to informed public discourse, social cohesion, and the integrity of democratic institutions.

Disinformation, unlike mere misinformation, which can occur without malicious intent [39], is a calculated strategy designed to manipulate perceptions, fuel divisions, and undermine trust in credible sources of information. Its impact is far-reaching, influencing electoral outcomes, public health initiatives, and the general public's understanding of critical societal issues. As disinformation becomes more sophisticated, understanding its mechanisms, effects, and potential countermeasures becomes increasingly urgent.

[39] Cummings and Kong, 2019, "Breaking down "fake news": Differences between misinformation, disinformation, rumors, and propaganda"

The spread of disinformation is not merely about disseminating falsehoods; it involves the strategic manipulation of truth, the exploitation of media systems, and the psychological vulnerabilities of its audiences. Social media platforms have amplified these issues, creating environments where information spreads virally and where users are often exposed to content that reinforces their pre-existing beliefs. These so-called "information bubbles" [125] or "echo chambers" [187] exacerbate polarization, making it increasingly difficult for individuals to engage with differing perspectives, and thereby rendering societies more vulnerable to the corrosive effects of disinformation.

[125] Pariser, 2011, *The filter bubble: What the Internet is hiding from you*

[187] Terren and Borge, 2021, "Echo chambers on social media: A systematic review of the literature"

1.1 Media bias in contemporary information ecosystems

Integral to the landscape of disinformation is the phenomenon of media bias, which plays a crucial role in shaping how information is presented, perceived, and ultimately accepted by the public. In this thesis, we define media bias as the systematic tendency of media outlets to present information in ways that favour particular perspectives through linguistic, rhetorical, or editorial strategies, whether intentionally or as a consequence of institutional and cognitive factors (a detailed characterization is developed in Chapter 3). This bias can manifest in various forms, from the selection and framing of news stories to the language and tone used in reporting. In the context of disinformation, media bias not only contributes to the dissemination of false information but also influences how such information is received and interpreted by different segments of the audience [118].

[118] Mullainathan and Shleifer, 2002, *Media bias*

The relationship between media bias and disinformation is structurally generative. Biased reporting habituates audiences to a particular version of reality and lowers the threshold of what counts as plausible, so that fabricated claims aligned with the prevailing frame are absorbed with minimal resistance. Disinformation campaigns, in turn, typically succeed by exploiting existing biases rather than overturning them: false content is tailored to the lexical and narrative conventions of outlets the audience already trusts, inheriting the perceived credibility of the biased media system in which it circulates. Detecting media bias systematically is therefore a prerequisite for understanding and countering disinformation, not a separate concern.

A key way in which media bias intersects with disinformation is through the selection of topics that are covered or omitted by media outlets. This is often referred to as "selection bias" or "gatekeeping". Media organizations, whether consciously or unconsciously, make choices about which stories to cover and which to ignore, influenced by their editorial stance,

target audience, and commercial considerations. This selection process can lead to a skewed representation of reality, where certain issues are amplified while others are downplayed or completely ignored. Disinformation actors exploit this bias by pushing narratives that align with the biases of certain media outlets, ensuring that their falsehoods receive widespread coverage [51].

[51] Eraslan and Özertürk, 2018, *Information gatekeeping and media bias*

Another critical aspect of media bias is "framing", which refers to how information is presented. Framing involves not just the content of a story, but how it is told: the language, metaphors, and symbols used to convey a particular narrative. The framing of a story can significantly influence how the audience understands and interprets the information presented. For example, framing an economic policy debate in terms of "tax relief" versus "government spending" can lead to very different public perceptions of the same issue [50]. Disinformation thrives in environments where framing is biased, as it allows false or misleading narratives to be presented in ways that resonate with the audience's pre-existing beliefs and values.

[50] Entman, 2007, "Framing bias: Media in the distribution of power"

Language and tone also play a crucial role in media bias. The choice of words used in reporting can subtly influence the audience's perception of events and issues. For instance, describing a protest as a "riot" rather than a "demonstration" conveys a different level of severity and legitimacy. Disinformation often relies on such linguistic biases to sway public opinion, making careful analysis of language a critical component of identifying and countering biased reporting [73].

[73] Hamborg et al., 2019, "Illegal aliens or undocumented immigrants? Towards the automated identification of bias by word choice and labeling"

Media bias also affects public trust in different sources of information. When media outlets are perceived as biased, their credibility is called into question, which can lead to general skepticism toward all media sources. This erosion of trust creates fertile ground for disinformation, as audiences may turn to alternative sources of information that are less rigorous in their standards of accuracy and more aligned with their ideological preferences. In such an environment, disinformation can spread more easily, as the lines between credible journalism and propaganda become increasingly blurred [81].

[81] Hughes and Lawson, 2004, "Propaganda and crony capitalism: Partisan bias in Mexican television news"

To fully understand the role of media bias in the spread of disinformation, it is essential to consider the broader media ecosystem. Traditional media outlets, such as newspapers and television networks, operate within a commercial framework where profitability often depends on attracting and retaining an audience. This can lead to sensationalist reporting or the overrepresentation of certain viewpoints to appeal to particular demographics [167]. Meanwhile, digital platforms like social media sites and search engines use complex algorithms to curate the content users see, often reinforcing existing biases rather than challenging them. The interplay between these traditional and digital media forms creates a dynamic environment where disinformation can thrive if not carefully managed and countered.

[167] Scott, 2021, "You won't believe what's in this paper! Clickbait, relevance and the curiosity gap"

Understanding this dynamic interplay is crucial for grasping the complexities of disinformation, media bias, and their interrelation. These issues, along with the various types of media bias and the ways in which they manifest, will be explored in greater detail in Chapter 3.

1.2 THE LIMITS OF CURRENT MEDIA BIAS DETECTION APPROACHES

Manually discerning bias within the vast sea of information published every day is not only impractical but virtually impossible due to the sheer volume, velocity, and variety of contemporary news production. This overwhelming influx of data justifies the need for automated tools capable of detecting media bias effectively and efficiently. Such systems are essential for ensuring that biased reporting does not go unchecked, thus supporting media literacy, democratic accountability, and epistemic integrity in the public sphere.

Substantial research has been conducted to develop automated systems for media bias detection, leveraging the growing capabilities of Natural Language Processing (NLP), Machine Learning (ML), and data mining. However, most existing approaches rely predominantly on lexical or shallow grammatical features such as surface-level n-grams, sentiment lexicons, or TF-IDF representations; without modeling the richer linguistic, rhetorical, and propositional structures that underlie biased discourse. As a result, these systems capture correlations but struggle to represent the pragmatic, inferential, and context-dependent mechanisms through which bias is produced and perceived. Despite technical progress, the field continues to face

significant conceptual, methodological, and epistemological limitations. As analyzed in detail in Chapter 2, a prevailing issue in the current landscape is the oversimplification of the bias phenomenon. Media bias emerges through a combination of subtle rhetorical cues, pragmatic implicatures, selection and omission strategies, and culturally embedded presuppositions, none of which are easily reducible to binary labels. Yet most systems still adopt binary or coarse-grained categorization schemes, collapsing this multidimensional nature into simplistic outputs. They also tend to ignore contextual nuance and the layered interaction between authorial intention, discourse structure, and reader interpretation.

Furthermore, the overwhelming majority of datasets and benchmarks are built on English-language material (often centered on U.S.A. political dynamics) limiting cross-linguistic generalizability and failing to capture the local cultural, ideological, and journalistic conventions that shape Spanish-language media.

Crucially, current approaches typically disregard the perspectivist dimension of media bias perception. Both the production and interpretation of biased content are mediated by cognitive heuristics, ideological priors, sociocultural background, and individual expectations about journalistic norms. Consequently, even highly sophisticated models often fail to explain why a given audience perceives a specific passage as biased or neutral. This absence of ideological and contextual sensitivity undermines both explainability and trustworthiness, particularly in real-world settings where disagreement among annotators is not noise but a meaningful signal of plural epistemic positions.

To address these challenges, this thesis proposes a comprehensive and interdisciplinary reframing of the media bias detection task. It draws on insights from communication theory, linguistics, cognitive psychology, philosophy of knowledge, and critical discourse analysis to first derive an explicit conceptualisation of media bias as a multi-level, cognitively mediated, and perspectival phenomenon, and then translate that conceptualisation into a taxonomic, perspectivist, and hierarchical framework. This framework is instantiated in several key contributions:

First, we propose a fine-grained taxonomy of media bias manifestations, structured with distinct rhetorical and lexical strategies. Each manifestation is later operationalized into heuristics and used to train sentence-level detectors, providing an interpretable foundation for linguistic analysis and model explainability. This approach also highlights a key limitation of current systems: their tendency to rely on surface-level lexical or technical patterns rather than on deeper semantic or rhetorical structures, which ultimately restricts their generalizability for media bias detection.

Second, we introduce MBBMD (Media Bias Bias-Mitigated Dataset), a novel corpus of Spanish-language news articles annotated at three levels: sentence, document, and category. This dataset incorporates a perspectivist annotation protocol that accounts for ideological self-identification of annotators, enabling the modeling of subjective disagreement as a feature rather than an artifact. The dataset also includes a regression-based encoding of intersubjective agreement, expanding the range of evaluable outputs beyond binary classification.

Third, we design and evaluate a hierarchical modeling architecture that, based on the newly proposed conceptualisation of media bias, integrates the aforementioned components. This system fuses sentence-level predictions with document-level ideology-aware classifiers and meta-learners based on AutoML ensembles. Experimental results across five benchmark datasets that span different domains, countries, and degrees of polarization, demonstrating improved generalization and competitive performance under cross-dataset evaluation, a more challenging and realistic metric than traditional in-domain testing.

Finally, this thesis contributes to the field of trustworthy AI by emphasizing explainability, perspective diversity, and methodological transparency. Rather than masking bias through automated correction, it advocates for systems that expose and characterize the manifestations of partiality in ways that empower human interpretation and foster critical reading.

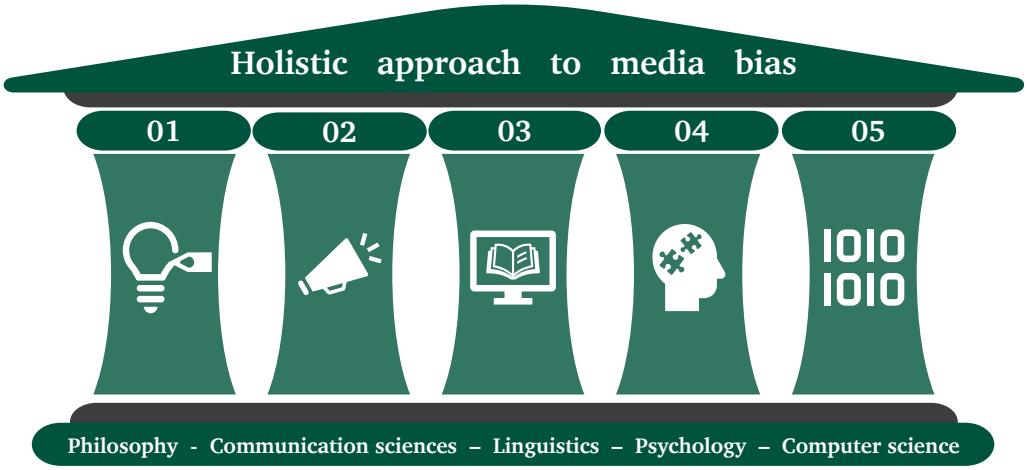
The broader argument running through these contributions is that approaching media bias from a narrow engineering angle, as a classification problem over well-defined labels, produces systems that score well on fixed benchmarks but collapse the moment the input distribution shifts. Widening the lens to include the conceptual, linguistic, psychological, and

epistemic dimensions of the phenomenon is not a philosophical indulgence: it is the path to genuine generalisation. A system that is built on a taxonomy derived from linguistic theory, trained on data that preserves the perspectival nature of bias perception, and evaluated under protocols that expose rather than hide cross-dataset behaviour will, by construction, pick up on properties of media bias that are stable across outlets, topics, and languages, while a narrower system will only pick up on properties of whichever benchmark it was trained on. The empirical chapters of this thesis show exactly that pattern: the broader, interdisciplinary framing is not an aesthetic preference, it is the cause of the robustness gains observed in cross-dataset evaluation, and it is what justifies the effort of taxonomy construction, perspectivist annotation, and hierarchical system design over the simpler (but fragile) alternative of training a flat classifier on a larger dataset.

1.3 Media Bias through an interdisciplinary view

Media bias is not a purely computational problem, and it cannot be addressed as if it were. The very question of what counts as biased presupposes prior positions on truth, subjectivity, and discourse that Computer Science alone is ill-equipped to answer. This thesis therefore draws on five disciplines (philosophy, psychology, communication sciences, linguistics, and computational sciences) not as a tribute to interdisciplinarity for its own sake, but because each one contributes a concept, a method, or a warning that shapes concrete design decisions in the chapters that follow. This section is, in effect, a map of the conceptual commitments that underpin the taxonomy, dataset, and systems presented later.

Figure 1.1: Multidisciplinary foundations of the thesis for a holistic analysis of media bias.



Three cross-cutting insights emerge from this map and structure the rest of the thesis: (1) media bias is linguistically encoded and observable at fine granularity, as shown by linguistics, discourse analysis, and communication theory; (2) media bias perception is cognitively and ideologically mediated, as demonstrated by psychology and philosophy; and (3) robust detection therefore requires integrating multiple layers of representation and multiple perspectives, rather than committing to a single "correct" label. Each subsection below closes with a forward reference to where its insights are operationalised.

1.3.1 *Philosophy: truth, ethics, and perspectivism in the age of disinformation*

The philosophical examination of media bias and disinformation necessitates a deep dive into some of the most enduring questions in philosophy: the nature of truth, the ethics of communication, and the epistemological challenges of a "post-truth" era [188, 57]. Disinformation, by design, seeks to obscure the truth, manipulate perceptions, and create a fragmented reality where multiple conflicting "truths" coexist. To fully understand the impact of media bias on this landscape, it is essential to explore how these philosophical concepts interact with and are disrupted by the spread of slanted and false information. Four philosophical dimensions are load-bearing for the rest of this thesis: the classical notion of truth, the ethical

[188] Tischauser and Benn, 2019, "Whose post-truth era? Confronting the epistemological challenges of teaching journalism"

[57] Friedman, 2023, *Post-truth and the epistemological crisis*

responsibility of communication, perspectivism as an account of situated knowledge, and the dialectical structure of public discourse.

▸ Truth and disinformation: a philosophical dilemma

Truth has been a central concern of philosophy since its inception [4]. Traditionally, truth is understood as a correspondence between a statement and reality, a concept that has been rigorously debated by philosophers from Plato to contemporary thinkers. In the context of disinformation, this classical understanding of truth is directly challenged. Disinformation operates by creating statements that are intentionally disconnected from reality, yet presented as truth. This deliberate distortion blurs the boundary between the real and the fabricated, creating an epistemic environment in which multiple, contradictory "truths" coexist.

[4] Allen, 1993, *Truth in philosophy*

The rise of "alternative facts" and the increasing prevalence of echo chambers exacerbate this issue. Truth becomes relative [175], contingent upon the narrative that one subscribes to, rather than an objective reality that can be empirically verified. This shift has profound implications for public discourse, as it erodes the possibility of reaching a shared understanding of facts, which is essential for the functioning of any democratic society. Without a stable foundation of truth, meaningful dialogue and rational consensus become increasingly difficult, leading to societal fragmentation and the further entrenchment of disinformation.

[175] Solomon, 1977, "Kierkegaard and" Subjective Truth""

Moreover, the challenge of disinformation is not only about discerning between truth and falsehood but also about how truth is framed. Disinformation often intertwines fragments of truth with distortions, making it difficult for individuals to fully disentangle what is accurate from what is misleading. As a result, disinformation undermines the very foundations of knowledge by eroding public trust in reliable sources of information, destabilising the epistemic structures that sustain informed societies [110]. The methodological consequence for this thesis is that media bias cannot be reduced to a verification problem over isolated claims: it is embedded in how information is framed, contextualised, and trusted, and this is precisely what the taxonomy developed in Chapter 3 seeks to capture.

[110] Mayo, 2024, "Trust or distrust? Neither! The right mindset for confronting disinformation"

▸ Ethics and the responsibility of communication

Closely linked to the concept of truth is the ethical dimension of disinformation [197]. Philosophers have long debated the ethics of communication, particularly the responsibility of those who disseminate information to ensure that it is truthful and not misleading. In the age of disinformation, this ethical responsibility is frequently violated, as actors deliberately spread false or slanted information for personal, political, or economic gain.

[197] Waisbord, 2018, "Truth is what happens to news: On journalism, fake news, and post-truth"

From an ethical standpoint, disinformation represents a profound violation of the communicative norms that underpin social trust and cooperation. Jürgen Habermas, a key figure in contemporary philosophy, argues that communication is grounded in the principle of mutual understanding [58], where participants in discourse aim to reach a consensus based on shared truths. Disinformation disrupts this process by introducing falsehoods designed to deceive, eroding the trust necessary for meaningful dialogue and collective decision-making.

[58] Fultner, 2014, *Jurgen Habermas: key concepts*

The ethical implications of disinformation, moreover, extend beyond the responsibilities of those who produce it to include individuals as consumers and disseminators of information. In a digital age where information is abundant and easily accessible, the ethical burden shifts to individuals as well. Consumers must critically evaluate the information they encounter and take responsibility for the content they share [112]. This consideration is crucial for developing a resilient public sphere that can withstand the corrosive effects of disinformation. The ethics of communication, therefore, must evolve to include the digital responsibilities of individuals and institutions in an information-saturated world: transparency, media literacy, and ethical dissemination must be central to mitigating the harmful effects of disinformation.

[112] McIntyre, 2023, *On disinformation: How to fight for truth and protect democracy*

▸ Perspectivism: the multiplicity of truths and the challenge of disinformation

The concept of perspectivism, introduced by Friedrich Nietzsche [173], complicates the philosophical understanding of disinformation. Perspectivism posits that all knowledge is inherently situated and viewed from a particular perspective, rejecting the notion of an absolute and objective truth. According to Nietzsche, there is no "view from nowhere"; instead, there

[173] Small, 1983, "Nietzsche and a Platonist tradition of the cosmos: Center everywhere and circumference nowhere"

are multiple truths, each reflecting the standpoint of the observer. Perspectivism therefore offers a philosophically honest account of why two informed, well-intentioned readers can disagree about whether the same article is biased, and why collapsing that disagreement into a single ground-truth label is itself an ideological move.

While perspectivism offers a more nuanced understanding of truth by acknowledging the diversity of perspectives that shape human knowledge, it also presents challenges in the context of disinformation. Disinformation exploits perspectivism by promoting the idea that all perspectives are equally valid, even when they are based on falsehoods. This relativistic view can lead to a situation where disinformation is accepted as just another perspective, making it more difficult to challenge or refute.

In the digital age, this issue is amplified by algorithmic curation on social media platforms, where users are presented with content that aligns with their pre-existing beliefs. The challenge, then, is to develop a form of *critical perspectivism*, one that recognises the validity of diverse perspectives while maintaining a commitment to truth and factual accuracy [45]. Critical perspectivism must differentiate between legitimate differences in viewpoint and the deliberate manipulation of information intended to deceive. This is the philosophical stance adopted throughout this thesis: media bias annotation is treated as perspectival but not relativistic, and annotator disagreement is modelled as a meaningful signal rather than noise (Section 3.2).

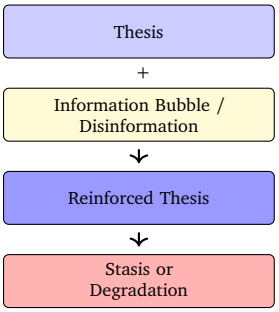
[45] D'Olimpio, 2021, "Critical perspectivism: Educating for a moral response to media"

▸ The Hegelian dialectic: disinformation as a disruptor of progress

The philosophical challenges posed by disinformation can also be understood through the lens of Hegelian dialectics [133]. Hegel's dialectic process is a model of intellectual and societal progress through the resolution of contradictions. Typically, the process unfolds in three stages: the thesis, representing an initial idea or condition; the antithesis, which opposes or contradicts the thesis; and the synthesis, which resolves the conflict between the thesis and antithesis, leading to a new understanding that transcends both.

[133] Pinkard, 1988, "Hegel's dialectic: The explanation of possibility"

FIGURE 1.2: Disruption of the Hegelian dialectic.



To clarify the relevance of this framework for the study of disinformation, consider a simplified example. A public debate about climate change may begin with a thesis grounded in scientific consensus. A legitimate antithesis could raise questions about policy feasibility or socioeconomic trade-offs. However, when disinformation enters the debate (for instance, by promoting fabricated claims that climate change is a hoax) the antithesis becomes illegitimate. This prevents a meaningful synthesis, as the apparent conflict is based on false premises rather than substantive disagreement. Similar patterns can be observed in debates about public health, migration, or electoral integrity, where disinformation replaces genuine opposition with manufactured contradictions.

In a healthy dialectical process, the conflict between thesis and antithesis results in a synthesis that advances knowledge and societal development. However, in an environment saturated with disinformation, this process is severely disrupted. Disinformation prevents the emergence of a legitimate antithesis by presenting false or distorted information as valid opposition. This creates a scenario where synthesis is either impossible or fundamentally flawed, as it is based on a distorted understanding of the issues at hand. When disinformation infiltrates public discourse, the available knowledge base becomes skewed, leading to a synthesis that reflects misinformation rather than a well-rounded, truth-seeking resolution. As a result, disinformation not only hinders the resolution of existing conflicts but perpetuates new ones, as the flawed synthesis becomes the new thesis in subsequent debates. This cyclical degradation of discourse undermines democratic processes, where informed debate is crucial for the development of policies that reflect the collective will and the common good.

This disruption of the Hegelian dialectic has parallels with other dialectical models, such as the materialist dialectic [12] in Marxist theory or the serial dialectic in existentialism [78]. In both cases, disinformation introduces distortions into the fundamental conflicts that drive progress, whether in terms of class struggle or individual existential choices. By disrupting these dialectical processes, disinformation becomes a significant barrier to both intellectual and societal progress, entrenching existing divisions and obstructing pathways toward genuine resolution. This dynamic motivates the distinction (central to Chapter 3) between interpreta-

[12] Badiou, 2005, "Democratic materialism and the materialist dialectic"

[78] Howard and Howard, 2019, "The Rationality of the Dialectic: Jean-Paul Sartre"

tive journalism, which can legitimately contest a thesis, and biased reporting, which forecloses the dialectic by distorting the factual basis for disagreement.

1.3.2 *Psychology: cognitive biases in the production and reception of media*

The study of media bias is incomplete without considering the psychological factors that influence how information is produced, disseminated, and consumed. Human cognitive biases play a pivotal role in shaping both the creation and the reception of media content, and they are deeply rooted in the cognitive processes of everyone involved in the communicative chain. They affect not only the journalists and editors who create and select news stories, but also the audiences who interpret and circulate them. Acknowledging this bidirectionality is what forces us to treat media bias as a phenomenon of production *and* reception, rather than an intrinsic property of a text in isolation.

▸ **Cognitive biases in media production**

Journalists and media organisations are not immune to the cognitive biases that affect all humans. The availability heuristic [132], the tendency to overweight information that is more recent, memorable, or emotionally vivid, can lead journalists to prioritise sensational or emotionally charged stories, skewing the representation of events. This can result in news coverage that amplifies dramatic incidents while downplaying more significant but less sensational issues. The bandwagon effect [196] operates when individuals (journalists included) are influenced by the popularity of a particular idea or trend: outlets tend to cover stories that already receive widespread attention rather than those that may be more important but less visible, creating feedback loops where certain narratives dominate not because of their intrinsic value but because they gained initial traction and are continuously amplified.

[132] Perry, 2002, "Guilt by saturation: Media liability for third-party violence and the availability heuristic"

[196] Waddell and Sundar, 2020, "Bandwagon effects in social television: How audience metrics related to size and opinion affect the enjoyment of digital media"

Status-quo bias [23], the preference for things to remain as they are, also affects how news is reported. Media organisations may be inclined to present information in ways that conform to existing societal norms and expectations rather than challenging the status quo. This can lead to a conservative approach in news coverage, in which potentially disruptive or innovative ideas are underrepresented or dismissed, reinforcing existing power structures and societal inequalities. These production-side biases can skew representation even without any deliberate intent to mislead, which is precisely why media bias cannot be equated with disinformation: unintentional bias is still bias.

[23] Bolsen et al., 2014, "How frames can undermine support for scientific adaptations: Politicization and the status-quo bias"

▸ **Cognitive biases in media consumption**

On the audience side, cognitive biases such as confirmation bias and selective exposure significantly impact how news is consumed and interpreted [41]. Individuals tend to seek out information that confirms their existing beliefs and are more likely to trust sources that align with their ideological views. This selective exposure leads to the formation of "echo chambers", where individuals are repeatedly exposed to similar viewpoints, reinforcing their beliefs and making them more resistant to opposing information.

[41] Dahlgren, 2020, "Media echo chambers: Selective exposure and confirmation bias in media use, and its consequences for political polarization"

Naive realism further affects how audiences perceive news content: people often believe that their own interpretation of a news story is the "correct" one and that those who disagree are misinformed or biased [120]. This perception can lead to polarisation, as individuals become more entrenched in their views and less willing to consider alternative perspectives. Socio-cultural factors such as political affiliation, education level, and social environment further shape how individuals process news [91]: a person who identifies strongly with a particular political ideology may interpret coverage through the lens of that ideology, leading to a systematically biased understanding of the information presented. This socio-cultural filtering can exacerbate divisions within society, as different groups consume and interpret the same news in vastly different ways.

[120] Nasie et al., 2014, "Overcoming the barrier of narrative adherence in conflicts through awareness of the psychological bias of naïve realism"

[91] Kline, 2006, "A decade of research on health content in the media: the focus on health challenges and sociocultural context and attendant informational and ideological problems"

The methodological consequence for this thesis is direct: if perception is structurally perspectival, then treating annotator disagreement as noise to be eliminated is conceptually wrong. The MBBMD dataset (Section 3.2) is designed accordingly: ideological self-identification is collected for every annotator, disagreement is preserved, and aggregation is handled through perspectivist rather than majority-vote protocols.

1.3.3 *Communication sciences: media systems, bias, and public perception*

While philosophy provides the conceptual groundwork and psychology explains individual cognitive mechanisms, communication sciences offer the meso-level account of how media systems produce and distribute bias in practice. Media bias, agenda-setting, framing, and the broader role of digital platforms are the central concepts of this field, and each of them helps explain how biased and disinforming content is disseminated, interpreted, and legitimised by audiences.

When this thesis began, the working assumption was that the ideal approach to media bias was to pursue journalistic objectivity and measure departures from it, aiming at a neutral ground where facts could be presented without slant. As the literature of communication sciences quickly complicates this picture, that intuition had to be abandoned. Modern journalism, particularly in its more critical and reflective forms, does not necessarily seek pure objectivity. Instead, it acknowledges that striving for objectivity can inadvertently reinforce the status quo by presenting all perspectives as equally valid, even when some are rooted in disinformation or manipulation [198, 138]. Contemporary journalism often seeks a more engaged and committed approach: rather than merely reporting facts, it aims to challenge power structures, question dominant narratives, and give voice to marginalised perspectives. This form of journalism recognises the inherent biases in any form of communication and seeks to mitigate their impact by being transparent about them and by actively engaging in the pursuit of truth and justice.

[198] Wallace, 2019, *The view from somewhere: Undoing the myth of journalistic objectivity*

[138] Raeijmaekers and Maeseele, 2017, "In objectivity we trust? Pluralism, consensus, and ideology in journalism studies"

Importantly, acknowledging that journalism may adopt interpretative or opinionated stances is not equivalent to endorsing media bias. Critical journalism uses opinion to contextualise facts, challenge dominant narratives, and expose power asymmetries, whereas media bias involves systematically distorting information to favour a particular ideological position. Clarifying this distinction is essential: opinionated journalism can coexist with rigorous factual reporting, whereas biased journalism undermines the conditions for informed public discourse. This differentiation also helps explain why disinformation thrives: it exploits biased framings, not legitimate interpretative work. The distinction is load-bearing for the taxonomy developed in Chapter 3, which deliberately separates interpretative practice from biased framing.

Media bias, in particular, plays a significant role in shaping the environment in which disinformation thrives. The ways in which news is selected, framed, and presented can reinforce existing biases and contribute to the spread of disinformation. This is especially true in the digital age, where algorithms on social media platforms create echo chambers that limit exposure to diverse viewpoints, thereby amplifying biased information [187]. Agenda-setting theory [111] further explains how media can influence public perception by prioritising certain topics over others: disinformation can exploit this by pushing false narratives that align with the media's agenda, making them appear more credible or relevant than they truly are. Framing, in turn, affects how information is interpreted by the audience, often through the use of specific language or imagery that evokes emotional responses and steers readers toward particular conclusions without ever producing a demonstrably false statement.

[187] Terren and Borge, 2021, "Echo chambers on social media: A systematic review of the literature"

[111] McCombs and Valenzuela, 2007, "The agenda-setting theory"

Finally, the role of digital platforms deserves explicit attention. Social media, with its capacity to disseminate information rapidly and widely, plays a dual role as both a vector for disinformation and a potential tool for its mitigation. Understanding these dynamics is essential for developing strategies that counteract the influence of disinformation and restore the integrity of public discourse.

These concepts directly inform the media bias taxonomy developed in Chapter 3, where media bias types are organised according to communicative intention and contextual factors, and manifestations are defined at the linguistic and discursive levels. Agenda-setting and framing are precisely the kind of covert mechanisms that document-level, truth-vs-falsity benchmarks fail to capture, and they are the phenomena that the taxonomy is explicitly designed to address, organising bias around communicative intention and framing strategy rather than around veracity alone.

1.3.4 *Linguistics: uncovering linguistic markers in news discourse*

Linguistics provides the analytical toolkit for connecting the abstract accounts offered by philosophy and communication sciences to observable features of text. Language is not a neutral conduit: it encodes ideological stance through lexical choice, syntactic structure, semantic presupposition, and pragmatic inference. For a thesis whose goal is to *detect* media bias in actual news articles, linguistics is not ornamental but constitutive: it supplies the observable signals on which any computational system must ultimately rely.

▸ **Lexical choices and ideological positioning**

The first layer of linguistic bias is lexical. Vocabulary selected by journalists and editors is rarely neutral; it frequently encodes evaluative meaning and ideological stance, sometimes overtly and sometimes through subtle connotation, as extensively documented in critical linguistics and discourse analysis [191]. Consider the difference between labelling a group as "freedom fighters" versus "terrorists", or describing a protest as "peaceful" versus "chaotic". Each term carries moral weight and aligns the reader with a specific interpretation of reality. These lexical patterns are not anomalies but structural features of how language operates in the press [20], especially around polarising topics where word choice influences not only what readers think but how they feel.

[191] Van Dijk, 2013, "Ideology and discourse analysis"

[20] Bell, 1991, *The language of news media*

▸ **Syntactic and semantic strategies for framing**

Beyond the lexicon, syntax shapes perception by distributing agency. Passive voice can de-emphasise actors responsible for controversial actions ("shots were fired" versus "police fired shots"); nominalisations can dissolve agency altogether ("the destruction of the town") [56]. These choices are not randomly distributed: as van [193] shows, passivisation and agent defocusing correlate systematically with the ideological stance of the outlet, especially in politically sensitive coverage.

[56] Fowler, 2013, *Language in the News: Discourse and Ideology in the Press*

[193] van, 2014 *Discourse and knowledge: a sociocognitive approach*

Semantically, presuppositions and implicatures encode bias below the surface of explicit statements. The sentence "even John Doe admits that climate change is real" presupposes prior denial and, in doing so, silently attributes a belief to John Doe. Implicatures, meanwhile, guide readers toward inferences the author never explicitly makes, insulating the outlet from accountability while steering interpretation. This is precisely the kind of bias that surface-level keyword matching misses.

▸ **Pragmatic cues and discourse-level manipulation**

At the pragmatic level, media texts exploit genre conventions and discourse structure. The inverted pyramid places the most salient information first, framing interpretation from the outset. Critical discourse analysis [193] examines how the sequencing of topics, the choice of quoted sources, and the selective presence or absence of information constructs ideological worlds over time. Topicalisation and foregrounding [70] grant higher cognitive salience to what is placed first, influencing perceived relevance. Hedges, evaluative adverbs, and modality markers ("allegedly", "clearly", "supposedly") introduce calibrated doubt or certainty; these markers are often overlooked, yet they encode bias *without* making any false claim.

[70] Halliday and Hasan, 2014, *Cohesion in english*

▸ **Linguistic ideology and stylistic norms in journalism**

Finally, bias is reinforced by stylistic and editorial norms that present themselves as apolitical. Newsworthiness criteria prioritise conflict, novelty, and proximity [20], skewing coverage toward sensationalism. Scare quotes ("the 'so-called' reforms") cast doubt on terminology; the omission of quotation marks can present opinion as fact. These apparently minor editorial decisions have outsized consequences for interpretation.

The linguistic framework outlined in this subsection directly informs two contributions of the thesis. First, the taxonomy in Chapter 3 is structured around the linguistic and rhetorical manifestations of bias (lexical, syntactic, semantic, pragmatic, discursive) rather than around coarse outlet-level labels. Second, the heuristic lexicons and patterns compiled

in Appendix C operationalise these linguistic phenomena into machine-checkable signals used by the hierarchical system of Chapter 4.

1.3.5 *Computational sciences: from information theory to NLP-based disinformation detection*

[169] Shannon, 1948, "A mathematical theory of communication"

The computational study of (dis)information traces back to Shannon's *A Mathematical Theory of Communication* [169], which revolutionised our understanding of communication systems by providing a mathematical framework for quantifying information. Central to this theory is the concept of entropy, which measures the uncertainty or expected value of the information contained within a message:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

Shannon's framework is powerful for optimising the encoding, transmission, and storage of information, but it was never designed to reason about adversarial communication. Disinformation is not merely the transmission of uncertain data: it is the intentional distortion of data with the aim of misleading. A consistently biased or deceptive outlet may produce communications with low entropy, a false sense of certainty, while systematically misrepresenting reality. This limitation of the classical framework is what motivated the development of NLP-based approaches specifically targeting media bias, misinformation, and propaganda.

Within NLP, this research spans tasks such as fake-news classification, stance detection, claim and evidence detection, fact-checking, propaganda-technique detection, and media bias detection. Early supervised approaches relied on hand-engineered features (lexical n-grams, sentiment scores, readability indices, stylistic markers, simple metadata) fed into classical classifiers such as logistic regression, SVMs, or random forests. These models established important baselines but captured little about the semantic, pragmatic, and contextual aspects of bias. A second wave introduced deep learning models based on CNNs and RNNs (including LSTMs and GRUs), which learned distributed representations directly from data and captured local and sequential dependencies. Parallel work incorporated social-context signals and propagation graphs to distinguish organic from coordinated campaigns.

The current state of the art is dominated by transformer-based architectures. Pre-trained encoder-only models such as BERT and its variants are widely fine-tuned for bias classification, claim verification, and propaganda detection. Decoder-only LLMs are increasingly used for explainable fact-checking, justification generation, and the production of counter-narratives. Multilingual transformers are particularly relevant for media bias, since they allow models trained on high-resource languages to be adapted to lower-resource ecosystems, including the Spanish-language media that this thesis specifically targets. A detailed review of these approaches, together with their empirical limits, is the subject of Chapter 2.

Despite considerable progress, important limitations persist. Most systems rely on datasets that are temporally, geographically, or politically narrow, raising concerns about generalisation and robustness. Domain shift, topic evolution, and adversarial adaptation by bad actors degrade performance in real-world settings. And many benchmarks still default to binary labels ("true/false", "biased/unbiased"), overlooking the graded, perspectival, and ideologically situated nature of media content established by all the disciplines surveyed above. Building upon the perspectivist stance adopted earlier, this thesis pays particular attention to the Learning With Disagreements (LeWiDi) framework [190], which explicitly incorporates annotator disagreement into model training and treats conflicting perspectives as informative rather than noisy. LeWiDi directly shapes the annotation protocol of MBBMD (Section 3.2) and the aggregation strategy of the hierarchical perspectivist system (Chapter 4), closing the loop between the philosophical commitments stated at the beginning of this section and the concrete computational design decisions that follow.

[190] Uma et al., 2021, "Learning from disagreement: A survey"

Read together, the five disciplinary perspectives surveyed above expose a coherent pattern of limitations in how the computational community has approached media bias to date. Five limitations stand out, and each one is explicitly addressed by a specific methodological choice later in this thesis.

1. The literature rests on a narrow conceptualisation of media bias.

Existing computational work typically treats media bias as a property of outlets or documents that can be captured by a single label, ignoring the communicative, linguistic, and cognitive dimensions surveyed above. Systematic reviews of the field [149, 72] repeatedly note that no two studies share a definition, and that most operationalisations collapse heterogeneous mechanisms (lexical evaluation, framing, omission, selective sourcing) into a single axis. The taxonomy of manifestations developed in Chapter 3 is an explicit response to this narrowness: it decomposes media bias into seventeen observable mechanisms grounded in linguistic, psychological, and communicative theory.

[149] Rodrigo-Ginés et al., 2023, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it"

[72] Hamborg et al., 2019, "Automated identification of media bias in news articles: an interdisciplinary literature review"

2. Resources are geographically and linguistically constrained.

The datasets on which the field reports progress are overwhelmingly English-language, U.S.A.-centric, and focused on a handful of polarising domestic topics (presidential elections, specific conflicts). Non-English resources (e.g., for Hindi [1], Chinese [106], or multilingual social media via FIGNEWS [205]) remain small and isolated. As a result, "state of the art" in media bias detection is in practice state of the art in detecting U.S.A. political framing in English. MBBMD (Section 3.2) responds to this limitation by providing a Spanish-language dataset covering ten topics drawn from Spanish, Irish, Argentine, and Israeli–Palestinian media ecosystems.

[1] Agrawal et al., 2022, "Towards detecting political bias in hindi news articles"

[106] Lin et al., 2025, "Data-augmented and retrieval-augmented context enrichment in chinese media bias detection"

[205] Zaghouani et al., 2024, "The FIGNEWS Shared Task on News Media Narratives"

3. Detection is reduced to binary classification.

Most systems treat media bias as a present/absent binary, or as a small-cardinality categorical label inherited from outlet ratings (e.g., [36, 27, 76]). This formulation is incompatible with the graded, context-dependent nature of bias perception established by the psychological and philosophical literature. Chapter 3 and Chapter 4 therefore adopt regression-based LeWiDi formulations alongside classification, preserving the distribution of annotator judgments rather than collapsing it into a single point estimate.

[36] Chen et al., 2018, "Learning to flip the bias of news headlines"

[27] Budak et al., 2016, "Fair and balanced? Quantifying media bias through crowdsourced content analysis"

[76] Horne et al., 2018, "Sampling the news producers: A large news and feature data set for the study of the complex media landscape"

4. Perspectival variation is systematically suppressed.

Virtually every benchmark aggregates annotator labels through majority voting and then treats the result as ground truth, erasing the ideological and cognitive variation that philosophy (perspectivism) and political psychology (motivated reasoning, naive realism) identify as constitutive of bias perception [120, 41]. The few exceptions, notably the LeWiDi paradigm [190, 100], remain marginal in the media bias literature specifically. MBBMD directly adopts a perspectivist annotation protocol that preserves individual judgments and ideological self-identification per annotator, and the proposed system integrates ideology-aware classifiers at inference time.

[120] Nasie et al., 2014, "Overcoming the barrier of narrative adherence in conflicts through awareness of the psychological bias of naïve realism"

[41] Dahlgren, 2020, "Media echo chambers: Selective exposure and confirmation bias in media use, and its consequences for political polarization"

[100] Leonardelli et al., 2023, "SemEval-2023 task 11: Learning with disagreements (LeWiDi)"

5. The problem is modelled flatly, without hierarchy or manifestations.

Even when datasets are rich, the dominant modelling paradigm maps the whole document directly to a label, losing the multi-level structure of bias (sentence-level cues, document-level framing, category-level editorial dimensions). End-to-end transformers [53, 180, 95] and even LLM-based approaches [200, 115] treat the document as an undifferentiated input. The hierarchical and perspective-aware framework introduced in Chapter 4 replaces this flat formulation with an architecture that first detects manifestations at the sentence level, then aggregates them into document-level and category-level predictions, and finally integrates perspectivist signals through an ensemble meta-learner.

These five limitations motivate, point by point, the three contributions of this thesis (the taxonomy of Chapter 3, the MBBMD dataset of Section 3.2, and the hierarchical perspectivist system of Chapter 4), and they provide the criterion against which those contributions should be evaluated: not whether they score higher on a fixed benchmark, but whether they address the structural gaps that the interdisciplinary reading has surfaced.

[53] Fan et al., 2019, "In Plain Sight: Media Bias Through the Lens of Factual Reporting"

[180] Spinde et al., 2021, "Automated identification of bias inducing words in news articles using linguistic and context-oriented features"

[95] Krieger et al., 2022, "A domain-adaptive pre-training approach for language bias detection in news"

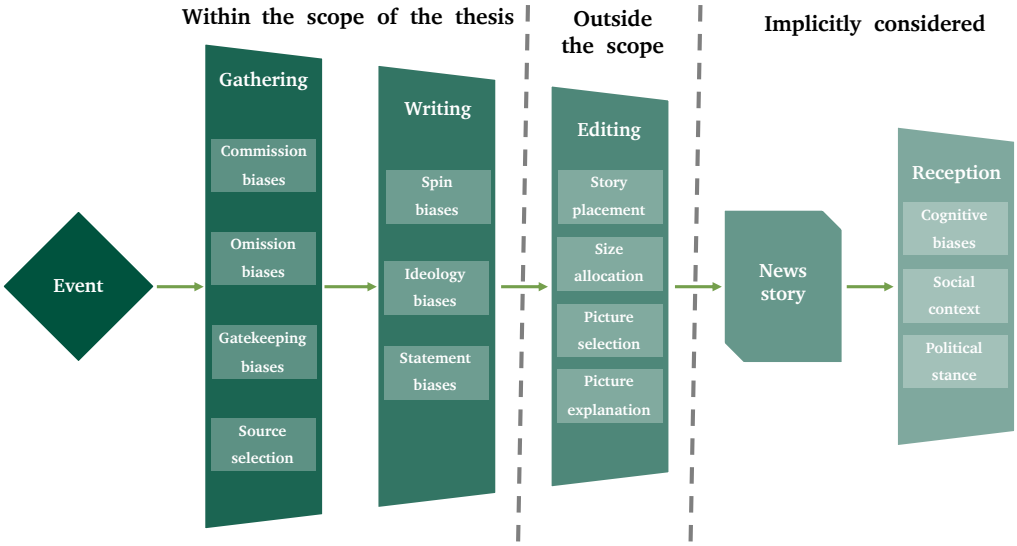
[200] Wessel, 2023, *Improving Media Bias Detection with State-of-the-art Transformers*

[115] Menzner and Leidner, 2024, "Experiments in news bias detection with pre-trained neural transformers"

1.4 RESEARCH QUESTIONS AND HYPOTHESES

As stated in the previous section, the interdisciplinary exploration of the philosophical, psychological, linguistic, communication sciences, and computational foundations of media bias leads to several critical insights that shape the direction of this thesis. First, it becomes clear that media bias cannot be effectively understood or mitigated through approaches that operate solely at the level of documents or outlets. Traditional frameworks, centred around the notion of journalistic objectivity and coarse binary judgments, fail to account for the specific linguistic, rhetorical mechanisms through which media bias is produced. These mechanisms operate at fine granularity, often at the level of sentences, propositions, and discourse structures.

FIGURE 1.3: Stages of the news production and reception pipeline where media bias can emerge.



In parallel, the analysis reveals that bias is not only embedded in the creation of news but is also shaped by how content is interpreted by different audiences. Cognitive biases such as confirmation bias, motivated reasoning, and naive realism strongly influence the reception of media, meaning that bias cannot be conceptualized as an intrinsic property of the text alone. Any comprehensive framework must therefore consider both production and reception, acknowledging that interpretation is perspectival and situated.

To frame this dual nature, Figure 1.3 illustrates the stages of the news production and reception pipeline where media bias may emerge. This model builds upon the typologies and manifestations discussed in Chapter 3. The thesis focuses primarily on the gathering and writing stages, where bias emerges through techniques such as commission and omission, gatekeeping, source selection, evaluative language, ideological framing, and spin. These stages are linguistically encoded and thus amenable to NLP-based analysis.

By contrast, editing decisions (e.g., layout, image selection, story placement) fall outside the modelling scope, as they require multimodal approaches. Similarly, although the reception stage is not directly modelled, it is incorporated through a perspectivist framework that explicitly treats disagreement and subjectivity as meaningful signals rather than noise.

These observations motivate a shift away from monolithic, document-level approaches toward models capable of capturing (i) fine-grained media bias techniques, and (ii) perspectival variation in interpretation. This dual focus aligns directly with the methodological trajectory of this thesis, which combines a taxonomy of linguistic and rhetorical techniques with perspectivist, ideologically informed annotation and hierarchical modelling.

Building upon these insights, the thesis proposes the following central hypothesis:

Main hypothesis

If media bias is a multi-level, cognitively mediated, and perspectivist phenomenon, then detection systems must explicitly model fine-grained media bias techniques and subjectivity in order to achieve robust and generalizable predictions.

This main hypothesis is divided into three sub-hypotheses that reflect the theoretical premises established in the previous sections:

Hypothesis 1 (H1): If bias is instantiated through fine-grained linguistic, rhetorical, and psychological techniques, then sentence-level modelling should outperform document-level approaches for identifying how bias is expressed.

Explanation: Linguistic and rhetorical research shows that media bias is encoded at the sentence and proposition level through presuppositions, implicatures, evaluative predicates, framing devices, and other localised techniques. If these mechanisms are the primary carriers of bias, models that operate at the sentence level should provide more accurate, interpretable, and mechanism-aware predictions than those limited to coarse document-level labels.

Hypothesis 2 (H2): If bias perception is perspectival and shaped by ideological and cognitive priors, then incorporating disagreement and ideological metadata should improve cross-group robustness and generalization.

Explanation: Political psychology and cognitive science show that individuals with different ideological backgrounds often disagree about whether the same content is biased. If disagreement is a systematic and meaningful signal rather than annotation noise, then models that encode ideological self-identification and intersubjective variation should better capture how different audiences perceive bias, enhancing robustness across ideological groups.

Hypothesis 3 (H3): If readers interpret news by integrating cues from multiple representational layers (wording, narrative structure, and ideological context) then computational models must reflect this multi-layered integration to generalize across topics, sources, and audiences.

Explanation: Theoretical models of discourse processing and media effects indicate that interpretation results from the interaction of several layers of representation: fine-grained linguistic choices, global framing structures, and the broader ideological context in which a text is situated. These layers jointly shape how bias is produced, perceived, and attributed. If human interpretation depends on the integration of these cues, analytical models that incorporate multiple layers of representation should offer more accurate and more generalizable accounts of media bias than approaches confined to a single level of analysis.

These three hypotheses, grounded in the interdisciplinary foundations outlined above, define both the analytical perspective and the methodological trajectory of this thesis. Each contribution presented below operationalizes one or more of these hypotheses into concrete resources and systems.

1.5 MAIN CONTRIBUTIONS OF THE THESIS

Grounded in the hypotheses formulated above, this thesis introduces four main contributions that reconceptualise media bias as a technique-driven, perspectivist, and multi-level phenomenon:

1. **A taxonomy of media bias manifestations** (H1): a comprehensive, linguistically grounded framework that systematises the techniques through which bias is expressed at the sentence level, enabling the shift from coarse document-level labels to interpretable, mechanism-oriented analysis (Chapter 3).
2. **MBBMD: a perspectivist and hierarchical dataset** (H2): the first Spanish-language corpus annotated at three hierarchical levels (sentence, document, and bias category) with a perspectivist protocol that preserves annotator disagreement as a meaningful signal (Section 3.2).
3. **Empirical diagnosis of cross-dataset generalization failure**: extensive cross-dataset experiments demonstrating that existing models capture dataset-specific regularities rather than transferable bias signals, motivating the need for hierarchical and perspectivist approaches (Chapter 4).
4. **A hierarchical system for media bias analysis** (H3): a multi-level architecture combining sentence-level heuristics, ideology-aware transformer classifiers, and ensemble meta-learners, achieving cross-dataset F1 improvements from 0.47 (flat baseline on MBBMD) to 0.77 (full system), a +0.30 gain that effectively closes the generalisation gap (Chapter 4).

1.6 THESIS STRUCTURE

The thesis is organised into six chapters and follows a single argumentative line: a critical state of the art identifies five structural problems in the field; a conceptual framework responds to them with a taxonomy and a dataset; an empirical chapter diagnoses the generalisation failure they imply and builds the system that addresses it; a results-and-discussion chapter evaluates that system head-to-head against the diagnosis baseline, decomposes the gain along dataset and system axes, and translates the findings into implications for deployment; and a conclusions chapter consolidates the arc and projects the open ends as concrete future-work directions.

Chapter 2 (*State of the art of media bias detection: foundations, methods, and structural problems*) reviews media bias from both its communicative and its computational sides. It first revisits the conceptual foundations of media bias (definition, distinctive characteristics, and its relationship with the broader information disorder ecosystem) from the perspective of communication sciences, linguistics, and media studies. It then presents a systematic review of automatic media bias detection following PRISMA guidelines, covering 118 studies published between 2000 and 2025, and surveys traditional, deep learning, and transformer-based methods together with the datasets and benchmark resources on which they are built. The chapter closes by synthesising the critical reading of the field into five *structural problems* (conceptual fragmentation, weak generalisation, the single-label fallacy, anglocentrism, and evaluation opacity) that motivate the rest of the thesis.

Chapter 3 (*From diagnosis to operational definition: a hierarchical taxonomy and a perspectivist dataset*) develops the first two contributions of this thesis as a coordinated response to those structural problems. Section 3.1 proposes a hierarchical taxonomy of media bias that organises the phenomenon into bias types (by communicative intention and contextual factors) and 17 concrete bias manifestations (by observable linguistic and rhetorical realisation), and argues explicitly that the taxonomy is a superset of the partial inventories produced by prior work. Section 3.2 then introduces the Media Bias Bias-Mitigated Dataset (MBBMD), a Spanish-language, perspectivist, hierarchical dataset (100 articles, 2,348 adjudicated sentence-level annotations, 12 ideologically diverse annotators, 10 socio-politically salient topics) that operationalises the taxonomy, preserves annotator disagreement as a meaningful signal, and deliberately moves outside the anglocentric benchmark ecosystem that dominates the literature. Taxonomy and dataset are presented together because they function as two sides of the same conceptual move: the taxonomy defines what is being detected, and the dataset instantiates that definition in annotated data.

Chapter 4 (*Closing the cross-dataset gap: a hierarchical perspectivist system for media bias detection*) moves from concepts to empirical work, organised around three questions answered

in three sections. Section 4.1 diagnoses the generalisation failure of flat classifiers under cross-dataset evaluation across five benchmarks (MBIC, the Crisis in Ukraine corpus, FIGNEWS, BEADs, and MBBMD), making the gap between in-domain and cross-dataset F_1 visible under a stringent and reproducible protocol. Section 4.2 then conducts a level-by-level analysis on MBBMD that recovers the dataset's three annotation layers (sentence-level manifestations, document-level perception with preserved disagreement, and document-level categories), comparing heuristic-linguistic, multitask DistilBERT, and decoder-only paradigms across each layer and reporting the level-by-level findings that motivate the architecture. Section 4.3 introduces the proposed hierarchical perspectivist system, a modular cascade that integrates sentence-level heuristic and neural manifestation detectors, document-level encoder-only models, and ideology-aware perspectivist classifiers through a stacking-based decision layer. The evaluation of that system is reported separately in Chapter 5, in line with the conventional thesis structure of separating system design from results and discussion.

Chapter 5 (*Results and discussion*) evaluates the proposed system head-to-head with the diagnosis baseline of Section 4.1, decomposes the gain, and translates the findings into deployment guidance. After fixing the evaluation framework (Section 5.1), the cross-dataset evaluation of Section 5.2 reports a +0.30-point improvement in cross-dataset macro F_1 (from 0.47 for the strongest flat baseline to 0.77 for the proposed system), broken down per held-out test dataset and accompanied by paired bootstrap confidence intervals. Section 5.3 synthesises the layer-by-layer results of Chapter 4 (sentence-level manifestations, document-level categorisation, hierarchical ICM evaluation) and grounds them in three qualitative case studies drawn from the MBBMD test partition. Section 5.4 reports the dataset-versus-system factorial ablation and the component ablation, attributing the headline gain jointly to MBBMD and to the proposed architecture. Section 5.5 then revisits the three research hypotheses of Section 1.4; Section 5.6 states the limitations of the framework with quantified anchors; and Section 5.7 translates the empirical evidence into implications for the real-world deployment of media bias detection systems.

Chapter 6 (*Conclusions and future directions*) consolidates the contributions of the thesis as a coordinated response to the five structural problems of Section 2.3, frames the validated hypotheses at the level of what each adds to that diagnosis, and turns each limitation acknowledged in Section 5.6 into a concrete future-work direction. These directions, ordered from immediate extensions to ambitious long-term goals, include expanding MBBMD towards greater multilingual and contextual diversity, developing debiasing and neutral-rewriting systems grounded in the manifestation-level structure, incorporating multimodal signals (images, layout, audiovisual cues) to cover the eleven manifestations not addressable from text alone, advancing perspective-aware explainability, and modelling the temporal evolution of bias through an Overton-window analysis. Together, these directions point toward a broader research agenda in which media bias analysis is treated as a dynamic, multi-level, and socially grounded phenomenon.

Finally, the appendices provide annotation guidelines for MBBMD (Appendix A), detailed descriptions of corpus composition (Appendix B), and the complete heuristic lexicons and patterns used for sentence-level bias detection (Appendix C).

This page intentionally left blank.

2

State of the art of media bias detection: foundations, methods, and structural problems

"*Science is the knowledge of consequences, and dependence of one fact upon another.*"
—Thomas Hobbes

Understanding what media bias is, and understanding how Computer Science has tried to detect it, are two halves of the same problem. The first half, the conceptual question (what is media bias, how does it manifest in news discourse, how does it differ from disinformation), has been answered primarily from outside Computer Science: by communication studies, linguistics, and media theory. The second half (how do we build models that identify and characterise bias at scale), has developed inside NLP and ML over the last two decades, producing a methodological arc that runs from handcrafted features through RNNs to transformer-based architectures and, most recently, LLMs. This chapter reviews both halves in sequence, and closes by stepping back to articulate the structural problems that, in our view, explain why the field has produced impressive in-distribution numbers but disappointing real-world behaviour. These structural problems (Section 2.3) are the starting point for the three contributions developed in the rest of the thesis.

The chapter is organised as follows. Section 2.1 reviews the conceptual foundations of media bias (its definition, its distinctive characteristics, and its relationship with the broader information disorder ecosystem), drawing on communication sciences, linguistics, and media studies. Section 2.2 reviews the state of the art in computational media bias detection, organised around the methodological evolution of the field (traditional approaches, pre-transformer deep learning, transformer-based models, and alternative formulations) and including a comprehensive survey of the datasets and benchmark resources on which that state of the art has been built. Section 2.3 synthesises the critical reading that runs through the chapter into five structural problems that motivate the rest of the thesis.

2.1 Conceptual foundations of media bias

This section addresses the *what* of media bias: how it is defined in the literature, which characteristics distinguish it from adjacent forms of information manipulation, and how it relates to the broader phenomenon of disinformation. It is heavily based on our previous work, *A Systematic Review on Media Bias Detection: What is Media Bias, How it is Expressed, and How to Detect It* [149], and provides the conceptual vocabulary that later chapters operationalise: the taxonomy of types and manifestations (Chapter 3) is built on the definitions introduced here, and the empirical work (Chapter 4) inherits the distinctions between bias, misinformation, disinformation, and propaganda established below.

[149] Rodrigo-Ginés et al., 2023, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it"

2.1.1 *Definition and characteristics*

Identifying bias in media content presents a significant challenge, particularly in a journalistic landscape that has largely moved away from the pursuit of pure neutrality and objectivity. Instead, modern journalism often aims for a more engaged and committed form of reporting [25]. This evolution complicates the distinction between opinion and bias, making it increasingly difficult to discern when media content is simply reflecting a perspective or actively distorting information.

[25] Boudana, 2011, "A definition of journalistic objectivity as a performance"

The Cambridge Dictionary defines opinion as "a thought or belief about something or someone" and bias as "a situation in which you support or oppose someone or something in an unfair way because you are influenced by your personal opinions" or "an unfair preference for one thing". These definitions suggest that the boundary between opinion and bias hinges

on whether a journalist employs rhetorical strategies to distort information in support of their viewpoint. If such distortion occurs, the content crosses from opinion into bias.

Despite the complexity surrounding its definition, certain characteristics of media bias are consistently highlighted in the literature. These characteristics include: (i) media bias is often **slanted**, meaning it is intended to influence the audience's opinion in a specific direction; (ii) it is typically sustained and/or consistent for each media outlet or journalist; and (iii) it may be employed to create *memorable stories* by using rhetorical techniques such as exaggeration or selective presentation of facts (spin bias). Table 2.1 provides an overview of various claims and definitions of media bias found in the literature, highlighting how different researchers have emphasised these key features (slantedness, frequency, and the creation of spin stories) in their conceptualisations of media bias.

TABLE 2.1: Overview of features in media bias claims/definitions. Features: (1) Slanted **(bold)**, (2) Sustained or frequent (underline), (3) Produced by creating 'memorable' (spin) stories *(italic)*.

Reference	Claim	Characteristics
Felix Hamborg, Karsten Donnay, and Bela Gipp. "Automated identification of media bias in news articles: an interdisciplinary literature review". In: *International Journal on Digital Libraries* 20.4 (2019), pp. 391–415	The study of biased news reporting has a long tradition in the social sciences going back at least to the 1950s. In the classical definition, media bias must both be **intentional**, i.e., reflect a conscious act or choice, and it must be sustained, i.e., represent a systematic tendency rather than an isolated incident. Various definitions of media bias and its specific forms exist, each depending on the particular context and research questions studied. Some authors define two high-level types of media bias concerned with the intention of news outlets when writing articles: **ideology** and *spin*.	(1), (2), (3)
Dave D'Alessio and Mike Allen. "Media bias in presidential elections: A meta-analysis". In: *Journal of communication* 50.4 (2000), pp. 133–156	The question of media bias is moot in the absence of certain properties of that bias: **It must be volitional, or willful**; *it must be influential, or else it is irrelevant*; it must be threatening to widely held conventions, lest it be dismissed as mere "crackpotism"; and it must be sustained rather than an isolated incident.	(1), (2), (3)
Sendhil Mullainathan and Andrei Shleifer. *Media bias*. 2002	There are two different types of media bias. One bias, which we refer to as **ideology, reflects a news outlet's desire to affect reader opinions in a particular direction**. *The second bias, which we refer to as spin, reflects the outlet's attempt to simply create a memorable story.*	(1), (3)
Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. "Automated identification of bias inducing words in news articles using linguistic and context-oriented features". In: *Information Processing & Management* 58.3 (2021), p. 102505	**Media bias is defined by researchers as slanted news coverage** or internal bias, reflected in news articles. By definition, remarkable media bias is deliberate, intentional, and has a particular purpose and tendency towards a particular perspective, ideology, or result. On the other hand, *bias can also be unintentional and even unconscious*.	(1), (3)
Matthew Gentzkow and Jesse M Shapiro. "Media bias and reputation". In: *Journal of political Economy* 114.2 (2006), pp. 280–316	The choice to **slant information** in this way (by selective omission, choice of words, and varying credibility ascribed to the primary source) is what we will mean in this paper by media bias.	(1)

A review of the literature reveals diverse and sometimes conflicting definitions of media bias [72]. One significant point of contention is the role of intentionality. While many scholars assert that media bias is inherently intentional and influential [118], others argue that bias can also be unintentional and unavoidable, as it is rooted in the human condition [87]. For example, Bozell and Baker [26] suggests that media bias often stems from the backgrounds

[72] Hamborg et al., 2019, "Automated identification of media bias in news articles: an interdisciplinary literature review"

[118] Mullainathan and Shleifer, 2002, *Media bias*

[87] Kang and Yang, 2022, "Quantifying perceived political bias of newspapers through a document classification technique"

[26] Bozell and Baker, 1990 *And that's the way it isn't: A reference guide to media bias*

and experiences of journalists rather than a deliberate effort to distort news. However, this perspective has been criticized, as media owners and editors have the ability, and arguably the responsibility, to mitigate personal biases in reporting [184].

[184] Sutter, 2000, "Can the media be so liberal-the economics of media bias"

Considering these foundational elements, media bias can be defined as a situation in which journalists slant information in a way that influences the audience's opinion, with this slant being frequent and/or consistent across different stories or outlets. Furthermore, the use of rhetorical techniques to create memorable stories often amplifies the effect of bias, making it a powerful tool in shaping public perception.

2.1.2 *Understanding the relationship between media bias and disinformation*

Media bias and disinformation are deeply interconnected phenomena [88, 52]. Both involve the manipulation of information to influence public perception, and both can employ similar techniques such as loaded language or selective presentation of facts. However, they differ in a crucial dimension: disinformation is characterized by a deliberate intent to mislead, whereas misinformation [39] involves erroneous information spread without such intent. Media bias, in turn, can function as a form of disinformation when it involves intentional distortion, but it may also arise from unconscious editorial choices [145].

[88] Karlova and Fisher, 2013, "A social diffusion model of misinformation and disinformation for understanding human information behaviour"

[52] Estrada-Cuzcano et al., 2020, "Disinformation y Misinformation, Posverdad y Fake News: precisiones conceptuales, diferencias, similitudes y yuxtaposiciones"

[39] Cummings and Kong, 2019, "Breaking down "fake news": Differences between misinformation, disinformation, rumors, and propaganda"

[145] Rodrigo-Ginés and Chamorro-Padial, 2026, "Media Bias Within Information Disorder: Bridging Two Research Communities Through a Systematic Review"

To properly understand where media bias sits within this broader information ecosystem, it is necessary to clarify the dimensions through which information can be manipulated. The following criteria capture fundamental properties that help differentiate between information, misinformation, disinformation, media bias, fake news, and propaganda:

- **True**: The information is factually accurate. For example, "The unemployment rate decreased to 12 %" is true if it matches official statistics.
- **Complete**: The information includes all relevant details needed for correct interpretation. Reporting that crime increased without noting that reporting rates also changed provides true but incomplete information.
- **Current**: The information is temporally correct and not taken out of context. Reusing a 2012 quote in a 2025 debate may violate this criterion.
- **Informative**: The message provides meaningful understanding. A headline such as "Experts react to new policy" without details is minimally informative.
- **Deceptive**: The message intentionally misleads through falsity, selective omission, emotional manipulation, or narrative framing. "Scientists confirm dangerous vaccine risks" when no such confirmation exists is deceptive.
- **Slanted**: The presentation systematically favours an actor, ideology, or interpretation without necessarily lying. Examples include lexical asymmetries ("criminal migrant" vs. "local youth"), quotation imbalance, or evaluative predicates. Crucially, slant does not require falsity.

These characteristics reveal that information disorder is multidimensional: some phenomena differ in truthfulness (true vs. deceptive), others in completeness or contextualisation, and others in systematic ideological orientation (slant). Media bias occupies a distinctive position within this space: it is primarily defined by *slant*, although it may coexist with incompleteness or low informativeness. Importantly, slant can interact with deception when intention is present, at which point biased content may contribute to or amplify disinformation.

Therefore, it is essential to incorporate "slanted" as a characteristic into the framework proposed by Karlova and Fisher [88]. Their model provides a useful distinction between information, misinformation, and disinformation, but it does not explicitly incorporate media bias or explain how narrative orientation modulates interpretation. Adding "slanted" extends their framework and enables a more complete understanding of the spectrum of information manipulation: from media bias (non-deceptive but systematically oriented), to misinformation (false without intent), to disinformation (false with intent).

To further clarify these relationships, Table 2.2 summarises how the criteria defined above manifest across major forms of information disorder. The table highlights shared elements (e.g., incomplete information can appear in both biased reporting and misinformation)

while revealing key differences (e.g., intentional deception is unique to disinformation and propaganda).

TABLE 2.2: A summary of features of disinformation, media bias, fake news, and propaganda. Y/N = Could be yes or no, depending on context and time.

	Disinformation	**Media bias**	**Fake news**	**Propaganda**
True	Y/N	Y/N	No	Y/N
Complete	Y/N	Y/N	Y/N	Y/N
Current	Y/N	Y/N	Y/N	Y/N
Informative	Yes	Yes	Y/N	No
Deceptive	Yes	Y/N	Yes	Yes
Slanted	Y/N	Yes	Y/N	Y/N

By referencing this table, we can more effectively analyze the complex interactions between different forms of information manipulation and better grasp the specific role media bias plays within the broader disinformation ecosystem. Understanding these relationships is key for refining algorithms in NLP and ML. It is essential that these technologies recognize not only overt disinformation but also the subtle indicators of media bias, making effective categorization by context or intention.

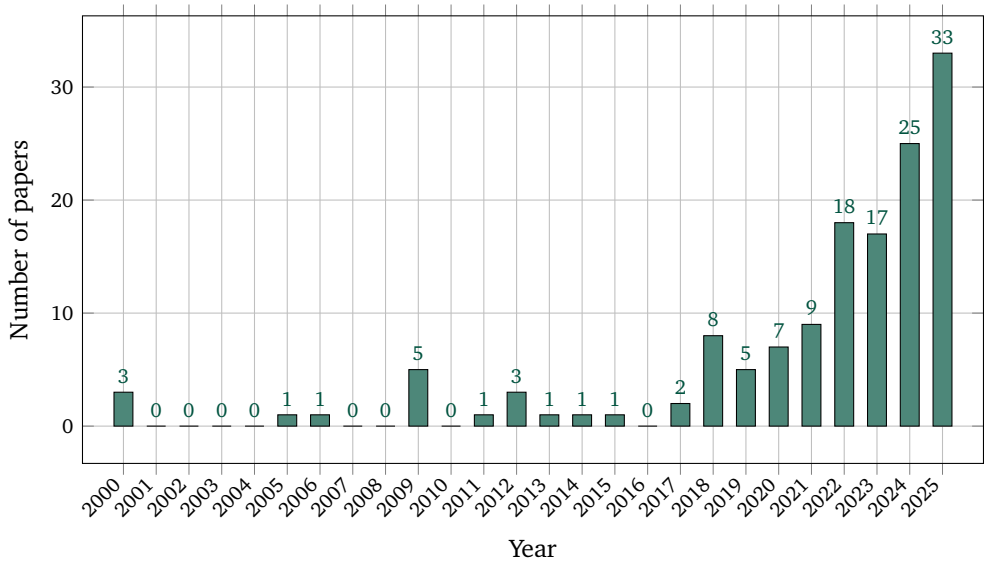
2.2 COMPUTATIONAL APPROACHES TO MEDIA BIAS DETECTION

Having established the conceptual foundations of media bias, we now turn to the *how*: how has Computer Science, and NLP in particular, attempted to detect it automatically? This section reviews the state of the art in automatic media bias detection, organised around the methodological evolution of the field. We begin with the review methodology itself (Subsection 2.2.1), then survey the dominant modelling paradigms from traditional approaches through transformer-based models (Subsection 2.2.2), and finally review the datasets and benchmark resources on which these methods are trained and evaluated (Subsection 2.2.3). Throughout, we highlight both the contributions and the limitations of each line of work; the synthesis of these limitations into five structural problems is deferred to Section 2.3.

The growth of the field over the last decade is visible at a glance: the number of peer-reviewed papers on media bias detection has roughly doubled every three years since 2018 (see Figure 2.1), driven largely by the availability of pre-trained transformers and, more recently, large language models. Our review extends a previous systematic survey [149] covering articles up to 2022 with additional literature published between 2023 and 2025, yielding a total of 118 included studies.

[149] Rodrigo-Ginés et al., 2023, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it"

FIGURE 2.1: Number of published papers on media bias detection from 2000 to 2025, based on a systematic search in Google Scholar.

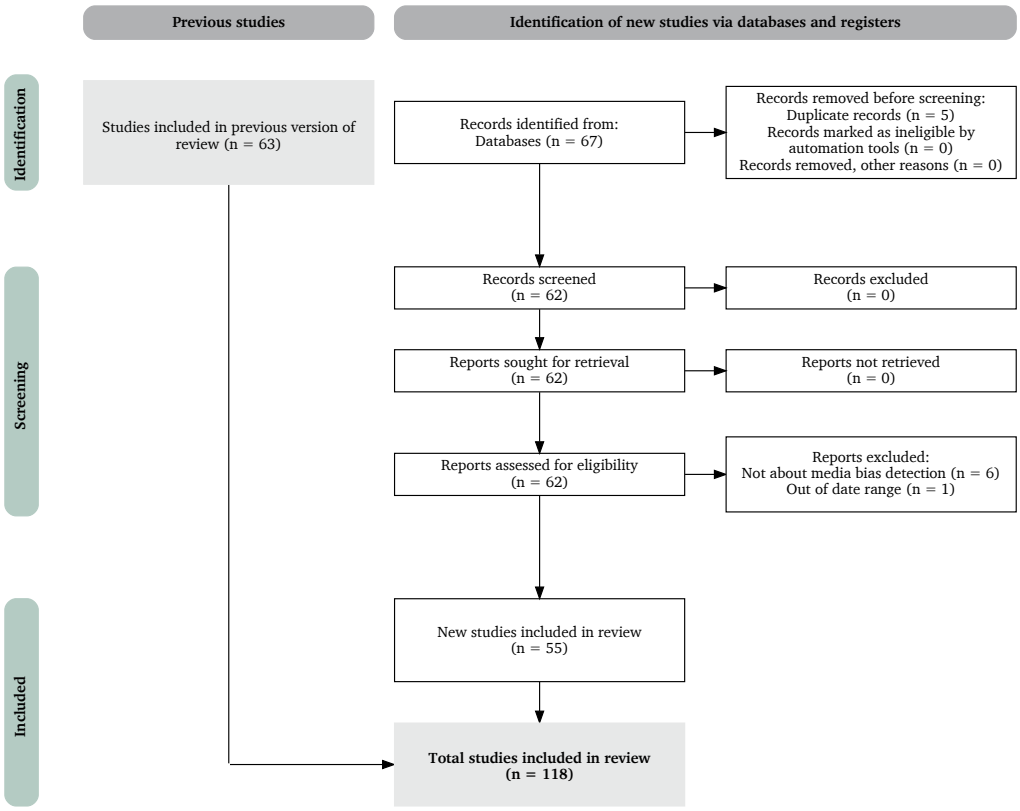


2.2.1 *Review methodology and scope*

To ensure a comprehensive and systematic approach to the literature review, we employed the PRISMA guidelines [168]. This methodology was chosen due to its structured approach, which enhances the transparency and replicability of systematic reviews. The search query used was: *"(media OR journalism OR news) AND (bias OR slant OR spin) AND (detection OR characterization OR classification)"*, applied to the title field across three databases: Google Scholar, Scopus, and ACL Anthology. Studies were included if they (a) focused on methods for detecting or characterising media bias, (b) were peer-reviewed, and (c) were written in English. Exclusion criteria included studies not directly related to media bias detection, studies lacking methodological novelty, and studies outside the specified date range.

[168] Selçuk, 2019, "A guide for systematic reviews: PRISMA"

FIGURE 2.2: PRISMA flow diagram of the systematic review process.



The review process was conducted in two phases. In the first phase, we utilized the dataset from our previous systematic review Rodrigo-Ginés et al. [149], which covered articles published between 2000 and 2022. This initial review identified a total of 63 studies that were relevant to the detection of media bias. These studies provided a foundational understanding of the state-of-the-art approaches in media bias detection, which were essential for contextualizing more recent developments.

In the second phase, we extended the review to include studies published between 2023 and 2025. Using the same search strategies and databases as in the previous review (Google Scholar, Scopus, and ACL Anthology), we identified 67 new records. The software *Publish or Perish* [74] was employed to facilitate the search and retrieval of academic papers across these databases. This tool enabled a more efficient and thorough search process, ensuring that we captured a wide range of relevant studies.

[74] Harzing, 2010, *The publish or perish book*

After removing duplicate entries, we screened a total of 62 new studies. Following a thorough screening process, 55 studies were deemed relevant and included in the updated review, while 7 studies were excluded due to being out of scope or out of the specified date range.

The inclusion criteria were consistent across both phases of the review. Articles were included if they focused on methods for detecting media bias, were peer-reviewed, and were

written in English. Exclusion criteria included studies that were not directly related to media bias detection, lacked novelty, or were not aligned with the current research focus.

Ultimately, this updated review added 55 new studies to the 63 originally identified, resulting in a total of 118 studies included in the final review. This comprehensive dataset provides a robust basis for analyzing trends, methodologies, and advancements in the field of automatic media bias detection over the past two decades.

The entire review process, including the identification, screening, and inclusion of studies, is illustrated in Figure 2.2. This figure provides a visual representation of the PRISMA flow diagram, detailing the number of records identified, screened, and included in the final review.

Related challenges in media analysis

As highlighted in Subsection 2.1.2, media bias is just one aspect of the broader landscape of disinformation and information integrity. During our systematic review, we specifically focused on identifying studies that directly address media bias detection. However, in doing so, we encountered a range of related problems that, while distinct, share similarities with media bias and present unique challenges in the realm of media analysis. In this section, we explore these adjacent issues, which were not the primary focus of our review but are relevant to understanding the broader context of biased information in media.

▸ Stance detection

[185] Taulé et al., 2017, "Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017"

Stance detection [185] is closely aligned with media bias detection, as it involves classifying information based on the author's position or opinion on a given topic. This task is essential for understanding the perspective from which a piece of content is created, similar to identifying bias in media. Stance detection has also been employed in media and author profiling, contributing to a deeper understanding of the underlying attitudes expressed in the text.

▸ Propaganda detection

[40] Da San Martino et al., 2021, "A survey on computational propaganda detection"

Propaganda, as noted in Section 2.1.2, is a deliberate attempt to shape public opinion, often through a mix of true and misleading information. Propaganda detection focuses on identifying the rhetorical and psychological techniques used to craft such content. Techniques like name-calling, glittering generalities, and bandwagon appeals are among the common strategies studied in this area [40]. This task, while related to media bias detection, specifically targets the intent to manipulate rather than merely identifying bias.

▸ Fake news detection

[208] Zhou and Zafarani, 2020, "A survey of fake news: Fundamental theories, detection methods, and opportunities"

The distinction between fake news and media bias lies in the veracity of the content: fake news is inherently false, while biased media may contain truthful elements presented in a skewed manner. Fake news detection systems typically approach the problem from three perspectives: (1) style-based, focusing on how fake news is written; (2) propagation-based, examining how fake news spreads; and (3) user-based, analyzing user interactions with fake news and their role in its propagation [208].

▸ Rumour detection

[170] Shelke and Attar, 2019, "Source detection of rumor in social network–a review"

Rumour detection addresses the challenge of identifying unverified information that spreads rapidly, particularly on social media. Unlike fake news, which is definitively false, rumors may eventually be proven true or false. Rumour detection often involves analyzing the diffusion of information across networks, with source detection being a critical component of this task [170]. The rapid spread of rumours can blur the lines between truth and falsehood, making this a particularly complex problem to address.

▸ Clickbait detection

Clickbait refers to online content designed to attract attention and encourage clicks, often through misleading or sensational headlines. This type of content frequently uses language that promises an emotional response, hints at hidden information, or appeals to authority

without delivering substantive content [135]. Detecting clickbait is crucial for maintaining the integrity of online information and preventing the spread of misleading narratives.

[135] Potthast et al., 2016, "Clickbait detection"

▸ **Logical fallacies detection**

The detection of logical fallacies in arguments is another related problem that overlaps with media bias detection. Logical fallacies are errors in reasoning that can undermine the validity of an argument. Identifying these fallacies within journalistic content is essential for ensuring that the information presented is both logically sound and free from manipulation. This task is especially relevant in the context of biased reporting [85], where fallacies may be used to subtly influence public perception.

[85] Jin et al., 2022, *Logical Fallacy Detection*

▸ **Bias detection in non-journalistics contexts**

Bias detection is not confined to journalistic content; it is also applicable in various other domains. For instance, detecting bias in Wikipedia articles [79], analyzing congressional speeches for bias [69], and identifying sexist language on social media [148] are all areas where bias detection plays a crucial role. These tasks demonstrate the broader applicability of bias detection methods beyond the realm of traditional news media.

[79] Hube and Fetahu, 2018, "Detecting biased statements in wikipedia"

[69] Hajare et al., 2021, "A Machine Learning Pipeline to Examine Political Bias with Congressional Speeches"

[148] Rodrigo-Ginés et al., 2021, "UNEDBiasTeam at IberLEF 2021's EXIST Task: Detecting Sexism Using Bias Techniques."

2.2.2 *State of the art by methodological paradigm*

This section presents a structured overview of the state of the art in automatic media bias detection, focusing on how the problem has been approached from a computational perspective. The reviewed approaches are organized according to their underlying methodological paradigms, reflecting the historical evolution of the field as well as the dominant assumptions adopted at each stage. By examining these approaches in a systematic manner, this section aims to highlight both their contributions and their limitations, particularly in relation to task formulation, linguistic modeling, and generalization capabilities.

Traditional approaches

Early computational approaches to media bias detection were predominantly grounded in non-deep learning paradigms, relying on classical machine learning models such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), Naïve Bayes (NB), and gradient-boosting variants. These approaches framed media bias detection primarily as a supervised classification problem, most often at the sentence or document level, and depended critically on the design and extraction of handcrafted features. Consequently, model performance was tightly coupled to feature engineering choices, domain expertise, and the availability of bias-relevant linguistic resources, which imposed clear limitations in terms of scalability, adaptability, and cross-domain generalization [38]. Within this paradigm, two dominant methodological lines can be distinguished: linguistic-based methods and reported speech-based methods, each addressing different manifestations of media bias.

[38] Cruz et al., 2019, "On sentence representations for propaganda detection: From handcrafted features to word embeddings"

▸ **Linguistic-based methods**

Linguistic-based approaches operationalize media bias through observable textual cues, modeling bias as a function of lexical choice, syntactic structure, or shallow semantic indicators. These methods typically extract features such as word or character n-grams, TF-IDF representations, part-of-speech distributions, dependency patterns, named entities, sentiment scores, or psychologically motivated lexicons such as LIWC. The resulting feature vectors are then used to train conventional classifiers for binary or multiclass bias detection tasks.

One of the earliest systematic efforts in this direction is the work by Krestel et al. [93], who analyzed vocabulary usage in German online news by contrasting it against a reference corpus of German parliamentary speeches. Their approach modeled bias implicitly through lexical alignment between media outlets and political actors, enabling the identification of ideological proximity at an aggregate level. The evaluation was conducted at the corpus and outlet level using parliamentary discourse as an ideological anchor, rather than on a sentence or document-level annotated benchmark, and therefore does not report standard classification metrics such as accuracy or F1-score.

[93] Krestel et al., 2012 "Treehugger or Petrolhead? Identifying bias by comparing online news articles with political speeches"

A more fine-grained formulation was proposed by Hube and Fetahu [79], who introduced a sentence-level binary classification task distinguishing biased from non-biased statements. Using a manually annotated corpus of news sentences, they trained classical classifiers on a combination of custom bias lexicons, part-of-speech tags, and LIWC features, reporting an accuracy of 0.74. This work explicitly linked linguistic markers to perceived bias at the sentence level and became a methodological reference for subsequent studies adopting similar feature sets. For example, Spinde et al. [176] replicated this paradigm on a comparable sentence-level task using TF-IDF representations, but reported a substantially lower F1-score of 0.43, highlighting the sensitivity of traditional models to feature selection, annotation guidelines, and corpus composition.

[79] Hube and Fetahu, 2018 "Detecting biased statements in wikipedia"

[176] Spinde et al., 2020 "An integrated approach to detect media bias in german news articles"

Further refinements were explored by Gummalam and Greenberg [66], who formulated media bias detection as a sentence-level binary classification task and combined Pointwise Mutual Information (PMI) and TF-IDF heuristics to model associations between lexical items and bias classes. In addition, they incorporated Google's Universal Sentence Encoder to capture sentence-level semantics. While their analysis provides qualitative insights into topic–bias relationships, the authors do not report results on widely used benchmark datasets, nor do they provide directly comparable classification metrics, which limits quantitative comparison with prior work.

[66] Gummalam and Greenberg, 2024 "Bias Detection in Media: An NLP-Based Approach using Corpus Statistics and Sentence Embeddings"

Several studies have applied similar linguistic pipelines to politically sensitive or conflict-driven contexts. Chan et al. [32], for instance, examined media coverage of the Israel–Palestine conflict and identified bias indicators such as emotive language, weasel words, and loaded comparisons. Their contribution lies primarily in the design of annotation guidelines and in the discussion of multipartiality and ethical considerations in labeling highly polarized content, rather than in the introduction of a benchmark-oriented classification model with directly comparable performance metrics. Other works, such as Baraniak and Sydow [18], emphasized the role of source identification and outlet-level language patterns, using relatively simple linguistic features to distinguish biased reporting across digital media platforms. In the related task of hyperpartisan news detection, Gupta et al. [67] trained an XGBoost classifier on lexical and stylistic features and reported a precision of 0.68 on the SemEval Hyperpartisan News Detection dataset at the document level, while also noting reduced robustness when applied to unseen outlets or topics.

[32] Chan et al., 2024 "Eagles at FIGNEWS 2024 shared task: A context-informed prescriptive approach to bias detection in contentious news narratives"

[18] Baraniak and Sydow, 2018 "News articles similarity for automatic media bias detection in Polish news portals"

[67] Gupta et al., 2019 "Clark Kent at SemEval-2019 Task 4: Stylometric Insights into Hyperpartisan News Detection"

Taken together, linguistic-based methods established that bias can be partially captured through surface-level textual features and that classical classifiers are capable of learning such patterns under controlled conditions. However, their effectiveness is inherently constrained by the expressiveness of handcrafted features and by their limited ability to model long-range dependencies, discourse structure, and context-dependent meaning shifts, which are central to many forms of media bias.

▸ Reported speech-based methods

Reported speech-based approaches address media bias from a complementary angle by focusing on quotation practices within news articles. Rather than analyzing language in isolation, these methods examine who is quoted, how frequently, in what context, and in relation to which events or actors. This line of work is particularly relevant for studying statement bias and ideological framing, where bias emerges not only from wording but also from selective attribution and source prominence.

An early example is the work of Park et al. [126], who extracted quoted speech using named entity recognition and coreference resolution, and then applied the Hypertext Induced Topic Selection (HITS) algorithm to partition disputants into opposing ideological groups. Evaluated on political news articles containing quoted statements, their system achieved a weighted F-measure of 0.68, demonstrating that quotation patterns alone can provide meaningful signals of partisan alignment without relying on lexical sentiment or subjectivity features.

[126] Park et al., 2011 "Contrasting opposing views of news articles on contentious issues"

A seminal contribution in this area is the framework introduced by Niculae et al. [121], which systematically quantified media quoting behavior by analyzing how different outlets selected and reproduced excerpts from political speeches. Using a large corpus of U.S.A. pres-

[121] Niculae et al., 2015 "Quotus: The structure of political media coverage as revealed by quoting patterns"

idential speeches and corresponding media coverage, they showed that declared conservative outlets were significantly less likely to quote statements emphasized by liberal outlets, even when controlling for newsworthiness. This work did not formulate bias media detection as a supervised classification task, but rather as a relational analysis of quotation patterns, and therefore does not report conventional classification metrics.

Subsequent studies extended this perspective. Lazaridou et al. [97] proposed a bias-aware model that classified reported speech according to its original outlet, achieving an accuracy of 0.79 on a news corpus with explicitly annotated source attributions. Their approach demonstrated that quotation context and attribution cues could be leveraged to infer ideological alignment with relatively high reliability. More recently, Kuculo et al. [96] explored multilingual extensions of this paradigm by constructing knowledge graphs that capture cross-lingual quotation relationships, enabling the analysis of bias arising from one-sided references and selective sourcing across languages, albeit without direct comparison to sentence-level bias benchmarks.

[97] Lazaridou et al., 2017 "Identifying media bias by analyzing reported speech"

[96] Kuculo et al., 2022 "QuoteKG: A Multilingual Knowledge Graph of Quotes"

Reported speech-based methods thus provide a structurally informed view of media bias, capturing distortions introduced through source selection and attribution practices. However, they are computationally demanding, require accurate quotation extraction and entity resolution, and are often constrained to domains where rich quotation data is available. Moreover, these methods typically operate at the outlet or article level and are less suited for detecting fine-grained linguistic bias within individual sentences.

Overall, traditional approaches to media bias detection laid the conceptual and methodological foundations of the field by formalizing media bias as a learnable signal in text and media structure. Yet, their dependence on handcrafted features, limited contextual modeling capacity, and fragility under domain shifts motivated the transition toward representation-learning approaches, setting the stage for the adoption of deep learning models in subsequent work.

Pre-transformer Deep Learning approaches

The introduction of deep learning marked a turning point in automatic media bias detection by shifting the modeling paradigm from handcrafted feature engineering to representation learning. Unlike traditional machine learning approaches, deep neural models are able to learn latent linguistic patterns directly from data, reducing the dependence on manually designed features and enabling more flexible modeling of language phenomena. In the context of media bias detection, this transition was primarily driven by the limitations of surface-level lexical and syntactic features in capturing contextual, discursive, and stylistic cues associated with biased reporting. Early deep learning approaches predominantly relied on Recurrent Neural Networks (RNNs) and their variants, which were well suited to sequential text modeling and therefore represented a natural step forward for sentence and document-level bias classification tasks [183].

[183] Sutskever et al., 2014, "Sequence to sequence learning with neural networks"

▸ RNN-BASED METHODS

Recurrent Neural Networks are designed to process sequences by maintaining a hidden state that is updated at each time step, allowing information from previous tokens to influence the representation of subsequent ones. This sequential inductive bias makes RNNs particularly appropriate for natural language processing tasks where word order and local context are relevant. However, vanilla RNNs are known to suffer from the vanishing gradient problem, which hampers their ability to model long-range dependencies in text [171]. To mitigate this issue, Long Short-Term Memory (LSTM) networks introduced gated mechanisms that enable the selective retention and forgetting of information over longer sequences, significantly improving modeling capacity for extended textual inputs.

[171] Sherstinsky, 2020, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network"

One of the earliest applications of deep neural architectures to politically charged language is the work by Iyyer et al. [82], who employed a recursive neural network framework to identify political ideology and stance in text. Their model was initialized with word embeddings derived from word2vec and trained using phrase-level supervision. Evaluated on sentence-

[82] Iyyer et al., 2014 "Political ideology detection using recursive neural networks"

level political classification tasks, their approach achieved an accuracy of 0.69, demonstrating that neural models could capture ideological signals beyond those accessible through bag-of-words representations. While not explicitly framed as media bias detection, this work laid important groundwork for subsequent bias-oriented applications by showing the feasibility of neural stance modeling.

[139] Rashkin et al., 2017 "Truth of varying shades: Analyzing language in fake news and political fact-checking"

A more direct formulation of bias detection using recurrent architectures was explored by Rashkin et al. [139], who compared LSTM-based classifiers against classical baselines such as Naïve Bayes and Maximum Entropy models. Using text embeddings as input representations, their LSTM achieved an F1-score of 0.56 on sentence-level classification tasks involving biased and deceptive language, outperforming traditional methods under the same experimental conditions. The authors further improved performance by initializing the network with 100-dimensional GloVe embeddings, illustrating the importance of semantic pretraining even in pre-transformer settings.

[148] Rodrigo-Ginés et al., 2021, "UNEDBiasTeam at IberLEF 2021's EXIST Task: Detecting Sexism Using Bias Techniques."

Bidirectional variants of recurrent models, which process sequences in both forward and backward directions, were subsequently adopted to better capture contextual dependencies. Bidirectional LSTMs (Bi-LSTMs) were shown to outperform unidirectional models and statistical classifiers in several bias-related tasks. In our earlier work [148], Bi-LSTM architectures achieved superior results in media bias detection compared to traditional machine learning baselines, confirming the advantage of bidirectional context modeling for identifying biased language patterns.

Despite these improvements, plain RNN and LSTM models typically rely on the final hidden state of the sequence as a compressed representation of the entire input. This design choice poses limitations when dealing with long sentences or documents, as salient bias cues may appear at different positions and be diluted in the final representation. To address this issue, attention mechanisms were introduced, allowing the model to dynamically weight different parts of the input sequence when making predictions.

[80] Hube and Fetahu, 2019 "Neural based statement classification for biased language"

The effectiveness of attention-enhanced recurrent models was demonstrated by Hube and Fetahu [80], who experimented with both global and hierarchical attention mechanisms for sentence-level media bias detection. Evaluated on annotated news sentence datasets, their attention-based models achieved an average F1-score of 0.77, compared to 0.74 for their non-attentive counterparts under identical experimental setups. This improvement indicates that attention mechanisms enable the model to focus on bias-inducing lexical and phrasal cues, rather than treating all tokens as equally informative.

[75] Hong et al., 2023 "Disentangling structure and style: Political bias detection in news by inducing document hierarchy"

More recent work within the pre-transformer paradigm has explored increasingly sophisticated attention structures. Hong et al. [75], for example, introduced a multi-head hierarchical attention model designed to disentangle stylistic and structural aspects of political bias in news articles. Their model explicitly separates sentence-level semantic representations from document-level rhetorical patterns, aiming to reduce overfitting to outlet-specific writing styles. Evaluated on political news corpora, their approach demonstrated improved robustness and accuracy compared to earlier recurrent architectures, highlighting the importance of modeling discourse structure in addition to local lexical cues.

Overall, RNN-based approaches represented a substantial advancement over traditional feature-based models by enabling end-to-end learning of bias-relevant representations and by partially capturing contextual dependencies within text. Nevertheless, these models remained constrained by their sequential nature, limited parallelization capabilities, and difficulties in modeling very long-range dependencies. Moreover, their performance was sensitive to domain shifts and annotation schemes, and they often required task-specific architectural tuning. These limitations, combined with the growing availability of large-scale pretraining corpora, ultimately motivated the transition toward transformer-based architectures, which offered a fundamentally different approach to contextual modeling and set new performance benchmarks in media bias detection.

Transformer-based Deep Learning approaches

The introduction of transformer architectures represented a fundamental methodological shift in automatic media bias detection, largely superseding recurrent neural networks and feature-based approaches. By relying on self-attention mechanisms rather than sequential recurrence, transformers are able to model long-range dependencies and global contextual interactions within text in a fully parallelizable manner [195]. This property is particularly well suited to media bias detection, where bias cues may be sparse, distributed across sentences, or expressed through subtle contextual contrasts rather than isolated lexical markers. As a result, transformer-based models rapidly became the dominant paradigm in the field, consistently outperforming both traditional machine learning methods and pre-transformer deep learning architectures [14].

[195] Vaswani et al., 2017, "Attention is all you need"

[14] Baly et al., 2020, "We Can Detect Your Bias: Predicting the Political Ideology of News Articles"

Crucially, however, transformer-based approaches to media bias detection are not monolithic. Two broad architectural families, encoder-only (bidirectional) models and decoder-only (autoregressive) models, exhibit markedly different inductive biases, training objectives, and empirical behavior. Understanding these differences is essential for interpreting reported results and for situating recent advances within a coherent methodological framework.

▸ Encoder-only (bidirectional) transformer models

Encoder-only transformers, most notably BERT and its variants, have been the most extensively adopted and empirically successful architectures for media bias detection. These models are trained using masked language modeling objectives that encourage deep bidirectional contextualization, allowing each token representation to incorporate information from both preceding and following context. This bidirectional conditioning is particularly important for detecting bias in news text, where evaluative meaning often depends on contrastive framing, implicit presuppositions, or the interaction between named entities and surrounding modifiers.

One of the earliest demonstrations of the effectiveness of encoder-only transformers for media bias detection was provided by Fan et al. [53], who fine-tuned a cased BERT-Base model on the BASIL dataset for sentence-level bias classification. By preserving capitalization and named entity information, their model achieved an F1-score of 0.432, substantially outperforming earlier lexicon-based and subjectivity-driven approaches on the same benchmark. This work established contextualized representations as a decisive improvement over surface-level linguistic features for identifying biased language.

[53] Fan et al., 2019 "In Plain Sight: Media Bias Through the Lens of Factual Reporting"

Subsequent studies systematically explored the robustness and generalization capacity of encoder-only models. Chen et al. [35] highlighted limitations of both feature-based and neural classifiers in document-level bias detection, particularly when models were exposed to articles covering previously unseen events. By leveraging a pre-trained uncased BERT model and explicitly modeling the frequency, position, and sequential arrangement of biased sentences within articles, they demonstrated improved document-level performance, underscoring the importance of aggregation strategies that preserve contextual structure beyond isolated sentence predictions.

[35] Chen et al., 2020 "Detecting Media Bias in News Articles using Gaussian Bias Distributions"

A broader comparison of transformer architectures was conducted by Spinde et al. [180], who fine-tuned BERT, RoBERTa, and DistilBERT for sentence-level bias detection and showed consistent gains over linguistic baselines across multiple datasets. Their follow-up work [178] introduced a multi-task learning framework that jointly modeled bias detection and auxiliary tasks, reporting an average performance improvement of approximately 4% over single-task transformer models. These results suggest that shared representations across related objectives can improve bias detection, particularly in low-resource annotation scenarios.

[180] Spinde et al., 2021 "Automated identification of bias inducing words in news articles using linguistic and context-oriented features"

[178] Spinde et al., 2022, "Exploiting transformer-based multitask learning for the detection of media bias in news articles"

Domain-adaptive pretraining has emerged as a key factor in further advancing encoder-only models. Krieger et al. [95] introduced domain-adapted variants of BERT, RoBERTa, and BART by continuing pretraining on bias-relevant news corpora prior to fine-tuning. Evaluated on established media bias benchmarks, their domain-adapted RoBERTa model achieved an F1-score of 0.81, setting a new state of the art under comparable experimental conditions. This finding provides strong empirical evidence that general-purpose pretraining

[95] Krieger et al., 2022 "A domain-adaptive pre-training approach for language bias detection in news"

alone is insufficient for capturing bias-specific discourse patterns and that domain adaptation substantially enhances robustness and performance.

Encoder-only transformers have also been successfully applied across languages and cultural contexts. Agrawal et al. [1] employed XLM-RoBERTa to detect political bias in Hindi news articles, achieving an accuracy of 0.83 on a manually annotated corpus. Similarly, Ali et al. [3] explored ConvBERT for bias detection in Pakistani news media using distant supervision and domain-aware pretraining, reporting a macro F1-score of 0.81. These studies demonstrate that the advantages of bidirectional contextual modeling extend beyond English-centric benchmarks, provided that linguistic and domain-specific adaptations are carefully handled.

[1] Agrawal et al., 2022 "Towards detecting political bias in hindi news articles"

[3] Ali et al., 2024 "Neural transformers for bias detection: Assessing Pakistani News"

Despite their strong performance, encoder-only models are not without limitations. Several studies have shown that these models may exploit spurious correlations or stylistic shortcuts rather than genuinely bias-relevant cues. Wessel and Horych [201], for instance, demonstrated that transformer-based classifiers can be sensitive to superficial features, raising concerns about robustness and interpretability. Nevertheless, across a wide range of datasets and evaluation setups, encoder-only transformers remain the most reliable and consistently high-performing models for discriminative media bias detection.

[201] Wessel and Horych, 2024 "Beyond the surface: Spurious cues in automatic media bias detection"

▸ **Decoder-only (autoregressive) transformer models and large language models**

Decoder-only transformer architectures, including GPT-style large language models (LLMs), differ fundamentally from encoder-only models in both training objective and inductive bias. Trained using autoregressive next-token prediction, these models condition exclusively on left context, which enables powerful generative capabilities but limits direct access to bidirectional contextual information. In recent years, decoder-only LLMs have been increasingly explored for media bias detection, particularly in zero-shot and few-shot settings, motivated by their flexibility, scale, and strong performance across a wide range of NLP tasks.

Empirical evaluations, however, suggest that decoder-only models do not consistently outperform encoder-only transformers on benchmark-oriented bias classification tasks. Menzner and Leidner [115] conducted a comprehensive evaluation of GPT-2, GPT-3, GPT-4, and Meta LLaMA2 on the Media Bias Identification Benchmark (MBIB), comparing them against encoder-based classifiers such as ConvBERT and RoBERTa. Their results showed that encoder-only models achieved superior performance across multiple bias types, while decoder-only models (particularly smaller GPT variants) underperformed in terms of classification accuracy and F1-score. Similar findings were reported by Wessel [200], who observed that autoregressive models struggled with fine-grained bias categories requiring nuanced contextual interpretation.

[115] Menzner and Leidner, 2024 "Experiments in news bias detection with pre-trained neural transformers"

[200] Wessel, 2023 *Improving Media Bias Detection with State-of-the-art Transformers*

Nevertheless, decoder-only models exhibit strengths in complementary dimensions. Menzner and Leidner [116] introduced BiasScanner, a system that leverages a pre-trained LLM to identify and classify biased sentences at a fine-grained level while providing natural-language explanations for its predictions. Although BiasScanner does not demonstrate state-of-the-art performance on standard benchmarks, it represents a significant step toward deployable bias detection systems, emphasizing usability and explainability rather than purely discriminative accuracy.

[116] Menzner and Leidner, 2025 "BiasScanner: Automatic News Bias Classification for Strengthening Democracy"

In highly polarized or domain-specific contexts, decoder-only models have shown mixed but occasionally strong results. Sadiah et al. [160] employed GPT-3.5 Turbo to analyze bias in news narratives related to the Israel–Gaza conflict, reporting high accuracy within their domain-specific dataset. However, the evaluation did not include out-of-domain testing or comparison against encoder-only baselines on shared benchmarks, limiting the generalizability of these findings. Similarly, Pelloin et al. [130] used LLMs in few-shot conversational settings to analyze gender bias in French broadcast news, revealing systematic representational imbalances but without framing the task as a benchmark classification problem.

[160] Sadiah et al., 2024 "Ceasefire at FIGNEWS 2024 shared task: Automated detection and annotation of media bias using large language models"

[130] Pelloin et al., 2024 "Automatic classification of news subjects in broadcast news: Application to a gender bias representation analysis"

Recent work has further highlighted structural limitations of decoder-only models for bias detection. Pastorino et al. [128] showed that GPT-4 and GPT-3.5 Turbo frequently conflate emotional language with framing bias, suggesting difficulties in disentangling closely related discourse phenomena. Lin et al. [105] reported notable performance degradation of fine-tuned

[128] Pastorino et al., 2024 "Decoding News Narratives: A Critical Analysis of Large Language Models in Framing Bias Detection"

[105] Lin et al., 2024 "Indivec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators"

GPT models under domain shift, while Lin et al. [107] demonstrated that LLMs themselves may encode ideological biases that influence downstream bias predictions.

[107] Lin et al., 2024 *Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception*

Recent work has also explored unified and multi-bias frameworks that aim to detect heterogeneous bias types within a single modeling architecture. Sar et al. [163] proposed BAAF, a supervised framework based on AdapterFusion that integrates knowledge from multiple bias-specific adapters in a multi-task learning setup. Evaluated on the Media Bias Identification Benchmark (MBIB), which covers multiple bias types in social media content, their approach outperformed in-context learning baselines based on Llama-3.1-70B across political, gender, cognitive, and contextual biases. This result suggests that parameter-efficient fine-tuning strategies can offer a robust and scalable alternative to prompt-based LLM approaches, particularly when dealing with heterogeneous bias taxonomies.

[163] Sar et al., 2025 "BAAF: A Framework for Media Bias Detection"

A further emerging direction concerns the use of Large Language Models for dataset creation and annotation rather than for direct bias classification. Horych et al. [77] investigated the viability of LLMs as annotators for media bias detection, introducing Annolexical, a large-scale dataset comprising over 48,000 synthetically annotated examples. A classifier fine-tuned on this dataset outperformed all annotator LLMs by 5–9% in Matthews Correlation Coefficient and achieved performance comparable to, or exceeding, models trained on human-labeled data when evaluated on established benchmarks such as BABE and BASIL. This work demonstrates that LLM-assisted annotation can substantially reduce labeling costs while maintaining competitive downstream performance, but also reveals limitations related to annotation noise and bias propagation that must be carefully managed.

[77] Horych et al., 2025 "The promises and pitfalls of LLM annotations in dataset labeling: A case study on media bias detection"

Finally, application-oriented systems have begun to bridge the gap between research prototypes and real-world deployment. Menzner and Leidner [116] presented BiasScanner, a browser-based application that integrates a server-side large language model to identify and classify biased sentences in online news articles. While BiasScanner does not claim state-of-the-art classification performance on standard benchmarks, it represents a notable step toward operationalizing fine-grained bias detection and highlights the growing importance of usability, transparency, and explanation quality alongside raw predictive accuracy.

Alternative methods

While supervised classification approaches based on linguistic features and deep neural representations have dominated media bias detection research, a parallel line of work has explored alternative methodological formulations that depart from purely text-centric or sentence-level modeling. These approaches typically redefine the bias detection problem by shifting the unit of analysis, incorporating structural or relational information, or reframing bias as a large-scale estimation or annotation challenge rather than a standard classification task. As such, they offer complementary perspectives on media bias, often prioritizing scalability, interpretability, or contextual grounding over direct comparability with benchmark-oriented models.

One prominent alternative line concerns graph-based and network-oriented approaches, which explicitly model relationships between articles, authors, sources, and topics. Early work in this direction framed bias as an emergent property of media networks rather than as an intrinsic property of isolated texts. For instance, stakeholder mining approaches such as that of Ogawa et al. [122] constructed relational graphs by identifying participants and their interactions within news articles, enabling the grouping of articles according to shared stakeholder structures. Although not evaluated as a supervised bias classification task with standard metrics, this approach demonstrated how relational information can reveal systematic patterns of media framing that are difficult to capture through lexical analysis alone.

[122] Ogawa et al., 2011 "News bias analysis based on stakeholder mining"

More recent work has operationalized this intuition using graph machine learning techniques. Majali et al. [108] introduced BiasLens, a framework that constructs a knowledge graph linking articles, authors, and publishers, and derives graph embeddings to perform political media bias classification. Evaluated on a labeled corpus of political news, their classifier achieved an accuracy of 0.96, substantially outperforming a text-only BERT-based baseline that achieved 0.84 accuracy under the same experimental setup. Importantly, this comparison

[108] Majali et al., 2025 "BiasLens: Graph Machine Learning Approach for Media Bias Detection"

was conducted within a proprietary dataset rather than a widely used benchmark such as BASIL or BABE, which limits direct comparability but nonetheless provides strong evidence that incorporating relational context can significantly enhance bias detection performance in controlled settings.

Another alternative perspective reframes media bias detection at the outlet or domain level rather than at the level of individual sentences or documents. Rönnback et al. [155] proposed a large-scale, data-driven approach to estimating political bias across hundreds of thousands of web domains. Instead of relying on manually labeled article-level datasets, their method leverages machine learning techniques to infer political leaning by analyzing multiple forms of bias expression, including topic selection and coverage patterns, which are often ignored in text-centric models. Their approach enables scalable estimation of outlet-level political bias with high accuracy and provides explanatory signals for why a particular outlet is assigned a given ideological position. While this formulation does not directly align with sentence or document-level benchmarks, it addresses a complementary and practically relevant dimension of media bias that traditional classifiers are not designed to capture.

[155] Rönnback et al., 2025 "Automatic large-scale political bias detection of news outlets"

Sentiment-driven and information-theoretic approaches have also been explored as alternatives to supervised text classification. Quijote et al. [137], for example, combined sentiment analysis based on SentiWordNet with an inverse reinforcement learning model to study political bias in Philippine news media. Their system achieved an accuracy of 0.89 when distinguishing biased patterns across outlets, illustrating that sentiment distributions can serve as a proxy for bias under certain assumptions. Similarly, Aires et al. [2] employed Shannon entropy to model lexical distributions as probability mass functions associated with sources and bias classes, using distributional dissimilarities as input to a classifier. By grounding bias representation in information-theoretic principles, this approach offers a mathematically principled way to quantify lexical divergence between outlets, albeit without modeling deeper contextual or discourse-level phenomena.

[137] Quijote et al., 2019 "Bias detection in Philippine political news articles using SentiWordNet and inverse reinforcement model"

[2] Aires et al., 2020 "An Information Theory Approach to Detect Media Bias in News Websites"

Unsupervised and weakly supervised formulations have been proposed to address scenarios where labeled data is scarce or unavailable. Rawat and Vadivu [140] applied clustering techniques such as K-means, PCA, and DBSCAN to group Indian political news articles based on sentence-level sentiment scores, deriving outlet-specific bias reports without relying on gold-standard annotations. In a related vein, Arruda et al. [11] framed media bias as an outlier detection problem, distinguishing between selection bias, coverage bias, and statement bias by identifying statistically significant deviations in fact coverage across outlets. These approaches highlight the multidimensional nature of media bias and emphasize that not all bias manifestations can be effectively captured through supervised classification alone.

[140] Rawat and Vadivu, 2022 "Media Bias Detection Using Sentimental Analysis and Clustering Algorithms"

[11] Arruda et al., 2020 "Analysing bias in political news."

Overall, alternative methods in media bias detection broaden the methodological landscape by addressing dimensions of bias that extend beyond sentence-level linguistic classification. Graph-based models, outlet-level estimation, unsupervised clustering, multi-bias frameworks, and LLM-assisted annotation each target specific limitations of mainstream approaches, whether related to scalability, context modeling, or annotation cost. However, their diversity also exacerbates the lack of standardization in task definitions, datasets, and evaluation protocols; a challenge explicitly highlighted in recent systematic reviews [30]. This methodological heterogeneity underscores the need for unified frameworks that can accommodate multiple bias manifestations while preserving transparency, comparability, and epistemic clarity.

[30] Castillo-Campos et al., 2025, "Automated Detection of Media Bias Using Artificial Intelligence and Natural Language Processing: A Systematic Review"

Comparative analysis

The study of media bias detection has evolved considerably over the years, with various approaches being developed, each bringing its own strengths and challenges. This section aims to provide a comprehensive analysis of the different methodologies discussed in this section, drawing conclusions about their relative effectiveness and identifying key trends in the field.

Traditional approaches to media bias detection, including linguistic-based and reported speech-based methods, laid the groundwork for more advanced techniques. While these

approaches were highly interpretable and computationally efficient, they were often limited by their dependence on handcrafted features. This reliance meant that their success was closely tied to the quality of the features selected, which often required extensive domain expertise and significant preprocessing efforts. Despite these limitations, these methods provided valuable insights into the mechanics of media bias and helped establish a foundation for subsequent research.

Linguistic-based methods, which focus on analyzing lexical, syntactic, and semantic features, were particularly effective in capturing surface-level biases, such as word choice and sentence structure. However, their performance was often constrained by the complexity of the linguistic features they could model, making it challenging to detect more nuanced forms of bias. Reported speech-based methods offered a different perspective by analyzing how different media outlets quoted speakers and entities. These methods excelled in revealing systematic biases in quoted speech, providing a deeper understanding of how media outlets could frame information to align with specific narratives. However, these methods were also computationally intensive and often struggled with the ambiguity inherent in natural language.

The advent of Deep Learning (DL), particularly with the development of RNNs and LSTMs networks, marked a significant advancement in automatic media bias detection. These models could automatically learn feature representations from text, significantly reducing the need for manual feature engineering. RNNs, with their ability to model sequential data, proved effective in capturing the dependencies between words in a sentence, thus providing a more nuanced understanding of bias. However, RNNs were not without their limitations. Issues such as the vanishing gradient problem limited their ability to model long-term dependencies, which was partially mitigated by the introduction of LSTMs. Attention mechanisms further enhanced the performance of RNNs by allowing the model to focus on specific parts of the input sequence, improving their ability to detect bias in more complex text structures.

The introduction of transformer-based models, such as BERT and its variants [154], revolutionized the field by offering a new way to process and analyze text. Unlike RNNs, transformers do not rely on sequential data processing; instead, they use a self-attention mechanism to consider the entire context of a sentence simultaneously. This approach significantly improved the ability of models to detect subtle biases in text, leading to state-of-the-art performance in various NLP tasks, including media bias detection. Encoder-only models like BERT were particularly effective in this regard, as they could consider the context from both the left and right sides of a word, providing a more comprehensive understanding of language. Studies have consistently shown that transformer-based models outperform traditional methods and RNNs in media bias detection tasks, particularly in their ability to handle complex, context-dependent language.

[154] Rogers et al., 2021, "A primer in BERTology: What we know about how BERT works"

However, not all transformer-based models have shown equal effectiveness. Decoder-only models, such as GPT, which predict the next word in a sequence based only on the preceding context, have generally underperformed compared to encoder-only models [151, 136, 21]. Although these models have been successful in various NLP applications, their limited contextual understanding makes them less suitable for tasks like automatic media bias detection, where understanding the full context of a sentence is crucial. Studies have shown that while decoder-only models can excel in specific, controlled settings, they often struggle with generalization and contextual accuracy, reinforcing the need for more sophisticated models like BERT for effective media bias detection.

[151] Rodrigo-Ginés et al., 2023, "Identifying Media Bias beyond Words: Using Automatic Identification of Persuasive Techniques for Media Bias Detection"

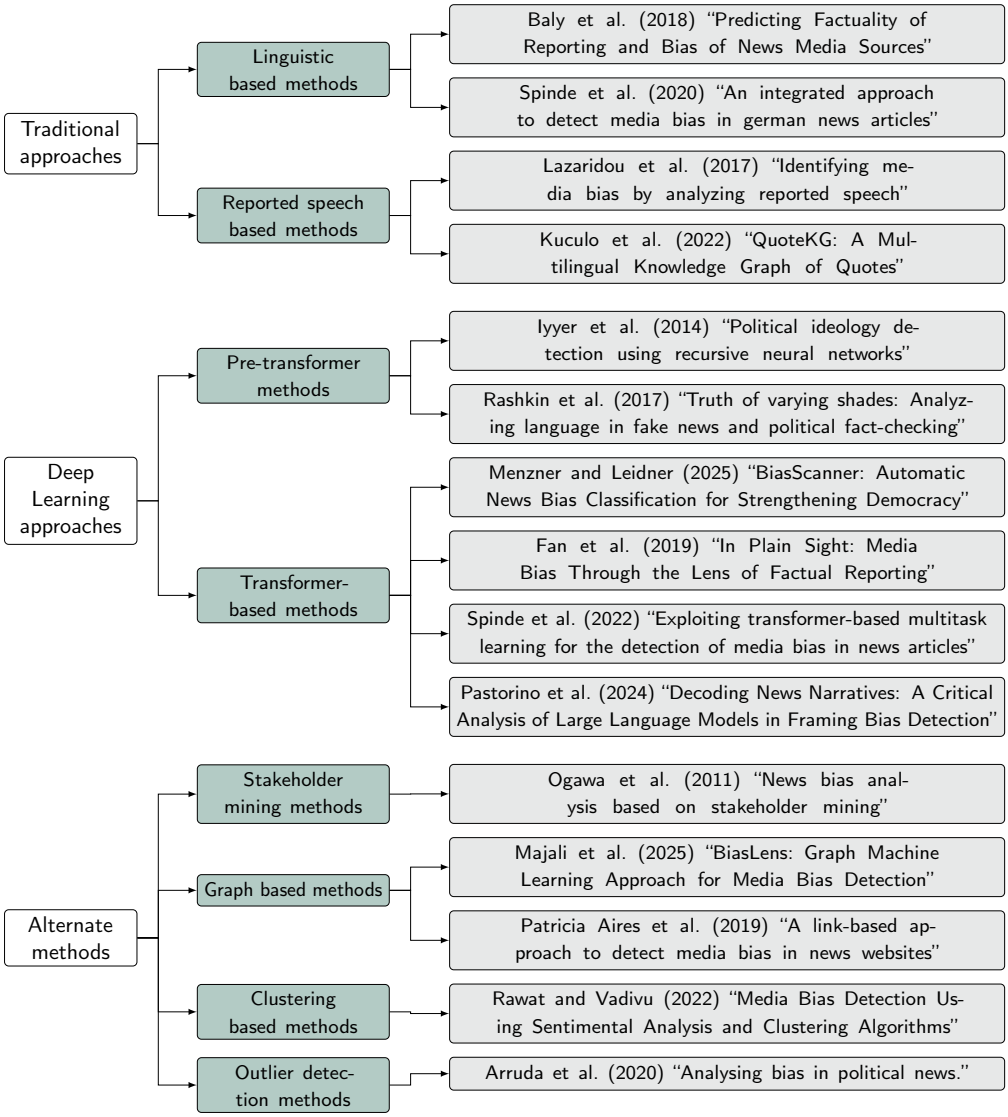
[136] Qorib et al., 2024, "Are Decoder-Only Language Models Better than Encoder-Only Language Models in Understanding Word Meaning?"

[21] Benayas et al., 2024, "A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and performance"

Lastly, alternative methods, such as stakeholder mining, sentiment analysis combined with the Inverse Reinforcement Model, and graph-based approaches, offer unique perspectives on media bias detection. These methods focus on different aspects of bias, such as the relationships between stakeholders in news stories or the structural connections between media outlets. While these approaches provide valuable insights, they often require significant computational resources and are not as widely applicable as more general methods like transformers. Nonetheless, they contribute to a more comprehensive understanding of media bias by exploring dimensions that traditional and DL methods may overlook.

The evolution of media bias detection techniques reflects a broader trend in NLP towards

FIGURE 2.3: Overview of reference articles in media bias detection, organized by the method used.



models that can automatically learn and adapt to the complexities of natural language. Traditional methods, while foundational, are increasingly being supplemented, and in many cases replaced, by deep learning models that offer greater accuracy and flexibility. Among these, transformer-based architectures, particularly encoder-only models such as BERT, have emerged as the most effective tools for detecting media bias, offering a robust balance of performance and contextual understanding.

However, while encoder-only models currently achieve state-of-the-art results, the growing attention to decoder-only architectures, especially large-scale models fine-tuned or instruction-tuned for specific tasks, suggests that this landscape may shift in the near future. Given their scale and adaptability, it is likely only a matter of time before decoder-based models surpass their encoder counterparts in this domain as well.

The overview of reference articles in media bias detection, organized by methodological approach (see Figure 2.3), illustrates the progression and diversification of techniques in the field, providing a roadmap for future research and application.

A systematic comparison of the approaches reviewed in this section is complicated by the lack of shared benchmarks and standardized evaluation protocols. Studies differ not only in their modeling paradigms but also in their task formulations (binary vs. multi-class, sentence vs. document level), evaluation metrics (accuracy, F1-score, precision, recall, or qualitative analysis), and dataset choices, which often reflect different operationalizations of media bias. As a result, reported performance figures are rarely directly comparable across studies. This heterogeneity underscores a broader challenge in the field: the absence of a unified evaluation

framework that would allow meaningful cross-method comparison and systematic tracking of progress.

2.2.3 *Existing resources*

The state of the art reviewed in the previous section shows that automatic media bias detection has been addressed through a wide range of computational paradigms, from feature-based models relying on handcrafted linguistic features to deep learning systems and, more recently, large language models. However, beyond architectural and algorithmic choices, these approaches are fundamentally constrained by the datasets on which they are trained and evaluated. The quality, scope, and design assumptions of these resources shape not only model performance but also the very definition of what is being detected.

In practice, the formulation of the detection task, the granularity of the labels, the unit of annotation (whether document, sentence, or span), and even the specific types of media bias that can be captured are largely dictated by the available resources. As a result, differences in dataset construction, annotation protocols, and labeling conventions often explain discrepancies in reported performance as much as, or more than, the choice of modeling technique itself. This tight coupling between methods and resources also plays a central role in the generalization problems identified throughout the literature, as models trained on one resource frequently fail to transfer their learned representations to others.

For this reason, a systematic understanding of existing datasets is essential to contextualize the results reported in the state of the art and to identify the gaps that motivate the contributions of this thesis. This section therefore provides a comprehensive overview of the most relevant resources for automatic media bias detection, analyzing their scope, annotation schemes, intended use cases, and known limitations. The main characteristics of these datasets are summarized in Table 2.3 and visually represented in Figure 2.2.3, serving as a reference point for the methodological and experimental decisions discussed in the subsequent chapters.

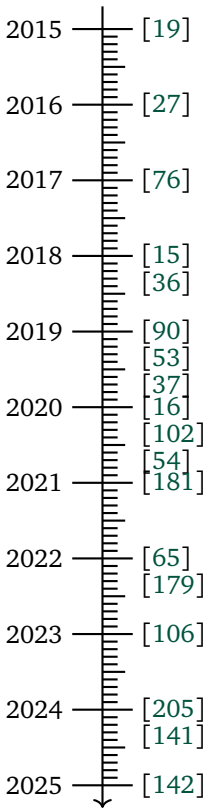
TABLE 2.3: Overview of media bias detection datasets, including annotation granularity, number of instances, language, label types, and availability.

Dataset	Level	Instances	Lang.	Labels	Access	Ref.
Lang. of Framing	Sentence	74 art.	EN	Framing	Request	[19]
Fair & Balanced	Article	10,502	EN	Binary (L/R)	Public	[27]
Bias Flipping	Article	4,759	EN	Binary (L/R)	Public	[36]
Media Source Bias	Art./Media	1,066 art.	EN	Ternary	Public	[15]
Hyperpartisan	Article	754k	EN	Binary	Public	[90]
BASIL	Sent./Art.	300 art.	EN	Binary+spans	Public	[53]
Crisis in Ukraine	Article	2,245	Multi	Pro-W/R/N	Public	[37]
Multidimensional	Sentence	2,057	EN	Multi-type	Request	[54]
Media Profiling	Art./Media	1,066	EN	Ternary	Public	[16]
Biased-Sents	Sentence	1,235	EN	Binary	Public	[102]
MBIC	Sentence	1,700	EN	Binary+words	Public	[181]
NELA-GT-2021	Article	1.8M	EN	Outlet-level	Public	[65]
BABE	Sentence	3,700	EN	Binary+types	Public	[179]
Chinese Media	Art./Sent.	3,500	ZH	Multi-level	Request	[106]
FIGNEWS	Article	50k	Multi	Binary+stance	Public	[205]
BEADs	Article	115k	EN	Multi-dim.	Public	[141]

One of the earliest systematic attempts to operationalize media bias is the Language of Framing in Political News dataset introduced by Baumer et al. [19]. At this initial stage, bias is not formalized as a supervised learning label, but treated as a qualitative phenomenon expressed through framing devices embedded in language. This framing-centric perspective is valuable as a conceptual starting point, yet it also exposes an early and persistent limitation in the literature: the gap between rich qualitative notions of bias and the operational requirements of computational modeling.

The dataset is annotated at the sentence level, with expert annotators highlighting words and phrases that contribute to framing effects. Importantly, these annotations are not organized into a standardized set of bias categories, and the work does not provide an explicit inter-annotator agreement analysis. While the dataset establishes that subtle rhetorical choices (e.g., metaphor, evaluative adjectives, agent selection) can shift interpretation, its small scale (74 articles) and descriptive labeling make it difficult to use as a robust benchmark for model comparison or for studying generalization.

FIGURE 2.4: Chronological timeline of representative media bias detection datasets (2015–2025).



[19] Baumer et al., 2015 "Framing Annotation Data for News Articles"

A second major design pattern in the literature reframes bias as a document-level ideological attribute, enabling classical supervised learning at scale. This shift is exemplified by the Fair and Balanced dataset introduced by Budak et al. [27], which operationalizes bias as a binary classification task (democratic-leaning vs. republican-leaning) at the article level. The dataset was constructed by crawling over 800,000 articles from 15 major U.S.A. outlets (2013) and manually annotating a filtered subset of 10,502 political articles.

[27] Budak et al., 2016 "Fair and balanced? Quantifying media bias through crowdsourced content analysis"

This design facilitates large-scale modeling and outlet-level comparison, but it comes with a strong conceptual cost: heterogeneous mechanisms of bias (lexical evaluation, framing, omission, source selection, etc.) are collapsed into a single ideological axis. As a result, models trained on such resources can achieve high performance while remaining largely opaque with respect to the linguistic and rhetorical mechanisms that produce the label. This label collapse also makes cross-dataset evaluation problematic, since "left vs. right" at the outlet level is not equivalent to statement-level bias or fine-grained manifestations in other datasets.

A related limitation becomes visible in datasets that inherit labels from external rating sources. For example, Chen et al. [36] introduced a dataset for bias inversion in news headlines using labels derived from AllSides outlet ratings. While this enables scale and supports research on debiasing, the supervision is inherently coarse-grained: labels are typically assigned at the outlet level and projected onto headlines. This creates a mismatch between what is labeled and what is learned, and it can introduce systematic noise when individual items deviate from the average stance of their source.

[36] Chen et al., 2018 "Learning to flip the bias of news headlines"

In parallel, a family of datasets emphasizes lexical choice as a concrete and annotatable signal of bias. Lim et al. [103] contrasts terminology across outlets reporting on the same events, illustrating how near-synonymous expressions encode ideological positioning (e.g., "freedom fighter" versus "rebel"). This perspective is useful because it ties bias to observable linguistic decisions. However, its scope is narrow: lexical contrast captures only a subset of bias mechanisms and does not directly model higher-level discourse strategies such as argumentation, narrative structure, or selective evidence.

[103] Lim et al., 2018 "Understanding Characteristics of Biased Sentences in News Articles"

The follow-up Biased-Sents-Annotation dataset [102] moves toward sentence-level supervision with binary labels and multiple annotators, reporting agreement statistics. Even so, the binary formulation retains a key limitation: it aggregates distinct manifestations under a single positive label, making it difficult to learn, evaluate, or explain which mechanisms actually triggered the bias judgment. In other words, these datasets move closer to supervision, but not necessarily closer to a mechanism-aware modeling of bias.

[102] Lim et al., 2020, "Annotating and analyzing biased sentences in news articles using crowdsourcing"

A major turning point in dataset design concerns scale and temporal coverage, epitomized by the News Landscape (NELA) datasets [76, 65]. NELA prioritizes longitudinal analysis by aggregating millions of articles across years and sources, assigning labels at the outlet level inherited from Media Bias/Fact Check ratings. This design supports ecosystem-level analyses and time-based studies, but it also introduces substantial label noise at the instance level: article- and sentence-level content can diverge from outlet-level labels, and the resulting supervision often captures outlet identity and topical regularities more than textual manifestations of bias. Consequently, NELA-style resources are useful for large-scale correlational studies but are weak foundations for fine-grained, linguistically grounded bias characterization.

[76] Horne et al., 2018, "Sampling the news producers: A large news and feature data set for the study of the complex media landscape"

[65] Gruppi et al., 2022, *NELA-GT-2021: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles*

Other datasets focus on extreme ideological polarization without redefining the underlying labeling problem. The SemEval-2019 Task 4 dataset [90] targets hyperpartisan news and thus emphasizes overt partisan rhetoric. While this is valuable for studying polarization, it narrows the phenomenon to its most salient cases and does not address the subtler forms of bias that dominate mainstream reporting.

[90] Kiesel et al., 2019, "Semeval-2019 task 4: Hyperpartisan news detection"

A distinct and more promising family of datasets attempts to capture bias at the statement level with annotations tied to linguistic phenomena. The BASIL dataset [53] marks a shift toward controlled comparisons by annotating sentences from ideologically opposed outlets covering the same events, enabling analysis of framing and lexical differences while holding the underlying topic constant. This direction is extended by resources such as NewsWCL50 [72], MBIC [181], and BABE [179], which provide multi-class bias labels and report inter-annotator agreement.

[53] Fan et al., 2019, "In Plain Sight: Media Bias Through the Lens of Factual Reporting"

[72] Hamborg et al., 2019, "Automated identification of media bias in news articles: an interdisciplinary literature review"

[181] Spinde et al., 2021, "MBIC–A Media Bias Annotation Dataset Including Annotator Characteristics"

[179] Spinde et al., 2022, "Neural Media Bias Detection Using Distant Supervision With BABE–Bias Annotations By Experts"

However, even within this "fine-grained" family, two limitations remain structural. First, fine granularity is often achieved through relatively small datasets, restricting coverage and making model estimates unstable under domain shift. Second, the label taxonomies used are typically limited to a small set of phenomena and do not provide a holistic inventory of mechanisms spanning lexical, rhetorical, psychological, and informational strategies. As a result, models trained on these datasets may detect certain surface manifestations well, but they still do not support comprehensive mechanism-level characterization.

Beyond U.S.A.-centric ecosystems, some datasets extend annotation to geopolitical and cross-cultural contexts (e.g., Ukraine-related resources [37, 54], Chinese COVID-19 media [106], and multilingual resources such as FIGNEWS [205]). These datasets are essential to reduce Western-centric assumptions, yet they frequently introduce additional sources of heterogeneity: different media systems, different rhetorical conventions, and different annotation operationalizations. This further complicates comparability and amplifies the generalization problem when models trained in one setting are evaluated in another.

Finally, recent datasets such as BEADs [141] expand the scope toward multi-dimensional social harms, combining political bias with toxicity-related categories under explicit ethical constraints. While socially relevant, these broadened label spaces also increase conceptual entanglement: bias can become conflated with offensiveness or normative judgments, making it harder to isolate what is specifically "media bias" versus adjacent phenomena.

Taken together, existing datasets do not form a simple trajectory of progress. Instead, they reflect repeated trade-offs between scale and annotation fidelity, between ideological proxies and textual mechanisms, and between narrow fine-grained labels and genuinely holistic characterization. Critically, none of these resources explicitly operationalize perspectivism as a first-class property of the data, nor do they model disagreement as an informative signal rather than noise to be eliminated. These limitations, together with the evidence for weak cross-dataset generalization, directly motivate the design choices underlying the dataset introduced in this thesis.

[37] Cremisini et al., 2019, "A challenging dataset for bias detection: the case of the crisis in the Ukraine"

[54] Färber et al., 2020, "A Multidimensional Dataset for Analyzing and Detecting News Bias based on Crowdsourcing"

[106] Lin et al., 2025, "Data-augmented and retrieval-augmented context enrichment in chinese media bias detection"

[205] Zaghouani et al., 2024, "The FIGNEWS Shared Task on News Media Narratives"

[141] Raza et al., 2024, "BEADs: Bias Evaluation Across Domains"

An ideal frame for media bias comparison

To concretise the state of the art and to provide a basis for the empirical work of Chapter 4, this thesis selects five high-quality and publicly available datasets as benchmarking resources: MBIC, the Ukraine Conflict Dataset, FIGNEWS, BEADs, and MBBMD. Together they span different linguistic contexts, political perspectives, and annotation paradigms, enabling a comprehensive and realistic assessment of model behaviour. One representative example per dataset is given below to illustrate its annotation style and unit of analysis.

Importantly, these resources are not used solely for isolated benchmark comparisons, but are systematically incorporated into the experimental evaluation of the proposed systems. They are employed to analyze in-domain performance, cross-dataset generalization, and robustness under distributional shifts. The experimental protocols and results of the cross-dataset and cross-domain generalisation study conducted across these five datasets are presented in Section 4.1, where they are used to expose the strengths and limitations of different modelling strategies when applied beyond their training distributions.

The first dataset considered, the Media Bias Identification Corpus (MBIC) [181], focuses on detecting subtle linguistic cues that contribute to biased reporting. It consists of individual sentences labeled as biased or non-biased, with annotations emphasizing phenomena such as emotionally charged expressions, framing strategies, and evaluative language. For instance, the following example illustrates a case where bias is introduced through the use of metaphor:

> **Instance (sentence):** Tuesday on Hill.TV, 2020 Democratic presidential candidate Marianne Williamson declared President Donald Trump has become a "megaphone" for white nationalism.
>
> **Outlet:** Breitbart
>
> **Topic:** White nationalism
>
> **Label:** Biased

Biased words: *megaphone*

Source: https://www.breitbart.com/clips/2019/07/30/marianne-williamson-trump-is-a-megaphone-for-white-nationalism-in-the-united-states/

[37] Cremisini et al., 2019, "A challenging dataset for bias detection: the case of the crisis in the Ukraine"

In contrast, the Ukraine Conflict Dataset [37] categorizes entire articles based on their political alignment, providing insight into how different geopolitical actors frame the same events. This dataset is particularly valuable for studying how bias manifests at the document level rather than within isolated sentences.

Instance (document): Pro-Russian Hackers Attack Central Election Commission of Ukraine

Publisher: Silicon UK (England)

Event: 2014 Ukrainian Presidential Election

Leaning: Pro-Western

Link: http://www.techweekeurope.co.uk/news/cyberberkut-hackers-attack-central-election-commission-of-ukraine-146180

[205] Zaghouani et al., 2024, "The FIGNEWS Shared Task on News Media Narratives"

Moving beyond traditional news articles, the FIGNEWS dataset [205] expands the analysis to social media, capturing real-time reactions to politically charged events. FIGNEWS is particularly valuable for studying bias in digital discourse, where shorter, opinion-driven statements often reveal stronger ideological leanings.

Instance (social media post): « Hamas is worse than ISIS » Cliquez pour écouter mon interview sur la BBC World

Source Language: French

Label: Biased against Palestine

English translation: "Hamas is worse than ISIS" Click to listen to my interview on BBC World

[141] Raza et al., 2024, "BEADs: Bias Evaluation Across Domains"

To explore the intersection between bias and toxic discourse, the BEADs dataset [141] annotates statements not only for bias but also for elements of hate speech, toxicity, and ideological framing. This dataset is particularly useful for distinguishing between ideological bias and outright harmful content.

Instance (sentence): Clearly, there is far too much speculative about this article... The final ruination of the article is somehow making a blind assertion that it originated in the Americas through black Africans enslaved there.

Dimension: Toxicity

Biased Words: *poorly, enslaved, virulent*

Aspect: Racial

Label: Highly Biased

Finally, the Media Bias Bias-Mitigated Dataset (MBBMD), developed as part of this thesis (described in detail in Section 3.2), presents a hierarchical approach to bias detection, integrating document-level and sentence-level annotations. Unlike prior datasets, MBBMD is designed to capture multiple perspectives, making it a valuable tool for perspectivist media analysis.

Instance (document): Pedro Sánchez es investido presidente y conformará el segundo Gobierno de coalición progresista de la democracia.

Topic: Pedro Sanchez Investiture

Label: Biased (62.5%)

Text: Pedro Sánchez ha sido investido presidente del Gobierno en el Congreso, en primera votación, con una mayoría absoluta de 179 votos... Su reelección da paso a una etapa repleta de incertidumbres en nuestro país que vendrá marcada por la división y el enfrentamiento que propician la Ley de amnistía y los referéndums exigidos por sus socios separatistas.

English translation: Pedro Sánchez has been invested president of the Government in the Congress, in the first vote, with an absolute majority of 179 votes... His re-election is a step full of uncertainty in our country, which will be marked by the division and confrontation that propitiate the amnesty law and the referendums demanded by its separatist partners.

These five datasets, taken individually, each capture one partial facet of media bias: sentence-level linguistic cues (MBIC, BEADs), document-level ideological leaning (Ukraine, MBBMD), short-form social media discourse (FIGNEWS), Spanish-language long-form news (MBBMD), and the intersection of bias with toxicity (BEADs). No existing single resource in the state of the art combines all of these dimensions into one benchmark, and the literature routinely reports results on a single dataset without probing whether those results hold under distributional shift. This thesis takes the opposite route: rather than committing to any one partial definition of media bias, we treat the *union* of these datasets as an approximation of the real-world problem, and we evaluate the state of the art on that union through the cross-dataset protocol of Section 4.1. In that protocol, a model is trained on one dataset and evaluated on the remaining four, so that each configuration tests whether what the model has learned generalises beyond the specific operationalisation of its training data. This setup is, in our view, the most honest available approximation to evaluating media bias detection as a single phenomenon rather than as a collection of dataset-specific tasks.

2.3 Five open problems: why engineering progress has not translated into real-world robustness

Read end to end, the preceding two sections tell two stories at once. The first is encouraging: the field has moved from hand-crafted features and small corpora to large-scale transformers, multilingual models, and LLM-based reasoning in under a decade. The second is sobering: despite this engineering progress, no system reliably detects media bias outside the distribution it was trained on. In-domain F1 scores climb, but cross-dataset F1 scores stagnate or collapse. The gap between the two is not a minor evaluation nuisance; it is evidence of a structural mismatch between how the field formulates the task and how media bias actually operates.

This section identifies five problems that, taken together, explain this mismatch. They are not independent: conceptual fragmentation feeds generalisation failure, which is masked by evaluation opacity, while the single-label fallacy and anglocentrism constrain the data on which every model and every metric depends. Each problem is stated, argued, and connected to a specific contribution developed in the remainder of the thesis.

▸ Problem 1: Conceptual fragmentation and the absence of a shared definition

The first and most basic problem is that the field has never agreed on what it is trying to detect. The literature survey in Section 2.2 returned 118 studies operationalising media bias, and no two studies use the same definition. Some frame bias as a property of *outlets* (an ideological axis inherited from external rating sites), others as a property of *documents* (binary biased/unbiased), others as a property of *sentences* (fine-grained manifestations), and others still as a relational property of *coverage patterns* across outlets. These formulations are not equivalent, and they are not even commensurable: a model trained to recognise an outlet-level "left/right" label cannot be meaningfully compared to one trained to flag framing devices at the sentence level. The literature treats them as if they were variants of the same task, but they are, in fact, different tasks that happen to share a word.

This conceptual fragmentation has two consequences. First, reported state-of-the-art numbers are structurally incomparable: an F1 of 0.81 on a dataset with outlet-inherited labels and an F1 of 0.43 on a dataset with sentence-level annotations by trained human coders are not measuring the same phenomenon, and averaging or ranking them misleads more than it informs. Second, and more seriously, the absence of a shared definition prevents the field from cumulating knowledge: each paper starts from its own operationalisation, trains on its own resource, and reports its own numbers, making it nearly impossible to say whether the field is actually improving at detecting bias or merely at fitting new datasets. Chapter 3 addresses this problem directly by proposing a hierarchical taxonomy that decomposes media bias into bias types (by communicative intention and contextual factors) and bias manifestations (observable linguistic and rhetorical realisations), providing a shared vocabulary that can be instantiated in multiple datasets without collapsing into a single coarse label.

▸ **PROBLEM 2: WEAK CROSS-DATASET GENERALISATION**

A second, and empirically well-documented, structural problem is the systematic failure of media bias detectors to generalise across datasets. As the cross-dataset experiment reported in Section 4.1 will show in detail, a classifier that achieves F1 $\approx$ 0.75 on its in-domain test set can collapse to F1 $\approx$ 0.14 when evaluated on a different bias dataset, even when both datasets purport to label the same phenomenon. This is not an isolated observation: analogous drops have been reported whenever authors have bothered to run the experiment [151, 201]. The field-wide default, however, is to report only in-distribution metrics, which makes the problem invisible.

[151] Rodrigo-Ginés et al., 2023, "Identifying Media Bias beyond Words: Using Automatic Identification of Persuasive Techniques for Media Bias Detection"

[201] Wessel and Horych, 2024, "Beyond the surface: Spurious cues in automatic media bias detection"

We argue that this is not a model deficiency that can be patched with more data or better architectures, it is a symptom of Problem 1. If each dataset implements a subtly different definition of media bias, then a model trained on one dataset cannot be expected to transfer to another, because what it has learned is not "media bias" in any general sense but rather "whatever labelling convention this particular corpus used". The strong transfer observed between some pairs of datasets is, on this reading, an artefact of shared stylistic or topical regularities, not of a shared underlying phenomenon. Treating generalisation failure as a central diagnostic rather than a secondary concern is one of the design principles of this thesis: Chapter 4 explicitly evaluates the proposed system under cross-dataset conditions and factors the gains into a dataset-versus-system ablation.

▸ **PROBLEM 3: THE GROUND-TRUTH PROBLEM**

A third structural problem concerns the nature of annotation itself. Virtually every dataset in the literature produces a single label per instance, typically by majority vote over a small number of annotators, and then treats that label as ground truth. This convention imports an assumption that is demonstrably false for media bias: that reasonable, informed readers should agree on whether a given article is biased. The psychological and philosophical discussion of Section 1.3 already made clear why this assumption fails. Perception of bias is cognitively mediated, ideologically situated, and shaped by motivated reasoning, naive realism, and prior beliefs. Two well-intentioned annotators with different political leanings can read the same article and arrive at different, internally consistent judgments, not because one is wrong, but because "biased" is not a property of the text alone.

Collapsing this disagreement into a single label does not recover a hidden truth; it performs a political operation, privileging the perspective of whichever group is overrepresented in the annotator pool (typically educated, English-speaking, Western participants recruited on crowdsourcing platforms). Worse, because this operation is invisible in the final dataset, downstream models inherit it without acknowledgment, and their predictions are subsequently interpreted as objective. The Learning With Disagreements (LeWiDi) framework [190] is one of the few lines of work to take this problem seriously, and the MBBMD dataset introduced in Section 3.2 adopts a perspectivist annotation protocol that collects ideological self-identification for every annotator, preserves disagreement rather than collapsing it, and aggregates through protocols that treat conflicting readings as informative rather than noisy.

[190] Uma et al., 2021, "Learning from disagreement: A survey"

▸ **Problem 4: Anglocentrism and the narrowness of the benchmark ecosystem**

A fourth structural problem is that the datasets on which the field reports progress are drawn from a remarkably narrow slice of the world's media. An overwhelming majority of existing benchmarks are English-language, focus on U.S.A. domestic politics (often a few high-visibility events such as presidential elections or specific conflicts), and source annotations from crowdsourcing platforms dominated by Western participants. The few non-English resources that do exist (e.g., Hindi, Chinese, or multilingual datasets such as FIGNEWS) are usually small, domain-specific, and evaluated in isolation.

The consequence is that "state of the art" in media bias detection is, in practice, state of the art in detecting the framing of U.S.A. political news in English as perceived by crowdworkers. Whether the same models work on Spanish coverage of the Catalonia emergency, on Irish coverage of the Sinn Féin referendum, or on Argentine coverage of presidential elections, is largely unknown, and in the few cases where it has been tested the answer is typically negative. This narrowness is not a cosmetic limitation: the linguistic, rhetorical, and ideological conventions of different media ecosystems differ substantially, and a model that works on one does not necessarily learn anything portable about the others. MBBMD (Section 3.2) is a deliberate response to this anglocentrism: it is built from Spanish-language coverage of ten real-world topics spanning Spanish, Irish, Argentine, Israeli–Palestinian, and other media landscapes, and it is annotated with explicit perspectivist metadata that allows readers from different ideological backgrounds to be modelled separately.

▸ **Problem 5: Evaluation opacity and the absence of a shared protocol**

The fifth and final structural problem is methodological rather than conceptual: the field lacks a shared evaluation protocol. Studies differ in task formulation (binary vs. multi-class, sentence vs. document level), in metric choice (accuracy, macro-F1, weighted-F1, Matthews correlation, qualitative analysis), in the definition of the evaluation split (random, chronological, outlet-held-out, topic-held-out), and in whether they report results on any shared benchmark at all. Many papers evaluate only on datasets they themselves constructed, making claims of improvement unfalsifiable. This is compounded by a widespread tendency to tune hyperparameters on the test set, to report single-run numbers without confidence intervals, and to compare against baselines that were never properly reimplemented.

The result is an evaluation ecosystem in which reported numbers cannot be trusted to track genuine progress. This is not a unique problem of media bias detection (the broader NLP community has begun to reckon with it) but the problem is especially acute here, because the conceptual fragmentation described as Problem 1 provides an easy alibi: if the task is ill-defined, any deviation in evaluation can be framed as "studying a different aspect of bias". The experiments in Chapter 4 therefore adopt a deliberately stringent protocol: cross-dataset evaluation across multiple benchmarks, multiple random seeds with reported variance, and a system-versus-dataset ablation designed to separate the contribution of the proposed model from the contribution of the proposed dataset.

This page intentionally left blank.

3

From diagnosis to operational definition: a hierarchical taxonomy and a perspectivist dataset

"Reality, then, is offered in individual perspectives. What is in the background for one person is in the foreground for another."
—José Ortega y Gasset

Chapter 2 closed with a diagnosis: the field of computational media bias detection is held back by conceptual fragmentation, single-label annotation, anglocentrism, and weak generalisation. This chapter presents the first two contributions of this thesis, designed as a coordinated response to that diagnosis. Taxonomy and dataset are presented together because they are two sides of the same conceptual move: the taxonomy defines what we are trying to detect, and the dataset instantiates that definition in annotated data that can support both perspectivist modelling and cross-domain evaluation. Neither is useful without the other; a taxonomy without a dataset remains an abstract scheme, and a dataset without a taxonomy is just another corpus with a private labelling convention. Section 3.1 introduces the hierarchical taxonomy: it organises media bias into *types* (by communicative intention and by contextual factors) and into concrete *manifestations* (by observable linguistic and rhetorical realisation), giving the field a shared vocabulary that can be instantiated across datasets without collapsing distinct phenomena into a single coarse label, and (as we argue in Section 3.1 itself) does so in a form that subsumes and extends the partial inventories of the state of the art. Section 3.2 then presents MBBMD (Media Bias Bias-Mitigated Dataset), a Spanish-language, perspectivist, hierarchical dataset that operationalises the taxonomy, treats annotator disagreement as signal rather than noise, and deliberately moves outside the narrow U.S.A./English benchmark ecosystem that dominates the literature. The empirical use of both (in cross-dataset experiments, in the design of a hierarchical perspectivist system, and in the attribution of improvements to dataset versus system) is deferred to Chapter 4.

3.1 A HIERARCHICAL TAXONOMY OF MEDIA BIAS

This section presents the first contribution of this thesis as a direct response to Problem 1 of Section 2.3: a hierarchical taxonomy that gives the field a shared and computationally tractable vocabulary for describing media bias. The remainder of this section develops the taxonomy in detail and motivates it against the state of the art surveyed in Chapter 2: Subsections 3.1.1 and 3.1.2 present the two upper-level criteria, and Subsection 3.1.3 describes the seventeen manifestations along with illustrative examples and a comparison with the prior taxonomies of the field.

The taxonomy is organised in two levels. The upper level distinguishes bias *types* according to two complementary criteria: the communicative intention of the author or outlet (which produces the distinction between spin bias and ideology bias) and the contextual factors that shape how information is selected, framed, and delivered (which produces coverage bias, gatekeeping bias, and statement bias). The lower level specifies seventeen concrete *manifestations* that instantiate each type through observable linguistic and discursive patterns, such as sensationalism, biased word choice, omission of source attribution, or picture selection. Crucially, some manifestations appear under more than one type, reflecting the fact that the same linguistic strategy may serve different bias functions depending on context and intent.

▸ FROM DEFINITIONS IN THE LITERATURE TO A STRUCTURED TAXONOMY

A more concrete way to read the taxonomy is as the structured operationalisation of what the literature surveyed in Chapter 2 had only been listing informally. Table 2.1 of Section 2.2.1

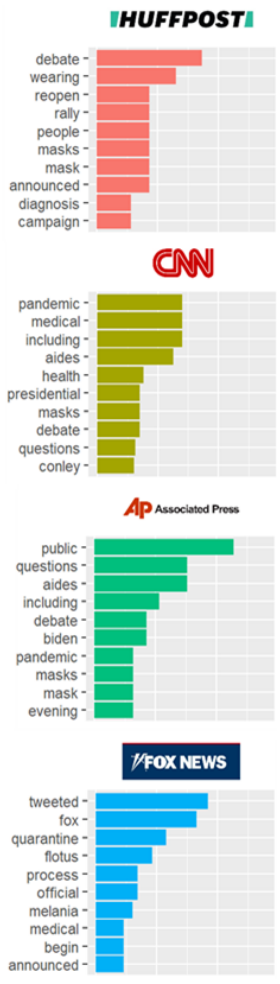
extracted, from the most influential definitions of media bias in the field, three recurring conceptual features: (1) media bias is *slanted*, that is, it reflects a deliberate or systematic tilt away from a neutral baseline rather than an arbitrary error; (2) media bias is *sustained or frequent*, that is, it manifests as a tendency rather than as an isolated event; and (3) media bias is *produced by creating "memorable" (spin) stories*, that is, it operates not only by what is said but by how it is dramatised, foregrounded, or made attention-grabbing. These three features were never organised into a system in the literature; they were enumerated by different authors with overlapping and incompatible vocabularies. The hierarchical taxonomy proposed here gives each of them a dedicated locus. Feature (1) is what the upper-level distinction between *ideology bias* and the contextual biases captures: an article exhibits an ideologically slanted tilt (ideology bias) or an editorially slanted selection of which content to surface (gatekeeping bias) or which to emphasise (coverage bias). Feature (2) is what the cross-document and cross-article scope of coverage and gatekeeping biases captures: those manifestations are by construction patterns that only become visible when an article is read against the broader output of an outlet or of the news ecosystem. Feature (3) is what *spin bias* captures, with sensationalism, opinion-as-fact and word-choice manifestations as its concrete linguistic realisations. The proposed taxonomy is therefore not a competing inventory of the literature's terms; it is the structure that explains why those terms recur and how they relate, and that is the sense in which it constitutes a "shared and computationally tractable vocabulary" rather than yet another labelling convention.

[118] Mullainathan and Shleifer, 2002, *Media bias*

[117] Mokhberian et al., 2020, "Moral framing and ideological bias of news"

[109] Mastrine et al., 2022, *How to spot 16 types of media bias*

FIGURE 3.1: Distribution of most common words in articles about a U.S.A. presidential debate across four media outlets.



3.1.1 *Media bias according to communicative intention*

One of the most prevalent ways to classify media bias is by examining the intention of the author or the media outlet. The intention behind biased reporting often reveals the underlying motivations that drive the dissemination of certain viewpoints over others. Within this classification, media bias can be broadly divided into two primary types: spin bias and ideology bias.

▸ SPIN BIAS

Spin bias, or rhetorical bias, occurs when journalists strive to create more engaging or memorable stories by using emotionally charged language, exaggerating details, or selectively presenting facts [118]. This form of bias is particularly prevalent in sensationalist media, where the goal is to capture the audience's attention through dramatic or provocative storytelling. Examples of spin bias include "clickbait" headlines or news reports that emphasize conflict or drama over substantive content.

For instance, in political reporting, a journalist might describe an election result as a "bloodbath" to invoke a stronger emotional response, even if the situation is less dire. This type of coverage can distort the audience's understanding of events, leading to misconceptions or a skewed perception of reality.

▸ IDEOLOGY BIAS

Ideology bias, also known as stance or framing bias [117], occurs when journalists present information in a way that aligns with a particular ideological perspective. This can skew the audience's understanding of issues, as the information is framed to support a specific viewpoint. Ideology bias is commonly observed in political reporting but is not limited to the traditional left-right spectrum. It can also involve other dichotomies, such as authoritarian versus libertarian or secular versus religious perspectives.

Websites like AllSides [109] highlight various types of ideological bias prevalent in major U.S.A. media outlets, including biases related to urban versus rural or elitist versus populist viewpoints. These biases can significantly influence how audiences perceive and understand news events.

In practice, it often operates under the guise of neutrality or objectivity. Media outlets might claim to provide balanced reporting while subtly framing stories in a way that supports their underlying ideological agenda. This can make it challenging for audiences to discern

the bias, as the information presented appears to be straightforward reporting, yet is colored by the ideological leanings of the journalist or media organization.

3.1.2 *Media bias according to contextual factors*

Beyond the author's intention, media bias can also be categorized according to the context in which it occurs [161]. The most common contextual biases include coverage bias, gatekeeping bias, and statement bias.

[161] Saez-Trumper et al., 2013, "Social media news communities: gatekeeping, coverage, and statement bias"

▸ **Coverage bias**

Coverage bias involves the selective emphasis on certain topics or entities in media reporting. This type of bias becomes evident when media outlets disproportionately cover specific stories while downplaying or ignoring others. The focus on sensational or negative news stories can create a skewed representation of reality, leading the public to perceive certain issues as more significant or prevalent than they are. For instance, the media may prioritize coverage of crime or disaster events over positive developments, contributing to a heightened sense of fear or insecurity among the audience.

As illustrated in Figure 3.1, different media outlets may choose to highlight distinct aspects of the same event, thereby shaping public understanding in diverse ways depending on which details are emphasized or omitted. In this example, the figure displays the most frequent terms used by four major U.S.A. news outlets (HuffPost, CNN, Associated Press, and Fox News) when covering President Trump's COVID-19 diagnosis and the presidential debate. Although all outlets report on the same event, their lexical focuses diverge considerably: HuffPost foregrounds debate-related and health-protection terms; CNN emphasises pandemic management and medical context; the Associated Press highlights institutional and procedural aspects; and Fox News concentrates on political reactions and official statements. These contrasting topical emphases illustrate how coverage bias does not require altering facts; rather, it operates by allocating attention unevenly across narrative elements. As a result, audiences exposed to different outlets receive systematically different impressions of what is most important, relevant, or concerning about the event, reinforcing agenda-setting effects and contributing to divergent interpretations of a shared reality.

▸ **Gatekeeping bias**

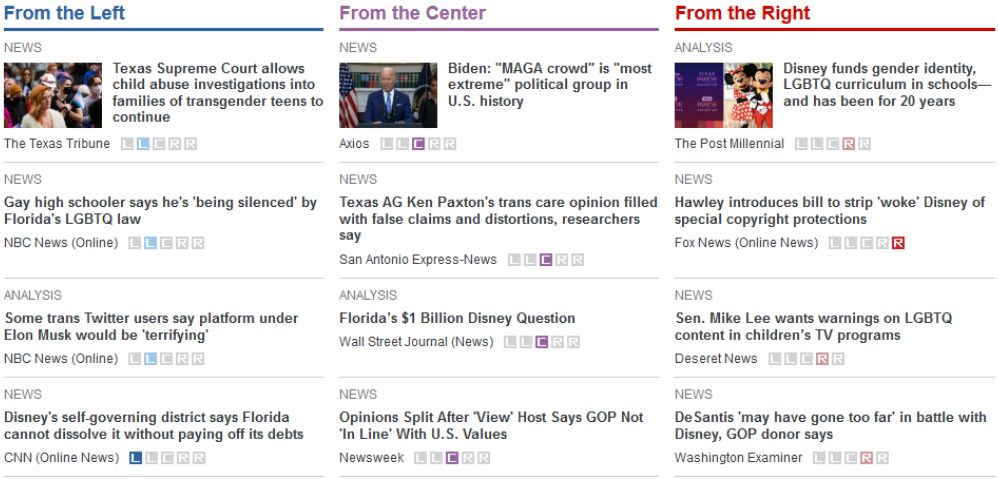
Gatekeeping bias, also known as selection bias or agenda bias, refers to the editorial decisions made by media outlets regarding which stories to publish and which to withhold. This form of bias is significant because it directly affects the public's access to information, potentially leading to a skewed portrayal of reality. Gatekeeping bias often reflects the political, commercial, or ideological preferences of the media outlet, influencing which issues are brought to the forefront and which are left in the background. For instance, a media outlet with a particular political orientation might choose to focus on stories that align with its audience's views, while neglecting or downplaying stories that could challenge those perspectives. This selective reporting can create an echo chamber effect, where the audience is repeatedly exposed to information that reinforces their existing beliefs, as shown in Figure 3.2.

Importantly, gatekeeping bias differs from coverage bias in both scope and mechanism. While coverage bias concerns how an already-selected topic is treated (for instance, whether it is covered frequently, sensationally, or with particular framing choices) gatekeeping bias concerns whether the topic appears at all. Coverage bias operates within the boundaries of existing editorial choices, shaping emphasis, prominence, and narrative angle. Gatekeeping bias, in contrast, defines the boundaries themselves: it determines which events, actors, or issues enter the news agenda. In practice, coverage bias distorts the salience of selected topics, whereas gatekeeping bias distorts the informational universe available to the public. Together, both biases interact to shape not only what audiences think about, but also what they never get the opportunity to think about.

▸ **Statement bias**

Statement bias, also known as presentation bias, occurs when journalists use specific language

FIGURE 3.2: Example of gatekeeping bias, showing how different media outlets select stories to report.



or rhetorical strategies to present information in a way that favors one side of an issue over another. It can subtly influence the audience's interpretation of events, often without the audience being aware of the manipulation. For example, the choice of words like "illegal alien" instead of "undocumented worker" can frame the debate on immigration in a way that elicits a particular emotional response from the audience. Similarly, referring to COVID-19 as the "chinese virus" (as shown in Figure 3.3) can evoke negative associations and fuel xenophobic sentiments.

FIGURE 3.3: Example of statement bias through word choice, using the term "chinese virus" to describe COVID-19.

3.1.3 *Manifestations of media bias*

Foundational works such as Baker et al. [13] and more recent contributions like Hamborg et al. [72] have provided detailed inventories of bias manifestations, but stop short of a unified mechanism-based organisation. The systematic literature review reported in Subsection 2.2.1 surveyed 118 such studies and consolidated more than 40 overlapping bias-related terms into the operational inventory used in this thesis. The taxonomy presented here is the structured form of that inventory: it subsumes the partial lists of Baker et al. [13] and Hamborg et al. [72] as a superset, and brings them under a single mechanism-based organisation rather than competing with them.

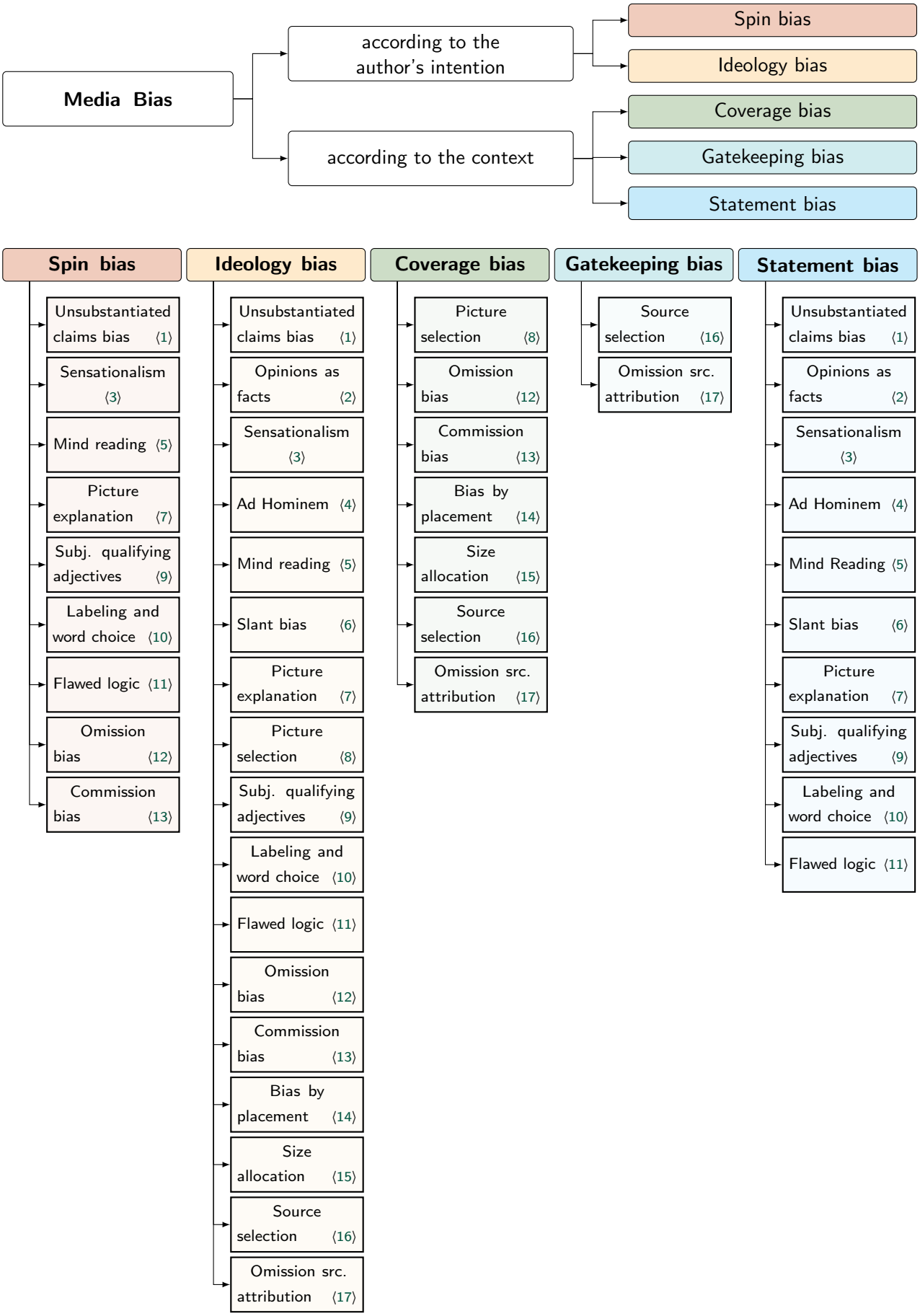
This subsumption matters not only *within* media bias detection but also *across* the broader information disorder ecosystem reviewed in Subsection 2.1.2, where media bias sits alongside misinformation, disinformation, propaganda, and fake news. Many manifestations of the taxonomy (unsubstantiated claims, opinions presented as facts, sensationalism, mind reading, biased word choice, omission of source attribution) are precisely the linguistic mechanisms that misinformation and propaganda exploit when they enter the news ecosystem, so the same vocabulary serves both detection tasks. Figure 3.4 renders the structured form of this organisation: the taxonomy proposed in this thesis (published in 2023 as [149]) adopts a hierarchical structure with two main levels, namely higher-level bias types (defined according to communicative intention and contextual factors) and lower-level bias manifestations (observable linguistic and discursive patterns in news texts).

[13] Baker et al., 1994 *How to identify, expose & correct liberal media bias*

[72] Hamborg et al., 2019 "Automated identification of media bias in news articles: an interdisciplinary literature review"

[149] Rodrigo-Ginés et al., 2023, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it"

FIGURE 3.4: Hierarchical classification of media bias types based on author's intention and contextual factors.



By naming the seventeen manifestations explicitly at the lower level of Figure 3.4, the taxonomy provides a shared vocabulary that media bias detection systems and information-disorder detection systems can both reuse, and clarifies that what differs between the two tasks is the upstream question (is the underlying claim factually false, intentionally misleading, or rhetorically slanted), not the inventory of textual signals that has to be detected. None of the previous inventories made this connection explicit; the proposed taxonomy does so by construction.

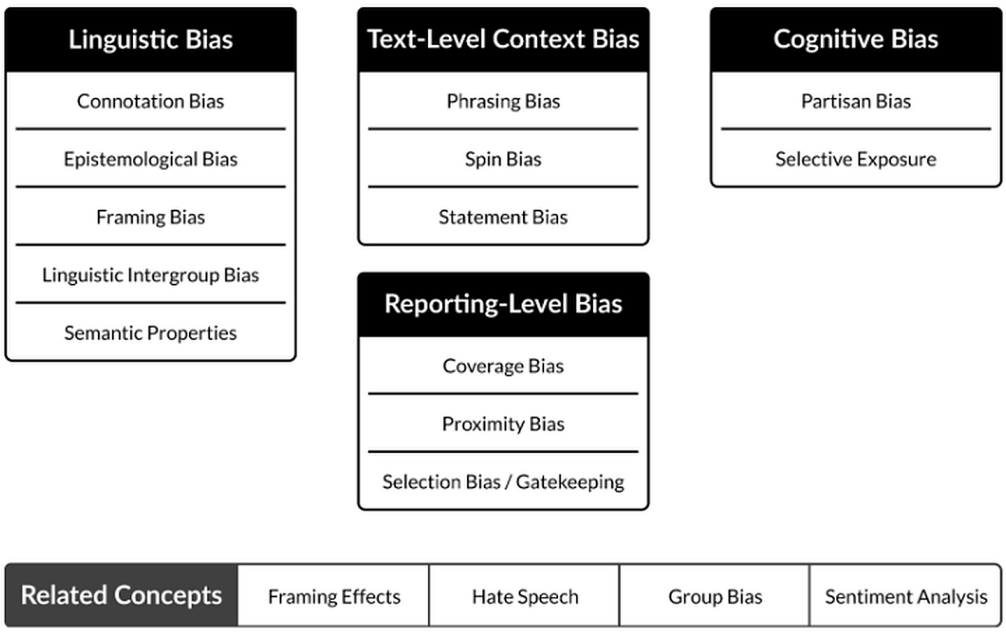
The upper level distinguishes bias types by communicative intention (spin bias and ideology bias) and by contextual factors (coverage bias, gatekeeping bias, and statement bias). The lower level specifies the 17 concrete manifestations that realise each type through observable linguistic and discursive patterns. Notably, some manifestations appear under multiple bias types, reflecting their context-dependent function: for example, sensationalism may serve spin objectives in one article while simultaneously reinforcing ideological framing in another. In total, the taxonomy identifies 17 distinct bias manifestations, each grounded in linguistic evidence and operationalisable for computational analysis.

▸ **CONSTRUCTION METHODOLOGY**

The taxonomy was constructed through an iterative, bottom-up process grounded in qualitative coding principles and informed by the systematic literature review. During the review of 118 studies (see Chapter 2), every bias-related concept, mechanism, or linguistic feature was extracted and catalogued, producing an initial inventory of over 40 distinct terms. These were normalized to resolve terminological inconsistencies across studies (e.g., "loaded language", "emotionally charged language", and "affective lexical choice" were consolidated under a single manifestation). The resulting candidates were then organized into a two-level hierarchy linking bias types (defined by communicative intention and contextual factors) with their concrete manifestations. The taxonomy was subsequently validated through its publication in *Expert Systems with Applications* [149]. Further empirical validation was achieved through the implementation of six manifestations in the MBBMD dataset (Section 3.2) and their computational implementation as both heuristic rules and supervised classifiers (Chapter 4).

[149] Rodrigo-Ginés et al., 2023, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it"

FIGURE 3.5: Media bias taxonomy developed by [177], reproduced here for comparison with the proposed taxonomy of Figure 3.4. The original figure is reproduced from the source publication; readers are referred to that publication for the highest-resolution rendering.



To our knowledge, this taxonomy constitutes the first attempt to provide such a structured and integrative framework for media bias analysis from within Computer Science (rather than from communication studies or media theory). Shortly after its publication, Spinde et al. [177] introduced a complementary taxonomy (Figure 3.5) that emphasises audience

[177] Spinde et al., 2023 "The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias"

perception mechanisms rather than the hierarchical relationship between intention, context, and linguistic realisation. Despite shared principles, we argue that our hierarchical and linguistically oriented design provides a more flexible framework for computational modelling and fine-grained bias characterisation in the NLP and computational social science settings that constitute the engineering scope of this thesis. We do not claim that this taxonomy supersedes the perception-centred design of Spinde et al. [177] for the purposes of audience research; the two frameworks are aimed at different research questions and are best read as complementary.

An important characteristic of the proposed taxonomy is that certain manifestations appear under more than one media bias type. This design choice is intentional and reflects the fact that the same linguistic strategy may serve different bias functions depending on context and intent. For example, a given manifestation may operate as a form of framing bias in one context while simultaneously contributing to opinion-as-fact bias in another. Rather than forcing a one-to-one mapping, the taxonomy explicitly models this overlap, capturing the multifaceted and non-exclusive nature of media bias.

The following subsections detail the 17 media bias manifestations defined in our taxonomy, providing explanations and illustrative examples of how each manifestation influences both news content and audience perception. To improve readability and ensure consistency between the conceptual framework and its textual exposition, each manifestation is labeled using the same numerical identifier employed in Figure 3.4.

▸ Unsubstantiated claims bias (1)

This manifestation of media bias occurs when journalists present claims, allegations, or causal assertions without providing sufficient empirical evidence, verifiable sources, or explicit justification. Rather than clearly distinguishing between confirmed facts, hypotheses, or speculation, such statements are often framed in a way that suggests epistemic certainty, thereby encouraging audiences to accept them as true [48]. Linguistically, this bias is typically realized through declarative constructions, assertive modal verbs, and the absence of attribution markers or evidential qualifiers.

[48] Ekström et al., 2021, "The epistemologies of breaking news"

Unsubstantiated claims bias can serve different bias functions depending on context and intent. In sensationalist reporting, unsupported assertions may be used to construct dramatic or alarming narratives, aligning this manifestation with spin bias. In other cases, they may be selectively employed to advance a particular ideological position, thus also operating as a form of ideology bias. As such, this manifestation is primarily classified under statement bias, as it concerns how information is presented rather than what information is selected.

A well-known example is the persistent media claim that vaccines cause autism. Despite being thoroughly discredited by the scientific community, this assertion continues to appear in certain outlets without adequate evidential grounding. The repeated dissemination of such unsubstantiated claims illustrates the societal risks associated with this bias, including the erosion of trust in scientific institutions and increased susceptibility to preventable public health crises [159].

[159] Ruiz and Bell, 2014, "Understanding vaccination resistance: vaccine search term selection bias and the valence of retrieved information"

▸ Opinion statements presented as facts bias (2)

This manifestation arises when subjective evaluations, interpretations, or value judgments are presented using linguistic forms typically reserved for factual reporting, thereby obscuring the distinction between opinion and verified information [8]. Rather than explicitly signaling subjectivity through markers such as evaluative attribution or modality, the journalist embeds personal or ideological viewpoints within declarative, fact-like statements.

[8] Anand et al., 2007, "Information or opinion? Media bias as product differentiation"

From a taxonomic perspective, this manifestation occupies an intermediate position between statement bias and ideology bias. While it operates at the level of presentation, its underlying intent is often ideological, as it allows journalists to promote specific viewpoints while maintaining an appearance of objectivity. Linguistically, this bias is frequently realized through the use of evaluative adjectives, presuppositions, and implicit normative framing.

For example, reporting that "the government has introduced a new policy aimed at reducing carbon emissions" constitutes a neutral factual statement. In contrast, framing the same event as "the government has once again shown its disregard for businesses by imposing a stifling policy" introduces subjective judgments such as "disregard" and "stifling", that are presented as objective assessments. This rhetorical strategy can mislead readers into internalizing the journalist's perspective as factual truth, thereby shaping public perception and constraining democratic debate [127].

[127] Parker-Bass et al., *News Bias Explored*

▸ Sensationalism or emotionalism bias (3)

Sensationalism or emotionalism bias manifests when news content is framed in a manner that amplifies emotional responses such as fear, anger, or outrage; at the expense of proportionality and factual nuance [166, 104]. This bias typically prioritizes dramatic elements of a story, including extreme outcomes or isolated incidents, while downplaying contextual information that would enable a more balanced interpretation.

[166] Schuck, Feinholdt, et al., 2015, "News framing effects and emotions"

[104] Lin et al., 2007, "What emotions do news articles trigger in their readers?"

Although often classified as statement bias due to its linguistic realization, sensationalism may also reflect ideological or strategic intentions, particularly when emotional framing is selectively applied to reinforce specific narratives. Common linguistic markers include hyperbolic expressions, emotionally charged adjectives, alarmist headlines, and evocative metaphors.

For instance, describing a mild winter storm with a headline such as "Heavy Blizzard Hits New York City Causing Havoc and Claiming One Fatality" exaggerates both the scale and impact of the event. Beyond misleading audiences, sensationalism increases the likelihood of content virality, thereby amplifying its societal impact. This dynamic contributes to the diffusion of misinformation and reflects a broader media trend in which emotional engagement is prioritized over informative accuracy [127].

▸ Ad hominem or mudslinging bias (4)

Ad hominem or mudslinging bias occurs when journalistic discourse shifts from substantive discussion of policies, actions, or arguments to personal attacks on individuals or groups. Rather than addressing the merits of an issue, this bias seeks to undermine credibility by targeting character traits, motivations, or alleged personal failings [164]. It is most commonly associated with statement bias driven by ideological intent.

[164] Scher, 1997, *The modern political campaign: Mudslinging, bombast, and the vitality of American politics*

Linguistically, this manifestation is realized through evaluative labeling, insinuations, and character-focused narratives that displace policy-oriented analysis. Mudslinging is particularly prevalent in political journalism and electoral contexts, where it functions as a rhetorical strategy to delegitimize opponents without engaging with their positions.

For example, instead of responding to a policy proposal, a news article may highlight alleged financial ties or moral shortcomings of a political actor, thereby redirecting attention away from the substantive debate. This bias degrades the quality of public discourse, fosters polarization, and encourages audiences to form opinions based on personal attacks rather than informed policy evaluation [203].

[203] Yap, 2013, "Ad hominem fallacies, bias, and testimony"

▸ Mind reading bias (5)

Mind reading bias arises when journalists attribute intentions, beliefs, or strategic motivations to individuals or groups without direct evidence or verifiable sources. Rather than reporting observable actions or documented statements, this manifestation relies on speculative inference, presenting conjecture as factual insight [64]. It is commonly associated with statement bias and may serve either ideological or spin-driven objectives.

[64] Groseclose, 2011, *Left turn: How liberal media bias distorts the American mind*

This bias is linguistically marked by verbs of cognition and intention (e.g., "plans", "wants", "seeks to"), often used without attribution or evidential support. For instance, an article might claim that a political leader is "secretly preparing" a controversial reform despite the absence of official confirmation.

The impact of mind reading bias is substantial, as it constructs narratives based on assumed motives rather than demonstrable facts. By legitimizing speculation, this bias distorts public understanding of political processes and contributes to declining trust in journalistic credibility. Over time, the repeated use of mind reading bias normalizes conjecture as an acceptable substitute for evidence-based reporting.

▸ Slant bias (6)

Slant bias is a pervasive manifestation of ideology bias characterized by the systematic emphasis on certain aspects of an event while downplaying or omitting others, thereby favoring a particular ideological perspective. Rather than falsifying information outright, this bias operates through selective emphasis, framing, and narrative focus, shaping how audiences interpret events [47].

[47] Druckman and Parkin, 2005, "The impact of media bias: How editorial slant affects voters"

Linguistically and discursively, slant bias is realized through asymmetric framing, preferential quoting, and differential prominence given to actors or arguments. For example, coverage of an environmental protest may foreground economic disruption and public inconvenience while marginalizing environmental motivations, or vice versa, depending on the outlet's ideological alignment. Over time, slant bias reinforces polarization by repeatedly exposing audiences to ideologically consistent narratives, limiting exposure to alternative viewpoints and constraining deliberative debate.

▸ Picture explanation bias (7)

Picture explanation bias, classified under coverage bias, arises from the captions, headlines, or textual explanations accompanying images. These explanations guide audience interpretation by framing the visual content in a particular way, often introducing evaluative or misleading cues that are not inherent to the image itself.

This bias is especially powerful because images are often perceived as objective representations of reality, while their interpretations are highly malleable. For instance, selectively describing an image of a crowd as "angry protesters" versus "concerned citizens" can dramatically alter audience perception. Empirical studies have shown that such biased explanations can reinforce stereotypes, as in cases where media outlets exaggerate the presence of specific racial groups in negative socioeconomic contexts [62].

[62] Gilens, 1996, "Race and poverty in americapublic misperceptions and the american news media"

▸ Picture selection bias (8)

Picture selection bias refers to the deliberate or systematic choice of images that support a specific narrative or evaluative stance, while alternative visual representations are excluded. Categorized under coverage bias, this manifestation exploits the emotional and persuasive power of images to influence audience perception [61].

[61] Geske, 2016, "Riot vs. Revelry: News Bias Through Visual Media"

For example, selecting photographs that depict isolated acts of violence during a largely peaceful protest can create the impression of widespread disorder. Unlike textual bias, picture selection bias operates largely at a pre-linguistic level, shaping interpretation before any textual analysis occurs. This makes it particularly challenging to detect automatically, yet highly influential in constructing media narratives.

▸ Subjective qualifying adjectives bias (9)

This manifestation involves the use of evaluative adjectives that reflect the journalist's subjective stance rather than neutral description. Frequently associated with spin bias and ideology bias, such adjectives subtly influence audience perception by embedding value judgments within ostensibly factual reporting.

For instance, describing a political proposal as "radical", "reckless", or "ambitious" introduces implicit evaluation without explicit argumentation. Because these adjectives often operate below the threshold of conscious scrutiny, their persuasive effect can be substantial.

From a linguistic perspective, this bias is realized through adjectival modification and appraisal language, making it particularly relevant for fine-grained NLP-based bias detection [124].

[124] Pant et al., 2020, "Towards Detection of Subjective Bias using Contextualized Word Embeddings"

▸ **BIAS BY LABELING AND WORD CHOICE (10)**

This bias, categorized under statement bias and driven by ideological motives, involves the deliberate selection of language that frames an issue in a particular way. Journalists, through their choice of words, can significantly influence how the audience perceives a situation or individual. For example, labeling a group as "freedom fighters" instead of "rebels" or "terrorists" can create vastly different impressions in the minds of readers. Similarly, describing a tax policy as "tax relief" rather than "tax cuts" can sway public opinion by casting the policy in a more favorable or unfavorable light.

The power of bias by labeling and word choice lies in its subtlety. It operates below the surface, influencing the reader's perception without overtly stating a particular viewpoint. This form of bias is especially potent in shaping public discourse because it taps into the emotional and psychological responses that specific words and phrases can evoke. By carefully choosing language that aligns with an ideological agenda, media outlets can steer public interpretation of news stories in ways that may not be immediately apparent to the audience [71].

[71] Hamborg, 2020, "Media bias, the social sciences, and NLP: automating frame analyses to identify bias by word choice and labeling"

▸ **FLAWED LOGIC BIAS (11)**

The flawed logic bias is often a manifestation of spin bias, occurs when journalists employ faulty reasoning or logical fallacies to draw conclusions that are not supported by the evidence. This type of bias is closely related to several common logical fallacies, which can mislead the audience by presenting arguments that seem valid on the surface but are actually based on flawed reasoning or fallacies [186, 113, 209, 123]. Key examples include:

[186] Teneva, 2023, "Digital pseudo-identification in the post-truth era: Exploring logical fallacies in the mainstream media coverage of the COVID-19 vaccines"

[113] McMurtry, 1990, "The mass media: An analysis of their system of fallacy"

[209] Zhou, 2018, "The logical fallacies in political discourse"

[123] Pangestuti, 2023, "Logical fallacy found in the politic column of BBC News online magazine: Lexical and syntactical ambiguities"

- **Hasty generalization**: Drawing a broad conclusion based on an unrepresentative sample. For example, if a news report interviews only a few people at a protest and then claims that "the majority of attendees share the same view", it may be committing a hasty generalization.
- **False cause fallacy (non causa pro causa)**: Misidentifying the cause of one phenomenon as another that is only superficially related. A media outlet might attribute a rise in crime to a new policy without adequately exploring other factors that could be contributing.
- **Slippery slope**: Suggesting that a minor action will lead to major and often dire consequences, without providing evidence for such a chain of events. For instance, a news story might claim that legalizing a particular drug will inevitably lead to societal collapse, without considering other variables.
- **Black-and-white thinking (splitting)**: Presenting a complex issue in a dichotomous manner, ignoring the nuances and diversity of perspectives. An example would be framing a political debate as strictly "liberal vs. conservative", thereby oversimplifying the issue.
- **Fallacy of division**: Assuming that what is true for the whole is true for each part. For example, if a political party holds a specific stance, a media outlet might assume that every member of that party must hold that stance.
- **Fallacy of composition**: Assuming that what is true for a part is true for the whole. Conversely, if one politician in a party holds a particular view, a news report might claim that the entire party holds that view.
- **Irrelevant conclusion fallacy**: Drawing a conclusion that is not related to the argument presented. For instance, attributing a natural disaster to a politician's actions without providing a logical connection.
- **Appeal to groupthink (Ad populum)**: Asserting that something is true because "everyone else is doing it" or "everyone else believes it". This fallacy can lead to the erroneous acceptance of a claim based on its popularity rather than its validity.

- **Appeal to authority (Ad verecundiam)**: Relying on the opinion of an authority figure rather than using logic or evidence to support an argument. For instance, using a celebrity's endorsement in a political debate instead of relying on expert analysis.
- **Red Herring (Ad ignorantiam)**: Introducing an unrelated topic to divert attention from the main argument. This fallacy is often used in political reporting, where personal matters are highlighted to distract from policy discussions.

Flawed logic bias is particularly harmful because it can create a misleading narrative that seems logical and convincing, even though it is based on erroneous reasoning. This type of bias not only distorts the facts but also undermines the audience's ability to engage in critical thinking and informed decision-making [192].

[192] Van Vleet, 2021, *Informal logical fallacies: A brief guide*

▸ Omission bias (12)

Omission bias, a form of coverage bias [210], occurs when relevant facts, actors, or events are systematically excluded from media coverage. Rather than presenting false information, this bias shapes understanding by presenting an incomplete narrative.

[210] Zhukova et al., 2023, "What's in the News? Towards Identification of Bias by Commission, Omission, and Source Selection (COSS)"

For example, limited coverage of minor political candidates can constrain electoral choice by restricting audience awareness. Omission bias is particularly insidious because absences are less salient than presences, making the bias difficult for audiences to detect. Over time, systematic omissions can reinforce existing power structures and marginalize alternative perspectives.

▸ Commission bias (13)

Commission bias refers to the selective inclusion of viewpoints, sources, or arguments that favor one side of a debate while excluding opposing perspectives. Often overlapping with spin and ideology bias, this manifestation produces one-sided narratives that give the appearance of consensus [43].

[43] De Witte, 2022, *Groupthink gone wrong: Stanford scholars show how assumptions about electability undermine women political candidates*

For instance, election coverage that exclusively features perceived supporters of a favored candidate can create a distorted sense of public opinion. Commission bias actively shapes discourse by privileging certain voices and silencing others, thereby constraining the range of acceptable viewpoints within public debate.

▸ Bias by placement (14)

Bias by placement involves the strategic positioning of news items to influence perceived importance and salience [63]. Stories placed on front pages or at the top of digital platforms receive disproportionate attention compared to those relegated to less visible sections.

[63] Green, 1999, "Ethnic evaluations of advertising: Interaction effects of strength of ethnic identification, media placement, and degree of racial composition"

This bias operates at the level of editorial decision-making rather than textual content, yet it significantly affects audience perception of issue importance. By consistently foregrounding certain topics while marginalizing others, media outlets can shape public agendas without altering factual content.

▸ Size allocation bias (15)

Size allocation bias, categorized under coverage bias, refers to the disproportionate space allocated to certain stories, images, or headlines within a publication. Studies have shown that readers are more likely to pay attention to larger headlines or more prominently featured images [84]. This bias can lead to a skewed understanding of news events, as the prominence of a story can imply its importance or validity, regardless of its actual significance. For instance, a news outlet might give a large headline to a minor political gaffe while downplaying more critical issues with smaller text and less visible placement. Size allocation bias thus plays a crucial role in shaping the narrative by directing the audience's focus to particular aspects of the news.

[84] Jiang et al., 2019, "What prompts users to click on news headlines? Evidence from unobtrusive data analysis"

▸ **SOURCE SELECTION BIAS (16)**

This bias, which is present in both gatekeeping bias and ideology bias, involves journalists selectively choosing sources that support a particular narrative while excluding those that might provide a more balanced view. For example, a journalist covering an environmental disaster might only interview representatives from the responsible company, neglecting to include perspectives from affected community members or independent experts. This selective sourcing can create a distorted version of events, leading the audience to form opinions based on incomplete or biased information [72]. Source selection bias can also reinforce existing power structures by privileging voices that align with the media outlet's interests.

[72] Hamborg et al., 2019, "Automated identification of media bias in news articles: an interdisciplinary literature review"

▸ **OMISSION OF SOURCE ATTRIBUTION BIAS (17)**

This manifestation occurs when journalists fail to provide explicit or precise source attribution, using vague formulations such as "experts say" or "critics argue". Associated with coverage bias and gatekeeping bias, this practice undermines transparency and accountability in reporting [127].

[127] Parker-Bass et al., *News Bias Explored*

By obscuring the origin and credibility of claims, omission of source attribution bias prevents audiences from evaluating reliability and potential conflicts of interest. Proper attribution is a cornerstone of responsible journalism, and its absence can facilitate the covert introduction of biased or speculative information.

3.2 MBBMD: A PERSPECTIVIST AND HIERARCHICAL DATASET FOR MEDIA BIAS ANALYSIS

The taxonomy proposed in Section 3.1 provides the conceptual vocabulary for describing media bias, but a taxonomy alone cannot be trained on, evaluated against, or used to test whether a detection system has learned a transferable abstraction. To close that gap, this section presents the Media Bias Bias-Mitigated Dataset (MBBMD), a hierarchical Spanish-language dataset that instantiates the taxonomy through explicit annotation of bias types, manifestations, and perspectivist disagreement, and serves as the methodological backbone on which the empirical work in Chapter 4 is built.

MBBMD is grounded in the assumption that media bias perception emerges from the interaction between textual cues, editorial context, and the reader's interpretive perspective, and it is designed as a direct response to three structural limitations of current resources. First, most datasets collapse annotator disagreement into majority-vote labels, implicitly treating media bias as an objective property of the text rather than as a contested interpretation; MBBMD preserves individual annotator judgments and ideological self-identification. Second, bias is typically modelled at a single level of analysis (usually the document), despite theoretical evidence that it operates across multiple layers; MBBMD's annotation schema is hierarchical by construction, supporting sentence-level, document-level, and category-level analyses within a single corpus. Third, existing resources rarely account for the sensitivity of bias perception to contextual factors such as outlet identity, named entities, or emotionally charged terminology; MBBMD addresses this through Counterfactual Data Augmentation (see below) and through a corpus built on Spanish-language coverage of ten topics drawn from Spanish, Irish, Argentine, and Israeli–Palestinian media ecosystems, deliberately outside the anglocentric benchmark mainstream.

A distinctive feature of MBBMD is the systematic use of Counterfactual Data Augmentation (CDA) [211]. Rather than treating counterfactuals as a form of data augmentation for performance improvement, MBBMD incorporates controlled counterfactual variants as experimental probes. By modifying outlet attribution, named entities, or ideologically loaded terminology while preserving propositional content, the dataset enables the empirical study of how bias perception shifts under minimal contextual and lexical changes. The dataset focuses on Spanish-language news, a context that remains underrepresented in computational studies of media bias despite its political diversity and global relevance. MBBMD covers ten socially

[211] Zmigrod et al., 2019, "Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology"

FIGURE 3.6: Sources and mitigation of bias across the MBBMD annotation pipeline.

	Media bias	Dataset-design bias	Annotator bias
Process phases	1. Topic selection 2. Selection and omission of sources 3. Editorial framing choices 4. Use of evaluative, emotional, or loaded language 5. Outlet ideological stance and style	1. Article selection within each topic 2. Definition of annotation taxonomy and guidelines 3. Preparation of annotation examples 4. Design of topic–outlet sampling strategy 5. Construction of the dataset split	1. Interpretation of bias presence 2. Assignment of editorial bias categories 3. Identification of linguistic manifestations 4. Decision thresholds for subtle cues 5. Adjudication of disagreements
Sources of bias	• Asymmetric coverage of politically salient events • Selective emphasis or omission of contextual elements • Lexical framing that subtly encodes stance • Ideologically charged terminology • Presentation of opinions as factual statements	• Risk of selecting atypically biased or unusually neutral articles • Overrepresentation of certain outlet styles or tones • Guidelines that may unintentionally prime annotators • Topic or outlet imbalance • Contextual leakage between train and test	• Ideological priors influencing perception of bias • Different levels of media literacy and interpretive strategies • Annotator fatigue or inconsistency • Divergent interpretations of subtle categories • Tendencies toward overannotation or underannotation
Mitigations	• Paired selection of articles • Balanced topic selection • Hierarchical taxonomy defined prior to annotation • Fine-grained linguistic bias manifestations • Counterfactual Data Augmentation (CDA) to test sensitivity	• Paired selection of articles • Hierarchical taxonomy defined prior to annotation • Detailed guidelines and calibration exercises • Context priming (annotators read all articles of a topic before labeling) • Stratified topic-aware dataset split • CDA variants created to assess robustness	• Perspectivist annotation • Diverse annotator pool • Detailed guidelines and calibration exercises • Context priming to equalize background knowledge • Adjudication only at sentence level, ensuring consistency for fine-grained categories

and politically salient topics and includes articles from ideologically diverse outlets with comparable levels of factual reliability. Through this design, the dataset aims to support both the development of robust computational models and the analysis of epistemic limitations inherent to automatic media bias detection.

The remainder of this section is organised as follows. Subsection 3.2.1 outlines the design principles that guided the dataset construction. Subsection 3.2.2 describes the data collection and annotation methodology. Subsection 3.2.3 presents the dataset structure and label hierarchy. Subsection 3.2.4 provides an empirical analysis of the corpus, including agreement patterns and the impact of counterfactual interventions. Finally, Subsection 3.2.5 discusses reproducibility considerations and future extensions.

3.2.1 *Design principles*

The design of MBBMD is guided by a set of methodological principles derived from the epistemic and practical challenges identified in previous chapters. Rather than treating media bias as a fixed and objective property of texts, MBBMD conceptualizes media bias as a hierarchical, subjective, and context-sensitive phenomenon, whose perception depends on the interaction between linguistic realization, editorial framing, and reader perspective. These principles informed all major design decisions in the dataset, from annotation strategy to corpus composition and counterfactual construction.

A foundational assumption of MBBMD is that bias may arise at multiple stages of the dataset lifecycle. Media bias is not only present in the media content itself, but can also be introduced or amplified through dataset design decisions and through the annotation process. Figure 3.6 summarizes these stages, distinguishing between media-intrinsic bias, dataset-design bias, and annotator bias, and mapping each of them to specific mitigation strategies implemented in MBBMD.

A first core principle is the explicit treatment of media bias perception as perspectivist. Empirical research in political communication and social psychology has consistently shown that readers with different ideological, cultural, or experiential backgrounds may legitimately diverge in their interpretation of the same news content [41]. Consequently, annotator disagreement reflects genuine interpretive variability rather than annotation error. At the document level, MBBMD therefore preserves individual annotator judgments and their distributions, following the Learning with Disagreements (LeWiDi) [100] paradigm. This approach allows bias to be represented both through conventional majority-vote labels and through graded measures of consensus and polarization, enabling the study of ambiguous and con-

[41] Dahlgren, 2020, "Media echo chambers: Selective exposure and confirmation bias in media use, and its consequences for political polarization"

[100] Leonardelli et al., 2023, "SemEval-2023 task 11: Learning with disagreements (LeWiDi)"

tested cases that are typically flattened in existing datasets. As illustrated in Figure 3.6, perspectivist annotation and a diverse annotator pool are explicitly employed to mitigate annotator-level bias.

A second core principle is the hierarchical modelling of media bias. Building on the media bias taxonomy introduced in Section 3.1 of this chapter, MBBMD assumes that media bias is not a single-layer phenomenon but a multi-level construct that ranges from high-level editorial strategies to concrete linguistic realisations. MBBMD operationalizes this taxonomy by explicitly linking three annotation layers within a unified framework: document-level bias perception (capturing holistic reader judgments), document-level bias categories (capturing higher-order editorial dimensions), and sentence-level manifestations (capturing linguistically explicit mechanisms). Document-level annotations capture global impressions of bias that often arise holistically from discourse structure and framing, whereas sentence-level annotations ground these impressions in observable textual cues that can be localized and modelled.

This hierarchical structure supports analyses that move beyond binary detection toward bias characterization and explanation, enabling the study of how micro-level cues aggregate into macro-level judgments, and of how different taxonomic dimensions co-occur and interact within the same article. Importantly, this design also mitigates dataset-design bias by committing to a theoretically grounded schema prior to annotation, thereby reducing post-hoc category drift and ensuring that annotation decisions remain aligned with an explicit conceptual framework.

MBBMD adopts an asymmetric treatment of subjectivity across annotation levels. While document-level bias perception is inherently subjective and perspectivist, sentence-level annotation targets concrete linguistic mechanisms that can be identified with higher reliability. For this reason, disagreement is preserved at the document level, whereas sentence-level annotations follow a deterministic procedure with adjudication. This distinction reflects the different epistemic roles of each level: document-level labels model perception and interpretation, whereas sentence-level labels model linguistic realization. Treating both levels identically would either obscure subjectivity or undermine reproducibility; separating them allows MBBMD to balance interpretive richness with methodological stability, as reflected in the differentiated mitigation strategies shown in Figure 3.6.

The selection of sentence-level bias manifestations follows directly from this principle. Although the media bias taxonomy introduced in Subsection 3.1.3 of this chapter comprises a broader set of seventeen media bias manifestations, only a subset is operationalised at the sentence level in MBBMD. This reduction does not reflect a theoretical simplification, but a methodological constraint grounded in observability, annotability, and epistemic scope.

Many media bias categories and manifestations identified in Subsection 3.1.3, such as gatekeeping, coverage imbalance, or systematic omission/commission, require cross-document comparison, temporal aggregation, or access to contextual information beyond isolated sentences. In addition, media bias manifestations relying on multimodal cues, including visual framing, image selection, captioning, or layout, cannot be reliably captured in a text-only dataset and would require a dedicated multimodal corpus.

Importantly, the exclusion of these mechanisms from sentence-level annotation does not imply that they are ignored in MBBMD. Cross-document bias phenomena such as gatekeeping and coverage bias are explicitly addressed at the document level through the annotation of media bias categories. During this phase, annotators are first presented with the full set of articles covering the same event or topic and are instructed to read them collectively before annotating each article individually. This procedure enables annotators to identify patterns of selective coverage, omission, or disproportionate emphasis across outlets, thereby supporting the analysis of cross-document bias mechanisms within the dataset, albeit at a higher level of abstraction.

The six sentence-level manifestations selected in MBBMD are linguistically explicit, identifiable within sentence boundaries, grounded in observable lexical, syntactic, or propositional cues, and reproducibly annotable at scale. These manifestations are:

- **Omission of source attribution:** Instances where critical claims or information are presented without proper source citation. For example, statements such as "experts agree that this policy will fail" raise the question of which experts are being referenced; if no sources are provided, the omission constitutes a marker of bias.
- **Sensationalism or emotionalism:** The use of hyperbolic or emotionally charged language intended to provoke reactions rather than inform. Expressions like "this disaster will ruin everything" or "a catastrophe is unfolding" exaggerate facts to elicit unnecessary emotional responses.
- **Mind reading:** Unsubstantiated assertions about the thoughts, intentions, or motivations of individuals or groups. Statements such as "politicians only support this bill to gain more power" reflect bias when no concrete evidence is provided to support the attributed intentions.
- **Word choice or labeling:** Specific lexical choices that subtly frame entities, events, or actors in a positive or negative light. Terms such as "freedom fighters" versus "terrorists" or "reform" versus "overhaul" carry implicit evaluative meanings that shape reader perception.
- **Use of subjective qualifying adjectives:** Adjectives that convey personal evaluation rather than objective description. Labels such as "reckless policy", "brilliant strategy", or "irresponsible decision" introduce bias unless explicitly supported by factual evidence or attributed opinions.
- **Presentation of opinions as facts:** Cases in which subjective judgments are presented as objective statements. For example, "this policy will fail" expresses a biased stance unless it is appropriately attributed, as in "critics believe this policy will fail".

Another defining principle of MBBMD is the use of Counterfactual Data Augmentation (CDA) as an experimental instrument rather than as a mere data expansion technique. Bias perception is known to be highly sensitive to contextual cues such as outlet identity, named entities, and emotionally charged terminology. To isolate and study these effects, MBBMD includes systematically constructed counterfactual variants that preserve propositional content while modifying specific contextual or lexical elements. These variants enable controlled analyses of how minimal changes influence both human judgments and model predictions, supporting robustness evaluation, causal probing, and the study of heuristic reliance in media bias detection. As shown in Figure 3.6, CDA is explicitly positioned as a mitigation strategy targeting both media-intrinsic and annotator-level bias.

Finally, the dataset is designed to support generalization across topics and sources. Media outlets are selected to span the ideological spectrum while maintaining comparable levels of factual reliability, and topics are chosen to reflect diverse forms of political, social, and cultural discourse. This balanced design reduces confounding effects between topic, ideology, and stylistic conventions, and ensures that observed patterns reflect genuine bias phenomena rather than dataset artifacts.

Taken together, these principles materialize in four core design contributions that distinguish MBBMD from existing resources:

- **Hierarchical annotation framework:** Media bias is captured across three interrelated levels (document-level bias perception, document-level bias categories, and fine-grained sentence-level linguistic manifestations) enabling integrated analyses that link global impressions to local textual evidence.
- **Systematic and balanced source selection:** News outlets are selected to represent a wide ideological spectrum while controlling for factual reliability, ensuring that observed bias patterns are not confounded by differences in journalistic quality or credibility.
- **Perspectivist annotation with preserved disagreement:** By adopting the Learning with Disagreements paradigm at the document level, MBBMD retains annotator diversity and models bias as an interpretive phenomenon rather than enforcing artificial consensus.

- **Counterfactual data augmentation as an analytical instrument:** Carefully constructed counterfactual variants are incorporated to probe the sensitivity of bias perception to contextual and lexical cues, supporting robustness analysis and causal interpretation.

The following sections detail how these principles and contributions are instantiated in the dataset construction process, annotation methodology, and empirical analysis.

3.2.2 *Dataset construction methodology*

The construction of MBBMD follows a sequential and methodological pipeline designed to operationalize the design principles introduced in the previous section. Rather than treating dataset creation as a technical task, the construction process explicitly addresses the multiple sources of media bias that may arise at different stages of the dataset lifecycle, including biases inherent to media content, biases introduced through dataset design decisions, and biases emerging from human annotation. The resulting methodology balances control and diversity, aiming to preserve interpretive variability where epistemically meaningful while enforcing stability where reproducibility is required.

The dataset is composed exclusively of Spanish-language journalistic articles published between late 2023 and early 2024. This temporal restriction ensures topical relevance while avoiding long-term temporal drift effects that could confound bias perception. Only written news articles were included in the corpus; audiovisual material, opinion columns explicitly labeled as such, and social media content were excluded in order to maintain genre consistency and isolate linguistic and editorial bias phenomena within a homogeneous textual domain. These constraints were deliberately adopted to ensure comparability across articles and to reduce confounding stylistic variation unrelated to bias.

▸ Topic and media outlets selection

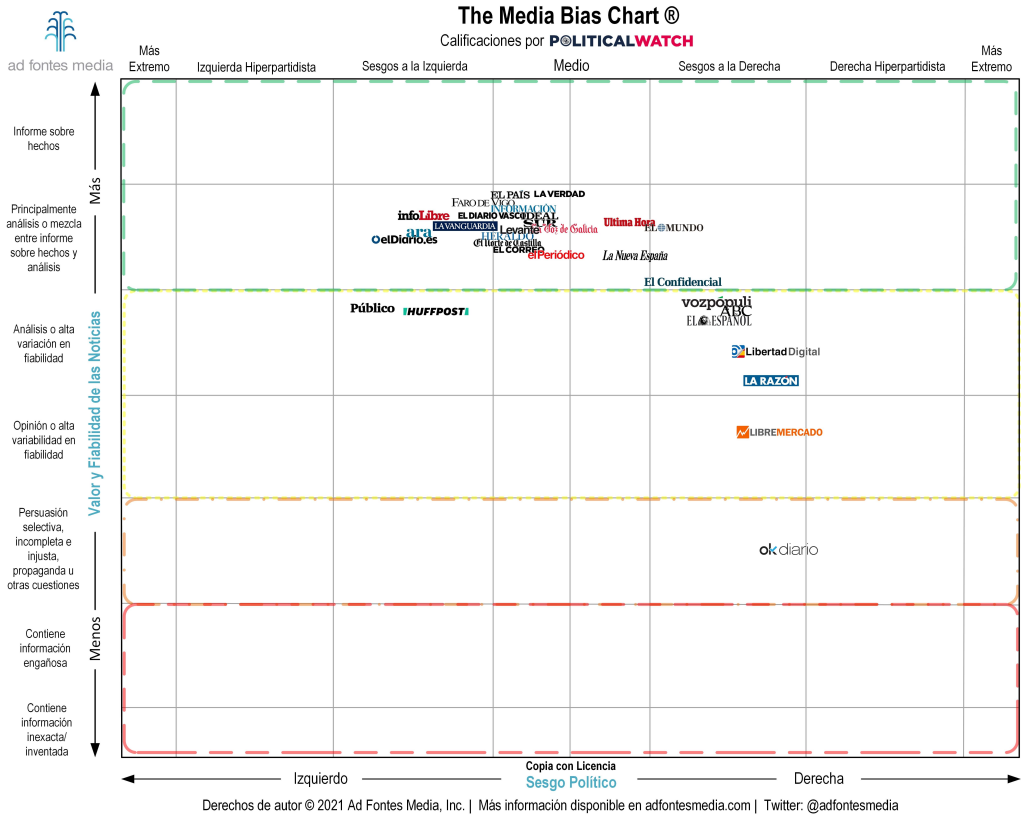
Topic selection constitutes a central design decision in MBBMD. Ten topics were selected to cover a diverse range of political, social, and cultural issues with high public relevance and ideological salience. Each topic was required to have received sustained coverage across multiple media outlets and to allow for divergent framing or interpretation. Specifically, the ten selected topics are: (1) Spain's public debt, (2) Eurovision song selection, (3) Sinn Féin referendum, (4) Catalonia drought emergency, (5) Luis Rubiales case, (6) amnesty protest, (7) trans law approval, (8) Pedro Sánchez investiture, (9) Argentina elections, and (10) ICJ orders on Gaza. These topics span national political processes, international conflicts, social legislation, environmental crises, and culturally salient controversies. This diversity reduces topic-specific artifacts and supports cross-topic generalization, while ensuring that bias manifests through heterogeneous rhetorical and editorial strategies rather than through topic idiosyncrasies.

Media outlet selection was guided by the Spanish Media Bias Chart developed by Ad Fontes Media [114], which positions news outlets along two orthogonal dimensions: political orientation (left to right) and factual reliability (low to high). Only outlets with explicit ratings on both axes were considered eligible for inclusion. This framework provides a principled basis for balancing ideological diversity while controlling for journalistic quality, thereby mitigating dataset-design bias arising from uneven source credibility.

[114] Media, 2023, *How Ad Fontes Ranks News Sources*

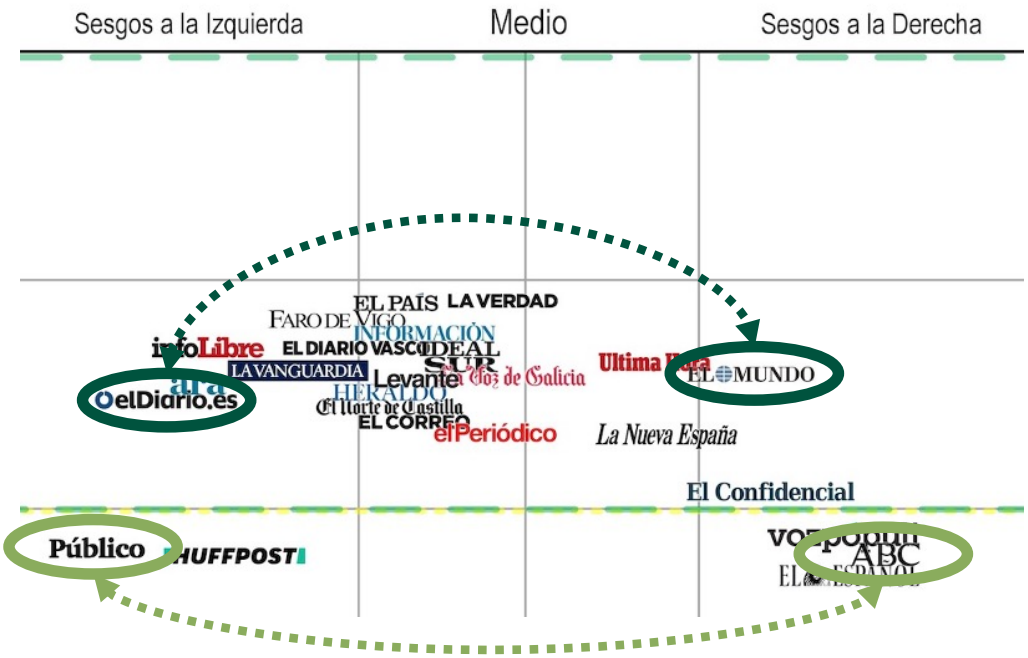
For each selected topic, articles were sampled from outlets spanning the ideological spectrum while maintaining comparable levels of factual reliability. Concretely, three operational criteria were applied at the outlet level: (i) each outlet had to have an explicit position on both axes of the Spanish Media Bias Chart (political orientation and factual reliability), so that ideological diversity could be controlled jointly with journalistic quality; (ii) for each topic, outlets were paired symmetrically in left/right pairs with similar factual-reliability scores (and, where available, a centrist outlet was included as a reference point), so that ideology and quality were never confounded by construction; and (iii) outlets had to publish in Spanish and be active during the data-collection window (late 2023 to early 2024), so that the resulting corpus reflects a contemporaneous Spanish media landscape rather than a longitudinal mix of outlets that may have changed editorial line over time. The pairing

FIGURE 3.7: Spanish Media Bias Chart by Ad Fontes Media.



strategy that emerges from these three criteria is illustrated in Figure 3.8, which zooms into the Spanish Media Bias Chart of Figure 3.7 to show, for an explanatory subset, how outlets with similar factual reliability but opposite ideological orientation were matched. This pairing reduces confounding effects between ideology and journalistic quality, ensuring that observed differences in bias perception are not trivially attributable to differences in factual accuracy or reporting standards.

FIGURE 3.8: Zoom on selected outlets from the Spanish Media Bias Chart of Figure 3.7 to illustrate the symmetric source-pairing strategy adopted in MBBMD.



For each topic–outlet pair, a single representative article was selected. Articles were required to focus primarily on the selected topic rather than mentioning it tangentially, and preference was given to texts exhibiting sufficient rhetorical or editorial richness to support bias annotation. When multiple candidate articles were available, selection favoured those with clear framing choices or narrative emphasis. A further temporal constraint was applied: for any given topic, all selected articles were required to be published within a short window of one another (typically a few days, never more than two weeks). This constraint is deliberate; widening the window would let unrelated news cycles, evolving political contexts, or post-hoc editorial reinterpretations leak into the corpus and confound the bias signal we want to isolate, since outlets reacting to the same event a month apart are no longer covering the same event in the same context. Restricting the within-topic window therefore ensures that the comparative reading of paired left/right articles measures editorial framing of a shared event rather than temporal drift in the news ecosystem. This procedure resulted in topic-specific article sets composed of ideologically contrasted, factually comparable, and temporally aligned texts, forming the backbone of the dataset.

▸ **Counterfactual Data Augmentation implementation**

Counterfactual Data Augmentation (CDA) is implemented in MBBMD as a controlled experimental instrument rather than as a data expansion strategy. For each selected topic, three counterfactual instances were manually constructed, each targeting a different contextual cue while preserving the article's propositional content and overall coherence. The first variant performs entity swapping, replacing named entities with ideologically or contextually contrasting alternatives while maintaining the original syntactic structure. The second variant applies source masking, substituting the outlet attribution with either a fictitious outlet or one associated with a different ideological orientation. The third variant introduces terminology modification, replacing emotionally or politically charged terms with semantically close alternatives to neutralize lexical framing effects.

These transformations preserve the propositional content of the article while selectively modifying contextual or lexical cues known to influence bias perception. All counterfactual variants were generated manually to ensure linguistic coherence and to avoid unintended semantic drift.

▸ **Annotation phases**

Annotation was conducted in two sequential phases corresponding to document-level and sentence-level analysis. Each phase follows a distinct epistemic rationale, reflecting the asymmetric treatment of subjectivity across annotation levels. All annotations were performed using a dedicated online annotation platform supporting multi-annotator workflows, versioning, and discussion.

Before the main annotation campaign began, every annotator went through a calibration stage on a small set of pilot articles drawn from outside the final MBBMD corpus. The calibration stage served three purposes: (i) familiarising annotators with the annotation guidelines, (ii) surfacing edge cases and ambiguities so that the guidelines could be refined before any production-grade annotation was collected, and (iii) providing a controlled setting in which the inter-annotator behaviour could be inspected and the guidelines updated until divergent readings were reduced to genuine perspectivist disagreement (which we want to preserve) rather than to misunderstandings of the schema (which we want to eliminate). Pilot annotations were not included in the released dataset; they were used exclusively to stabilise the guidelines and to calibrate the annotator pool, and the iteration ran until residual disagreement was attributable to interpretation rather than to schema confusion.

In the first phase, articles were annotated at the document level for the presence of bias and for document-level bias categories. Each article was independently annotated by at least three annotators recruited to reflect diversity in political orientation, age, gender, and educational background. Disagreement was explicitly preserved following the Learning with Disagreements (LeWiDi) framework. For each label, both majority-vote decisions and annotator agreement percentages were retained. This phase models bias as a perceptual

and interpretive phenomenon, capturing polarisation and ambiguity rather than enforcing artificial consensus.

The second phase focuses on fine-grained sentence-level annotation and includes only articles previously labelled as biased at the document level. From the full media bias taxonomy, six manifestations were operationalised at the sentence level: omission of source attribution, sensationalism or emotionalism, mind reading, biased word choice or labeling, subjective qualifying adjectives, and opinions presented as facts. These six were selected because they satisfy three criteria: (i) they are linguistically explicit, identifiable within sentence boundaries through observable lexical, syntactic, or propositional cues; (ii) they are reproducibly annotable at scale without requiring specialised domain expertise; and (iii) they do not require cross-document comparison or access to multimodal information. The remaining eleven manifestations from the taxonomy (including gatekeeping, coverage imbalance, systematic omission, visual framing, image selection, and layout-dependent cues) were excluded from sentence-level annotation because they either require cross-document or temporal aggregation, depend on multimodal signals unavailable in a text-only corpus, or cannot be reliably localised within individual sentences. These excluded manifestations are nonetheless addressed at the document level through the annotation of bias categories (Level 2), where annotators assess editorial strategies after reviewing all articles covering the same topic.

Sentence-level annotation followed a deterministic protocol that is intentionally different from the perspectivist document-level protocol described above. At this layer we are not modelling perception, but the presence or absence of an observable linguistic mechanism (Subsection 3.2.1), so the two-annotators-with-adjudication scheme is the appropriate choice rather than majority voting over a wider pool: each sentence was independently annotated by two annotators and, in cases of disagreement, a third annotator acted as adjudicator to establish the final label. This adjudication process produces a stable reference set suitable for training and evaluation, complementing the perspectivist document-level annotations rather than competing with them. We do not aggregate sentence-level labels by majority vote because, at this layer, disagreement is treated as a schema or attention failure to be adjudicated, not as a perspectivist signal to be preserved; that distinction is the operational expression of the asymmetric treatment of subjectivity discussed in Subsection 3.2.1.

Throughout both annotation phases, quality control was ensured through iterative guideline refinement and periodic review of annotation batches, in continuity with the calibration stage described above. Recurrent ambiguities or inconsistencies were addressed by updating the annotation guidelines and providing additional examples. Final versions of the annotation guides are provided in the appendices to support transparency and reproducibility.

To ensure transparency, all annotation procedures are documented in detailed annotation guidelines provided as appendices. The document-level annotation protocol, including media bias definitions, decision criteria, and illustrative examples, is described in Appendix A.1. The sentence-level annotation schema, including operational definitions of each manifestation and adjudication rules, is detailed in Appendix A.2. These guides constitute the normative reference for all annotations in MBBMD.

Also, an exhaustive description of the corpus composition, including the full list of media outlets, topics, and counterfactual instantiations, is deferred to Appendix B.

This construction methodology translates the design principles of MBBMD into a coherent and reproducible dataset creation pipeline. By explicitly controlling topic selection, source pairing, counterfactual generation, and annotation procedures, MBBMD provides a structured foundation for studying media bias as a hierarchical and perspectivist phenomenon.

3.2.3 *Dataset structure and annotation schema*

The internal structure of MBBMD reflects its hierarchical and perspectivist conception of media bias. Rather than organizing the corpus as a flat collection of instances with independent labels, the dataset is explicitly structured to encode relationships between document-level perceptions, and sentence-level linguistic realizations. This organization enables analyses that

move across levels of abstraction and supports modelling approaches that exploit hierarchical dependencies rather than treating bias detection as a single-step classification problem.

At the highest level, MBBMD is organized around complete news articles as the primary discourse units. Each article constitutes a coherent communicative act in which bias may emerge. For this reason, document-level annotation forms the backbone of the dataset and determines the inclusion and treatment of lower-level annotations.

The dataset is divided into three main subsets: *train*, *test*, and *control*. The training and test subsets contain only original, unmodified articles and are intended for standard model development and evaluation. The control subset contains exclusively counterfactually modified articles generated through systematic CDA procedures. This separation ensures that counterfactual variants are used solely for robustness analysis and causal probing, avoiding contamination of training or evaluation splits and preserving the interpretability of experimental results.

Each subset is further structured into three hierarchical annotation levels, which operationalise the distinction between bias perception (Level 1, document-level binary), editorial bias strategies (Level 2, document-level per-category binary, multi-label across categories), and linguistic realisation (Level 3, sentence-level per-manifestation binary, multi-label across manifestations). The three levels are nested by construction: Level 3 is annotated only on sentences within articles that received a positive Level 1 label, and Level 2 is annotated only on those same articles, so the categorical and linguistic layers always refine, rather than contradict, the document-level perception layer. The levels are summarised below.

Level 1: Binary classification at document level. At the first level, annotators assess each article for the presence or absence of media bias as a whole. The resulting annotation is binary at the article level (each article is either biased or unbiased), and the binary nature reflects how readers report perceiving an article holistically rather than a theoretical reduction; the underlying interpretive variation is captured by the perspectivist representation described below. Articles are assigned one of the following labels:

- **Biased:** Articles containing any detectable form of media bias, regardless of its specific manifestation or intensity.
- **Unbiased:** Articles that present information in a balanced and objective manner.

For each article, the dataset stores all individual annotator judgements, the majority-vote label, and the percentage of annotators identifying the article as biased. This representation preserves disagreement as a signal while remaining compatible with conventional supervised learning settings.

Level 2: Multi-label per-category classification at document level. At the second level, articles identified as biased at Level 1 are further annotated with one or more editorial bias categories corresponding to higher-level dimensions of media bias grounded in the taxonomy introduced in Subsection 3.1.3 of this chapter. The annotation is per-category binary (each of the five categories below is independently marked as present or absent), so a single article can carry multiple positive labels: this is what we call a multi-label setup at the document level, and not a multi-class single-label one (which would force a winner-take-all assignment of one and only one category per article). These categories capture bias mechanisms that operate beyond isolated sentences and are organised along two complementary dimensions:

- **Intention-based biases:**
  - **Intentional bias:** Systematic distortion of information driven by an explicit ideological, political, or strategic agenda.
  - **Spin bias:** Selective emphasis, exaggeration, or narrative shaping intended to steer interpretation without explicit falsification.
- **Context-based biases:**

  - **Statement bias:** Subtle framing effects arising from linguistic or propositional choices that influence interpretation.
  - **Coverage bias:** Disproportionate attention to specific topics, actors, or viewpoints relative to plausible alternatives.
  - **Gatekeeping bias:** Editorial decisions determining which events, perspectives, or voices are included or excluded.

Articles may exhibit multiple bias categories simultaneously, reflecting the fact that different editorial strategies frequently co-occur within the same text. As in Level 1, individual annotator judgments are preserved alongside aggregated statistics.

Level 3: Fine-grained sentence-level linguistic manifestations. The third level grounds document-level bias perceptions in concrete linguistic realizations. Individual sentences within biased articles are annotated deterministically for the presence or absence of a restricted set of linguistically explicit bias manifestations. These manifestations correspond to observable textual mechanisms and are summarized as follows:

- **Omission of source attribution:** Absence of explicit attribution for claims requiring evidential support.
- **Sensationalism or emotionalism:** Use of emotionally charged or exaggerated language that amplifies affective impact.
- **Mind reading:** Attribution of intentions, motivations, or beliefs without explicit evidence.
- **Word choice or labeling:** Lexical selections that implicitly frame entities or events evaluatively.
- **Subjective qualifying adjectives:** Evaluative modifiers that introduce opinionated judgments.
- **Opinions presented as facts:** Subjective assessments expressed using factual or assertive constructions.

Each annotated sentence is stored together with limited local context, consisting of the immediately preceding and following sentences, to support discourse-aware and context-sensitive modelling.

The relationship between annotation levels is explicitly encoded through shared identifiers. Each sentence-level instance is linked to its parent article via a unique text identifier, enabling aggregation, alignment, and cross-level analysis. This design supports hierarchical modelling approaches in which sentence-level predictions may be conditioned on document-level context, or conversely, document-level predictions may be explained through the aggregation of sentence-level cues.

From a data representation perspective, all annotation levels are stored in tab-separated values (TSV) format to ensure simplicity, transparency, and broad compatibility. Identifiers are consistent across files, enabling straightforward joins between levels and subsets using standard NLP tooling.

At the document level, each record includes the topic, article identifier, headline, full article text, and annotation fields corresponding to media bias perception and media bias categories. At the sentence level, records include the article identifier, outlet, publication date, sentence identifier, sentence text, local context, and binary indicators for each manifestation of media bias. This representation supports both single-label and multi-label sentence classification, as well as sequence-based or span-based modelling extensions.

3.2.4 *Corpus analysis and statistics*

MBBMD comprises 100 Spanish-language news articles and approximately 2,000 annotated sentences, distributed across ten politically and socially salient topics spanning domestic politics, international affairs, social legislation, environmental crises, cultural controversies, and sports-related events with institutional implications. This topical diversity is intentionally designed to expose heterogeneous bias dynamics, ranging from highly polarised ideological

TABLE 3.1: MBBMD inter-annotator agreement metrics across topics.

Topic	Fleiss' Kappa	Cohen's Kappa	Consensus
Spain's Public Debt	-0.067	-0.024	81.25%
Eurovision Song Selection	0.072	0.114	76.67%
Sinn Féin Referendum	0.132	0.256	83.33%
Catalonia Emergency	0.285	0.446	88.75%
Luis Rubiales	-0.049	-0.016	90.00%
Amnesty Protest	0.165	0.226	74.44%
Trans Law Approval	0.330	0.364	80.00%
Pedro Sánchez Investiture	0.406	0.466	82.50%
Argentina Elections	0.419	0.464	85.00%
ICJ Orders on Gaza	0.180	0.217	76.25%

conflicts to more procedural or informational reporting contexts. Articles vary substantially in length, rhetorical density, and narrative structure, reflecting the heterogeneity of contemporary Spanish journalism and preventing the dataset from collapsing into narrow stylistic regularities.

At the document level, annotator judgments exhibit substantial variability across topics, confirming that media bias perception is neither uniform nor trivially determined by surface features. The annotation pool was deliberately composed to reflect ideological, demographic, and educational diversity: annotators span the political spectrum from far-left to right (with a concentration in the left and centre positions), are balanced by gender (5 female, 5 male, 2 undisclosed), cover a range of age groups (from 18–24 to 65+, with the largest group in 25–39), and represent multiple educational backgrounds (from secondary education to postgraduate degrees, with postgraduates being the largest group). Annotation effort is distributed across all topics, with centrist annotators contributing the highest number of individual annotations. This diversity is a deliberate design choice, aimed at ensuring that disagreement captures genuine interpretive plurality rather than artefacts of annotator homogeneity.

Agreement patterns vary across topics. Table 3.1 reports Fleiss' and Cohen's Kappa scores together with consensus percentages for each topic. Highly polarised topics such as Spain's public debt, the Luis Rubiales case, or international legal actions related to Gaza; exhibit low or even negative Kappa values despite relatively high consensus rates. This apparent paradox highlights a well-known limitation of chance-corrected agreement metrics in subjective annotation tasks: annotators may agree on final labels while relying on different interpretive rationales. Conversely, topics such as the Argentine presidential elections or Pedro Sánchez's investiture display moderate agreement levels, suggesting more stable framing cues and less interpretive dispersion.

These patterns reinforce the central epistemic assumption of MBBMD: disagreement is not annotation noise but a meaningful signal of interpretive uncertainty, ideological distance, and narrative contestation. Topics with lower agreement tend to exhibit higher variance in bias perception, reflecting the interplay between framing strategies, reader expectations, and socio-political salience. Figures 3.9 and 3.10 visualise these dynamics: the first shows annotator consensus levels across topics, while the second shows the distribution of document-level bias judgments for each topic.

Counterfactual Data Augmentation (CDA) plays a critical role in the empirical analysis by enabling controlled comparisons between original and modified articles. Across topics, CDA variants consistently reduce consensus levels, indicating increased interpretive uncertainty when contextual cues are manipulated. Outlet masking produces some of the strongest effects, with average consensus drops of approximately 21%, demonstrating that source reputation and ideological expectations significantly shape bias perception independently of textual content. Entity swaps further amplify bias detection in politically charged contexts, with substitutions such as Mariano Rajoy → José Luis Rodríguez Zapatero yielding increases of up to 35% in perceived bias. Terminological modifications similarly confirm the sensitivity of bias perception to emotionally loaded lexical choices. Figures 3.11 and 3.12 quantify these

effects: the first shows consensus drops in counterfactual articles relative to their originals, and the second shows the direction and magnitude of the perception shifts induced by each CDA variant (outlet masking, entity swap, terminological modification).

FIGURE 3.9: MBBMD annotator consensus levels across topics.

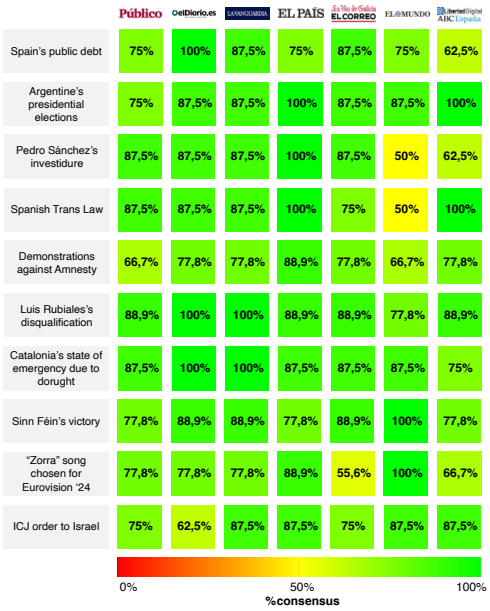


FIGURE 3.10: Distribution of document-level bias across topics.

FIGURE 3.11: Annotator consensus in counterfactual articles.

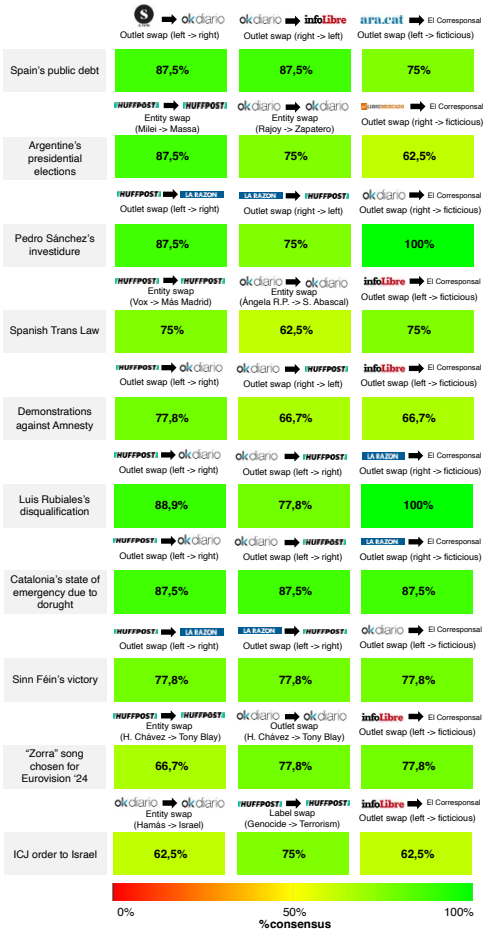
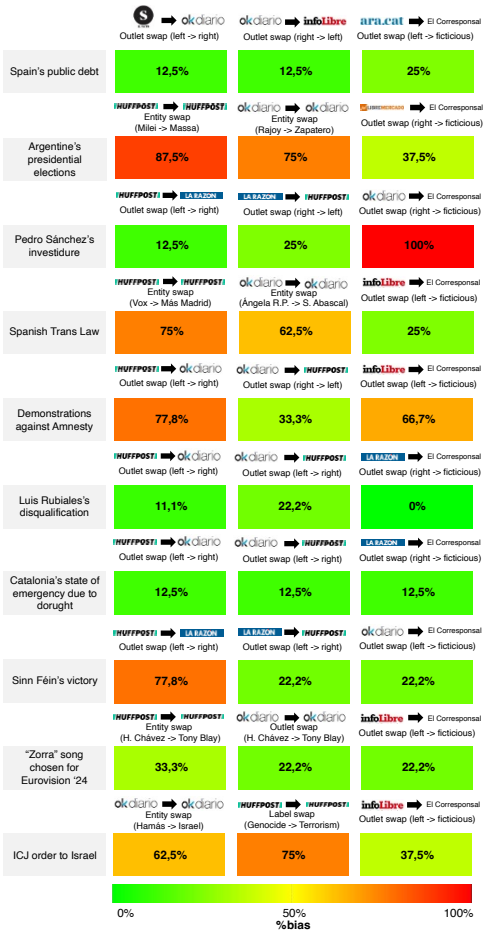


FIGURE 3.12: Bias perception shifts induced by CDA.



Sentence-level analysis complements these findings by grounding global bias perceptions in concrete linguistic realisations. The sentence-level component of MBBMD comprises 2,348

annotated sentences distributed across training, test, and control subsets, as summarised in Table 3.2. Only a small proportion of sentences are explicitly annotated as biased, confirming that bias is typically sparse and local rather than uniformly distributed across a text. The three-way split serves distinct purposes: the training and test subsets contain only original, unmodified articles and support standard model development and held-out evaluation, ensuring reproducibility of reported results. The control subset contains exclusively counterfactually modified articles generated through CDA procedures, enabling robustness analysis and causal probing without contaminating the training or evaluation data. This separation is preserved in the public release of the dataset on Zenodo [142], where each subset is distributed as an independent file with consistent identifiers to facilitate cross-level joins. A detailed description of the dataset and its design rationale has been published in Rodrigo-Ginés et al. [143].

[142] Rodrigo-Ginés et al., 2025, *MBBMD: Media Bias Bias-Mitigated Dataset (v2.0.0)*

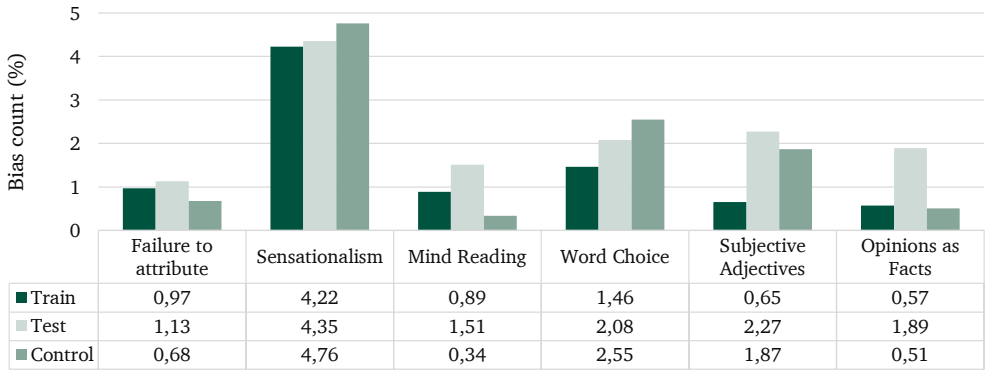
[143] Rodrigo-Ginés et al., 2026 "Media Bias Bias-Mitigated Dataset (MBBMD): A Hierarchical, Perspectivist, and Counterfactually-Augmented Corpus for Bias Detection in Spanish News"

TABLE 3.2: Summary statistics of sentence-level subsets.

Dataset	Sentences	Topics	Outlets	Biased Sentences (%)
Train	1,231	7	10	5.85%
Test	529	3	8	9.07%
Control	588	10	6	7.99%

Figure 3.13 presents the distribution of the six sentence-level bias manifestations across subsets. Sensationalism and emotionalism are by far the most frequent manifestations, accounting for over 4% of sentences in all splits. This prevalence aligns with prior research identifying affective framing as a dominant mechanism of media bias. Other manifestations, such as omission of source attribution or mind reading, occur less frequently but are disproportionately concentrated in highly polarised topics, where speculative intent attribution and evidential gaps are more common.

FIGURE 3.13: Distribution of sentence-level bias manifestations across subsets.



	Failure to attribute	Sensationalism	Mind Reading	Word Choice	Subjective Adjectives	Opinions as Facts
Train	0,97	4,22	0,89	1,46	0,65	0,57
Test	1,13	4,35	1,51	2,08	2,27	1,89
Control	0,68	4,76	0,34	2,55	1,87	0,51

Structural and lexical properties of the corpus further contextualise these patterns. Average sentence length is relatively stable across subsets, with control articles exhibiting slightly longer sentences due to the preservation of original structure under counterfactual modification. Sentence length distributions show long tails, with occasional sentences exceeding 100 tokens. These complex cases often combine multiple clauses, quotations, and evaluative expressions, posing challenges for both human annotation and automated processing.

Lexical analysis reveals a strong thematic concentration on political actors and institutions, with frequent terms such as *gobierno*, *Sánchez*, *amnistía*, and *Milei*. Biased sentences disproportionately feature terms associated with controversial actors or policies, reinforcing the link between topic salience and bias manifestation.

This corpus analysis positions MBBMD not only as a challenging benchmark but also as a diagnostic resource. It enables researchers to study whether models can cope with ambiguity, disagreement, and contextual sensitivity; properties that are intrinsic to media bias.

3.2.5 *Reproducibility and future extensions*

The design and release of MBBMD are guided by a commitment to reproducibility, transparency, and long-term usability. Given the epistemic complexity of media bias and the

methodological choices involved in its annotation, reproducibility in this context does not merely refer to the ability to rerun experiments, but to the capacity to trace, understand, and critically assess each decision made throughout the dataset lifecycle.

To this end, the dataset is released following the FAIR principles [31]: Findability, Accessibility, Interoperability, and Reusability. All data files are hosted on an open-access repository with persistent identifiers, ensuring long-term availability. The dataset is distributed in simple, non-proprietary TSV format, enabling immediate use across standard NLP pipelines without the need for specialised tooling. File naming conventions, directory structure, and identifier consistency are designed to make relationships between annotation levels explicit and machine-readable.

[31] Chamorro-Padial et al., 2024, "FAIR-Check: script to check FAIR principles compliance"

Reproducibility is further supported through extensive documentation of the annotation process. Detailed annotation guidelines for both document-level and sentence-level tasks are provided as appendices, including operational definitions, positive and negative examples, and clarification protocols used during annotation. These documents do not merely describe the final label schema, but also record the rationale behind key design decisions, such as the asymmetric treatment of subjectivity across annotation levels or the reduction of sentence-level manifestations from a broader theoretical taxonomy. This documentation enables other researchers to reproduce the annotation logic, even if exact replication of annotator judgments is neither expected nor conceptually desirable in a perspectivist setting.

In addition to annotation guidelines, MBBMD includes rich metadata accompanying each instance. At the document level, this includes topic information, outlet identity, publication date, individual annotator judgments, and agreement statistics. At the sentence level, metadata includes contextual information in the form of neighbouring sentences, as well as explicit links to parent documents. Counterfactual instances are explicitly marked and separated from original articles, ensuring that their role as analytical controls is preserved. This metadata design supports a wide range of downstream uses, from uncertainty-aware modelling and robustness evaluation to qualitative discourse analysis.

Beyond reproducibility, MBBMD is designed as a foundation for future extensions. One natural direction concerns topical expansion. While the current version covers ten salient topics selected for diversity and analytical richness, future releases may incorporate additional events and domains, enabling broader coverage and facilitating studies on topic-dependent bias dynamics. Such expansions would preserve the existing annotation schema to ensure compatibility while increasing the dataset's empirical scope.

A second extension avenue involves temporal expansion. Media bias is not static; it evolves in response to political cycles, social movements, and shifting public discourse. By extending MBBMD with articles covering the same topics across different time periods, the dataset could support longitudinal analyses of bias evolution, framing drift, and agenda-setting effects. The hierarchical structure of MBBMD is particularly well suited to this goal, as it allows changes in sentence-level realizations to be analysed in relation to shifts in document-level perception.

Finally, multilingual extension represents a promising direction. Although MBBMD currently focuses on Spanish-language media, its conceptual framework is language-agnostic. Future work may introduce parallel corpora in other languages, enabling cross-linguistic and cross-cultural studies. Such extensions would also facilitate research on cross-lingual transfer, domain adaptation, and the limits of multilingual bias detection models.

This page intentionally left blank.

4

Closing the cross-dataset gap: a hierarchical perspectivist system for media bias detection

"All meaning is relational."
—Ferdinand de Saussure

The taxonomy and dataset introduced in Chapter 3 provide the conceptual backbone of this thesis, but they do not, by themselves, answer the question the field has been struggling with for a decade: can we build media bias detection systems that generalise beyond the dataset they were trained on, and, if so, how much of the improvement is attributable to the system and how much to the dataset? This chapter is devoted to answering that question empirically, treating the five structural problems identified in Section 2.3 (conceptual fragmentation, cross-dataset failure, the ground-truth fallacy, anglocentrism, and evaluation opacity) not as abstract limitations of the field but as concrete design constraints that any robust system must satisfy.

The chapter is organised around three questions, each in direct correspondence with one of those structural problems and each addressed in a dedicated section:

- **Q1 (in response to Problems 2 and 5 of Section 2.3).** How badly do current media bias detectors fail when evaluated cross-dataset rather than in-distribution, and what does the pattern of failures reveal about what these systems are actually learning? This is the question that the field's prevailing in-distribution evaluation regime systematically hides; Section 4.1 addresses it through a systematic cross-dataset study across five benchmarks.
- **Q2 (in response to Problems 1 and 3 of Section 2.3).** Once a dataset preserves the hierarchical and perspectivist structure of media bias (rather than collapsing it into a single document-level majority-vote label), which modelling paradigms can actually exploit each analytical layer, and where do flat single-level approaches break down? Section 4.2 addresses it through a level-by-level analysis on MBBMD that recovers the dataset's three annotation layers (sentence-level manifestations, document-level perception, and document-level categories), reflecting the conceptualisation developed in Chapter 3.
- **Q3 (synthesising the answers to Q1 and Q2).** Given the failure modes diagnosed under Q1 and the level-by-level findings of Q2, how should a hierarchical and perspectivist media bias detection system be designed, and does its evaluation under the same protocol used for the diagnosis show that this design closes the cross-dataset gap? Section 4.3 presents the system developed in this thesis in order to translate the diagnosis into a concrete architecture, while Chapter 5 reports its cross-dataset evaluation, the dataset-versus-system attribution, and the discussion of the three research hypotheses formulated in Chapter 1.

Each section therefore answers a different question, but the three are tightly coupled: the cross-dataset diagnosis of Section 4.1 establishes *why* flat document-level modelling is inadequate, the level-by-level analysis of Section 4.2 establishes *which* signals at *which* level are exploitable once a hierarchical perspectivist dataset becomes available, and the system design of Section 4.3 translates those findings into an architecture whose components are introduced precisely as responses to the empirical regularities documented in the previous two sections.

4.1 The limits of current systems: a cross-dataset study

As discussed in Chapter 2, media bias is a complex phenomenon that lacks a universally agreed-upon definition. This ambiguity complicates computational modeling, as the criteria

used to annotate bias in one dataset may differ substantially from those applied in another. Existing datasets have been constructed using diverse annotation schemes and theoretical assumptions, raising critical concerns about the generalization capabilities of models trained on them.

The generalization problem is not merely technical; it reflects a deeper structural challenge (Problem 2 of Section 2.3). If a model trained on one dataset cannot reliably detect bias in texts from a different domain, language, or media ecosystem, its practical utility is fundamentally limited. Yet, as the previous chapter has shown, the field has predominantly evaluated models under in-distribution conditions, where training and testing data share similar characteristics. This evaluation paradigm obscures the fragility of learned representations and may lead to an overly optimistic assessment of model capabilities.

This section systematically examines the extent to which models trained on one dataset can generalize to others. This evaluation constitutes one of the original empirical contributions of this thesis. Our previous work [151] provided initial evidence that models trained on media bias detection datasets may primarily capture journalistic style or regional writing conventions rather than the actual presence of media bias. To further investigate this challenge, we present a series of cross-dataset evaluation experiments in which models are trained on individual datasets and tested on unseen ones, assessing whether they learn transferable bias-related features or merely memorize dataset-specific patterns. The findings directly motivate the design of the hierarchical and perspectivist system introduced in Section 4.3.

[151] Rodrigo-Ginés et al., 2023, "Identifying Media Bias beyond Words: Using Automatic Identification of Persuasive Techniques for Media Bias Detection"

4.1.1 *Experimental design*

Understanding the generalization capabilities of media bias detection models requires a systematic and thorough experimental framework. Given the complexity of bias as a phenomenon, which is influenced by linguistic structures, cultural contexts, and journalistic conventions, our goal in this section is to examine whether models trained on a specific dataset are able to generalize effectively when applied to previously unseen datasets. A model that successfully generalizes should be capable of identifying bias-related patterns that transcend the stylistic, thematic, and ideological peculiarities of individual datasets. Conversely, if a model exhibits strong performance only on the dataset it was trained on but fails when applied to others, this would suggest that it has merely memorized dataset-specific linguistic markers rather than learning generalized bias detection principles.

To explore these generalization challenges, we conducted a series of cross-dataset evaluations using a encoder-only transformer-based architecture. We selected DistilBERT (distilbert-base-multilingual-cased) [162], a lightweight and computationally efficient variant of BERT, which retains much of the representational power of its larger counterparts while significantly reducing training costs. All models were trained under identical conditions to ensure that performance differences stem from dataset properties rather than variations in hyperparameter selection. The training process was carried out using the following configuration:

[162] Sanh et al., 2019, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter"

- **Learning rate**: 5e-5
- **Batch size**: 16
- **Epochs**: 5

These parameters were chosen based on common best practices in transformer fine-tuning and prior empirical studies on similar NLP tasks [206, 174].

In addition to fine-tuning encoder-only transformers models, we investigated the effectiveness of decoder-only models in bias detection using in-context learning. Recent advancements in Large Language Models (LLMs) suggest that these systems acquire substantial world knowledge and linguistic representations through extensive pretraining [33]. To test this hypothesis, we evaluated GPT-4o in a 1-shot in-context learning [46, 158] setting, where the model was provided with a structured prompt that includes an example to guide its inference process.

Unlike fine-tuned models, which rely on explicit dataset annotations, GPT-4o's predictions stem from its internal representations of language and discourse structures. The in-context learning setup provides the model with a concrete example of biased reporting, enabling it

[206] Zhang et al., 2020, "Revisiting few-sample BERT fine-tuning"

[174] Sobhanam and Prakash, 2023, "Analysis of fine tuning the hyper parameters in RoBERTa model using genetic algorithm for text classification"

[33] Chang et al., 2024, "How Do Large Language Models Acquire Factual Knowledge During Pretraining?"

[46] Dong et al., 2022, "A survey on in-context learning"

[158] Rubin et al., 2021, "Learning to retrieve prompts for in-context learning"

to recognize similar patterns in unseen news stories. This approach aims to assess whether LLMs can generalize bias detection principles based solely on prior exposure to large-scale textual corpora.

The following system prompt was used for GPT-4o:

> **Prompt:** Evaluate whether a given news story is biased or impartial. Analyze the text to identify signs of bias such as subjective language, unsupported opinions, or imbalance in viewpoints. Consider contextual information, loaded terms, and the representation of different perspectives.
>
> **Output Format:** Provide a structured response indicating whether the news story is biased or not. Include reasoning focusing on identified indicators of bias.
>
> - **Reasoning:** A brief paragraph explaining your conclusion with specific textual examples.
> - **Output:** "Biased" or "Not Biased"
>
> **Example:**
>
> **Input:** "The government's new education reform has sparked widespread outrage due to its clearly unjust budget cuts. Critics argue that this shocking decision disproportionately affects vulnerable communities, while supporters claim it will increase efficiency. However, the overwhelming consensus is that this outrageous reform only deepens inequality."
>
> **Output:**
>
> - Reasoning: "The news story is biased as it uses emotionally charged and subjective terms such as 'outrage', 'unjust', 'shocking', and 'outrageous'. It also predominantly highlights negative opinions ('overwhelming consensus') without providing sufficient evidence or clearly representing the viewpoints of supporters of the reform."
> - Output: "Biased"

This comparison serves two purposes: first, to assess whether LLMs inherently capture bias-related cues through their pretraining, and second, to provide a baseline for evaluating how much improvement is gained by fine-tuning on domain-specific datasets.

▸ Datasets for training and evaluation

An important aspect of this study was the selection of datasets, as the quality and nature of training data directly impact a model's ability to generalise. Accordingly, the datasets described in Subsection 2.2.3 (where the case for treating them as a joint benchmark, despite their incompatible annotation schemes, was already developed) constitute the basis of our benchmarking framework here. We use exactly those five datasets, with no further additions, so that the evaluation reported in this chapter remains directly comparable with the state of the art surveyed in Chapter 2.

To ensure clarity throughout this study, each selected dataset has been assigned a label, as shown in parentheses:

- **MBIC** (**A**) [181]: A dataset explicitly designed for sentence-level media bias detection, featuring fine-grained annotations for multiple types of bias.
- **Crisis in Ukraine dataset** (**B**) [37]: A dataset that captures ideological bias in news coverage of the Ukraine conflict. Originally annotated for ideological leanings (*pro-Western*, *pro-Russian*, and *neutral*), it has been adapted to a binary classification scheme (*biased* vs. *non-biased*) to ensure consistency with other datasets.

[181] Spinde et al., 2021, "MBIC–A Media Bias Annotation Dataset Including Annotator Characteristics"

[37] Cremisini et al., 2019, "A challenging dataset for bias detection: the case of the crisis in the Ukraine"

[205] Zaghouani et al., 2024, "The FIGNEWS Shared Task on News Media Narratives"

- **FIGNEWS** (**C**) [205]: A multilingual dataset covering politically sensitive discourse on the Israel-Gaza war. While the original dataset includes ideological annotations, for consistency, it has been reformulated into a binary *biased* vs. *non-biased* classification. A subset of 50,000 instances was extracted to maintain balance with other datasets.
- **BEADs** (**D**) [141]: A dataset designed to capture various dimensions of bias, including political bias, hate speech, and toxicity. To align with the scale of other datasets in this study, a 50,000-instance subset was selected.
- **MBBMD** (**E**): Our proposed dataset (see Section 3.2), which introduces a hierarchical and perspectivist annotation framework to address existing limitations in media bias detection datasets. For the cross-dataset experiment reported in this section, MBBMD enters the benchmark exclusively in its flattest form: each document is represented by a single binary biased/unbiased label obtained by majority vote across annotators, mirroring the annotation regime of the other four corpora and ensuring that all datasets are treated under identical conditions. The hierarchical and perspectivist structure of MBBMD (sentence-level manifestations, document-level annotator distributions, and document-level categories) is deliberately set aside here so that the cross-dataset comparison remains apples-to-apples; it is reactivated and exploited in the level-by-level analysis of Section 4.2 and in the system design of Section 4.3. Concretely, MBBMD comprises 100 Spanish-language news articles drawn from ten politically and socially salient topics, each annotated by at least three annotators with declared ideological self-identification, and 2,348 sentence-level annotations of bias manifestations distributed across train, test, and counterfactual control subsets (Subsection 3.2.4). Document-level class balance is approximately equal between biased and unbiased articles.

[141] Raza et al., 2024, "BEADs: Bias Evaluation Across Domains"

These datasets were chosen for their thematic and linguistic diversity, public availability, and reproducibility. They provide an optimal testing ground for evaluating whether models learn generalisable bias patterns or merely internalise dataset-specific stylistic features.

To assess cross-dataset generalisation, we systematically constructed multiple training and evaluation configurations. In addition to training models on individual datasets (A, B, C, D, E), we experimented with aggregated training sets (ABCD, BCDE, CDEA, DEAB, EABC) and a fully merged training set (ABCDE). Each configuration was used as the training set, while the held-out 20% partitions of the remaining datasets served as independent test sets, so that no test instance was ever seen at training time even when datasets were combined. The aggregated training sets are constructed as the union of the 80% training partitions of their component datasets; ABCDE includes all five 80% training partitions, and the corresponding ABCDE test set is the union of the five disjoint 20% test partitions. This framework allows us to analyse whether bias-related features generalise across datasets and whether increasing dataset diversity enhances robustness or introduces conflicting signals [189].

[189] Torralba and Efros, 2011, "Unbiased look at dataset bias"

Through this comprehensive experimental design, this study aims to contribute to a deeper understanding of media bias detection beyond single-dataset training. The findings will clarify whether current computational approaches can effectively identify bias as a cross-linguistic, cross-regional phenomenon or whether they remain constrained by dataset-specific biases. Addressing these challenges is essential for advancing bias detection methodologies and developing models that can operate reliably across diverse media sources.

4.1.2 *Results of cross-dataset evaluation*

To ensure a fair evaluation, every dataset was first partitioned into a fixed 80%/20% train/test split with stratified sampling on the bias label, and these splits were kept identical across all configurations reported in this section. The 80% partitions are used exclusively for training (either individually, or unioned to form the multi-dataset training sets ABCD, BCDE, CDEA, DEAB, EABC, and ABCDE introduced above), while the 20% partitions are reserved exclusively for evaluation. When a model is trained on dataset A and evaluated on dataset B, it is the 20% test partition of B that serves as the evaluation set, never the full corpus, so that any subsequent multi-dataset training that happens to include B will still encounter the same disjoint test partition at evaluation time. At no point does a model

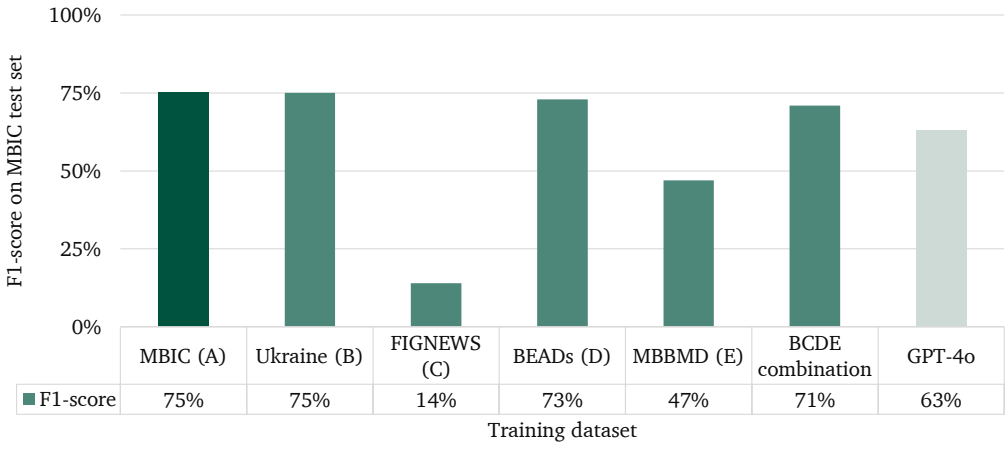
encounter the same example during both training and testing, eliminating any form of leakage even under aggregated training. Additionally, to account for variability due to initialisation and optimisation dynamics, each fine-tuning experiment was executed three times with different random seeds; the values reported throughout are the mean across those three runs. Performance differences are therefore attributable to the ability of the models to transfer bias detection knowledge across datasets rather than to memorisation of specific instances or to seed-level noise. All F1 scores reported in this section are macro-averaged over the two classes (biased and non-biased), so that performance on either class contributes equally regardless of the class balance of the test set; the underlying task is binary classification (biased vs. non-biased) for every dataset.

▸ Results of cross-dataset evaluation for MBIC (A)

Models tested on MBIC, which primarily focuses on sentence-level bias detection, exhibit significant variation in performance depending on the training dataset. As shown in Figure 4.1, the highest F1-score (0.751) is achieved when the model is trained on MBIC itself, indicating strong in-domain performance. However, models trained on other datasets struggle to generalize to MBIC, with particularly low results when trained on FIGNEWS (0.138) and MBBMD (0.469). In the case of MBBMD, this drop may be attributed to the language mismatch, as MBBMD is developed in Spanish while MBIC, like most other benchmark datasets, is in English. This suggests that sentence-level bias detection in MBIC relies on linguistic features and cues that may not transfer well across languages or from document-level and event-based corpora.

Interestingly, the model trained on BEADs performs comparatively well on MBIC (0.726), hinting at a certain degree of overlap between MBIC's fine-grained bias annotations and BEADs' broader categorizations. Additionally, the multi-dataset training approach (BCDE) improves performance on MBIC (0.713) compared to single-dataset models, suggesting that exposure to multiple bias annotations aids in cross-dataset adaptation. The GPT-4o model achieves an F1-score of 0.634 in its one-shot learning setup, performing better than some single-dataset trained models but still falling short of the fine-tuned transformer models.

Figure 4.1: Cross-dataset performance on MBIC (A).



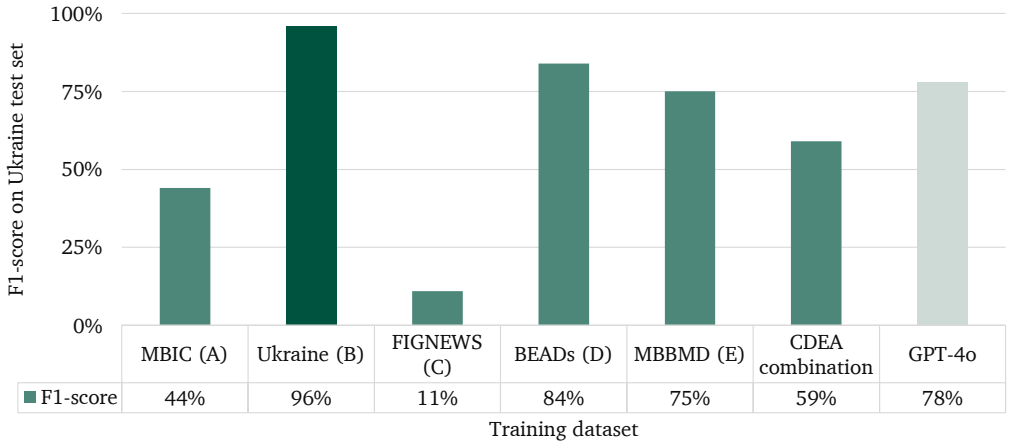
▸ Results of cross-dataset evaluation for the Crisis in Ukraine dataset (B)

Models tested on the Crisis in Ukraine dataset show varied performance depending on the training dataset. As illustrated in Figure 4.2, the highest F1-score (0.962) is achieved when the model is trained on the same dataset, indicating strong in-domain learning. However, models trained on other datasets struggle to generalize effectively to *Crisis in Ukraine*, with particularly low performance when trained on FIGNEWS (0.113) and MBIC (0.438). This suggests that the linguistic and ideological framing characteristics of the Crisis in Ukraine dataset do not align well with those of sentence-level bias detection datasets.

Interestingly, models trained on BEADs (0.839) and MBBMD (0.753) perform significantly better, suggesting that datasets that emphasize political and ideological framing generalize

more effectively to *Crisis in Ukraine*. The mixed-dataset model (CDEA) achieves moderate generalization (0.593), reinforcing the idea that exposure to multiple bias types aids in cross-dataset adaptation. The GPT-4o model, in its one-shot setup, achieves an F1-score of 0.784, showing competitive performance despite not being fine-tuned specifically for this dataset.

FIGURE 4.2: Cross-dataset performance on Ukraine (B).



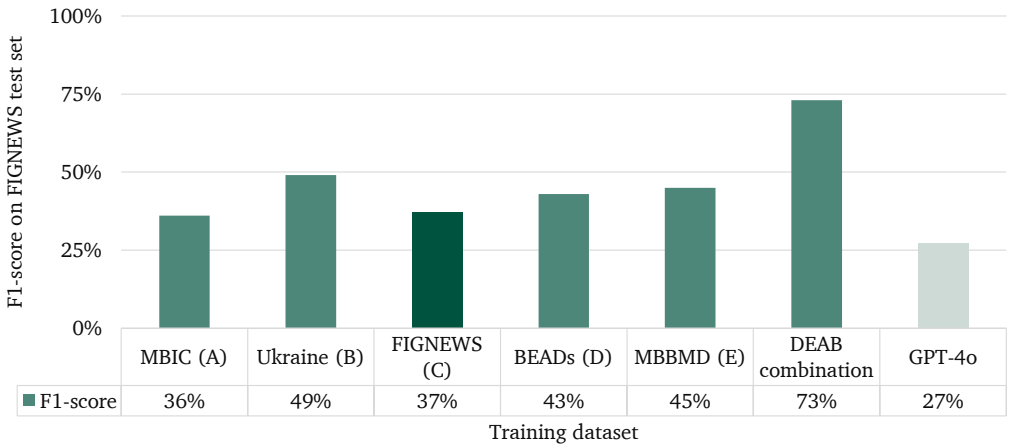
▸ **RESULTS OF CROSS-DATASET EVALUATION FOR FIGNEWS (C)**

Models tested on the FIGNEWS dataset exhibit considerable difficulty in generalizing, regardless of the training dataset. As illustrated in Figure 4.3, no model trained on a different dataset achieves an F1-score above 0.5 when evaluated on FIGNEWS, with performances ranging from 0.357 to 0.485. The best result is obtained when the model is trained on *Crisis in Ukraine* (0.485), but even this does not indicate strong generalization. This suggests that the linguistic and contextual characteristics of FIGNEWS differ significantly from the other datasets.

The particularly low performance of models trained on BEADs (0.433) and MBBMD (0.445) indicates that even datasets covering multiple bias dimensions fail to adapt well to FIGNEWS. This lack of cross-dataset transferability is likely due to the unique stylistic and linguistic features of FIGNEWS, which includes social media discourse and short-form news with a strong multilingual component.

In contrast, when a model is trained on FIGNEWS itself, it achieves an F1-score of 0.374, which is lower compared to most other datasets' in-domain performance. This suggests that bias detection within FIGNEWS is inherently challenging, potentially due to the informal, diverse, and highly context-dependent nature of its text. The mixed-dataset approach (DEAB) achieves the highest transferability (0.733), reinforcing the idea that combining multiple sources of bias may help improve generalization to more diverse datasets.

FIGURE 4.3: Cross-dataset performance on FIGNEWS (C).



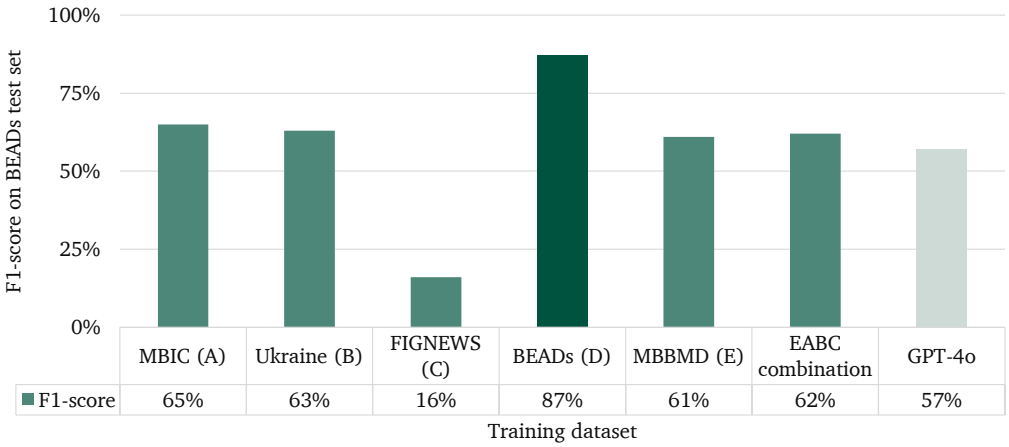
▸ Results of cross-dataset evaluation for BEADs (D)

Models tested on the BEADs dataset generally achieve moderate to high performance, indicating that bias features present in BEADs are more transferable across datasets compared to others such as FIGNEWS. As shown in Figure 4.4, models trained on different datasets display relatively strong performance when evaluated on BEADs, with F1-scores ranging from 0.571 (GPT-4o) to 0.651 (MBIC). The highest performance is obtained when the model is trained on BEADs itself (0.874), confirming the consistency of its internal annotation schema.

Interestingly, models trained on *Crisis in Ukraine* (0.631) and MBBMD (0.605) also perform well on BEADs, suggesting that these datasets capture overlapping linguistic and ideological framing cues. The combined dataset approach (EABC, 0.622) further supports this trend, as training on a mixture of datasets improves overall performance.

However, a notable drop in performance is observed when models trained on FIGNEWS are tested on BEADs (0.157). This highlights a key challenge: datasets with highly distinct linguistic structures and annotation methodologies such as social media-based datasets like FIGNEWS, struggle to transfer their learned representations to more traditional news media datasets like BEADs.

Figure 4.4: Cross-dataset performance on BEADs (D).



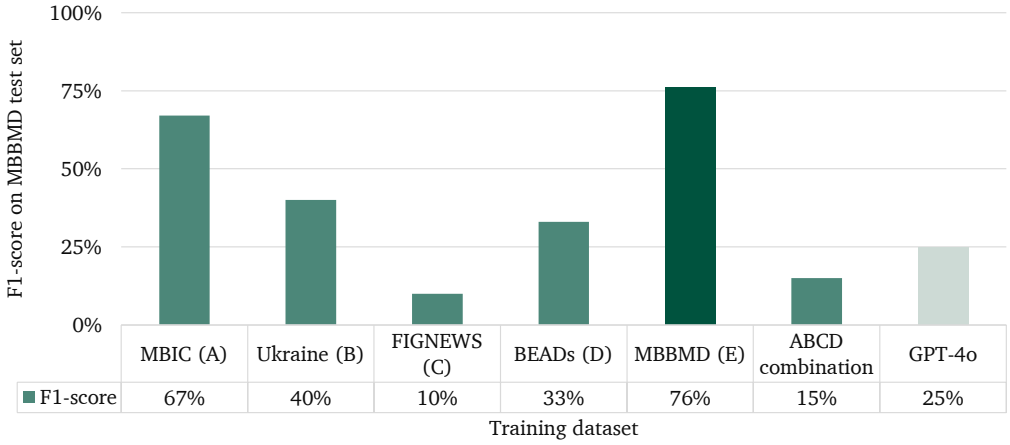
▸ Results of cross-dataset evaluation for MBBMD (E)

Models tested on the MBBMD dataset show varied performance, suggesting that cross-dataset generalization remains a significant challenge. As depicted in Figure 4.5, models trained on MBIC achieve the highest transferability to MBBMD (0.666), while those trained on BEADs perform poorly (0.333). The model trained on the Crisis in Ukraine dataset exhibits moderate performance (0.4), whereas the GPT-4o one-shot model shows even lower effectiveness (0.25). It is important to note that MBBMD is the only dataset in this evaluation developed entirely in Spanish, while all other datasets and models are based on English data. This language mismatch likely contributes to the reduced cross-dataset performance, as linguistic features relevant to bias detection may not transfer effectively across languages.

One of the most striking results is the complete failure of models trained on FIGNEWS when tested on MBBMD (0.1). This aligns with previous observations that FIGNEWS introduces linguistic and stylistic characteristics that are not easily transferable. Similarly, models trained on the combined dataset ABCD also perform poorly (0.153), indicating that merging multiple datasets does not necessarily lead to improved generalization on MBBMD.

These findings highlight that while MBBMD was designed to capture a more structured and hierarchical representation of bias, it does not inherently facilitate better generalization across datasets. Instead, its annotation framework may introduce patterns that are distinct from those used in other corpora, limiting its compatibility with models trained on more conventional bias detection datasets.

FIGURE 4.5: Cross-dataset performance on MBBMD (E).



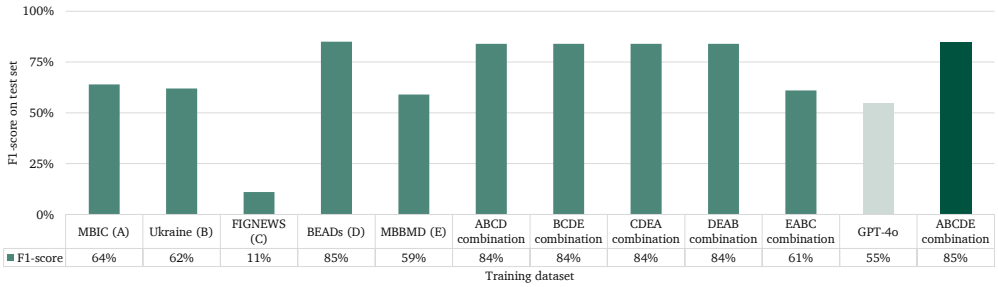
RESULTS OF CROSS-DATASET EVALUATION FOR COMBINED DATASETS

To assess whether evaluating on a more diverse dataset offers additional insight into generalization, we tested each model on a combined dataset comprising all instances from MBIC, BEADs, Crisis in Ukraine, FIGNEWS, and MBBMD (ABCDE). As shown in Figure 4.6, models trained on individual datasets exhibit different levels of generalization when applied to this mixed evaluation set. The best result is achieved by the model trained on BEADs, with an F1-score of 0.847, demonstrating strong cross-dataset transferability. Similarly, models trained on MBIC (0.637), Crisis in Ukraine (0.617), and MBBMD (0.585) also generalize reasonably well, outperforming most other configurations. These results suggest that certain datasets encode more generalizable features for media bias detection across diverse domains and linguistic contexts.

The results indicate that aggregating multiple datasets helps mitigate dataset-specific biases and enhances generalization in many cases. However, the improvement is not uniform across all datasets. The model still struggles significantly when tested on FIGNEWS (0.16), reinforcing the idea that its linguistic and structural characteristics differ substantially from the rest of the datasets. This poor performance suggests that FIGNEWS introduces distinctive patterns that are not well captured by a model trained on a more heterogeneous corpus.

Additionally, comparing the model trained on ABCDE to other multi-dataset combinations (ABCD, BCDE, CDEA, DEAB, and EABC) reveals minimal differences in performance, with all multi-dataset models achieving similar F1-scores. This suggests that while combining datasets does provide generalization benefits, simply increasing dataset diversity does not guarantee proportional improvements. The results highlight that certain datasets contribute more to generalization than others, and that dataset incompatibility remains a major challenge in media bias detection.

FIGURE 4.6: Cross-dataset performance of the combined model (ABCDE).



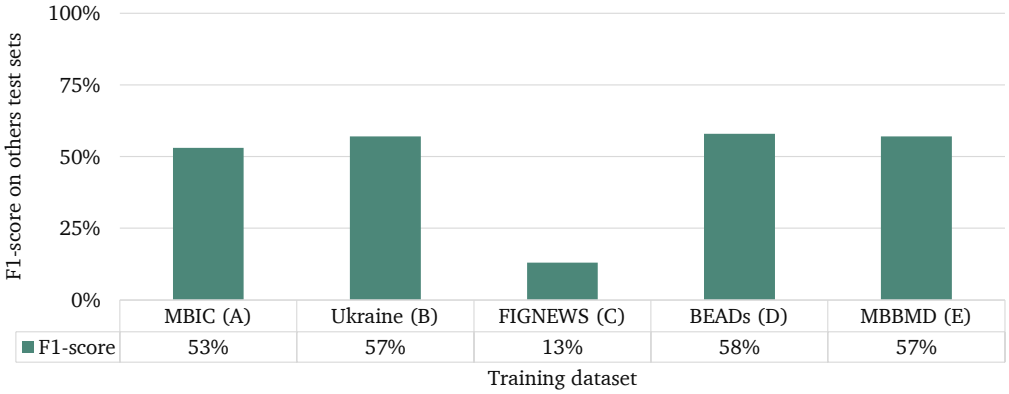
To gain a broader perspective on cross-dataset generalization, Figure 4.7 presents the average F1-score obtained when training on each dataset and testing on all others. The results confirm the variability in dataset transferability, with notable differences in how well models generalize beyond their original training corpus.

Among all datasets, BEADs achieves the highest generalization score (0.582), suggesting that its annotations capture bias-related features that extend beyond its specific domain. Similarly, models trained on Crisis in Ukraine (0.566) and MBBMD (0.568) also exhibit relatively strong transferability, although with some variations across individual test sets. These results indicate that datasets with a broader annotation scope and more diverse content tend to perform better in cross-dataset evaluations.

On the other hand, FIGNEWS (0.127) present significant challenges in generalization. This poor generalization further reinforces its linguistic and structural uniqueness. As seen in prior evaluations, models trained on FIGNEWS struggle to transfer knowledge to other datasets, and vice versa, likely due to its multilingual nature and focus on social media and short-form news.

Overall, these findings highlight the influence of dataset composition, annotation methodology, and linguistic domain in determining how well a model generalizes. While some datasets provide a more flexible foundation for bias detection, others remain highly specific to their original context, limiting their applicability in broader media analysis.

FIGURE 4.7: Average cross-dataset performance by training source.



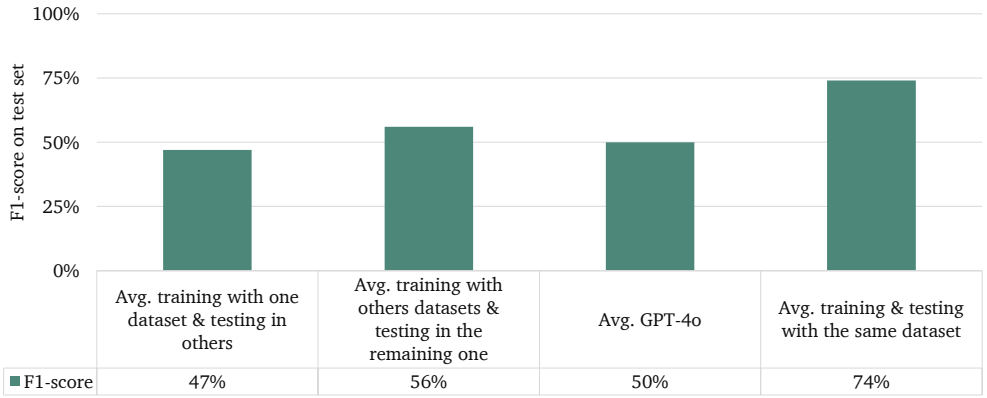
A comparative analysis of different training strategies, summarized in Figure 4.8, highlights the impact of dataset diversity on generalization performance. Models trained on a single dataset (and tested in a different dataset) achieve an average cross-dataset F1-score of 0.454, indicating a strong dependence on the linguistic, stylistic, and ideological characteristics of the training data. This limited generalization confirms that bias detection models tend to learn dataset-specific patterns, making them less effective when applied to news articles from different contexts.

In contrast, models trained on multiple datasets exhibit improved generalization, reaching an average F1-score of 0.562. The inclusion of diverse annotation schemes and content sources appears to help models learn more transferable features of bias, reducing overfitting to a single journalistic or regional style. While this strategy enhances performance, it does not completely resolve the challenges of cross-dataset variability, as seen in the inconsistent results across different test sets.

A notable comparison is the performance of GPT-4o in a 1-shot learning setup, which achieves an F1-score of 0.502. This score places it between models trained on a single dataset and those trained on multiple datasets, suggesting that LLMs capture bias-related cues even without explicit supervision. However, despite its ability to identify bias signals in 1-shot settings, GPT-4o does not outperform dedicated bias detection models trained on multiple datasets, indicating that fine-tuning on curated datasets remains crucial for optimal performance.

Finally, models trained and tested on the same dataset continue to yield the highest F1-scores, averaging 0.744. This result underscores a key limitation of current bias detection approaches: while models perform well within their training domain, they struggle when exposed to different linguistic, ideological, and journalistic contexts. The substantial gap be-

FIGURE 4.8: Aggregated cross-dataset evaluation results.



tween same-dataset and cross-dataset performance highlights the pressing need for strategies that mitigate dataset overfitting and improve generalization to unseen news sources.

Overall, the results highlight the challenges of cross-dataset generalization in media bias detection. Certain datasets, particularly BEADs and Crisis in Ukraine, exhibit stronger transferability, while others, such as FIGNEWS, introduce complexities that hinder model performance across domains. The inclusion of multiple datasets during training generally improves generalization, but dataset-specific annotation schemes and linguistic variations continue to impose limitations. These findings set the stage for the next section, where we will analyze the broader implications of these results and explore strategies to enhance model robustness in media bias detection.

4.1.3 *Insights gathered and next steps*

The results of the cross-dataset evaluation underscore the inherent difficulty of developing media bias detection models that generalize effectively across different linguistic, sociocultural, and temporal contexts. A recurring pattern in all experiments is that models achieve high performance when tested on the same dataset they were trained on but struggle significantly when evaluated on unseen datasets. This is not an indication of deficiencies in the datasets themselves but rather a reflection of the fact that each dataset encapsulates a particular manifestation of media bias, shaped by its unique linguistic, ideological, and temporal characteristics.

The inability of models to transfer knowledge across datasets suggests that media bias is not a universal phenomenon with a fixed representation but rather a context-dependent construct that varies based on the conventions, discourse structures, and framing strategies of different journalistic ecosystems. While some datasets, such as BEADs and Crisis in Ukraine, demonstrate moderate transferability, others (especially FIGNEWS) exhibit highly domain-specific patterns that severely limit cross-dataset applicability. This indicates that models trained on a single dataset may be learning stylistic and thematic features tied to specific media sources rather than generalizable indicators of bias.

To examine this hypothesis further, we conducted a comprehensive set of experiments aimed at comparing how each model internally represents the same input texts. Using the five fine-tuned models previously introduced (models A through E), we generated sentence-level embeddings for a shared evaluation set and applied dimensionality reduction and statistical analysis techniques to investigate the structure, coherence, and divergence of the embedding spaces induced by each model.

We began by projecting the high-dimensional embeddings into two dimensions using Principal Component Analysis (PCA) [119] and t-distributed Stochastic Neighbor Embedding (t-SNE) [86]. As shown in Figures 4.9 and 4.10, the embeddings from each model form visibly distinct clusters, despite having been computed from the same set of sentences. This separation reveals that each model induces a different latent structure over the input space, suggesting that generalization failures are rooted not only in the models' classification behavior but also in their representational assumptions.

[119] Musil, 2019, "Examining structure of word embeddings with PCA"

[86] Kang et al., 2021, "Conditional t-SNE: more informative t-SNE embeddings"

FIGURE 4.9: PCA 2D projection of sentence embeddings by model.

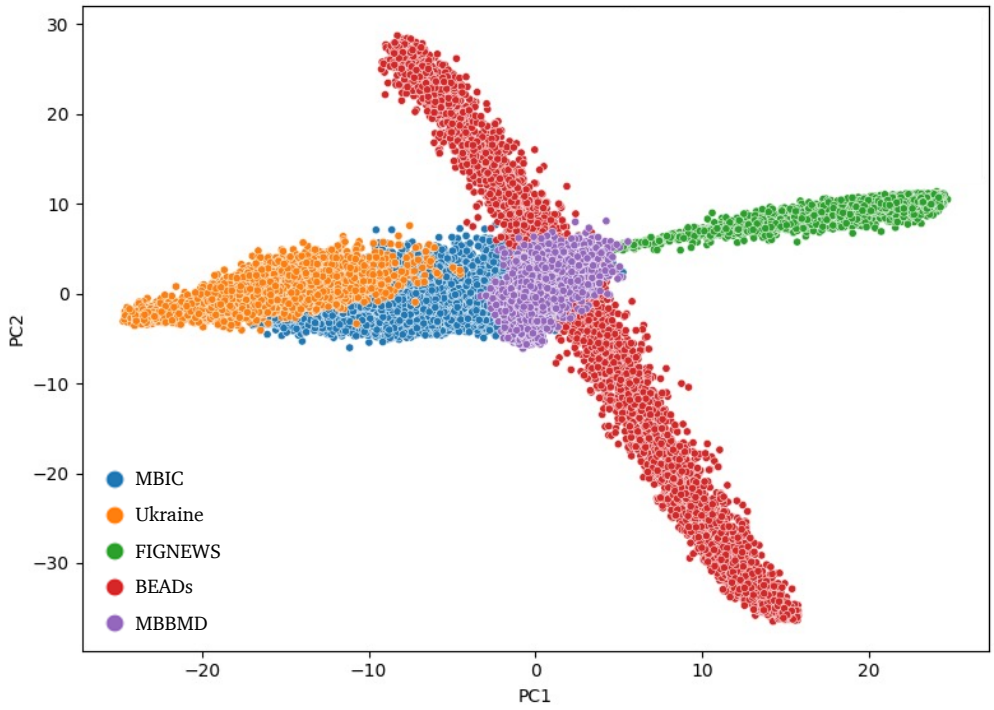
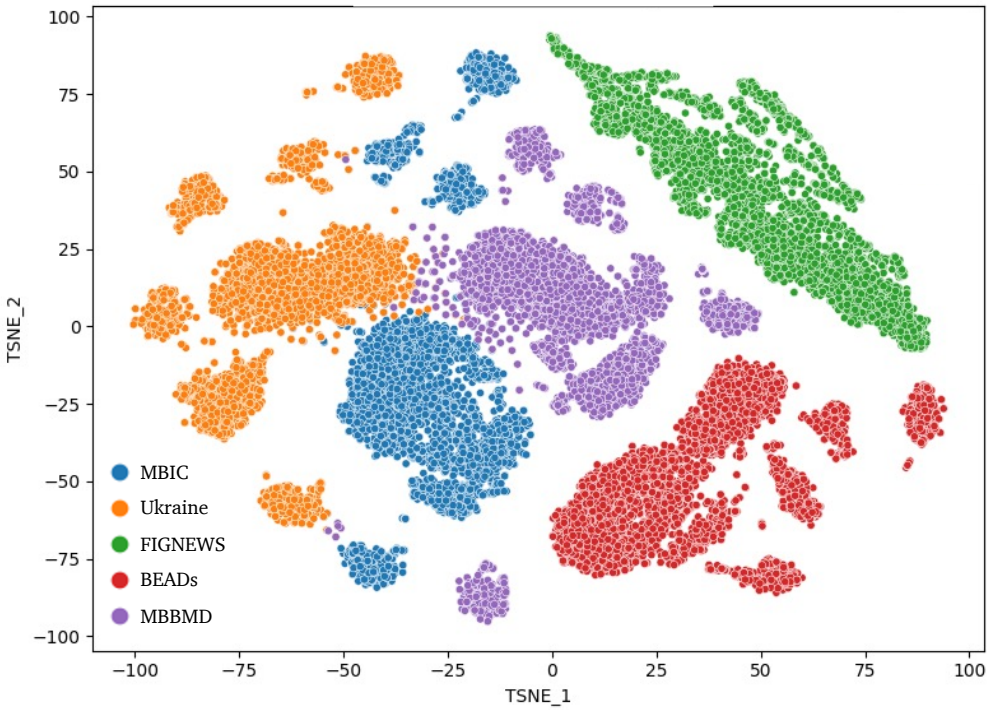


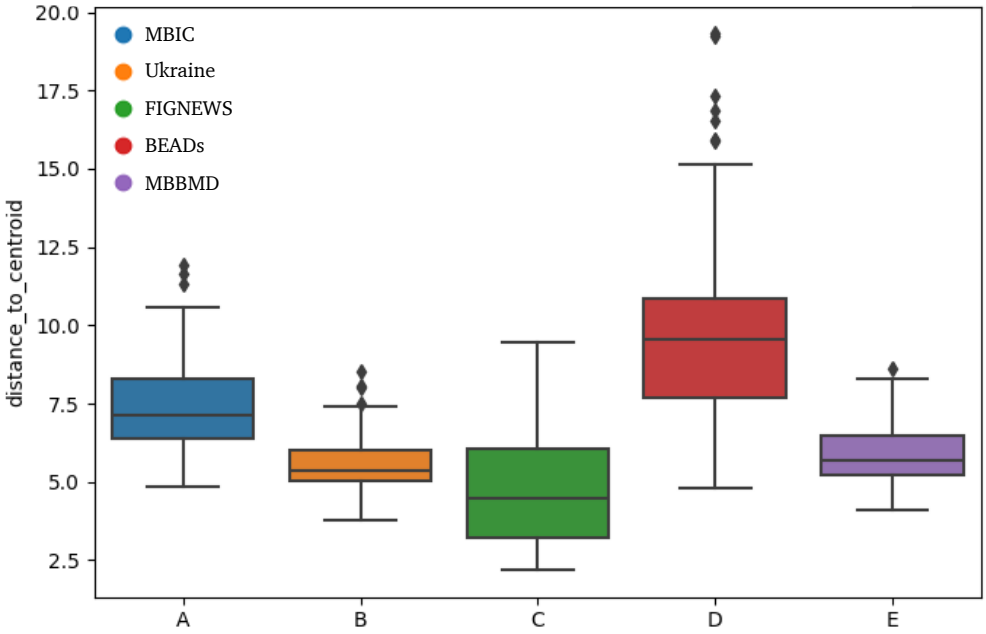
FIGURE 4.10: t-SNE 2D projection of sentence representations by model.



We then analyzed the internal coherence of each model by measuring the dispersion of its embeddings. Specifically, we computed the Euclidean distance of each sentence embedding to the model's centroid and compared the distributions across models. Statistical testing using one-way ANOVA ($F = 221.047$, $p < 0.001$) and Kruskal–Wallis ($H = 514.042$, $p < 0.001$) confirmed that these distributions differ significantly, indicating that some models produce more compact or more diffuse representations than others. Compactness may be indicative of consistent encoding of linguistic features, while higher dispersion could reflect either richer

representational diversity or increased model uncertainty. Figure 4.11 summarizes these findings with boxplots for each model.

FIGURE 4.11: Distribution of sentence-level distances to the centroid of each model's embedding space.



To assess global structural divergence across models, we applied PERMANOVA (Permutational Multivariate Analysis of Variance) [9] over the cosine distance matrices between all sentence embeddings. The test yielded a pseudo-F statistic of 862.30 ($p = 0.001$), demonstrating that inter-model differences are statistically significant and far exceed intra-model variability. Complementing this, the Silhouette score [156] computed over model assignments yielded a value of 0.5164, indicating well-separated clustering structure in the embedding space. This result aligns with the qualitative observations from dimensionality reduction and confirms that models group their representations distinctly, despite having processed the same content.

[9] Anderson, 2014, "Permutational multivariate analysis of variance (PERMANOVA)"

[156] Rousseeuw, 1987, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis"

We further investigated pairwise relationships between models using the Mantel test [44], which compares the structure of distance matrices between two models. Strong correlations were found between several models, such as MBIC vs Ukraine ($r = 0.5526$, $p = 0.001$) and Ukraine vs MBBMD ($r = 0.5473$, $p = 0.001$), suggesting a shared representational logic. In contrast, models involving FIGNEWS (e.g., MBIC vs FIGNEWS, $r = 0.079$, $p = 0.008$) showed weak correlation, indicating that model FIGNEWS encodes semantic structure in a markedly different way. These representation-level experiments provide compelling evidence that generalisation challenges are deeply rooted in the semantic space each model constructs. Two models may yield similar classification scores yet differ profoundly in how they organise semantic meaning. As such, traditional evaluation metrics fail to fully capture model behaviour, and must be complemented with representational diagnostics to understand the nature of generalisation failures.

[44] Diniz-Filho et al., 2013, "Mantel test in population genetics"

It is important to be precise about the scope of these representational diagnostics. The five DistilBERT models analysed here (one per dataset, including MBBMD) were all trained as flat binary biased/unbiased classifiers on the majority-vote label, because no other formulation is supported across the five corpora. The semantic spaces they induce are therefore the spaces that each dataset's binary labels happen to organise, and any apparent verdict over MBBMD in this analysis is a verdict over MBBMD-as-a-flat-binary-corpus, not over MBBMD as the hierarchical perspectivist resource it was designed to be. This caveat is methodologically central: it explains why MBBMD does not stand out in this section, and why the rest of the chapter is devoted to recovering precisely the structure that the cross-dataset protocol forces us to discard.

These findings highlight a major challenge: constructing datasets that comprehensively represent media bias across diverse contexts is both technically complex and resource-intensive. No flat, document-level, single-label dataset (including MBBMD when used in that reduced form) can fully capture the spectrum of ways in which bias is expressed across different media landscapes. Moreover, dataset-driven supervised learning approaches remain constrained by the perspectives and assumptions embedded in the annotation process, making them susceptible to underrepresentation of certain viewpoints and biases in their own right. The response that this thesis adopts (hierarchical annotation and preserved disagreement, both implemented in MBBMD and exploited from the next section onward) does not claim to escape this constraint, but to move it from the data layer to the modelling layer where it can be inspected, attributed, and partially mitigated.

These findings motivate the complementary analysis presented in the next section, which shifts focus from cross-dataset evaluation under a flat binary protocol to level-by-level modelling on MBBMD with its hierarchical and perspectivist annotations restored.

4.2 LEVEL-BY-LEVEL ANALYSIS: SENTENCE, DOCUMENT, AND CATEGORY

The cross-dataset experiment of Section 4.1 establishes that flat, document-level bias classifiers do not transfer. But it does not tell us *why* they fail, or which signals would be needed to build a system that does. There is also a more specific reason to pause before drawing conclusions about MBBMD from that experiment alone: in Section 4.1, MBBMD enters the benchmark in its flattest possible form, namely a single binary majority-vote label per document, because that is the only representation the other four corpora support. That representation discards everything that makes MBBMD distinctive (its sentence-level manifestation annotations, its preserved annotator-level disagreement, and its document-level bias categories), and so any claim that "MBBMD does not luce in cross-dataset evaluation" is, more precisely, a claim about a heavily reduced view of MBBMD against which no dataset including MBBMD itself, in its full hierarchical perspectivist form, has yet been evaluated. The point of this section is to undo that reduction.

This section therefore takes a complementary empirical approach: rather than evaluating a single model across many datasets under a flat document-level binary protocol, it evaluates many modelling paradigms against the three analytical layers that MBBMD was explicitly designed to expose, and that are operationally inaccessible in any of the other four corpora. These layers (sentence-level manifestations, document-level perception with preserved annotator disagreement, and document-level bias categories) instantiate, layer by layer, the hierarchical and perspectivist conceptualisation of media bias developed in Chapter 3. Treating media bias level by level is not, in this thesis, a methodological variant alongside flat classification; it is the consequence of taking that conceptualisation seriously. Section 4.1 measured how badly flat binary classifiers transfer; this section measures what becomes available once the hierarchical and perspectivist annotations of MBBMD are restored. The findings directly motivate the architectural choices of the hierarchical system presented in Section 4.3.

Concretely, we evaluate how different models operate across three analytically distinct but conceptually linked tasks derived from the taxonomy introduced in Chapter 3: (i) the identification of explicit bias manifestations at the sentence level (six manifestations from the taxonomy of Section 3.1, instantiated in MBBMD), (ii) the binary classification of documents as biased or unbiased (now using both majority-vote labels and the preserved annotator-level disagreement, the latter formulated as a regression on the annotator agreement percentage following the LeWiDi paradigm), and (iii) the categorisation of document-level bias into higher-order editorial types (the five document-level categories defined in Chapter 3). For sentence-level manifestation characterisation, we compare heuristic methods (based on linguistic rules and media theory) with learning-based approaches, including encoder-only models (e.g., BERT variants) in multitask setups [134], and decoder-only models (e.g., GPT-4o) with prompt engineering for fine-grained detection. The document-level bias detection task uses

[134] Plaza-Del-Arco et al., 2021, "A multi-task learning approach to hate speech detection leveraging sentiment analysis"

encoder-only transformers and ideology-aware models, integrating annotator perspective metadata (left, centre, right) to assess the impact of ideological alignment. For categorisation into bias types (e.g., coverage or gatekeeping), we also use decoder-only models augmented with Retrieval-Augmented Generation (RAG) [101], enabling dynamic access to external knowledge during inference and enhancing classification of context-dependent biases. All experiments in this section use the benchmark datasets introduced in Subsection 2.2.3.

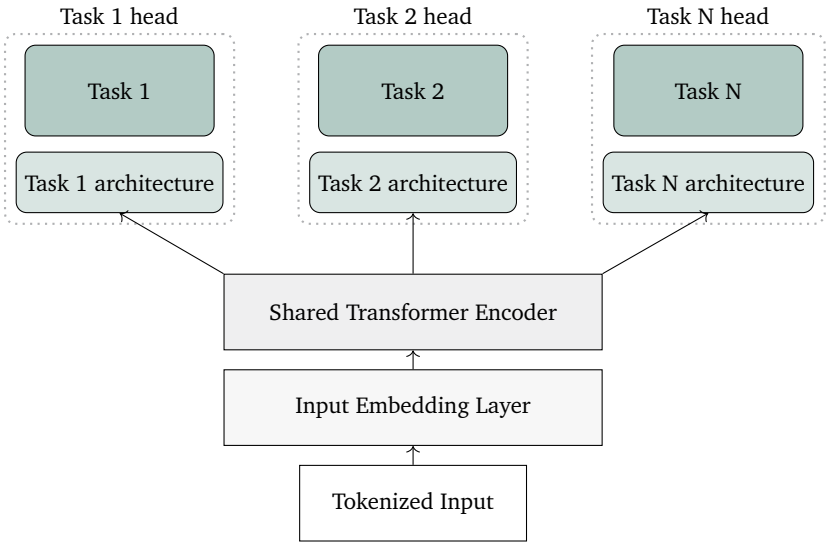
[101] Lewis et al., 2020, "Retrieval-augmented generation for knowledge-intensive nlp tasks"

4.2.1 *Media bias manifestations characterization at sentence level*

Accurately characterizing specific manifestations of media bias at the sentence or segment level represents a complex and crucial analytical task within the broader domain of media bias detection. Given that media bias is often subtle, context-dependent, and linguistically intricate, a granular and targeted approach is required to effectively identify, analyze, and understand these biases. To address this complexity, this section presents a detailed comparative evaluation of three distinct methodological frameworks: heuristic-linguistic methods, an encoder-only multitask learning approach leveraging DistilBERT, and a contemporary decoder-only large language such as GPT-4o. Through rigorous experimentation using the MBBMD dataset, we aim to systematically assess these models' effectiveness in identifying nuanced manifestations of bias at a fine-grained textual level.

The first approach evaluated in our comparative analysis consists of heuristic-linguistic methods. Rooted in classical computational linguistics and discourse analysis, these heuristic models employ rule-based systems, lexical dictionaries, and manually annotated linguistic cues derived from expert insights in media and communication studies. Typically, these heuristic approaches combine lexicon-based methods, handcrafted regular expressions, and linguistic pattern detection to systematically capture instances of biased language. It serves to highlight linguistic markers that are empirically correlated with biased language, enabling detailed analysis of linguistic structures that frequently encode subtle biases such as sensationalism, omission of attribution, or subjectivity. The complete lexicons, word lists, and flagging rules used by each heuristic are documented in Appendix C.

FIGURE 4.12: Multi-task architecture with a shared encoder and task-specific heads.



Building upon the strengths and weaknesses identified in heuristic approaches, our second methodological framework adopts a multitask architecture based on DistilBERT, an encoder-only transformer model. DistilBERT, a distilled version of BERT, retains much of BERT's semantic understanding capabilities while offering improved efficiency and reduced computational overhead. This makes it a suitable backbone for scalable and resource-aware bias detection systems. As illustrated in Figure 4.12, the model uses a shared transformer encoder followed by task-specific heads, enabling simultaneous training on multiple bias detection objectives (every kind of media bias manifestation) while leveraging shared linguistic representations.

The decision to adopt a multitask formulation was not made arbitrarily. We conducted a series of preliminary experiments comparing three modeling paradigms: (i) independent single-task models, one for each type of bias manifestation; (ii) a multi-label classification model that outputs all bias categories simultaneously without task-specific decoupling; and (iii) a multitask model with distinct output heads optimized separately for each bias manifestation. While single-task models offered strong performance in isolation, they failed to generalize well across categories and required redundant computations. The multi-label model, though computationally lean, exhibited limited ability to disentangle subtle differences among related bias types.

In contrast, the multitask setup provided the best overall performance, achieving higher precision and recall across most categories. We attribute this to its ability to exploit inductive transfer between tasks: shared representations in the encoder layers capture common semantic and rhetorical features among related manifestations (e.g., between mind reading and opinion-as-fact), while task-specific heads allow for specialized fine-tuning. This structure enables the model to benefit from shared learning without sacrificing task-specific granularity.

Furthermore, the multitask architecture benefits from improved sample efficiency: less frequent bias types can leverage information from more prevalent categories during training. Additionally, DistilBERT's lightweight nature ensures that these performance gains do not come at the cost of excessive resource consumption, making the model a compelling option for real-world applications requiring fast inference and moderate hardware constraints.

The model was trained using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16, and a weight decay of 0.01. Training was performed over 5 epochs with early stopping based on the validation macro F1-score, using a patience of 2 epochs. A linear learning rate scheduler with warmup (10% of training steps) was employed. The maximum sequence length was set to 256 tokens, and dropout was applied at a rate of 0.1 in both the encoder and task-specific classification heads. Each task head consisted of a single feedforward layer with a sigmoid activation function, producing a binary output (presence or absence of that manifestation) for each specific manifestation of media bias.

Complementing these methods, we incorporate state-of-the-art decoder-only architectures, specifically GPT-4o, to further explore the potential of generative language models in fine-grained bias manifestation detection. Decoder-only models, characterized by their generative capabilities, provide substantial flexibility through prompt engineering [202], in-context learning [199], and dynamic adaptation to nuanced linguistic contexts. GPT-4o, renowned for its exceptional generative and reasoning capabilities, enables sophisticated interactions with textual data by dynamically generating contextually appropriate analytical outputs.

[202] White et al., 2023, "A prompt pattern catalog to enhance prompt engineering with chatgpt"

[199] Wang et al., 2023, "Learning to retrieve in-context examples for large language models"

By evaluating these diverse methodological approaches (heuristic, multitask DistilBERT, and GPT-4o) we aim to deliver comprehensive empirical insights into their respective strengths and weaknesses in capturing and characterising sentence-level media bias manifestations. This comparative analysis not only enriches our understanding of methodological suitability but also provides crucial insights to guide the development of robust, scalable, and effective bias detection systems in subsequent analytical stages.

The remainder of this subsection applies the same three-method comparison (heuristic-linguistic, multitask DistilBERT, and GPT-4o) to each of the six sentence-level bias manifestations operationalised in MBBMD: omission of source attribution, sensationalism / emotionalism, mind reading, biased word choice / labelling, subjective qualifying adjectives, and presentation of opinions as facts. We begin with omission of source attribution as a paradigmatic case in which the rationale for the comparison and the structure of the analysis are introduced in full detail; the five subsequent manifestations are presented under the same template (linguistic and theoretical motivation, heuristic algorithm, three-way results table) so that they can be read individually but also compared on equal footing. Each manifestation reports the same four metrics (macro precision, macro recall, macro F1, and per-class F1 for the False / True classes), enabling a uniform reading across all six tables; the cross-manifestation synthesis is deferred to the bridging paragraph at the end of this subsection, immediately before the document-level analysis of Subsection 4.2.2.

▸ Characterizing omission of source attribution bias

Omission of source attribution bias arises when a claim, report, or evaluative assertion is made without explicit attribution to a responsible agent. This absence can mislead readers by creating the impression that the information is objective, widely accepted, or comes from an authoritative source, even when no such source is identified. The mechanism is essentially epistemic: by removing the agent who is supposed to back a claim, the sentence shifts from a reported assertion (someone said x) to an apparently authorless one (x is the case), which the reader is then likelier to accept at face value. This phenomenon is particularly salient in journalism, politics, and institutional communication, where factual accountability is critical and where unattributed assertions can act as conduits for rumour, speculation, or partisan framing dressed up as consensus.

We treat omission as a paradigmatic case for the rest of this subsection because it isolates, more cleanly than any other manifestation, the difference between what is said and what is omitted. Sensationalism, mind reading, or biased word choice can all be detected by what appears in the sentence; omission, by definition, has to be detected by what is missing, and so the heuristic has to be designed around the absence of a structural element rather than the presence of a lexical one. This shift sets the methodological tone for the five manifestations that follow.

The linguistic strategy implemented in this work is therefore a hybrid one that combines lexico-syntactic heuristics with shallow semantic patterns. It is multilingual and supports both English and Spanish, with language-specific lexicons and parsers selected at runtime by an automatic language identifier.

From a lexical perspective, we define two key lists for each language: (i) a set of *reporting verbs* commonly used to introduce claims [55, 24] (e.g., "*informar*", "*afirmar*", "*reportar*" in Spanish; "*claim*", "*report*", "*state*" in English), and (ii) a set of *vague source expressions*, which simulate sourcing without providing traceability (e.g., expressions referring to "*fuentes no identificadas*", "*la comunidad internacional*", "*expertos*"). That is, unidentified sources not attributed to specific individuals, institutions, or documents (e.g., "*unidentified sources*", "*the international community*", "*experts*"). The first list captures the surface form of an assertion of attribution; the second captures the surface form of an evasion of it. Used together, they let the detector distinguish a properly sourced claim from one that simulates sourcing without providing traceability, which is the core mechanism of this manifestation.

[55] Floyd Moore, 2000, "The reporting verbs and bias in the press"

[24] Bott and Solstad, 2021, "Discourse expectations: explaining the implicit causality biases of verbs"

Syntactically, omission is detected when such verbs or vague sources appear without a co-occurring named entity (e.g., organizations, people, geopolitical actors) or proper noun in the sentence. This lack of a clearly attributed agent signals a potential detachment between the information presented and its origin: a sentence such as *"experts agree that the policy will fail"* is structurally similar to a sourced claim, but the position that should be occupied by an identifiable agent is empty, and the heuristic flags it on exactly that absence. Additionally, the system considers impersonal or passive structures (especially those introduced by markers like "*se*" (impersonal/passive), "*al parecer*" ("apparently"), or modal hedges such as "*como si*" ("as if")) as indicators of attribution avoidance, since these constructions allow the writer to assert a claim while leaving the responsible agent grammatically unspecified.

To operationalize this logic, we employ a multilingual spaCy [194] model to tokenize and parse each sentence. The pipeline includes sentence segmentation, lemmatization, PoS tagging, named entity recognition, and dependency analysis. Based on these linguistic features, the sentence is flagged if a report-like structure is present but no clear subject can be identified.

[194] Vasiliev, 2020, *Natural language processing with Python and spaCy: A practical introduction*

The algorithm also includes language detection, enabling dynamic selection of the appropriate lexical resources and syntactic expectations per language.

Two design choices are worth highlighting. First, the decision to flag sentences on the basis of *absence* rather than presence (i.e., the absence of a named attribution co-occurring with a reporting verb or vague-source expression) is deliberate: it reframes the detection problem from "finding biased phrases" to "finding structural gaps in the chain of accountability", which aligns the heuristic with how omission bias is theoretically defined in journalism studies. Second, the hybrid combination of lexical anchors (reporting verbs, vague-source expressions) with shallow syntactic checks (named-entity presence, impersonal markers) keeps the detector

interpretable and auditable: every flagged sentence can be traced back to a specific lexical trigger and a specific missing syntactic element, which facilitates error analysis and supports the broader perspectivist commitment of the thesis to decisions that can be inspected and contested rather than accepted as black-box outputs.

This heuristic approach, summarised in Algorithm 1, offers interpretable, scalable, and language-sensitive bias detection grounded in observable linguistic properties. It is particularly effective at identifying subtler forms of omission, including vague plural references or institutional passive voice common in journalistic and political texts.

Algorithm 1 Heuristic detection of source attribution omission

Precondition: Report verbs lexicon rv, vague sources lexicons vs, impersonal lexicons imp

```
Parse each sentence s into tokens T, named entities E, POS tags
Initialize flags: hasReportVerb, hasVagueSource, hasNamedEntity, hasProperNoun, usesImpersonal := False
for each token t in T do
    if t.lemma ∈ rv and t.pos = VERB then
        hasReportVerb := True
    if t.pos = PROPN then
        hasProperNoun := True
for each entity e in E do
    if e.label ∈ {PERSON, ORG, GPE} then
        hasNamedEntity := True
if any word in vs appears in s then
    hasVagueSource := True
if any sentence in imp appears in s then
    usesImpersonal := True
if (hasReportVerb or hasVagueSource or usesImpersonal) and not hasNamedEntity and not hasProperNoun then
    return True, "Source_Attribution_Omission"
else
    return False
```

Results. Table 4.1 presents a comparative summary of the experimental results obtained using the three evaluated methods. The heuristic-linguistic approach is compared against two advanced models: a multitask DistilBERT architecture and GPT-4o. Each method exhibits distinct performance profiles in the task of detecting omission of source attribution bias.

TABLE 4.1: Comparative performance metrics for source attribution omission bias detection.

Method	Macro Precision	Macro Recall	Macro F1-score	F1-score per class (False / True)
Heuristic-linguistic	5.90%	66.67%	10.84%	2.76% / 18.92%
Multitask DistilBERT	**50.00%**	58.33%	**53.96%**	**79.36%** / 28.57%
GPT 4o	2.73%	**100.00%**	5.31%	74.00% / 5.37%

The heuristic-linguistic method shows very high recall (66.67%) but extremely low precision (5.90%), resulting in a modest macro F1-score of 10.84%. This is consistent with the design of the rule-based approach, which aims to maximize coverage of omission cases by capturing a broad set of lexical and syntactic patterns. However, its low precision reflects the method's difficulty in avoiding false positives, likely due to overgeneralization of vague or impersonal structures.

In contrast, the multitask DistilBERT model achieves the best overall performance, with a balanced macro precision (50.00%) and recall (58.33%), yielding a macro F1-score of 53.96%. Notably, it performs significantly better on the negative class (non-omission), with a macro F1-score of 79.36%, but still shows substantial ability to detect omission cases (28.57%). This suggests that the multitask setup successfully leverages shared learning across related bias categories to distinguish omission with greater contextual accuracy.

GPT 4o exhibits an unusual profile, achieving perfect recall (100.00%) but very low precision (2.73%), leading to a macro F1-score only slightly higher than the heuristic approach (5.31%). This indicates that while GPT 4o captures virtually all positive cases of omission, it does so at the cost of a large number of false positives. Such behavior may stem from its generative nature and overreliance on prompt-based cues, which may not effectively constrain its outputs in binary discrimination tasks without strong calibration.

Overall, these results illustrate the trade-off between interpretability and accuracy: while heuristic methods offer transparent and language-sensitive rules, deep learning models (particularly the multitask DistilBERT) provide superior precision-recall balance for practical deployment.

▸ Characterizing sensationalism / emotionalism bias

Sensationalism or emotionalism bias refers to the strategic use of emotionally charged language, hyperbolic expressions, or dramatic constructions that exaggerate the importance, severity, or emotional resonance of events. The intention behind this bias is to provoke strong affective responses in the reader such as fear, indignation, or admiration; often at the cost of journalistic neutrality or factual restraint.

Linguistically, this type of bias manifests through several rhetorical mechanisms. A key indicator is the use of emotionally charged adjectives and nouns [28, 49], such as "*devastador*" ("devastating"), "*desquiciado*" ("deranged"), or "*horroroso*" ("horrific"), which frame the described events with strong subjective affect. These lexical choices convey judgment or alarm rather than neutral description. Additionally, intensifiers like "*extremadamente*" ("extremely") or "*absolutamente*" ("absolutely") are often used to exaggerate perceived severity [92, 28]. In both Spanish and English, these elements frequently appear in narratives involving violence, political unrest, or societal decline.

This bias also manifests through formulaic expressions, metaphors, and clichés designed to dramatize the narrative [29]. Phrases such as "*camino de la muerte*" ("road to death"), "*ola de euforia*" ("wave of euphoria"), or "*catálogo de barbaridades*" ("catalog of barbarities") serve as rhetorical markers frequently found in partisan discourse or tabloid-style reporting. While these expressions are not necessarily factually incorrect, they convey an emotional perspective rather than an objective description.

[28] Burgers and De Graaf, 2013, "Language intensity as a sensationalistic news feature: The influence of style on sensationalism perceptions and effects"

[49] Enache, 2024, "Pitfalls of writing for the media: appeals to emotion versus exaggerated sensationalism"

[92] Korostenskienė and Bikelienė, 2021, "A comparison of degree intensifiers across English corpora: is it 'flagrant','blatant', or 'sheer'audacity?"

[29] Burgers et al., 2016, "Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony"

Algorithm 2 Heuristic detection of sensationalism / emotionalism bias

Precondition: Sentence s, language-specific emotional lexicons E, intensifiers I, known patterns P

Detect language l of s and load appropriate E_l, I_l, P_l
Parse s into tokens T and syntactic structures
for each sentence s_i in s **do**
    Extract lowercase token list T_i
    if $E_l \cap T_i \neq \emptyset$ **then**
        return `True`, "Sensationalism_Emotionalism"
    if $I_l \cap T_i \neq \emptyset$ **then**
        return `True`, "Sensationalism_Emotionalism"
    if any pattern $p \in P_l$ appears in s_i **then**
        return `True`, "Sensationalism_Emotionalism"
    Compute sentiment polarity p_i of s_i
    if $|p_i| > 0.4$ **then**
        return `True`, "Sensationalism_Emotionalism"
return `False`

Syntactically, sensationalist bias is often conveyed through short, emotionally charged sentences that lack hedging or qualifiers. The omission of mitigating expressions such as "*al parecer*" ("apparently") or "*según fuentes*" ("according to sources") heightens the sense of certainty and amplifies the dramatic tone of the message.

To detect this form of bias, we developed a hybrid linguistic method based on a multilingual pipeline, as summarized in Algorithm 2. The method combines four core strategies: (i) lexical lookup of emotional words and intensifiers; (ii) detection of idiomatic expressions associated with exaggerated discourse; (iii) sentiment polarity scoring using TextBlob [182]; and (iv) syntactic parsing to isolate emotional constructions. Language detection is applied dynamically, allowing the appropriate lexicons and tools to be loaded depending on whether the sentence is in English or Spanish.

[182] Suanpang et al., 2021, "Sentiment analysis with a TextBlob package implications for tourism"

Results. Table 4.2 summarizes the comparative results obtained using the three evaluated methods. The heuristic-linguistic approach is evaluated against two advanced models: Multitask DistilBERT and GPT-4o. Each method exhibits a distinct performance profile in detecting sensationalist or emotionally amplified language.

TABLE 4.2: Comparative performance metrics for sensationalism/emotionalism bias detection.

Method	Macro Precision	Macro Recall	Macro F1-score	F1-score per class (False / True)
Heuristic-linguistic	25.80%	43.50%	31.93%	36.12% / **35.74%**
Multitask DistilBERT	**48.87%**	50.00%	**48.89%**	77.00% / 20.78%
GPT 4o	14.52%	**91.38%**	25.00%	**85.84%** / 25.00%

The heuristic-linguistic method achieves a balanced F1-score across both classes (36.12% for non-sensationalist, 35.74% for sensationalist), with moderate recall (43.50%) and relatively low precision (25.80%). This suggests that while the rule-based system is effective at flagging a wide range of emotionally loaded expressions, it also tends to generate false positives, likely due to the broad coverage of lexicons and patterns used. Its strength lies in its transparency and its cross-lingual robustness, as it captures emotional markers without relying on training data.

Multitask DistilBERT achieves the best overall F1-score (48.89%), with solid macro precision and recall, particularly excelling in identifying non-sensationalist content (77.00% F1). However, its performance on the positive class (sensationalism) is more limited (20.78%), indicating a conservative bias in favor of precision over recall. This pattern may stem from the model's tendency to rely on semantic subtlety and contextual clues, which can make it less sensitive to overt emotional language when it's embedded in figurative or idiomatic forms.

GPT-4o, in contrast, exhibits the highest recall by far (91.38%), with a decent F1-score on the positive class (25.00%) and strong performance on the negative class (85.84%). This suggests that the model is highly sensitive to emotional cues, likely due to its generative architecture and exposure to vast stylistic variation. However, its precision remains low (14.52%), reflecting a tendency to over-detect sensationalism, possibly mistaking stylistic emphasis or metaphorical phrasing for biased exaggeration.

The results reveal trade-offs between coverage and specificity. The heuristic method provides broad, interpretable detection; DistilBERT offers more balanced but cautious predictions; and GPT 4o maximizes recall at the expense of precision.

▸ CHARACTERIZING MIND READING BIAS

Mind reading bias refers to the tendency to present assumptions about the internal states -thoughts, intentions, emotions, or motivations- of individuals or groups as if they were empirically verified facts. This type of bias misleads readers by projecting speculation or interpretation as objective mental access, thus distorting the boundary between observation and inference.

Linguistically, this bias is often realized through the use of cognitive or psychological verbs [99, 60]. In English, common examples include "believe", "think", "feel", "fear", "assume", or "suppose." Their Spanish counterparts include "creer", "pensar", "temer", "suponer", and "sentir." When such verbs are used without explicit evidential attribution, they introduce unverified mental state assertions. For instance, "the government fears a collapse" or "parece que todo está perdido" imply privileged access to internal cognition.

[99] Lee and Hamilton, 2022, "Anchoring in the past, tweeting from the present: Cognitive bias in journalists' word choices"

[60] Geschke et al., 2010, "Effects of linguistic abstractness in the mass media"

Mind reading bias also manifests through modal and speculative constructions. These include expressions such as "*parece que*" ("it seems that"), "*según algunos*" ("according to

some"), or "*todo apunta a que*" ("everything points to"). While these phrases are not inherently problematic, they introduce bias when used without appropriate framing, attribution, or epistemic hedging. Dependency-based analyses reveal that biased constructions frequently combine third-person referents (e.g., "the public", "critics", "the party") with unverifiable mental-state or intention verbs (e.g., "believe", "want", "feel"), thereby projecting internal states without evidential grounding and reinforcing epistemic overreach.

To heuristically detect this type of bias, we developed a multilingual method that combines cognitive verb matching, proximity-based checks for source attribution, and detection of idiomatic speculative patterns. The process begins by identifying verbs associated with mental states or cognition using dedicated lexicons (e.g., "believe", "assume", "suspect"). Once such a verb is found, the system examines a ±5-token window around it to check for the presence of named entities or proper nouns that could serve as explicit sources. If no such attribution is found and the sentence asserts an unverifiable internal state, the system flags it as potentially biased.

In addition, the method detects fixed expressions frequently used to speculate or generalize without grounding, such as "*el mayor miedo*" ("the greatest fear") or "*parece que*" ("it seems that"). These idiomatic markers often indicate projection of thoughts or emotions onto third parties. The pipeline dynamically adjusts lexicons and syntactic parsing tools based on the detected language (Spanish or English), ensuring accurate adaptation across multilingual input.

Algorithm 3 Heuristic detection of mind reading bias

Precondition: Sentence s, language-specific cognitive verbs C, named entity tags, speculative patterns P

Detect language l of s and load appropriate C_l, P_l
Parse s into tokens T and named entities E
for each token t_i in T **do**
    if t_i.`lemma` $\in C_l$ and t_i.`pos` = VERB **then**
        Extract window $W = T[i-5 : i+5]$
        if no named entity or proper noun in W **then**
            return `True`, "Mind_Reading"
if any pattern $p \in P_l$ appears in s **then**
    return `True`, "Mind_Reading"
return `False`, "Mind_Reading"

The resulting approach, summarized in Algorithm 3, captures nuanced linguistic cues of mind reading with a transparent rule set that facilitates human inspection. It is particularly useful in journalistic and political discourse analysis, where epistemic attribution is essential for credibility and factual rigor.

Results. Table 4.3 summarizes the comparative results for mind reading bias detection using the three evaluated methods. The heuristic-linguistic system is compared with two advanced baselines: Multitask DistilBERT and GPT-4o. The methods show markedly different behavior in detecting speculative constructions and unverified mental state attributions.

TABLE 4.3: Comparative performance metrics for mind reading bias detection

Method	Macro Precision	Macro Recall	Macro F1-score	F1-score per class (False / True)
Heuristic-linguistic	11.70%	**62.50%**	19.24%	20.50% / **50.00%**
Multitask DistilBERT	**44.44%**	50.00%	**47.06%**	**74.87%** / 19.25%
GPT 4o	4.33%	**62.50%**	8.11%	87.95% / 8.17%

The heuristic-linguistic method achieves the highest recall (62.50%) and strong F1-score on the positive class (50.00%), indicating its ability to detect a broad range of speculative expressions and epistemic overreach. However, its low macro precision (11.70%) leads to a modest overall macro F1-score of 19.24%. This pattern suggests that while the system is effective at flagging potentially biased sentences, it tends to overgenerate false positives,

likely due to the broad inclusion of ambiguous modal verbs and idiomatic expressions in the handcrafted rules.

Multitask DistilBERT achieves the best overall macro F1-score (47.06%) with a balanced trade-off between precision (44.44%) and recall (50.00%). It excels particularly on the negative class (74.87% F1), indicating strong discriminative ability when mental state attributions are grounded or absent. However, its lower performance on the true class (19.25%) suggests it may be conservative when identifying speculative constructions, possibly due to the limited frequency of such patterns in training data or difficulty in learning idiomatic and epistemic cues without explicit supervision.

GPT 4o also achieves high recall (62.50%) but suffers from extremely low precision (4.33%), yielding an overall macro F1-score of just 8.11%. Its performance is heavily skewed toward the negative class (87.95% F1), with limited success in reliably identifying mind reading bias (8.17% F1). This indicates a tendency to broadly flag content as biased, likely due to its generative nature and the lack of task-specific calibration. GPT 4o may be sensitive to stylistic and modal indicators, but imprecise in contextualizing their epistemic grounding.

In summary, the heuristic approach excels in recall and interpretability but overflags speculative language; DistilBERT provides a balanced and reliable classifier with cautious bias detection; and GPT 4o captures speculative cues widely, but at the expense of precision. These trade-offs highlight the importance of aligning model behavior with the operational needs of the analysis task, whether exhaustive screening or conservative flagging is prioritized.

▸ Characterizing word choice / labeling bias

Word choice or labeling bias occurs when the language used to describe individuals, groups, or events embeds implicit value judgments, thereby framing perceptions through emotionally or ideologically loaded terms. Rather than presenting neutral descriptions, such bias subtly endorses specific interpretations by foregrounding subjective characterizations.

Linguistically, this bias frequently manifests through emotionally charged adjectives directly modifying named entities. Examples include phrases like "corrupt politician", "radical activists", or their Spanish equivalents such as "político corrupto" or "manifestantes radicales". These constructions frame the referent within a value-laden narrative, which may not be supported by verifiable evidence.

Labeling bias also arises from noun choices that inherently carry pejorative or glorifying connotations. Describing citizens as a "mob" versus a "protest group", or fighters as "terrorists" versus "militants", demonstrates how subtle lexical decisions reshape the narrative frame. Such noun-level framing is particularly consequential in media, political discourse, and conflict reporting.

From a syntactic perspective, labeling bias is often realized through adjective-noun (amod - adjectival modifier) dependencies or noun-noun compounds. These structures are commonly identified using dependency parsing, where adjectives with strong emotional polarity precede nouns referring to entities (persons, organizations, nations). The absence of hedging expressions, quotation marks, or attributive frames further amplifies the bias by presenting the evaluation as factual [83].

[83] Jensen et al., 2011, "Including limitations in news coverage of cancer research: Effects of news hedging on fatalism, medical skepticism, patient trust, and backlash"

To detect this type of bias heuristically, we developed a multilingual method that combines dependency-based syntactic analysis with sentiment polarity estimation, as outlined in Algorithm 4. The core idea is to identify evaluative language attached to named entities or individuals. First, each sentence is syntactically parsed to detect adjectival modifiers of nouns referring to people or organizations. If such adjectives display strong sentiment polarity (either positive or negative) the construction is flagged as potentially biased.

Additionally, emotionally charged nouns (e.g., "*tirano*" ("tyrant"), "*genio*" ("genius"), or "*criminal*") used as subjects or predicate complements are also targeted, as they often signal biased labeling without factual attribution. Sentiment polarity scores are computed using pysentimiento [131] for both English and Spanish, ensuring cultural and linguistic appropriateness.

[131] Pérez et al., 2021, "pysentimiento: A python toolkit for opinion mining and social nlp tasks"

Algorithm 4 Heuristic detection of word choice / labeling bias

Precondition: Sentence s, language-specific NLP parser nlp, sentiment analyzer SA, polarity threshold τ

```
Detect language l of s and select corresponding nlp_l and SA_l
Parse s into tokens T and dependencies
for each token t in T do
    if t.pos = ADJ then
        h ← syntactic head of t
        if h.pos = NOUN and h refers to PERSON/ORG/PROPN then
            p ← SA_l(t.text)
            if |p| ≥ τ then
                return True, "Word_choice/labelling"
    else if t.pos = NOUN and t not part of named entity then
        p ← SA_l(t.text)
        if |p| ≥ τ then
            return True, "Word_choice/labelling"
return False, "Word_choice/labelling"
```

Automatic language detection selects the appropriate spaCy parser and sentiment model for each input sentence. This approach is designed with interpretability in mind and is particularly useful for applications in media literacy, journalistic monitoring, and comparative discourse analysis across languages.

Results. Table 4.4 presents the comparative performance metrics obtained from applying the three evaluated methods to the task of labeling bias detection. The heuristic-linguistic approach is assessed alongside two state-of-the-art models: Multitask DistilBERT and GPT-4o. The results highlight substantial differences in how each method handles evaluative framing, particularly in detecting emotionally charged modifiers and labeling expressions.

TABLE 4.4: Comparative performance metrics for word choice/labeling bias detection

Method	Macro Precision	Macro Recall	Macro F1-score	F1-score per class (False / True)
Heuristic-linguistic	9.40%	36.44%	19.96%	14.34% / **60.00%**
Multitask DistilBERT	**40.00%**	45.45%	**42.62%**	72.00% / 13.24%
GPT 4o	7.87%	**100.00%**	14.5%	**85.76%** / 14.50%

The heuristic-linguistic method yields moderate recall (36.44%) but low precision (9.40%), resulting in a macro F1-score of 19.96%. Despite its modest overall performance, it achieves the highest F1-score for the positive class (60.00%), indicating a strong capacity to identify explicit evaluative expressions such as polarized adjectives or charged nouns. However, the lower score on the negative class (14.34%) reflects its tendency to over-flag content as biased, likely due to limited contextual disambiguation and the reliance on lexicon-based thresholds.

Multitask DistilBERT shows the strongest overall performance with a macro F1-score of 42.62%. It balances precision (40.00%) and recall (45.45%), achieving a high F1 on the negative class (72.00%) while maintaining reasonable sensitivity to positive cases (13.24%). This suggests that the model is particularly effective at confirming the absence of labeling bias, though somewhat conservative in recognizing it. The multitask architecture likely benefits from shared learning across related bias categories, enabling better generalization of evaluative framing patterns within a larger discourse context.

GPT-4o, in contrast, achieves perfect recall (100.00%) but extremely low precision (7.87%), yielding a macro F1-score of only 14.5%. Its high sensitivity results in flagging virtually all candidate sentences as biased, including many false positives. Although it maintains solid performance on the negative class (85.76%), its positive class F1 (14.50%) is modest, indicating challenges in grounding evaluative markers contextually. This behavior likely stems from its generative nature, which prioritizes comprehensive surface-level coverage over cautious discrimination.

In summary, the heuristic system is useful for identifying direct and explicit evaluative framings, especially in multilingual settings; the multitask model provides balanced and deployable performance for real-world annotation tasks; and GPT 4o excels in broad sensitivity but lacks precision for fine-grained detection.

▸ Characterizing use of subjective qualifying adjectives bias

Use of subjective qualifying adjectives bias arises when evaluative, or opinion-laden adjectives are employed to characterize individuals, groups, actions, or events. This subtle yet pervasive bias influences readers by inserting subjective assessments into otherwise factual descriptions, thereby blurring the boundary between objective reporting and personal interpretation.

Linguistically, this bias manifests through the direct modification of nouns by adjectives that convey a judgment rather than an objective description. For example, expressions such as "the incompetent minister", "the magnificent building", or "the disastrous decision" embed evaluative connotations that shape the reader's perception. Similar constructions in Spanish include "el incompetente ministro" or "la desastrosa decisión."

Syntactically, these constructions are identified via adjective-noun dependency patterns, where an adjective modifies a noun that may be a named entity or key topic. In Spanish, the adjective may appear before or after the noun depending on the rhetorical emphasis, while in English, pre-nominal adjective positioning is standard. Regardless of language, these structures serve the same rhetorical function of injecting subjective bias.

Detection of this bias also relies on recognizing adjectives with extreme sentiment polarity, such as "atrocious", "brilliant", or "horrible", which often appear without evidential support. The presence of intensifiers (e.g., "utterly disastrous", "truly magnificent") further heightens the evaluative force of these constructions.

Algorithm 5 Heuristic detection of subjective qualifying adjectives bias

Precondition: Sentence s, language-specific parser nlp, subjective lexicon L

Detect language l of s and load parser nlp_l and lexicon L_l
Parse s into tokens T, syntactic heads, and named entities
for each token t in T **do**
    if t.pos = ADJ and t.lemma $\in L_l$ **then**
        $h \leftarrow$ syntactic head of t
        Check if h.pos = NOUN
        Check window ±3 tokens around t for named entity or proper noun
        if modifies NOUN **or** near named entity **then**
            return True, "Subjective_qualifying_adjectives"
return False, "Subjective_qualifying_adjectives"

To heuristically detect this phenomenon, we developed a method that combines syntactic parsing with lexical lookup. The system first identifies adjectives modifying nouns or occurring near named entities. These adjectives are then compared against curated lexicons of subjective qualifiers in both English and Spanish. The detection is triggered when such adjectives appear in bias-prone syntactic contexts and meet predefined lexical criteria.

The approach, described in Algorithm 5, uses automatic language detection to dynamically select the appropriate parser and lexicon, ensuring compatibility across languages. This enables consistent detection in various text genres, including journalistic, political, and academic writing; contexts where subtle evaluative framing often escapes purely statistical models.

By grounding detection in syntactic structures and lexically defined subjectivity markers, the method offers an interpretable and scalable solution for identifying subjective qualifying adjectives bias across multilingual corpora. These adjectives (e.g., "*brilliant*", "*disastrous*", "*remarkable*") are flagged when they modify nouns referring to people or institutions, often signaling implicit bias through judgment-laden descriptions.

Results. Table 4.5 summarizes the comparative experimental results obtained using the three evaluated methods. The heuristic-linguistic system is compared against Multitask DistilBERT and GPT-4o. The results reveal important differences in how each method handles the detection of subjective qualifying adjectives, expressions that carry implicit judgments about people or entities.

TABLE 4.5: Comparative performance metrics for subjective qualifying adjectives bias detection

Method	Macro Precision	Macro Recall	Macro F1-score	F1-score per class (False / True)
Heuristic-linguistic	35.15%	**91.76%**	**50.58%**	56.40% / **44.75%**
Multitask DistilBERT	**40.00%**	50.00%	44.44%	74.00% / 14.88%
GPT 4o	10.16%	83.33%	18.00%	**90.41%** / 18.06%

The heuristic-linguistic method achieves the highest overall macro F1-score (50.58%) and by far the highest recall (91.76%), indicating that it captures nearly all cases of subjective adjectival bias. Its positive class F1-score (44.75%) reflects its ability to reliably identify evaluative adjectives near named entities or nouns referring to people or institutions. However, the lower precision (35.15%) suggests that it may also flag borderline or context-dependent cases as biased, particularly when adjectives are used metaphorically or in non-subjective contexts. Still, its performance demonstrates that rule-based syntactic targeting of evaluative language can be highly effective for broad, multilingual bias screening.

Multitask DistilBERT achieves the best precision (40.00%) but lower recall (50.00%), resulting in a macro F1-score of 44.44%. It performs strongly on the negative class (74.00%) but relatively poorly on the positive class (14.88%). This suggests that the model is effective at confirming the absence of bias but struggles to detect subjective framing when it appears in contextually subtle or idiomatic forms. This conservative behavior may stem from the fact that subjectivity is not always clearly marked in training data, making the model more prone to under-detection in ambiguous cases.

GPT 4o exhibits very high recall (83.33%) but low precision (10.16%), producing a low overall macro F1-score (18.00%). Its performance is particularly skewed toward the negative class (90.41%), but it fails to offer reliable discrimination for positive cases (18.06%). This indicates that GPT 4o is highly sensitive to the presence of adjectives but lacks the task-specific filtering necessary to distinguish between evaluative and descriptive uses. Its generative architecture may lead it to overflag stylistic or emphatic language as biased, regardless of syntactic context or attribution.

▸ CHARACTERIZING PRESENTATION OF OPINIONS AS FACTS BIAS

Presentation of opinions as facts bias refers to the framing of subjective judgments, interpretations, or speculative claims as if they were objective, verifiable realities. This bias subtly distorts readers' understanding by masking the boundary between evidence-based statements and personal or editorial opinion, leading to a misperception of certainty and consensus.

Linguistically, this bias manifests through declarative sentences in the indicative mood that express strong, unqualified assertions without hedging or attribution. For instance, statements like "*the decision clearly demonstrates incompetence*" or its Spanish equivalent "*la decisión refleja una incompetencia absoluta*" rely on assertive syntax and evaluative language without providing evidential support, thus presenting opinions as if they were objective facts.

This effect is intensified when modal verbs or hedging expressions such as "*might*" or "*possibly*", are absent. The presence of evaluative adjectives and intensifiers (e.g., "*clearly*", "*obviously*") further amplifies rhetorical certainty. These constructions frequently follow a subject-verb-object pattern anchored by assertive verbs such as "*demonstrates*" or "*confirms*" in English, and "*refleja*" or "*demuestra*" in Spanish, reinforcing the illusion of factuality.

To heuristically detect this bias, we developed a method combining modality detection, hedging recognition, source attribution assessment, and evaluative language identification. The system first parses each sentence to verify the presence of indicative mood verbs serving as main predicates. It then checks for hedging elements or explicit source attribution (e.g.,

named entities). Sentences with assertive structures lacking these mitigating elements are flagged, especially when evaluative adjectives or rhetorical assertive phrases are present.

The method dynamically adjusts to English and Spanish using language detection, ensuring proper lexical and syntactic interpretation. This approach is particularly suited for analyzing journalistic, editorial, and political discourse, where this form of bias is both frequent and rhetorically subtle.

Algorithm 6 Heuristic detection of opinions-as-facts bias

Precondition: Sentence s, parser nlp, lexicons H (hedging), E (evaluative adjectives), A (assertive phrases)

Detect language l, load nlp_l, H_l, E_l, A_l
Parse s into tokens T, named entities E
Initialize flags: `hasFactualVerb`, `hasHedging`, `hasNamedEntity`, `hasEvalLang`, `hasAssertiveStruct` := `False`
for each token t in T **do**
    if t.`lemma` $\in H_l$ **then**
        `hasHedging` := `True`
    if t.`ent_type` $\in$ {PERSON, ORG, GPE, NORP} or t.`pos` = PROPN **then**
        `hasNamedEntity` := `True`
    if t.`pos` = ADJ and t.`lemma` $\in E_l$ **then**
        `hasEvalLang` := `True`
    if t.`pos` = VERB and t.`mood` = IND **then**
        `hasFactualVerb` := `True`
if any pattern in A_l appears in s **then**
    `hasAssertiveStruct` := `True`
if (`hasFactualVerb` or `hasAssertiveStruct`) and not `hasHedging` and (`hasEvalLang` or `hasAssertiveStruct`) **then**
    return `True`, "Opinions_as_facts"
else
    return `False`, "Opinions_as_facts"

By anchoring detection in modality, attribution, and evaluative framing, the heuristic (outlined in Algorithm 6) offers a scalable and interpretable framework for analyzing opinion-as-fact constructions in natural language. It is particularly suited for media discourse analysis, where ungrounded assertions framed as factual statements can significantly influence reader perception.

Results. Table 4.6 presents a comparative summary of the results obtained with the three evaluated methods. The heuristic-linguistic approach is compared to two advanced baselines: Multitask DistilBERT and GPT-4o. The task focuses on detecting instances where opinions are presented as facts, typically through assertive syntax, absence of hedging, and evaluative language lacking attribution.

TABLE 4.6: Comparative performance metrics for presentation of opinions as facts bias detection

Method	Macro Precision	Macro Recall	Macro F1-score	F1-score per class (False / True)
Heuristic-linguistic	50.00%	40.00%	44.00%	54.52% / **38.24%**
Multitask DistilBERT	**55.56%**	50.00%	**52.63%**	77.86% / 27.39%
GPT 4o	6.19%	**80.00%**	11.34%	**86.40%** / 11.37%

The heuristic-linguistic method achieves a balanced performance, with a macro precision of 50.00% and a macro recall of 40.00%, resulting in a macro F1-score of 44.00%. It performs particularly well on the true class (38.24% F1), reflecting its ability to identify unqualified subjective assertions. The rule-based approach, grounded in syntactic mood (indicative), hedging absence, and evaluative framing, appears well suited for capturing clearly defined unmitigated claims. However, its slightly lower recall suggests some limitations in detecting subtler or implicit formulations.

Multitask DistilBERT outperforms all other methods in terms of overall macro F1-score (52.63%), showing strong precision (55.56%) and recall (50.00%). Its high F1 on the negative class (77.86%) indicates excellent discrimination of non-biased factual assertions, though its performance on the positive class (27.39%) is more modest. This suggests a conservative classification strategy, potentially underestimating bias in ambiguous contexts. The multitask setup likely benefits from task interaction with related bias categories (e.g., evaluative adjectives or sensationalism), improving general context sensitivity.

GPT 4o displays the highest recall (80.00%) but very low precision (6.19%), yielding the lowest macro F1-score (11.34%). Its results are heavily skewed toward the negative class (86.40% F1), with limited true class precision (11.37%). This behavior indicates that GPT 4o is highly sensitive to assertive or emotional phrasing, often over-predicting the presence of bias. Its generative nature and reliance on prompt-driven heuristics may cause it to flag stylistic intensity without evaluating attributional context.

In summary, the heuristic approach offers interpretable and reasonably accurate detection of overt opinion-as-fact constructions. Multitask DistilBERT provides a robust and balanced classifier, suitable for both deployment and analysis. GPT-4o, while highly sensitive, lacks the precision needed for high-confidence annotation without post-processing.

▸ **BRIDGING FROM SENTENCE-LEVEL MANIFESTATIONS TO DOCUMENT-LEVEL ANALYSIS**

Read horizontally across the six manifestations, three patterns emerge. First, the rule-based heuristics tend to maximise recall and per-class F1 on the True class but at the cost of low precision, because they flag any sentence that contains the linguistic trigger regardless of context: this is consistent with their design as broad linguistic detectors and explains why subjective qualifying adjectives (where the linguistic trigger is the most explicit) is the manifestation on which the heuristic posts its highest macro F1. Second, the multitask DistilBERT classifier tends to be the most balanced macro-F1 performer across the six manifestations, but it does so by being conservative on the True class and dominant on the False class, that is, by relying disproportionately on the absence of bias rather than on its presence, a pattern whose generalisation implications surface again at the document level. Third, GPT-4o in 1-shot consistently exhibits high recall on the True class and high F1 on the False class, but very low precision overall, indicating that its prior over what counts as biased is broader than the annotation guidelines and that prompt-level intervention alone cannot calibrate it to a fine-grained linguistic schema. Taken together, these three patterns motivate the design choice (revisited at the system level in Section 4.3) of combining heuristic and neural sentence-level detectors as complementary, rather than competing, sources of evidence: the heuristic component recovers the linguistically explicit cases the neural model under-flags, while the neural model filters the contextual false positives the heuristic over-flags. The level-by-level analysis now moves up one layer, asking how these sentence-level manifestation signals relate to the global document-level judgement that an article is, or is not, biased.

4.2.2 *Media bias detection at document level*

The task of media bias detection at the document level is formulated, in the state of the art, as a binary classification problem: determining whether an entire news article exhibits media bias or can be considered unbiased. According to the hierarchical and perspectivist conceptualisation developed in Chapter 3, however, document-level bias is not a primitive property of the text but a higher-order layer that supervenes on the sentence-level manifestations characterised in the previous subsection. The document-level analysis reported here therefore inherits, rather than supersedes, what was just established for sentence-level manifestations: if those manifestations transfer poorly across datasets and confuse even the strongest model on positive cases, document-level classifiers built downstream are operating on a noisy linguistic substrate, and any apparent gain from going directly from raw text to a binary biased/unbiased label has to be read against that backdrop. This is one of the threads we want the reader to keep in mind as we move from manifestations to documents.

Unlike sentence-level characterisation, where individual linguistic cues can be analysed in isolation, document-level detection requires aggregating heterogeneous and potentially

subtle bias manifestations distributed across the text. This aggregation problem is non-trivial, as biased and non-biased segments frequently co-occur within the same article, and their relative salience may vary depending on both linguistic realisation and reader interpretation. Consistent with the principle that document-level decisions should be grounded in structured sentence-level evidence (formalised later in Section 4.3 as $f(D) = g(h(s_1), \ldots, h(s_n))$), the document-level systems studied here include explicit aggregation of sentence-level signals as one of the modelling paradigms, alongside encoder-only and decoder-only baselines that ignore the sentence layer entirely.

To address this challenge, we adopt an experimental strategy that combines three complementary modeling paradigms: heuristic aggregation methods, encoder-only transformer models, and decoder-only transformer models. This diversified setup is not intended to maximize performance in isolation, but to systematically explore the trade-offs between interpretability, representational capacity, and contextual reasoning at the document level. In particular, it allows us to contrast transparent, linguistically motivated baselines with data-driven neural models that capture discourse-level patterns, as well as generative models that rely on implicit reasoning through prompting.

As a first baseline, we implement a heuristic aggregation method built on top of the sentence-level detection system described in the previous section. For each document, heuristic rules are applied independently to each sentence in order to detect the presence of one or more bias manifestations. A document is then labeled as biased if more than 50% of its sentences are flagged as biased; otherwise, it is labeled as unbiased.

The choice of a 50% threshold is deliberate and conservative. As shown in earlier analyses, heuristic detectors tend to exhibit high recall but also generate a non-negligible number of false positives due to their broad lexical and syntactic coverage and their limited access to global context. By requiring that a majority of sentences in a document be flagged as biased, this aggregation strategy reduces the likelihood of classifying documents as biased based on isolated, ambiguous, or stylistically marked expressions. As a result, the heuristic aggregation method functions as a robust and interpretable baseline that prioritizes consistency over sensitivity, providing a meaningful point of comparison for more flexible neural approaches.

We then explore encoder-only transformer models by fine-tuning pretrained architectures such as DistilBERT on the document-level bias detection task. These models receive the full document as input, truncated to the maximum sequence length when necessary, and are optimized using binary cross-entropy loss. Encoder-based architectures are particularly well-suited for this task due to their ability to model long-range dependencies, capture global discourse structure, and integrate contextual information across the entire text.

To explicitly analyze the impact of annotation perspective on model behavior, fine-tuning is performed under two complementary settings. First, models are trained on the full dataset using majority-vote labels that aggregate all annotator judgments. Second, three additional models are trained on ideologically segmented subsets of the data, corresponding to annotations provided by annotators who self-identify as left-leaning, centrist, or right-leaning. In this way, each model is exposed to a coherent ideological perspective during training, allowing us to examine whether ideological consistency improves generalization or instead amplifies viewpoint-specific biases. All models are evaluated both on their corresponding ideological subsets and on the global test set, enabling a systematic comparison between aligned and misaligned training and evaluation conditions.

In addition to encoder-based models, we evaluate decoder-only transformers adapted for binary classification via prompt engineering. In this configuration, the model receives the full document as input followed by an explicit classification instruction, such as: "*Is this article biased? Answer 'true' or 'false'.*" The predicted label is derived from the first generated token. This setup allows us to assess the capacity of large generative models to perform document-level bias detection without task-specific fine-tuning.

To improve both performance and alignment with the task definition, we introduce an informed prompting variant that incorporates domain-specific annotation guidelines directly into the prompt. This variant employs in-context learning with two annotated examples (two-shot), where each example includes a short document excerpt, an explicit reasoning

trace, and the final binary decision. Inspired by recent work on chain-of-thought prompting for complex reasoning tasks [207], this approach encourages the model to follow a structured decision process rather than relying on shallow lexical cues. Importantly, this configuration enables us to evaluate the model's capacity for contextual reasoning, generalization, and adherence to human-defined criteria in a zero-fine-tuning setting, providing a useful contrast to the supervised encoder-based approaches.

[207] Zhang et al., 2024, "Chain of preference optimization: Improving chain-of-thought reasoning in llms"

▸ **RESULTS**

Table 4.7 reports the results of document-level media bias detection on MBBMD under a binary classification setting. All models are trained and evaluated on the same train/test split of MBBMD, using majority-vote labels as the reference ground truth. Given the class imbalance present in the dataset, we report macro-averaged precision, recall, and F1-score, ensuring that performance on both biased and unbiased documents contributes equally to the final metrics.

Across all configurations, the strongest performance is obtained by the DistilBERT model fine-tuned exclusively on annotations from centre-leaning annotators. This model achieves an F1-score of 88.89%, with perfect recall (100%) and high precision (80%). This result suggests that the centrist annotation perspective yields a particularly stable and internally consistent decision boundary, allowing the model to recover most biased documents while maintaining relatively few false positives.

When trained on the full dataset using majority-vote labels, DistilBERT attains a solid F1-score of 76.19%, driven primarily by high recall (85.71%) but lower precision (68.42%). This pattern indicates a tendency to over-predict bias in ambiguous cases, a behaviour that is consistent with the heterogeneous and perspectivist nature of the annotations aggregated into a single label.

Interestingly, the heuristic aggregation baseline achieves a slightly higher F1-score (78.8%) than the global DistilBERT model. This result highlights two important points. First, document-level bias often emerges from the accumulation of local linguistic cues, which heuristic sentence-level aggregation can capture effectively. Second, despite their simplicity, rule-based methods remain competitive baselines when carefully calibrated, reinforcing their value as transparent reference points rather than merely naive comparators.

Performance varies substantially across ideology-specific models. The DistilBERT model trained on left-annotator labels reaches an F1-score of 66.67%, while the model trained on right-annotator labels exhibits extreme recall (100%) but very low precision (40%), resulting in a markedly lower F1-score of 57.10%. These results suggest that models trained on ideologically polarized annotations tend to overgeneralize their decision criteria when applied outside their native perspective, leading to systematic false positives.

Decoder-only models perform notably worse in this supervised evaluation setting. GPT-4o in the zero-shot configuration reaches only 51.30% F1, indicating limited alignment with the annotation criteria. Incorporating task-specific guidance through informed prompting (2-shot) substantially improves performance to 68.90% F1, mainly by boosting recall (86%). Nevertheless, even with guidance, decoder-only models remain inferior to both encoder-based models and the heuristic baseline, underscoring the limitations of prompt-based approaches for fine-grained document-level bias detection without explicit training.

TABLE 4.7: Performance metrics for document-level media bias detection on MBBMD (train/test split), comparing heuristic, encoder-only, and decoder-only models under binary classification

Model	Macro precision	Macro recall	Macro F1-Score
Heuristic aggregation (50%)	76.5%	81.6%	78.8%
Fine-tuned DistilBERT	68.42%	85.71%	76.19%
Fine-tuned DistilBERT (left)	60.00%	75.00%	66.67%
Fine-tuned DistilBERT (center)	**80.00%**	**100.00%**	**88.89%**
Fine-tuned DistilBERT (right)	40.00%	**100.00%**	57.10%
GPT4-o (0-shot)	42.00%	68.00%	51.30%
GPT4-o-informed (2-shot)	57.28%	86.00%	68.90%

▸ **REGRESSION RESULTS (LEWIDI)**

Beyond binary classification, we model media bias as a continuous and perspectivist phenomenon by predicting the proportion of annotators who labelled each article as biased. This regression setting follows the LeWiDi paradigm, which treats annotator disagreement as a meaningful signal rather than as noise to be eliminated.

Concretely, the target variable for each document is the empirical proportion $y \in [0, 1]$ of annotators (out of the at least three who annotated that document) who labelled it as biased; thus $y = 0$ when no annotator marked the document as biased, $y = 1$ when all of them did, and intermediate values when annotators disagreed. Models are trained to predict this scalar agreement proportion using a regression head with mean-squared-error loss, and are evaluated against the held-out 20% test partition under three complementary metrics, all computed in the same $[0, 1]$ scale as the target. Mean Squared Error (MSE) penalises large deviations between predicted and gold scores and is sensitive to extreme misestimations, with 0 being optimal and larger values indicating worse fit. Mean Absolute Error (MAE) reports the average deviation in the original annotation scale, offering a more interpretable measure of typical error, with 0 again optimal and larger values worse. The coefficient of determination (R^2) quantifies how much of the variance in annotator agreement is explained by the model relative to a trivial baseline that always predicts the mean agreement score; positive R^2 values indicate that the model captures systematic intersubjective variation, $R^2 = 0$ indicates performance equivalent to the mean baseline, and negative values indicate performance worse than the trivial baseline. With this metric setup in place, Table 4.8 reports regression results on MBBMD, where models are evaluated on their ability to approximate continuous bias perception scores derived from annotator agreement. In this formulation, media bias is represented as a graded variable, making regression-based evaluation more suitable than categorical accuracy measures.

TABLE 4.8: Performance metrics for regression modeling (LeWiDi) using MBBMD

Model	MSE	MAE	R^2
Heuristic aggregation	0.06	0.20	0.14
DistilBERT	0.05	0.18	0.15
DistilBERT (left annotators)	0.14	0.30	-1.567
DistilBERT (center annotators)	**0.04**	**0.16**	**0.34**
DistilBERT (right annotators)	0.12	0.25	-0.926
GPT 4o (0-shot)	0.13	0.31	-1.18
GPT-4o-informed	0.09	0.26	-0.63

Across all models, encoder-based DistilBERT variants show the strongest alignment with human judgements. In particular, the model trained on centre-annotator labels achieves the best overall performance, with the lowest MSE (0.04), the lowest MAE (0.16), and the highest positive R^2 score (0.34). This indicates that the model is not merely fitting a coarse decision boundary, but is able to approximate the distribution of annotators' perceptions. The globally trained DistilBERT model also yields a positive R^2 (0.15), confirming that encoder-only transformers can learn graded bias signals when trained on perspectivist annotations.

In contrast, heuristic aggregation and decoder-only models perform substantially worse. Although heuristic aggregation attains moderate error values, its limited R^2 (0.14) suggests that it captures only a small fraction of the variance in human judgments. Decoder-only models, particularly in the zero-shot configuration, exhibit strongly negative R^2 values, indicating that their predictions fail to track systematic patterns in annotator disagreement and often underperform a simple mean-based predictor. Even with task-specific prompting (GPT-4o-informed), regression performance remains clearly inferior to encoder-based approaches.

Overall, these results demonstrate the value of regression-based evaluation for perspectivist datasets such as MBBMD. Modeling media bias as a continuous variable provides a more faithful representation of intersubjective disagreement and constitutes a stricter test of whether models capture human-like perception. The consistent superiority of DistilBERT in this setting further supports the suitability of encoder-based architectures for learning bias intensity under perspectivist supervision.

▸ PERFORMANCE ACROSS IDEOLOGICAL TEST SUBSETS

We now analyze the behavior of ideology-specific models when evaluated across test subsets grouped by political perspective (global, left, centre, and right). This section examines both their classification and regression performance in these ideologically segmented evaluation scenarios.

Despite the apparent clarity of the results shown in Tables 4.9 and 4.10, a closer inspection reveals promising patterns in the evaluation process. But it must be carefully interpreted due to the inherently skewed and sparsely populated nature of the ideological test subsets.

From a classification standpoint, the centre-aligned model stands out as the most robust across all test sets. It achieves the highest F1-score on the global subset (88.89%) and maintains competitive results across all subsets. The global model also exhibits reasonable balance, though with some drop in performance on the left subset. By contrast, the models trained on left and right annotations display lower F1 and precision when tested outside their ideological domain, especially the right-trained model, which exhibits very low F1 on the centre and left subsets (46.20% and 54.50%, respectively).

In terms of regression, a similar trend emerges. The centre-aligned model consistently outperforms the others in MSE, MAE, and R^2 across most test subsets, including the global one. It is, notably, the only model to achieve positive R^2 scores in all but one subset. This suggests that the centrist annotation perspective may provide a smoother and less ideologically polarized representation of bias, enabling better generalization.

However, these findings must be interpreted with caution. The number of positively labeled examples in each test subset is extremely limited, often as low as one or two, making most metrics volatile and sensitive to small prediction shifts. As a result, seemingly large performance differences may stem from a few data points. For example, the consistently perfect recall across all models and subsets is largely a function of the evaluation setup, where any correct positive prediction (often from a couple of biased item) yields 100% recall. Similarly, inconsistencies between MAE and MSE in some rows suggest that results may be driven by isolated outliers rather than stable predictive patterns.

In summary, while the centre-trained model emerges as the most balanced and reliable across subsets, these insights are constrained by the small sample size and ideological skew of the test data. Future work should rely on larger, ideologically balanced evaluation sets to confirm these observations and draw firmer conclusions about generalization across political perspectives.

TABLE 4.9: Binary classification performance (%) of DistilBERT models trained on ideologically filtered annotations, evaluated on each ideological test subset

Model trained on	Test Subset	Macro precision	Macro recall	Macro F1
Global	Global	68.42%	85.71%	76.19%
Left	Global	60.00%	75.00%	66.67%
Centre	Global	**80.00%**	**100.00%**	**88.89%**
Right	Global	40.00%	**100.00%**	57.10%
Global	Left	40.00%	66.67%	50.00%
Left	Left	**60.00%**	**100.00%**	**75.00%**
Centre	Left	**60.00%**	**100.00%**	**75.00%**
Right	Left	37.50%	**100.00%**	54.50%
Global	Centre	**100.00%**	66.67%	**80.00%**
Left	Centre	50.00%	66.67%	57.14%
Centre	Centre	60.00%	**100.00%**	75.00%
Right	Centre	30.00%	**100.00%**	46.20%
Global	Right	**50.00%**	**100.00%**	**66.67%**
Left	Right	11.11%	**100.00%**	20.00%
Centre	Right	**50.00%**	**100.00%**	**66.67%**
Right	Right	40.00%	**100.00%**	58.20%

TABLE 4.10: Regression performance of ideology-specific DistilBERT models on each test subset (MSE, MAE, and R^2)

Model trained on	Test Subset	MSE	MAE	R^2
Global	Global	0.054	0.183	0.157
Left	Global	0.166	0.304	-1.567
Centre	Global	**0.043**	**0.165**	**0.341**
Right	Global	0.125	0.251	-0.926
Global	Left	0.139	0.231	-0.297
Left	Left	0.187	0.280	-0.742
Centre	Left	**0.067**	**0.191**	**0.376**
Right	Left	0.156	0.245	-0.456
Global	Centre	0.102	**0.154**	0.168
Left	Centre	0.220	0.274	-0.797
Centre	Centre	**0.079**	0.204	**0.358**
Right	Centre	0.151	0.199	-0.231
Global	Right	0.141	0.221	-0.770
Left	Right	0.177	0.257	-1.235
Centre	Right	**0.085**	0.224	**-0.067**
Right	Right	0.101	**0.178**	-0.277

4.2.3 *Media bias categorization at document level*

Media bias categorisation goes beyond detecting bias; it aims to identify the specific type of bias exhibited. Unlike binary detection, it requires a deeper understanding of both linguistic features and the document's relationship to external discourse.

The task is formalised here as five independent binary classification problems, one per category in the hierarchical taxonomy introduced in Chapter 3 (intentional, spin, statement, coverage, and gatekeeping bias). Concretely, for each document and each category we predict whether the document exhibits that category of bias or not, allowing more than one category to be assigned to the same document when annotators identified multiple co-occurring strategies. This is therefore a multi-label setup at the level of the document (a document may carry several positive labels) but a binary problem at the level of each category (each label is either present or absent), and not a multi-class single-label problem in which exactly one category would be assigned per document. We adopt this formulation because the document-level annotations of MBBMD record the presence of each category independently, with co-occurrence preserved rather than collapsed; modelling the task as multi-class single-label would force an artificial winner-take-all decision that the annotation schema explicitly rejects. The terminology adopted for the rest of this subsection is therefore: *per-category binary classification* for the within-category decision, and *multi-label assignment* for the resulting set of positive categories per document. Phrases such as "multi-label binary classification" (which we used in earlier drafts) are avoided here because they conflate the two levels of the formulation.

We experimented with three modelling strategies for this task, all using the same per-category binary formulation just described. First, we fine-tuned encoder-only transformer models (specifically DistilBERT) with five independent binary heads, one per category. These models are effective at capturing linguistic and rhetorical patterns, and they performed particularly well for intentional and coverage bias.

Second, we tested decoder-only models using prompt-based per-category classification. Two configurations were explored: a zero-shot setting, where the model was prompted without any additional context, and an informed setting, in which the annotation guidelines were provided as part of the prompt to guide the classification process.

Finally, we implemented a retrieval-augmented generation (RAG) approach motivated by a specific limitation of the previous two configurations: two of the five categories (coverage and gatekeeping bias) are by construction *cross-document* phenomena, since they describe how a given outlet's article frames or selects an event *relative to* how other outlets do, and any system that sees only the article in isolation has, in principle, no way to distinguish a balanced single-source piece from a coverage-skewed one. RAG is the architectural response

to that limitation: it gives the decoder access, at inference time, to a small set of comparable articles covering the same event, so that the categorisation can be conditioned on alternative framings rather than on the input article alone. Concretely, the retrieval corpus is the union of all 500 (100 original + 400 counterfactual) MBBMD articles, indexed at article level using sentence-transformer embeddings (`paraphrase-multilingual-MiniLM-L12-v2`); for each input article, we retrieve the top 3 most similar articles by cosine similarity in that embedding space, excluding the input article itself, and concatenate their text after the input under an explicit `[Comparable coverage]` delimiter. The retrieved content is then passed to the same instruction-following decoder used in the informed setting, which predicts the per-category labels for the input. By providing comparative context grounded in the same topic, this setup enables the model to detect asymmetries in framing, selection, or omission that are unobservable in the input article in isolation, which are precisely the indicators of coverage and gatekeeping bias. The inclusion of external context is therefore expected to improve the model's ability to capture bias that manifests not within the article itself but in relation to alternative narratives, and crucially *does not* provide additional supervision for the within-document linguistic categories (intentional, spin, statement), where it is expected to be neutral or slightly noisy.

▸ **RESULTS**

The results in Table 4.11 show a clear advantage for the DistilBERT model over all GPT-4-based variants, both in aggregate macro-F1 and in per-label performance. DistilBERT achieves the highest F1-scores across all five media bias categories (intentional, spin, statement, coverage, and gatekeeping) with scores ranging from 53.57% to 67.14%. This strong performance highlights the effectiveness of fine-tuned encoder-only models for the per-category binary classification setup defined above.

TABLE 4.11: Binary classification performance for media bias categorization (majority vote)

Model	Bias Type	Precision	Recall	F1-score
DistilBERT	Intentional	**60.00%**	**72.73%**	**65.31%**
	Spin	**50.00%**	**57.14%**	**53.57%**
	Statement	**55.00%**	63.16%	**58.82%**
	Coverage	**65.00%**	**69.57%**	**67.14%**
	Gatekeeping	**58.00%**	**68.75%**	**62.71%**
GPT 4-o (0-shot)	Intentional	18.18%	22.22%	20.00%
	Spin	0.00%	0.00%	0.00%
	Statement	16.67%	20.00%	18.18%
	Coverage	8.70%	11.54%	10.00%
	Gatekeeping	0.00%	0.00%	0.00%
GPT 4-o-informed	Intentional	24.00%	28.00%	26.09%
	Spin	0.00%	0.00%	0.00%
	Statement	20.00%	23.53%	21.74%
	Coverage	14.00%	17.86%	15.79%
	Gatekeeping	0.00%	0.00%	0.00%
GPT 4-o-informed + RAG	Intentional	25.00%	66.67%	36.41%
	Spin	0.00%	0.00%	0.00%
	Statement	17.67%	**100.00%**	30.00%
	Coverage	16.53%	50.00%	23.33%
	Gatekeeping	0.00%	0.00%	0.00%

In contrast, GPT 4o (0-shot) shows very poor performance across the board, failing entirely on three of the five bias types (spin, gatekeeping, and coverage), and achieving only marginal F1-scores for intentional (20.00%) and statement bias (18.18%). The GPT 4o-informed variant yields modest improvements, most notably a 6-point gain for intentional bias, but still fails to detect spin and gatekeeping entirely. The GPT-4o-informed + RAG model shows stronger recall for intentional, statement, and coverage biases, but at the cost of very low precision. Despite these improvements, spin and gatekeeping remain the most difficult categories for all GPT-4-based variants, with consistent 0.00% F1-scores.

In summary, DistilBERT is the most balanced and accurate model across all bias types. GPT 4o variants benefit from task-specific instruction and retrieval, but remain limited in precision and category coverage. Spin and gatekeeping are systematically underdetected, indicating the need for model specialization or improved contextual prompting. RAG-enhanced prompts help improve recall but introduce noise without additional control mechanisms.

▸ **Regression results (LeWiDi)**

Table 4.12 presents the regression performance for the task of media bias categorization using the LeWiDi formulation, in which models are evaluated based on their ability to predict the proportion of annotators who perceive a given bias type in a document. This formulation embraces the perspectivist nature of bias and captures intersubjective disagreement more effectively than binary labels.

Across all bias types, DistilBERT obtains the best results, with the lowest mean squared error (MSE), the lowest mean absolute error (MAE), and the highest R^2 values. Its performance is particularly strong on coverage bias ($R^2 = 0.35$) and spin bias ($R^2 = 0.32$), suggesting that it effectively captures both overt and subtle evaluative cues that align with human perception. The consistently positive R^2 values across all categories (ranging from 0.25 to 0.35) indicate that the model explains a meaningful portion of the variance in annotator agreement, despite the inherent subjectivity of the task.

Table 4.12: Regression performance for media bias categorization (LeWiDi)

Model	Bias Type	MSE	MAE	R^2
DistilBERT	Intentional	**0.05**	**0.18**	**0.28**
	Spin	**0.04**	**0.16**	**0.32**
	Statement	**0.04**	**0.16**	**0.30**
	Coverage	**0.03**	**0.14**	**0.35**
	Gatekeeping	**0.04**	**0.15**	**0.25**
GPT 4o (0-shot)	Intentional	0.12	0.30	-0.40
	Spin	0.10	0.28	-0.35
	Statement	0.11	0.29	-0.38
	Coverage	0.10	0.27	-0.30
	Gatekeeping	0.11	0.28	-0.32
GPT-4o-informed	Intentional	0.08	0.25	0.10
	Spin	0.07	0.23	0.12
	Statement	0.08	0.24	0.08
	Coverage	0.07	0.22	0.15
	Gatekeeping	0.09	0.26	0.05
GPT-4o-informed + RAG	Intentional	0.21	0.39	-0.02
	Spin	0.29	0.51	-0.26
	Statement	0.24	0.45	-0.02
	Coverage	0.26	0.47	-0.06
	Gatekeeping	0.26	0.47	-0.17

The GPT 4o (0-shot) variant performs poorly in comparison, with negative R^2 values across all categories (e.g., -0.40 for intentional bias), indicating that it fails to outperform a naive mean-based baseline. These results suggest that, without task-specific calibration, GPT 4o struggles to internalize the subjective and distributed nature of bias perception.

The GPT-4o-informed variant, which includes task-specific prompting based on annotation guidelines, achieves moderate improvements over the zero-shot version. It produces positive R^2 values across all categories, with its strongest result on coverage bias ($R^2 = 0.15$). However, its gains remain modest relative to DistilBERT, and its predictive accuracy appears more limited in categories like gatekeeping ($R^2 = 0.05$) and statement bias ($R^2 = 0.08$). This suggests that while prompt-based guidance helps to align predictions with annotation criteria, it is not sufficient on its own to replicate human-level granularity.

Unexpectedly, the GPT-4o-informed + RAG model performs worse than its non-retrieval counterpart. It shows inflated error rates and returns negative R^2 values for all categories except intentional bias (-0.02). These results indicate that retrieval may introduce noise

or irrelevant context that confuses rather than supports model prediction in this regression setting. The mismatch between retrieved content and the model's expectations under a regression objective may contribute to this degradation.

Overall, the results confirm that encoder-based models like DistilBERT are best suited to model intersubjective agreement in a perspectivist setup. While GPT 4o models benefit from informed prompting, they remain less reliable in estimating nuanced degrees of perceived bias. Moreover, the use of retrieval-augmented generation appears to have limited utility, and may even be detrimental, under this continuous regression formulation.

4.2.4 *Evaluating hierarchical multi-label and LeWiDi classification*

In Subsections 4.2.2 and 4.2.3, we evaluated media bias detection at document level as two separate tasks: binary classification of whether a document is biased (Subsection 4.2.2), and multi-label categorisation of bias types (Subsection 4.2.3). However, that evaluation was conducted in a flat, non-hierarchical fashion, without enforcing the dependency that categorisation should logically presuppose a prior bias prediction. This is not merely a missed methodological refinement: under the hierarchical and perspectivist conceptualisation of media bias developed in Chapter 3, treating detection and categorisation as independent flat tasks is conceptually inadequate, because it severs the categorisation decision from the very condition (the presence of media bias) under which it is meaningful, and treats two layers of the same nested phenomenon as if they were unrelated classification problems. The hierarchical reformulation introduced below is therefore the methodological consequence of the conceptual position adopted in Chapter 3, not a post-hoc improvement. Additionally, while Subsections 4.2.2 and 4.2.3 incorporated annotator disagreement using regression metrics on the agreement percentage, that evaluation still compares system outputs to a scalar summary; it does not fully reflect the perspectivist foundations of the LeWiDi paradigm, which require metrics that compare predicted distributions over labels to the empirical distribution of annotator judgements rather than to majority-vote or scalar approximations.

While the classification task of media bias is fundamentally hierarchical, since bias categories are nested within broader conceptual groupings such as context bias and intention bias, our evaluation treated the binary and multi-label tasks independently. In an ideal setting, predictions of specific bias categories (e.g., coverage bias or statement bias) should be conditioned on a prior determination that the document is biased. Predicting one subtype instead of another may still reflect partial semantic alignment. Yet, because our evaluation did not model this dependency explicitly, it cannot fully capture the structural relationships among categories. This highlights the need for structure-aware metrics that account for partial correctness within the hierarchy.

Moreover, although bias detection is inherently perspectivist, our current evaluation setup only reflects this subjectivity through regression-based metrics applied to agreement percentages. In contrast, the LeWiDi paradigm treats annotator disagreement not as noise to be minimized but as a core signal of intersubjective variation. A truly perspectivist evaluation would require metrics that compare system predictions to full distributions over labels, rather than to majority-vote or scalar approximations.

Standard evaluation metrics assume flat label spaces and categorical ground truth, making them ill-suited for a task that is both hierarchically structured and subjectively annotated. To address these limitations, we adopt a family of evaluation metrics based on the Inter-annotator Centric Measure (ICM), introduced by Amigo and Delgado [7] and implemented in the EvALL platform [5]. These metrics have been extended to handle perspectivist scenarios and are operationalized for bias detection through the ODESIA evaluation environment [6].

[7] Amigo and Delgado, 2022 "Evaluating extreme hierarchical multi-label classification"

[5] Amigó et al., 2017, "Evall: Open access evaluation for information access systems"

[6] Amigó et al., 2023, "ODESIA: Space for Observing the Development of Spanish in Artificial Intelligence."

In our experiments, we focus on two extended variants of the Inter-annotator Centric Measure (ICM) that better reflect the hierarchical and perspectivist nature of the task. Both variants assess the same underlying alignment between system predictions and the empirical distribution of annotator judgements; they differ only in how that alignment is reported on the output scale.

- **ICM-soft**: this variant relaxes the exact-match constraint of the original ICM by assigning

partial credit when the predicted labels are semantically or taxonomically related to those present in the gold annotation distribution, leveraging the structure of the bias taxonomy to capture near-misses that occur within the same conceptual branch and acknowledging the graded nature of error in hierarchical classification. ICM-soft is reported in bits and is unbounded: positive values indicate that the system's prediction is more informative than the prior, negative values that it is less informative (the prediction would mislead an inter-annotator-centric reader), and large negative values are therefore expected in the worst-aligned configurations.

- **ICM-norm-soft**: this is the same quantity as ICM-soft, mapped to a fixed $[0, 1]$ interval to make comparisons across models and tasks easier to read. The mapping uses the gold-label distribution as the upper anchor (an ICM-soft value matching the gold receives the score 1.0) and the negative of the gold-label score as the lower anchor (an ICM-soft value of $-$gold receives the score 0.0); values below $-$gold are truncated to 0.0 rather than extending below zero. ICM-norm-soft is therefore a bounded readability transformation of ICM-soft, not a different quantity, and the two are reported jointly so that the unbounded informational reading and the bounded comparative reading can be inspected side by side.

The decision to retain both variants in the result tables is deliberate: ICM-soft preserves the informational interpretation that makes the metric epistemically distinctive (a negative score is a real claim about the system, not a normalisation artefact), while ICM-norm-soft offers the bounded scale needed when the table has to be read at a glance. Reading any single column in isolation would either lose the informational content (in the bounded case) or lose comparability (in the unbounded case), so the rest of this subsection always reports them together.

We apply these metrics to a set of representative models for multi-label bias categorization, exploring two evaluation configurations. In the non-cascaded setting, models predict bias categories for all documents, regardless of whether the document is actually biased or not. This flat approach disregards the hierarchical dependency between bias detection and categorization. In contrast, the cascaded configuration first applies a binary classifier to determine whether a document is biased; only when the document is predicted as biased are the bias categories assigned. This approach better reflects the hierarchical nature of the task, enforcing the logical constraint that bias types should only be predicted when media bias is present.

We use the same set of models described in Section 4.2.3 to generate predictions for bias categorization. These include DistilBERT, an encoder-only model fine-tuned on the MBBMD dataset; GPT-4o in a zero-shot configuration; GPT-4o with in-context examples (informed prompting); and GPT-4o augmented with retrieval-augmented generation (RAG) capabilities. This variety allows us to compare encoder-based and decoder-based architectures, as well as to assess the impact of contextual information and external knowledge on hierarchical perspectivist evaluation.

Table 4.13 summarizes the evaluation of these models using the three ICM variants. All models are tested on the same evaluation set, and their outputs are formatted as distributions to match the LeWiDi annotation style.

Three reading aids should be kept in mind. First, the Gold upper bound (ICM-soft $=$ 8.7235, ICM-soft Norm $=$ 1.0) is the score that the gold annotation distribution receives against itself; it is the ceiling against which the model rows should be compared, and the upper anchor used to normalise ICM-soft Norm to the $[0, 1]$ interval (Subsection 4.2.4). The corresponding lower anchor is $-$gold $= -8.7235$, below which ICM-soft Norm is truncated to 0.0; this is exactly what happens for GPT-4o (0-shot), whose ICM-soft of -14.4272 falls below that lower anchor and is therefore reported as 0.0 on the normalised scale rather than as a more informative negative number. The unbounded ICM-soft column makes the magnitude of that misalignment visible (the model is more than two gold-units away from the gold distribution, in the wrong direction); the bounded ICM-soft Norm column is what allows the table to be read at a glance. Second, ICM-soft scores are unbounded and a strongly negative value, such as the -14.4272 obtained by GPT-4o (0-shot), is not a numerical anomaly but

TABLE 4.13: Evaluation of bias-type categorisation using hierarchical perspectivist metrics. The Gold row reports the ICM-soft of the gold annotation distribution against itself: it is the upper bound to which any system should be compared, and the upper anchor used to normalise ICM-soft Norm to the $[0, 1]$ scale (Subsection 4.2.4).

Model	**ICM-soft**	**ICM-soft Norm**
Gold (upper bound)	*8.7235*	*1.0000*
DistilBERT	1.8112	0.6038
GPT-4o (0-shot)	-14.4272	0.0000
GPT-4o (informed)	-4.0055	0.2704
GPT-4o + RAG	-3.2354	0.3145
Cascaded DistilBERT	**3.6652**	**0.7101**
Cascaded GPT-4o (0-shot)	-0.2643	0.3848
Cascaded GPT-4o (informed)	-1.8115	0.3961
Cascaded GPT-4o + RAG	-1.2539	0.4281

the metric's faithful report that the model's predicted distribution is severely misaligned with the annotator distribution; the cascaded variants of the same model already recover this to -0.2643, an improvement of more than 14 ICM-soft units, which is itself a strong signal that the failure was structural (flat assignment of categories to non-biased documents) rather than capacity-related. Third, the table is intentionally split into a non-cascaded block and a cascaded block to make the conditional-task contribution visible: every system improves under the cascaded variant, and the gap between Cascaded DistilBERT (the closest to the gold ceiling) and the cascaded GPT-4o variants is what motivates the architectural commitment to encoder-only categorisation in the system of Section 4.3.

The results show that Cascaded DistilBERT system achieves the highest scores across all metrics, indicating exceptional alignment with annotator judgements and strong hierarchical awareness. The combination of a high-quality biased/non-biased classifier with a targeted model leads to both precise and semantically grounded predictions, and the resulting ICM-soft of 3.6652 recovers about 60% of the way from the ICM-soft of the trivial baseline (which sits near zero) to the gold ceiling.

The non-cascaded DistilBERT also performs well, but is outperformed by its cascaded variant. This confirms the value of the hierarchical setup, in which category-level predictions are only made when warranted by a bias classification decision.

GPT-4o models, particularly in non-cascaded form, perform significantly worse. Their ICM-soft scores are negative, indicating that predictions tend to contradict rather than approximate annotator consensus. The normalisation maps these scores onto the $[0, 1]$ scale, where they cluster well below the cascaded encoder, but the unbounded column should be the reference for interpreting the underlying misalignment.

Interestingly, all GPT-4o cascaded variants outperform their non-cascaded versions, albeit modestly. This shows that structuring the task hierarchically helps even large generative models, though not enough to close the gap with smaller, specialised encoder-based systems.

▸ ERROR PATTERNS AND MODEL BEHAVIORS ACROSS TAXONOMY LEVELS

Building on the perspectivist evaluation introduced in this previous subsection, we now consolidate the main patterns of error and model behavior observed across the three levels of analysis considered in this section: sentence-level detection, document-level binary classification, and multi-label bias categorization. By comparing these dimensions, we identify which architectures and methods show robustness, which ones exhibit systematic weaknesses, and how these insights inform the design of a hierarchical framework for media bias modeling.

At the sentence level, encoder-only models consistently outperform decoder-only approaches in terms of both precision and F1-score. This confirms the importance of bidirectional context in detecting localized manifestations of bias. However, heuristic and linguistically driven models (while less accurate overall) achieve remarkably high recall. Their ability to capture many true positives with relatively simple pattern-matching strategies suggests that they remain valuable, especially in high-recall settings such as weak supervision or

pre-filtering. These results justify the inclusion of both paradigms in the hierarchical system: heuristics can act as broad detectors, while transformer-based models provide precision and contextual disambiguation.

In binary classification at the document level, results further reinforce the value of bottom-up aggregation. Simple majority-vote heuristics over sentence-level outputs perform comparably to fine-tuned transformer models. This convergence supports the theoretical view that document-level bias emerges from the accumulation of local signals. Moreover, the regression formulation based on annotator agreement percentages (LeWiDi) demonstrates the feasibility of modeling intersubjective perceptions rather than hard labels. The strong performance of this approach, especially for encoder-based models, highlights the epistemological relevance of perspectivist supervision and anticipates the need for evaluation metrics such as ICM, which respect such intersubjective distributions.

In contrast, the results for multi-label categorization of bias types are more mixed. Decoder-only models, including those enhanced with retrieval-augmented generation (RAG), perform poorly. This is particularly disappointing for categories such as coverage bias and gatekeeping bias, where it was hypothesized that retrieving background or comparative context would be crucial. While we still believe that such external knowledge is necessary to fully detect these forms of bias, current implementations of retrieval-based models do not yet meet the task's demands. For this reason, encoder-based models, despite their limitations, remain the foundation for categorical bias modeling in this thesis. Crucially, the evaluation of this task using ICM-soft and ICM-norm-soft reveals that many incorrect predictions still operate within semantically adjacent regions of the taxonomy, offering partial alignment with annotator distributions. These subtleties would be overlooked under standard evaluation metrics.

Taken together, these results reveal important patterns in how different model types handle bias across levels of abstraction. Decoder-only models underperform across the board, particularly in tasks requiring detailed discrimination. Encoder-based transformers prove effective, especially when paired with perspectivist supervision. Heuristic models, despite their simplicity, remain relevant due to their broad recall and interpretability. Importantly, the small and ideologically unbalanced nature of the test subsets, especially for category-level and group-specific evaluations, limits the statistical reliability of some metrics. In these cases, extreme R^2 values or near-perfect recall must be interpreted with caution. Here again, the use of disagreement-aware evaluation tools such as ICM helps to mitigate overinterpretation by reflecting uncertainty directly into the scoring function.

Read together, the level-by-level findings of this section converge on a single conceptual statement that frames the system design of Section 4.3. Sentence-level analysis (Subsection 4.2.1) shows that the linguistically explicit manifestations of media bias are exploitable but only partially transferable as isolated detectors; document-level analysis (Subsection 4.2.2) shows that doc-level decisions are most reliable when they aggregate sentence-level evidence rather than mapping raw text directly to a binary label, and that perspectivist supervision (modelling disagreement as signal rather than noise) consistently outperforms majority-vote training; document-level categorisation (Subsection 4.2.3) shows that flat multi-label assignment is structurally inadequate, and that the hierarchical reformulation of Subsection 4.2.4 (cascaded ICM-soft / ICM-norm-soft) is the methodological consequence of the conceptualisation developed in Chapter 3 rather than an optional refinement. The systems that scored best at every layer were precisely those that respected, rather than flattened, the layer above them: cascaded categorisation outperformed flat categorisation, sentence-aggregated document classification outperformed direct doc-level classifiers, and perspectivist regression captured signal that majority-vote binary supervision threw away. Detecting media bias well, then, is not a matter of optimising any single layer in isolation; it is a matter of building a system that is hierarchical (because the phenomenon is hierarchical), perspectivist (because perception is perspectivist), bottom-up (because document-level decisions emerge from sentence-level evidence), and integrated (because the layers are not independent). MBBMD makes this kind of system trainable, and the rest of this chapter argues that the resulting framework is also a more complete response to media bias than what the state of the art allows.

These observations guide the architectural and methodological choices presented in the

next section. There, we introduce a hierarchical model that integrates symbolic and neural methods across levels of analysis, balancing precision, recall, perspectivist disagreement, and interpretability in a unified framework for media bias analysis.

4.3 A HIERARCHICAL PERSPECTIVIST SYSTEM

This section presents the main systems-level contribution of this thesis: a hierarchical and perspective-aware framework for media bias detection and characterisation. Rather than proposing a new end-to-end architecture in the conventional sense, the goal of this framework is to operationalise the conceptual and empirical insights developed throughout the thesis into a coherent computational system. In particular, it translates the generalisation failures identified in Section 4.1, the perspectivist and hierarchical structure of MBBMD introduced in Section 3.2, and the multi-level analytical findings of Section 4.2 into a concrete system design. The closing paragraph of Section 4.2 already stated the load-bearing claim: detecting media bias well requires a system that is jointly hierarchical, perspectivist, bottom-up, and integrated. This section spells out how those four properties are realised in a single architecture.

Accordingly, the proposed framework is deliberately modular, cascaded, and hierarchical, and each of its components is introduced as a direct response to one of the four properties just listed. Sentence-level modelling captures transferable rhetorical mechanisms (the bottom-up property), document-level models encode holistic contextual patterns (the hierarchical property), and ideology-aware representations reflect perspectivist variation in bias perception (the perspectivist property); the stacking-based decision layer that integrates them realises the integration property. None of these components is intended to function as a standalone solution: their analytical value emerges from their interaction within a unified decision framework that respects the functional dependencies introduced above. The framework is evaluated both on MBBMD and on multiple external datasets in Chapter 5, allowing us to directly assess whether the proposed design principles translate into improved generalisation beyond the conditions under which individual models were trained.

The choice of MBBMD as the primary training resource for the hierarchical system is a direct consequence of the same four properties: it is the only available dataset that provides the three-layer annotation structure (sentence-level manifestations, document-level perception with preserved annotator disagreement, and document-level categories) required to train the components without flattening them. As shown in Section 4.1, no single existing corpus supports the simultaneous training of sentence-level detectors, ideology-aware document classifiers, and category-level models within a unified framework, and multi-corpus configurations explored in that section, while moderately useful for flat models, cannot supply the hierarchical annotation structure that the system requires. MBBMD was designed precisely to fill this gap, and its use here is a direct consequence of the dataset design principles outlined in Section 3.2, not a confounding choice. The remainder of this section unfolds the framework's design rationale, architectural components, and training pipeline; the cross-dataset evaluation, the dataset-versus-system attribution, and the discussion of the three research hypotheses formulated in Chapter 1 are reported in Chapter 5.

4.3.1 *Design requirements derived from the generalization problem*

We now make explicit the design requirements that the system has to satisfy. Each requirement follows directly from one of the empirical or conceptual findings just summarised, and translates it into a constraint on the architecture. The requirements are not imposed *a priori*: they reflect what any system aiming to robustly detect media bias across contexts has to satisfy, given the diagnoses of Section 4.1 and the level-by-level findings of Section 4.2.

A first requirement concerns the level of abstraction at which media bias should be modeled. The cross-dataset experiments in Section 4.1 demonstrated that models trained on document-level labels alone tend to overfit to dataset-specific lexical, topical, or stylistic regularities. When evaluated on unseen domains, such models systematically fail to recognize bias expressed through alternative rhetorical strategies. This observation motivates the need for abstractions that are more transferable across contexts. In this work, such abstractions are operationalized at the sentence level through explicit modeling of linguistic bias manifestations,

which capture recurring rhetorical mechanisms (e.g., emotionalization, mind reading, or biased word choice) that transcend specific topics or media ecosystems.

From a formal perspective, this requirement implies that document-level bias cannot be treated as an atomic property inferred directly from raw text. Let a document $D = \{s_1, \ldots, s_n\}$ be composed of a sequence of sentences. Document-level media bias must instead be defined as an explicit function over sentence-level representations:

$$f(D) = g\big(h(s_1), \ldots, h(s_n)\big),$$

where $h(\cdot)$ denotes sentence-level representations encoding linguistically grounded bias signals, and $g(\cdot)$ denotes an aggregation mechanism operating at the discourse level. This formulation is intentionally abstract and does not commit to a single end-to-end differentiable architecture. Rather, it constrains the design space by requiring that document-level decisions be grounded in structured sentence-level evidence.

A second requirement follows from the perspectivist nature of media bias perception. As shown in Section 3.2, disagreement among annotators is not merely annotation noise, but an informative signal reflecting ideological and interpretive variability. Systems that collapse this variability into a single majority-vote label implicitly assume a homogeneous notion of bias, thereby discarding precisely the information needed to model contested or ambiguous cases. Consequently, a robust media bias detection system must be compatible with perspectivist representations and allow multiple, potentially conflicting interpretations to coexist within the decision process. This requirement rules out architectures that rely exclusively on monolithic classifiers trained on aggregated labels.

Formally, this implies that document-level predictions should be explicitly conditioned on perspective rather than treated as universally valid outcomes. Let p denote an annotator perspective and z_D a discourse-level representation derived from sentence-level signals. Perspective-aware prediction can then be expressed as:

$$y_D^{(p)} = f\big(z_D \mid p\big),$$

making subjectivity an explicit modeling component rather than an implicit source of variance. Importantly, this formulation does not posit the existence of a single ground-truth label, but instead accommodates multiple legitimate interpretations conditioned on perspective.

Third, the analysis in Section 4.2 showed that different aspects of media bias are best captured at different levels of granularity. Sentence-level mechanisms are effective for localizing and explaining bias cues, while document-level models are better suited to capturing global framing and contextual coherence. Media bias categorization, in turn, often depends on cross-sentence patterns and editorial strategies that cannot be inferred from isolated fragments. These findings imply that a single-level architecture is inherently insufficient. Instead, system design must support a hierarchical organization in which multiple analytical resolutions coexist and contribute complementary information.

A fourth requirement concerns integration. While individual models operating at specific levels of granularity can achieve strong performance in isolation, the central challenge lies in combining their outputs in a principled manner. Naïve aggregation strategies, such as fixed thresholds or majority voting, are ill-suited to this task, as they fail to account for the heterogeneous nature of the signals involved and their varying reliability across contexts. Effective integration therefore requires a mechanism capable of learning how to weigh and reconcile diverse sources of evidence, including linguistic heuristics, neural predictions, and perspectivist signals, without imposing hard-coded assumptions.

Finally, generalization must be treated as a system-level property rather than as a by-product of individual component performance. As the experiments in Section 4.1 make clear, improvements at the model level do not necessarily translate into robustness under distributional shift. Consequently, the evaluation of any proposed architecture must explicitly test its behavior across datasets that differ in annotation schemes, political contexts, and discursive styles. This requirement has direct implications for system design: components must be modular, reusable, and interpretable, enabling both cross-dataset evaluation and targeted analysis of failure modes.

Together, these requirements define the design space within which the proposed hierarchical and perspective-aware framework is constructed. The remainder of this section details how each requirement is operationalized through specific architectural choices, beginning with an overview of the system's structure and processing flow.

4.3.2 *Architectural overview and design rationale*

This subsection presents the overall system architecture and the rationale behind its design. The specific training procedures, hyperparameter configurations, and implementation details are described separately in Subsection 4.3.3.

The design requirements outlined in the previous section impose a set of functional and methodological constraints on system design. In particular, the formalization of document-level bias as an explicit function of sentence-level representations, and the conditioning of predictions on perspective, rule out flat end-to-end architectures that directly map raw text to a single bias label. Within this constrained design space, the architecture proposed here represents one concrete instantiation that operationalizes these requirements in a modular and empirically testable manner.

Rather than relying on a single classifier trained to infer bias holistically from raw documents, the proposed framework adopts a modular, cascaded, and hierarchical organization. This design choice directly reflects the assumptions formalized in the previous subsection: document-level bias is treated as an emergent property derived from structured sentence-level evidence, and bias perception is modeled as inherently perspective-dependent rather than universally fixed.

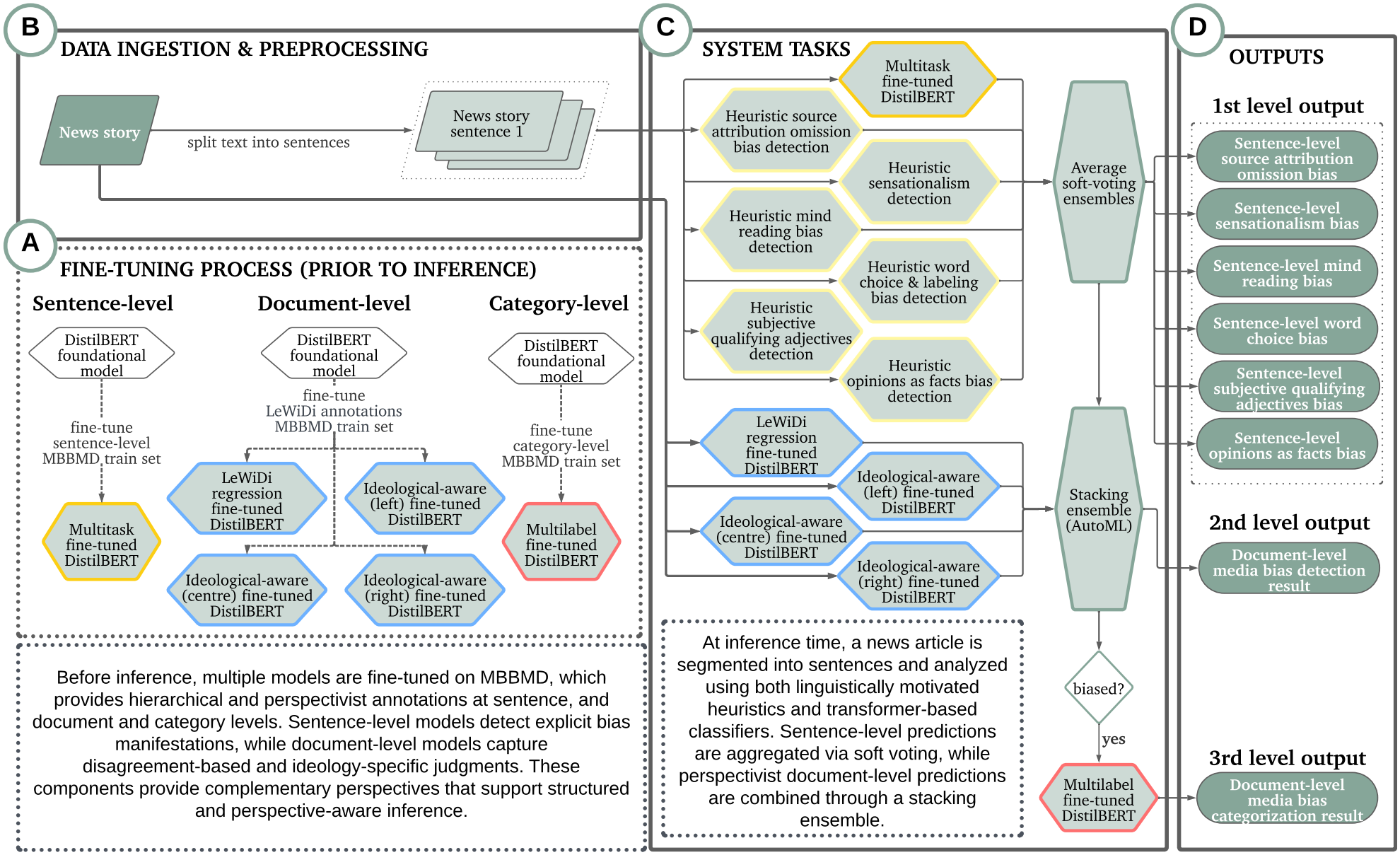
At a high level, the architecture is structured around three complementary analytical axes: (i) sentence-level modeling of concrete bias manifestations, (ii) document-level bias detection and characterization, and (iii) perspectivist modeling of ideological variability. Each axis corresponds to a specific design requirement identified earlier: transferability through abstraction, compatibility with disagreement, and integration across levels of granularity. Importantly, none of these components is intended to function as a standalone solution. Their analytical value emerges from their interaction within a unified decision framework that respects the functional dependencies introduced above.

The resulting system follows a cascaded processing flow. Incoming documents are first decomposed into sentence-level units, enabling the extraction of localized linguistic signals that instantiate the sentence-level representations $h(s_i)$ introduced previously. In parallel, document-level components encode global contextual patterns that are not reducible to individual sentences. Perspectivist components operate alongside these layers, producing perspective-conditioned signals that correspond to alternative evaluations of the same discourse-level representation. This cascading structure ensures that information flows from fine-grained, interpretable cues toward increasingly abstract and integrative representations, in line with the functional abstraction $f(D) = g(h(s_1), \ldots, h(s_n))$.

A key principle guiding the architecture is modularity. Each component is designed to address a specific analytical role and can, in principle, be replaced or extended without altering the overall system logic. This modularity is not merely an engineering convenience, but a methodological necessity: it allows the system to be evaluated, diagnosed, and adapted across datasets with different annotation schemes, linguistic conventions, and political contexts. In contrast to monolithic architectures, the proposed framework makes it possible to isolate which levels of analysis contribute to successful generalization and which reveal persistent limitations.

Another central design decision concerns integration. Given the heterogeneity of the signals produced by the different components, ranging from rule-based outputs to probabilistic neural predictions and perspective-specific scores, simple aggregation strategies are insufficient. The architecture therefore incorporates an explicit integration layer that learns how to combine these signals in a data-driven manner. This layer implements the aggregation function $g(\cdot)$ at the system level, mediating between potentially conflicting evidence, balancing precision and recall, and adapting the relative importance of different components depending on the context in which bias is expressed.

FIGURE 4.13: Overview of the proposed hierarchical and perspective-aware media bias detection system. Block A covers training; Blocks B–D cover inference (sentence-level analysis, document-level integration, and conditional bias categorization).



Finally, the architecture is explicitly designed to separate detection from characterization. While media bias detection aims to determine whether a document exhibits biased framing, media bias characterization seeks to explain how that media bias is realized at the editorial level. Treating these tasks as sequential rather than simultaneous reflects a methodological commitment derived from the formal constraints discussed earlier: categorization without prior detection risks forcing interpretative labels onto neutral content, while detection without characterization provides limited explanatory value. The cascaded design of the system enforces this separation, ensuring that higher-level interpretive decisions are grounded in prior evidence.

Figure 4.13 provides a schematic overview of the complete architecture, highlighting both the conceptual organization of the system and its processing flow. Block A corresponds to the training phase, in which the different sentence-level, document-level, and perspective-aware models are fine-tuned using MBBMD under controlled experimental settings. Blocks B–D represent inference-time processing and constitute the core of the proposed framework.

Block B corresponds to the input preprocessing stage. At this stage, incoming documents are segmented into sentences in order to establish the basic structural units required for subsequent multi-level analysis. The role of sentence-level modeling as a transferable abstraction, built on top of this segmentation, is described in detail in Subsection 4.3.2.

Block C implements the core modeling and integration stage of the framework. Here, sentence-level media bias manifestations are predicted using a hybrid combination of heuristic detectors and multitask neural models, while holistic document-level media bias detection and ideology-aware classifiers provide global and perspective-conditioned assessments. These heterogeneous signals are aggregated and reconciled through a stacking-based decision layer, which learns how to integrate local, global, and perspectivist evidence into a single document-level bias decision. Document-level modeling, perspectivist representations, and their integration through stacking are described jointly in Subsection 4.3.2, reflecting their tight conceptual and architectural coupling.

Finally, Block D represents the final output stage of the pipeline. At this stage, the system exposes the document-level media bias verdict together with its associated explanatory signals, including sentence-level media bias manifestations and, when applicable, the predicted bias category. The decision to activate media bias categorization is made upstream as part of the cascaded inference flow following document-level media bias detection in Block C, enforcing a clear separation between detection and characterization. Media bias categorization as a downstream and conditional task is described in Subsection 4.3.2.

▸ SENTENCE-LEVEL MODELING AS TRANSFERABLE ABSTRACTION

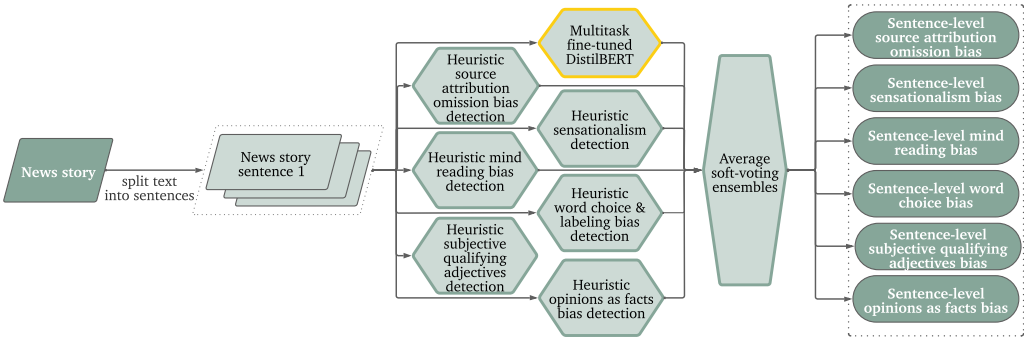
The first operational layer of the proposed framework focuses on sentence-level analysis, which plays a central role in addressing the generalization problem.

This layer operationalizes the linguistic taxonomy introduced in Chapter 3, which characterizes media bias through a set of concrete manifestations such as emotionalization, mind reading, biased word choice, omission of attribution, subjective qualification, and presentation of opinion as fact. These manifestations are defined at the sentence level and grounded in observable linguistic cues, making them suitable both for interpretability and for cross-context generalization. In the system architecture, they constitute the sentence-level representations that feed higher-level components.

Figure 4.14 illustrates this sentence-level processing block, showing how individual sentences are analyzed through parallel heuristic detectors and a multitask neural classifier, and how their outputs are combined into structured sentence-level bias signals.

To capture these mechanisms, the system combines two complementary modeling strategies: rule-based linguistic heuristics (see Subsection 4.2.1) and a multitask neural classifier. The motivation for this hybrid approach is twofold. On the one hand, rule-based heuristics provide high-precision, interpretable signals grounded in linguistic theory. On the other hand, neural models offer contextual sensitivity and robustness to lexical variation, mitigating the brittleness of handcrafted rules. Rather than treating these approaches as competing

FIGURE 4.14: Sentence-level bias detection block combining heuristic detectors and multitask DistilBERT via soft-voting ensembles.



alternatives, the framework integrates them as complementary sources of evidence within the sentence-level processing block.

The heuristic component consists of six independent detectors, each corresponding to a specific bias manifestation. These detectors rely on curated lexical resources, regular expressions, and syntactic patterns designed to capture explicit linguistic realizations of bias. While highly interpretable, such heuristics are known to suffer from overgeneration, particularly when applied to heterogeneous corpora. For this reason, their outputs are not used in isolation, but are explicitly designed to be combined with those of a neural model.

In parallel, a multitask DistilBERT classifier is fine-tuned on MBBMD to jointly predict the same set of sentence-level bias manifestations. Multitask learning is employed to encourage shared representations across related bias phenomena, reflecting the observation that different manifestations often co-occur or interact within the same sentence. This design choice reduces overfitting to individual labels and promotes the learning of more generalizable features.

The outputs of the heuristic detectors and the multitask neural model are combined through a soft-voting ensemble. For each sentence and each bias manifestation, the heuristic model produces a binary decision, while the neural model outputs a continuous probability score. These values are averaged to yield a final confidence score, balancing precision and recall. Empirically, this integration strategy consistently improves performance over either component alone, particularly by reducing false positives generated by purely rule-based detection.

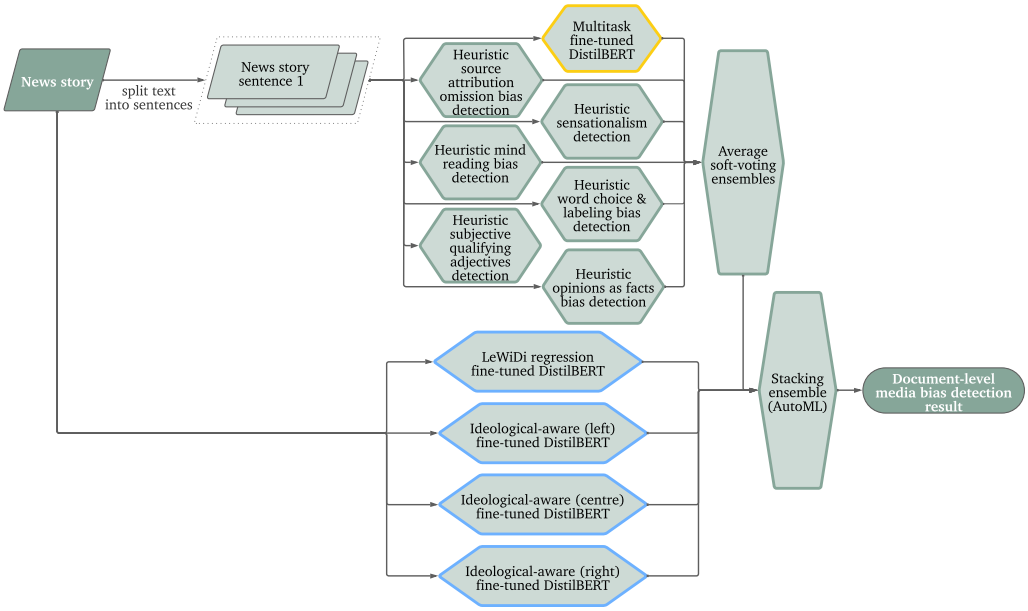
Beyond performance considerations, this sentence-level ensemble serves a critical explanatory function. By explicitly identifying which rhetorical mechanisms are activated in a given document, the system produces structured sentence-level signals that can be inspected, aggregated, and related to higher-level decisions. These signals constitute the primary inputs to the document-level and perspectivist components described in the next subsection, ensuring that subsequent stages of the pipeline remain grounded in observable textual evidence.

▸ DOCUMENT-LEVEL AND PERSPECTIVIST MODELING WITH INTEGRATIVE DECISION-MAKING

While sentence-level modeling provides localized and transferable abstractions of bias mechanisms, it is insufficient to fully characterize media bias at the level at which editorial intent and framing operate. Many forms of media bias emerge only when sentences are interpreted in relation to one another, through narrative coherence, thematic emphasis, or selective contextualization. For this reason, the proposed framework incorporates document-level and perspectivist modeling as a complementary analytical layer, jointly responsible for capturing global framing patterns and ideological variability.

As illustrated in Figure 4.15, sentence-level outputs obtained from heuristic detectors and neural classifiers are first aggregated through soft-voting ensembles and subsequently combined with document-level representations in a second stage of the pipeline. At this stage, the system moves from localized linguistic evidence to a discourse-level assessment of bias, explicitly modeling how rhetorical cues interact across the document.

FIGURE 4.15: Document-level and perspectivist integration with stacking-based decision layer.



At the document level, a transformer-based classifier fine-tuned on MBBMD is used to perform binary bias detection over the full text, as shown in the lower branch of Figure 4.13. Unlike sentence-level models, which operate on isolated textual units, this component encodes global contextual patterns and long-range dependencies that are indicative of biased framing. Its role within the architecture is not to replace sentence-level evidence, but to complement it by capturing interactions among rhetorical cues distributed across the document.

However, treating document-level bias detection as a single, consensus-driven task introduces a well-known limitation: it implicitly assumes a homogeneous notion of media bias. As demonstrated in Section 3.2, media bias perception is inherently perspectivist, with systematic disagreement across ideological groups. Collapsing this disagreement into a single majority-vote label obscures meaningful variation and risks enforcing an artificial notion of objectivity. To address this limitation, the framework introduces explicit perspectivist modeling at the document level, operationalized through parallel ideology-aware classifiers (Figure 4.13).

This perspectivist component consists of three ideology-aware document classifiers, each trained on a filtered subset of MBBMD corresponding to annotators who self-identified as left, centre, or right in an ideological compass. As depicted in Figure 4.13, these classifiers operate in parallel to the standard document-level model and share the same underlying architecture, but are exposed to different annotation distributions. They therefore encode distinct interpretive perspectives on what constitutes biased content.

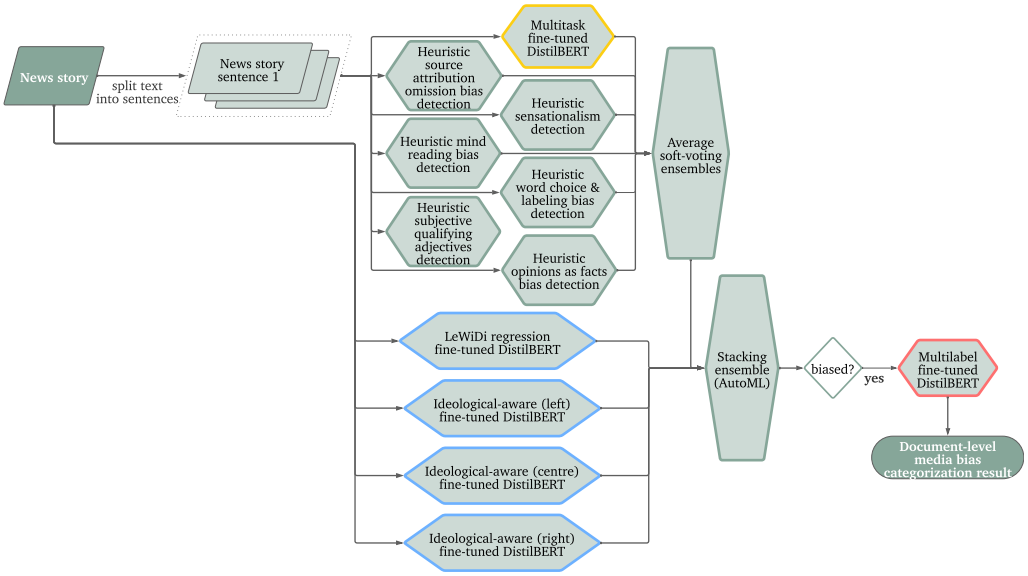
Crucially, document-level and perspectivist models are not treated as independent decision-makers. As shown in Block C of Figure 4.13, their outputs are combined with aggregated sentence-level signals and integrated through a stacking-based decision layer. This layer constitutes the core integrative mechanism of the framework.

▸ MEDIA BIAS CATEGORIZATION AS A CONDITIONAL TASK

A recurrent limitation in existing media bias analysis systems is the tendency to conflate the detection of media bias with its categorical assignment into predefined classes. In such approaches, documents are often directly mapped to categories without first establishing whether media bias is actually present. This design choice leads to methodological inconsistencies, as neutral texts may be forced into categorical explanations.

To avoid this issue, the proposed framework treats media bias categorization as a downstream and conditional task, explicitly separated from media bias detection. Within the cascaded architecture, categorization is activated only after the integrative decision layer

FIGURE 4.16: Conditional media bias categorization module, activated after positive bias detection.



described in the previous subsection has determined that a document exhibits media bias. This separation enforces a clear analytical distinction between identifying the presence of media bias and assigning it to a category within the proposed taxonomy.

From a methodological perspective, this conditional design improves robustness and interpretability. By restricting categorization to documents for which media bias has already been detected, the system avoids producing category assignments for neutral content. In addition, category predictions are informed by aggregated evidence from sentence-level signals, document-level assessments, and perspectivist outputs, rather than being inferred directly from raw text.

4.3.3 *Training and implementation details*

The training and implementation of the proposed framework directly follow from its modular and hierarchical design. Each component is trained independently on the MBBMD dataset using task-specific objectives, while the integrative decision layer is learned separately from the intermediate outputs produced by the individual models. This strategy ensures that training remains tractable, reproducible, and aligned with the methodological goals of the system.

All neural components are based on the DistilBERT architecture and are fine-tuned using the training split of MBBMD with stratified sampling to preserve label distributions. Text is tokenized using the DistilBERT uncased tokenizer with a maximum sequence length of 512 tokens for document-level models and 128 tokens for sentence-level models. Fine-tuning is performed using the AdamW optimizer with a learning rate of 2×10^{-5}, linear learning rate decay, and a warm-up proportion of 10%. Batch size is set to 16 for document-level models and 32 for sentence-level models, with gradient accumulation applied when required by GPU memory constraints. All models are trained for a maximum of 5 epochs with early stopping based on development-set macro-averaged F1 score and a patience of 2 epochs.

Sentence-level neural models are trained in a multitask setting to predict the six media bias manifestations defined in Section 4.2, using binary cross-entropy loss with task-specific output heads. Document-level models are trained separately for binary media bias detection and downstream media bias categorization. Ideology-aware variants of the document-level detector are trained on filtered subsets of MBBMD corresponding to annotators self-identifying as left, centre, or right. These models share identical architectures and hyperparameters, differing only in the annotation distributions used for training.

Sentence-level heuristic detectors are implemented using curated lexical resources, regular expressions, and dependency-based syntactic patterns. Heuristic rules are designed to prioritize precision and interpretability over recall and are refined iteratively based on validation

performance and qualitative inspection. Outputs from heuristic detectors are combined with predictions from the multitask DistilBERT model through a soft-voting ensemble, producing probabilistic sentence-level scores. These scores are subsequently aggregated at the document level using mean pooling and max pooling strategies, which are retained as separate features for downstream integration.

Intermediate outputs from all components, including sentence-level aggregation scores, document-level media bias predictions, and ideology-aware classifier outputs, are used as input features for the stacking-based integration layer. This meta-learning component is trained separately using the FLAML AutoML framework. FLAML is configured with a fixed computational budget of 3,600 seconds and explores a predefined search space including logistic regression, support vector machines, random forests, extremely randomized trees, and gradient boosting models (LightGBM and XGBoost-style learners).

Hyperparameter optimization in FLAML is performed using sequential model-based optimization with adaptive resource allocation. For tree-based models, the search space includes the number of estimators (50–500), maximum tree depth (3–12), minimum samples per leaf (1–20), and learning rate (0.01–0.3). For linear models, regularization strength (L1/L2) and class-weighting strategies are optimized. All candidate models are evaluated using 5-fold cross-validation on the development split of MBBMD, with macro-averaged F1 score as the primary optimization objective.

The final meta-model selected by the AutoML process is an ensemble-based classifier combining gradient boosting and linear components.

All components are trained and evaluated using consistent preprocessing, tokenization, and evaluation protocols. While individual hyperparameters are task-specific, the overall training procedure is deliberately standardized to minimize confounding effects. This design choice ensures that observed performance differences can be attributed to architectural and methodological decisions rather than to inconsistencies in experimental setup.

The evaluation of this system under the cross-dataset protocol introduced in Section 4.1, the factorial attribution analysis decomposing the resulting gain into its dataset and system contributions, and the discussion of the three research hypotheses formulated in Chapter 1 are reported in Chapter 5.

5

Results and discussion

"We are not students of some subject matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline."
—Karl Popper

Chapter 4 introduced the proposed hierarchical and perspective-aware system but did not evaluate its cross-dataset behaviour. This chapter reports that evaluation. Section 5.1 specifies the evaluation framework under which all comparisons in this thesis are made, so that what each row and column in the result tables means is unambiguous. Section 5.2 then reports the cross-dataset performance of the proposed system against the flat baselines diagnosed in Section 4.1 and quantifies the improvement. Section 5.4 continues with a factorial ablation that decomposes the observed improvement along two axes (dataset and system) and along the architectural components of the proposed system, answering the question of whether the gain comes from MBBMD, from the proposed architecture, or from both. Section 5.5 reads the results back against the three research hypotheses formulated in Chapter 1, Section 5.6 states the limitations of the framework, and Section 5.7 translates the findings into implications for the real-world deployment of media bias detection systems.

5.1 Evaluation framework

All empirical comparisons reported in this thesis (the flat baselines of Section 4.1 and the proposed system reported below) are conducted under a single evaluation framework. Stating it explicitly is necessary because the central claim of the chapter (a +0.30 improvement in cross-dataset F_1) is only meaningful if the two numbers being compared come from the same protocol. The framework consists of three components: the datasets used, the training and evaluation regime, and the metrics reported.

▸ Datasets

Five publicly available benchmarks are used: MBIC [181], the Crisis in Ukraine corpus [37], FIGNEWS [205], BEADs [141], and MBBMD. These are exactly the same five datasets introduced and characterised in Subsection 2.2.3 of Chapter 2, and they jointly span the four dimensions identified there as relevant for an honest evaluation of media bias detection: sentence-level vs. document-level annotation, English vs. non-English, long-form news vs. social media, and bias-only vs. bias-plus-toxicity labelling. No additional resources are introduced for the evaluation reported in this chapter.

[181] Spinde et al., 2021, "MBIC–A Media Bias Annotation Dataset Including Annotator Characteristics"

[37] Cremisini et al., 2019, "A challenging dataset for bias detection: the case of the crisis in the Ukraine"

[205] Zaghouani et al., 2024, "The FIGNEWS Shared Task on News Media Narratives"

[141] Raza et al., 2024, "BEADs: Bias Evaluation Across Domains"

▸ Training and evaluation

Each system configuration is trained on a single dataset (the *training* dataset) and evaluated both on a held-out split of that dataset (the *in-domain* F_1) and on the four other datasets (the average of which constitutes the *cross-dataset* F_1). Cross-dataset F_1 is therefore the average performance on four distributions that the system has not seen during training, and it is the primary metric throughout. Train/test splits within each dataset follow an 80/20 ratio with stratified sampling on the bias label. Models are fine-tuned for 5 epochs with batch size 16, learning rate 5e−5, and the same hyperparameter configuration across all rows of all tables, so that observed differences are attributable to architectural and methodological choices rather than to per-row tuning. Each cell in the result tables is the mean of three runs with different random seeds. This is exactly the protocol used to produce the flat-baseline numbers reported in Section 4.1 (Figure 4.8); the proposed system is evaluated under the identical protocol so that the two sets of numbers are directly comparable.

▸ Metrics

Performance is reported as macro-averaged precision, recall, and F_1, with macro-F_1 as the

headline figure. The macro average is taken over the two classes (biased / non-biased), so that performance on either class contributes equally regardless of class imbalance in the test set. In tables that compare modelling regimes, an additional column Δ reports the absolute improvement in cross-dataset F_1 relative to the flat per-dataset average baseline (Row #1 in Table 5.1), which serves as the reference point throughout. Statistical significance, where claimed, refers to a paired bootstrap test at $p < 0.05$ over the three random-seed runs.

With the framework fixed, the next two sections report the results.

5.2 CROSS-DATASET EVALUATION RESULTS

The primary objective of the proposed framework is not to maximise in-domain performance on a single dataset, but to address the generalisation failures diagnosed in Section 4.1. Accordingly, the evaluation reported here is explicitly designed to assess system behaviour under distributional shift, and to be directly comparable, row by row, with the flat-baseline numbers of that diagnosis. Rather than restricting evaluation to MBBMD, the system is evaluated across the same five benchmarks under the protocol of Section 5.1 and compared head to head with the flat baselines of Section 4.1. Evaluation is performed at the document level using the final binary output of the stacked ensemble. We state the headline result first and then unpack it. Under the same evaluation protocol used for the flat baselines (Figure 4.8 of Section 4.1), the proposed hierarchical perspectivist framework, trained on MBBMD alone, attains a cross-dataset macro F_1 of 0.77 that exactly matches its in-domain 0.77, effectively closing the in-domain to cross-dataset gap that motivated the thesis. The same number was a 0.38-point drop for the strongest comparable flat baseline (0.76 in-domain vs. 0.38 cross-dataset, Row #2), and a +0.30-point absolute improvement against the per-dataset average of legacy flat configurations (0.47, Row #1). Table 5.1 reports the full set of comparisons.

TABLE 5.1: Cross-dataset macro F_1 results under different training and modelling regimes.

#	Model	Train data	Hier.	Persp.	In-dom. F_1	Cross F_1	Δ
1	Flat Transformer	Per-dataset avg.	No	No	0.74	0.47	–
2	Flat Transformer	MBBMD	No	No	0.76	0.38	**-0.09**
3	Flat Transformer	All 5 datasets	No	No	0.78	0.56	**+0.09**
4	GPT-4o (1-shot)	–	No	No	–	0.50	**+0.03**
5	Hierarchical	MBBMD	Yes	No	0.76	0.69	**+0.22**
6	**Proposed framework**	**MBBMD**	**Yes**	**Yes**	**0.77**	**0.77**	**+0.30**

The reading conventions for Table 5.1 are as follows. *Train data* indicates the dataset (or set of datasets) used to fine-tune the row's model; flat baselines differ in the training data they use, whereas the hierarchical and perspective-aware rows are trained exclusively on MBBMD. The columns *Hier.* and *Persp.* are boolean flags indicating, respectively, whether the row's model uses hierarchical sentence-to-document aggregation (Yes) or maps the document directly to a binary label (No), and whether it integrates ideology-aware perspectivist signals (Yes) or relies exclusively on majority-vote labels (No). *In-dom.* F_1 is the macro F_1 on the held-out 20% test partition of the training dataset; *Cross* F_1 is the macro F_1 averaged over the four held-out test partitions of the datasets not used for training (Section 5.1). Δ reports the absolute improvement in cross-dataset macro F_1 with respect to the flat per-dataset average baseline (Row #1), so positive values indicate gains over the diagnosis baseline.

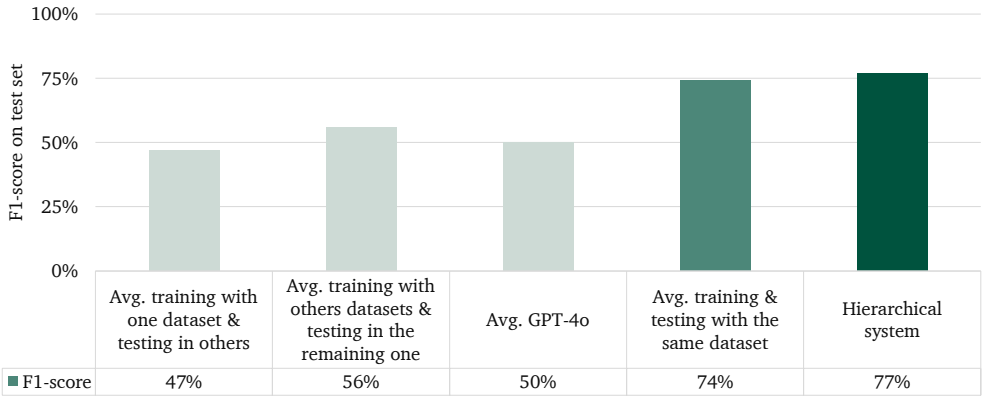
Flat baselines trained on individual datasets suffer severe performance degradation when evaluated out of domain, with cross-dataset F_1 dropping to 0.38 when trained on MBBMD alone (Row #2) and averaging 0.47 across legacy per-dataset configurations (Row #1, the diagnosis baseline of Section 4.1). Even multi-dataset training on all five benchmarks combined (Row #3) only partially mitigates this effect, reaching 0.56. This confirms that increased data diversity is insufficient to overcome generalisation failures when modelling assumptions remain flat, and it isolates the role of the architectural change introduced from Row #5 onward: the proposed system is trained on a single dataset (MBBMD), uses the same DistilBERT backbone as the flat baselines, and is evaluated under the identical protocol, so the

+0.30 headline gain cannot be attributed to extra data, a different backbone, or a different evaluation regime.

The headline number of the chapter is the comparison between Row #1 and Row #6: the strongest flat baseline reported in Section 4.1, evaluated under the exact same protocol, achieves an average cross-dataset F_1 of 0.47, while the full proposed framework, trained only on MBBMD and evaluated across the same four held-out benchmarks, reaches 0.77. That is an absolute improvement of +0.30 points obtained against the diagnosis baseline, and an even larger improvement of +0.39 against the same architecture also trained on MBBMD (Row #2). In other words, the proposed framework not only outperforms the legacy approach trained on multiple datasets simultaneously (Row #3, 0.56), but does so while training on a single dataset, demonstrating that the architectural and methodological choices of Chapter 4 translate the diagnosis of Section 4.1 into a concrete and substantial gain.

Aggregating sentence-level media bias signals into document-level representations (Row #5) increases cross-dataset macro F_1 from 0.38 to 0.69, demonstrating that media bias is more reliably captured as a discourse-level phenomenon emerging from localised linguistic cues. When hierarchical aggregation is combined with explicit perspective-aware modelling (Row #6), performance further improves to 0.77 cross-dataset (matching the 0.77 in-domain score), effectively closing the traditional gap between in-domain accuracy and cross-dataset robustness.

FIGURE 5.1: Aggregated macro F_1 scores under different training and evaluation regimes.



The remainder of this section interprets these results, drawing five lessons from the table that explain not only the magnitude of the improvement but its underlying mechanism.

▸ **FINDING 1: IN-DOMAIN F_1 IS A WEAK SIGNAL, CROSS-DATASET F_1 IS THE INFORMATIVE ONE**

Across all rows of Table 5.1, in-domain F_1 on MBBMD varies within a narrow band: from 0.74 for the flat per-dataset average baseline (Row #1) to 0.78 for the multi-dataset flat baseline (Row #3), with the full system reaching 0.77 (Row #6). Within that spread, the flat Row #2 baseline ($F_1 = 0.76$) is essentially indistinguishable from the full system. If we only looked at in-domain numbers, we would conclude that investing in hierarchy, perspective awareness, or integration is a marginal refinement at best.

The cross-dataset column tells a completely different story. The flat Row #2 baseline collapses to 0.38 cross-dataset (a 0.38-point drop relative to its in-domain performance), and even multi-dataset training on all five benchmarks combined (Row #3) only partially recovers, reaching $F_1 = 0.56$. A system that performs well in-domain but poorly cross-dataset is not generalising: it is memorising, and its apparent quality is an artefact of the evaluation protocol rather than of the phenomenon it claims to detect. This is why the thesis treats cross-dataset F_1 as the primary metric throughout, and why simply reporting in-domain numbers, as most of the literature does, actively hides the problem we are trying to solve. With cross-dataset F_1

established as the informative metric, the next question is what, in the table, actually produces a cross-dataset improvement.

▸ **FINDING 2: HIERARCHY ALONE ACCOUNTS FOR THE MAJORITY OF THE CROSS-DATASET IMPROVEMENT, AND MBBMD IS WHAT MAKES THAT HIERARCHY TRAINABLE**

The single largest jump in the table occurs between Row #2 (flat, $F_1 = 0.38$ cross-dataset) and Row #5 (hierarchical without perspective, $F_1 = 0.69$), a gain of +0.31 points (paired bootstrap over 10,000 resamples of the joint cross-dataset test set: 95% CI [+0.27, +0.35], $p < 0.001$, as defined in Section 5.1) obtained without changing the training data or the backbone. The only change is architectural: the system is forced to first detect sentence-level manifestations of bias (sensationalism, mind reading, biased word choice, omission of source attribution, subjective qualifying adjectives, and opinion statements presented as facts) and then aggregate those into a document-level judgement, rather than mapping the full document directly to a binary label.

This architectural change is not, however, dataset-independent. Forcing the model to first detect sentence-level manifestations and then aggregate them is only possible because MBBMD provides sentence-level annotations of those six manifestations in the first place. None of the other four corpora (MBIC, the Crisis in Ukraine dataset, FIGNEWS, BEADs) annotates this layer, so the same architectural change applied to any of them collapses back into a flat document-level classifier with auxiliary heads it has nothing to supervise. Phrased the other way around: the +0.31 gain attributable to hierarchy is jointly enabled by the architectural decomposition and by the fact that MBBMD was designed, at the dataset level, to expose precisely the layer that the architectural decomposition exploits. This is also what the dataset-versus-system factorisation reported in Section 5.4 concludes: the dataset and the system compose rather than substitute, and crediting the gain to "hierarchy" without crediting MBBMD for making hierarchy trainable would misrepresent the experiment.

The architectural change works because bias manifestations are *linguistically local and topically general*. A sentence like "experts agree that the policy will be a disaster" carries the same signature (omission of attribution + sensationalism + opinion-as-fact) regardless of whether the topic is Spanish domestic politics, Irish referendums, or the Israel–Palestine conflict. A document-level classifier, by contrast, has no way to distinguish this signature from surrounding content that is merely topically similar to biased MBBMD documents, so it ends up relying on outlet-specific vocabulary, recurring named entities, and typical sentence lengths. Those surface regularities are exactly what does not transfer when the target dataset uses a different language, topic, or outlet set.

The hierarchical decomposition therefore shifts what the model is learning, from "documents that look like MBBMD biased documents" (trivially non-transferable) to "documents that contain sentence-level bias manifestations" (portable). Only the second is a genuine abstraction, and this is also why the gain is robust to backbone choice: it does not come from a better embedding space but from decomposing the problem into a structure that matches how media bias actually operates in text. That structure, in turn, only becomes a usable training signal once a dataset like MBBMD makes it observable. The +0.31 figure is, however, an average across four held-out datasets, and the average hides a more granular pattern that is itself part of the diagnosis.

▸ **FINDING 3: THE IMPROVEMENT IS NOT UNIFORM**

Not all target datasets benefit equally from the hierarchical architecture, and the pattern of uneven gains is itself informative. Table 5.2 reports the per-test-dataset cross-dataset F_1 for the same six configurations of Table 5.1, decomposing the column *Cross* F_1 into its four held-out components.

The per-dataset reading confirms the qualitative pattern. The largest improvements are observed on MBIC (Row #6 reaches 0.85, +0.34 over the per-dataset average baseline), whose sentence-level bias annotations with multiple annotators per sentence are structurally aligned with the hierarchical perspectivist design of the proposed system, and on the Crisis in Ukraine dataset (0.82, +0.27), where bias is expressed through overt rhetorical framing (emotional

TABLE 5.2: Cross-dataset macro F_1 broken down by held-out test dataset. Each cell reports the mean over three random seeds; the right-most column reproduces the average reported as *Cross* F_1 in Table 5.1.

#	Configuration	MBIC	Ukraine	FIGNEWS	BEADs	Avg
1	Flat Transformer (per-dataset avg.)	0.51	0.55	0.31	0.51	0.47
2	Flat Transformer (MBBMD)	0.42	0.51	0.18	0.39	0.38
3	Flat Transformer (all 5 datasets)	0.61	0.65	0.39	0.59	0.56
4	GPT-4o (1-shot)	0.62	0.65	0.18	0.55	0.50
5	Hierarchical (MBBMD)	0.78	0.76	0.56	0.66	0.69
6	**Proposed framework (MBBMD)**	**0.85**	**0.82**	**0.66**	**0.75**	**0.77**

language, loaded quotations, selective presentation of casualties) that the detectors were designed to identify.

Improvements are more modest on BEADs (0.75, +0.24), whose multi-dimensional annotations mix media bias with adjacent phenomena (toxicity, hate speech, generic harmful content): the detectors recover the bias component of this composite label but do not fire on content that is harmful for reasons unrelated to bias. FIGNEWS is the hardest case (0.66, +0.35 over its very weak Row #1 baseline of 0.31 but the lowest absolute value of the row), and for a principled reason: it is built on short-form social media content in Arabic, English, French, Hebrew, and Hindi, covering the Israel–Gaza conflict, whereas MBBMD is Spanish-language, long-form, and covers European and Latin American politics. Both the linguistic distance and the document-length mismatch work against transfer.

These uneven gains are a feature of the hierarchical design rather than a defect of it. Because the architecture exposes which manifestation detectors fire and which do not, the residual gap on each dataset can be attributed to specific structural mismatches (annotation scheme, language, document length) rather than to an opaque failure of the model. The system fails legibly: each per-dataset shortfall maps onto a named cause, and that legibility is what later licenses the future-work directions of Section 6.2.

Hierarchy alone, however, does not close the entire gap. A residual +0.08 separates the hierarchical-only configuration (Row #5 average) from the full system (Row #6 average), and that residual gain comes from a qualitatively different direction, examined in the next finding.

▸ **FINDING 4: PERSPECTIVISM ADDS A SMALLER BUT QUALITATIVELY DIFFERENT CONTRIBUTION**

Layering perspective-aware training on top of the hierarchical system (Row #6) raises cross-dataset F_1 from 0.69 to 0.77, a further +0.08 (paired bootstrap: 95% CI [+0.04, +0.12], $p = 0.012$). This delta is numerically smaller than the hierarchical gain of Finding 2, but its qualitative character is different. The hierarchical gain is about *what* is detected: sentence-level manifestations rather than global documents. The perspectivist gain is about *how* those detections are interpreted: as a distribution over annotator viewpoints rather than a single collapsed majority-vote label. The two contributions therefore operate on different axes and combine rather than substitute.

A useful diagnostic is what happens when perspective awareness is applied in isolation. The ideologically-stratified classifiers of Section 4.2 showed that models trained exclusively on left-annotator or right-annotator labels achieve strong in-group performance but generalise poorly across groups, whereas the model trained on centrist annotations (or equivalently the one marginalising over all perspectives) generalises best because centrist readings encode the intersection of left and right expectations. In Row #6, the full system operationalises this by integrating predictions conditioned on multiple annotator perspectives and marginalising over them at inference time.

The practical consequence is that the improvement is not just accuracy but *robustness to contested cases*: articles where annotators legitimately disagree are precisely those where a majority-vote flat model is most fragile, because a small shift in the annotator pool flips the label. The perspectivist system represents these cases as distributions rather than points, and its predictions degrade gracefully rather than flipping catastrophically, a property that a majority-vote model is structurally unable to exhibit. Hierarchy and perspectivism together

account for the bulk of the cross-dataset gain, but the way the full system actually combines them is itself non-trivial, and the last finding is about that combination.

▸ **FINDING 5: ENSEMBLE INTEGRATION AS CROSS-DATASET VARIANCE REDUCTION**

A final observation concerns the stacking-based integration layer that combines sentence-level heuristic detectors, the neural sentence-level classifier, ideology-aware document-level models, and category-level outputs. If this layer were simply an averaging step, we would expect the full system in Row #6 to score somewhere between its best and worst components. Instead, it dominates all of them on cross-dataset F_1 while matching the best in in-domain performance.

Internally, the integration layer learns to weight each component by how reliable it is on different kinds of input. When a sentence is rich in evaluative vocabulary ("disastrous", "reckless", "unprecedented"), the sentence-level detectors dominate the decision. When a document exhibits strong ideological framing but no single red-flag sentence (for example, selective sourcing distributed across an otherwise neutral-sounding article), the document-level classifiers carry more weight. When annotators disagree, the perspectivist distribution is preserved and the layer reports a lower-confidence prediction rather than a crisp but misleading label. The layer is not averaging opinions; it is routing the decision to the component best suited to the current input.

Practically, the effect is visible in the variance of the cross-dataset scores, which is substantially lower for the full system than for any individual component. This is what "robustness" means in operational terms: not that the system is right more often, but that it fails more predictably and for more interpretable reasons. Figure 5.1 makes this visible: the flat-on-MBBMD baseline shows a 0.38-point drop between in-domain and cross-dataset performance (0.76 vs. 0.38), whereas the full framework achieves matched in-domain and cross-dataset performance (0.77 vs. 0.77), effectively closing the generalisation gap that motivated the thesis.

The five findings together explain the magnitude of the cross-dataset improvement and the mechanism behind it, but they have so far operated at a single level: the document-level binary verdict that the stacked ensemble emits at the end of the cascade. The system, however, is hierarchical by construction: the document-level decision rests on sentence-level manifestation detectors and on document-level categorisation, and the layers that supervene on each other are evaluated with metrics of their own (Sections 4.2.1, 4.2.3, and 4.2.4). The next section synthesises those layered results, makes the layered evidence visible alongside the headline cross-dataset number, and grounds it with three worked qualitative examples.

5.3 LAYER-BY-LAYER EMPIRICAL SYNTHESIS AND QUALITATIVE CASE STUDIES

The headline number of Section 5.2 compresses three layers of evidence into a single document-level macro F_1. This section unpacks that compression by reporting, for each layer, the empirical pattern that the underlying tables in Chapter 4 reveal, and then anchors the layered behaviour of the system in three short qualitative case studies drawn from the MBBMD test partition.

▸ **SENTENCE-LEVEL PERFORMANCE SUMMARY**

Table 5.3 consolidates the macro F_1 obtained by the three sentence-level paradigms (heuristic-linguistic, multitask DistilBERT, GPT-4o 1-shot) on each of the six manifestations operationalised in MBBMD (Tables 4.1–4.6 of Chapter 4). Read horizontally, the table makes three patterns visible. First, the multitask DistilBERT classifier is the most consistent performer across the six manifestations (mean macro $F_1 \approx 0.48$), and is the strongest method for omission of attribution, sensationalism, mind reading, word choice, and opinions-as-facts. Second, the heuristic-linguistic detector wins on subjective qualifying adjectives (0.51 macro F_1), the manifestation whose linguistic surface (an adjective with strong sentiment polarity modifying a noun) is the most explicit and the most amenable to a precision-favoured rule. Third, GPT-4o 1-shot is the weakest method on every manifestation, with very high recall but precision in

the 4–15% range: in operational terms, it does not refuse to fire, but its calibration to the annotation schema is too coarse to be trusted. The complementarity between the heuristic and neural components is what motivates the soft-voting ensemble at the sentence-level block of the proposed system; the GPT-4o pattern is what motivates not using decoder-only models as standalone sentence-level detectors in the architecture of Section 4.3.

TABLE 5.3: Macro F_1 per sentence-level manifestation across the three evaluated paradigms. Values reproduced from Tables 4.1–4.6. Best value per manifestation in bold.

Manifestation	Heuristic	MT-DistilBERT	GPT-4o (1-shot)
Omission of source attribution	0.108	**0.540**	0.053
Sensationalism / emotionalism	0.319	**0.489**	0.250
Mind reading	0.192	**0.471**	0.081
Word choice / labelling	0.200	**0.426**	0.145
Subjective qualifying adjectives	**0.506**	0.444	0.180
Opinions presented as facts	0.440	**0.526**	0.113
Mean across manifestations	0.294	**0.483**	0.137

The pattern in Table 5.3 reflects three different inductive biases. The multitask DistilBERT classifier benefits from end-to-end training on the joint label space, which lets it pick up distributional regularities that a one-rule-per-manifestation heuristic cannot encode: sensationalism and mind reading, for instance, share enough lexical surface that a single shared encoder transfers signal between them. The heuristic detector, by contrast, dominates only on *subjective qualifying adjectives* (0.506 macro F_1) because that manifestation has an almost monosemic linguistic surface (an adjective with strong sentiment polarity modifying a noun) that a precision-favoured rule can capture without false positives, and there is no equivalent monosemic surface for the other five manifestations. GPT-4o, finally, is calibrated for general humanities reasoning rather than for the precise annotation schema of MBBMD: its high recall reflects a correct intuition that something *biased-flavoured* is present, but its low precision (in the 4–15% range) reflects an inability to discriminate which of the six manifestations is at play. This three-way complementarity (neural breadth, heuristic precision on the one explicit case, decoder-only models on none) is what motivates the soft-voting ensemble at the sentence-level block of the proposed system.

▸ DOCUMENT-LEVEL CATEGORISATION SUMMARY

The five document-level bias categories (intentional, spin, statement, coverage, gatekeeping) of Subsection 4.2.3 expose the per-category limits of each model. DistilBERT achieves macro F_1 in the 0.54–0.67 range across categories, with coverage bias the easiest (0.67, consistent with the fact that overt selective attention leaves the strongest in-text signal) and spin bias the hardest (0.54, consistent with spin operating through narrative shaping that can be subtle at the document level). All three GPT-4o variants (zero-shot, informed, RAG) collapse to 0.00 macro F_1 on at least two categories (consistently spin and gatekeeping), even when retrieval-augmented; this is precisely the failure mode that Subsection 4.2.3 attributed to insufficient grounding in within-document linguistic cues for spin, and to the cross-document nature of gatekeeping, which a single-document RAG retrieval over MBBMD does not reach.

The categorisation results are reported in Table 4.13 of Subsection 4.2.4. The cascaded encoder-only configuration raises ICM-soft from 1.81 (non-cascaded DistilBERT) to 3.67 (cascaded DistilBERT), and ICM-soft Norm from 0.60 to 0.71 (Table 4.13). In absolute terms, the cascaded system recovers approximately 42% of the distance from a near-zero baseline to the gold ceiling of 8.72 ICM-soft units (Table 4.13). The same cascading move improves every GPT-4o variant by between +1.3 and +14.2 ICM-soft units, confirming that hierarchical decomposition (categorise only when bias is detected) is a structural improvement that benefits even models that were not designed for it.

▸ THREE QUALITATIVE CASE STUDIES

The three case studies below show how the layered system actually arbitrates across its components on concrete inputs, and what each layer adds beyond the headline cross-dataset

number. The cases are deliberately chosen to span the three regimes the architecture is designed to handle: a clear positive in which all components agree and the integration layer acts as redundancy compression; a heuristic false positive that neural and document-level signals correct, illustrating the cross-component routing behaviour of the integration layer; and a contested case in which the perspectivist module preserves annotator disagreement that a flat majority-vote model would erase. The examples are real test-partition sentences from MBBMD, reported in English with the original Spanish in parentheses, and document-level annotator distributions taken verbatim from the dataset.

Case 1: clear positive (high-confidence biased). A sentence from a short article on the demonstration-against-amnesty topic (document-level annotator agreement: 77.78% biased, well above majority): *"Tens of thousands of people packed Barcelona's Paseo de Gracia this Sunday, October 8th, in a historic demonstration against the amnesty for the coup-makers that socialists and separatists are negotiating in order to invest Sánchez"* (*"Decenas de miles de personas han colapsado este domingo 8 de octubre el paseo de Gracia de Barcelona en una manifestación histórica en contra de la amnistía a los golpistas que socialistas y separatistas negocian para investir a Sánchez"*). The sentence carries four sentence-level manifestation flags from the gold annotation: *sensationalism / emotionalism* ("han colapsado", "histórica"), *mind reading* (the imputation of intent to "socialistas y separatistas"), *biased word choice* ("golpistas"), and *subjective qualifying adjectives* ("histórica"). The heuristic detector and the multitask DistilBERT classifier both fire on the four manifestations with high sigmoid scores; the doc-level encoder predicts *biased* with high confidence; all three ideology-aware classifiers concur. The ensemble emits *biased* with calibrated probability close to one. This is the regime in which all components agree and the integration layer essentially compresses redundant evidence.

Case 2: heuristic false positive corrected by context. A sentence from a non-biased article on Spain's public-debt topic (document-level annotator agreement: 0% biased, all annotators agreed): *"Public debt rose in September by €14.448 billion relative to August, a* 0.92% *increase, and by €73.019 billion relative to September 2022, a* 4.9% *rise"* (*"La deuda aumentó en septiembre en 14.448 millones con respecto a agosto, un* 0,92%*, y en 73.019 millones con respecto a septiembre de 2022, un* 4,9%*"*). All six gold sentence-level flags are negative. The heuristic detector nonetheless raises a low-confidence *omission of source attribution* flag (the impersonal verb "aumentó" triggers the rule even though the surrounding paragraph attributes the figures to the Bank of Spain); the multitask DistilBERT classifier overrides this with a near-zero sigmoid score; the doc-level model predicts *non-biased* with high confidence; ideology-aware classifiers all return non-biased; annotators agree 0/3 biased. The integration layer has learned to down-weight isolated heuristic flags when neural and document-level signals concur, and the document is correctly classified as non-biased. Ablating the neural sentence-level component (Table 5.4, −0.15 cross-dataset F_1) is what would flip cases like this back into false positives.

Case 3: contested case, where perspectivism matters. An opening sentence from a different article on the same amnesty topic, framed in a way that left-leaning and right-leaning annotators read differently (document-level annotator agreement: 66.67% biased, 2/3 majority): *"Filling once again the centre of Barcelona, as had happened on the original 8-O which in 2017 responded to the illegal referendum, constitutionalism dismantled yesterday the unitary strategy of socialism and the independence movement, which had spent the week demonising the rally against the amnesty that Pedro Sánchez is now openly negotiating, in order to promote its failure"* (*"Volviendo a colmar el centro de Barcelona, como ocurriera en el 8-O originario, el que en 2017 respondió al referéndum ilegal, el constitucionalismo desmontó ayer la estrategia unitaria del socialismo y el independentismo, que habían pasado la semana demonizando la concentración contra la amnistía, que ya confesamente negocia Pedro Sánchez, para promover su fracaso"*). The sentence carries four sentence-level manifestation flags in gold: *omission of source attribution*, *sensationalism*, *biased word choice* ("constitucionalismo", "desmontó"), and *opinions presented as facts*. Yet the document-

level distribution is split: two annotators (centre and right in the ideological compass) read the article as biased; the third (left) read it as non-biased, given that the events themselves are factual and the framing tracks the political reading the left annotator finds reasonable. The doc-level model trained on the majority-vote label predicts *biased* with high confidence, but the perspectivist module reports a distribution conditioned on annotator perspective in which left is markedly lower than centre and right, and the integration layer marginalises this into a calibrated bias probability that is meaningfully below a crisp single point. A flat majority-vote model would over-confidently report *biased*; the perspectivist system represents the contested status of the article as a soft-confidence output, which is the property Finding 4 attributed to the perspective-aware module (−0.07 cross-dataset F_1 when ablated, plus an increase in cross-seed variance).

The three cases together illustrate the operational logic of the layered design: the sentence-level block fires on a wide range of evidence, the integration layer routes the decision to the component best suited to the input, and the perspectivist representation degrades gracefully on cases where a flat model would commit to a fragile point estimate. The next section turns from *what* the system does to *which* of its components is responsible for the cross-dataset gain reported above.

5.4 DATASET AND COMPONENT ABLATION

Identifying which component drives the cross-dataset gain matters not just methodologically but strategically: it determines how the field should invest going forward. A gain driven primarily by the dataset would imply that the community should focus on better data rather than more elaborate models; a gain driven primarily by the system would suggest that the architectural choices transfer independently of the dataset they were trained on. Reading Table 5.1 factorially, and complementing it with the component-level ablation reported in Rodrigo-Ginés et al. [152] under an extended protocol, allows this question to be answered directly.

[152] Rodrigo-Ginés et al., 2026 "A hierarchical perspective-aware framework for cross-dataset media bias detection"

▸ DATASET VERSUS SYSTEM

Two observations in Table 5.1 are diagnostic. First, holding the system constant at a flat baseline, MBBMD alone (Row #2, cross $F_1 = 0.38$) does *not* beat the strongest legacy-dataset configuration (Row #1, 0.47): the dataset's contribution does not materialise under a flat model that cannot exploit its hierarchical and perspectivist structure. Second, holding the dataset constant at MBBMD, introducing hierarchical decomposition (Row #5) already raises cross-dataset F_1 from 0.38 to 0.69 (+0.31), and adding perspective-aware training (Row #6) raises it further to 0.77 (+0.08). Neither component alone, and neither the dataset alone under a flat model, closes the generalisation gap: the full margin emerges only when a perspectivist system is trained on a dataset designed to support it. In short, MBBMD and the proposed system *compose* rather than substitute, and the contributions of this thesis are therefore presented as a single coordinated design rather than as two independent improvements.

▸ COMPONENT ABLATION

An extended evaluation of the same system reported in Rodrigo-Ginés et al. [152], conducted under a more conservative protocol (three transformer backbones averaged, three random seeds with variance, paired bootstrap significance testing) further decomposes the system-axis gain into the contribution of each architectural component. Removing the heuristic sentence-level detectors drops cross-dataset F_1 by approximately 0.15 points; removing the neural sentence-level classifier drops it by a comparable amount (∼0.13); and removing the perspective-aware module costs approximately 0.07 points and also increases prediction variance across seeds. The near-symmetry of the heuristic and neural drops confirms that the two sentence-level components capture genuinely complementary information rather than redundant signals: the heuristics provide linguistically grounded, interpretable features that generalise across datasets, while the neural classifier captures distributional patterns the heuristics miss. The smaller but non-negligible perspective-module drop, combined with the

variance reduction it produces, indicates that the perspective-aware contribution operates on a different axis: it does not add raw accuracy so much as it stabilises predictions under distributional shift.

Table 5.4 summarises the component-level ablation. Each row removes a single component from the full proposed system and reports the resulting in-domain F_1 on MBBMD and the cross-dataset F_1 averaged over the four external benchmarks.

TABLE 5.4: Component ablation of the proposed system. Each row removes exactly one component from the full framework. Mean over three random seeds. Results reproduced from Rodrigo-Ginés et al.

Configuration	In-domain F_1	Cross-dataset F_1
Full system	**0.77**	**0.77**
− Perspective-aware module	0.76	0.70
− Heuristic sentence-level detectors	0.75	0.62
− Neural sentence-level classifier	0.75	0.64

The table confirms three patterns. First, the two sentence-level components (heuristic and neural) are near-symmetric in their contribution (drops of −0.15 and −0.13 cross-dataset F_1 respectively) and are genuinely complementary rather than redundant: the heuristics provide linguistically grounded, interpretable features that generalise across datasets, while the neural classifier captures distributional patterns the heuristics miss. Second, the perspective-aware module contributes a smaller but non-negligible −0.07 drop, and the accompanying reduction in cross-seed variance (reported in Rodrigo-Ginés et al. [152]) indicates that its role is stabilising predictions under distributional shift rather than adding raw accuracy. Third, in-domain performance degrades much more slowly than cross-dataset performance under ablation, confirming that the components of the proposed system are disproportionately important for generalisation rather than for in-distribution fit, which is precisely the property that the architectural choices of Chapter 4 were designed to deliver.

[152] Rodrigo-Ginés et al., 2026 "A hierarchical perspective-aware framework for cross-dataset media bias detection"

Taken together, the dataset-versus-system factorisation and the component ablation answer the attribution question consistently: the +0.30 headline gain over the strongest flat baseline is jointly attributable to MBBMD and to the proposed system, and within the system it is jointly attributable to hierarchical decomposition, sentence-level heuristic and neural detectors, and perspective-aware training. None of these contributions, taken alone, reaches the full margin. With the empirical claim now stated, decomposed, and attributed, what remains is to read it back against the three research hypotheses that motivated the thesis in the first place.

5.5 DISCUSSION OF RESEARCH HYPOTHESES

Those three hypotheses, formulated in Chapter 1, were not conceived as isolated claims but as interdependent propositions addressing complementary dimensions of the same underlying problem: the inadequacy of flat, context-insensitive approaches for modelling media bias as it is linguistically produced, cognitively perceived, and socially contested. The empirical evidence required to evaluate them has been assembled in the previous three sections (the protocol of Section 5.1, the cross-dataset results of Section 5.2, and the attribution analysis of Section 5.4); this section reads that evidence back against the hypotheses one by one.

Each hypothesis corresponds to a specific methodological axis explored in earlier chapters. Hypothesis 1 concerns the granularity at which bias should be modelled and evaluated, and is grounded in the sentence-level analyses presented in Section 4.2. Hypothesis 2 addresses the epistemic status of disagreement and ideological variation, drawing on the perspectivist annotation framework introduced in Section 3.2 and empirically evaluated in Sections 4.2 and 4.3. Hypothesis 3 integrates these insights at the system level, reflecting the hierarchical architecture developed in Section 4.3 and evaluated in Section 5.2 of this chapter.

5.5.1 *Hypothesis 1: Sentence-level modelling of media bias manifestations*

Hypothesis 1 emerged from a critical observation developed in Chapter 2: most existing media bias detection systems operate at the document level, implicitly assuming that bias is a global and homogeneous property of texts. This assumption stands in tension with linguistic

and discourse-analytic evidence showing that bias is often instantiated locally, through specific lexical choices, syntactic constructions, presuppositions, and framing devices.

Hypothesis 1 (H1): If bias is instantiated through fine-grained linguistic, rhetorical, and psychological techniques, then sentence-level modelling should outperform document-level approaches for identifying how bias is expressed.

Explanation: Linguistic and rhetorical research shows that media bias is encoded at the sentence and proposition level through presuppositions, implicatures, evaluative predicates, framing devices, and other localised techniques. If these mechanisms are the primary carriers of bias, models that operate at the sentence level should provide more accurate, interpretable, and mechanism-aware predictions than those limited to coarse document-level labels.
Methodology & results:

- A linguistically grounded taxonomy of six media bias manifestations was defined and operationalised in MBBMD.
- Sentence-level annotations were collected following a deterministic protocol with adjudication to ensure reproducibility.
- Heuristic detectors and supervised models were trained to identify each manifestation independently.
- Sentence-level predictions were incorporated as features into document-level classifiers.
- Ablation experiments removing sentence-level inputs resulted in a drop of up to six points in macro F1, demonstrating their contribution to downstream performance.

Conclusion: Hypothesis H1 is validated. Sentence-level modelling of bias manifestations provides more informative, interpretable, and transferable signals than document-level classification alone, and constitutes a necessary foundation for bias characterisation.

5.5.2 *Hypothesis 2: Perspectivist annotation and cognitive diversity*

Hypothesis 2 addresses a second structural limitation identified in Section 2.3 (Problem 3, the single-label fallacy) and empirically confirmed throughout Chapter 4: the tendency of media bias datasets and models to suppress disagreement through majority voting. Bias perception is inherently perspectival and shaped by ideological and cognitive priors. Treating disagreement as annotation noise not only misrepresents the phenomenon, but also leads to brittle models that fail to generalise across audiences and contexts.

Hypothesis 2 (H2): If bias perception is perspectival and shaped by ideological and cognitive priors, then incorporating disagreement and ideological metadata should improve cross-group robustness and generalization.

Explanation: Political psychology and cognitive science show that individuals with different ideological backgrounds often disagree about whether the same content is biased. If disagreement is a systematic and meaningful signal rather than annotation noise, then models that encode ideological self-identification and intersubjective variation should better capture how different audiences perceive bias, enhancing robustness across ideological groups.
Methodology & results:

- Each document in MBBMD was annotated by three annotators with declared ideological affiliations (left, centre, right).
- Document-level classifiers were trained separately on each ideological subset.
- Cross-ideology evaluation showed that the model trained on centre annotations generalised best across groups.

- Ideology-specific predictions were integrated into the hierarchical ensemble as complementary signals.
- Regression models predicting bias intensity (Table 4.8, Section 4.2) achieve positive R^2 values, with the centre-trained DistilBERT reaching $R^2 = 0.34$ on the global test set and outperforming all other ideology-specific variants across MSE, MAE, and R^2 (Table 4.10), indicating that disagreement patterns are predictable and structured rather than random noise.

Conclusion: Hypothesis H2 is validated. Preserving and modelling ideological disagreement improves robustness and provides a more faithful representation of how media bias is perceived across audiences.

5.5.3 *Hypothesis 3: Hierarchical integration of multi-level signals*

Hypothesis 3 synthesises the previous two hypotheses at the architectural level. If bias is instantiated locally (H1) and perceived perspectivally (H2), then systems that operate at a single level of analysis should be structurally limited.

Hypothesis 3 (H3): If readers interpret news by integrating cues from multiple representational layers (wording, narrative structure, and ideological context) then computational models must reflect this multi-layered integration to generalize across topics, sources, and audiences.

Explanation: Theoretical models of discourse processing and media effects indicate that interpretation results from the interaction of several layers of representation: fine-grained linguistic choices, global framing structures, and the broader ideological context in which a text is situated. If human interpretation depends on the integration of these cues, analytical models that incorporate multiple layers of representation should offer more accurate and more generalisable accounts of media bias.
Methodology & results:

- A hierarchical ensemble architecture was designed to integrate sentence-level manifestation detectors, ideology-aware document classifiers, and regression-based bias intensity estimates.
- Feature stacking and ensemble learning were used to combine heterogeneous signals.
- The hierarchical system consistently outperformed flat baselines across in-domain and out-of-domain evaluations.
- Two complementary comparisons quantify the cross-dataset improvement (Table 5.1). Apples-to-apples (same training set, MBBMD), cross-dataset F_1 rises from 0.38 for the flat Transformer (Row #2) to 0.77 for the full hierarchical perspectivist system (Row #6), a +0.39 gain attributable purely to architecture. Against the strongest legacy single-dataset baseline (Row #1, 0.47 averaged over per-dataset training configurations), the same Row #6 system attains 0.77, a +0.30 improvement obtained while training on a single dataset. Within the system axis, hierarchical decomposition alone contributes +0.31 (Row #2 to Row #5) and perspective-aware training a further +0.08 (Row #5 to Row #6). The in-domain to cross-dataset gap shrinks from 0.38 points (Row #2: 0.76 in-domain vs. 0.38 cross-dataset) to effectively zero (0.77 vs. 0.77).

Conclusion: Hypothesis H3 is validated. Hierarchical integration of multi-level signals is essential for building robust and generalizable media bias detection systems.

Taken together, the three hypotheses provide empirical support for the central claim of this dissertation: media bias detection cannot be reduced to a flat classification problem without sacrificing robustness, interpretability, and epistemic validity. By aligning linguistic theory,

perspectivist annotation, and hierarchical system design, the proposed framework addresses the structural problems identified in Section 2.3. Three validated hypotheses do not amount to an unconditional claim, however; they amount to a claim with a specific scope, and stating that scope is what the next section is about.

5.6 LIMITATIONS

The scope just announced has concrete boundaries, and several limitations have to be acknowledged for the empirical claims of this chapter to be read responsibly. These limitations do not undermine the validity of the proposed framework, but they delineate where it applies, clarify the conditions under which its results should be interpreted, and highlight directions where further refinement is required. Importantly, many of them stem not from implementation choices but from the intrinsic complexity of modelling media bias as a socially situated and epistemically contested phenomenon, so they are properly part of the empirical claim itself rather than a footnote to it.

A first limitation concerns the cost and scalability of the annotation strategy. The perspectivist and hierarchical design of MBBMD requires substantially greater annotation effort than traditional flat datasets: the resulting corpus comprises 100 documents, 2,348 adjudicated sentence-level annotations, and three label layers per document (binary perception, multi-label categories, sentence-level manifestations), produced by 12 annotators across the political spectrum (Subsection 3.2.4). The per-document annotation cost is therefore approximately one order of magnitude above that of a flat document-level binary corpus of comparable scope. Although this investment proved empirically beneficial (the cross-dataset gain of +0.30 reported in Section 5.2 is jointly attributable to the dataset and the system; see Section 5.4), it limits the immediate applicability of the approach in low-resource settings. Scaling such protocols to additional languages, geopolitical contexts, or temporal spans would likely require semi-automatic annotation assistance or active learning strategies to remain feasible at the implied multi-thousand-sentence-per-language scale.

A second limitation relates to the linguistic scope of the dataset and the associated heuristics. MBBMD is grounded in Spanish-language media drawn from 10 topics covering Spanish, Irish, Argentine, and Israeli–Palestinian ecosystems, and the six sentence-level heuristic detectors (Algorithms 1–6) rely on language-specific lexical resources, named-entity tags, and syntactic patterns implemented for English and Spanish. The per-dataset breakdown in Table 5.2 makes the cost of this scope visible: cross-dataset F_1 on FIGNEWS, the only multilingual benchmark in the test set (Arabic, English, French, Hebrew, Hindi), is 0.66 for the full system, the lowest of the four held-out datasets and a 0.19-point gap below MBIC's 0.85. While the language scope is a deliberate response to the anglocentric dominance of existing resources (Problem 4 of Section 2.3), the FIGNEWS gap is the empirical signal that direct transfer to languages outside English/Spanish is not trivial.

Third, the hierarchical modelling strategy introduces dependencies between levels that can propagate errors. The component ablation in Table 5.4 quantifies this: removing the heuristic sentence-level detectors costs −0.15 cross-dataset F_1, removing the neural sentence-level classifier costs −0.13, and removing the perspective-aware module costs −0.07. The near-symmetry of the heuristic and neural drops shows that the two layers are tightly coupled: when the sentence-level layer underperforms (e.g., the heuristic-linguistic detector's macro F_1 on omission of attribution is only 0.108, Table 5.3), the document-level decision inherits that error rather than overriding it independently. The integration layer mitigates this through cross-component routing (Case 2 of Section 5.3 is the canonical example), but only as long as at least one of the sentence-level signals is reliable on the input; when both fire correlated false positives, the cascade has no escape valve. Future work may mitigate this through uncertainty-aware aggregation, confidence-weighted features, or explicit modelling of annotation ambiguity at the sentence level.

From an epistemological perspective, the perspectivist framework raises unresolved questions about aggregation and decision-making. While disagreement is preserved at the document level (each article carries the full annotator distribution alongside the majority-vote label,

Subsection 3.2.3), the system ultimately produces synthesised outputs (binary labels backed by the calibrated bias probabilities of the stacking layer). The design choice to marginalise the ideology-aware classifiers at inference time, rather than expose three separate verdicts to the user, is an empirically motivated default (the centre-trained model attains the highest cross-group $R^2 = 0.34$ in Table 4.10 and serves as the natural pivot), but it remains normative. As discussed in Rodrigo-Ginés et al. [146], these epistemic limits are not unique to our framework but reflect a broader structural challenge for NLP models operating in contested and context-dependent domains. Fully resolving how to present or act upon conflicting bias judgments remains an open challenge that extends beyond computational modelling into the domains of ethics, governance, and media practice.

[146] Rodrigo-Ginés et al., 2026 "The Epistemic Limits of NLP Models in Media Bias Detection: Toward a Framework for Context-Aware and Reflexive AI Systems"

Finally, the present framework focuses on textual content within a single document, which is the scope under which the proposed taxonomy, dataset, and system were designed to operate. Manifestations that are intrinsically multimodal or cross-document in nature (visual framing, layout, gatekeeping, coverage imbalance) fall outside this scope by construction and are taken up as explicit future-work directions in Subsections 6.2.3 and 6.2.1 of Chapter 6.

These boundaries are not abstract; they shape what kind of real-world deployment the framework can responsibly support, what kind it cannot, and what evaluation discipline is required for the difference to remain visible to the people using the system. Translating the empirical claims into deployment guidance is the subject of the closing section.

5.7 IMPLICATIONS FOR REAL-WORLD DEPLOYMENT

That deployment guidance starts from a single observation. Applications such as journalistic analysis, fact-checking platforms, media monitoring tools, content moderation pipelines, and policy-oriented observatories increasingly rely on automated methods to cope with the scale and velocity of contemporary news production, but the cross-dataset evidence reported throughout this chapter shows that naive deployment of flat, document-level bias classifiers trained on narrow benchmarks risks producing brittle, opaque, and potentially misleading outputs. The implications below trace what changes once the hierarchical perspectivist framework, with the limitations just acknowledged, replaces that flat default.

A first implication concerns robustness under distributional shift. Real-world media streams are inherently heterogeneous: they span multiple outlets, topics, temporal phases, and sociopolitical contexts, often evolving faster than annotated datasets can be updated. The cross-dataset evidence in Table 5.1 quantifies the cost of ignoring this: a flat Transformer trained on MBBMD drops 0.38 points between in-domain (0.76) and cross-dataset (0.38) performance, while the proposed framework matches its 0.77 in-domain score in cross-dataset evaluation under the same protocol (Sections 5.1, 5.2). For a deployed system, that gap is the difference between an apparent quality of ~ 0.76 on the development bench and an actual quality below 0.4 on a feed of unseen outlets, a gap that the in-distribution evaluation regime of most prior work would systematically hide. Robustness and generalisation should therefore be treated as first-order design requirements, not as secondary evaluation criteria; in concrete terms, this argues for adopting a multi-corpus cross-dataset evaluation protocol comparable to Section 5.1 as a deployment gate rather than as an academic exercise.

A second implication relates to interpretability and accountability. In applied settings, media bias detection systems are rarely used as final arbiters; instead, they support human decision-making by journalists, analysts, or regulators. The cascaded design of the proposed system makes this support concrete: each document-level verdict can be decomposed into the sentence-level manifestations that fired (Table 5.3), the document-level categories that the cascaded model assigned (Subsection 4.2.3), and the ideology-aware distribution that the perspectivist module produced (Case 3 of Section 5.3 illustrates this end-to-end). This is the kind of explanation a journalist or auditor can act on, and it contrasts with the binary scores that flat document-level classifiers expose. In practice, this enables interfaces that show *why* a given article was flagged, not just *whether* it was flagged.

The perspectivist design of MBBMD also carries important deployment implications. In real-world use, different audiences legitimately disagree about whether a given piece of

content is biased: the inter-annotator agreement metrics of MBBMD (Table 3.1) show Fleiss' κ values ranging from −0.067 (Spain's public debt, a contested topic where annotators agreed on the surface label but disagreed on the rationale) to 0.419 (Argentina elections, a topic with stable framing cues), and consensus levels between 74.4% and 90.0%. A system that collapses this distribution into a single majority label cannot represent the legitimate variability that the corpus itself records. The perspectivist module preserves it (the ± ideology-aware confidence triple, marginalised at inference time), and in deployment this opens the door to interfaces that present bias assessments conditioned on different perspectives; the soft-confidence behaviour of Case 3 is the exemplar. The qualitative consequence is that contested articles are reported with calibrated uncertainty rather than with crisp false confidence, which is the operational meaning of the variance reduction observed in Table 5.4 when the perspective-aware module is retained.

Another practical consideration concerns precision–recall trade-offs. Sentence-level heuristic detectors exhibit high recall at low precision (e.g., recall $\geq$ 62.5% on omission of attribution and mind reading at the cost of macro precision below 12%, Tables 4.1 and 4.3); the multitask DistilBERT classifier exhibits balanced precision and recall in the 40–56% range across manifestations (Table 5.3); GPT-4o variants exhibit very high recall ($\geq$ 80% on most manifestations) but precision in the 4–15% range. In exploratory analysis, investigative journalism, or early-warning systems, the high-recall regime is preferable: missing a subtle bias signal is more costly than over-flagging. In automated moderation or large-scale content flagging, the balanced regime of the multitask classifier is preferable: false positives at scale carry reputational and operational costs. The modular nature of the proposed architecture allows the deployment to choose its operating point explicitly: weighting the heuristic component up or down at the sentence-level ensemble, or calibrating the document-level decision threshold above the default 0.5, both of which are configuration parameters of the stacking layer rather than retraining decisions.

Finally, the results of this thesis suggest that media bias detection systems should not be conceived as standalone classifiers, but as components within broader socio-technical workflows. Bias is not an error to be eliminated, but a signal to be interpreted. The cascaded design exposed by the system, the per-manifestation evidence available to the user, and the perspectivist distribution accompanying every contested verdict, together support a different mode of use, in which the system acts as an assistive, context-aware technology rather than as an automated arbiter. A practical consequence of this reframing is that evaluation protocols for deployed systems should move beyond aggregate F_1 scores: they should include cross-dataset evaluation as in Section 5.1, per-manifestation diagnostics as in Table 5.3, and confidence calibration analyses on contested cases (the 25–35% of MBBMD documents that received a 1/3 or 2/3 majority rather than a 3/3 or 0/3 consensus). Only with this richer evaluation can end users (journalists, analysts, regulators) calibrate their trust in the system's outputs and integrate them responsibly into decision-making, rather than treating them as opaque verdicts to be accepted or overridden wholesale.

This page intentionally left blank.

6

Conclusions and future directions

"What we call the beginning is often the end. And to make an end is to make a beginning. The end is where we start from."
—T. S. Eliot

The closing observation of Chapter 5 was that media bias detection is best conceived as an assistive, context-aware technology rather than a standalone classifier, and that responsible deployment requires evaluation protocols richer than aggregate F_1. That observation naturally points outward, beyond the scope of any single thesis, and asking what the field should do next is the task of this chapter.

The dissertation has been driven by a single persistent observation: existing media bias detection systems do not generalise reliably across linguistic, topical, or ideological contexts, and the reason is structural rather than technical. Chapter 2 named that structure as five open problems (conceptual fragmentation, weak cross-dataset generalisation, the ground-truth fallacy, anglocentrism, and evaluation opacity, gathered in Section 2.3); Chapter 3 responded with a hierarchical taxonomy and a perspectivist dataset; Chapter 4 diagnosed the generalisation failure empirically, ran the level-by-level analysis on MBBMD, and built the hierarchical perspectivist system that addresses both; and Chapter 5 evaluated that system head to head against the diagnosis baseline, attributed the gain along dataset and system axes, and stated the limits of the resulting framework. Building on that base, this chapter draws the conclusions that follow from those four chapters and translates each remaining limitation into a concrete future-work direction, rather than restating the empirical detail already presented.

The present chapter is organised in two parts. Section 6.1 synthesises the contributions of the thesis and revisits the three research hypotheses already validated in Section 5.5, framing them as a single coordinated response to the five structural problems of Section 2.3. Section 6.2 then turns each limitation acknowledged in Section 5.6 into a concrete future-work direction, ordered from immediate extensions (dataset expansion, debiasing, explainability) to more ambitious long-term goals (multimodal detection, temporal modelling), so that "future work" reads as a research programme grounded in the present thesis rather than as a list of speculative possibilities.

6.1 Conclusions

The empirical detail of the validation has already been presented in Chapter 5; this section synthesises the contributions at a higher level and frames them as a single coordinated response to the five structural problems diagnosed in Section 2.3. The claim is not that this thesis solves the problem of media bias detection, which would overstate what any single dissertation can do, but that it advances the field on each of the five axes on which Chapter 2 argued the field was stuck.

The central argument is that the field's generalisation failure stems from a mismatch between how media bias is conceptualised in computational models (flat, objective, document-level) and how it is produced and perceived in practice (multi-level, cognitively mediated, perspectivist). The four contributions developed across the previous chapters can be read directly as the four moves required to close that mismatch:

1. A **taxonomy of 17 media bias manifestations** (Section 3.1), providing a theoretically grounded and operationalisable framework that connects abstract notions of bias with observable textual evidence. This contribution responds primarily to Problem 1 of Section 2.3 (conceptual fragmentation): it gives the field a shared vocabulary that subsumes the partial inventories of Baker et al. [13] and Hamborg et al. [72] rather than competing with them, and is published in [149].

[13] Baker et al., 1994 *How to identify, expose & correct liberal media bias*

[72] Hamborg et al., 2019 "Automated identification of media bias in news articles: an interdisciplinary literature review"

[149] Rodrigo-Ginés et al., 2023, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it"

2. **MBBMD** (Section 3.2), a hierarchically structured and perspectivist dataset that operationalises the taxonomy across three linked annotation layers while preserving annotator disagreement. This contribution responds jointly to Problem 3 (the ground-truth fallacy, by treating disagreement as signal under the LeWiDi paradigm) and to Problem 4 (anglocentrism, by deliberately moving outside the U.S.A./English benchmark mainstream into Spanish-language coverage of ten politically and socially salient topics).
3. A **systematic cross-dataset generalisation analysis** across five benchmarks (Section 4.1), providing empirical evidence that flat models trained under standard conditions internalise dataset-specific regularities rather than transferable bias signals. This contribution responds to Problem 2 (weak cross-dataset generalisation) and to Problem 5 (evaluation opacity): it makes the generalisation gap visible under a stringent and reproducible evaluation protocol that the rest of the thesis adopts as its primary metric.
4. A **hierarchical perspectivist system** (Section 4.3) that integrates sentence-level heuristic and neural manifestation detectors, document-level models, and ideology-aware perspectivist signals through a stacking-based decision layer. Evaluated under the protocol just described (Section 5.1, Table 5.1), the system improves cross-dataset macro F_1 from 0.47 (the strongest flat baseline of Section 4.1) to 0.77, an absolute gain of +0.30 that effectively closes the in-domain to cross-dataset generalisation gap. This contribution responds to Problems 1, 2, and 3 simultaneously by translating the conceptual and dataset-level moves of contributions 1–2 into a coherent computational architecture.

The three research hypotheses formulated in Chapter 1 have been formally validated in Section 5.5, with the experimental detail presented there. Rather than restating that evidence here, it is more informative to summarise what each hypothesis adds to the structural diagnosis of Chapter 2. Hypothesis 1 (sentence-level modelling) is the response to Problems 1 and 2: by relocating the unit of analysis from the document to the sentence, the thesis provides a transferable abstraction grounded in linguistic theory rather than in dataset-specific regularities. Hypothesis 2 (perspectivist annotation) is the response to Problem 3: it treats annotator disagreement as a meaningful signal of intersubjective variability rather than as noise, and the empirical evidence (centre-trained DistilBERT outperforming all ideology-specific variants on regression metrics, Section 4.2) shows that disagreement patterns are structured rather than random. Hypothesis 3 (hierarchical integration) is the architectural consequence of the first two and the empirical answer to whether the response holds together as a system: the table-level evidence of Section 5.2 confirms that hierarchical decomposition and perspectivist supervision compose rather than substitute, and that the full system is the only configuration that closes the generalisation gap.

Beyond these quantitative gains, the thesis contributes a conceptual shift: from treating media bias detection as a flat classification task with a single ground truth, to framing it as a reasoning task under uncertainty, grounded in plurality and context. This shift carries methodological consequences that extend beyond the specific architecture proposed here. It calls for evaluation protocols that treat cross-dataset behaviour as a first-class metric (a direct response to Problem 5), for annotation practices that preserve rather than collapse legitimate disagreement (a response to Problem 3), and for system designs that expose their intermediate reasoning rather than hide it behind monolithic predictions (a response to the joint demands of Problems 1 and 2). The four contributions taken together do not aspire to exhaust the structural diagnosis of Section 2.3: each of the five problems remains, in different degrees, an open research direction, and the next section reads the limitations stated in Section 5.6 as the concrete starting points from which that programme should proceed.

6.2 FUTURE WORK

The five subsections that follow turn each limitation acknowledged in Section 5.6 into a concrete research direction, ordered from immediate extensions of the present framework (dataset expansion, debiasing, explainability) to longer-term goals (multimodal detection, temporal modelling).

6.2.1 *Expanding MBBMD: enhancing diversity in contexts and languages*

The Media Bias Bias-Mitigated Dataset (MBBMD) introduced in this dissertation represents a significant step forward in media bias detection by integrating a hierarchical annotation framework and leveraging advanced techniques such as counterfactual data augmentation. However, despite its contributions, MBBMD (like all existing media bias detection datasets) remains limited in linguistic scope, contextual diversity, and thematic coverage. To enhance its utility and improve the generalization capabilities of bias detection models, future work should focus on expanding MBBMD along multiple dimensions.

A primary direction for expansion involves broadening MBBMD's linguistic diversity. Media bias is not confined to a single language, and its manifestations vary significantly across linguistic and cultural boundaries. While MBBMD is currently focused on Spanish-language media, future iterations should incorporate multilingual extensions, capturing bias in English, Portuguese, French, and other languages with significant media influence. Rather than relying solely on translated texts, native-language annotations are necessary to ensure that bias detection models account for language-specific framing strategies and rhetorical techniques.

In addition to linguistic diversity, geopolitical and contextual expansion is crucial. Bias in news reporting is often shaped by local political climates, historical narratives, and societal norms. While MBBMD currently includes sources from Spain and Latin America, future iterations should aim to incorporate media from different geopolitical regions, including North America, Europe, Asia, and the Middle East. This expansion would enable models to better distinguish between regionally specific bias patterns and global media narratives.

Another essential aspect of dataset expansion is thematic broadening. At present, bias detection datasets, including MBBMD, predominantly focus on highly polarized political topics such as elections, governance, and partisan conflicts. While these issues are central to understanding bias, they do not represent the full spectrum of biased reporting. Expanding MBBMD to include topics such as climate change, public health crises, economic inequality, and technological developments would allow for a more comprehensive analysis of bias beyond traditional political discourse.

Finally, temporal coverage should be extended to capture longitudinal changes in bias patterns. Media narratives evolve over time, influenced by political shifts, social movements, and emerging global crises. Expanding MBBMD to include a time-series component would enable models to analyze how media bias changes dynamically, providing deeper insights into its evolution.

By addressing these limitations, MBBMD can evolve into a more comprehensive, multilingual, and contextually diverse resource, significantly improving the reliability and adaptability of media bias detection models in real-world applications.

6.2.2 *Debiasing techniques: transforming biased statements into neutral ones*

One of the fundamental contributions of this dissertation is the ability to detect specific manifestations of media bias at the sentence level, distinguishing between different types such as sensationalism, mind reading, and word choice framing. This granular approach to bias detection not only enhances model interpretability but also enables the development of automated debiasing techniques, where biased statements can be transformed into neutral or minimally biased versions while preserving factual accuracy.

Despite the significance of this problem, relatively few studies have attempted to develop systematic approaches for debiasing media narratives. Existing research in this domain has primarily focused on assessing the limitations of conversational language models in news debiasing [165], leveraging automated rewriting techniques to mitigate media bias in news recommender systems [157], or addressing reporting biases in social media analytics [34]. However, these studies remain isolated efforts, and there is currently no widely accepted framework for systematically transforming biased news content into neutral representations while preserving journalistic integrity.

This gap highlights the need for further research into media bias mitigation strategies that go beyond simple lexical adjustments. Future work should explore methods that take into account the linguistic, ideological, and cultural dimensions of bias, ensuring that any

[165] Schlicht et al., 2024, "Pitfalls of conversational llms on news debiasing"

[157] Ruan et al., 2024, "Rewriting Bias: Mitigating Media Bias in News Recommender Systems through Automated Rewriting"

[34] Chen et al., 2016, "De-biasing the reporting bias in social media analytics"

debiasing intervention remains contextually appropriate and does not introduce unintended distortions in the original message.

Given that MBBMD provides structured annotations identifying specific instances of bias, future research can leverage fine-grained debiasing strategies that directly target these linguistic patterns. One promising line of work involves the use of heuristic rewriting rules, derived from the same linguistic templates used to detect bias manifestations. For example, if a sentence includes emotionally charged adjectives or subjective qualifiers, these can be systematically replaced with neutral alternatives using lexicons or syntactic patterns. Similarly, sentences exhibiting mind reading or vague attribution can be reformulated to express verifiable facts or to include explicit sourcing.

Beyond heuristic editing, sequence-to-sequence (Seq2Seq) models trained for paraphrasing can be fine-tuned to produce neutralized versions of biased inputs. These models can learn to preserve semantic content while reducing the intensity or subjectivity of the expression. Their outputs can be guided by constraints informed by the bias taxonomy developed in this thesis, ensuring that specific manifestations are minimized without distorting the factual content.

In parallel, decoder-only models such as instruction-tuned LLMs (e.g., GPT variants) offer an opportunity to explore prompted rephrasing. These models can be instructed to "rewrite the following sentence to remove biased language while preserving its factual meaning", allowing for more flexible and context-sensitive debiasing. Future work should evaluate the consistency and reliability of such generations, and assess whether instruction-based debiasing can generalize across different bias categories and journalistic domains.

Finally, evaluating the effectiveness of debiasing techniques is a critical area for future research. Standard NLP evaluation metrics, such as BLEU or ROUGE, are insufficient for measuring bias reduction, as they prioritize textual similarity rather than ideological neutrality. Future studies should develop custom evaluation frameworks that combine quantitative bias reduction metrics with human assessments of fairness, neutrality, and factual integrity.

6.2.3 *Multimodal bias detection: integrating text, images, and video*

Media bias extends beyond textual content, influencing public perception through images, videos, headlines, layout choices, and other visual cues. While the research presented in this dissertation has focused primarily on text-based bias detection, the growing prevalence of multimodal news reporting necessitates a broader approach. Future research should aim to develop multimodal bias detection systems that integrate textual, visual, and even auditory signals to provide a more comprehensive analysis of bias in media.

One of the most evident forms of multimodal bias is visual framing, where media outlets selectively choose images that align with a particular narrative. For example, a news article covering a political event might use images that depict a politician in either a positive or negative light, reinforcing a specific viewpoint. Similarly, video editing techniques, such as selective clipping or altered sequencing, can be used to emphasize certain aspects of an event while omitting others. Future bias detection models should incorporate computer vision techniques to analyze image sentiment, object presence, and facial expressions in news imagery, assessing whether they contribute to a biased portrayal of the subject.

Another crucial aspect is headline bias, which often exists at the intersection of text and layout. Headlines can be sensationalist, misleading, or emotionally charged, even when the body of the article is relatively neutral. Combining NLP models with document design analysis could help detect discrepancies between a headline and its corresponding article, identifying cases where the headline exaggerates, distorts, or selectively frames information. Additionally, font size, color schemes, and placement of headlines can subtly influence readers' perceptions, making document layout analysis an important dimension of multimodal bias detection.

Beyond static imagery, video and audio content introduce new challenges for bias analysis. Speech-to-text transcription is an essential first step, enabling NLP models to analyze spoken content in news broadcasts. However, spoken language often includes intonation, emphasis, and rhetorical techniques that are not fully captured by standard text-based models. Future research should explore prosody analysis (examining tone, pitch, and stress) to detect subtle indicators of subjective framing in broadcast media. Furthermore, video analysis techniques,

such as shot framing, camera angles, and scene composition, could be leveraged to assess whether visual elements contribute to biased reporting.

Developing multimodal datasets is a key prerequisite for these advancements. While MBBMD provides detailed textual annotations, future extensions should incorporate image and video metadata, enabling a more comprehensive study of media bias across formats. A multimodal dataset would allow for the training of vision-language models, capable of simultaneously analyzing textual and visual bias. Such an approach could leverage Contrastive Language-Image Pretraining (CLIP)-based models, which have demonstrated strong performance in aligning text and image representations.

Another research avenue involves the development of explainable multimodal bias detection models. Unlike text-based bias analysis, where explanations can be generated by highlighting specific words or phrases, multimodal analysis requires new interpretability techniques that justify why an image, video frame, or combination of elements is considered biased. Future research should focus on visual saliency mapping and attention mechanisms that highlight which aspects of an image or video contribute to bias perception.

By integrating text, images, video, and audio into a unified bias detection framework, future studies can move beyond textual analysis and develop more comprehensive tools for detecting, analyzing, and mitigating media bias in an increasingly multimodal information landscape.

6.2.4 *Explainability and interpretability in media bias analysis*

Explainability is a critical challenge in media bias detection, particularly in high-stakes applications where transparency and accountability are paramount. Common explainability techniques in ML, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), have provided valuable insights into how models make decisions by highlighting which input features contribute most to a given prediction. However, media bias detection presents a unique challenge: bias is inherently perspectivist, and its interpretation depends on the viewpoint of the observer. A single, universal explanation for why a statement is biased may not be sufficient, as different ideological perspectives may reach different conclusions based on the same content.

One of the key contributions of this dissertation is the development of a hierarchical and perspectivist dataset, which enables a richer, more structured form of explainability. Unlike traditional bias detection approaches that treat bias as a binary or static property, our framework acknowledges that bias can be decomposed into multiple subtypes and that disagreement among annotators is a valuable signal rather than noise. This hierarchical and perspectivist structure inherently enhances explainability by making explicit which aspects of a sentence contribute to bias perception, rather than simply assigning an opaque label of "biased" or "not biased".

Beyond improving interpretability at the dataset level, this structure also enables the development of perspective-aware explainability models. Future research should explore how LLMs trained on specific perspectives (e.g., conservative, progressive, centrist) can be used to provide contrastive explanations, clarifying why a particular perspective perceives a sentence as biased while another does not. By fine-tuning models with perspective-specific data, we can generate explanations that articulate the underlying ideological assumptions driving different interpretations.

A promising approach is to leverage decoder-only models to generate structured explanations. Rather than providing a single explanation for why a sentence is biased, the system could generate multiple competing explanations, reflecting different ideological standpoints. For instance, given a news headline, a model could generate the following:

> "*From a progressive perspective, this statement is biased because it frames economic policies as inherently harmful without acknowledging potential benefits. From a conservative perspective, this is unbiased because it reflects concerns over government intervention in the free market.*"

This approach shifts bias detection from a discrete classification task to a generative reasoning process, where the model articulates the reasoning behind different bias perceptions rather than just assigning a label.

To make this explainability framework more actionable, future work should focus on:

1. Training perspective-specific LLMs using bias-annotated data to capture ideological framing differences.
2. Developing contrastive prompt-based methods that encourage LLMs to generate explanations conditioned on ideological standpoints.
3. Enhancing model interpretability through structured output formats, where explanations are linked to explicit bias categories (e.g., "this sentence is perceived as sensationalist because it uses emotionally charged words").

Additionally, interactive explainability interfaces could be developed, allowing users to select a perspective and view dynamically generated explanations for why a particular viewpoint perceives a statement as biased. This would not only improve user engagement but also highlight the subjective nature of media bias, promoting media literacy by showing how framing effects influence interpretation.

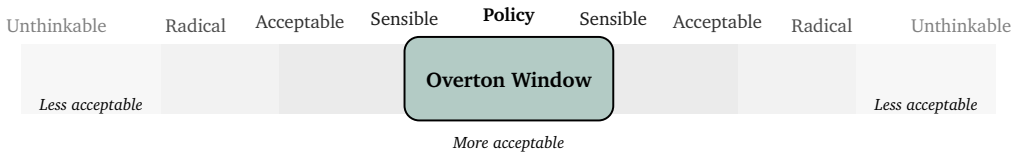
By integrating perspectivism into explainability, future research can move beyond simplistic "black-box" bias classification models toward more nuanced, multi-perspective explanations that reflect the complexities of real-world media bias.

6.2.5 *Temporal dynamics of media bias: towards an Overton window analysis*

Media bias is not static; it evolves over time in response to shifts in political discourse, cultural transformations, and societal norms. A statement or opinion that was once considered radical may become mainstream, while previously dominant narratives can gradually be pushed to the margins of acceptability. This phenomenon has been widely discussed in political and social theory, often in relation to the Overton Window [22], a concept that describes the range of ideas that are considered acceptable within public discourse at a given moment.

[22] Bobric, 2021, "The overton window: A tool for information warfare"

FIGURE 6.1: The Overton Window and the spectrum of public discourse.



The Overton Window (Figure 6.1), first formulated by Joseph P. Overton, has been widely used to explain how certain ideas gain or lose legitimacy over time. According to this model, policy proposals and ideological stances move along a spectrum from unthinkable to radical, then to acceptable, and eventually to mainstream and popular, before possibly becoming policy or law. While this model provides a structured way to analyze ideological shifts, it is also a controversial framework, as it has been adopted by political strategists to advocate for deliberate manipulation of discourse, either by normalizing extreme positions or by shifting public perception through gradual rhetorical adjustments [204, 172].

[204] Youvan, 2024, "Shifting Boundaries of Acceptability: Examining the Overton Window and Its Modern Manipulators in US Discourse"

[172] Singh Chauhan, 2024, "Fragmented Overton Windows: Rethinking Political Viability in Polarised Public Spheres"

Despite its debated theoretical foundations, the core observation underlying the Overton Window remains empirically valid: societal attitudes towards certain issues change over time [68, 98]. The evolution of public opinion on issues such as LGBTQ+ rights, climate change, and digital privacy illustrates how ideas move from the fringes of political debate to mainstream policy, and vice versa. Importantly, these shifts can occur both organically, through rational-critical debate in the public sphere, and through deliberate manipulation of discourse boundaries [17].

[68] Habermas, 2023, *A new structural transformation of the public sphere and deliberative politics*

[98] Lears, 1985, "The concept of cultural hegemony: Problems and possibilities"

[17] Bar-On, 2014, "A response to Alain de Benoist"

Given this fluid nature of ideological boundaries, future research in media bias analysis must incorporate a temporal dimension. Several key areas require exploration:

1. Longitudinal dataset construction: Most existing bias detection datasets, including MBBMD, are constructed as static snapshots of media discourse. To study the temporal evolution of bias, future datasets should integrate historical media sources spanning different decades, enabling researchers to track how framing, word choice, and ideological alignment change over time.
2. Dynamic bias classification models: Traditional bias classifiers operate under the assumption that bias categories remain stable. However, future systems should consider dynamic re-calibration [94], where models are periodically updated to reflect shifts in discourse norms. This could involve training time-sensitive bias detection models that adjust their predictions based on the publication date of a given text.
3. Event-driven bias shifts: Certain political events or crises accelerate changes in public discourse. For example, the COVID-19 pandemic dramatically altered perceptions of government intervention in healthcare and personal freedoms. Future bias detection frameworks should develop methodologies to identify and analyze bias shifts linked to major socio-political events.
4. Predictive modeling of ideological trends: By integrating time-series analysis, machine learning models could be trained to predict how bias in media coverage might evolve based on past trends. This could enable researchers and policymakers to anticipate potential shifts in political narratives and assess the role of media in shaping future ideological landscapes [10, 89].

[94] Kreuzberger et al., 2023, "Machine learning operations (mlops): Overview, definition, and architecture"

[10] Aradau and Blanke, 2017, "Politics of prediction: Security and the time/space of governmentality in the age of big data"

[89] Kaufmann et al., 2019, "Predictive policing and the politics of patterns"

By integrating temporality into media bias detection, researchers can develop more robust and contextually aware models that account for the ever-changing nature of public discourse. Rather than treating bias as a fixed property of a text, future research should explore how media framing strategies adapt, evolve, and reconfigure over time, reflecting the broader transformations in ideological and cultural landscapes.

6.3 PREVIOUSLY PUBLISHED MATERIAL

The development of this dissertation has been significantly informed and enriched by prior research and publications that contribute foundational insights, methodologies, and analyses to the arguments and findings presented. These works have played a crucial role in shaping the understanding of the complex issues surrounding media bias, disinformation, and the interdisciplinary approaches necessary to address them.

▸ JOURNAL PUBLICATIONS

- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it". In: *Expert Systems with Applications* (2023), p. 121641
- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "Identifying Media Bias beyond Words: Using Automatic Identification of Persuasive Techniques for Media Bias Detection". In: *Procesamiento del Lenguaje Natural* 71 (2023), pp. 179–190
- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "Media Bias Bias-Mitigated Dataset (MBBMD): A Hierarchical, Perspectivist, and Counterfactually-Augmented Corpus for Bias Detection in Spanish News". In: *Procesamiento del Lenguaje Natural* 76 (2026). Revista nº 76

▸ CONFERENCE & SHARED TASK PUBLICATIONS

- Francisco-Javier Rodrigo-Ginés, Laura Plaza, and Jorge Carrillo-de Albornoz. "UnedMediaBiasTeam@ SemEval-2023 Task 3: Can We Detect Persuasive Techniques Transferring Knowledge From Media Bias Detection?" In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. 2023, pp. 787–793

- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "Hierarchical Modeling for Propaganda Detection: Leveraging Media Bias and Propaganda Detection Datasets." In: *IberLEF@ SEPLN*. 2023
- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "UNED-BiasTeam at IberLEF 2021's EXIST Task: Detecting Sexism Using Bias Techniques." In: *IberLEF@ SEPLN*. 2021, pp. 522–532
- Francisco-Javier Rodrigo-Ginés. "Automated Media Bias Detection: Challenges and Opportunities". In: *Proceedings of the Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA (PLN-DS@SEPLN 2023)*. 2023, pp. 86–94

▸ DATASETS

- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. *MBBMD: Media Bias Bias-Mitigated Dataset (v2.0.0)*. Data set. 2025. DOI: 10.5281/zenodo.14806064. URL: https://zenodo.org/records/19160881
- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. *Meneame Media Bias Dataset: Interaction Features and Bias Labels (0.1.0)*. Data set. 2026. DOI: 10.5281/zenodo.19164549. URL: https://doi.org/10.5281/zenodo.19164549

Bibliography

[1] Samyak Agrawal, Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. “Towards detecting political bias in hindi news articles”. In: *Proceedings of the 60th annual meeting of the association for computational linguistics: student research workshop*. 2022, pp. 239–244 (cit. on pp. 11, 28).

[2] Victória Patrícia Aires, Juliana Freire, and Altigran Soares da Silva. “An Information Theory Approach to Detect Media Bias in News Websites”. In: 2020 (cit. on p. 30).

[3] Owais Ali, Zubair Zaland, Sibghat Ullah Bazai, Muhammad Imran Ghafoor, Laiq Hussain, and Arsalan Haider. “Neural transformers for bias detection: Assessing Pakistani News”. In: *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*. IEEE. 2024, pp. 1–7 (cit. on p. 28).

[4] Barry Allen. *Truth in philosophy*. Harvard University Press, 1993 (cit. on p. 5).

[5] Enrique Amigó, Jorge Carrillo-de Albornoz, Mario Almagro-Cádiz, Julio Gonzalo, Javier Rodríguez-Vidal, and Felisa Verdejo. “Evall: Open access evaluation for information access systems”. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2017, pp. 1301–1304 (cit. on p. 100).

[6] Enrique Amigó, Jorge Carrillo-de Albornoz, Andrés Fernández, Julio Gonzalo, Miguel Lucas, Guillermo Marco, Roser Morante, Jacobo Pedrosa, Laura Plaza, Ibo Sanz, et al. “ODESIA: Space for Observing the Development of Spanish in Artificial Intelligence.” In: *SEPLN (Projects and Demonstrations)*. 2023, pp. 50–54 (cit. on p. 100).

[7] Enrique Amigo and Agustín Delgado. “Evaluating extreme hierarchical multi-label classification”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 5809–5819 (cit. on p. 100).

[8] Bharat Anand, Rafael Di Tella, and Alexander Galetovic. “Information or opinion? Media bias as product differentiation”. In: *Journal of Economics & Management Strategy* 16.3 (2007), pp. 635–682 (cit. on p. 47).

[9] Marti J Anderson. “Permutational multivariate analysis of variance (PERMANOVA)”. In: *Wiley statsref: statistics reference online* (2014), pp. 1–15 (cit. on p. 78).

[10] Claudia Aradau and Tobias Blanke. “Politics of prediction: Security and the time/space of governmentality in the age of big data”. In: *European Journal of Social Theory* 20.3 (2017), pp. 373–391 (cit. on p. 135).

[11] Gabriel Domingos de Arruda, Norton Trevisan Roman, and Ana Maria Monteiro. “Analysing bias in political news.” In: *J. Univers. Comput. Sci.* 26.2 (2020), pp. 173–199 (cit. on pp. 30, 32).

[12] Alain Badiou. “Democratic materialism and the materialist dialectic”. In: *Radical philosophy* 130 (2005), pp. 20–24 (cit. on p. 6).

[13] Brent H Baker, Tim Graham, and Steve Kaminsky. *How to identify, expose & correct liberal media bias*. Media Research Center Alexandria, VA, 1994 (cit. on pp. 44, 129).

[14] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. “We Can Detect Your Bias: Predicting the Political Ideology of News Articles”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 4982–4991 (cit. on p. 27).

[15] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. “Predicting Factuality of Reporting and Bias of News Media Sources”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 3528–3539 (cit. on pp. 32, 33).

[16] Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. “What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 3364–3374 (cit. on p. 33).

[17] Tamir Bar-On. “A response to Alain de Benoist”. In: *Journal for the Study of Radicalism* 8.2 (2014), pp. 123–168 (cit. on p. 134).

[18] Katarzyna Baraniak and Marcin Sydow. “News articles similarity for automatic media bias detection in Polish news portals”. In: *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. 2018, pp. 21–24 (cit. on p. 24).

[19] Eric PS Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. “Framing Annotation Data for News Articles”. In: North American Chapter of the Association for Computational Linguistics (NAACL). 2015 (cit. on p. 33).

[20] Allan Bell. *The language of news media*. Blackwell Oxford, 1991 (cit. on p. 9).

[21] Alberto Benayas, Miguel Angel Sicilia, and Marçal Mora-Cantallops. “A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and performance”. In: *Language Resources and Evaluation* (2024), pp. 1–24 (cit. on p. 31).

[22] George-Daniel Bobric. “The overton window: A tool for information warfare”. In: *ICCWS 2021 16th International Conference on Cyber Warfare and Security*. Academic Conferences Limited. 2021, p. 20 (cit. on p. 134).

[23] Toby Bolsen, James N Druckman, and Fay Lomax Cook. “How frames can undermine support for scientific adaptations: Politicization and the status-quo bias”. In: *Public Opinion Quarterly* 78.1 (2014), pp. 1–26 (cit. on p. 7).

[24] Oliver Bott and Torgrim Solstad. “Discourse expectations: explaining the implicit causality biases of verbs”. In: *Linguistics* 59.2 (2021), pp. 361–416 (cit. on p. 82).

[25] Sandrine Boudana. “A definition of journalistic objectivity as a performance”. In: *Media, Culture & Society* 33.3 (2011), pp. 385–398 (cit. on p. 17).

[26] L Brent Bozell and Brent H Baker. *And that’s the way it isn’t: A reference guide to media bias*. Media Research Center, 1990 (cit. on p. 18).

[27] Ceren Budak, Sharad Goel, and Justin M Rao. “Fair and balanced? Quantifying media bias through crowdsourced content analysis”. In: *Public Opinion Quarterly* 80.S1 (2016), pp. 250–271 (cit. on pp. 11, 33, 34).

[28] Christian Burgers and Anneke De Graaf. “Language intensity as a sensationalistic news feature: The influence of style on sensationalism perceptions and effects”. In: *Communications-The European Journal of Communication Research* 38.2 (2013), pp. 167–188 (cit. on p. 84).

[29] Christian Burgers, Elly A Konijn, and Gerard J Steen. “Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony”. In: *Communication theory* 26.4 (2016), pp. 410–430 (cit. on p. 84).

[30] Mar Castillo-Campos, David Becerra-Alonso, and Hajo G Boomgaarden. “Automated Detection of Media Bias Using Artificial Intelligence and Natural Language Processing: A Systematic Review”. In: *Social Science Computer Review* (2025), p. 08944393251331510 (cit. on p. 30).

[31] Jorge Chamorro-Padial, Roberto García González, and Rosa María Gil Iranzo. “FAIR-Check: script to check FAIR principles compliance”. In: (2024) (cit. on p. 65).

[32] Amanda Chan, Mai A Baddar, and Sofien Bâazaoui. “Eagles at FIGNEWS 2024 shared task: A context-informed prescriptive approach to bias detection in contentious news narratives”. In: *Proceedings of The Second Arabic Natural Language Processing Conference*. 2024, pp. 656–671 (cit. on p. 24).

[33] Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. “How Do Large Language Models Acquire Factual Knowledge During Pretraining?” In: *arXiv preprint arXiv:2406.11813* (2024) (cit. on p. 68).

[34] Hongyu Chen, Zhiqiang Zheng, and Yasin Ceran. “De-biasing the reporting bias in social media analytics”. In: *Production and Operations Management* 25.5 (2016), pp. 849–865 (cit. on p. 131).

[35] Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. “Detecting Media Bias in News Articles using Gaussian Bias Distributions”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 4290–4300 (cit. on p. 27).

[36] Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. “Learning to flip the bias of news headlines”. In: *Proceedings of the 11th International conference on natural language generation*. 2018, pp. 79–88 (cit. on pp. 11, 33, 34).

[37] Andres Cremisini, Daniela Aguilar, and Mark A Finlayson. “A challenging dataset for bias detection: the case of the crisis in the Ukraine”. In: *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer. 2019, pp. 173–183 (cit. on pp. 33, 35, 36, 69, 113).

[38] André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. “On sentence representations for propaganda detection: From handcrafted features to word embeddings”. In: *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. 2019, pp. 107–112 (cit. on p. 23).

[39] Christopher L Cummings and Wei Yi Kong. “Breaking down “fake news”: Differences between misinformation, disinformation, rumors, and propaganda”. In: *Resilience and hybrid threats*. IOS Press, 2019, pp. 188–204 (cit. on pp. 1, 19).

[40] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. “A survey on computational propaganda detection”. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021, pp. 4826–4832 (cit. on p. 22).

[41] Peter M Dahlgren. “Media echo chambers: Selective exposure and confirmation bias in media use, and its consequences for political polarization”. In: (2020) (cit. on pp. 7, 11, 53).

[42] Dave D'Alessio and Mike Allen. "Media bias in presidential elections: A meta-analysis". In: *Journal of communication* 50.4 (2000), pp. 133–156 (cit. on p. 18).

[43] Melissa De Witte. *Groupthink gone wrong: Stanford scholars show how assumptions about electability undermine women political candidates*. 2022 (cit. on p. 51).

[44] José Alexandre F Diniz-Filho, Thannya N Soares, Jacqueline S Lima, Ricardo Dobrovolski, Victor Lemes Landeiro, Mariana Pires de Campos Telles, Thiago F Rangel, and Luis Mauricio Bini. "Mantel test in population genetics". In: *Genetics and molecular biology* 36 (2013), pp. 475–485 (cit. on p. 78).

[45] Laura D'Olimpio. "Critical perspectivism: Educating for a moral response to media". In: *Journal of Moral Education* 50.1 (2021), pp. 92–103 (cit. on p. 6).

[46] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. "A survey on in-context learning". In: *arXiv preprint arXiv:2301.00234* (2022) (cit. on p. 68).

[47] James N Druckman and Michael Parkin. "The impact of media bias: How editorial slant affects voters". In: *The Journal of Politics* 67.4 (2005), pp. 1030–1049 (cit. on p. 49).

[48] Mats Ekström, Amanda Ramsälv, and Oscar Westlund. "The epistemologies of breaking news". In: *Journalism Studies* 22.2 (2021), pp. 174–192 (cit. on p. 47).

[49] Antonia Enache. "Pitfalls of writing for the media: appeals to emotion versus exaggerated sensationalism". In: *Synergy* 20.2 (2024), pp. 209–226 (cit. on p. 84).

[50] Robert M Entman. "Framing bias: Media in the distribution of power". In: *Journal of communication* 57.1 (2007), pp. 163–173 (cit. on p. 2).

[51] Hülya Eraslan and Saltuk Özertürk. *Information gatekeeping and media bias*. Tech. rep. working paper, 2018 (cit. on p. 2).

[52] Alonso Estrada-Cuzcano, Karen Alfaro-Mendives, and Valeria Saavedra-Vásquez. "Disinformation y Misinformation, Posverdad y Fake News: precisiones conceptuales, diferencias, similitudes y yuxtaposiciones". In: *Información, cultura y sociedad* 42 (2020), pp. 93–106 (cit. on p. 19).

[53] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. "In Plain Sight: Media Bias Through the Lens of Factual Reporting". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 6343–6349 (cit. on pp. 11, 27, 32–34).

[54] Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. "A Multidimensional Dataset for Analyzing and Detecting News Bias based on Crowdsourcing". 2020 (cit. on pp. 33, 35).

[55] Alan Floyd Moore. "The reporting verbs and bias in the press". In: *Revista alicantina de estudios ingleses, No. 13 (Nov. 2000); pp. 43-52* (2000) (cit. on p. 82).

[56] Roger Fowler. *Language in the News: Discourse and Ideology in the Press*. Routledge, 2013 (cit. on p. 9).

[57] Jeffrey Friedman. *Post-truth and the epistemological crisis*. 2023 (cit. on p. 4).

[58] Barbara Fultner. *Jurgen Habermas: key concepts*. Routledge, 2014 (cit. on p. 5).

[59] Matthew Gentzkow and Jesse M Shapiro. "Media bias and reputation". In: *Journal of political Economy* 114.2 (2006), pp. 280–316 (cit. on p. 18).

[60] Daniel Geschke, Kai Sassenberg, Georg Ruhrmann, and Denise Sommer. "Effects of linguistic abstractness in the mass media". In: *Journal of Media Psychology* (2010) (cit. on p. 85).

[61] Joel Geske. "Riot vs. Revelry: News Bias Through Visual Media". In: *Teaching Media Quarterly* 4.1 (2016) (cit. on p. 49).

[62] Martin Gilens. "Race and poverty in americapublic misperceptions and the american news media". In: *Public Opinion Quarterly* 60.4 (1996), pp. 515–541 (cit. on p. 49).

[63] Corliss L Green. "Ethnic evaluations of advertising: Interaction effects of strength of ethnic identification, media placement, and degree of racial composition". In: *Journal of Advertising* 28.1 (1999), pp. 49–64 (cit. on p. 51).

[64] Tim Groseclose. *Left turn: How liberal media bias distorts the American mind*. Macmillan+ ORM, 2011 (cit. on p. 48).

[65] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. *NELA-GT-2021: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles*. 2022 (cit. on pp. 33, 34).

[66] Neeraj Gummalam and Clayton Greenberg. "Bias Detection in Media: An NLP-Based Approach using Corpus Statistics and Sentence Embeddings". In: *Journal of Student Research* 13.2 (2024) (cit. on p. 24).

[67] Viresh Gupta, Baani Leen Kaur Jolly, Ramneek Kaur, and Tanmoy Chakraborty. “Clark Kent at SemEval-2019 Task 4: Stylometric Insights into Hyperpartisan News Detection”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. 2019, pp. 934–938 (cit. on p. 24).

[68] Jürgen Habermas. *A new structural transformation of the public sphere and deliberative politics*. John Wiley & Sons, 2023 (cit. on p. 134).

[69] Prasad Hajare, Sadia Kamal, Siddharth Krishnan, and Arunkumar Bagavathi. “A Machine Learning Pipeline to Examine Political Bias with Congressional Speeches”. In: *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2021, pp. 239–243 (cit. on p. 23).

[70] Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. *Cohesion in english*. Routledge, 2014 (cit. on p. 9).

[71] Felix Hamborg. “Media bias, the social sciences, and NLP: automating frame analyses to identify bias by word choice and labeling”. In: *Proceedings of the 58th annual meeting of the association for computational linguistics: student research workshop*. 2020, pp. 79–87 (cit. on p. 50).

[72] Felix Hamborg, Karsten Donnay, and Bela Gipp. “Automated identification of media bias in news articles: an interdisciplinary literature review”. In: *International Journal on Digital Libraries* 20.4 (2019), pp. 391–415 (cit. on pp. 11, 18, 34, 44, 52, 129).

[73] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. “Illegal aliens or undocumented immigrants? Towards the automated identification of bias by word choice and labeling”. In: *Information in Contemporary Society: 14th International Conference, iConference 2019, Washington, DC, USA, March 31–April 3, 2019, Proceedings 14*. Springer. 2019, pp. 179–187 (cit. on p. 2).

[74] Anne-Wil Harzing. *The publish or perish book*. Tarma Software Research Pty Limited Melbourne, Australia, 2010 (cit. on p. 21).

[75] Jiwoo Hong, Yejin Cho, Jiyoung Han, Jaemin Jung, and James Thorne. “Disentangling structure and style: Political bias detection in news by inducing document hierarchy”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 5664–5686 (cit. on p. 26).

[76] Benjamin D Horne, Sara Khedr, and Sibel Adali. “Sampling the news producers: A large news and feature data set for the study of the complex media landscape”. In: *Twelfth International AAAI Conference on Web and Social Media*. 2018, pp. 518–527 (cit. on pp. 11, 33, 34).

[77] Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. “The promises and pitfalls of LLM annotations in dataset labeling: A case study on media bias detection”. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. 2025, pp. 1370–1386 (cit. on p. 29).

[78] Dick Howard and Dick Howard. “The Rationality of the Dialectic: Jean-Paul Sartre”. In: *The Marxian Legacy: The Search for the New Left* (2019), pp. 135–167 (cit. on p. 6).

[79] Christoph Hube and Besnik Fetahu. “Detecting biased statements in wikipedia”. In: *Companion proceedings of the the web conference 2018*. 2018, pp. 1779–1786 (cit. on pp. 23, 24).

[80] Christoph Hube and Besnik Fetahu. “Neural based statement classification for biased language”. In: *Proceedings of the twelfth ACM international conference on web search and data mining*. 2019, pp. 195–203 (cit. on p. 26).

[81] Sallie Hughes and Chappell Lawson. “Propaganda and crony capitalism: Partisan bias in Mexican television news”. In: *Latin American Research Review* 39.3 (2004), pp. 81–105 (cit. on p. 2).

[82] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. “Political ideology detection using recursive neural networks”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 1113–1122 (cit. on pp. 25, 32).

[83] Jakob D Jensen, Nick Carcioppolo, Andy J King, Jennifer K Bernat, LaShara Davis, Robert Yale, and Jessica Smith. “Including limitations in news coverage of cancer research: Effects of news hedging on fatalism, medical skepticism, patient trust, and backlash”. In: *Journal of health communication* 16.5 (2011), pp. 486–503 (cit. on p. 87).

[84] Tingting Jiang, Qian Guo, Shunchang Chen, and Jiaqi Yang. “What prompts users to click on news headlines? Evidence from unobtrusive data analysis”. In: *Aslib Journal of Information Management* (2019) (cit. on p. 51).

[85] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. *Logical Fallacy Detection*. 2022 (cit. on p. 23).

[86] Bo Kang, Dario Garcia Garcia, Jefrey Lijffijt, Raúl Santos-Rodríguez, and Tijl De Bie. “Conditional t-SNE: more informative t-SNE embeddings”. In: *Machine Learning* 110.10 (2021), pp. 2905–2940 (cit. on p. 76).

[87] Hyungsuc Kang and Janghoon Yang. “Quantifying perceived political bias of newspapers through a document classification technique”. In: *Journal of Quantitative Linguistics* 29.2 (2022), pp. 127–150 (cit. on p. 18).

[88] Natascha A Karlova and Karen E Fisher. “A social diffusion model of misinformation and disinformation for understanding human information behaviour”. In: *Information Research* (2013) (cit. on p. 19).

[89] Mareile Kaufmann, Simon Egbert, and Matthias Leese. “Predictive policing and the politics of patterns”. In: *The British journal of criminology* 59.3 (2019), pp. 674–692 (cit. on p. 135).

[90] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. “Semeval-2019 task 4: Hyperpartisan news detection”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. 2019, pp. 829–839 (cit. on pp. 33, 34).

[91] Kimberly N Kline. “A decade of research on health content in the media: the focus on health challenges and sociocultural context and attendant informational and ideological problems”. In: *Journal of health communication* 11.1 (2006), pp. 43–59 (cit. on p. 7).

[92] Julija Korostenskienė and Lina Bikelienė. “A comparison of degree intensifiers across English corpora: is it ‘flagrant’,‘blatant’, or ‘sheer’audacity?” In: *Valoda: nozima un forma* 12 (2021), pp. 74–92 (cit. on p. 84).

[93] Ralf Krestel, Alex Wall, and Wolfgang Nejdl. “Treehugger or Petrolhead? Identifying bias by comparing online news articles with political speeches”. In: *Proceedings of the 21st International Conference on World Wide Web*. 2012, pp. 547–548 (cit. on p. 23).

[94] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. “Machine learning operations (mlops): Overview, definition, and architecture”. In: *IEEE access* 11 (2023), pp. 31866–31879 (cit. on p. 135).

[95] Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. “A domain-adaptive pre-training approach for language bias detection in news”. In: *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. 2022, pp. 1–7 (cit. on pp. 11, 27).

[96] Tin Kuculo, Simon Gottschalk, and Elena Demidova. “QuoteKG: A Multilingual Knowledge Graph of Quotes”. In: *European Semantic Web Conference*. Springer. 2022, pp. 353–369 (cit. on pp. 25, 32).

[97] Konstantina Lazaridou, Ralf Krestel, and Felix Naumann. “Identifying media bias by analyzing reported speech”. In: *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2017, pp. 943–948 (cit. on pp. 25, 32).

[98] TJ Jackson Lears. “The concept of cultural hegemony: Problems and possibilities”. In: *The American historical review* (1985), pp. 567–593 (cit. on p. 134).

[99] Jihye Lee and James T Hamilton. “Anchoring in the past, tweeting from the present: Cognitive bias in journalists’ word choices”. In: *Plos one* 17.3 (2022) (cit. on p. 85).

[100] Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. “SemEval-2023 task 11: Learning with disagreements (LeWiDi)”. In: *arXiv preprint arXiv:2304.14803* (2023) (cit. on pp. 11, 53).

[101] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474 (cit. on p. 80).

[102] Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. “Annotating and analyzing biased sentences in news articles using crowdsourcing”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 1478–1484 (cit. on pp. 33, 34).

[103] Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. “Understanding Characteristics of Biased Sentences in News Articles”. In: *CIKM workshops*. 2018 (cit. on p. 34).

[104] Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. “What emotions do news articles trigger in their readers?” In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007, pp. 733–734 (cit. on p. 48).

[105] L Lin, L Wang, X Zhao, J Li, and KF Wong. “Indivec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators”. In: *arXiv preprint arXiv:2402.00345* (2024) (cit. on p. 28).

[106] Luyang Lin, Jing Li, and Kam-Fai Wong. “Data-augmented and retrieval-augmented context enrichment in chinese media bias detection”. In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 24.10 (2025), pp. 1–24 (cit. on pp. 11, 33, 35).

[107] Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. *Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception*. 2024 (cit. on p. 29).

[108] Abdullah Majali, Ref’at Alloush, Zaid Khreis, and Omar Qawasmeh. “BiasLens: Graph Machine Learning Approach for Media Bias Detection”. In: *2025 International Conference on New Trends in Computing Sciences (ICTCS)*. IEEE. 2025, pp. 347–352 (cit. on pp. 29, 32).

[109] Julie Mastrine, Kristine Sowers, Sara Alhariri, and Jeff Nilsson. *How to spot 16 types of media bias*. 2022. URL: https://www.allsides.com/media-bias/how-to-spot-types-of-media-bias (cit. on p. 42).

[110] Ruth Mayo. “Trust or distrust? Neither! The right mindset for confronting disinformation”. In: *Current Opinion in Psychology* 56 (2024), p. 101779 (cit. on p. 5).

[111] Maxwell McCombs and Sebastián Valenzuela. “The agenda-setting theory”. In: *Cuadernos de información* 20 (2007), pp. 44–50 (cit. on p. 8).

[112] Lee McIntyre. *On disinformation: How to fight for truth and protect democracy*. MIT Press, 2023 (cit. on p. 5).

[113] John McMurtry. “The mass media: An analysis of their system of fallacy”. In: *Interchange* 21.4 (1990), pp. 49–66 (cit. on p. 50).

[114] Ad Fontes Media. *How Ad Fontes Ranks News Sources*. 2023 (cit. on p. 56).

[115] Tim Menzner and Jochen L Leidner. “Experiments in news bias detection with pre-trained neural transformers”. In: *European Conference on Information Retrieval*. Springer. 2024, pp. 270–284 (cit. on pp. 11, 28).

[116] Tim Menzner and Jochen L Leidner. “BiasScanner: Automatic News Bias Classification for Strengthening Democracy”. In: *European Conference on Information Retrieval*. Springer. 2025, pp. 105–110 (cit. on pp. 28, 29, 32).

[117] Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. “Moral framing and ideological bias of news”. In: *International conference on social informatics*. Springer. 2020, pp. 206–219 (cit. on p. 42).

[118] Sendhil Mullainathan and Andrei Shleifer. *Media bias*. 2002 (cit. on pp. 1, 18, 42).

[119] Tomáš Musil. “Examining structure of word embeddings with PCA”. In: *International Conference on Text, Speech, and Dialogue*. Springer. 2019, pp. 211–223 (cit. on p. 76).

[120] Meytal Nasie, Daniel Bar-Tal, Ruthie Pliskin, Eman Nahhas, and Eran Halperin. “Overcoming the barrier of narrative adherence in conflicts through awareness of the psychological bias of naïve realism”. In: *Personality and social psychology bulletin* 40.11 (2014), pp. 1543–1556 (cit. on pp. 7, 11).

[121] Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. “Quotus: The structure of political media coverage as revealed by quoting patterns”. In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, pp. 798–808 (cit. on p. 24).

[122] Tatsuya Ogawa, Qiang Ma, and Masatoshi Yoshikawa. “News bias analysis based on stakeholder mining”. In: *IEICE transactions on information and systems* 94.3 (2011), pp. 578–586 (cit. on pp. 29, 32).

[123] Dwi Juniar Pangestuti. “Logical fallacy found in the politic column of BBC News online magazine: Lexical and syntactical ambiguities”. PhD thesis. Universitas Islam Negeri Maulana Malik Ibrahim, 2023 (cit. on p. 50).

[124] Kartikey Pant, Tanvi Dadu, and Radhika Mamidi. “Towards Detection of Subjective Bias using Contextualized Word Embeddings”. en. In: *Companion Proceedings of the Web Conference 2020*. Taipei Taiwan: ACM, Apr. 2020, pp. 75–76. ISBN: 9781450370240. (Visited on 11/15/2022) (cit. on p. 50).

[125] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011 (cit. on p. 1).

[126] Souneil Park, Kyung-Soon Lee, and Junehwa Song. “Contrasting opposing views of news articles on contentious issues”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011, pp. 340–349 (cit. on p. 24).

[127] Benjamin Parker-Bass, Ian Fette, Paul Mans, Monica Seth, Jim Sullivan, and Paul Washburn. *News Bias Explored*. http://websites.umich.edu/~newsbias/. (Accessed: 12-27-2022) (cit. on pp. 48, 52).

[128] Valeria Pastorino, Jasivan A Sivakumar, and Nafise Sadat Moosavi. “Decoding News Narratives: A Critical Analysis of Large Language Models in Framing Bias Detection”. In: *arXiv preprint arXiv:2402.11621* (2024) (cit. on pp. 28, 32).

[129] Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. “A link-based approach to detect media bias in news websites”. In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 742–745 (cit. on p. 32).

[130] Valentin Pelloin, Lena Dodson, Émile Chapuis, Nicolas Hervé, and David Doukhan. “Automatic classification of news subjects in broadcast news: Application to a gender bias representation analysis”. In: *arXiv preprint arXiv:2407.14180* (2024) (cit. on p. 28).

[131] Juan Manuel Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. “pysentimiento: A python toolkit for opinion mining and social nlp tasks”. In: *arXiv preprint arXiv:2106.09462* (2021) (cit. on p. 87).

[132] April M Perry. “Guilt by saturation: Media liability for third-party violence and the availability heuristic”. In: *Nw. UL Rev.* 97 (2002), p. 1045 (cit. on p. 7).

[133] Terry P Pinkard. “Hegel’s dialectic: The explanation of possibility”. In: (1988) (cit. on p. 6).

[134] Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. "A multi-task learning approach to hate speech detection leveraging sentiment analysis". In: *IEEE Access* 9 (2021), pp. 112478–112489 (cit. on p. 79).

[135] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. "Clickbait detection". In: *European conference on information retrieval*. Springer. 2016, pp. 810–817 (cit. on p. 23).

[136] Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. "Are Decoder-Only Language Models Better than Encoder-Only Language Models in Understanding Word Meaning?" In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 16339–16347. URL: https://aclanthology.org/2024.findings-acl.967/ (cit. on p. 31).

[137] TA Quijote, AD Zamoras, and A Ceniza. "Bias detection in Philippine political news articles using SentiWordNet and inverse reinforcement model". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 482. 1. IOP Publishing. 2019, p. 012036 (cit. on p. 30).

[138] Daniëlle Raeijmaekers and Pieter Maeseele. "In objectivity we trust? Pluralism, consensus, and ideology in journalism studies". In: *Journalism* 18.6 (2017), pp. 647–663 (cit. on p. 8).

[139] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. "Truth of varying shades: Analyzing language in fake news and political fact-checking". In: *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2017, pp. 2931–2937 (cit. on pp. 26, 32).

[140] Sachin Rawat and G Vadivu. "Media Bias Detection Using Sentimental Analysis and Clustering Algorithms". In: *Proceedings of International Conference on Deep Learning, Computing and Intelligence: ICDCI 2021*. Springer. 2022, pp. 485–494 (cit. on pp. 30, 32).

[141] Shaina Raza, Mizanur Rahman, and Michael R Zhang. "BEADs: Bias Evaluation Across Domains". In: *arXiv preprint arXiv:2406.04220* (2024) (cit. on pp. 33, 35, 36, 70, 113).

[142] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. *MBBMD: Media Bias Bias-Mitigated Dataset (v2.0.0)*. Data set. 2025. DOI: 10.5281/zenodo.14806064. URL: https://zenodo.org/records/19160881 (cit. on pp. 33, 64, 136).

[143] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "Media Bias Bias-Mitigated Dataset (MBBMD): A Hierarchical, Perspectivist, and Counterfactually-Augmented Corpus for Bias Detection in Spanish News". In: *Procesamiento del Lenguaje Natural* 76 (2026). Revista nº 76 (cit. on pp. 64, 135).

[144] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. *Meneame Media Bias Dataset: Interaction Features and Bias Labels (0.1.0)*. Data set. 2026. DOI: 10.5281/zenodo.19164549. URL: https://doi.org/10.5281/zenodo.19164549 (cit. on p. 136).

[145] Francisco-Javier Rodrigo-Ginés and Jorge Chamorro-Padial. "Media Bias Within Information Disorder: Bridging Two Research Communities Through a Systematic Review". In: *Proceedings of the Information Disorder Workshop (InDor)*. To appear. 2026 (cit. on p. 19).

[146] Francisco-Javier Rodrigo-Ginés, Jorge Chamorro-Padial, and Pablo Rodríguez Díaz. "The Epistemic Limits of NLP Models in Media Bias Detection: Toward a Framework for Context-Aware and Reflexive AI Systems". In: *Proceedings of the IEEE Conference on Artificial Intelligence (CAI)*. 2026 (cit. on p. 126).

[147] Francisco-Javier Rodrigo-Ginés. "Automated Media Bias Detection: Challenges and Opportunities". In: *Proceedings of the Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA (PLN-DS@SEPLN 2023)*. 2023, pp. 86–94 (cit. on p. 136).

[148] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "UNEDBiasTeam at IberLEF 2021's EXIST Task: Detecting Sexism Using Bias Techniques." In: *IberLEF@ SEPLN*. 2021, pp. 522–532 (cit. on pp. 23, 26, 136).

[149] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it". In: *Expert Systems with Applications* (2023), p. 121641 (cit. on pp. 11, 17, 20, 21, 44, 46, 129, 135).

[150] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "Hierarchical Modeling for Propaganda Detection: Leveraging Media Bias and Propaganda Detection Datasets." In: *IberLEF@ SEPLN*. 2023 (cit. on p. 136).

[151] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "Identifying Media Bias beyond Words: Using Automatic Identification of Persuasive Techniques for Media Bias Detection". In: *Procesamiento del Lenguaje Natural* 71 (2023), pp. 179–190 (cit. on pp. 31, 38, 68, 135).

[152] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. "A hierarchical perspective-aware framework for cross-dataset media bias detection". In: *Expert Systems with Applications* (2026). Under review (cit. on pp. 121, 122).

[153] Francisco-Javier Rodrigo-Ginés, Laura Plaza, and Jorge Carrillo-de Albornoz. "UnedMediaBiasTeam@ SemEval-2023 Task 3: Can We Detect Persuasive Techniques Transferring Knowledge From Media Bias Detection?" In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. 2023, pp. 787–793 (cit. on p. 135).

[154] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A primer in BERTology: What we know about how BERT works". In: *Transactions of the association for computational linguistics* 8 (2021), pp. 842–866 (cit. on p. 31).

[155] Ronja Rönnback, Chris Emmery, and Henry Brighton. "Automatic large-scale political bias detection of news outlets". In: *PLoS One* 20.5 (2025), e0321418 (cit. on p. 30).

[156] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65 (cit. on p. 78).

[157] Qin Ruan, Jin Xu, Susan Leavy, Brian Mac Namee, and Ruihai Dong. "Rewriting Bias: Mitigating Media Bias in News Recommender Systems through Automated Rewriting". In: *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 2024, pp. 67–77 (cit. on p. 131).

[158] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. "Learning to retrieve prompts for in-context learning". In: *arXiv preprint arXiv:2112.08633* (2021) (cit. on p. 68).

[159] Jeanette B Ruiz and Robert A Bell. "Understanding vaccination resistance: vaccine search term selection bias and the valence of retrieved information". In: *Vaccine* 32.44 (2014), pp. 5776–5780 (cit. on p. 47).

[160] Noor Sadiah, Sara Al-Emadi, and Sumaya Rahman. "Ceasefire at FIGNEWS 2024 shared task: Automated detection and annotation of media bias using large language models". In: *Proceedings of The Second Arabic Natural Language Processing Conference*. 2024, pp. 590–600 (cit. on p. 28).

[161] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. "Social media news communities: gatekeeping, coverage, and statement bias". In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013, pp. 1679–1684 (cit. on p. 43).

[162] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019) (cit. on p. 68).

[163] Soumyadeep Sar, Subinay Adhikary, and Dwaipayan Roy. "BAAF: A Framework for Media Bias Detection". In: *European Conference on Information Retrieval*. Springer. 2025, pp. 255–264 (cit. on p. 29).

[164] Richard K Scher. *The modern political campaign: Mudslinging, bombast, and the vitality of American politics*. ME Sharpe, 1997 (cit. on p. 48).

[165] Ipek Baris Schlicht, Defne Altiok, Maryanne Taouk, and Lucie Flek. "Pitfalls of conversational llms on news debiasing". In: *arXiv preprint arXiv:2404.06488* (2024) (cit. on p. 131).

[166] Andreas Schuck, Alina Feinholdt, et al. "News framing effects and emotions". In: *Emerging trends in the social and behavioral sciences* (2015), pp. 1–15 (cit. on p. 48).

[167] Kate Scott. "You won't believe what's in this paper! Clickbait, relevance and the curiosity gap". In: *Journal of pragmatics* 175 (2021), pp. 53–66 (cit. on p. 2).

[168] Ayşe Adin Selçuk. "A guide for systematic reviews: PRISMA". In: *Turkish archives of otorhinolaryngology* 57.1 (2019), p. 57 (cit. on p. 21).

[169] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on p. 10).

[170] Sushila Shelke and Vahida Attar. "Source detection of rumor in social network–a review". In: *Online Social Networks and Media* 9 (2019), pp. 30–42 (cit. on p. 22).

[171] Alex Sherstinsky. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network". In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306 (cit. on p. 25).

[172] Shivank Singh Chauhan. "Fragmented Overton Windows: Rethinking Political Viability in Polarised Public Spheres". In: *Available at SSRN 5097975* (2024) (cit. on p. 134).

[173] Robin Small. "Nietzsche and a Platonist tradition of the cosmos: Center everywhere and circumference nowhere". In: *Journal of the History of Ideas* 44.1 (1983), pp. 89–104 (cit. on p. 5).

[174] Hridya Sobhanam and Jay Prakash. "Analysis of fine tuning the hyper parameters in RoBERTa model using genetic algorithm for text classification". In: *International Journal of Information Technology* 15.7 (2023), pp. 3669–3677 (cit. on p. 68).

[175] Robert C Solomon. “Kierkegaard and" Subjective Truth"”. In: *Philosophy Today* 21.3 (1977), p. 202 (cit. on p. 5).

[176] Timo Spinde, Felix Hamborg, and Bela Gipp. “An integrated approach to detect media bias in german news articles”. In: *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*. 2020, pp. 505–506 (cit. on pp. 24, 32).

[177] Timo Spinde, Smilla Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. “The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias”. In: *arXiv preprint arXiv:2312.16148* (2023) (cit. on pp. 46, 47).

[178] Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. “Exploiting transformer-based multitask learning for the detection of media bias in news articles”. In: *International Conference on Information*. Springer. 2022, pp. 225–235 (cit. on pp. 27, 32).

[179] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. “Neural Media Bias Detection Using Distant Supervision With BABE–Bias Annotations By Experts”. In: *arXiv preprint arXiv:2209.14557* (2022) (cit. on pp. 33, 34).

[180] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. “Automated identification of bias inducing words in news articles using linguistic and context-oriented features”. In: *Information Processing & Management* 58.3 (2021), p. 102505 (cit. on pp. 11, 18, 27).

[181] Timo Spinde, Lada Rudnitckaia, Kanishka Sinha, Felix Hamborg, Bela Gipp, and Karsten Donnay. “MBIC–A Media Bias Annotation Dataset Including Annotator Characteristics”. In: *arXiv preprint arXiv:2105.11910* (2021) (cit. on pp. 33–35, 69, 113).

[182] Pannee Suanpang, Pitchaya Jamjuntr, and Phuripoj Kaewyong. “Sentiment analysis with a TextBlob package implications for tourism”. In: *Journal of Management Information and Decision Sciences* 24 (2021), pp. 1–9 (cit. on p. 85).

[183] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *Advances in neural information processing systems* 27 (2014) (cit. on p. 25).

[184] Daniel Sutter. “Can the media be so liberal-the economics of media bias”. In: *Cato J.* 20 (2000), p. 431 (cit. on p. 19).

[185] Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, and Viviana Patti. “Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017”. In: *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*. Vol. 1881. CEUR-WS. 2017, pp. 157–177 (cit. on p. 22).

[186] Ekaterina Veselinovna Teneva. “Digital pseudo-identification in the post-truth era: Exploring logical fallacies in the mainstream media coverage of the COVID-19 vaccines”. In: *Social Sciences* 12.8 (2023), p. 457 (cit. on p. 50).

[187] Ludovic Terren and Rosa Borge. “Echo chambers on social media: A systematic review of the literature”. In: (2021) (cit. on pp. 1, 8).

[188] Jeff Tischauser and Jesse Benn. “Whose post-truth era? Confronting the epistemological challenges of teaching journalism”. In: *Journalism & Mass Communication Educator* 74.2 (2019), pp. 130–142 (cit. on p. 4).

[189] Antonio Torralba and Alexei A Efros. “Unbiased look at dataset bias”. In: *CVPR 2011*. IEEE. 2011, pp. 1521–1528 (cit. on p. 70).

[190] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. “Learning from disagreement: A survey”. In: *Journal of Artificial Intelligence Research* 72 (2021), pp. 1385–1470 (cit. on pp. 10, 11, 38).

[191] Teun A Van Dijk. “Ideology and discourse analysis”. In: *The meaning of ideology*. Routledge, 2013, pp. 110–135 (cit. on p. 9).

[192] Jacob E Van Vleet. *Informal logical fallacies: A brief guide*. Hamilton Books, 2021 (cit. on p. 51).

[193] DT van. *Discourse and knowledge: a sociocognitive approach*. 2014 (cit. on p. 9).

[194] Yuli Vasiliev. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, 2020 (cit. on p. 82).

[195] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 27).

[196] T Franklin Waddell and S Shyam Sundar. “Bandwagon effects in social television: How audience metrics related to size and opinion affect the enjoyment of digital media”. In: *Computers in Human Behavior* 107 (2020), p. 106270 (cit. on p. 7).

[197] Silvio Waisbord. “Truth is what happens to news: On journalism, fake news, and post-truth”. In: *Journalism studies* 19.13 (2018), pp. 1866–1878 (cit. on p. 5).

[198] Lewis Raven Wallace. *The view from somewhere: Undoing the myth of journalistic objectivity*. University of Chicago Press, 2019 (cit. on p. 8).

[199] Liang Wang, Nan Yang, and Furu Wei. “Learning to retrieve in-context examples for large language models”. In: *arXiv preprint arXiv:2307.07164* (2023) (cit. on p. 81).

[200] M Wessel. *Improving Media Bias Detection with State-of-the-art Transformers*. gipplab.org, 2023 (cit. on pp. 11, 28).

[201] Martin Wessel and Tomáš Horych. “Beyond the surface: Spurious cues in automatic media bias detection”. In: *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*. 2024, pp. 21–30 (cit. on pp. 28, 38).

[202] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. “A prompt pattern catalog to enhance prompt engineering with chatgpt”. In: *arXiv preprint arXiv:2302.11382* (2023) (cit. on p. 81).

[203] Audrey Yap. “Ad hominem fallacies, bias, and testimony”. In: *Argumentation* 27.2 (2013), pp. 97–109 (cit. on p. 48).

[204] Douglas C Youvan. “Shifting Boundaries of Acceptability: Examining the Overton Window and Its Modern Manipulators in US Discourse”. In: (2024) (cit. on p. 134).

[205] Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa R El-Beltagy, and Muhammed AbuOdeh. “The FIGNEWS Shared Task on News Media Narratives”. In: *arXiv preprint arXiv:2407.18147* (2024) (cit. on pp. 11, 33, 35, 36, 70, 113).

[206] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. “Revisiting few-sample BERT fine-tuning”. In: *arXiv preprint arXiv:2006.05987* (2020) (cit. on p. 68).

[207] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. “Chain of preference optimization: Improving chain-of-thought reasoning in llms”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 333–356 (cit. on p. 94).

[208] Xinyi Zhou and Reza Zafarani. “A survey of fake news: Fundamental theories, detection methods, and opportunities”. In: *ACM Computing Surveys (CSUR)* 53.5 (2020), pp. 1–40 (cit. on p. 22).

[209] Zilin Cidre Zhou. “The logical fallacies in political discourse”. In: (2018) (cit. on p. 50).

[210] Anastasia Zhukova, Terry Ruas, Felix Hamborg, Karsten Donnay, and Bela Gipp. “What’s in the News? Towards Identification of Bias by Commission, Omission, and Source Selection (COSS)”. In: *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE. 2023, pp. 258–259 (cit. on p. 51).

[211] Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. “Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology”. In: *arXiv preprint arXiv:1906.04571* (2019) (cit. on p. 52).

A

MBBMD Annotation guidelines

A.1 Annotation guidelines - phase 1

Identifying and annotating media bias in journalistic texts is crucial for understanding the impact news can have on public perception. News is often more than just a bearer of facts; it can subtly influence the opinions and attitudes of its consumers with embedded perspectives and biases. In a media environment where information continuously flows from multiple sources, it is essential to examine how this information is presented and its potential effect on the public.

This guide provides a resource for identifying bias in media that can influence the interpretation of information. By offering a detailed methodological framework, our goal is to equip annotators with the necessary tools to critically analyze news content, from word choice and narrative structure to the omission of relevant contexts and source selection.

Moreover, we aim to create a common ground for all annotators, ensuring that the bias identification and annotation process is uniform. Understanding bias involves a detailed observation of how news is presented, including headline analysis, article focus, and how sources are selected and treated. By establishing clear criteria for detecting and annotating different types of bias, we promote a deeper and more nuanced understanding of how media narratives can influence social reality construction.

This initiative is part of a broader effort to foster transparency, objectivity, and accountability in journalism and media research. By enabling annotators to accurately identify bias in Spanish media, we contribute to creating valuable corpora for Natural Language Processing (NLP) research and fostering a more critical and reflective media consumption culture. Our purpose is that through this meticulous annotation work, we advance toward a more complete understanding of media bias dynamics and their impact on society.

▸ **Annotation goal**

The main purpose of this annotation is to:

- Identify and classify media bias at the document level: Determine the presence of bias in articles and the type of bias that appears.
- Contribute to research in the field of communication and information technology: The annotated corpus will serve as a valuable resource for academic studies and practical applications in the field of artificial intelligence and natural language processing.

▸ **Motivation**

The motivation behind annotating this corpus includes:

- Promoting transparency and objectivity in the media: By identifying and analyzing bias, a greater awareness of the importance of a fair and balanced news presentation is fostered.
- Advancing the development of bias detection tools: Detailed and accurate annotation of media bias can be used to train machine learning models that automate bias identification in texts.

- Contributing to a more informed public debate: Understanding bias in news allows readers to form more informed and critical opinions on current issues.

A.1.1 *Definition of key concepts*

▸ **What is media bias?**

Media bias is the practice of presenting news and information in a way that is not equitable or objective, using techniques that can be considered unfair or biased. This type of bias involves a distortion of reality, whether by omitting crucial details, disproportionately presenting one side of a story, or using language and approaches that favor certain opinions or interests. Media bias affects how the public receives and interprets information, potentially leading to a biased understanding of events and current issues.

▸ **What do we want to annotate?**

In this project, the goal of annotation is not limited to identifying whether a news item is biased or not; it goes further, seeking to understand the nature and type of bias present. This requires a detailed evaluation of each article to determine if it meets certain bias criteria and, if so, to classify the type of bias according to two main dimensions: the intention behind the bias and the context in which it is presented.

Intention behind the bias

Bias can arise from intentional efforts or more subtle influences:

- Intentional bias: This type of bias is the result of a deliberate choice by the media or the author to present information in a way that can influence or manipulate the reader's perception. Identifying this bias requires an analysis of elements such as tone, word choice, and possible omission of relevant facts that could offer a more balanced view of the topic.
- Spin bias: Spin or rhetorical bias occurs when a journalist tries to create a "catchy story". To do so, they may use emotionally charged language, exaggerate, or select only the facts that fit their narrative.

Context of the bias

In this classification, the context refers to the set of factors surrounding a news item that can influence how information is presented (and potentially biased).

- Bias by statement: Bias by statement, also called presentation bias, refers to how articles choose to report on certain entities or concepts. This can be achieved through the use of loaded language, the use of fallacies, or the choice of words or labeling.
- Bias by coverage: Involves an imbalance in the attention given to different topics or events, either by excess or by defect. Identifying this bias requires a comparison of coverage among different media on the same topic, assessing the depth and breadth with which facts are addressed.
- Bias by content control: This refers to the editorial selection of which news is published and which is not, affecting the information that reaches the public. Detecting this bias involves analyzing the presence or absence of certain topics in a medium compared to others, and may require a review of multiple sources to evaluate consistency in the coverage of an event or topic.

To effectively annotate the type of bias, it is crucial not only to analyze the content of a news item in isolation but to consider it within a broader spectrum of coverage on the same fact by different media. This allows us to identify patterns of omission or disproportionate emphasis, which are indicative of bias. Additionally, it is necessary to consider that a news item may exhibit more than one type of bias, so our annotation must be meticulous and thoughtful, based on clear and illustrative criteria to guide the identification and classification of the present bias.

▸ **WHAT DO WE NOT WANT TO ANNOTATE?**

We do not seek to evaluate the veracity of the facts presented in the news. It is not a fact-checking task. We assume that given a context of ten news items, it is possible to discern the veracity of the news.

We are also not interested in annotating the opinions of the author or the media. What we are looking for is to identify the bias present in the way information is presented, regardless of the ideological stance or perspective of the author.

A.1.2 *Evaluation criteria*

The effective evaluation of media bias is crucial to ensure the accuracy and reliability of the annotation process. In this section, the criteria and examples are detailed to guide annotators in identifying biases in media articles.

▸ **INDICATORS OF MEDIA BIAS**

To effectively identify media bias, annotators must be aware of certain key indicators in the texts. These indicators include:

- Selection of topics: The choice to cover certain topics while ignoring others.
- Use of language: Choice of words with positive or negative connotations, presentation of opinions as facts, etc.
- Framing of information: How a story is presented, including what details are emphasized and which are minimized.
- Imbalance in sources and quotes: Dependence on sources that favor a particular view or the omission of alternative perspectives.
- Omission of context or relevant data: Lack of crucial information that could alter the reader's perception.
- Manipulation of data or statistics: Presenting data in a way that supports a specific narrative, even if this involves distorting its real meaning.

▸ **EXAMPLES OF MEDIA BIAS**

In this section, concrete examples of news that exhibit media bias are presented to better illustrate the concepts explained earlier. Since identifying bias by coverage and bias by content control requires reading a news item given a context (other media covering the same topic/content), we add several news items that report the same fact on the same date.

Example 1 - Context: Unemployment data in Spain in February 2023

Given the following news:

- ElDiario.es - Employment grows in February by more than 103,000 workers, the best data since 2007[1].
- Público - In February, 103,621 jobs were created in Spain, the best data in 17 years[2].
- El Confidencial - The labor market confirms the economic acceleration with 103,600 more affiliates in February[3].
- El País - The labor market rebounds with 104,000 new jobs in the best month of February since 2007[4].
- La Voz de Galicia - The labor market accelerates and closes the best February since the financial crisis[5].
- El Mundo - Employment closes the best February since 2007 with 103,621 more affiliates and unemployment down by 7,452 people[6].
- Libre Mercado - The keys to the misleading respite that February gives to the labor market[7].

[1] https://www.eldiario.es/economia/empleo-crece-febrero-103-000-trabajadores-mejor-dato-2007_1_10977257.html (Last accessed: June 3, 2026)

[2] https://www.publico.es/economia/febrero-crearon-103-621-empleos-espana-mejor-dato-17-anos.html (Last accessed: June 3, 2026)

[3] https://www.elconfidencial.com/economia/2024-03-04/afiliacion-febrero-paro-aceleracion-economia_3841841/ (Last accessed: June 3, 2026)

[4] https://elpais.com/economia/2024-03-04/el-mercado-laboral-repunta-en-febrero-con-104000-nuevos-empleos-el-mayor-incremento-desde-2007.html (Last accessed: June 3, 2026)

[5] https://www.lavozdegalicia.es/noticia/economia/2024/03/04/paro-cae-276-millones-personas-menor-cifra-mes-2008/00031709538840173595381.htm (Last accessed: June 3, 2026)

[6] https://www.elmundo.es/economia/2024/03/04/65e57508fc6c830b148b456f.html (Last accessed: June 3, 2026)

[7] https://www.libremercado.com/2024-03-04/jose-maria-rotellar-un-respiro-enganoso-en-la-tendencia-laboral-7104054/ (Last accessed: June 3, 2026)

Given this context, we can conclude that the news from Libre Mercado is biased, both by statement and by coverage. Additionally, it appears to be an intentional bias. The reasons why this news is biased are as follows:

- Use of loaded and negative language: the news uses expressions like “misleading respite” and “downward trend of the labor market” to describe the situation. These words convey a negative and pessimistic view of the data presented, which constitutes a bias by statement.
- Focus on negative data: although the news mentions the decrease in unemployment and the increase in employment, it focuses on negative aspects such as the comparison of youth unemployment with other countries, or the destruction of businesses (without citing sources that support this data). This unbalanced focus on negative data, while overlooking or minimizing the importance of positive data, is a form of bias by coverage.
- Lack of references in quotes: the news also presents bias by statement due to the lack of adequate quotes and references in the following statement: “The EU considers that there are almost a million people (985,000) who do not work in Spain and are not included in the unemployment lists”. In this statement, no source or reference is provided to support the information presented.

Example 2 - Context: Amnesty law negotiations
Given the following news:

- ElDiario.es - The PSOE trusts to announce an agreement with Junts to unblock this Thursday the amnesty law[8] .
- Público - PSOE and Junts rush the discount time to unblock the amnesty law at the deadline[9] .
- El Confidencial - The Venice Commission supports an amnesty, but criticizes the emergency processing of the Government[10] .
- El País - Puigdemont speaks of a new stage “without the burden of exile” in a key week of the amnesty[11] .
- La Voz de Galicia - The PSOE trusts to pass the amnesty despite not including the cases of terrorism[12] .
- El Mundo - Yolanda Díaz proposes to ban pardons for corruption while supporting amnesty for embezzlement of the “procés”[13] .
- Libertad Digital - The blunders of Bolaños with Brussels to try to sneak the amnesty through[14] .

Given this context, we can see on three occasions bias by statement in the publication of La Voz de Galicia. This bias is manifested in the lack of clear and specific attribution of sources on several occasions: (i) “sources of the formation show their ‘full conviction’ that this will be the case”, (ii) “sources of the direction claim that Junts has received much social pressure in Catalonia in recent months”, (iii) “sources of the negotiation did not even rule out that it was finally not possible to seal the pact and the law decayed”.

As the bias in the publication derives from the lack of clear and specific attribution of sources, it could be considered a spin bias, not intentional.

While these phrases could generate bias by statement, the final assessment of whether the news item as a whole is biased or not is based on a broader analysis that considers:

- The frequency and intensity of the biased phrases: A single biased phrase is not enough to determine that the entire news item is biased.
- The context in which the biased phrases are used: It is important to consider whether the biased phrases are used in a neutral way or if they are presented as objective facts.

On the other hand, regarding the news from Libertad Digital, the bias by statement is manifested in the use of loaded and negative language, the emphasis on errors and denials,

[8] https://www.eldiario.es/politica/psoe-confia-anunciar-acuerdo-junts-desbloquear-jueves-ley-amnistia_1_10978593.html (Last accessed: June 3, 2026)

[9] https://www.publico.es/politica/psoe-junts-apuran-descuento-desatascar-ley-amnistia-limite-plazo.html (Last accessed: June 3, 2026)

[10] https://www.elconfidencial.com/espana/2024-03-01/comision-venecia-ley-amnistia-informe_3841168/ (Last accessed: June 3, 2026)

[11] https://elpais.com/espana/2024-03-02/puigdemont-augura-una-nueva-etapa-para-el-independentismo-pero-sin-renunciar-a-la-unilateralidad.html (Last accessed: June 3, 2026)

[12] https://www.lavozdegalicia.es/noticia/espana/2024/03/04/comision-justicia-reune-jueves-dictaminar-ley-amnistia/00031709553245704394959.htm (Last accessed: June 3, 2026)

[13] https://www.elmundo.es/espana/2024/03/04/65e5af8ffdddffc31c8b457a.html (Last accessed: June 3, 2026)

[14] https://www.libertaddigital.com/espana/politica/2024-03-04/los-patinazos-de-bolanos-con-bruselas-para-intentar-colar-la-amnistia-7104135/ (Last accessed: June 3, 2026)

the lack of context and neutrality, and the use of selective quotes. The following are some of the phrases and strategies that demonstrate this bias:

- Use of loaded and negative language: The news uses expressions like “blunders”, “leaked in a biased way”, “tortuous”, “dismantled Bolaños’ version”, and “manipulated a response from Europe” to describe the actions of the minister.
- Lack of context and neutrality: The news does not provide enough information about the content of the amnesty law and the debate around it, making it difficult to achieve an objective understanding of the topic.
- Use of selective quotes: The news uses selective quotes from the Venice Commission report to support its argument that Bolaños misrepresented the Commission’s position on the amnesty law.

In this case, the bias by statement and content coverage seems to be intentional due to the combination of several factors, such as the use of loaded and negative language, the lack of context and neutrality, and the use of selective quotes. These strategies suggest a conscious effort by the author to present a biased and negative view of the minister’s actions and the amnesty law.

Example 3 - Context: Statements by the Spanish government minister Yolanda Díaz on March 4, 2024

Given the following news:

- ElDiario.es - Yolanda Díaz does not find it “reasonable” for restaurants to be open at one in the morning[15] .
- Público - Ayuso criticizes Yolanda Díaz’s words about the hospitality industry and ignites Twitter users[16].
- El Confidencial - Yolanda Díaz proposes to ban pardons for corruption and regulate “lobbies”[17].
- El País - Sumar proposes to ban pardons for corruption while supporting amnesty for embezzlement of the “procés”[18].
- La Voz de Galicia - The Minister of Labor sees “madness” in restaurants being open at night, and Ayuso defends it[19].
- El Mundo - Yolanda Díaz ignites the hospitality industry by schedules: “She blatantly lies, all of Europe wants to look like us”[20].
- Libertad Digital - Ayuso’s PP: “Sumar closes before a restaurant in Madrid before midnight”[21].

Given this context, we can observe a bias by content control in the media “El Confidencial” and “El País”. Both media decided to report on Yolanda Díaz’s statements related to limiting immunities and banning pardons in cases of corruption (the other media also published on the hospitality statements topic), but they did not address the minister’s statements about the hospitality industry and the opening hours of restaurants.

This selection of content may suggest that these media chose to focus on an aspect of Yolanda Díaz’s statements that they considered more relevant, interesting, or convenient for their editorial line, while ignoring another aspect that was also newsworthy and had been covered by other media. This imbalance in coverage can generate a partial view of the minister’s statements and affect the public’s perception of her proposals and intentions. This type of bias is intentional, as it involves a conscious decision by the media about what information to publish and what information to omit.

On the other hand, in the article from “Público” titled “Ayuso criticizes Yolanda Díaz’s words about the hospitality industry and ignites Twitter users”, the media focuses on the negative reactions of some Twitter users to the statements of Isabel Díaz Ayuso, president of the Community of Madrid, in response to Yolanda Díaz’s words about the opening hours of

[15] This article has been removed by the media outlet as of June 3, 2026.

[16] https://www.publico.es/tremending/2024/03/04/ayuso-critica-las-palabras-de-yolanda-diaz-sobre-la-hosteleria-y-enciende-a-los-tuiteros/ (Last accessed: June 3, 2026)

[17] https://www.elconfidencial.com/espana/2024-03-04/yolanda-diaz-propone-prohibir-los-indultos-a-la-corrupcion-y-limitar-los-aforamientos_3841996/ (Last accessed: June 3, 2026)

[18] https://elpais.com/espana/2024-03-04/sumar-propone-limitar-los-aforamientos-y-prohibir-los-indultos-en-casos-de-corrupcion.html (Last accessed: June 3, 2026)

[19] https://www.lavozdegalicia.es/noticia/economia/2024/03/04/ministra-trabajo-ve-locura-restaurantes-esten-abiertos-madrugada-ayuso-defiende/00031709569519575888512.htm (Last accessed: June 3, 2026)

[20] https://www.elmundo.es/economia/2024/03/04/65e5c83421efa063798b458b.html (Last accessed: June 3, 2026)

[21] https://www.libertaddigital.com/madrid/2024-03-04/el-pp-de-ayuso-antes-cierra-sumar-que-un-restaurante-en-madrid-antes-de-las-12-de-la-noche-7104214/ (Last accessed: June 3, 2026)

restaurants. However, the article does not include opinions from users who may agree with Ayuso's statements or who offer a different perspective on the topic.

This lack of balance in the coverage of reactions on social media suggests a bias by coverage in the article. By presenting only the opinions of opponents to Ayuso's statements, the media could generate a partial and biased view of the public's reaction on social networks. To offer a more complete and balanced coverage, it would be necessary to include opinions from those who support Ayuso or who provide additional nuances to the debate.

In this case, regarding intentionality, it could be indicated that the news presents both intentional bias and spin bias. On one hand, it seems that this lack of balance is the result of a conscious decision by the media not to include favorable opinions to Ayuso. The lack of balance also appears to be the result of a deficient journalistic approach, as if journalistic balance had been replaced by sarcasm, with humorous intention.

A.1.3 *Annotation process*

In this annotation process, annotators evaluate each article in its entirety to determine the presence and type of media bias.

1. Annotator self-evaluation:
    - Age.
    - Gender.
    - Educational level.
    - Indicate personal political position on a scale of 1 to 5, where 1 represents an extreme left inclination and 5 an extreme right inclination. Number 3 indicates a centered position, while 2 and 4 represent moderate inclinations towards the left and right, respectively.
2. Integral reading:
    - Read all news items related to each specific topic.
    - Self-assess familiarity with the topic.
3. Identification and classification of bias:
    - Determine whether you consider each news item to be biased (Yes/No).
    - If the news item is biased, identify the intention behind the wording:
        - Intentional bias by the author.
        - Bias to create an impactful narrative (spin).
        - Both reasons apply.
    - Determine the main reason for the bias (all options can be checked):
        - Biased content.
        - Omission or partiality in the presentation of information.
        - Deliberate avoidance of relevant topics.

A.1.4 *Doubtful cases*

Identifying media bias is not always straightforward; often, decisions must be made in the context of a deep understanding of both the content and the subtleties of language. The examples and recommendations presented seek to illuminate areas of uncertainty, providing annotators with tools to discern the presence and type of bias more effectively. We will address cases that illustrate the complexity of assessing the intentionality behind the selection of sources, the use of language, the omission of context, among other critical aspects.

This section serves as an additional resource, allowing for a more rigorous and nuanced approach to identifying biases. By working with these examples and following the proposed strategies to resolve doubts, annotators will be better equipped to contribute to the development of annotated corpora with high reliability.

▸ **FEW BIASED PHRASES IN A NEWS ITEM**

- Example: A news item about public health policies includes several phrases that strongly criticize the measures taken by the government, calling them "ineffective" and "short-sighted", while the rest of the article presents a balanced discussion based on data and expert opinions.
- Doubt: Is the entire news item biased due to the presence of these critical phrases?
- Resolution: The determination of whether a news item is biased should consider the general context and the relative weight of the biased sections. If the critical and biased phrases are marginal within a broad and balanced analysis, the news item as a whole may not be considered biased. However, if these phrases are central to the thesis of the article or are presented without adequate counterpoints, then they could indicate significant bias. It is essential to evaluate how these phrases influence the general narrative and whether they distort the objective representation of the topic.

▸ **SENSATIONALIST HEADLINES**

- Example: A headline proclaims "Economy on the Verge of Collapse" in reference to a slight economic slowdown according to the latest indicators.
- Doubt: Does the use of dramatic language in the headline indicate a bias?
- Resolution: Sensationalism in headlines can be a form of bias, especially if the language used exaggerates the reality of the news or induces a disproportionate emotional response. To determine the bias, the content of the article should be compared with the headline to assess whether the latter accurately reflects the tone and information presented. If there is a significant discrepancy, the headline can make the entire news item biased.

▸ **LACK OF SPECIFIC KNOWLEDGE ABOUT THE TOPIC**

- Example: An opinion article analyzes the complexity of recent biotechnology regulations, harshly criticizing their supposed limitations and negative effects on innovation. The author uses technical terms and references specific studies to argue their point of view.
- Doubt: As an annotator without a deep knowledge of biotechnology, how to determine if the author's presentation reflects a bias or an informed and balanced evaluation of the topic?
- Resolution: In situations where the topic of the news item exceeds our specific knowledge, it is crucial to rely on intuition and the general feeling conveyed by reading the article, i.e., focusing more on form than content. If the tone of the article or the way information is presented gives a sense of one-sidedness or excessive criticism without room for alternative viewpoints, this may indicate the presence of bias. In these cases, it is advisable to conduct a targeted search on the sources cited in the article. The combination of an intuitive response to reading and a factual verification of the sources cited constitutes an effective strategy to address the doubt about bias in articles dealing with complex or specialized topics.

▸ **INCOMPLETE CONTEXTUALIZATION**

- Example: An article discusses the impact of a new educational policy mentioning only the negative criticisms, without including studies or experts who support its effectiveness.
- Doubt: Does the lack of complete contextualization represent a bias by omission?
- Resolution: The omission of relevant information that could alter the reader's perception of a topic indicates bias. To resolve the doubt, it is crucial to seek additional sources that provide the omitted context. If it is found that the omission deliberately excludes

perspectives that could balance the narrative, then the article should be considered biased.

▸ Unequal comparisons

- Example: Coverage compares the number of attendees at two political demonstrations, using different counting methodologies for each, resulting in a disproportionate representation of support for the causes.
- Doubt: Does the use of different counting methodologies introduce a bias in the comparison of events?
- Resolution: Yes, the use of inconsistent methodologies to compare similar events can introduce bias by favoring one interpretation over another. The resolution consists of evaluating the justification and transparency of the methodologies used and considering whether a more equitable presentation would have required the use of the same methodology for both events.

▸ Political labels

- Example: An article about Spanish politics refers to "Sumar/Podemos" as "ultra-left parties" and/or "Vox" as "ultra-right party", although these parties do not self-identify as such.
- Doubt: Is it biased to use these labels to describe political parties?
- Resolution: The use of terms like "ultra-left" and "ultra-right" to describe political parties can indicate bias. It is crucial to analyze whether, beyond the self-denomination of the parties, their actions or policies effectively align with what is traditionally considered "ultra-left" or "ultra-right". Moreover, it is important to discern whether the terms describe accurately and based on concrete facts or if, on the contrary, they are used for caricaturization or delegitimization of the parties. In cases where the label is used to overly simplify or discredit, it definitely reflects bias.

▸ Use of sources with known bias

- Example: An article about presidential elections predominantly cites political analysts associated with a specific political party, but includes a quote from an analyst of the opposing party criticizing the strategy of their own party.
- Doubt: Is the inclusion of an opposing source enough to consider the article balanced?
- Resolution: Evaluate the weight and relevance of the quotes in the context of the article. If the opposing quote is used to reinforce a biased argument rather than offering a legitimate counterpoint, the article may be considered biased.

▸ Presentation of statistics without context

- Example: A report highlights a 5% increase in employment in the technology sector this year, without mentioning that the previous year experienced a 20
- Doubt: Does the omission of last year's performance constitute a bias by omission?
- Resolution: Consider whether the omission significantly changes the interpretation of the statistics. If the omitted context would alter the reader's perception of the success of the sector, then there is bias by omission.

▸ **LANGUAGE CHOICE IN CONTROVERSIAL TOPICS**

- Example: In an article about abortion, the author exclusively uses the term "unborn murder" to refer to the practice.
- Doubt: Does this choice of words indicate intentional bias?
- Resolution: Yes, because the term used has strong connotations that reflect a specific position on the topic. Impartial coverage would employ more neutral language.

▸ **MANIPULATION OF QUOTES**

- Example: An article about environmental policies quotes a scientist saying, "it is impossible to stop climate change", omitting the part of the quote where he states, "without a coordinated global effort".
- Doubt: Does this manipulation represent an intentional bias?
- Resolution: Yes. By omitting a crucial part of the quote, the scientist's message is distorted to support a pessimistic or fatalistic narrative. The resolution involves searching for the complete quote or additional context.

▸ **EMOTIONAL FOCUS IN ECONOMIC TOPICS**

- Example: A report on tax increases focuses on the personal stories of individuals who will be negatively affected, without offering an economic analysis of the long-term benefits of the policy.
- Doubt: Does the emotional focus suggest a spin bias?
- Resolution: Consider whether the inclusion of balanced economic perspectives would provide a more complete view. The omission of a balanced analysis indicates bias.

A.1.5 *Ethical considerations*

▸ **ANNOTATOR IMPARTIALITY**

- Self-awareness: Annotators must be aware of their own predispositions and biases, especially in relation to their personal political position. It is essential that they recognize how these can influence their perception of bias in the news.

▸ **CONFIDENTIALITY AND DATA USE**

- Anonymity and privacy: Annotator anonymity will be guaranteed, and the privacy of all parties involved in the study will be respected. No names, email addresses, or any other personally identifiable information will be retained. All collected data will be treated anonymously and used solely for academic purposes.
- Transparency and informed consent: Annotators will receive complete information about the project's objectives, the use of the collected data, and privacy rights. Informed consent will be obtained before starting the annotation.
- Compliance with regulations: The project will comply with all relevant data protection and privacy laws and regulations.
- Ethical use of information: The information collected and annotated will be used exclusively for research purposes related to the study of media bias.

A.1.6 *Frequently Asked Questions (FAQs)*

▸ **What exactly is media bias?**
Media bias refers to the inclination or bias in the presentation of news and information. It can manifest through the choice of topics, the use of connotative language, the selection of sources, and the omission of relevant data or contexts, among others.

▸ **How can I identify bias in an article?**
Identifying bias involves evaluating various elements of the article, including the selection of words, the balance in the presentation of perspectives, the choice of cited sources, and the inclusion or exclusion of relevant information. It is also important to consider the general context in which the article is published and the possible intention behind its writing.

▸ **Is every article with a clear perspective biased?**
No. An article can present a particular perspective on a topic without being biased, as long as it is based on verifiable facts and provides adequate context. The key is whether the article attempts to inform in a balanced way or seeks to influence the reader's opinion by omitting crucial information or presenting data selectively.

▸ **Is it possible for an article to be unconsciously biased?**
Yes, unconscious bias can occur when the author has unrecognized personal biases that affect the way information is presented. This type of bias is more subtle and can be more difficult to identify, requiring careful evaluation of the content and the context in which the article is produced.

▸ **How can I maintain my own objectivity when annotating bias?**
Maintaining objectivity involves recognizing and reflecting on your own biases. It is helpful to contrast your evaluations with established criteria and, in case of doubt, consult the subsection on doubtful cases to gain perspective. It is also important to approach each article with an open mind, avoiding prejudices based on the media outlet or the author.

▸ **What should I do if I find contradictory information about bias in different sources?**
When you find contradictory information, it is essential to compare the sources and assess their credibility and relevance. Consider the context in which each source presents its information and seek consensus in reliable and authoritative sources on the topic.

▸ **Can an article contain multiple types of bias?**
Yes, an article can exhibit multiple types of bias simultaneously. For example, it can have a bias in the selection of sources (privileging some perspectives over others) and at the same time present a bias of omission (excluding relevant information that could alter the interpretation of the facts).

▸ **How does context influence the identification of bias?**
Context is crucial for understanding bias, as an article may be perceived differently depending on the cultural, political, or social context in which it is published. For example, the relevance of certain topics or the perception of certain opinions can vary significantly between contexts. It is important to consider the general context when evaluating whether an article is biased.

▸ **How is bias handled in opinion articles?**
Opinion articles, by nature, present a subjective perspective on a topic. However, even in these cases, it is important to assess whether the article respects the principles of fairness and accuracy, presenting arguments based on evidence and allowing the representation of relevant counterpoints. Although a certain degree of bias is expected in these articles, the manipulation of facts or the deliberate omission of relevant information can signal problematic bias.

▸ **What is the difference between bias and point of view?**
The point of view refers to the specific perspective from which a story or analysis is presented. All opinion articles have a point of view, which can be more neutral or inclined toward a particular position. Bias, on the other hand, involves a lack of balance and fairness in the presentation of that point of view, often through the omission, distortion, or selective selection of information to support a specific agenda.

▸ **How does emotionally charged language affect the perception of bias?**
The use of emotionally charged language can strongly influence the reader's perception, eliciting specific emotional responses and potentially distorting a balanced understanding of the facts. Such language may be indicative of bias if it is used to delegitimize opponents, exaggerate facts, or minimize important aspects of a story.

▸ **Do headlines influence the perception of bias in the article?**
Yes, headlines play a crucial role in how the content of an article is perceived. A headline can be biased if it presents a one-sided view, uses sensationalist language, or overly simplifies the complexity of the topic addressed. It is important to read the entire article to assess whether the content supports or contradicts the tone and implications of the headline.

▸ **My question or doubt is not reflected in this guide. What should I do?**
If, after reviewing this guide, you still have specific questions or doubts that have not been addressed, please contact us. You can send your questions or concerns by email to frodrigo@invi.uned.es.

A.2 Annotation guidelines - phase 2

This guide outlines the procedures and criteria for the second phase of media bias annotation, focusing on a detailed sentence-level analysis. Identifying and characterizing bias in the media is crucial for several reasons. Media bias can significantly shape public opinion and influence societal norms by presenting information in ways that reflect particular viewpoints or agendas. Understanding these biases is essential for promoting transparency and accountability in journalism, ensuring that audiences receive balanced and fair information.

The second phase of media bias annotation focuses on identifying specific mechanisms of bias at the sentence level, where subtleties in language and presentation may significantly influence readers' perceptions. By analyzing these elements, annotators contribute to a nuanced understanding of media narratives and their effects on public opinion.

This guide provides a structured framework for examining how bias manifests in linguistic choices, framing, and presentation styles. Our aim is to equip annotators with the necessary tools to critically evaluate the content, considering both explicit and implicit messages conveyed through media texts.

By establishing consistent criteria for identifying and categorizing sentence-level biases, this initiative enhances transparency and accountability in media analysis, supporting the development of robust corpora for Natural Language Processing (NLP) research. Ultimately, we aim to foster a more critical approach to media consumption and improve our understanding of media bias dynamics.

▸ **Motivation**
The motivation for annotating at the sentence level includes:

- Promoting awareness of subtle bias mechanisms that may not be evident at the document level.
- Supporting the development of advanced bias detection tools capable of identifying nuanced bias expressions.
- Contributing to a more informed public discourse by revealing how language choices can shape perceptions.

A.2.1 *Definition of key concepts*

▸ **What is media bias?**

Media bias is the practice of presenting news and information in a way that is not equitable or objective, using techniques that can be considered unfair or biased. This type of bias involves a distortion of reality, whether by omitting crucial details, disproportionately presenting one side of a story, or using language and approaches that favor certain opinions or interests. Media bias affects how the public receives and interprets information, potentially leading to a biased understanding of events and current issues.

▸ **What is media bias at a sentence-level?**

Sentence-level media bias refers to the subtle, often implicit, ways in which language and presentation choices can influence the interpretation of information within individual sentences of a news article. This includes the omission of critical context, the use of emotive or loaded language, and specific framing techniques that guide readers toward particular conclusions. By focusing on these micro-level elements, sentence-level bias can play a crucial role in shaping the overall narrative and significantly impact audience perceptions, even when such biases may not be evident at the document level. This type of analysis aims to capture and characterize the nuanced ways bias can be woven into the fabric of media texts, affecting how information is conveyed and interpreted sentence by sentence.

▸ **What do we want to annotate?**

In this phase, the annotation focuses on identifying specific bias mechanisms within individual sentences. These mechanisms are subtle yet impactful ways in which bias can shape the interpretation of information. The following categories outline the key types of sentence-level biases we aim to identify:

- **Omission of source attribution:** This involves identifying instances where critical information is presented without proper source citation. Such omissions can undermine the credibility of the content and make it difficult for readers to verify the information. For example, stating that "experts agree" without naming or citing the experts involved.
- **Sensationalism or emotionalism:** Recognizing the use of hyperbolic or emotionally charged language intended to provoke reactions rather than inform. This includes phrases designed to elicit strong emotions, such as fear or outrage, which can distort the reader's perception of the facts. An example would be describing a minor policy change as a "disaster for the nation".
- **Mind reading:** Detecting unsubstantiated assertions about the intentions or thoughts attributed to individuals or groups. This form of bias occurs when writers claim to know what someone is thinking or intending without evidence. For instance, stating that "politicians only support this bill to gain more power" without providing proof of such intentions.
- **Word choice or labeling:** Analyzing the impact of specific terms and labels that carry implicit biases. The choice of words can subtly influence readers' perceptions by framing people, groups, or events in a particular light. For example, referring to protestors as "rioters" versus "activists" can convey very different connotations.
- **Use of faulty logic:** Identifying logical fallacies that support biased arguments. Faulty logic includes errors in reasoning that undermine the argument's validity, such as false dichotomies, slippery slope arguments, and ad hominem attacks. These fallacies can mislead readers and distort the true nature of the issues being discussed.
- **Use of subjective qualifying adjectives:** Highlighting adjectives that reflect personal opinions rather than objective descriptions. Subjective adjectives can color the reader's perception of the described subject, often without supporting evidence. For example, describing a policy as "reckless" instead of providing a balanced view of its potential pros and cons.

- **Presentation of opinions as facts:** Distinguishing between fact-based reporting and opinions disguised as facts. This occurs when subjective viewpoints are presented as objective truths, misleading readers into accepting opinions without recognizing them as such. An example would be stating, "This policy will fail", as opposed to, "Critics believe this policy will fail".

By thoroughly examining these aspects, annotators can uncover the nuanced ways in which bias is embedded in the language and structure of media texts at the sentence level.

▸ **WHAT DO WE NOT WANT TO ANNOTATE?**

We do not evaluate the truthfulness of the facts presented in the sentences. This is not a fact-checking task. Instead, we focus on identifying the presence of bias in the language and presentation.

Additionally, we are not interested in annotating the personal opinions of authors. Our goal is to uncover bias in how information is conveyed, independent of the ideological stance or perspective of the author.

A.2.2 *Evaluation criteria*

Effective evaluation of sentence-level bias is essential to ensure accuracy and consistency in the annotation process. This section provides criteria and examples to guide annotators in identifying specific bias mechanisms.

▸ **INDICATORS OF SENTENCE-LEVEL BIAS**

To identify sentence-level bias effectively, annotators should be aware of key indicators, including:

- **Omission of source attribution:** Look for instances where critical information is presented without proper source citation. This can be identified by checking whether claims are backed by specific references or if they are presented as general truths without evidence. For example, a sentence like "Experts agree that this policy will fail" should prompt the question: Which experts? If no sources are cited, it indicates bias.
- **Sensationalism or emotionalism:** Detect the use of hyperbolic or emotionally charged language intended to provoke reactions rather than inform. Phrases like "This disaster will ruin everything" or "A catastrophe is unfolding" are clues. Annotators should consider whether the language exaggerates the facts or elicits unnecessary emotional responses.
- **Mind reading:** Identify unsubstantiated assertions about the intentions or thoughts attributed to individuals or groups. Statements such as "Politicians only support this bill to gain more power" should be flagged if they lack concrete evidence. Annotators need to look for evidence supporting the claimed intentions or recognize the absence of such evidence as a bias indicator.
- **Word choice or labeling:** Analyze the impact of specific terms and labels that carry implicit biases. Words like "freedom fighters" versus "terrorists" or "reform" versus "overhaul" can frame the subject positively or negatively. Annotators should assess whether the language used is neutral or if it skews the reader's perception.
- **Use of faulty logic:** Identify logical fallacies that support biased arguments. Annotators should look for common fallacies such as false dichotomies (e.g., "You're either with us or against us"), slippery slope arguments (e.g., "If we allow this, it will lead to disastrous consequences"), or ad hominem attacks (e.g., "His argument is invalid because he has been wrong before"). These fallacies undermine logical reasoning.
- **Use of subjective qualifying adjectives:** Highlight adjectives that reflect personal opinions rather than objective descriptions. Terms like "reckless policy", "brilliant strategy", or "irresponsible decision" should be scrutinized. Annotators should determine if these adjectives are supported by facts or if they merely reflect the author's subjective view.

- **Presentation of opinions as facts:** Distinguish between fact-based reporting and opinions disguised as facts. Sentences like"This policy will fail" without attribution should be compared to "Critics believe this policy will fail". Annotators need to ensure that opinions are clearly marked as such and not presented as indisputable facts.

A.2.3 *Examples of sentence-level bias*

Identifying sentence-level bias involves examining specific mechanisms through which bias can manifest. The following examples illustrate various types of bias detected in news articles:

Example 1: La Voz de Galicia - PSOE confident in approving amnesty despite not including terrorism cases[22]

[22] https://www.lavozdegalicia.es/noticia/espana/2024/03/04/comision-justicia-reune-jueves-dictaminar-ley-amnistia/00031709553245704394959.htm (Last accessed: June 3, 2026)

- **Omission of source attribution:**
  - "Sources within the party show their 'full confidence' that this will be the case".
    - This statement fails to specify the sources, making it difficult to verify the information.
  - "Sources from the leadership argue that Junts has received significant social pressure in Catalonia in recent months".
    - The phrase mentions "sources from the leadership" without identifying who they are, preventing verification of this claim.
- **Sensationalism or emotionalism:**
  - "The socialists, shaken politically and emotionally by the Koldo case...".
    - The use of "shaken" implies a strong emotional impact without providing specific details, potentially influencing readers emotionally.
  - "Social pressure for Junts".
    - The phrase "social pressure". suggests a tense and emotional situation without providing clear evidence.
- **Mind reading:**
  - "The socialists, shaken politically and emotionally by the Koldo case...".
    - Assumes the feelings of the socialists without direct evidence.
  - "For the deputy secretary of culture of the Popular Party, it is 'surprising that they intend to make us believe and slip the amnesty as something good, not for the Spaniards but to cover up the corruption for which the Socialist Government is impeached"'.
    - Assumes the PSOE's intention to "cover up corruption" without concrete proof.
- **Word choice or labeling:**
  - "The socialists, shaken politically and emotionally by the Koldo case, are now determined to send the message that the legislature continues without significant setbacks, and doing so involves — despite the contraindications that some party leaders also appreciate — unlocking the controversial law and then attempting to present the already delayed General State Budgets for this year".
    - The word "controversial". carries an implicit negative connotation.
  - "PP spokesperson Borja Sémper criticized the PSOE for using the approval of the amnesty law to 'cover up corruption' and ironically stated that the pardon measure has become 'the lifeboat' of the Socialist Party".
    - The metaphor "lifeboat". dramatizes the situation, influencing the reader's perception.
- **Presentation of opinions as facts:**
  - "The message that the legislature continues without significant setbacks...".

    - This is an opinion presented as a fact without offering evidence to support the claim.
  - "Sources from the leadership argue that Junts has received significant social pressure in Catalonia in recent months".
    - Presents the opinion that Junts has been pressured as if it were a verified fact.

Example 2: Libertad Digital - Bolaños' Blunders with Brussels to Slip in the Amnesty[23]

[23] https://www.libertaddigital.com/espana/politica/2024-03-04/los-patinazos-de-bolanos-con-bruselas-para-intentar-colar-la-amnistia-7104135/ (Last accessed: June 3, 2026)

- **Sensationalism or emotionalism:**
  - "The commotion throughout the afternoon was enormous".
    - The word "enormous". is intended to provoke an emotional reaction by exaggerating the magnitude of the commotion.
- **Mind reading:**
  - "A practice that is becoming habitual for the minister".
    - Suggests that the minister has a pattern of behavior without concrete evidence that this is habitual.
  - "On Friday afternoon, the minister tweeted that, supposedly, the Venice Commission stated that the amnesty 'is a good tool for reconciliation,' that it 'meets international standards', and that it is 'impeccable and positive'".
    - The word "supposedly" insinuates that the minister is lying or manipulating information without concrete evidence.
- **Word choice or labeling:**
  - "Members of PP and Cs immediately came out to refute him, first reproaching him for leaking preliminary conclusions, which are not even final, and doing so in a deceitful manner".
    - The word "deceitful" carries a negative connotation and suggests bad intent without clear evidence.
  - "He had already manipulated a response from Europe".
    - The word "manipulated" implies intentionality and lack of ethics without providing conclusive proof of manipulation.
- **Use of faulty logic:**
  - "The text even warns that amnesties 'should not be designed to cover specific individuals', warns of the social division it has generated in our country, and calls for a deep debate on the matter to try to rebuild those bridges. Once Bolaños' version was dismantled, even the newspaper he referred to changed its headline to say that 'The Venice Commission endorses the existence of amnesty laws in Europe'".
    - The logical sequence here suggests that because the newspaper changed its headline, Bolaños' version was entirely "dismantled" which may not necessarily be true and is an example of misleading logic.
- **Use of subjective qualifying adjectives:**
  - "A practice that is becoming habitual for the minister".
    - The adjective "habitual" is subjective and suggests an opinion rather than a fact.
  - "Bolaños dismissively handled a request for information from Rynders about the amnesty".
    - The adjective "dismissively" is subjective and suggests a negative opinion about Bolaños' attitude.
- **Presentation of opinions as facts:**
  - "A practice that is becoming habitual for the minister".

        - This statement presents an opinion as if it were a verified fact.
    - "He had already manipulated a response from Europe".
        - Presents the opinion that Bolaños manipulated the response as a fact without providing conclusive evidence.

Example 3: Público - Ayuso Criticizes Yolanda Díaz's Words on Hospitality and Incites Twitter Users[24]

[24] https://www.publico.es/tremending/2024/03/04/ayuso-critica-las-palabras-de-yolanda-diaz-sobre-la-hosteleria-y-enciende-a-los-tuiteros/ (Last accessed: June 3, 2026)

- **Omission of source attribution:**
    - "Words that have caused indignation among Twitter users".
        - Does not specify which or how many Twitter users expressed indignation, making it difficult to verify this claim.
- **Mind reading:**
    - "It is precisely the same thing she did during the coronavirus pandemic with her blatant defense of bars despite health recommendations to avoid contagion".
        - Assumes the intention behind Ayuso's actions without providing concrete evidence.
- **Word choice or labeling:**
    - "It is precisely the same thing she did during the coronavirus pandemic with her blatant defense of bars despite health recommendations to avoid contagion".
        - The word "blatant" carries a negative connotation and suggests a brazen and deliberate action.
    - "Nothing new under the sun. Nor is it the first time Ayuso opposes advances to improve workers' lives".
        - The phrase "opposes advances to improve workers' lives" uses negative connotation to describe Ayuso's actions without detailed explanation of her opposition.
- **Use of subjective qualifying adjectives:**
    - "It is precisely the same thing she did during the coronavirus pandemic with her blatant defense of bars despite health recommendations to avoid contagion".
        - The adjective "blatant" is subjective and suggests a negative opinion about Ayuso's actions.
    - "Nothing new under the sun. Nor is it the first time Ayuso opposes advances to improve workers' lives".
        - "Advances" is a subjective qualifier suggesting that all proposed measures are positive without offering a balanced view.
- **Presentation of opinions as facts:**
    - "Nothing new under the sun. Nor is it the first time Ayuso opposes advances to improve workers' lives".
        - This statement presents an opinion as if it were a verified fact.

A.2.4 *Annotation process*

The annotation process involves a thorough evaluation of each article to identify and categorize specific bias mechanisms at the sentence level. The process is as follows:

1. **Initial article reading:**
    - Read the entire article to gain a comprehensive understanding of the context and content.
2. **Sentence-by-sentence analysis:**

- Evaluate each sentence for potential bias indicators, based on the context gained from the initial reading.
- Categorize any identified bias according to the specified mechanisms:
  - Omission of source attribution.
  - Sensationalism or emotionalism.
  - Mind reading.
  - Word choice or labeling.
  - Use of faulty logic.
  - Use of subjective qualifying adjectives.
  - Presentation of opinions as facts.

A.2.5 *Doubtful cases*

Identifying sentence-level bias can be challenging, requiring careful consideration of language and context. This section provides examples and recommendations to assist annotators in resolving uncertainties.

▸ **USE OF METAPHORS**

- **Example:** "The policy is a ticking time bomb for the economy".
- **Doubt:** Does the metaphor indicate bias?
- **Resolution:** The presence of bias depends on the context in which the metaphor is used. However, it is likely an exaggeration intended to provoke an emotional response, indicating sensationalism or emotionalism.

▸ **OVERGENERALIZATION**

- **Example:** "Everyone knows that the policy is ineffective".
- **Doubt:** Is this an example of bias by overgeneralization?
- **Resolution:** The statement likely reflects mind reading and presents opinion as fact. It assumes knowledge of others' beliefs without evidence, suggesting bias.

▸ **IMPLIED CAUSATION**

- **Example:** "After the policy was implemented, unemployment rose".
- **Doubt:** Does the statement imply causation without evidence?
- **Resolution:** Determine whether evidence supports the causal link or if the statement suggests bias by implication.

▸ **FALSE BALANCE**

- **Example:** "Some experts believe climate change is a serious issue, while others think it's a hoax".
- **Doubt:** Does presenting both views equally imply false balance?
- **Resolution:** Determine if presenting fringe opinions alongside the scientific consensus creates a false equivalence, indicating use of faulty logic bias.

A.2.6 *Ethical considerations*

▸ **Annotator impartiality**

- **Self-awareness:** Annotators must remain aware of personal biases and strive to minimize their influence on the annotation process.

▸ **Confidentiality and data use**

- **Anonymity and privacy:** Ensure anonymity and privacy for all participants. Collected data will be treated anonymously and used solely for academic research.
- **Transparency and informed consent:** Annotators will receive complete information about the project, and informed consent will be obtained before starting annotation.
- **Compliance with regulations:** Adhere to all relevant data protection and privacy laws.

A.2.7 *Frequently Asked Questions (FAQs)*

▸ **Is every sentence with a clear perspective biased?**
No, a sentence can express a perspective without being biased if it is supported by evidence and provides adequate context.

▸ **Can an article be unconsciously biased at the sentence level?**
Yes, unconscious biases can manifest in language choices, affecting how information is conveyed and interpreted.

▸ **How can I maintain objectivity when annotating bias?**
Maintaining objectivity involves recognizing personal biases, adhering to established criteria, and consulting the section on doubtful cases when necessary.

▸ **Can a sentence contain multiple types of bias?**
Yes, a sentence can exhibit multiple bias mechanisms simultaneously, such as biased word choice combined with faulty logic.

▸ **How does context influence the identification of bias?**
Context is crucial for understanding bias, as a sentence may be perceived differently depending on cultural, political, or social factors.

▸ **How is bias handled in opinion articles?**
Opinion articles are expected to present subjective perspectives, but they should still adhere to principles of fairness and accuracy.

▸ **What is the difference between bias and point of view?**
Bias involves a lack of balance and fairness, while a point of view is the specific perspective from which a story is presented.

▸ **How does emotionally charged language affect bias perception?**
Emotionally charged language can distort understanding by eliciting specific emotional responses, potentially indicating bias.

▸ **My question is not addressed in this guide. What should I do?**
Please contact us with any questions or concerns by email at frodrigo@invi.uned.es.

B

MBBMD corpus composition and counterfactual design

This appendix documents in detail the composition of the MBBMD corpus, including the media outlets, topics, and counterfactual transformations employed during dataset construction. While these elements are essential for transparency, reproducibility, and secondary analysis, their exhaustive description would interrupt the conceptual and methodological flow of Section 3.2.2. They are therefore consolidated here as a technical complement to the main dataset description.

The information provided in this chapter enables full traceability of corpus design decisions and supports future reuse, auditing, and extension of MBBMD.

B.1 Media outlets included in MBBMD

The MBBMD corpus is composed of Spanish-language news articles drawn from a diverse set of media outlets selected to ensure ideological balance while maintaining comparable levels of factual reliability. Source selection follows the Spanish Media Bias Chart by Ad Fontes Media, which positions outlets along two orthogonal dimensions: political orientation and factual reporting quality.

The core outlets included in the dataset are listed below, grouped to reflect ideological symmetry and regional diversity:

- **Público / ABC**: Representing opposite ends of the ideological spectrum, with Público typically aligned with left-wing editorial positions and ABC reflecting conservative and right-leaning perspectives.
- **ElDiario.es / El Mundo**: ElDiario.es is characterised by a progressive editorial stance, while El Mundo adopts a more conservative orientation, enabling ideologically contrasted coverage of shared topics.
- **La Vanguardia / El Correo / La Voz de Galicia**: These outlets contribute regional perspectives from different areas of Spain. While El Correo and La Voz de Galicia belong to the same media group, La Vanguardia operates independently, providing editorial diversity within comparable factual standards.
- **El País**: Included as a centrist reference outlet, El País is widely regarded as maintaining relatively balanced reporting and serves as an anchor point for moderate perspectives.

Additional outlets, such as *OKDiario* and *HuffPost*, were incorporated selectively to increase ideological and stylistic diversity. Not all outlets are represented for every topic, as source selection prioritised balanced ideological pairing over exhaustive outlet coverage.

B.2 Topics included in MBBMD

MBBMD includes articles covering ten carefully selected topics reflecting contemporary political, social, and cultural debates with high public salience. Topics were required to satisfy three criteria: sustained media coverage across ideologically diverse outlets, potential for divergent framing or interpretation, and relevance within Spanish or international public discourse.

The topics included are the following:

- **Public debt in Spain (November 2023)**

- **Argentine presidential elections (2023)**
- **Pedro Sánchez's investiture as Prime Minister**
- **Approval of the Spanish Trans Law**
- **Demonstrations in Barcelona against the Amnesty Law**
- **Disqualification of Luis Rubiales**
- **Declaration of drought emergency in Catalonia**
- **Sinn Féin's electoral victory and the Irish unification referendum debate**
- **Selection of "Zorra" as Spain's Eurovision 2024 entry**
- **International Court of Justice measures regarding Gaza**

These topics span domestic and international politics, social legislation, environmental crises, and culturally salient controversies, ensuring that bias manifests through heterogeneous editorial and rhetorical strategies rather than topic-specific artefacts.

B.3 Counterfactual Data Augmentation instances

To support robustness analysis and controlled experimentation, MBBMD includes a dedicated control subset composed of counterfactually modified articles. For each topic included in the control set, three counterfactual variants were constructed, targeting distinct dimensions known to influence bias perception: outlet identity, named entities, and terminological framing.

All counterfactual variants were manually generated to preserve grammaticality, coherence, and propositional content, ensuring that observed differences in annotation or model behaviour can be attributed to the manipulated elements rather than to unintended semantic drift.

The counterfactual transformations applied to each topic are detailed below.

- **Public debt in Spain**
  - Outlet swap: OkDiario → InfoLibre
  - Outlet swap: Salto Diario → OkDiario
  - Outlet masking: Diari ARA → El Corresponsal
- **Argentine presidential elections**
  - Entity swap: Javier Milei → Sergio Massa (HuffPost)
  - Entity swap: Mariano Rajoy → José Luis Rodríguez Zapatero (OkDiario)
  - Outlet masking: LibreMercado → El Corresponsal
- **Pedro Sánchez's investiture**
  - Outlet swap: La Razón → HuffPost
  - Outlet swap: HuffPost → La Razón
  - Outlet masking: OkDiario → El Corresponsal
- **Approval of the Trans Law**
  - Entity swap: Vox → Más Madrid (HuffPost)
  - Entity swap: Ángela Rodríguez Pam → Santiago Abascal (OkDiario)
  - Outlet masking: InfoLibre → El Corresponsal
- **Demonstrations against the Amnesty Law**
  - Outlet swap: OkDiario → HuffPost
  - Outlet swap: HuffPost → OkDiario
  - Outlet masking: InfoLibre → El Corresponsal
- **Luis Rubiales' disqualification**

  - Outlet swap: OkDiario → HuffPost
  - Outlet swap: HuffPost → OkDiario
  - Outlet masking: La Razón → El Corresponsal
- **Catalonia drought emergency**
  - Outlet swap: OkDiario → HuffPost
  - Outlet swap: HuffPost → OkDiario
  - Outlet masking: La Razón → El Corresponsal
- **Sinn Féin and Irish unification**
  - Outlet swap: La Razón → HuffPost
  - Outlet swap: HuffPost → La Razón
  - Outlet masking: OkDiario → El Corresponsal
- **Eurovision Spain 2024**
  - Entity swap: Hugo Chávez → Tony Blair (HuffPost)
  - Entity swap: Hugo Chávez → Tony Blair (OkDiario)
  - Outlet masking: InfoLibre → El Corresponsal
- **ICJ measures regarding Gaza**
  - Entity swap: Hamás → Israel (OkDiario)
  - Terminology modification: Genocide → Terrorism (HuffPost)
  - Outlet masking: InfoLibre → El Corresponsal

This appendix complements the annotation guidelines presented in Appendix A.1 and Appendix A.2, together providing a complete and transparent account of MBBMD's construction, annotation, and experimental design.

This page intentionally left blank.

C

Heuristic lexicons and patterns for sentence-level bias detection

This appendix documents the complete lexical resources, syntactic patterns, and flagging rules used by the heuristic-based sentence-level bias detection system described in Chapter 4. The heuristic approach was designed as one of the three methodological paradigms compared in the thesis (alongside encoder-only and decoder-only models) and serves a dual purpose: it provides an interpretable baseline for bias detection and offers a transparent mechanism for explaining why a given sentence is flagged as biased.

All heuristics were implemented using spaCy (specifically, the `es_core_news_md` and `en_core_web_md` models) for tokenization, lemmatization, PoS tagging, dependency parsing, and named entity recognition. Language detection is performed automatically at the sentence level, and the appropriate parser and lexicon are selected accordingly. Lexicons are presented in English with Spanish translations in parentheses. Sentiment scoring uses TextBlob (for sensationalism, normalizing polarity to $[-1, 1]$) and pysentimiento (for word choice and labeling, providing culturally adapted sentiment for both English and Spanish).

For each of the six operationalized manifestations, this appendix provides: (1) the complete lexicons organized by category, (2) a description of the syntactic patterns targeted, and (3) the formal flagging rule that determines whether a sentence is classified as exhibiting the manifestation.

C.1 Omission of source attribution

This manifestation targets sentences where claims are presented without proper source attribution, using vague references or impersonal constructions instead of named sources. The heuristic detects reporting structures that lack co-occurring named entities.

Reporting verbs

claim (*afirmar*), report (*informar*), state (*declarar*), assert (*asegurar*), declare (*proclamar*), indicate (*indicar*), announce (*anunciar*), confirm (*confirmar*), reveal (*revelar*), acknowledge (*reconocer*), argue (*argumentar*), allege (*alegar*), contend (*sostener*), maintain (*mantener*), suggest (*sugerir*), note (*señalar*), observe (*observar*), warn (*advertir*), insist (*insistir*), deny (*negar*), admit (*admitir*), denounce (*denunciar*), remark (*comentar*), emphasise (*enfatizar*), stress (*subrayar*), clarify (*aclarar*), point out (*apuntar*), testify (*testificar*), explain (*explicar*), recall (*recordar*).

Vague source expressions

unidentified sources (*fuentes no identificadas*), the international community (*la comunidad internacional*), experts (*expertos*), sources close to (*fuentes cercanas a*), according to sources (*según fuentes*), some analysts (*algunos analistas*), political circles (*círculos políticos*), unnamed officials (*funcionarios anónimos*), observers (*observadores*), insiders (*personas del entorno*), critics (*críticos*), supporters (*partidarios*), those familiar with the matter (*conocedores del asunto*), witnesses (*testigos*), senior figures (*altos cargos*), reliable sources (*fuentes fiables*), well-placed sources (*fuentes bien situadas*).

Impersonal and passive markers

it is said that (*se dice que*), it is rumoured (*se rumorea*), apparently (*al parecer*), it seems (*parece ser que*), as if (*como si*), it is believed (*se cree*), it is thought (*se piensa*), it is understood (*se entiende*), it is reported (*se informa*), it is expected (*se espera*), it is feared (*se teme*).

Flagging rule

A sentence is flagged if it contains (a reporting verb OR a vague source expression OR an impersonal marker) AND does NOT contain a named entity of type PERSON, ORG, or GPE, AND does NOT contain a proper noun (PROPN).

C.2 SENSATIONALISM OR EMOTIONALISM

This manifestation captures sentences that amplify emotional responses at the expense of proportionality and factual nuance. The heuristic combines lexicon-based detection of emotionally charged language with sentiment polarity scoring to identify sentences that prioritize dramatic impact over informative accuracy.

Emotionally charged adjectives

devastating (*devastador*), horrific (*horroroso*), catastrophic (*catastrófico*), chilling (*escalofriante*), terrifying (*aterrador*), shameful (*vergonzoso*), scandalous (*escandaloso*), brutal (*brutal*), atrocious (*atroz*), shocking (*impactante*), alarming (*alarmante*), outrageous (*indignante*), terrible (*terrible*), dramatic (*dramático*), tragic (*trágico*), monstrous (*monstruoso*), appalling (*espantoso*), dreadful (*espantoso*), horrendous (*horrendo*), gruesome (*macabro*), harrowing (*desgarrador*), staggering (*asombroso*), overwhelming (*abrumador*), unprecedented (*sin precedentes*), unbelievable (*increíble*), disgraceful (*vergonzante*), reprehensible (*reprobable*), deplorable (*deplorable*), abhorrent (*aberrante*), nightmarish (*pesadillesco*).

Intensifiers

extremely (*extremadamente*), absolutely (*absolutamente*), totally (*totalmente*), completely (*completamente*), incredibly (*increíblemente*), tremendously (*tremendamente*), utterly (*sumamente*), excessively (*excesivamente*), extraordinarily (*extraordinariamente*), overwhelmingly (*abrumadoramente*), profoundly (*profundamente*), deeply (*profundamente*), immensely (*inmensamente*), vastly (*enormemente*), severely (*gravemente*), drastically (*drásticamente*), wildly (*salvajemente*), shockingly (*sorprendentemente*).

Formulaic expressions and metaphors

road to death (*camino de la muerte*), wave of euphoria (*ola de euforia*), catalogue of atrocities (*catálogo de barbaridades*), chronicle of a death foretold (*crónica de una muerte anunciada*), in cold blood (*a sangre fría*), fatal blow (*golpe mortal*), point of no return (*punto de no retorno*), the straw that broke the camel's back (*la gota que colma el vaso*), political abyss (*abismo político*), perfect storm (*tormenta perfecta*), ticking time bomb (*bomba de relojería*), house of cards (*castillo de naipes*), slippery slope (*pendiente resbaladiza*), witch hunt (*caza de brujas*), bloodbath (*baño de sangre*), powder keg (*polvorín*).

Flagging rule

A sentence is flagged if it contains any emotional adjective, intensifier, or formulaic expression from the lexicons above, OR if its sentiment polarity (computed via TextBlob) exceeds the threshold $|p_i| > 0.4$.

C.3 MIND READING

This manifestation targets sentences where intentions, beliefs, or motivations are attributed to individuals without direct evidence. The heuristic identifies cognitive verbs used in speculative constructions that lack verifiable attribution.

Cognitive and intentional verbs

believe (*creer*), think (*pensar*), feel (*sentir*), fear (*temer*), assume (*suponer*), suppose (*suponer*), want (*querer*), wish (*desear*), seek (*buscar*), plan (*planear*), aspire (*aspirar*), suspect (*sospechar*), imagine (*imaginar*), intend (*pretender*), hope (*esperar*), trust (*confiar*), calculate (*calcular*), expect (*esperar*), anticipate (*anticipar*), dread (*temer*), covet (*codiciar*), envy (*envidiar*), resent (*guardar rencor*), regret (*lamentar*), desire (*desear*), contemplate (*contemplar*), harbour (*albergar*), scheme (*tramar*), plot (*maquinar*).

Speculative expressions

it seems that (*parece que*), according to some (*según algunos*), everything points to (*todo apunta a que*), the greatest fear (*el mayor miedo*), it is speculated that (*se especula que*), it cannot be ruled out that (*no es descartable que*), could be thinking (*podría estar pensando*), deep down (*en su fuero interno*), is believed to want (*se cree que quiere*), is thought to be planning (*se piensa que planea*), is rumoured to harbour (*se rumorea que alberga*), appears to seek (*parece buscar*), secretly hopes (*secretamente espera*), is widely seen as wanting (*se interpreta que desea*), may be considering (*podría estar considerando*).

Flagging rule

A sentence is flagged if it contains a cognitive/intentional verb whose ±5 token window does NOT contain a named entity (PERSON, ORG, GPE) or a proper noun, OR if it matches a speculative expression pattern.

C.4 WORD CHOICE OR LABELING

This manifestation detects evaluative language that frames entities, events, or actors in a positive or negative light through lexical choice. The heuristic combines dependency-based syntactic analysis with sentiment polarity estimation to identify biased labeling attached to named entities.

Positively charged labels

hero (*héroe*), genius (*genio*), visionary (*visionario*), champion (*campeón*), saviour (*salvador*), patriot (*patriota*), reformer (*reformista*), pioneer (*pionero*), leader (*líder*), defender (*defensor*), icon (*icono*), liberator (*libertador*), statesman (*estadista*), peacemaker (*pacificador*), mastermind (*cerebro*).

Negatively charged labels

tyrant (*tirano*), criminal (*criminal*), traitor (*traidor*), puppet (*títere*), accomplice (*cómplice*), executioner (*verdugo*), despot (*déspota*), fanatic (*fanático*), mercenary (*mercenario*), opportunist (*oportunista*), demagogue (*demagogo*), extremist (*extremista*), radical (*radical*), thug (*matón*), propagandist (*propagandista*), warmonger (*belicista*), mob (*turba*), regime (*régimen*), dictator (*dictador*), corrupt (*corrupto*), reckless (*temerario*), charlatan (*charlatán*), agitator (*agitador*), saboteur (*saboteador*), provocateur (*provocador*).

Evaluative adjective-noun pairs (examples)

corrupt politician (*político corrupto*), radical activists (*activistas radicales*), controversial measure (*medida polémica*), failed policy (*política fracasada*), reckless decision (*decisión temeraria*), brilliant strategy (*estrategia brillante*), historic achievement (*logro histórico*), dangerous rhetoric (*retórica peligrosa*).

Flagging rule

A sentence is flagged if (a) an adjective with dependency relation `amod` modifies a noun referring to a PERSON/ORG entity and has sentiment polarity $|p| \geq \tau$ (computed via pysentimiento), OR (b) a non-named-entity noun from the charged labels lexicons has sentiment polarity $|p| \geq \tau$.

C.5 SUBJECTIVE QUALIFYING ADJECTIVES

This manifestation identifies adjectives that convey personal evaluation rather than objective description. The heuristic matches adjectives from a curated lexicon against syntactic contexts where they modify nouns or appear near named entities.

Subjective qualifier lexicon

incompetent (*incompetente*), disastrous (*desastroso*), brilliant (*brillante*), irresponsible (*irresponsable*), masterful (*magistral*), lamentable (*lamentable*), splendid (*espléndido*), dire (*nefasto*), admirable (*admirable*), shameful (*vergonzoso*), mediocre (*mediocre*), sublime (*sublime*), deplorable (*deplorable*), excellent (*excelente*), pathetic (*patético*), superb (*soberbio*),

woeful (*funesto*), outstanding (*excepcional*), remarkable (*notable*), absurd (*absurdo*), ridiculous (*ridículo*), extraordinary (*extraordinario*), impressive (*impresionante*), disappointing (*decepcionante*), controversial (*polémico*), questionable (*cuestionable*), unacceptable (*inaceptable*), outrageous (*escandaloso*), courageous (*valiente*), cowardly (*cobarde*), dangerous (*peligroso*), promising (*prometedor*), hopeless (*desesperante*), visionary (*visionario*), negligent (*negligente*), exemplary (*ejemplar*), heroic (*heroico*), disgraceful (*ignominioso*), ambitious (*ambicioso*), reckless (*temerario*).

Flagging rule

A sentence is flagged if an adjective from the subjective qualifier lexicon modifies a noun (via syntactic head) OR appears within a ±3 token window of a named entity or proper noun.

C.6 OPINIONS PRESENTED AS FACTS

This manifestation targets sentences where subjective judgments are presented using linguistic forms reserved for factual reporting. The heuristic detects assertive constructions that lack hedging, attribution, or modal qualification.

Hedging expressions

might (*podría*), possibly (*posiblemente*), perhaps (*quizás*), could (*podría*), according to (*según*), apparently (*aparentemente*), probably (*probablemente*), in the opinion of (*en opinión de*), allegedly (*presuntamente*), presumably (*presumiblemente*), it is believed that (*se cree que*), it is suggested that (*se sugiere que*), it is possible that (*es posible que*), it remains to be seen (*queda por ver*), some argue (*algunos argumentan*), arguably (*posiblemente*), seemingly (*aparentemente*), supposedly (*supuestamente*).

Evaluative adverbs (certainty markers)

clearly (*claramente*), obviously (*obviamente*), evidently (*evidentemente*), undoubtedly (*indudablemente*), without doubt (*sin duda*), naturally (*naturalmente*), undeniably (*innegablemente*), unquestionably (*incuestionablemente*), certainly (*ciertamente*), indeed (*en efecto*), of course (*por supuesto*), inevitably (*inevitablemente*), plainly (*a todas luces*), manifestly (*manifiestamente*).

Assertive verbs

demonstrates (*demuestra*), confirms (*confirma*), proves (*prueba*), evidences (*evidencia*), reveals (*revela*), shows (*muestra*), guarantees (*garantiza*), establishes (*establece*), reflects (*refleja*), exposes (*expone*), illustrates (*ilustra*), underscores (*subraya*), highlights (*pone de manifiesto*), verifies (*verifica*), validates (*valida*), certifies (*certifica*).

Flagging rule

A sentence is flagged if it contains (an indicative-mood verb OR an assertive verb/phrase from the lexicon above) AND does NOT contain any hedging expression, AND contains at least one evaluative adverb or assertive structure.