# Automated Media Bias Detection: Challenges and Opportunities

Francisco-Javier Rodrigo-Ginés[1]

[1]NLP & IR Group, UNED / Madrid, 28040, Spain

Abstract

The presence of bias in news reporting poses a significant threat to informed decision making and democratic processes. Traditional manual methods of detecting bias are limited in scalability and efficiency, necessitating automated approaches. This paper proposes research aimed at developing robust automated methods for media bias detection. The paper highlights the limitations of current approaches, emphasises the need for high-quality datasets, and explores various methodological avenues, including linguistic-based models, neural network models, and alternative approaches such as stakeholder mining and non-verbal bias analysis. The creation of a comprehensive and diverse dataset for bias detection is a key focus, addressing the limitations of existing datasets. The proposed methodology involves data collection, annotation, and validation, with a particular emphasis on addressing biases that may arise during the annotation process. The proposed experiments include training and evaluating bias detection models using the new dataset, exploring multilingual bias detection, and assessing the impact of annotator context on bias annotations. Overall, the research aims to advance automated media bias detection, contribute to media transparency, and support the functioning of democratic societies.

**Keywords**

Media bias detection, Natural Language Processing, Disinformation

## 1. Justification of the Proposed Research

The media's role in shaping public opinion and decision-making is pivotal in democratic societies, necessitating the delivery of unbiased and accurate information. The presence of bias—whether political, ideological, or commercial—can distort public perception and undermine democratic processes [1]. The digital age has exacerbated this issue, with the rapid dissemination of news via the internet and social media leading to an increase in the spread and influence of media bias.

Traditional methods of bias detection are manual, relying on expert analysis, which is time-consuming and impractical given the vast amount of content generated daily. This highlights the need for automated, efficient, and objective methods capable of handling the volume and velocity of contemporary media output. Automation promises consistency and the potential for real-time bias monitoring, which is crucial for maintaining the integrity of news dissemination.

Developing automated media bias detection systems is complex, requiring diverse datasets and sophisticated computational techniques. Approaches range from non-neural network

*Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA, 28 September 2023, Jaén, Spain.*

✉ frodrigo@invi.uned.es (F. Rodrigo-Ginés)
iD 0000-0001-6235-6860 (F. Rodrigo-Ginés)

models, employing statistical methods and handcrafted linguistic features, to neural network models like RNNs and transformers, which are adept at learning textual representations and sequences [2]. However, these systems often lack generalizability and fail to fully account for the linguistic nuances and journalistic context of bias.

Our research seeks to bridge this gap by integrating computational techniques with a deep understanding of linguistic and journalistic elements, focusing on identifying persuasive tactics and potential biases within texts. This approach not only detects bias but also provides insights into its nature and objectives, enhancing the model's applicability across various contexts and domains.

Ultimately, this work aims to bolster media transparency, accuracy, and fairness. By advancing automated bias detection, we contribute to safeguarding informed public discourse and the integrity of democratic decision-making processes.

## 2. Background and Related Work

Media bias research spans various fields such as communication science, political science, journalism, and more recently, computer science, and natural language processing. This research has primarily focused on understanding, detecting, and measuring media bias.

In the field of communication and political science, a substantial body of research has focused on understanding the causes and consequences of media bias. Research by [3] and [4] are seminal works that provided frameworks to understand the nature of media bias and its implications on society and politics. These studies employed manual, expert-based methods to analyze and measure bias in media outlets.

The field of journalism also contributed significantly to our understanding of media bias. Researchers have provided valuable insights into the nature of bias, how it manifests in news content, and the various forms it can take, such as story selection bias, framing bias, and more [5]. Manual content analysis methods were the main tools used in these studies.

In recent years, the focus has shifted towards automated detection of media bias, primarily within the field of machine learning. Early works in this area applied basic techniques such as bag-of-words representations of news content [6]. These Traditional approaches in media bias detection have employed statistical learning or machine learning techniques, such as logistic regression, support vector machines, random forests, and naive Bayes [7]. These methods usually also involve extracting handcrafted linguistic features from the text, which are then used as input to a classification algorithm trained on labeled data.

Linguistic-based methods rely on lexical, syntactic, and semantic features extracted from the text itself. Lexical features include n-grams, custom lexicons, and topics, while syntactic features encompass Part of Speech (PoS) tags and Named Entity Recognition (NER). Semantic features capture psychological cues, such as emotion, using tools such as Linguistic Inquiry and Word Count (LIWC) [8]. Reported speech-based methods focus on analysing sources quoted in the text as a means to detect media bias. By examining the use of reported speech, it is possible to identify patterns that indicate biased reporting [9]. These methods, though somewhat effective, struggled to capture the subtle and complex nature of media bias.

Deep learning methods, particularly Recurrent Neural Networks (RNNs) and Transformers, have gained prominence in media bias detection due to their ability to automatically learn feature representations from text and model the sequential structure of sentences. RNNs, including traditional RNNs and Long Short-Term Memory (LSTM) RNNs, have been widely used for media bias detection. Traditional RNNs suffer from the problem of vanishing gradients, limiting their ability to model long-term dependencies. LSTM RNNs address this issue by incorporating an internal memory [10]. Transformers, a type of neural network architecture based on self-attention mechanisms, have shown superior performance in modelling sequential structures compared to RNNs [11]. Transformers have been increasingly used in media bias detection, outperforming linguistic-based methods and RNNs [12, 13, 14].

Beyond traditional and neural network models, other research directions have explored alternative approaches to media bias detection. Some studies have investigated stakeholder mining, analysing the relationships between stakeholders and their interests [15]. Others have focused on detecting influential nodes in media outlets by building network graphs and applying community detection algorithms [16]. Additionally, nonverbal bias has been explored by analyzing emotions conveyed through news images [17].

However, the effectiveness of these methods is highly dependent on high-quality and diverse datasets. Existing datasets vary in size, diversity, and annotation levels, and there is a growing emphasis on analysing bias at the claim or sentence level. The availability and creation of robust datasets play a crucial role in advancing automated media bias detection research.

## 3. Dataset Availability

The development of high-quality, diverse datasets is essential for the effectiveness of media bias detection methods, as they depend on such datasets for training and evaluation [18, 8, 19]. These datasets, which vary in size, diversity, content nature, coverage period, and annotations, are shifting towards more granular analyses at the claim or sentence level. We draw particular inspiration from the novel DIPROMATS dataset [20], which is distinguished by its innovative taxonomy for propaganda detection in news and social media. This taxonomy provides a nuanced framework for identifying and categorizing propaganda techniques, making it a valuable model for our dataset development aimed at enhancing the detection and understanding of media bias through similar detailed linguistic and rhetorical classifications.

## 4. Description of the Proposed Research

Our research is primarily aimed at addressing the limitations present in current approaches to automatic media bias detection, with a particular emphasis on the role and impact of datasets. To accomplish this, we have embarked on an extensive journey in both understanding and operationalizing media bias for the purposes of automatic detection.

In the preliminary stages of our research, we recognized the necessity of a profound exploration into the concept of media bias. We delved into the vast body of literature across communication science, political science, and journalism to synthesize a comprehensive definition. This effort included examining the phenomenon’s complexity, its various taxonomies developed in other areas, and its interconnection with current research themes such as the detection of fallacies, cognitive biases, and the proliferation of fake news. For instance, media bias can often be a precursor or a contributing factor to the spread of disinformation, shaping narratives in a way that aligns with certain fallacies or cognitive biases.

Building on this enriched understanding, we proceeded to enumerate all possible forms of media bias, guided by the extensive body of media bias literature. This catalog of media bias forms is instrumental for the subsequent stages of our research. It serves as a guidepost, helping us create future comprehensive and diverse datasets, and informing the development of our detection algorithms.

Our definition of media bias is not limited to the overt presentation of skewed perspectives but also includes subtle and indirect forms. These encompass a wide range of manifestations, such as unsubstantiated claims bias, opinion statements presented as facts, sensationalism or emotionalism, ad hominem attacks, mind reading, slant bias, and biases in picture selection and explanation. We also consider linguistic nuances like subjective qualifying adjectives, labeling, word choice, flawed logic, and biases stemming from omission, commission, placement, size allocation, source selection, and the lack of source attribution. By embracing this comprehensive understanding of media bias, we lay the foundation for an inclusive and holistic approach to automatic bias detection.

Our latest achievement is the conceptualization and execution of a novel media bias detection cascade model. This model is engineered to address the deficiencies in existing automated media bias detection mechanisms, particularly their narrow concentration on bias at the article level. Deviating from this traditional approach, our cascade model has the capability to detect bias across multiple layers—from the broad spectrum of the entire article down to the specific sentence and claim level. The model operates by initially evaluating bias on a larger, article-wide scale, before descending into a more meticulous examination at the sentence and claim level. This innovative methodology not only coheres with our more nuanced understanding of media bias but also bolsters the accuracy and exhaustiveness of the bias detection process. This model has proved the feasibility of our idea of using the detection of linguistic and rhetorical techniques to generalize bias detection. Although we have incorporated just a handful of these techniques, expanding on them promises to result in substantial advancements in the field. While the focus on article and sentence levels might seem less pertinent, it has been explored to a lesser extent, yet provides valuable insights for future research.

Looking forward, our next major step involves the creation of a biases-mitigated media bias dataset. Despite the variety of existing datasets, they predominantly suffer from a lack of diversity and size, a narrow focus on highly-engaged and political news, and a significant language bias, with most datasets solely containing English language news. Our goal is to create a dataset that transcends these limitations.

We aim to create a dataset that is rich in diversity, encompassing a wide range of news domains beyond politics, from numerous sources across the globe, and covering a multitude of distinct events. We also intend to include news articles in languages other than English (at least Spanish), to account for the significant non-English speaking global population. Importantly, our dataset will aim to capture the variety of bias forms we have identified, and include annotations at the article, sentence, and claim level in alignment with our Media Bias Detection Cascade model.

In creating this dataset, we aim to address the limitations of current datasets, provide a robust resource for the development and evaluation of future media bias detection algorithms, and most importantly, contribute towards a more comprehensive understanding of media bias in our increasingly global and digital news landscape.

In recognition of the dynamic and evolving nature of societal norms and language, our future work will delve into the historical and temporal context of media bias. We acknowledge that what was once deemed ”normal” or socially acceptable can shift dramatically over time, with these changes mirrored in the language and presentation of news and events. To this end, we propose the development of adaptive datasets and models that are not only reflective of current linguistic and societal standards but are also designed to accommodate paradigm shifts over time.

We envision a framework where our datasets and detection models are periodically updated to align with the evolving societal consensus, ensuring their relevance and accuracy. This could involve the implementation of machine learning algorithms that are capable of learning from new data and adjusting the parameters of bias detection accordingly.

Furthermore, we are intrigued by the prospect of a retrospective comparative study. Such an endeavor would involve collecting and analyzing historical data, juxtaposing it with contemporary datasets to trace the evolution of language and bias. This comparative analysis would not only provide valuable insights into the progression of media narratives but also serve as a testament to the changing landscape of societal norms and values.

By incorporating these temporal dimensions into our research, we aim to contribute to a more nuanced and historically informed understanding of media bias, one that recognizes the fluidity of language and societal attitudes and prepares our models to adapt to the continuous march of time and social change.

## 5. Methodology and Proposed Experiments

The development of our comprehensive, biases-mitigated dataset for automatic media bias detection involves several key steps and methodological considerations. Our process begins with data collection, followed by annotation and validation. To ensure the successful implementation of our research, we have also outlined a series of proposed experiments.

### 5.1. Data Collection

In the data collection phase, we aim to gather news articles from a diverse range of sources, covering various domains and events, and written in Spanish. To achieve this diversity, we will utilize web scraping tools to extract news articles from online news platforms.

However, it is important to note that in the process of collecting a textual dataset, bias can be generated if certain precautions are not taken into account. Bias can be a distortion of reality introduced in the dataset due to various factors, including the preferences or arbitrary selection made by the dataset creator. Therefore, it is crucial to ensure that the dataset is representative of reality to avoid training and adapting the model to biased data, which can result in poor generalization to the real world [21].

To mitigate bias in our dataset, we will implement several strategies. Firstly, we will carefully select a wide range of news sources that represent diverse perspectives, ideologies, and viewpoints. This will help ensure a balanced representation of different narratives and minimize the influence of any particular bias. Additionally, we will establish clear guidelines and criteria for data collection, ensuring that articles are selected based on objective criteria such as relevance, credibility, and popularity rather than subjective judgments. Furthermore, we will prioritize transparency by providing detailed information about the sources and collection process, allowing researchers to assess and evaluate the dataset's potential biases.

### 5.2. Data Annotation and Validation

Following the data collection phase, the subsequent step involves annotation, where we recognize the potential influence of the annotator's context, including sociocultural factors, gender, age, and political identity, on their perception of media bias [22]. It is crucial to address this issue to avoid introducing biases into the annotation process, such as gender bias, ethnic bias, religious bias, and other forms of bias.

To tackle this challenge, we will conduct a comprehensive survey with annotators representing diverse profiles, and using the counterfactual data augmentation technique [23]. The survey aims to examine how the context of annotators influences their annotations. Instances to be tagged will be randomly presented to the annotators, who will label the instances based on their perception of bias, considering both the media source and the entities discussed. By analyzing the survey results, we can gain valuable insights into potential biases that may emerge during the annotation process and understand the impact of the annotator's context on the annotations.

After the annotation stage, a crucial validation phase will be initiated to confirm the robustness and precision of our dataset. This will include assessing the inter-annotator agreement, taking into account the diversity of annotator profiles, and using this disagreement as a learning [24] opportunity to refine our models and annotation guidelines.

By conducting the validation phase, we aim to provide a robust and trustworthy dataset for studying media bias detection. The validation process will ensure the dataset's quality, enhance its utility for the research community, and facilitate the development of more accurate and effective bias detection techniques.

### 5.3. Proposed Experiments

Once our dataset is ready, we propose a series of experiments to evaluate its utility and robustness. Initially, we will use the dataset to train various machine learning models for automatic media bias detection. This will help us evaluate the performance of existing models on our dataset and compare it with their performance on older datasets.

Furthermore, we will explore the potential of our dataset in enabling multilingual media bias detection. This will involve training and testing models across different languages present in our dataset. By studying bias detection in multiple languages, we can assess the generalizability of the models and the effectiveness of our dataset across diverse linguistic contexts.

By conducting these experiments, we hope to validate the utility of our dataset and detection model, understand the influence of annotator context on bias annotations, explore bias mitigation techniques, and contribute to the advancement of automatic media bias detection research.

# 6. Specific Research Elements Proposed for Discussion

There are several elements of our research that merit further discussion, both in terms of our proposed methodology as well as the broader implications of our work. Here are the key areas that we suggest for further exploration:

## 6.1. Spanish Media Bias Detection

The proposed dataset includes news articles in multiple languages, thereby creating an opportunity to extend the media bias detection studies to languages other than English. This addition introduces several new challenges such as language-specific nuances and cultural factors influencing media bias. It also requires discussion on how to adapt or develop bias detection methods that can handle multiple languages.

## 6.2. Annotation Approach

Our methodology includes manual annotation at the article, sentence, and claim level. This comprehensive approach provides a rich source of information for detecting bias, but also raises questions about the feasibility of such extensive annotation and the reliability of the results, particularly when scaled up. We propose a discussion on the efficiency of this approach, and potential strategies for improving the annotation process.

## 6.3. Balance between Bias Forms

In our research, we aim to capture a wide range of bias forms in our dataset. However, striking a balance between these different forms could be a challenging task. Too much focus on one type of bias could skew the dataset and influence the performance of detection models. Thus, a discussion on how to maintain a balanced representation of different bias forms in the dataset would be beneficial.

## 6.4. Media Bias Detection Cascade Model

Our proposed cascade model for media bias detection represents a pioneering approach designed to enhance the precision of bias identification across various media content levels. To elucidate the model's functionality and its adeptness in addressing the intricacies of bias detection, we shall consider an illustrative example.

Imagine an article that discusses a political figure's policy. At the article level, the model initially assesses the overall tone and direction of the narrative. It might detect a slant bias if the article predominantly highlights positive aspects of the policy while neglecting any criticism. Moving to the sentence level, the model scrutinizes the language used, such as the presence of subjective adjectives or ad hominem attacks that could indicate bias in individual statements. Finally, at the claim level, the model evaluates the factual basis of the claims made, checking for unsubstantiated assertions or flawed logic.

### 6.5. Generalizability of the Dataset

While our dataset aims to be as diverse as possible in terms of sources, languages, and topics, the question remains about how representative it is of global media. We encourage a discussion on how well our dataset captures the breadth and depth of media bias across different regions and cultures.

### 6.6. Ethical Considerations

Finally, any study on media bias should also consider the ethical implications of its findings. This includes a discussion on how bias detection models might be used or misused, the responsibility of researchers in communicating their findings, and the potential for such tools to influence public opinion or be exploited for propaganda.

Through discussing these elements, we aim to shed light on the challenges and opportunities in the field of automatic media bias detection, and to inspire new directions for future research.

## Acknowledgments

This research is funded by FAIRTRANSNLP-DIAGNOSIS: Measuring and quantifying bias and fairness in NLP systems, grant PID2021-124361OB-C32, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. This work has been also funded by the Ministry of Universities and the European Union through the Europeamn NextGenerationUE funds and the ”Plan de Recuperación, Transformación y Resiliencia”.

## References

[1] J. Strömbäck, In search of a standard: Four models of democracy and their normative implications for journalism, Journalism studies 6 (2005) 331–345.

[2] A. Cremisini, D. Aguilar, M. 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.

[3] T. Groseclose, J. Milyo, A measure of media bias, The quarterly journal of economics 120 (2005) 1191–1237.

[4] M. Gentzkow, J. M. Shapiro, What drives media slant? evidence from us daily newspapers, Econometrica 78 (2010) 35–71.

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

[6] C. Lin, Y. He, Joint sentiment/topic model for sentiment analysis, in: Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 375–384.

[7] A. F. Cruz, G. Rocha, H. L. 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.

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

[9] K. Lazaridou, R. Krestel, F. Naumann, Identifying media bias by analyzing reported speech, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 943–948.

[10] A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network, Physica D: Nonlinear Phenomena 404 (2020) 132306.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[12] K. Tangri, Using natural language to predict bias and factuality in media with a study on rationalization, Ph.D. thesis, Massachusetts Institute of Technology, 2021.

[13] M. Sinha, T. Dasgupta, Determining subjective bias in text through linguistically informed transformer based multi-task network, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3418–3422.

[14] T. Spinde, L. Rudnitckaia, J. Mitrović, F. Hamborg, M. Granitzer, B. Gipp, K. Donnay, Automated identification of bias inducing words in news articles using linguistic and context-oriented features, Information Processing & Management 58 (2021) 102505.

[15] T. Ogawa, Q. Ma, M. Yoshikawa, News bias analysis based on stakeholder mining, IEICE transactions on information and systems 94 (2011) 578–586.

[16] V. Patricia Aires, F. G. Nakamura, E. 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.

[17] L. Boxell, Slanted images: Measuring nonverbal media bias (2018).

[18] B. D. Horne, S. Khedr, S. 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.

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

[20] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de-Albornoz, Iván Gonzalo-Verdugo, Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 71 (2023).

[21] M. Geva, Y. Goldberg, J. Berant, Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets, arXiv preprint arXiv:1908.07898 (2019).

[22] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, L. A. Ureña-López, A survey on bias in deep nlp, Applied Sciences 11 (2021) 3184.

[23] R. Zmigrod, S. J. Mielke, H. Wallach, R. Cotterell, Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology, arXiv preprint arXiv:1906.04571 (2019).

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