Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 53-64 recibido 30-11-2025 revisado 20-01-2026 aceptado 12-02-2026

Media Bias Bias-Mitigated Dataset (MBBMD): A Hierarchical, Perspectivist, and Counterfactually-Augmented Corpus for Bias Detection in Spanish News

Media Bias Bias-Mitigated Dataset (MBBMD): un corpus jerárquico, perspectivista y con aumentos contrafactuales para la detección de sesgo en noticias en español

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de-Albornoz, Laura Plaza
NLP & IR UNED, 28040 Madrid, Spain
frodrigo@invi.uned.es, {jcalbornoz, lplaza}@lsi.uned.es

Abstract: Media bias manifests through subtle editorial, discursive, and linguistic mechanisms that shape public perception without explicit falsehoods. Research on automatic media bias detection has focused largely on English resources and on binary, document-level labels, overlooking the hierarchical and perspectivist nature of bias. This article presents the Media Bias Bias-Mitigated Dataset (MBBMD), a new corpus designed to address these limitations. MBBMD integrates three annotation levels: binary and multilabel document-level bias, and fine-grained sentence-level manifestations. The annotation process combines a perspectivist document-level scheme, preserving annotator disagreement, with a deterministic sentence-level procedure. The dataset also incorporates systematic Counterfactual Data Augmentation (CDA), enabling analyses of how such factors influence perceived bias. The sentence-level component comprises 2,348 annotated sentences covering six linguistic manifestations of media bias.
Keywords: disinformation, media bias detection, perspectivist annotation, counterfactual augmentation.

Resumen: El sesgo mediático se manifiesta mediante sutiles mecanismos que moldean la percepción pública. La investigación se ha centrado mayoritariamente en el nivel documental y en recursos en inglés, pasando por alto la naturaleza jerárquica y perspectivista del sesgo. Este artículo presenta el Media Bias Bias-Mitigated Dataset (MBBMD), un nuevo corpus que aborda estas limitaciones. MBBMD integra tres niveles de anotación: binario y multietiqueta a nivel de documento, y manifestaciones de sesgo a nivel de oración. El proceso combina un enfoque perspectivista a nivel documental, preservando el desacuerdo entre anotadores, con un esquema determinista a nivel de oración. Además, incorpora técnicas sistemáticas de Counterfactual Data Augmentation (CDA), que permiten analizar cómo distintos factores influyen en la percepción del sesgo. El corpus contiene 2.348 frases anotadas que cubren seis manifestaciones lingüísticas de sesgo mediático.
Palabras clave: desinformación, detección de sesgo mediático, anotación perspectivista, aumento contrafactual.

1 Introduction

The digital revolution has transformed how news is produced, disseminated, and consumed, democratizing access while exacerbating challenges such as disinformation and media bias (Fallis, 2015). Unlike overt falsehoods, media bias subtly distorts public discourse through framing, selective omission, and rhetorical techniques (Rodrigo-Ginés, 2023), hindering policy-making and eroding trust in institutions. Recent European and international initiatives on online disinformation and media literacy underscore the need for reliable tools to detect such phenomena. However, computational resour-

ISSN 1135-5948 DOI 10.26342/2026-76-4 ©2026 Sociedad Española para el Procesamiento del Lenguaje Natural

ces remain heavily centred on English, limiting their applicability to other contexts.

Spanish-language media presents specific challenges for bias detection. Its ecosystems are shaped by political polarization, regional variation, and heterogeneous editorial practices, as observed in Spain, Argentina, and Venezuela (Gutiérrez, Sapiezynska, and Sánchez, 2014; Ordaz, 2016; Lodola and Kitzberger, 2017). Models trained on English-centric corpora often fail to generalize to these contexts, highlighting the need for datasets that capture the particularities of Spanish news. Existing resources also rely primarily on document-level majority-vote labels, masking the inherently subjective and perspectivist nature of media bias.

Recent taxonomies and systematic reviews (Rodrigo-Ginés, Carrillo-de Albornoz, and Plaza, 2024; Spinde et al., 2023) advocate hierarchical, multi-level conceptualizations of media bias. These frameworks stress that bias operates simultaneously at macro-discursive levels (e.g., intention, framing, coverage) and micro-textual ones (e.g., omission, subjective adjectives, emotionally charged language), and emphasize preserving annotator disagreement as a core element of perspectivist annotation. Despite these advances, resources that operationalize such taxonomies for Spanish news, integrate perspectivist supervision, and explicitly address annotator bias remain scarce. Outlet identity, political connotations of entities, and charged terminology all influence perceived bias, yet few datasets systematically evaluate these effects.

To address these gaps, this article introduces the *Media Bias Bias-Mitigated Dataset* (MBBMD), a hierarchical and perspectivist corpus for media bias detection in Spanish news media in Spain. MBBMD integrates three annotation layers: (i) binary document-level bias, (ii) multilabel bias categories, and (iii) fine-grained sentence-level manifestations. The process combines perspectivist document-level labels with adjudicated sentence-level supervision. Additionally, systematic Counterfactual Data Augmentation (CDA) (Zmigrod et al., 2019), including entity swaps, outlet masking, and terminology modifications, enables controlled analysis of contextual and lexical influences on perceived bias. As shown in Figure 1, the design identifies potential bias sources and applies mitigation strategies throughout the pipeline.

MBBMD comprises 100 articles across 10 salient topics, selected through a balanced methodology ensuring ideological diversity and comparable factual reliability. It includes 2,348 annotated sentences covering six linguistic manifestations of bias. By combining hierarchical annotation, perspectivist principles, and counterfactual robustness, MBBMD provides a structured resource for advancing media bias detection in Spanish and supports research on subjectivity, generalization, and fine-grained bias analysis.

2 Related Work

Media bias has been studied across political communication, journalism studies, and computational linguistics. Classical work defined it as a deviation from balanced reporting driven by editorial choices, selective coverage, or linguistic framing (Entman, 2007). Early typologies differentiated content, coverage, and statement bias (Floyd Moore, 2000), while later research highlighted subtler discursive mechanisms shaping interpretation beyond explicit falsehoods (Rodrigo-Ginés, Carrillo-de Albornoz, and Plaza, 2023).

Computational research has expanded rapidly, though mostly in English. Key datasets include AllSides, Media Bias Fact Check–derived corpora, hyperpartisan news collections, and finer-grained resources such as BASIL (Fan et al., 2019), NewsWCL50 (Hamborg, Donnay, and Gipp, 2019), and MBIC (Spinde et al., 2021). SemEval tasks on persuasion and propaganda (Piskorski et al., 2023) further enriched work on argumentation and framing. Yet these resources are monolingual, limit cross-cultural generalization, and rarely encode hierarchical relations between editorial-level phenomena (e.g., gatekeeping, selection, spin) and linguistic cues (evaluative language, framing). Most rely on majority-vote labels, collapsing disagreement and obscuring the subjective nature of bias.

Methodological trends mirror this evolution. Early systems relied on classical ML models (LR, SVM, RF) with handcrafted features such as n-grams (Qorib, Moon, and Ng, 2024), sentiment lexicons (Quijote, Zamoras, and Ceniza, 2019), LIWC categories (Hube and Fetahu, 2018), POS patterns, or quoted-speech structures (Cruz, Rocha, and Cardoso, 2019). Neural architectures (RNNs, LSTMs, attention mecha-

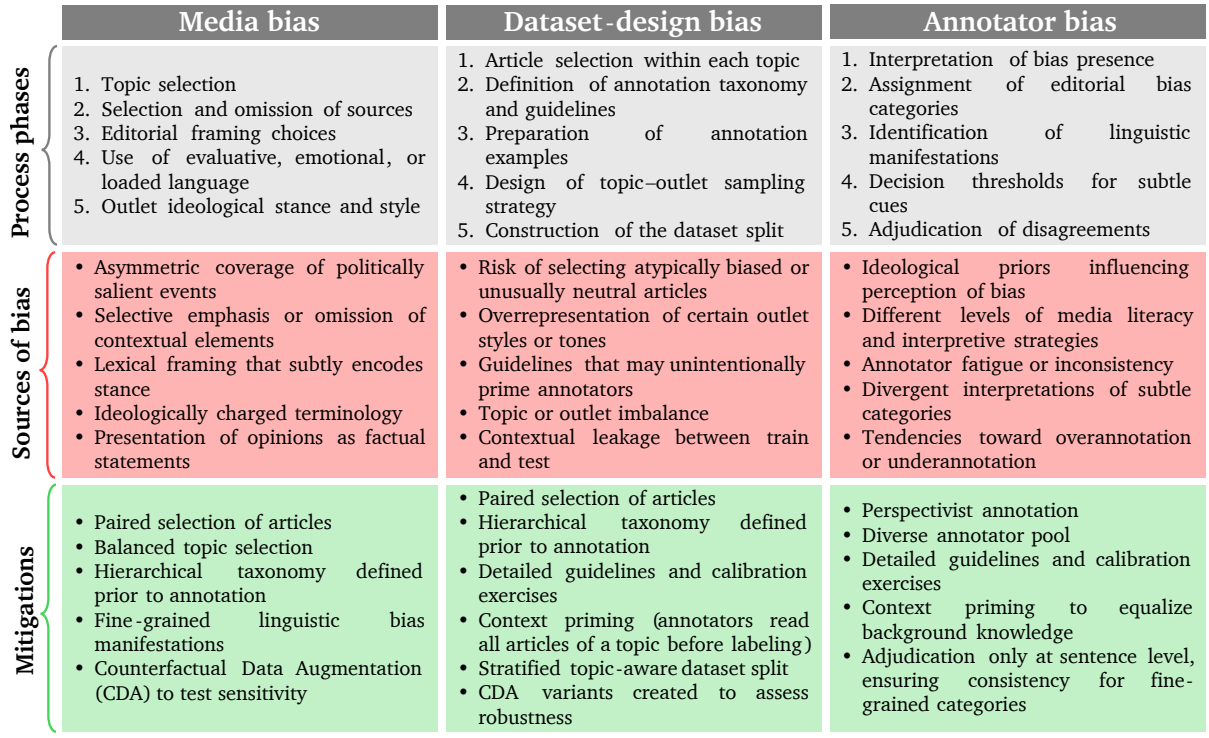

Figura 1: Sources and mitigation of bias across the MBBMD annotation pipeline. The figure summarises the stages at which bias may arise, the specific sources of bias associated with each stage, and the mitigation strategies implemented in MBBMD.

nisms) improved sentence-level representations but struggled with long-range discourse. Transformer-based encoders brought major advances in framing and bias categorization (Rakhecha et al., 2023; Ali et al., 2024), and domain-adapted or multitask variants jointly modeled framing categories and linguistic cues. Multilingual encoders extended cross-lingual applicability. Decoder-only LLMs enabled strong few-shot performance but remain unstable, often misclassifying figurative language as bias (Trhlik and Stenetorp, 2024).

Despite these advances, current approaches remain limited by datasets that fail to capture the hierarchical and perspectivist nature of media bias. Most corpora collapse disagreement into majority labels, provide coarse framing or stance categories, and do not model relationships between document-level judgments, sentence-level cues, and ideological perspectives. Recent work calls for multitask and multidimensional modelling (Horych et al., 2024; Wessel et al., 2023) and for jointly capturing editorial and linguistic strategies (Liu et al., 2023).

In Spanish, resources are scarce and typically focus on adjacent tasks such as fake news, sentiment, hate speech, stance, or political opinion mining. They do not encode bias taxonomies, rhetorical mechanisms, or ideological perspectives, nor do they establish hierarchical links between document- and sentence-level annotations. Cross-lingual transfer performs poorly due to cultural and rhetorical differences in Spanish-language journalism, reinforcing the need for native datasets.

Recent taxonomies conceptualize media bias as a hierarchical construct spanning macro-level editorial dimensions (intentional bias, spin, coverage, gatekeeping) and micro-level cues (omission of attribution, sensationalism, evaluative adjectives, labeling) (Rodrigo-Ginés, Carrillo-de Albornoz, and Plaza, 2024). These frameworks stress the importance of integrating complementary annotation layers, preserving disagreement as perspectivist signal, and incorporating robustness mechanisms sensitive to entities, sources, and charged terminology.

Research on subjectivity supports this view: politically charged tasks such as hate speech, persuasion, stance, and framing bene-

fit from multi-annotator settings that retain disagreement distributions (Uma et al., 2021; Leonardelli et al., 2023). The Learning with Disagreements paradigm formalizes this, showing that preserving annotator diversity improves robustness and fairness; however, no Spanish media bias dataset adopts this approach.

Counterfactual Data Augmentation has gained relevance for probing and mitigating bias, particularly in sentiment, toxicity, and stance classification (Vallecillo-Rodríguez et al., 2024). Yet its application to media bias remains limited, despite the established influence of outlet reputation, actor identity, and framing on perceived bias.

To our knowledge, no prior Spanish dataset combines hierarchical annotation, perspectivist principles, and systematic counterfactual augmentation within a unified framework. MBBMD fills this gap by operationalizing recent taxonomies, preserving annotator disagreement, and incorporating controlled CDA variants to analyse how entities, sources, and terminology modulate perceived media bias.

3 Corpus Design

The Media Bias Bias-Mitigated Dataset (MBBMD) has been created to address key limitations in existing media bias resources, which typically focus on document-level annotations, collapse subjectivity through majority-vote labels, and rarely operationalize hierarchical taxonomies or fine-grained linguistic manifestations. They also lack mechanisms to analyze how outlet identity, named entities, or emotionally charged terminology influence annotators' perceptions. MBBMD responds to these gaps through a unified hierarchical annotation framework incorporating perspectivist document-level labels, deterministic sentence-level adjudication, and a systematic CDA component.

Media bias operates across multiple layers of interpretation. Recent taxonomies distinguish between macro-level editorial categories (e.g., intentional bias, spin, statement bias, coverage, gatekeeping) and micro-level linguistic cues. The hierarchical structure adopted in MBBMD follows this view by explicitly separating *why* bias emerges from *how* it is linguistically realized in the text.

Figure 2 summarizes the conceptual taxonomy that guides the document-level annotations. Bias can be categorized according to the author's intention and according to contextual dimensions. These high-level categories are then linked to sentence-level manifestations such as sensationalism or omission of attribution, operationalized as concrete labels in the corpus.

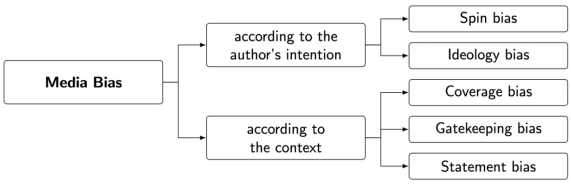

Figura 2: Hierarchical classification of media bias types based on author's intention and contextual factors. Based in (Rodrigo-Ginés, Carrillo-de Albornoz, and Plaza, 2024).

Media bias perception is also inherently subjective (Vallone, Ross, and Lepper, 1985; Elejalde, Ferres, and Herder, 2018). Annotators with different ideological, cultural, or professional backgrounds often diverge in subtle or contested cases, making perspectivist annotation necessary to preserve interpretive variability. Moreover, studies show that outlet reputation, political actors, and charged terminology significantly modulate perceived bias (Gentzkow and Shapiro, 2006; León, 2022). To account for these phenomena, MBBMD includes controlled counterfactual variants that probe the stability of human and model judgments under systematic perturbations.

3.1 Hierarchical annotation levels

MBBMD implements a hierarchical framework that mirrors the structure illustrated in Figure 2, where media bias is divided into two overarching dimensions: the author's intention and the context of presentation. The corpus operationalizes this hierarchy through three complementary annotation layers that combine perspectivist document-level labels with a deterministic sentence-level scheme.

Before annotation began, all annotators received a clear operational definition of media bias to ensure consistency across judgments. Media bias was defined as any linguistic or editorial choice that systematically favours one interpretation, actor, or viewpoint over plausible alternatives, even in the absence of factual inaccuracies. Annotators were instructed that bias may arise through se-

lective emphasis, omission of contextual elements, evaluative or emotionally loaded wording, or the presentation of opinions as factual statements.

To ensure that annotators evaluated each article with full situational awareness, the annotation process incorporated a context-priming step: before labeling any specific article, annotators were shown the complete set of ten news articles describing the same event. Only after reading them, later annotated individually, were annotators allowed to begin labeling. This procedure reduced anchoring effects and ensured that judgments reflected perceived bias in the writing rather than differences in background knowledge or article ordering, contributing to greater contextual consistency across annotators.

3.1.1 Level 1: Binary document-level bias (perspectivist)

At the first level, each article was annotated by a pool of twelve annotators with diverse ideological backgrounds. Annotators independently judged whether the article contained any form of bias (*biased / not biased*). No consensus was enforced: individual labels were preserved alongside the proportion of annotators marking the article as biased. This follows the *Learning with Disagreements* (LeWiDi) paradigm, which treats disagreement not as noise but as a signal of genuine subjective variation. The resulting distributions allow downstream users to work either with majority-vote labels or with full perspectivist information.

3.1.2 Level 2: Document-level bias categories (multilabel, perspectivist)

For any article identified as biased by at least one annotator, a second layer captures the specific editorial dimensions involved. Categories reflect the two high-level branches of the underlying taxonomy:

Author's intention:

- **Intentional bias**: perceived ideological alignment or goal-driven slant.
- **Spin bias**: rhetorical intensification, emphasis manipulation, or emotionally loaded framing.

Context of presentation:

- **Statement bias**: linguistic or discursive phrasing that subtly favours one interpretation.
- **Coverage bias**: disproportionate emphasis on specific aspects of the issue.
- **Gatekeeping bias**: systematic omission or downplaying of relevant actors, events, or viewpoints.

As in Level 1, all individual labels and their distributions are preserved, enabling analysis of agreement, ideological polarization, and topic-specific variability.

3.1.3 Level 3: Sentence-level linguistic manifestations (majority vote)

The third level focuses on fine-grained linguistic cues through a deterministic annotation process. A subset of 2,348 sentences was annotated for six manifestations of media bias. These six categories were selected based on their recurrence in the preceding analyses, as well as their prominence in contemporary taxonomic work on discourse-based bias. Together, they represent empirically observable mechanisms through which editorial and contextual bias materialise at sentence level.

1. **Omission of attribution.** Claims, evaluations, or allegations are presented without specifying who says or believes them (e.g., "Experts warn that…" without identifying the experts).
2. **Sensationalism or emotionalism.** Use of exaggerated, dramatic, or affectively charged language that amplifies the perceived gravity, urgency, or emotional impact of an event. Examples include hyperbolic verbs, catastrophic imagery, or theatrical framing.
3. **Mind reading (unwarranted inference of intentions).** Attributing motivations, intentions, or internal states to individuals or groups without evidence (e.g., "The government seeks to silence critics").
4. **Biased word choice or labeling.** Use of ideologically loaded terms or evaluative labels to describe actors, groups, or events (e.g., "regime", "radicals").
5. **Subjective qualifying adjectives.** Adjectives that encode personal or ideological judgment rather than verifiable

properties, such as "irresponsible decision" or "brilliant reform".

6. **Opinions presented as facts.** Statements that express subjective interpretation or evaluative stance while being phrased as factual reporting (e.g., "The policy is a failure").

Each sentence was independently annotated by two annotators, and disagreements were resolved through a final curation and adjudication step to produce a single gold-standard label per sentence. Unlike the document-level annotation, where perspectivism is meaningful, sentence-level labeling prioritizes reliability and consistency, making this layer suitable for training and evaluating fine-grained classification models.

3.2 Source selection

To ensure ideological balance and comparable factual reliability, the dataset uses the Spanish Media Bias Chart (Ad-Fontes-Media, 2023) as a reference for outlet selection. Sources are paired symmetrically across the ideological spectrum and matched by factual reliability so that, whenever possible, each left-leaning outlet has a right-leaning counterpart with a similar factual score. Core outlets include:

- *Público*, *ElDiario.es*, *El País*
- *ABC*, *El Mundo*, *La Vanguardia*
- *El Correo*, *La Voz de Galicia*

Additional outlets such as *HuffPost* and *OKDiario* expand stylistic and ideological diversity. A fictitious outlet, *El Corresponsal*, is used exclusively in CDA instances to decouple content from outlet reputation. Although some outlets have regional origins, article selection was restricted to nationally or internationally relevant events to avoid introducing local-topic bias.

Figure 3.2 illustrates how symmetric pairing is applied in practice. The figure zooms into a subset of outlets from the Spanish Media Bias Chart and highlights symmetric pairs (e.g., *Público–ABC*, *ElDiario.es–El Mundo*) that occupy comparable positions in terms of factual reliability but opposite ideological orientation. For each topic, articles are selected so that coverage from left and right-leaning outlets is balanced.

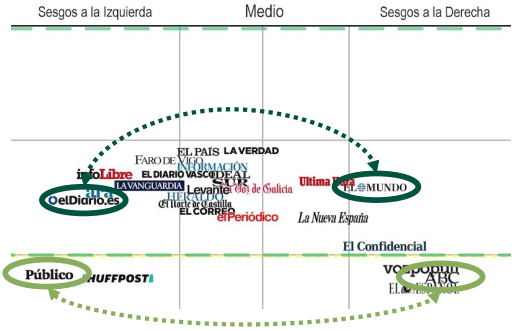

Figura 3: Some of the selected outlets to illustrate symmetric source pairing. Sources with similar factual reliability but opposite ideological orientation are paired.

The corpus includes ten politically and socially salient topics reflecting varied editorial pressures and discourse styles: Spain's public debt (2023), Argentina's 2023 presidential elections, Pedro Sánchez's investiture, the Spanish Trans Law, the Barcelona demonstrations against the Amnesty Law, Luis Rubiales' disqualification, the Catalonia drought emergency, Sinn Féin's victory and Irish unification prospects, Spain's 2024 Eurovision entry, and the ICJ ruling on Gaza.

Topics were selected to cover domestic and international politics, social policy, sports, and cultural events, and to include both highly polarized issues (e.g., Amnesty Law, Gaza) and more descriptive or policy-focused topics (e.g., public debt, drought). For each topic, at least one article is selected from a left-leaning outlet and one from a right-leaning outlet, following the symmetric pairing strategy described above.

3.3 Counterfactual Data Augmentation (CDA)

CDA is applied to systematically test the influence of outlet identity, entities, and terminology on perceived bias. For each topic, three CDA variants are created using one of the following transformations:

- **Entity swapping**, in which political actors or organizations are replaced with ideologically contrasted counterparts selected based on publicly documented party alignment and positioning in the Spanish political spectrum, and validated independently by two researchers. (e.g., *Pedro Sánchez* → *Santiago Abascal*; *Javier Milei* → *Sergio Massa*).

- **Outlet masking**, in which the original outlet is replaced either with its ideological opposite or with the fictitious *El Corresponsal*, allowing researchers to isolate the effect of source reputation.
- **Terminology modification**, in which emotionally loaded or highly charged terms are substituted with more neutral alternatives (e.g., *genocide* → *conflict*).

These variants are included only in the control subset to support robustness analyses and causal probing without contaminating the main training and test splits.

3.4 Annotation process

Annotation proceeded in two phases. Document-level annotation was carried out by a pool of twelve annotators with diverse ideological, demographic, and educational backgrounds. Articles were annotated independently, and disagreement patterns were preserved rather than resolved. This aligns with perspectivist principles and recent approaches such as Learning with Disagreements, and reflects the inherent subjectivity of high-level bias perception.

Sentence-level annotation followed a deterministic scheme. Two annotators labeled each sentence for the six linguistic manifestations of bias, and a third annotator adjudicated any discrepancies. Annotators were encouraged to provide brief rationales for decisions, creating an additional layer that can be exploited in future interpretability research.

While annotator bias cannot be fully eliminated in politically charged tasks, several mitigation strategies were implemented at the design and annotation levels. First, the annotator pool was ideologically diverse and self-reported across the political spectrum. Second, all annotators followed a structured training and calibration protocol based on shared guidelines and practice examples. Third, a context-priming step ensured that annotators read all ten articles covering the same event before labeling any individual text, reducing anchoring effects. Finally, document-level disagreement was preserved rather than collapsed into majority vote, treating interpretive variability as signal rather than noise.

Overall, this combined perspectivist–deterministic methodology reflects both the subjectivity of macro-level bias judgments and the need for reliable fine-grained supervision at the sentence level, providing a coherent framework for studying media bias hierarchically.

4 Corpus Analysis

This section provides an extended analysis of the corpus, examining its structure, the demographic and ideological composition of annotators, patterns of agreement and disagreement, and the empirical impact of the counterfactual variants. We also present baseline and comparative experiments that demonstrate how hierarchical information, perspectivist labels, and controlled counterfactuals influence model behaviour. Together, these analyses highlight the linguistic complexity and epistemic richness encoded in MBBMD.

4.1 Corpus statistics

MBBMD comprises 100 original news articles across ten politically and socially salient topics, with 2,348 manually annotated sentences (23.5 per article on average, SD = 5.2). The corpus spans diverse journalistic styles, providing a suitable context for analysing how macro-level editorial strategies align with micro-level linguistic realisations of bias.

The distribution of sentence-level manifestations is strongly asymmetric. Biased lexical choices and subjective adjectives dominate, reflecting the prominence of lexical framing in Spanish news writing. Less frequent cues such as omission of attribution or mind reading nonetheless show high discriminative value, especially in political and conflict-oriented articles. This imbalance mirrors real usage patterns: subtle rhetorical strategies typically emerge through lexical nuance rather than overt argumentative structure.

These tendencies vary across topics. Polarised events (e.g., Amnesty Law, ICJ ruling on Gaza, partisan demonstrations) exhibit a higher density of evaluative language and framing cues, whereas procedural topics (e.g., drought management, inflation indicators) rely more on descriptive structures and contain fewer explicitly biased terms. Such variation underscores the need for models capable of adapting to topic- and genre-specific linguistic behaviour.

The counterfactual component further enriches the dataset. Thirty additional articles were generated through entity swapping, outlet masking, and terminology modifica-

tion. These controlled interventions preserve propositional meaning while altering contextual or lexical cues. Preliminary analysis shows that even minimal changes systematically shift perceived bias, confirming that bias emerges not only from factual content but through its contextual framing, source identity, and affective realisation. This duality makes MBBMD valuable for robustness evaluation, causal inference, and the study of annotation-level variability.

4.2 Annotator composition

A key feature of MBBMD is its perspectivist orientation: individual judgments are preserved rather than collapsed into majority-vote labels. This approach requires a heterogeneous annotator pool capable of reflecting genuine interpretive diversity. The twelve annotators recruited for the task span a broad ideological spectrum (one far-left, six left-leaning, three centrist, two right-leaning), together with substantial demographic variability: 42 % female, ages ranging from 18–24 to 65+, and educational backgrounds covering secondary studies (17 %), vocational training (17 %), bachelor's degrees (42 %), and postgraduate studies (25 %). Annotator political orientation was self-reported using a standard left–right scale and was used exclusively for descriptive and analytical purposes.

All annotators followed a structured training protocol consisting of guideline review, practice examples, and calibration exercises based on representative cases. Training emphasised the distinction between propositional content and linguistic realisation and clarified that annotations should not involve factual verification but the identification of bias-relevant cues. Annotators also received instruction on the hierarchical taxonomy and the sentence-level manifestations to ensure consistent interpretation of the labels.

To minimise priming effects and ensure comparable situational awareness, annotators worked independently, article assignment was randomised, and topic–outlet combinations were balanced. Critically, before annotating any individual article, annotators were shown the full set of ten articles describing the same event. This "context-priming" step ensured shared background knowledge and mitigated discrepancies in interpretation arising from uneven prior exposure or familiarity with the topic.

4.3 Document-level annotation agreement

Since the goal of perspectivist annotation is not to enforce consensus but to represent interpretive variability, high agreement is neither expected nor required. Nevertheless, agreement statistics provide insight into task difficulty. For the binary bias detection task, annotators reached unanimity in 62 % of articles, corresponding to a Fleiss' κ of 0.41. For multilabel editorial categories, unanimity decreased to 48 % ($\kappa = 0{,}37$), values comparable to those reported in other ideologically charged tasks such as framing, persuasion, or hate-speech severity, where κ values typically range between 0.30 and 0.50 (Lu et al., 2025; Ruiz García, 2025). Unlike stance detection tasks (e.g., (Ding et al., 2025)), which involve explicit target positions, media bias perception requires interpretive judgment over implicit framing strategies.

Agreement varies markedly across topics. Highly polarised events, such as the ICJ ruling on Gaza or the demonstrations related to the Amnesty Law, produce the greatest divergence: the average proportion of annotators marking these articles as biased is 54–63 %, with standard deviations above 22 points. In contrast, procedural and fact-focused topics such as drought management or inflation indicators yield substantially higher consensus, with "not biased" agreement rates above 75 % and variability below 10 points. This pattern indicates that disagreement arises from socio-political salience and interpretive openness rather than annotation noise.

Analyses also reveal systematic links between document-level judgments and linguistic cues. Articles judged as biased by a large proportion of annotators contain nearly twice as many evaluative adjectives and biased labels (mean = 4.8 vs. 2.6 per article) and show a 35 % increase in loaded terminology compared to articles widely judged as "not biased". Conversely, high-consensus "not biased" articles contain proportionally more descriptive structures and exhibit a 40–45 % reduction in subjective expressions. This correlation validates the design of MBBMD: sentence-level manifestations provide the linguistic grounding of document-level perceptions.

Finally, perspectivist distributions uncover patterns that would be lost under majority-vote aggregation. Several articles

(12 % of the corpus) display bimodal distributions in which annotators split between "biased" and "not biased" in roughly equal proportions (45–55 %). These cases correspond to texts where rhetorical subtlety, framing choices, or ideological proximity legitimately allow for divergent readings. Such examples are essential for evaluating computational models intended to approximate human-like interpretive variability, as they expose epistemic complexity that deterministic labels would obscure.

4.4 CDA statistics and impact

The CDA component introduces three controlled transformations per topic (entity swapping, outlet masking, and terminology modification) each designed to preserve propositional content while altering contextual or lexical cues. These interventions generate 30 counterfactual articles (three per topic) and provide an experimental setting to isolate which dimensions of the text most strongly influence perceived bias.

Entity swapping produces the largest perceptual shifts. Across topics, articles originally classified as neutral exhibit an average increase of 17 % in perceived bias after replacing a political actor with an ideologically opposed counterpart, whereas substitutions involving an aligned actor reduce perceived bias by approximately 8 points. This asymmetry aligns with well-documented ingroup–outgroup effects: readers tend to judge discourse involving opposed actors as more biased, even when propositional content remains unchanged (Brewer, 1999).

Outlet masking also generates substantial variation. When an article is reattributed to an outlet perceived as ideologically aligned with the reader, perceived bias decreases by 7 %; when attributed to an opposed outlet, it increases by 11 %. These effects reproduce the classic source-credibility heuristic: identical content is interpreted differently depending on the political alignment and perceived trustworthiness of the source.

Terminology modification yields more moderate but consistent shifts. Replacing emotionally loaded expressions with neutral paraphrases reduces perceived bias by roughly 12 % on average. However, these effects are highly topic-dependent. In high-salience domains such as immigration, security, or international conflict, lexical substitutions account for changes of up to 18 points, while in procedural or fact-reporting topics (e.g., drought management, inflation indicators), the impact remains below 6 %.

The CDA results confirm that bias perception is not determined solely by propositional content. Instead, it emerges from the interaction between contextual cues, ideological expectations, and the linguistic realisation of the text. By encoding counterfactual variants directly in the dataset, MBBMD enables controlled causal analysis and robustness testing that are rarely feasible in existing media bias corpora.

4.5 Baseline and comparative modelling results

The modelling experiments assess the difficulty of MBBMD and the contribution of its hierarchical and perspectivist design. We evaluate two core tasks: (i) binary document-level bias detection and (ii) multiclass sentence-level prediction of six linguistic manifestations, which differ sharply in granularity and linguistic subtlety. The 100 original articles were split into 70/30 train–test partitions, stratified by topic to prevent topical leakage. CDA variants were excluded from training and used exclusively for robustness evaluation. For document-level experiments, both majority-vote labels and full annotator distributions (soft labels) were used. Sentence-level models were trained on adjudicated gold labels.

Classical models (LR, SVM, RF) trained with TF–IDF features reach 64–69 % accuracy on the document-level task but degrade to near-chance performance at the sentence level (macro-F1 $<$ 0,20), indicating that surface lexical cues are insufficient for fine-grained bias detection. Transformer encoders perform substantially better, achieving 0.74–0.77 accuracy and 0.44–0.48 macro-F1, as summarised in Table 1.

Models exploiting MBBMD's hierarchical structure outperform those trained on each level independently: a joint document–sentence XLM-RoBERTa model reaches 0.81 accuracy and 0.52 macro-F1, showing that editorial cues provide useful context for interpreting linguistic manifestations, and vice versa.

Perspectivist supervision also yields clear benefits. Soft-label models that use full annotator distributions outperform majority-

Model	Doc. Acc.	Sent. F1
Classical ML		
LR	0.64	0.18
SVM	0.67	0.19
RF	0.69	0.17
Transformer Encoders		
DistilBERT (base)	0.74	0.44
XLM-RoBERTa	0.77	0.46
Hierarchical / Perspectivist		
Hier. XLM-RoBERTa	**0.81**	0.52
Soft-label XLM-RoBERTa	0.79	**0.53**
Cross-Lingual		
XLM-RoBERTa (EN-only)	0.59	0.28
Zero-shot LLM		
GPT-5 (zero-shot)	0.66	0.32

Tabla 1: Baseline and comparative modelling results on MBBMD.

vote variants, particularly on ambiguous cases, and show higher robustness under CDA perturbations, preserving 10–15 % more correct predictions. This indicates that disagreement-aware supervision regularises decision boundaries and reduces overfitting to annotator-specific priors.

CDA stress tests reveal notable fragility in transformer models. Under entity swapping, models misclassify 20–35 % of articles differing only in one actor; under outlet masking, predictions shift by up to 15 % depending on the outlet identity. These behaviours mirror human heuristics and highlight the need for datasets that explicitly encode robustness-oriented phenomena.

Cross-lingual experiments further show that media bias does not transfer easily across cultural and linguistic contexts. XLM-RoBERTa models fine-tuned on English corpora lose up to 18 macro-F1 points when evaluated on MBBMD, performing substantially worse than Spanish-trained counterparts. This reinforces the need for native resources like MBBMD for reliable evaluation and training. Models fine-tuned on MBBMD inevitably internalize patterns present in the dataset, including its ideological distributions and annotation tendencies. While the perspectivist design mitigates over-reliance on majority labels, future work should further investigate cross-dataset generalization and robustness beyond corpus-specific patterns.

5 Discussion and Conclusions

MBBMD provides the first hierarchical, perspectivist, and counterfactually augmented resource for media bias analysis in Spanish news media in Spain. Its design is grounded in three observations: (i) media bias is multi-dimensional and cannot be reduced to a single label; (ii) annotator judgments are shaped by ideological, cultural, and experiential factors, requiring a perspectivist approach; and (iii) both humans and models are highly sensitive to surface cues such as outlet identity, named entities, and charged terminology, making CDA essential for robustness analysis.

A key contribution of MBBMD is the integration of document-level and sentence-level perspectives within a unified hierarchical framework. Unlike prior datasets that focus solely on editorial categories or fine-grained cues, MBBMD links both levels, enabling analyses of how macro-level strategies manifest through lexical framing, omission of attribution, and evaluative language. The combination of perspectivist document-level distributions and adjudicated sentence-level labels preserves interpretive variability while providing stable supervision for hierarchical and multi-task modelling.

The CDA component further strengthens the dataset. Controlled manipulations of outlets, political actors, and terminology isolate causal drivers of bias perception and expose the fragility of both human and model judgments. Even minimal changes can produce substantial perceptual shifts, highlighting models' sensitivity to surface cues. CDA also enables robustness testing and counterfactual comparisons between natural and perturbed variants.

Despite these contributions, limitations remain. The corpus covers ten salient topics within a specific socio-political context, and automated counterfactuals cannot fully capture deeper narrative or socio-pragmatic dimensions.

Overall, MBBMD provides an empirically grounded benchmark for hierarchical modelling, fine-grained bias analysis, and robustness evaluation. Its perspectivist and counterfactual design offers a structured approach to modelling media bias as a contextual and subjective phenomenon, with methodological principles transferable to other domains.

Acknowledgments

This work was supported by the Spanish Ministry of Science, Innovation and Universities (project ANNOTATE (PID2024-156022OB-C31)) funded by MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+).

References

Ad-Fontes-Media. 2023. How ad fontes ranks news sources.

Ali, O., Z. Zaland, S. U. Bazai, M. I. Ghafoor, L. Hussain, and A. Haider. 2024. Neural transformers for bias detection: Assessing pakistani news. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–7. IEEE.

Brewer, M. B. 1999. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444.

Cruz, A. F., G. Rocha, and H. L. Cardoso. 2019. 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*, pages 107–112.

Ding, Y., K. He, B. Li, L. Zheng, H. He, F. Li, C. Teng, and D. Ji. 2025. Zero-shot conversational stance detection: Dataset and approaches. *arXiv preprint arXiv:2506.17693*.

Elejalde, E., L. Ferres, and E. Herder. 2018. On the nature of real and perceived bias in the mainstream media. *PloS one*, 13(3):e0193765.

Entman, R. M. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Fallis, D. 2015. What is disinformation? *Library trends*, 63(3):401–426.

Fan, L., M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang. 2019. 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)*, pages 6343–6349.

Floyd Moore, A. 2000. The reporting verbs and bias in the press. *Revista alicantina de estudios ingleses, No. 13 (Nov. 2000); pp. 43-52*.

Gentzkow, M. and J. M. Shapiro. 2006. Media bias and reputation. *Journal of political Economy*, 114(2):280–316.

Gutiérrez, F. C., E. Sapiezynska, and R. Sánchez. 2014. Venezuela, en la prensa internacional: una cobertura sesgada. *Revista Latina de Comunicación Social*, (69):368–389.

Hamborg, F., K. Donnay, and B. Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Horych, T., M. Wessel, J. P. Wahle, T. Ruas, J. Waßmuth, A. Greiner-Petter, A. Aizawa, B. Gipp, and T. Spinde. 2024. Magpie: Multi-task media-bias analysis generalization for pre-trained identification of expressions. *arXiv preprint arXiv:2403.07910*.

Hube, C. and B. Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*, pages 1779–1786.

León, H. G. H. 2022. Media bias, competition and limited resources: The role of reputation. Master's thesis, Centro de Investigacion y Docencia Economicas (Mexico).

Leonardelli, E., A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, and M. Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). *arXiv preprint arXiv:2304.14803*.

Liu, Y., X. F. Zhang, K. Zou, R. Huang, N. Beauchamp, and L. Wang. 2023. All things considered: Detecting partisan events from news media with cross-article comparison. *arXiv preprint arXiv:2310.18827*.

Lodola, G. and P. Kitzberger. 2017. Politización y confianza en los medios de comunicación: Argentina durante el kirchnerismo. *Revista de ciencia política (Santiago)*, 37(3):635–658.

Lu, J., K. Ma, K. Wang, K. Xiao, R. K.-W. Lee, B. Xu, L. Yang, and H. Lin. 2025. Is llm an overconfident judge? unveiling the capabilities of llms in detecting offensive language with annotation disagreement. *arXiv preprint arXiv:2502.06207*.

Ordaz, L. V. 2016. El sesgo mediocéntrico del framing en españa: una revisión críti-

ca de la aplicación de la teoría del encuadre en los estudios de comunicación. *Zer: Revista de estudios de comunicación= Komunikazio ikasketen aldizkaria*, 21(41).

Piskorski, J., N. Stefanovitch, G. Da San Martino, and P. Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361.

Qorib, M., G. Moon, and H. T. Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16339–16347, Bangkok, Thailand, August. Association for Computational Linguistics.

Quijote, T., A. Zamoras, and A. Ceniza. 2019. Bias detection in philippine political news articles using sentiwordnet and inverse reinforcement model. In *IOP Conference Series: Materials Science and Engineering*, volume 482, page 012036. IOP Publishing.

Rakhecha, K., S. Rauniyar, M. Agrawal, and A. Bhatt. 2023. A survey on bias detection in online news using deep learning. In *2023 2nd international conference on applied artificial intelligence and computing (ICAAIC)*, pages 396–403. IEEE.

Rodrigo-Ginés, F.-J. 2023. Automated media bias detection: Challenges and opportunities. *PLN-DS@ SEPLN*, pages 86–94.

Rodrigo-Ginés, F.-J., J. Carrillo-de Albornoz, and L. Plaza. 2023. Identifying media bias beyond words: using automatic identification of persuasive techniques for media bias detection. *Procesamiento del Lenguaje Natural*, 71:179–190.

Rodrigo-Ginés, F.-J., J. Carrillo-de Albornoz, and L. Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Ruiz García, V. 2025. Dataset y sistemas basados en niveles de acuerdo para la detección de sexismo en tuits y memes.

Spinde, T., S. Hinterreiter, F. Haak, T. Ruas, H. Giese, N. Meuschke, and B. Gipp. 2023. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv preprint arXiv:2312.16148*.

Spinde, T., L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, and K. Donnay. 2021. Mbic–a media bias annotation dataset including annotator characteristics. *arXiv preprint arXiv:2105.11910*.

Trhlik, F. and P. Stenetorp. 2024. Quantifying generative media bias with a corpus of real-world and generated news articles. *arXiv preprint arXiv:2406.10773*.

Uma, A. N., T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Vallecillo-Rodríguez, M. E., M. V. Cantero-Romero, I. C. De Castro, A. Montejo-Ráez, and M.-T. Martín-Valdivia. 2024. Conan-mt-sp: A spanish corpus for counternarrative using gpt models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3677–3688.

Vallone, R. P., L. Ross, and M. R. Lepper. 1985. The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the beirut massacre. *Journal of personality and social psychology*, 49(3):577.

Wessel, M., T. Horych, T. Ruas, A. Aizawa, B. Gipp, and T. Spinde. 2023. Introducing mbib-the first media bias identification benchmark task and dataset collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2765–2774.

Zmigrod, R., S. J. Mielke, H. Wallach, and R. Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.