



Review

A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it

Francisco-Javier Rodrigo-Ginés^{*}, Jorge Carrillo-de-Albornoz, Laura Plaza

NLP & IR Group, UNED, Madrid, 28040, Spain

ARTICLE INFO

Keywords:

Natural Language Processing (NLP)
Media bias detection
Information theory
Disinformation

ABSTRACT

Media bias and the intolerance of media outlets and citizens to deal with opposing points of view pose a threat to the proper functioning of democratic processes. In this respect, we present a systematic review of the literature related to media bias detection, in order to characterize and classify the different types of media bias, and to explore the state-of-the-art of automatic media bias detection systems. The main objectives of this paper were twofold. First, we framed information, misinformation and disinformation within a theoretical framework that allows us to differentiate the different existing misinformation problems such as us media bias, fake news, or propaganda. Second, we studied the state of the art of automatic media bias detection systems: analyzing the most recently used techniques and their results, listing the available resources and the most relevant datasets, and establishing a discussion about how to increase the maturity of this area. After doing a comprehensive literature review, we have identified and selected a total of 17 forms of media bias that can be classified depending on the context (e.g., coverage bias, gatekeeping bias, or statement bias), and on the author's intention (e.g., spin bias, or ideology bias). We also reviewed, following the PRISMA methodology, the main automatic media bias detection systems that have been developed so far, selecting 63 relevant articles, from which we extracted the most used techniques; including non-deep learning methods (e.g., linguistic-based methods, and reported speech-based methods), and deep learning methods (e.g., RNNs-based methods, and transformers-based methods). Additionally, we listed and summarized 18 available datasets for the task of automatic media bias detection. In conclusion, the current methods for automatic media bias detection are still in their infancy and there is still a lot of potential for improvement in terms of accuracy and robustness. We have proposed some future research lines that could potentially contribute to the development of more advanced techniques.

1. Introduction

In our current digital age, the abundance of information sources, both professional and non-professional, accessible via a mere internet connection is unparalleled. Yet, this vast ocean of information does not necessarily translate to a better informed society. Instead, many users often become entangled in the webs of disinformation, a significant portion of which can be attributed to *media bias*.

Two cognitive biases inherent in human cognition further exacerbate the challenge of discerning genuine information from falsehoods: (i) *naïve realism*, the predisposition to perceive our interpretations of information as objective, leading us to conclude that those who hold differing opinions are either misinformed or biased (Ross & Ward, 1996); and (ii) *confirmation bias*, which compels us towards information that resonates with our existing beliefs (Nickerson, 1998).

This phenomenon is aggravated by the fact that users are more and more isolated within what are known as *information bubbles*. These

bubbles cause the user to only access or receive information according to his personality and interests, thus enlarging the problem of confirmation bias: “an information bubble is like your own unique universe of information in which you live online. What there is depends on who you are and what you do. You do not decide what comes in and you do not see what is left out” (Pariser, 2011).

Information bubbles not only limit the information that reaches the user, but also generate group polarization (Sunstein, 2009), that is, the user not only strengthens his opinion, but it becomes increasingly polarized. This group pressure makes the user choose socially safe options when consuming and sharing information, regardless of whether it is true or not (Asch, 1951).

This growing polarization in public discourse, and the growing intolerance of citizens to deal with opposing political points of view (Bennett & Iyengar, 2008), pose a threat to the proper functioning of

^{*} Corresponding author.

E-mail addresses: frodrigo@invi.uned.es (F.-J. Rodrigo-Ginés), jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz), lpplaza@lsi.uned.es (L. Plaza).

democratic processes. Journalism is a basic key to measure the health of a democracy. In fact, the relationship between democracy and the media is often understood in terms of a social contract (Strömbäck, 2005). Given this symbiotic relationship, ensuring that media remains unbiased and true to its fundamental principles is paramount. As media serves as the primary source of information for the majority, any bias can distort the democratic dialogue.

In today's information-driven society, the significance of detecting and understanding media biases cannot be understated. Expert systems, especially those entrenched in the domain of Natural Language Processing (NLP), have emerged as critical components in the quest to dissect and understand the labyrinth of media content. These advanced computational platforms harness sophisticated algorithmic techniques, enabling them to delve deep into the nuances of media sources to identify and classify potential biases.

Furthermore, the evolution and proliferation of these expert systems are not merely restricted to detection. They are progressively being equipped to rectify biases, ensuring that the information disseminated to the public is as unbiased and genuine as possible. This ability to cleanse media content can play a transformative role in shaping public perception, reducing the spread of misinformation, and fostering a more informed citizenry.

On the user's end, these systems serve as indispensable tools, bestowing upon readers the capability to differentiate between genuine information, misinformation, and outright disinformation. By providing such clarity, these expert systems amplify the discerning capabilities of the average user, ensuring that they are not easily swayed by biased or skewed narratives.

Moreover, these advancements in NLP-driven expert systems play a pivotal role in reinforcing the media's integrity. In an age where trust in media is waning, the presence of such systems can act as a buffer, ensuring that media retains its credibility and continues to function as a cornerstone of democracy. By systematically filtering out biases and ensuring the delivery of factual content, these systems not only restore faith in journalistic practices but also fortify the role of media in ensuring a balanced and democratic society.

Building upon the significant roles that these expert systems play, this systematic review takes a deep dive into the intricacies of media bias and its adjacent problems. We utilize and further expand the theoretical framework laid out by Karlova and Fisher (2013) to effectively categorize and differentiate various challenges like media bias, fake news, propaganda, and the like. Central to our study is the cataloging of media bias types that have been identified by current research trends within the realm of information sciences. An integral part of our exploration includes a comprehensive literature survey, shedding light on existing mechanisms and systems designed to detect media bias across diverse facets: from the broader scope of entire documents to specific claims, and even encompassing the overarching tendencies of media outlets.

The main objective of this review is to analyze the current state of media bias detection and to propose future research directions. This review may be of help to researchers who are interested in knowing which are the types of media bias and how it is usually expressed, as well as to those who want to know what are the different approaches to automatic media bias detection, and what are the existing datasets that may be used to train and test systems for this task.

The paper is divided into two main parts. The first part consists of a systematic review of the literature related to media bias as a misinformation problem, with the aim of characterizing and classifying the different forms and types of media bias. This part also includes a theoretical framework that allows us to differentiate between information, misinformation and disinformation.

The second part of the paper is focused on automatic media bias detection. The main objectives of this part are the following: (i) to analyze the main techniques recently used for automatic media bias detection and to explore their results, (ii) to list the available resources

and datasets for the task of media bias detection, and (iii) to discuss potential future research lines to increase the maturity of this area.

Media bias detection is a relatively new area in the field of natural language processing (NLP). The first attempts to automatically detect media bias date back to the early 2000s (Park, Lee, & Song, 2011). However, these early methods were very limited in terms of accuracy and robustness.

In the last decade, the development of deep learning techniques has had a great impact on the field of NLP, and media bias detection is no exception. The introduction of recurrent neural networks (RNNs) (Rashkin, Choi, Jang, Volkova, & Choi, 2017) and, more recently, transformer networks (Baly, Da San Martino, Glass, & Nakov, 2020), have allowed the development of advanced techniques for media bias detection that outperform traditional methods.

To our knowledge, the only similar review article related to media bias detection is the one published by Hamborg, Donnay, and Gipp (2019). This article is focused on the different forms of media bias, and provides a general overview of the existing systems until 2019.

In contrast, we focus not only on the different forms of media bias, but also on the most used techniques and datasets for automatic media bias detection. Additionally, we propose a theoretical framework for the different misinformation problems, and a discussion about how to increase the maturity of the existing techniques.

This systematic review is organized as follows: in Section 2, we elucidate the methodology employed in conducting this review, detailing our search strategy, criteria for selection, and the process of screening and selection. In Section 3, we define media bias and discuss the different types of media bias that exist, according to the current research trend. In Section 4, we present a theoretical framework to distinguish and differentiate problems such as fake news, misinformation, disinformation, and propaganda. In Section 5, we introduce the task of automatic media bias detection, and we compare it with similar tasks. In Section 6, we present an extensive literature review to find out what mechanisms and systems exist today to detect media bias at various levels: at a document level, at a sentence level, and at a media level. In Section 7, we analyze the available datasets for the task of automatic media bias detection, identifying certain problems such as the predominance of English language datasets, a heavy focus on the political domain, and a significant influence of US news and politics. Section 8 presents the discussion and future work, and Section 9 concludes the review.

2. Methodology

This systematic review is conducted in strict adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher, Altman, Liberati, & Tetzlaff, 2011). The PRISMA guidelines provide a structured checklist and flow diagram, which are essential tools for ensuring the comprehensive and transparent reporting of systematic reviews and meta-analyses.

2.1. Search strategy

In order to conduct a thorough and systematic exploration of the literature related to media bias detection, a multi-faceted search strategy was developed. Recognizing the interdisciplinary nature of media bias detection, which bridges the domains of journalism and natural language processing (NLP), the search was concentrated on three primary academic databases: Google Scholar, Scopus, and ACL Anthology. The paper selection process, in alignment with the PRISMA methodology, is visually represented in Fig. 1.

For the searches within Google Scholar and Scopus, the capabilities of the *Publish or Perish* software (Harzing, 2010) were employed. This software facilitated the automation and optimization of the search processes. The query, formulated to encapsulate the core objectives of this research, was articulated as:

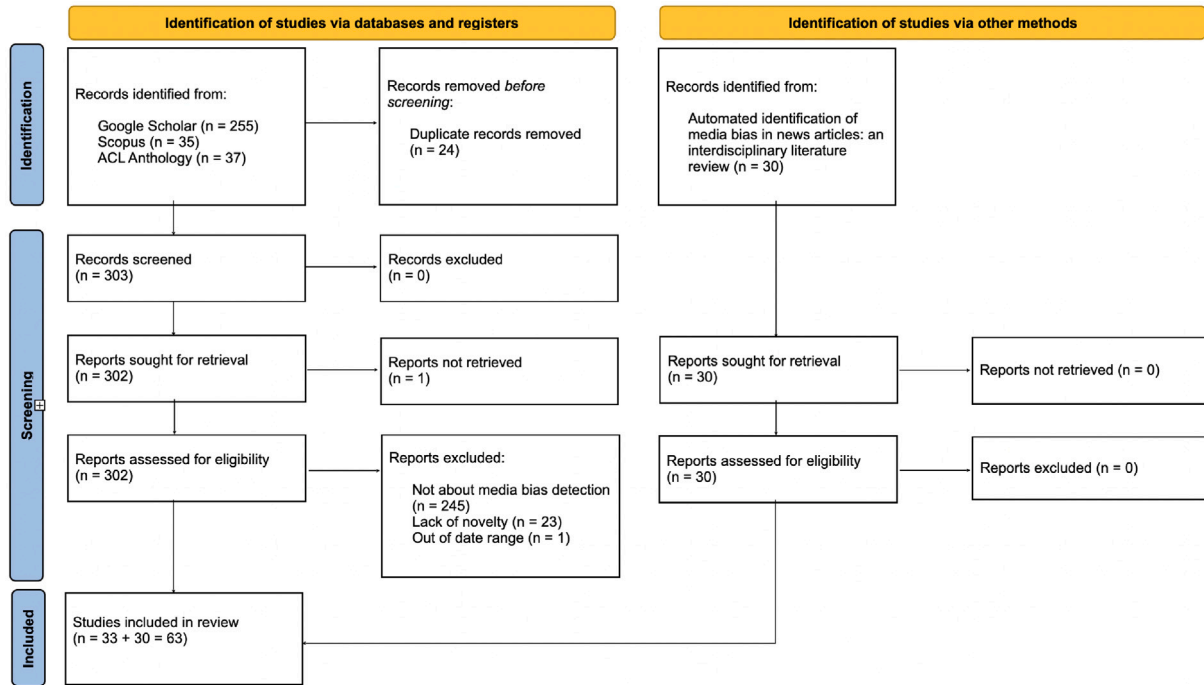


Fig. 1. A PRISMA flowchart illustrating the systematic paper selection process undertaken in this study.

(media OR journalism OR news) AND (bias OR slant OR spin) AND (detection OR characterization OR classification)

This query is the result of a meticulous process of refining and validating search terms to ensure they are both comprehensive and relevant to the domain of media bias detection, especially within the context of NLP. The formulation of this query was not arbitrary; rather, it was the culmination of a series of deliberate steps, each aimed at enhancing the precision and relevance of our search:

- 1. Preliminary research:** Our journey began with an informal review of the existing literature in the domain of media bias detection. This initial exploration was instrumental in familiarizing ourselves with the terminologies and concepts that are frequently associated with media bias detection, especially when viewed through the lens of NLP.
- 2. Iterative searches:** Armed with insights from our preliminary research, we embarked on a series of iterative searches. Each iteration was an exercise in refinement. We meticulously evaluated the results of each search, refining our terms based on their efficacy in yielding relevant results. Terms that did not produce pertinent results were either modified or discarded. This iterative process, though time-consuming, was pivotal in ensuring that our final search terms struck the right balance between comprehensiveness and specificity.
- 3. Cross-checking with known literature:** To bolster the validity of our search terms, we undertook a cross-referencing exercise. We compared the results produced by our query with a curated list of seminal articles and papers in the domain of media bias detection. This step was not merely a validation exercise but also a means to ensure that our search terms were adept at capturing the vast expanse of the field, both in terms of its breadth and depth.

The search was confined to publications from the period 2000 to 2022. Within Google Scholar, both keywords and titles were examined, whereas the search within Scopus was restricted to titles.

The ACL Anthology presented certain challenges due to its lack of support for date-specific searches. To address this, a manual collection and filtration process was undertaken, ensuring the inclusion of relevant papers within the specified date range.

2.2. Criteria for selection

To maintain the rigor of this review, explicit criteria for the inclusion and exclusion of papers were established. These criteria were pivotal in navigating the extensive literature, ensuring the selection of studies that were not only germane to the research question but also met rigorous academic standards.

2.2.1. Inclusion criteria

The criteria followed for the inclusion of papers are listed below:

- Papers with a primary focus on the detection or characterization of media bias.
- Papers written in English or Spanish.
- Publications from the period 2000 to 2022.

2.2.2. Exclusion criteria

The following exclusion criteria were adopted:

- Papers addressing related but distinct topics, such as fake news detection, stance detection, or political bias in social networks.
- Literature reviews or studies without significant novel contributions.

2.3. Screening and selection process

The initial database search yielded 255 papers from Google Scholar, 35 from Scopus, and 37 from the ACL Anthology. After eliminating 24 duplicates, a total of 303 papers were screened. Of these, one was inaccessible, 245 were deemed not directly relevant to media bias detection, 23 lacked significant novelty, and one was outside the stipulated date range.

Additionally, insights were incorporated from the survey (Hamborg, Donnay et al., 2019), that provides an overview of media bias systems up to 2019. In total, from this survey we included 30 papers.

Table 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	Characteristic
Hamborg, Donnay et al. (2019)	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 <u>sustained</u> , 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 <i>spin</i> .	(1), (2), (3)
D'Alessio and Allen (2000)	The question of media bias is moot in the absence of certain properties of that bias: It must be volitional, or willful ; <i>it must be influential, or else it is irrelevant</i> ; it must be threatening to widely held conventions, lest it be dismissed as mere “crackpotism”; and <u>it must be sustained</u> rather than an isolated incident.	(1), (2), (3)
Mullainathan and Shleifer (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 . <i>The second bias, which we refer to as spin, reflects the outlet’s attempt to simply create a memorable story.</i>	(1), (3)
Spinde, Rudnitckaia et al. (2021)	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, <i>bias can also be unintentional and even unconscious.</i>	(1), (3)
Gentzkow and Shapiro (2006)	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)

2.4. Ensuring transparency and replicability

In line with best academic practices, transparency and replicability were prioritized. A GitHub repository has been established, containing the results and the exact queries used in Publish or Perish, ensuring that other researchers can replicate the methodology and validate the findings. The repository can be accessed at [GitHub-A-systematic-review-on-media-bias-detection-PRISMArepository](#).

3. Media bias

3.1. Definition and characteristics

Differentiating in the media whether a story is biased or not is a complicated task. Journalism long ago abandoned the idea of seeking only neutrality and objectivity in pursuit of creating a more committed journalism (Boudana, 2011), which makes it more difficult to differentiate between opinion and bias.

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”. According to these definitions, we can infer that the line that separates bias from opinion depends on whether the journalist uses rhetorical artifacts that distort the information to support his opinion, or not.

We reviewed the literature in search of definitions or statements about media bias, and we found various, even opposite claims (Hamborg, Donnay et al., 2019). The biggest difference can be found in the intentionality, in this sense, most authors claim that media bias should be solely influential and intentional (Hamborg, Donnay et al., 2019; Mullainathan & Shleifer, 2002) or else it is irrelevant (D'Alessio & Allen, 2000), while other authors argue that certain degree of bias is unintentional and unavoidable as bias is a natural part of the human condition (Kang & Yang, 2022). For Bozell and Baker (1990), the media bias is not intentional, but rather depends on the background of the journalists: “though bias in the media exists, it is rarely a conscious attempt to distort the news. It stems from the fact that most members of the media elite have little contact with conservatives and make little effort to understand the conservative viewpoint”. This vision, opposed

to the basic principles of self-criticism of the deontological codes of journalism, has been criticized as media owners and their agents could take steps to prevent journalists’ personal views from biasing their reporting (Sutter, 2000), that is, even if the content was not biased in its conception, it was in its publication.

In addition to the fact that media bias always has a journalistic format, the most repeated characteristics in the definitions of media bias that we can find in the literature are: (1) that is slanted, that is, it has the intention of influencing the opinion of the receiver in a particular direction; (2) that the same bias is frequent and/or consistent for each media/journalist; and (3) that journalists may add bias in their attempt to create memorable stories.

In Table 1 we can see the different claims about media bias existing in the literature. For each definition, we highlight fragments that correspond to one of the three characteristics that we have identified.

With these basic and consensual components of bias in mind, we can define media bias as a situation in which journalists slant information in a story to influence the opinion of the receiver, that is frequent and/or consistent for each media outlet or journalist.

Also, just as we can find different definitions of media bias in the literature, we also find different classifications of it. The most popular classify media bias according to (1) the author’s intention, and (2) the context in which it occurs.

3.2. Types of media bias according to the author’s intention

There are two categories of media bias according to author’s intention: Spin bias and ideology bias.

3.2.1. Spin bias

The spin or *rhetoric bias* occurs when the journalist tries to create a “memorable story” (Mullainathan & Shleifer, 2002). The journalist may use loaded or emotional language, exaggerate, or select only certain facts that fit the story in order to make it more interesting. Some examples of this bias include “clickbait” headlines and stories that focus on drama instead of substance.

The spin bias can often be found in media coverage of controversial topics or events. For example, a journalist may choose to focus on the most dramatic aspects of a story, such as violence or conflict, in order to make it more attention-grabbing. This type of coverage can



The skinny version: There are more than a hundred Republican-held congressional districts across the country that have a narrower margin than 17. If seats that look like this one in Pennsylvania are toss-ups in November, it's going to be a bloodbath.

Fig. 2. Example of spin bias.
Source: BBC as cited in all-sides.com.

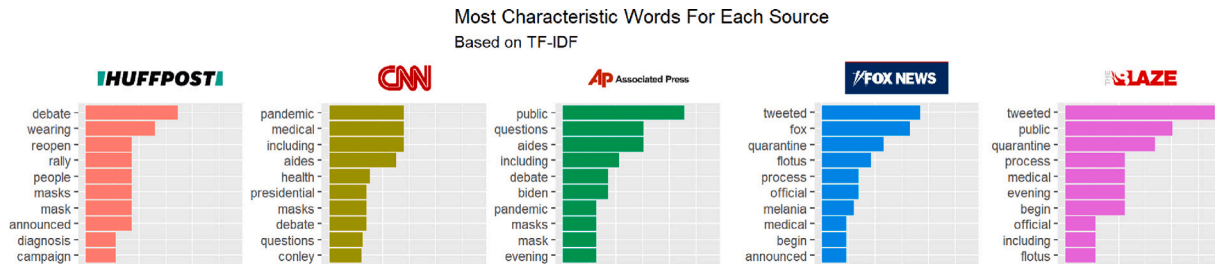


Fig. 3. Distribution of most common words in articles about US presidential debate in five media outlets.
Source: Law (2020).

often distort the reality of a situation and lead to misunderstanding or misinformation.

In the example below (Fig. 2) we can see spin bias by defining a possible election result as a bloodbath.

3.2.2. Ideology bias

The ideology bias, also known as *stance* or *framing bias*, occurs when the issuer presents the information in an partial way. The author may be biased towards a certain ideology, which can influence the way they present information in their work. This can make it difficult for readers to assess the accuracy and impartiality of the information presented. The ideology bias is commonly detected in political issues, but it goes beyond the typical political compass (left v. right). The website AllSides (Mastrine, Sowers, Alhariri, & Nilsson, 2022) lists 14 common types of ideological bias in the main US media outlets: authoritarian v. libertarian, individualist v. collectivist, secular v. religious, traditionalist v. progressive, elitist v. populist, rural v. urban, and nationalist/localist v. globalist.

3.3. Types of media bias according to the context

Regarding the context in which the media bias occurs, there are three categories of bias: coverage bias, gatekeeping bias, and statement bias (Saez-Trumper, Castillo, & Lalmas, 2013).

3.3.1. Coverage bias

Coverage bias refers to the (quantitative and qualitative) visibility of topics or entities in media coverage. It is related to the tendency of the media to cover some stories and not others. This can be due to a variety of reasons, such as the media's focus on negative stories, the media's focus on stories that will generate a lot of attention, or the media's focus on stories that fit a particular narrative. One example of coverage bias is the tendency for the media to focus on negative stories. This can lead to a distorted view of reality, as the media is more likely to cover stories that are sensational or have a high potential for conflict. This can also lead to a sense of fear or anxiety among the public, as they may believe that the world is a much more dangerous place than it actually is (Stafford, 2014). In Fig. 3, we can see how different media cover the same event in different ways, focusing on different topics or aspects of the event. The Huffington Post focuses more on the electoral debate, while CNN mentions more the management of the pandemic; and Fox News and LAZE publish information about something posted on Twitter.

3.3.2. Gatekeeping bias

Gatekeeping bias, also called *selection bias* or *agenda bias*, relates to the stories that the media select or reject to report. This can lead to a biased portrayal of events, as some stories may be deemed more important than others, regardless of their actual importance. We can perceive selection bias by looking at how the media covers a certain story or person (Saez-Trumper et al., 2013). In Fig. 4 we can see how media outlets with different political stances, in the same time space, decide which news to tell and which not on the same topic, in this case, LGBTQ issues.

3.3.3. Statement bias

Statement bias, also called *presentation bias*, refers to how articles choose to inform about certain entities/concepts. This can be done through the use of loaded language or by presenting one side of an issue as the only side.

Within presentation bias, journalists can use different writing styles to bias the news. Most common examples are *labelling* and *word choice* (Hamborg, Zhukova, & Gipp, 2019). Labelling is a form of statement bias where the media outlet uses certain words or phrases to describe an individual, event, or organization in a way that conveys a particular opinion or perspective. For example, referring to someone as an "illegal immigrant" rather than using the term "undocumented worker" creates a negative connotation and implies that the person is doing something wrong.

Word choice can also be used to create statement bias. This is when specific words are chosen in order to make an argument more persuasive. For example, choosing to use the word "abortion" instead of "choice" or "reproductive rights" is more likely to create an emotional response in the reader and make them more likely to agree with the writer's position. Both labelling and word choice are concrete examples of forms of media bias that can be used to influence the way readers perceive a certain issue.

In Fig. 5 we can see an example of statement bias by word choice. In this case, journalists call COVID-19 the "Chinese virus".

3.4. Forms of media bias

Just as there are several types of media bias, it can manifest itself in various ways. Bypassing articles listing media bias manifestations






















From the Left	From the Center	From the Right
<p>NEWS</p> <p> Texas Supreme Court allows child abuse investigations into families of transgender teens to continue</p> <p>The Texas Tribune </p>	<p>NEWS</p> <p> Biden: "MAGA crowd" is "most extreme" political group in U.S. history</p> <p>Axios </p>	<p>ANALYSIS</p> <p> Disney funds gender identity, LGBTQ curriculum in schools—and has been for 20 years</p> <p>The Post Millennial </p>
<p>NEWS</p> <p>Gay high schooler says he's 'being silenced' by Florida's LGBTQ law</p> <p>NBC News (Online) </p>	<p>NEWS</p> <p>Texas AG Ken Paxton's trans care opinion filled with false claims and distortions, researchers say</p> <p>San Antonio Express-News </p>	<p>NEWS</p> <p>Hawley introduces bill to strip 'woke' Disney of special copyright protections</p> <p>Fox News (Online News) </p>
<p>ANALYSIS</p> <p>Some trans Twitter users say platform under Elon Musk would be 'terrifying'</p> <p>NBC News (Online) </p>	<p>ANALYSIS</p> <p>Florida's \$1 Billion Disney Question</p> <p>Wall Street Journal (News) </p>	<p>NEWS</p> <p>Sen. Mike Lee wants warnings on LGBTQ content in children's TV programs</p> <p>Deseret News </p>
<p>NEWS</p> <p>Disney's self-governing district says Florida cannot dissolve it without paying off its debts</p> <p>CNN (Online News) </p>	<p>NEWS</p> <p>Opinions Split After 'View' Host Says GOP Not 'In Line' With U.S. Values</p> <p>Newsweek </p>	<p>NEWS</p> <p>DeSantis 'may have gone too far' in battle with Disney, GOP donor says</p> <p>Washington Examiner </p>
<p>NEWS</p> <p>Jen Psaki in tears during interview on Republican anti-LGBTQ 'cruelty'</p> <p>The Guardian </p>	<p>DATA</p> <p>Deep partisan divide on whether greater acceptance of transgender people is good for society</p> <p>Pew Research Center </p>	<p>NEWS</p> <p>Randi Weingarten says parental rights bills are 'the way in which wars start'</p> <p>Fox News (Online News) </p>
<p>ANALYSIS</p> <p>Talk of race, sex in schools divides Americans: AP-NORC poll</p> <p>Associated Press Politics & Fact Check </p>	<p>NEWS</p> <p>Florida set to strip Disney of self-governing status in dispute over LGBTQ law</p> <p>Reuters </p>	<p>OPINION</p> <p>Gov. DeSantis Is Right To Attack Disney. Republicans Everywhere Should Follow His Lead</p> <p>The Federalist </p>

Fig. 4. Example of gatekeeping bias, each media outlet decides which stories to tell or not.



Fig. 5. Example of statement bias by word choice.

heavily based on US politics, in the literature we find similar classifications. Baker, Graham, and Kaminsky (1996) list seven different forms, including: bias by commission, bias by omission, bias by story selection, bias by placement, bias by the selection of sources, bias by spin, and bias by labeling. Hamborg, Donnay et al. (2019) add three more forms to this list, including biases that take into account not only the text, but also the accompanying image (size allocation, picture selection, and picture explanation). In addition, the AllSides website (Mastrine et al., 2022) establishes a list of 16 media bias forms very similar to the one established by the two previous papers (spin, unsubstantiated claims, opinion statements presented as facts, sensationalism/emotionalism, mudslinging/ad hominem, mind reading, slant, flawed logic, bias by omission, omission of source attribution, bias by story choice and placement, subjective qualifying adjectives, word choice, negativity bias, photo bias, elite v. populist bias).

After analyzing all forms of media bias explained in the literature, we have merged duplicities from these sources, and classified each form of media bias according to the two types of bias seen in the previous section, resulting in 17 forms of media bias. Fig. 6 shows these forms

and their classification according to the author's intention and to the context.

3.4.1. Unsubstantiated claims bias

Unsubstantiated claims bias is a type of bias that occurs when journalists make claims that are not backed up by evidence. This can lead to people believing things that are not true, or at least not as true as they might be if there was evidence to support the claims. It can be related to the fake news concept (Kohlmeier, 2018).

One example of unsubstantiated claims bias is when a journalist claim that vaccinations cause autism (Ruiz & Bell, 2014). There is no scientific evidence to support this claim, but some people continue to believe it. This can lead to people not vaccinating their children, which can put them at risk for serious illnesses.

3.4.2. Opinion statements presented as facts bias

Opinion statements presented as facts bias is based on the use of subjective language or statements under the guise of reporting objectively. Subjective language is often used in editorials and opinions, as these are designed to persuade the reader to a particular viewpoint.

For example, consider a news report on a newly implemented policy. An unbiased report might state, "The government has introduced a new policy aimed at reducing carbon emissions over the next decade". On the other hand, a biased report using opinionated language might state, "The government has once again shown its disregard for businesses by imposing a new, stifling policy that threatens to undermine our economy, all in the name of reducing carbon emissions". In the latter statement, subjective terms like "disregard" and "stifling" present a negative opinion as an objective fact, potentially leading readers to perceive the policy unfavorably without examining its actual merits or details.

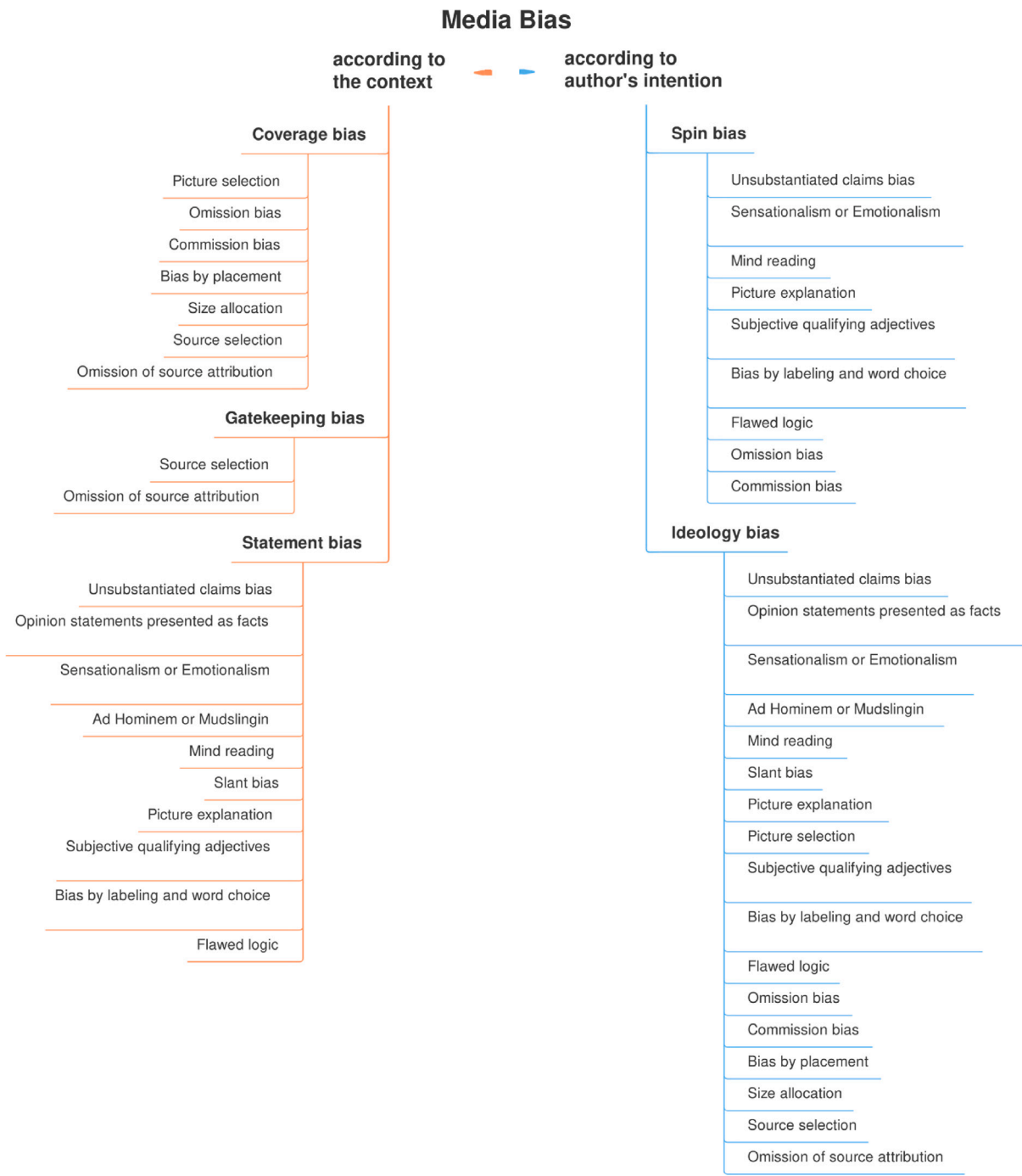


Fig. 6. Hierarchical Classification of Media Bias Types Based on Author's Intention and Contextual Factors.

3.4.3. Sensationalism/emotionalism bias

Sensationalism is a type of bias where information is exaggerated in order to create a more emotional reaction. This can be done by selectively choosing information that supports a certain view, while leaving out information that may contradict it. Sensationalism is often used in the media to increase viewership or readership.

For instance, imagine a city experiencing its first snowfall of the season. An unbiased report might state, “The city experienced its first snowfall today, marking the start of the winter season with a light blanket of snow”. However, a sensationalized report might declare, “Residents were in shock as an unexpected and intense blizzard wreaked havoc across the city, leaving many questioning if they are prepared for the winter’s fury ahead!”. While both statements address

the snowfall, the latter exaggerates the event’s severity and potential consequences, triggering heightened emotional responses from readers. This kind of reporting, while captivating, can lead to misinformed perceptions and unnecessary panic among the audience.

3.4.4. Ad hominem/mudslingin bias

The Ad Hominem bias is when a journalist attacks another person instead of their argument, while mudslinging bias happens when people attack each other’s character instead of debating the issue (Yap, 2013).

For example, consider a televised debate on healthcare reform. Instead of discussing the merits and drawbacks of a particular healthcare proposal, one debater might resort to Ad Hominem attacks by saying, “Well, of course you’d support that proposal; you’ve been known to

take donations from big pharmaceutical companies”. Similarly, in a heated political campaign, an opponent might use mudslinging tactics by airing ads that delve into a candidate’s past indiscretions or mistakes rather than addressing their policy positions. Such attacks divert attention away from the substantive issues and focus it on personal characteristics or actions, potentially leading the public to form opinions based on character assaults rather than policy strengths or weaknesses.

3.4.5. Mind reading bias

Mind reading is a type of media bias that occurs in journalism when a journalist assumes that he or she knows what another person thinks, feels, or intends without actually speaking to that person.

One example of mind reading in journalism can be seen in the way some writers cover political stories. They may make assumptions about a politician’s motives or intentions without actually speaking to the politician herself. This can lead to inaccuracies or even outright falsehoods being reported as fact.

3.4.6. Slant bias

Slant bias is a type of bias that occurs when someone has preferences for one thing over another. It can include cherry-picking information or data to support one side. This can be due to a person’s own personal preferences or experiences, or it can be due to outside influences, such as media portrayals. Slant bias can lead to people making judgments about others or situations without all of the facts, or it can cause them to misinterpret information.

For instance, let us consider the coverage of an environmental protest. If a news outlet has a slant bias in favor of industrial growth and against environmental activism, their report might focus on the traffic disruptions caused by the protest and any minor infractions committed by protestors. They might use terms like “inconvenienced commuters” or “rowdy protestors”, emphasizing negative aspects. The report might also underplay or completely ignore the core environmental concerns that led to the protest in the first place, or the peaceful and constructive actions of the majority of protestors. In contrast, an article with an opposite slant bias might glorify the protestors as “eco-warriors” or “champions of the planet”, while downplaying any negative aspects of the protest. Both versions present skewed views of the event, and readers may not get a complete and balanced understanding of what transpired.

3.4.7. Omission bias

The bias by omission is a type of media bias in which media outlets choose not to cover certain stories, topics or aspects of stories. One example of bias by omission is the lack of coverage of certain political candidates by the news media. For instance, a candidate who is not well-known or who is considered to be a long shot for the nomination may not receive much coverage from the media. This can make it difficult for voters to make an informed decision about who to vote for.

3.4.8. Commission bias

Bias by commission is a type of media bias in which media coverage favors one particular party or side in a dispute, often to the exclusion of the other party or side.

One example of this type of bias is when a news organization only interviews people who support a particular political candidate, and does not interview anyone who supports the other candidates. This can create the impression that the candidate who is favored by the news organization is the only one who is qualified or has valid points, while the other candidates are not worth considering (De Witte, 2022). This type of bias can also occur when a news organization only covers one side of a controversial issue, such as abortion, and does not give any coverage to the other side.

3.4.9. Bias by labeling and word choice

Labeling and word choice are biases related to how the journalist choose the words to present the story (Hamborg, 2020). If a story is about a controversial issue, the journalist may use loaded language to make one side seem more favorable. This can be done by using words with a positive connotation to describe one side (i.e. “coalition forces”) or words with a negative connotation to describe the other side (i.e. “invading forces”) (Parker-Bass et al., 2022).

3.4.10. Flawed logic bias

Flawed logic or faulty reasoning bias consists on leading to conclusions that are not justified by the given evidence. It is related to these logical fallacies (Van Vleet, 2021):

- **Hasty generalization or over-generalization:** This happens when a claim is made based on evidence that is too small. For instance, if a media outlet were to interview only three people at a protest and then claim “The majority of attendees share the same view”, they would be making an over-generalization.
- **False cause fallacy (non causa pro causa):** This fallacy consists on mislocating the cause of one phenomenon in another that is only seemingly related. A media outlet may incorrectly link two events. For instance, they might attribute a rise in crime to a recent policy change without adequately exploring other factors that might have contributed.
- **Slippery slope:** This fallacy suggests that an action will initiate a chain of events culminating in a predictable later event, without establishing or quantifying the relevant contingencies. A news story might suggest that legalizing a particular drug will lead to an inevitable increase in addiction and societal breakdown, without providing evidence for this chain of events.
- **Black-and-white thinking (splitting):** Related to polarization, is the failure in a person’s thinking to bring together the dichotomy of both positive and negative qualities of the self and others into a realistic whole. Media may portray a complex issue in a dichotomous way, such as framing a political debate as purely “liberal vs conservative”, ignoring the nuances and diverse perspectives.
- **Fallacy of division:** Consists in inferring that something is true about one or more of the parts of a compound because it is true about the compound of which it is a part. A media outlet might assume that if a political party has a specific stance, then every member of that party must hold that stance.
- **Fallacy of composition:** Consists in inferring that something is true about a set or group just because it is true about one or more of its parts or components. Conversely, if one politician in a party holds a view, a news report might claim that the entire party holds that view.
- **Fallacy of accident:** It is committed by improperly applying a generalization to individual cases. A media outlet might take a generalization and improperly apply it to a specific case, such as applying broad economic trends to an individual’s financial situation.
- **Irrelevant conclusion fallacy:** This happens when a conclusion is drawn, and it is not related to the argument. A news report might conclude an argument that is not related to the facts presented, such as attributing a natural disaster to a controversial political figure’s actions.
- **Appeal to groupthink (Ad populum):** Consists in appealing to the fact that everyone else is doing it or everyone else believes it. An outlet may claim something is true simply because a majority believes it, without critically examining the evidence.
- **Appeal to authority (Ad verecundiam):** Appealing to an authority figure instead of using logic. Using a celebrity’s endorsement as evidence in a political argument, rather than relying on expert analysis or substantial facts.

- **Red Herring (Ad ignorantiam):** This fallacy consists in introducing a new topic that is unrelated to the argument. media report might introduce an unrelated issue to divert attention from the main argument, such as focusing on a politician's personal life instead of their policy decisions.

3.4.11. Bias by placement

Bias by story placement is when a story is placed in a position that is more likely to be seen by people. This can be a deliberate choice by the person who is placing the story, or it can be an accidental choice.

For instance, a newspaper editor might decide to feature a particular political scandal on the front page, while relegating another seemingly important story about healthcare reforms to a less prominent page. In doing so, the newspaper might be emphasizing the scandal over the healthcare reform, thereby influencing the readers' perception of which topic is more newsworthy or important.

3.4.12. Subjective qualifying adjectives bias

When a journalist uses qualifying adjectives, they may be introducing subjective bias into her writing. This can be done deliberately, in order to influence the reader's opinion, or inadvertently, due to the journalist's own personal beliefs (Pant, Dadu, & Mamidi, 2020). Qualifying adjectives can be used to cast doubt over facts, or to presuppose the truth of conclusions. In both cases, the use of these words and phrases can introduce bias into the reporting of events.

For instance, consider a situation where two political figures are engaged in a debate. One journalist might describe one of the figures as having given a "fiery" speech, while another might label it as "passionate". Here, "fiery" carries a connotation of being aggressive or possibly even unruly, while "passionate" evokes sentiments of enthusiasm and strong belief. Depending on the adjective used, readers may form different opinions about the political figure's demeanor and the content of the speech.

3.4.13. Size allocation bias

Both readers of a printed or digital medium are often selective about the information they choose to read. Some papers study the behavior of readers while they consume the content of a website. The idea behind these research is to study if the user is reading all the content or if he or she is only reading some parts of the articles (Holsanova, Rahm, & Holmqvist, 2006). Some studies suggest that a reader is more likely to read the headline of an article than the whole article itself (Jiang, Guo, Chen, & Yang, 2019). The headline is the first thing that a reader sees and it can influence the decision of whether to read the whole article or not.

3.4.14. Source selection bias

Source selection bias is the tendency of journalists to choose sources that will support their stories, instead of choosing sources that will give them an accurate account of what happened.

The effects of source selection bias are similar to the effects of commission and omission (Hamborg, Donnay et al., 2019) in that they can lead to a distorted or incomplete view of an event.

An example might be a journalist covering a local environmental disaster and only interviewing representatives from the responsible company, without giving a voice to affected community members or independent experts.

3.4.15. Omission of source attribution bias

Omission of source attribution happens when a journalist does not back up his or her claim with a source, or the source is diffuse, or unspecific. Some examples of omission of source attribution are phrases such as "according to a source", "critics say", or "experts believe". In some cases, the omission of source attribution can be intentional, in order to protect the source's anonymity.

3.4.16. Picture selection bias

This bias is similar to the word choice bias, but for images. Image selection bias is a type of cognitive bias that refers to the tendency for people to base their judgments on images that are presented to them, rather than on the actual content of the images. Picture selection bias can lead people to make judgments about something without considering all of the information that is available. For example, photos taken in the same manifestation can show totally different realities, they can already show violent or peaceful protesters, thus impacting the reader's opinion (Geske, 2016).

3.4.17. Picture explanation bias

Just as the selection of images can affect the opinion of the reader, the caption that is added next to the image can also add bias to the content. For example, some US media outlets tends to exaggerates the proportion of African Americans among the poor in their published photos (Gilens, 1996), causing the public's to associate race and poverty.

4. Media bias and disinformation

Information scientists have long debated the nature of information: what it is, where it comes from, its effect on society, etc. From its earliest stages, information science has sought to define information. Shannon and Weaver (1948) postulates that information can be quantified as bits of a signal. Their work does not help to understand disinformation since assimilating it to mere noise would ignore its possible informative nature (Karlova & Fisher, 2013), as we can see in the Table 2.

In recent years, the study of information has shifted from a focus on its definition to a focus on its function. This shift has been motivated in part by the recognition that information is not a static thing, but rather is always changing and always in flux. The study of information now emphasizes the ways in which information is produced, distributed, and consumed, and the ways in which it affects individuals, groups, and societies.

Estrada-Cuzcano, Alfaro-Mendives, and Saavedra-Vásquez (2020), Karlova and Fisher (2013), propose a theory in which they not only establish a difference between information and disinformation, but also decompose disinformation into two other concepts: misinformation and disinformation. Misinformation and disinformation present erroneous information (either because it is false, incomplete and/or out of date), but the characteristic that differentiates them is that the misinformation does not have any misleading intentions, and the disinformation does.

Karlova and Fisher (2013) characterize information, misinformation, and disinformation according to whether what they present is true (all the information that is presented corresponds to reality), complete (all the relevant information is presented), current (the information is presented in a timely manner), informative (the information is presented in a way that is useful to the recipient), and/or deceptive (the information is presented in a way that is intended to mislead the recipient).

If we take into account the description and characterization of media bias that we have made in the previous Section, it is possible to appreciate the relation between media bias and disinformation. In fact, media bias and disinformation share many common features: first, disinformation is often spread deliberately, with the intention of misleading people, as media bias can be deliberately introduced into news stories in order to influence the way that people perceive them; second, disinformation is often spread by people who have a vested interest in the outcome of an event. For example, a political candidate may spread disinformation about their opponent in order to make themselves look more favorable. This is similar to the way that media bias can be introduced by journalists into stories in order to favor one side over another; third, disinformation is often spread through the use of media, such as television, radio, and the internet,

Table 2

A summary of features of information, misinformation, disinformation, media bias, fake news, and propaganda. Y = Yes; N = No; Y/N = Could be Yes and No, depending on context and time. Own elaboration, based on [Karlova and Fisher \(2013\)](#) work.

	Information	Misinformation	Disinformation	Media bias	Fake news	Propaganda
True	Y	Y/N	Y/N	Y/N	N	Y/N
Complete	Y/N	Y/N	Y/N	Y/N	Y/N	Y/N
Current	Y	Y/N	Y/N	Y/N	Y/N	Y/N
Informative	Y	Y	Y	Y	Y/N	N
Deceptive	N	N	Y	Y/N	Y	Y
Slanted	N	N	Y/N	Y	Y/N	Y/N

the same media news stories are shared; and fourth, disinformation is often spread through the use of language, such as loaded words and phrases, and through the use of images, similar to the way that media bias can be introduced into news stories through the use of language and images.

That is why we think that is necessary to adapt and enhance the framework proposed by [Karlova and Fisher \(2013\)](#), adding *slanted* (information that is presented in a way that is partial, or one-sided) as a new characteristic which can help us to better understand the relationship between information and disinformation, as we showed in the [Table 2](#).

Also, adding this characteristic (*slanted*), not only we can add media bias to the framework, but also some related problems such as fake news, or propaganda. Fake news refers to the dissemination of information that has been purposely manipulated (i.e. fabricated) in order to cause damage or influence public opinion, and propaganda is a deliberate attempt to influence public opinion.

We need tools that help us to understand the problem of misinformation and disinformation, and help us to take action. This framework helps to understand the conceptual differences between information, misinformation, disinformation, media bias, fake news, and propaganda. In addition, we have proposed that the slanted characteristic should be added to the framework to complete it, as it is needed to understand the concept of media bias, and differentiate it from disinformation in some particular cases (see [Table 2](#)). It is important to note that media bias is not always disinformation. In this sense, the framework helps to understand when we are dealing with disinformation and when with media bias, and also with fake news and propaganda.

5. Automatic media bias detection

Media bias detection is a process of automatically identifying the presence of bias in a journalistic text. There are many ways to detect bias in journalistic texts, but most methods involve some combination of Natural Language Processing (NLP) and machine learning. Some common NLP techniques for detecting bias include sentiment analysis ([Lin, Bagrow, & Lazer, 2011](#)), topic modeling ([Best, van der Goot, Blackler, Garcia, & Horby, 2005](#)), and the study of lexical features ([Hube & Fetahu, 2018](#)). Machine learning techniques can be used to identify patterns in the text that indicate the presence of bias.

Bias detection is difficult because, as we already seen, there is no agreed-upon definition of what constitutes bias in journalism. Additionally, bias can be subtle and may not be easily detectable.

The detection of media bias can be addressed both as a binary classification problem or as a multi-class classification problem. In binary classification, the goal is to predict whether a piece of content is biased or not. In multi-class classification, the goal is more diverse as it can be: (1) the degree of bias or polarization (e.g. reliable, mixed, unreliable as in [Horne, Khedr, and Adali \(2018\)](#)), (2) the stance towards an event (e.g. pro-Russia, neutral, pro-Western as in [Cremisini, Aguilar, and Finlayson \(2019\)](#)), or (3) the political compass (e.g. extreme-left, left, center-left, center, center-right, right, extreme-right) as in [Baly, Karadzhov et al. \(2020\)](#). There are also specific cases in which the detection of media bias is a multi-label task, as in the case of [Budak,](#)

[Goel, and Rao \(2016\)](#) in which the bias towards both the US Republican Party and the Democratic Party is studied.

In addition, the study of media bias can be done at different levels: (1) at sentence-level, if the goal is to identify which sentences contain bias within a document; (2) at the article-level, if the whole document is analyzed; (3) at the journalist-level, to analyze that person's bias, and (4) at the media outlet-level, to study whether that media outlet is hyperpartisan.

5.1. Related problems

As discussed in [Section 3](#), the problem of media bias can be classified within an information v. disinformation framework, but it is not the only disinformation problem that occurs nowadays. In this section we will briefly cover some of the other tasks that are related to media bias detection.

Stance detection

The stance detection task is very similar to the detection of media bias, since aims at classifying information based on the stance of the author. Stance can be defined as the position or opinion of an author on a given topic. Stance detection has been also used for media and author profiling ([Taulé et al., 2017](#)).

Propaganda detection

As we stated in [Section 3](#), propaganda is a deliberate attempt to influence public opinion. Propaganda often uses a mixture of true and false information, and its detection is based on the study of the rhetorical and psychological techniques used for creating propaganda ([Da San Martino et al., 2021](#)). Some of these devices are: name calling, glittering generalities, transfer, testimonial, plain folks, card stacking, and bandwagon.

Fake news detection

The main difference between fake news and media bias is that fake news are always false, while media bias can be true or false. There are three different perspectives when creating fake news detection systems: (1) style-based: how fake news are written, (2) propagation-based: how fake news spread, and (3) users-based: how users engage with fake news and the role that users play (or can play) in fake news creation, propagation, and intervention ([Zhou & Zafarani, 2020](#)).

Rumor detection

A rumor is a piece of information that is spread without any confirmation of its truth. The problem is closely related to the problem of detecting fake news. The main difference between the two is that a rumor is a piece of information that is not verified yet, whereas a fake news is already considered false. In the context of social media, rumors spread very fast, and can often be difficult to distinguish from real news. In rumor detection, the diffusion through social networks is usually studied, and source detection is one of its main tasks ([Shelke & Attar, 2019](#)).

Clickbait detection

Clickbait is a type of online content that is designed to lure readers to click on a link. Clickbait often used language in which something unnamed is referred to, some emotional reaction is promised, some lack of knowledge is ascribed, or some authority is claimed (Potthast, Köpsel, Stein, & Hagen, 2016).

Bias detection in non-journalistic contexts

Apart from journalism, there are several other domains where the problem of finding bias in text has been studied. Some examples of detection of bias in non-journalistic contexts are: the detection of bias in Wikipedia (Hube & Fetahu, 2018), the detection of bias in congressional speeches (Hajare, Kamal, Krishnan, & Bagavathi, 2021), the use of bias features in order to detect sexist messages on social networks (Rodrigo-Ginés, Carrillo-de Albornoz, & Plaza, 2021), or the study of bias for monitoring the public opinions on real estate market policies (Cao, Xu, & Shang, 2021).

5.2. Approaches for detecting media bias

In this subsection, we will analyze the most used techniques in the task of detecting media bias. We will first analyze the traditional methods, i.e., those that are based on classical machine learning, and then move on to the more recent methods that are based on deep learning. Finally, we will list some researches that do not fit in any of these two categories.

5.2.1. Non-deep learning models

The pioneering methods for detecting media bias predominantly utilized machine learning techniques, such as logistic regression, support vector machines, random forests, and naïve bayes. In these approaches, features extracted from the text are input into a classification algorithm, trained to categorize articles/sentences by media outlets into predefined classes. A notable limitation of these traditional methods is their reliance on handcrafted features. Consequently, the performance of such a method is heavily influenced by feature selection (Cruz, Rocha, & Cardoso, 2019). In this Section, we differentiate between *language-based methods* and *reported speech-based methods*. Language-based methods study a range of features from the text: lexical features pertaining to word choice and usage, morphosyntactic features related to grammar and sentence structure, and semantic features dealing with meaning. On the other hand, reported speech-based methods focus on the analysis of sources quoted within the text.

Linguistic based methods

In this category of methods, classical machine learning models such as SVM, Logistic Regression, Random Forest, etc., are trained over linguistic features. These types of methods have been the most typically used for the media bias detection task, mostly due to the fact that most of the research papers on this topic were published from the early 2000s up to the early 2010s, before the recently proposed deep learning methods were popularized.

We can break down linguistic features into three types: *lexical* features such as n-grams, topics or custom lexicons; *syntactic features* such as the Part Of Speech (PoS); and finally, *semantic features* such as Linguistic Inquiry, the Named Entity Recognition (NER), or Word Count (LIWC), a common approach to identify linguistic cues related to psychological processes such as anger, sadness, or social wording.

In Krestel, Wall, and Nejd (2012), the authors present an automated method for identifying vocabulary biases in online content by comparing it to parliamentary speeches. Using 15 years of speeches from the German Bundestag, they employed the vector space model from information retrieval, with term weights based on TF-IDF scores. They found that vocabulary biases in national online newspapers and magazines align with those in political party speeches, demonstrating

the effectiveness of their approach in tracking political leanings in media content.

A clear example of a bias detection method using linguistic features is the work of Hube and Fetahu (2018). Some of the features they employ are related to the use of custom lexicons, PoS, and LIWC. One of the lexicons used contains biased terms, which has been shown a posteriori to not work well, since a word can be biased or not depending on the context (Hamborg, Donnay et al., 2019). In order to use biased word lexicons, you need to create them for the specific context you are analyzing, as in Rodrigo-Ginés et al. (2021). Their model is able to detect biased statements with an accuracy of 0.74.

The work of Hube and Fetahu (2018) had quite an impact as other authors replicated this method in other areas. Lim, Jatowt, and Yoshikawa (2018a, 2018b) used the same approach by extracting named entities from the text as well, obtaining an accuracy of 0.7. Spinde, Hamborg, and Gipp (2020) also trained their models with the text represented as TF-IDF features, achieving a F1-score performance of 0.43.

In Baraniak and Sydow (2018), the authors emphasize the significant impact of digital media on public opinion and the prevailing issue of information bias in news presentations. They highlight the need for tools capable of detecting and analyzing this bias. The study focuses on the automatic detection of articles discussing identical events or entities, which could be useful in comparative analysis or creating test/training sets. Three machine learning algorithms were tested for predicting article sources based solely on content, deliberately excluding explicit source attributes. Among the tested algorithms, which included naïve bayes, logistic regression, and support vector machines, the latter exhibited the best performance, underscoring the feasibility of source recognition through basic language analysis.

Al-Sarraj and Lubbad (2018) delve into the bias present in online mass media, particularly in its portrayal of politically charged events. Given the foundational expectation of neutrality in journalism, any inclination towards a particular viewpoint challenges the very ethos of press and media. One of the principal manifestations of such bias is the deployment of misleading terminologies. This study revolves around the coverage of the 2014 Israeli war on Gaza by Western media outlets. There is a widespread sentiment among the Palestinian populace suggesting an overt bias in Western media towards the Israeli narrative and vice versa. In this research endeavor, the authors conduct a text mining experiment on Western media content to decipher patterns indicating press orientation and, subsequently, any biases favoring one side over the other.

Harnessing text mining techniques and machine learning algorithms, the authors embarked on detecting biases within news articles. Their methodology comprised crawling articles from seven leading Western media outlets, preprocessing this content into a structured format amenable for analysis, and subsequently constructing sentiment classifiers to prognosticate the inherent bias in these articles. The research ventured into a comparative analysis of three supervised machine learning algorithms for sentiment classification, each paired with varying n-grams. Notably, the combination of SVM with bi-grams emerged as the most efficacious, boasting impressive performance metrics, including an accuracy of 0.91, a recall of 0.88, and an F-measure of 0.91.

Another interesting approach was presented by Gupta, Jolly, Kaur, and Chakraborty (2019), where they developed a system for news bias prediction as part of the SemEval 2019 task. This system was primarily based on the XGBoost algorithm and utilized character and word-level n-gram features. These features were represented through both TF-IDF and count vector-based correlation matrices. The goal of the model was to ascertain whether a given news article could be characterized as hyperpartisan. On testing, their model demonstrated a precision rate of 0.68 on the dataset provided by the competition organizers. Further evaluation on the BuzzFeed corpus revealed that

their XGBoost model, when coupled with simple character-level N-Gram embeddings, could reach impressive accuracies nearing 0.96. Despite these accomplishments, the authors acknowledged a significant limitation of their model. Their system showed a pronounced inability to identify a larger portion of relevant results, indicating a low recall.

In the backdrop of rising concerns about fake news, bias, and propaganda, Baly, Karadzhov, Saleh, Glass, and Nakov (2019) embarked on an investigation into two relatively less explored facets: (i) the trustworthiness of entire news media outlets measured on a 3-point scale, and (ii) the political ideology of the same, gauged on a 7-point scale ranging from extreme-left to extreme-right bias. Rather than focusing on individual articles, they aimed to evaluate the overarching stance of the news outlet itself. They put forth a multi-task ordinal regression framework that jointly models these two aforementioned problems. This endeavor was fueled by the insight that outlets exhibiting hyper-partisanship often sacrificed trustworthiness, resorting to emotional appeals rather than adhering to factual information. In stark contrast, media occupying a central position generally showcased a higher degree of impartiality and trustworthiness. The research heavily relied on the MBFC dataset (Baly, Karadzhov, Alexandrov, Glass, & Nakov, 2018), which incorporates annotations for 1,066 news media. These annotations, manually added, gauge the factuality and political bias of the media outlets on the respective scales previously mentioned.

Results from the study indicated the superiority of the multi-task ordinal regression model over the majority class baseline. Performance metrics showed a boost when auxiliary tasks were incorporated in the modeling process. For instance, for factuality prediction, the combination of understanding whether a medium is centric or hyper-partisan proved crucial. A medium devoid of strong political ideology was generally deemed more trustworthy than a heavily biased counterpart. On the other hand, for political bias prediction on a 7-point scale, the best model harnessed information at broader levels of granularity. This assisted in minimizing significant errors in the predictions.

Further contributing to this shared task, Palić et al. (2019) presented their approach to hyperpartisan news detection. Their system, which is reliant on the SVM model from the Python Scikit-Learn library, processed raw textual articles and determined their hyperpartisan nature. Demonstrating commendable prowess, they secured the 6th position out of 42 participating teams, boasting an accuracy rate of 0.79. On a related note, Färber, Qurdina, and Ahmed (2019) delved into classifying news articles based on their bias using a convolutional neural network.

Other authors have complemented the use of linguistic features with the use of word embeddings, such as word2Vec in the case of Chen, Wachsmuth, Al Khatib, and Stein (2018), Preotjuc-Pietro, Liu, Hopkins, and Ungar (2017), or doc2vec like Geng (2022).

Lastly, in Kameswari, Sravani, and Mamidi (2020), the authors address the subtle influence of presuppositions in news discourse. Presuppositions, by their nature, introduce information indirectly, making it less likely to be critically evaluated by readers. Drawing from discourse analysis and the Gricean perspective, this study seeks to link the type of knowledge presupposed in news articles with the underlying biases they may contain. They introduce guidelines for detecting different presuppositions in articles and provide a dataset of 1,050 annotated articles for bias and presupposition levels. By incorporating sophisticated linguistic features, their supervised classification method, particularly the Random Forest classifier, achieved an accuracy of 0.96 and an F1 score of 0.95, surpassing previous state-of-the-art models in political bias detection.

Reported speech based methods

In this category of methods, the features extracted are based on reported speech. Reported speech is defined by Oxford Languages as a phenomenon in which a speaker's words reported in subordinate clauses governed by a reporting verb, with the required changes of person and tense. Reported speech tells you what someone said, but

without using the person's actual words. It is an integral part of the news storytelling, and a common artifact in statement and ideology media bias.

The use of reported speech can be used to create a false sense of balance in news stories (Lazaridou, Krestel, & Naumann, 2017). This is done by including quotes from both sides of an issue, even if one side is clearly more credible or trustworthy than the other. This can lead to a biased portrayal of the issue, as the less credible side is given equal weight to the more credible side.

Some authors have studied and analyzed reported speech in order to create media bias detection models and have applied them to various news sources. Park et al. (2011) were among the first authors to analyze reported speech in search of media bias. To do this, they developed a system that extracted, through NER and coreference resolution, the subjects identified in phrases between quotation marks. Subsequently, they developed a key opponent-based partitioning method based on the HITS algorithm for disputant partitioning. The method first identifies two key opponents, each representing one side, and uses them as a pivot for partitioning other disputants. Finally, they classified the quoted phrases with an SVM model. They measured performance using precision weighted F-measure (wF) to aggregate the F-measure of three groups (the two opposing groups, and another group called "Other"). The best performance model got an overall average of the weighted F-measure of 0.68.

In 2015, Niculae, Suen, Zhang, Danescu-Niculescu-Mizil, and Leskovec (2015) published one of the most cited papers on the detection of media bias through the analysis of reported speech. In this work they proposed a framework for quantifying to what extent quoting political speeches follows systematic patterns that go beyond the relative importance (or newsworthiness) of the quotes. Niculae et al. (2015) manually annotated several media outlets into four categories: declared liberal, declared conservative, suspected liberal, and suspected conservative. They then created a corpus of presidential speeches, that they used for calculating the probability that a given quote would be cited by any of the media outlets annotated. Their best performing method obtained a F1-score of 0.28, and a Matthews Correlation Coefficient (MCC) of 0.27. One of its most interesting results is that declared conservative outlets are less likely to quote a statement that declared liberals media reported compared to a random quote.

Lazaridou and Krestel (2016) continued with the line of work of Niculae et al. (2015). In 2016, they analyzed different UK news, identifying the different subjects of the quoted sentences through NER and coreference resolution. Once the entities were identified, they analyzed the quoting patterns of different media outlets, realizing that Labour's quotes in 2004 are three times more recurrent than the ones from the Conservatives and twelve times more frequent than those of the Liberals.

In 2017, Lazaridou et al. (2017) published another article expanding on their previous research by developing a bias-aware model based on Random Forest to classify the reported speech to its original outlet and comparing this approach against a naive baseline that only leverages the content of reported speech, getting an average accuracy of 0.797.

Other papers have carried out a similar analysis, either to reveal community structure and interaction dynamics (Samory, Cappelleri, & Peserico, 2017) to measure information propagation in literary social networks (Sims & Bamman, 2020), or to analyze the difference between the secular media outlets and the religious media outlets in Turkey (Özge & Ercan, 2020).

Finally, Kuculo, Gottschalk, and Demidova (2022) published a knowledge graph in 2022 that can be used to analyze the bias potentially caused by one-sided quotes and references, that is, references that demonstrate one side of the picture of available evidence. They performed an evaluation of cross-lingual alignment for eight selected persons in English, German and Italian, obtaining an average F1-score of 0.99.

5.2.2. Deep learning models

In recent years, deep learning methods have been successfully applied to the task of detecting media bias. Using deep learning methods, one can automatically learn feature representation from text. In addition, deep learning methods are more capable of modeling the sequential structure of a sentence (Sutskever, Vinyals, & Le, 2014).

RNNs based methods

RNNs are a type of neural networks that are capable of modeling the sequential structure of a sentence. RNNs have an internal state (i.e. memory) that retains information about the previous words in a sentence. Thus, RNNs can model the sequential structure of a sentence (Sherstinsky, 2020).

There are two types of RNNs: (1) traditional RNNs, and (2) long short-term memory RNNs (LSTM). Traditional RNNs are very limited in modeling long-term dependencies in a sentence, this is because traditional RNNs have the so-called vanishing gradient problem. LSTM RNNs are capable of modeling long-term dependencies in a sentence by using an internal memory.

This type of model has been widely used in the task of automatic detection of media bias. For instance, in Iyyer, Enns, Boyd-Graber, and Resnik (2014), the authors apply a recursive deep learning framework to the task of identifying the political position evinced by a sentence. The RNN is initialized using a word embeddings (word2vec) and also includes annotated phrase labels in its training. Their best performing model for sentence-level bias detection got an accuracy of 0.693.

In 2017, Rashkin et al. (2017) showed that an LSTM-type RNN outperforms other classical machine learning models such as Naïve Bayes or Max-Entropy. They trained a LSTM model both with text (represented as TD-IDF features), and with text and LIWC features. The LSTM outperforms (F1-score = 0.56) the other models when only using text as input. The LSTM word embeddings are initialized with 100-dimension embeddings from GloVe. Bidirectional LSTM (Bi-LSTM) models have also been proved to perform better than classical statistic learning models (Rodrigo-Ginés et al., 2021).

A similar analysis was done in Baly, Da San Martino et al. (2020), obtaining better results with BERT-based (F1-score = 0.80, and MAE = 0.33) transformers than with an LSTM model (F1-score = 0.65, and MAE = 0.52).

One disadvantage of plain RNN models is that the classification is done based on the last hidden state. In the case of long sentences, this can be problematic as the weights from the different input sequences have to be correctly represented in the last state. Attention mechanisms have proven to be successful in circumventing this problem. The results show that this type of RNNs obtain better predictions (average F1-score of 0.77) than RNNs without attention models (average F1-score of 0.74). Hube and Fetahu (2019) experimented with RNNs with attention models, both with Global Attention and Hierarchical Attention, proving that RNNs are able to capture the important words and phrases that introduce bias in a statement, and that employing attention mechanisms (both global and hierarchical) can further improve the performance.

Transformers based methods

Transformers are a type of neural networks that have been shown to outperform RNNs in modeling the sequential structure of a sentence. The main difference between Transformers and RNNs is that Transformers do not have an internal state, instead, Transformers use the so-called self-attention mechanism to model the sequential structure of a sentence (Vaswani et al., 2017).

Transformers have been shown to outperform RNNs in several tasks such as machine translation, question answering, and text classification. Regarding the media bias detection task, models based on transformers are beginning to displace models based on linguistic features and RNNs, since they generally obtain better results (Baly, Da San Martino et al., 2020).

Fan et al. (2019) proposed a classifier based on a BERT-Base model. They used the “cased” version as it was useful for taking into account named entities, which are important for bias detection. They run BERT over the BASIL dataset at a sentence level and performed stratified 10-fold cross validation. The results improved using models based on transformers (F1 = 0.432) rather than using lexicons of polarity and subjectivity as in previous research (F1 = 0.26) (Wilson, Wiebe, & Hoffmann, 2005) (Choi & Wiebe, 2014).

In a related effort, Chen, Al Khatib, Stein, and Wachsmuth (2020) identified shortcomings in both feature-based and neural text classification techniques when it comes to detecting media bias. Notably, they realized that solely relying on the distribution of low-level lexical information was ineffective, especially for new event articles. Consequently, they proposed the use of second-order information about biased statements within an article to enhance detection efficiency. This was achieved by leveraging the probability distributions of the frequency, positions, and sequential order of lexical and informational sentence-level bias through a Gaussian Mixture Model. Their results, from an existing media bias dataset, indicated that the frequency and positions of biased statements play a significant role in influencing article-level bias. The sequential order, however, was found to be of secondary importance.

To demonstrate the superiority of their approach, Chen et al. (2020) utilized a pre-trained uncased BERT model. This was fine-tuned and optimized to handle both the beginning and end segments of articles, considering BERT's 512-token limit. Their findings emphasized the ineffectiveness of standard models for article-level bias detection, especially when devoid of features related to events. Classifiers primarily relying on style or structural features without specifically designed features for the task struggled with bias detection. Furthermore, in their experimentation with the Gaussian Mixture Model, they deduced that using 5 mixtures was generally optimal. As for the order of the Markov process they employed, a first-order Markov, which considers a position's dependency solely on its preceding position, proved most effective given the size of their dataset.

Other authors as Sinha and Dasgupta (2021), Tangri (2021) follow closely this approach. However, Sinha and Dasgupta (2021) also proved that augmenting linguistic features along with contextual embedding improves the performance of the model.

Spinde, Rudnitckaia et al. (2021) used some fine-tuned pre-trained models such as BERT, RoBERTa, and DistilBERT using Distant Supervision Learning, getting better results than linguistic-based models in the task of detecting bias at a sentence-level. They obtained a ROC AUC of 0.79, and a F1-score of 0.43 with their best performing model. Later, they improved their results in Spinde et al. (2022) using a Multi-task Learning (MTL) model. Their best-performing implementation achieves a F1-score of 0.78, performing the evaluations on the BABE dataset.

Krieger, Spinde, Ruas, Kulshrestha, and Gipp (2022) followed the Spinde, Rudnitckaia et al. (2021) research, presenting new state-of-the-art transformer-based models adapted to the media bias domain called DA-RoBERTa, DA-BERT, and DA-BART. They pre-trained their model with a bias domain corpus. These models outperformed (F1 = 0.81) the (Spinde, Rudnitckaia et al., 2021) models.

In the paper Agrawal, Gupta, Gautam, and Mamidi (2022), the authors address the escalating issue of media-driven political propaganda. They specifically focus on biased reporting that can shape misleading narratives, especially when favoring a specific political entity. Recognizing the lack of a dataset tailored for detecting political bias in Hindi news articles, the authors curated their own, encompassing 1,388 articles. These articles were categorized based on their inclination: biased towards, against, or neutral concerning the BJP, India's central ruling party at the time. Through various baseline approaches in machine learning and deep learning, the transformer-based model XLM-RoBERTa emerged as the top-performing method with an accuracy of 0.83, an F1-macro score of 0.76, and a MCC of 0.72.

In light of the growing polarization in society, [Lei, Huang, Wang, and Beauchamp \(2022\)](#) steered their focus towards sentence-level media bias analysis, rather than the more common article-level examination. Their motivation stemmed from the recognition that individual sentences within an article can differ substantially in their ideological slant. This paper proposed a method that taps into the inherent discourse structures within news to unveil and analyze these biases. Particularly, by analyzing a sentence's discourse role and its relation to adjacent sentences, the researchers could discern the ideological position of the author even in seemingly neutral sentences. The approach employed the functional news discourse structure and the Penn Discourse TreeBank (PDTB) discourse relations to guide bias identification at the sentence level. They distilled knowledge from these discourse structures into their system. RoBERTa was chosen as the underlying language model, with initial sentence embeddings derived from the sentence start token. Contextual information from the embeddings was then captured using a Bi-LSTM layer. Experiments revealed that integrating both global functional discourse and local rhetorical discourse relations led to notable improvements in recall (0.82–0.86) and precision (0.28–0.34) for bias sentence identification. It is worth noting that their work highlighted the scarcity of datasets dedicated to sentence-level bias detection, citing BASIL ([Fan et al., 2019](#)) and biased-sents ([Lim, Jatowt, Färber, & Yoshikawa, 2020](#)) as the sole available resources that annotate biased sentences within a news context.

[Kim and Johnson \(2022\)](#) introduce CLoSE, a multi-task learning model specifically designed for embedding indicators of frames in news articles for political bias prediction. The essence of framing lies in emphasizing particular aspects of societal issues to mold public sentiment. Detecting such framing constructs provides insights into the dissemination of biased narratives. At the heart of CLoSE is either a BERT-based or RoBERTa-based encoder, which culminates in a pooling layer to generate a sentence embedding, further used for political bias classification. The model harnesses the power of contrastive learning, ensuring that sentences with similar subframe indicators are proximate in the embedding space while maintaining distance from sentences of different subframes. Three potential pooling methods were considered: utilizing the output of the CLS token, averaging all output vectors, and maximizing over the output vectors. However, the mean pooling approach, which had previously showcased superior performance in textual similarity and inference tasks, was selected for CLoSE. Furthermore, the research underscores the model's flexibility and efficiency through its ability to adjust the emphasis between the contrastive learning and classification loss, with experimental results indicating optimal performance when incorporating both objectives. In essence, the integration of subframe indicators markedly boosts political bias classification.

[Cabot, Abadi, Fischer, and Shutova \(2021\)](#) delved into the computational modeling of political discourse, specifically analyzing populist rhetoric, using their “Us vs. Them” dataset. The dataset comprises 6,861 annotated Reddit comments that encapsulate populist attitudes. The study aimed to understand the interplay between populist mindsets, social group targeting (such as Immigrants, Refugees, Muslims, Jews, Liberals, and Conservatives), and associated emotions. For their model architecture, they employed the Robustly Optimized BERT Pre-training Approach (RoBERTa) in its BASE variant. Emphasizing the utility of multi-task learning, the research demonstrated the potential of using emotion and group identification as auxiliary tasks. Their Single-task Learning (STL) baseline for the “UsVsThem” task achieved a Pearson R score of 0.54. Incorporating emotion identification as an auxiliary task, the score marginally increased to 0.55. Further inclusion of group multi-task learning setup led to a higher score of 0.56.

5.2.3. Other methods

In this subsection we will show several lines of research that are different from those shown up to now, but which are also interesting.

The first one is the study of the media bias as a stakeholder mining problem ([Ogawa, Ma, & Yoshikawa, 2011](#)). In this work, the authors propose a network-based model to mine stakeholders (a participant in an event described in a news article) from a given dataset of news. In order to do so, they mined the stakeholders and their interests by analyzing the sentence structure, and developing a relation graph called RelationshipWordNet. The edges of the graph are the relation between two stakeholders. The goal of their approach is to find groups of news that share a common theme and to identify the stakeholders for those groups.

In [Quijote, Zamoras, and Ceniza \(2019\)](#), Quijote et al. examine the prevalence of bias in Philippine political news articles by employing sentiment analysis and the Inverse Reinforcement Model. Leveraging data obtained from popular Philippine news websites such as Inquirer.net, Philstar, Manila Bulletin, The Manila Times, and Journal Online, the articles were subjected to preprocessing to eliminate stop-words and achieve uniform casing. SentiWordNet was then used to assign scores reflecting positivity, negativity, and objectivity to words. Based on the word's highest score, each document was categorized as either positive or negative. Finally, the Inverse Reinforcement Model was deployed to compute deviation values for each outlet's articles, achieving an impressive accuracy of 0.89, precision of 1, recall of 0.60, and F-Measure of 0.75 in bias detection.

A related research line is the one focused on detecting influential nodes in media outlets. In this case, [Patricia Aires, G. Nakamura, and F. Nakamura \(2019\)](#) construct a graph representing the connections between news portals. In this graph, they apply a community detection algorithm, based on network topology, to identify the groups and check if they are composed of websites with similar political orientation.

These same authors have studied the problem of media bias using an information theory approach, using the Shannon entropy for calculating the importance of terms in the vocabulary ([Aires, Freire, & da Silva, 2020](#)). Once they calculate the entropy for each term, they represent the news portals and bias classes using a probability mass function (PMF), that they use in order to compute the dissimilarities between them using the weighted vocabulary. After calculating the dissimilarity matrix for the differences between the speech of each news portal and each class of bias, they fed a classifier with this matrix as features. This classifier uses these dissimilarity scores to distinguish the classes among each other.

In the paper [Rawat and Vadivu \(2022\)](#), the authors emphasize the significance of media bias and its impact on public perception. Though computer science models, notably those in NLP, offer scalable solutions to detect bias, they often lack depth in addressing key questions posed by traditional social science models. A critical challenge is the limited availability of labeled data, especially for Indian political news.

The authors' methodology involves collecting English political news articles from various Indian media outlets using web crawlers, and categorizing the news based on their political alignments. The project distinguishes between left-biased (favoring left ideology parties) and right-biased news. By utilizing clustering algorithms such as K-means, PCA, and DBSCAN, the news articles are grouped according to the political parties they seem to favor. Each article is then processed sentence-wise, with unnecessary content like punctuation and stop words removed. Sentiment analysis determines whether each sentence possesses a positive, negative, or neutral sentiment. A bias score for each article is then computed as:

$$\text{Bias Score} = \frac{\text{No. of positive/negative/neutral sentences}}{\text{Total no. of sentences in article}}$$

With this approach, the authors aim to generate comprehensive reports pinpointing which Indian media houses display bias towards specific political parties, thereby shedding light on the broader dynamics of media-driven polarization.

The paper by [de Arruda, Roman, and Monteiro \(2020\)](#) sheds light on the multi-dimensional nature of media bias. The researchers define bias

through a tripartite model encompassing selection bias, coverage bias, and statement bias. Their strategy hinges on outlier detection, positing bias as a conspicuous deviation from typical behavior. Their findings reveal the capability to not only discern bias in distinct outlets but also to understand its origins, intensity, and interactions with other dimensions, painting a holistic image of the examined phenomenon. One might question how we can ever ascertain the presence of coverage bias or even fathom the true distribution of events. A noteworthy approach, as exemplified by [de Arruda et al. \(2020\)](#), treats bias detection akin to an outlier detection problem. If outlets markedly diverge in their fact coverage, bias detection might be feasible. But a conundrum arises when all outlets exhibit aligned behavior. This raises pertinent questions on the feasibility of detecting various bias forms, considering the data at our disposal. Such concerns underscore the need for innovative methodologies and deeper insights into bias detection's multifaceted landscape.

In [Jiang, Wang, Song, and Maynard \(2020\)](#), the authors delve into the rising influence of commercial pressures on news media, resulting in more sensational and dramatized biases within newspaper articles. The subsequent bias can lead to polarized opinions, potentially misleading readers. This paper explores learning models tailored for news bias detection, particularly focusing on the integration of Latent Dirichlet Allocation (LDA) distributions. These distributions are anticipated to enhance the feature space by introducing word co-occurrence distribution and local topic probability for each document. In the proposed models, the LDA features are integrated both on the sentence and document levels. The experimental evaluations, conducted on a hyperpartisan newspaper article detection task, reveal that hierarchical models incorporating LDA features show superior performance compared to their non-hierarchical counterparts.

Lastly, it is worth mentioning the research of [Boxell \(2018\)](#). In this work, he creates a dataset of news images, extracts the emotions using Microsoft's Emotion API, and with these features trains a linear regression model. The idea behind this research is to better understand the degree to which nonverbal bias is present across online media, and how it impacts the political beliefs and feelings of the citizens.

5.2.4. Comparative analysis of media bias detection techniques

The field of media bias detection has seen rapid advancements in recent years, with a plethora of techniques being developed to tackle this complex issue. These techniques come with their own sets of advantages and limitations. The purpose of this subsection is to offer a detailed comparative analysis of the different methodologies, highlighting their respective strengths and weaknesses. This analysis aims to serve as a comprehensive guide for researchers, aiding them in choosing the most suitable techniques for their specific media bias detection projects.

Classical machine learning methods

Classical machine learning methods present the following strengths:

- **Interpretability:** Traditional machine learning algorithms such as decision trees, and logistic regression are highly interpretable. They provide valuable insights into the importance of different features, thereby aiding in the understanding of the model's decision-making process.
- **Computational Efficiency:** These algorithms are generally less computationally demanding, making them an ideal choice for projects with limited computational resources or for quick prototyping.

However, they also present the following weaknesses:

- **Limited Complexity:** While effective for simpler tasks, classical machine learning methods often struggle to capture the nuanced semantics and complexities inherent in natural language. This often necessitates extensive feature engineering to achieve satisfactory performance.

- **Dataset Sensitivity:** The performance of these methods can be highly sensitive to the quality and distribution of the dataset, often requiring careful preprocessing and feature selection to ensure robustness across different domains.

Reported speech methods

Reported speech methods present the following strengths:

- **Contextual Analysis:** Reported speech methods excel in analyzing the context in which statements are made, providing a nuanced understanding of bias.
- **Narrative Structure:** These methods can reveal the narrative structure of an article, helping to identify framing techniques that may indicate bias.

However, they also present the following weaknesses:

- **Complexity:** Parsing and understanding reported speech can be computationally intensive and may require advanced natural language processing techniques.
- **Ambiguity:** The interpretation of reported speech can sometimes be ambiguous, making it challenging to draw definitive conclusions about bias.

Deep learning methods

Deep learning methods present the following strengths:

- **Semantic Understanding:** Deep learning models, especially recurrent neural networks (RNNs) and transformers, have shown remarkable capabilities in understanding the intricate semantics of natural language. They often outperform classical methods in tasks that require a deep understanding of context and semantics.
- **Feature Learning:** One of the most significant advantages of deep learning models is their ability to automatically learn relevant features from the data, eliminating the need for manual feature engineering to a large extent.

However, they also present the following weaknesses:

- **Computational Cost:** The training and deployment of deep learning models often require specialized hardware and are computationally expensive, which can be a limiting factor for some projects.
- **Interpretability:** Deep learning models are often criticized for being "black boxes", as they offer limited interpretability compared to classical methods. This can be a significant drawback when the model's decision-making process needs to be fully understood.

Hybrid approaches

Hybrid approaches present the following strengths:

- **Balanced Performance:** Hybrid models that combine classical machine learning algorithms with deep learning techniques can offer a balanced approach. They leverage the interpretability and computational efficiency of classical methods while benefiting from the semantic understanding capabilities of deep learning models.

However, they also present the following weaknesses:

- **Complexity:** The process of integrating different types of models can introduce additional layers of complexity, both in terms of model architecture and the training process. This can make the model more challenging to optimize and interpret.

6. Datasets for media bias detection

There are various definitions of media bias in the literature, and as we have already seen, it can be detected at various levels. Therefore, the researchers have followed different approaches when it comes to retrieving and generating datasets for the task of automatic media bias detection. In this section, we chronologically list the currently available datasets. All of them are summarized in Table 3.

Budak et al. (2016) created a dataset with news of fifteen different US news outlets in order to investigate the selection and framing of political issues. They collected more than 800,000 news items published in 2013. They ran two machine learning models to eliminate news that were not political (for example, sports news, finance, technology, etc.). After applying these filters, the dataset was reduced to the 14 percent. Then, a random subset of 10,502 news instances were manually annotated using Amazon Mechanical Turk. Each instance includes two annotations, one on bias towards the Democratic Party (very positive, positive, somewhat positive, neutral, somewhat negative, negative, and very negative), and another on bias towards the Republican Party (very positive, positive, somewhat positive, neutral, somewhat negative, negative, and very negative).

The use of disinformation is prominent during a war conflict (Nimmo, 2015). Cremisini et al. (2019) collected and manually annotated 4,538 news articles that report on the situation in Ukraine in 2014–2015, with particular focus on the annexation of Crimea by Russia in 2014, the military conflicts in Southeastern Ukraine, and the Maidan protests. They categorized the bias of each article based on its country of origin, placing each country into pro-Western, Neutral, or pro-Russian bias classes on the basis of known geopolitical alliances.

In 2017, Horne et al. (2018) created NEwS LANDscape (NELA2017), a political news dataset, containing more than 136,000 articles from 92 news sources, collected data from around seven months in 2017. NELA2017 was created with the intention of having a large and diverse dataset for the detection of media bias. The dataset has been updated in later years, being able to access versions of the dataset with data from 2018 (NELA-GT-2018), 2019 (NELA-GT-2019), 2020 (NELA-GT-2020), and 2021 (NELA-GT-2020); the latter having more than 1,850,000 instances from 367 different sources (Nørregaard, Horne, & Adali, 2019) (Gruppi, Horne, & Adali, 2022). Also, they presented in 2022 NELA-Local, a dataset of over 1.4M online news articles from 313 local U.S. news outlets covering a geographically diverse set of communities across the United States (Horne et al., 2022). These datasets have both political news, news related to the US elections, as well as news about COVID. They annotate each instance of the dataset according to its veracity (reliable, mixed, and unreliable), obtaining the information from the Media Bias/Fact Check (MBFC) website.

In 2018, Baly et al. (2018) created a dataset of 1,066 news instances to detect fake news by studying media bias. To annotate the data, they labeled each news item according to the political bias classification (extreme-left, left, center-left, center, center-right, right, extreme-right) that the Media Bias/Fact Check (MBFC) website assigns to the medium that published the news. In 2020, the same authors Baly, Karadzhov et al. (2020) created a new dataset following the same methodology of 864 instances. Both datasets can be used for automatic media bias detection tasks, both at the article-level and at the medium-level.

In 2019, Fan et al. (2019) published BASIL (Bias Annotation Spans on the Informational Level), a dataset of 300 news articles annotated with 1,727 bias spans. The dataset uses 100 triplets of articles, each reporting the same event from three outlets of different political ideology. Annotations were conducted manually both on document-level and sentence-level.

BASIL was not the first dataset created for claim-level media bias detection, the first antecedent was published by Baumer, Elovic, Qin, Polletta, and Gay (2015). This dataset has 74 news items, including words and phrases that Amazon Mechanical Turk annotators annotated according to their perception of framing.

In 2019, Hamborg, Donnay et al. (2019) created NewsWCL50, an open dataset inspired by BASIL for the evaluation of methods to automatically identify bias by word choice and labeling (Hamborg, Donnay et al., 2019). It is a claim-level dataset that contains 50 articles that cover 10 political news events, each reported on by 5 online US news outlets. On average, each article has 170 manual annotations.

Another claim-level dataset manually annotated by 10 annotators is MBIC (Media Bias Including Characteristics) (Spinde, Rudnitckaia, Sinha et al., 2021). MBIC is a dataset that contains 1,700 phrases from 1,000 different articles that potentially contain bias by word choice. The scraped articles correspond to 8 US media outlets, and cover 14 controversial topics (abortion, coronavirus, elections 2020, environment, gender, gun control, immigration, Donald Trump's presidency, vaccines, white nationalism), and four not so controversial topics (student debt, international politics, and world news, middle class, sport).

They annotated the dataset manually via Amazon Mechanical Turk. Labels are: biased, non-biased, and no agreement. They created another dataset built on top of the MBIC data set called BABE with 3,700 instances balanced among topics and outlets, containing media bias labels on the word and sentence-level.

Lim et al. (2018b) studied word-level bias by comparing words across the content of different news articles that report the same news event. They collected articles from various news outlets using Google News, creating a dataset of 89 news articles with 1,235 sentences and 2,542 unique words from 83 news outlets. For the annotation process, they used a crowdsourcing platform called Figure Eight with 60 workers participating in the task.

In 2020, the same authors created another dataset (Biased-Sents-Annotation) (Lim et al., 2020) for fostering bias-detection studies on claim-level, with the objective of helping designing methods to detect bias triggers. They selected 4 topics covering issues on the English news reported between September 2017 and May 2018. The four topics that make up the dataset are: (1) "Trump Clashes With Sports World Over Player Protests", (2) "Facebook critics want regulation, investigation after data misuse", (3) "Tillerson says U.S. ready to talk to North Korea; Japan wants pressure", and (4) "Republican lawmaker commits suicide amid sexual molestation allegations". The news were again collected from Google news and annotated manually via Figure Eight workers. The resulting dataset consists of 371 articles for the Trump issue, 103 articles for the Facebook event, 39 articles for North Korea, and 44 news articles for the republican lawmaker event. The labels are: neutral, slightly biased, biased, and very biased.

Finally, for the SemEval-2014 conference, a dataset was created to detect hyperpartisan news (highly politically biased) (Kiesel et al., 2019). Each news item is labeled "no hyperpartisan content", "mostly unbiased", "non-hyperpartisan content", "not sure", "fair amount of hyperpartisan content", or "extreme hyperpartisan content". This dataset has 1,273 news items manually annotated by experts, and 75,400 instances automatically annotated according to the assessment obtained from the Media Bias/Fact Check (MBFC) website.

Analyzing the 17 datasets that we have listed, we can see how in recent years, thanks to datasets such as BASIL (Fan et al., 2019), NewsWCL50 (Hamborg, Donnay et al., 2019), MBIC (Spinde, Rudnitckaia, Sinha et al., 2021), and BABE (Spinde, Plank et al., 2021) we are able to study the problem of automatic analysis of media bias at the claim or sentence level, and not only at the article level.

We can also see how the problems identified in Horne et al. (2018) still remain: (1) the available datasets are small both in size and in sources, (2) the news from the available datasets reports few distinct events, and (3) the available datasets only collect news with a lot of engagement. To these three issues we can also add the fact that all the datasets are in English and none in another language, as well as the domain to which most of the news in the datasets belongs is the political domain, and that they are highly influenced by news and US politics.

Table 3
Overview of media bias detection datasets.

Dataset	Dataset size	Instance	Rating	Annotation	Year	Reference
Fair and Balanced dataset	10,502	Article-level	Multilabel (bias towards US political parties)	Manually (Crowd)	2016	Budak et al. (2016)
Bias of News Media Sources dataset	1,066	Article/Media-level	Political compass	Scraped from MB/FC website	2018	Baly et al. (2018)
News Media Profiling dataset	864	Article/Media-level	Political compass	Scraped from MB/FC website	2020	Baly, Karadzhov et al. (2020)
NELA2017	136k	Article-level	Veracity	Scraped from MB/FC website	2017	Baly et al. (2018)
NELA-GT-2018	731k	Article-level	Veracity	Scraped from MB/FC website	2018	Nørregaard et al. (2019)
NELA-GT-2019	1.12M	Article-level	Veracity	Scraped from MB/FC website	2019	Gruppi, Horne, and Adali (2020)
NELA-GT-2020	1.12M	Article-level	Veracity	Scraped from MB/FC website	2020	Gruppi, Horne, and Adali (2021)
NELA-GT-2021	1.8M	Article-level	Veracity	Scraped from MB/FC website	2021	Gruppi et al. (2022)
NELA-Local	1.4M	Article-level	Veracity	Scraped from MB/FC website	2022	Horne et al. (2022)
MBIC	1,700	Sentence-level	Multilabel (bias and political compass)	Manually (Crowd)	2021	Spinde, Rudnitskaia, Sinha et al. (2021)
Language of Framing in Political News dataset	74	Sentence-level	Annotated text	Manually (Crowd)	2015	Baumer et al. (2015)
Hyperpartisan News Detection	1,273/75,400	Article-level	Hyperpartisanship	Manual (Experts) + Automatic (MBFC)	2019	Kiesel et al. (2019)
BASIL	300	Sentence/Article-level	Annotated text/political compass	Manually	2019	Fan et al. (2019)
Crisis in the Ukraine dataset	4,538	Article-level	Stance	Manually (Expert)	2015	Cremisini et al. (2019)
Biased-Sents-Annotation	557	Sentence-level	Annotated text	Manually (Crowd)	2020	Lim et al. (2020)
Characteristics of biased sentences dataset	89	Word/Sentence-level	Annotated text	Manually (Crowd)	2018	Lim et al. (2018b)
NewsWCL50	8,656	Sentence-level	Annotated text	Manually	2020	Hamborg, Zhukova et al. (2019)
BABE	3,700	Sentence-level	Annotated text	Manually (Experts)	2021	Spinde, Plank et al. (2021)

Given these observations and the inherent limitations they impose on comprehensive research, we are firmly of the belief that there is an imperative need to curate new datasets. These datasets should not only be more expansive but also richly diverse, encompassing a multitude of languages, a broader spectrum of events, varied domains, and a range of socio-cultural contexts to ensure a holistic representation.

7. Discussion and future work

In this systematic review, we have presented a comprehensive overview of the state of the art in automatic media bias detection. We have seen that the task is far from being solved, as there is still a lot of work to be done in order to improve the performance of the models and to create more diverse datasets.

Several challenges persist in the realm of automatic media bias detection. The complexity of this task stems from its multi-dimensional nature; it involves the detection of political bias, factual accuracy, and veracity. Each of these aspects can be further segmented into different layers of analysis, such as document-level or sentence-level evaluations.

The prevalent methodologies currently employed lean heavily on machine learning and deep learning techniques. These often incorporate lexicons, feature engineering, and word embeddings. Despite these advancements, there remains an evident gap in creating models that effectively discern the nuanced elements of media bias.

A significant limitation is the current lack of diverse datasets dedicated to this task. Existing datasets frequently suffer from size constraints, are predominantly oriented towards particular domains like politics, and are primarily in English.

7.1. Limitations and considerations

This research endeavors to provide a comprehensive understanding of media bias detection techniques, especially from a machine learning perspective. However, it is crucial to highlight some limitations and considerations that come with our approach:

- Machine learning bias:** Primarily, our study is strongly focused on machine learning techniques. While machine learning offers powerful tools for media bias detection, there exist other methods which might not rely heavily on machine learning, or might not use it at all. The scope of this study largely stemmed from our database queries, which had a predisposition towards machine learning methods. It is vital to understand that while machine learning is a dominant tool in this field, it is not the exclusive tool.
- Exclusivity of methods:** This review delineates the media bias detection task by breaking down methodologies into non-neural and neural-based models, and other disparate methods. While this categorization helps in providing a structured overview, some methods might overlap between categories or could be a hybrid of multiple techniques, which might not be captured exclusively in our categorization.
- Contextual limitations:** Bias in media is multi-faceted and deeply contextual. Some of the described techniques, especially the linguistic-based methods, might not capture bias that arises from what is omitted from a report, rather than what is stated. The nuance of media bias, in many instances, requires deep

contextual understanding which might not always be feasible with algorithmic approaches.

4. **Database limitations:** Our search spanned databases like Google Scholar, Scopus, and ACL Anthology. While these are comprehensive repositories, there might be relevant works in other databases or publications that we might have missed due to our database selection.

7.2. Future work

Drawing from the current landscape of media bias detection research, we underline the following prospective research avenues to deepen our understanding of media bias and enhance automated detection techniques:

Develop a gold standard dataset

Our review reveals that there is a lack of publicly available benchmark datasets for automated media bias detection. This makes it difficult to compare and contrast the results from different studies and evaluate the generalizability of the proposed methods. We encourage future work to generate a gold standard dataset that is more diverse in terms of content type and the domain of media bias.

Develop datasets in multiple languages, or a multilingual dataset

Our review reveals that most of the existing datasets are English-only. The lack of multilingual datasets prevents generalizability of the proposed methods across different languages, especially for low-resource ones. We encourage future work to generate datasets in multiple languages or a multilingual dataset.

Develop unified evaluation metrics

We also find that there is no unified evaluation metric for automated media bias detection. Different studies adopt different metrics that may not be directly comparable. This makes it difficult to compare and contrast the results from different studies and evaluate the generalizability of the proposed methods.

Use of contextual information via knowledge bases

We identify that there is a lack of work using context information from knowledge bases such as DBpedia (Auer et al., 2007) and YAGO (Rebele et al., 2016). Using context information from knowledge bases can help in understanding the context of a news article and, in turn, enhance the performance of the proposed methods. Taking into account the context of a news article can help in understanding the purpose of a claim, for example, if we analyze some news reporting the 2022 Russian invasion of Ukraine, the context may provide information about the event, that would help to understand if the news report is biased or not. We encourage future work to use context information from knowledge bases.

Explore resources beyond text

Although a great deal of research has been conducted on automated media bias detection, most of the existing studies focus on text-based methods. Other types of resources such as images, videos, and social network data have been largely overlooked. Also, while the majority of existing studies focus on the automated detection of media bias in a single media type (e.g., text), we find that the majority of studies adopt a combination of multiple features to improve the performance of the automated systems. However, there is little research that focuses on cross-media methods that can utilize multiple types of resources (e.g., text, images, videos, and social network data) describing the same event/fact to improve the performance of the automated systems.

Explore explainable media bias detection methods

The majority of existing studies focus on the automated detection of media bias without analyzing the forms of media bias that occurs in the news, though we find that there is little research that focuses on explainable media bias detection methods. With explainable artificial intelligence (XAI) methods (Došilović, Brčić, & Hlupić, 2018), you could not only identify bias in news, but also give the user information about why that text is biased. Explaining the rationale for why an article is deemed biased and how the system made that determination can help users understand how to interpret that information. If users see that their own opinions are challenged and understand how that happened, they may be more likely to consider and question their own beliefs, which is a key goal of journalism, and a way to avoid falling into bubble filters.

There are several ways to make artificial intelligence systems explainable, e.g., through providing a justification of the decision that is made, or the use of human-friendly representations, such as natural language description. XAI systems are beginning to be created in the field of fake news detection (Madhu Kumar, Chacko, et al., 2022), but nothing has yet been developed for the detection of media bias.

8. Conclusions

Disinformation and media bias are problems that have become even more visible in recent years due to the spread of the social media, and pose a threat to the democratic processes of the states. It is important to have, in the hands of all citizens, tools that allow them to distinguish information designed to inform from other types that are deliberately biased. To achieve this, it is important to understand these problems, and the different cues that characterize them.

In this paper, we have presented a theoretical framework capable of comparing different disinformation problems such as media bias, propaganda or fake news; we have also defined, classified and characterized media bias; lastly, we have reviewed the current state of automated media bias detection research.

As we have already seen in the previous section, there is still a lot of work to be done towards the standardization of both datasets and metrics in order to have a common benchmark with which to compare different methods. While the current detection approaches are mainly deep learning methods, the future of media bias detection lies in the development of methods that are both explainable and transferable.

Finally, it should be noted that there is still no agreement on what is meant by media bias, and that the current approaches are mainly focused on the statement bias, which is just one of the several types of media bias that exist. This means that there is still a lot of work to be done to develop techniques capable of recognizing the other types of media bias defined in this paper.

CRediT authorship contribution statement

Francisco-Javier Rodrigo-Ginés: Conceptualization, Investigation, Resources, Writing – original draft. **Jorge Carrillo-de-Albornoz:** Conceptualization, Validation, Writing – review & editing, Supervision. **Laura Plaza:** Conceptualization, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article

Acknowledgments

This work has been financed by the European Union (NextGenerationEU funds) through the “Plan de Recuperación, Transformación y Resiliencia”, by the Ministry of Economic Affairs and Digital Transformation and by the UNED University. It has also been financed by the Spanish Ministry of Science and Innovation (project FairTransNLP (PID2021-124361OB-C32)) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. All authors have read and approved the final version of the manuscript.

References

- Agrawal, S., Gupta, K., Gautam, D., & Mamidi, R. (2022). Towards detecting political bias in Hindi news articles. In *Proceedings of the 60th annual meeting of the association for computational linguistics: student research workshop* (pp. 239–244).
- Aires, V. P., Freire, J., & da Silva, A. S. (2020). An information theory approach to detect media bias in news websites.
- Al-Sarraj, W. F., & Lubbad, H. M. (2018). Bias detection of Palestinian/Israeli conflict in western media: A sentiment analysis experimental study. In *2018 international conference on promising electronic technologies* (pp. 98–103). IEEE.
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational Influence Processes*, 58, 295–303.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, & P. Cudré-Mauroux (Eds.), *The semantic web* (pp. 722–735). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Baker, B. H., Graham, T., & Kaminsky, S. (1996). *How to identify, expose & correct liberal media bias* (2nd ed.). Alexandria, Va.: Media Research Center, OCLC 42464501.
- Baly, R., Da San Martino, G., Glass, J., & Nakov, P. (2020). 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* (pp. 4982–4991).
- Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3528–3539).
- Baly, R., Karadzhov, G., An, J., Kwak, H., Dinkov, Y., Ali, A., et al. (2020). 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* (pp. 3364–3374).
- Baly, R., Karadzhov, G., Saleh, A., Glass, J., & Nakov, P. (2019). Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. arXiv preprint arXiv:1904.00542.
- Baraniak, K., & Sydow, M. (2018). News articles similarity for automatic media bias detection in polish news portals. In *2018 federated conference on computer science and information systems* (pp. 21–24). IEEE.
- Baumer, E. P., Elovic, E., Qin, Y., Polletta, F., & Gay, G. (2015). *Framing annotation data for news articles*. North American Chapter of the Association for Computational Linguistics (NAACL).
- Bennett, W. L., & Iyengar, S. (2008). A new era of minimal effects? The changing foundations of political communication. *Journal of Communication*, 58(4), 707–731.
- Best, C., van der Goot, E., Blackler, K., Garcia, T., & Horby, D. (2005). Europe media monitor. In *Web intelligence action - technical report EUR221 73 EN*. European Commission.
- Boudana, S. (2011). A definition of journalistic objectivity as a performance. *Media, Culture & Society*, 33(3), 385–398.
- Boxell, L. (2018). Slanted images: Measuring nonverbal media bias.
- Bozell, L. B., & Baker, B. H. (1990). *And that's the way it isn't: A reference guide to media bias*. Media Research Center.
- Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1), 250–271.
- Cabot, P.-L. H., Abadi, D., Fischer, A., & Shutova, E. (2021). Us vs. Them: A dataset of populist attitudes, news bias and emotions. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume* (pp. 1921–1945).
- Cao, L., Xu, P., & Shang, W. (2021). A text-based mining approach for real estate policy impact monitoring and analysis. In *2021 IEEE international conference on big data (big data)* (pp. 1575–1581). IEEE.
- Chen, W.-F., Al Khatib, K., Stein, B., & Wachsmuth, H. (2020). Detecting media bias in news articles using Gaussian bias distributions. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 4290–4300).
- Chen, W.-F., Wachsmuth, H., Al Khatib, K., & Stein, B. (2018). Learning to flip the bias of news headlines. In *Proceedings of the 11th international conference on natural language generation* (pp. 79–88).
- Choi, Y., & Wiebe, J. (2014). +/-Effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1181–1191).
- Cremisini, A., Aguilar, D., & Finlayson, M. A. (2019). 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* (pp. 173–183). Springer.
- Cruz, A. F., Rocha, G., & Cardoso, H. L. (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* (pp. 107–112).
- Da San Martino, G., Cresci, S., Barrón-Cedeño, A., Yu, S., Di Pietro, R., & Nakov, P. (2021). A survey on computational propaganda detection. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 4826–4832).
- D'Alessio, D., & Allen, M. (2000). Media bias in presidential elections: A meta-analysis. *Journal of Communication*, 50(4), 133–156.
- de Arruda, G. D., Roman, N. T., & Monteiro, A. M. (2020). Analysing bias in political news. *Journal of Universal Computer Science*, 26(2), 173–199.
- De Witte, M. (2022). Groupthink gone wrong: Stanford scholars show how assumptions about electability undermine women political candidates. *Stanford News Service*, URL <https://news.stanford.edu/press-releases/2022/02/02/groupthink-gone-tical-candidates/>.
- Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st international convention on information and communication technology, electronics and microelectronics* (pp. 0210–0215). <http://dx.doi.org/10.23919/MIPRO.2018.8400040>.
- Estrada-Cuzcano, A., Alfaro-Mendives, K., & Saavedra-Vásquez, V. (2020). Disinformation y misinformation, posverdad y fake news: Precisiones conceptuales, diferencias, similitudes y yuxtaposiciones. *Información, Cultura Y Sociedad*, (42), 93–106.
- Fan, L., White, M., Sharma, E., Su, R., Choubey, P. K., Huang, R., et al. (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* (pp. 6343–6349).
- Färber, M., Qurdina, A., & Ahmedi, L. (2019). Team peter brinkmann at semeval-2019 task 4: Detecting biased news articles using convolutional neural networks. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 1032–1036).
- Geng, Y. (2022). Media bias detecting based on word embedding. *Highlights in Science, Engineering and Technology*, 12, 61–67.
- Gentzkow, M., & Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2), 280–316.
- Geske, J. (2016). Riot vs. Revelry: News bias through visual media. *Teaching Media Quarterly*, 4(1).
- Gilens, M. (1996). Race and poverty in americapublic misperceptions and the american news media. *Public Opinion Quarterly*, 60(4), 515–541.
- Gruppi, M., Horne, B. D., & Adali, S. (2020). NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles. <http://dx.doi.org/10.48550/ARXIV.2003.08444>, arXiv. URL <https://arxiv.org/abs/2003.08444>.
- Gruppi, M., Horne, B. D., & Adali, S. (2021). NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles. <http://dx.doi.org/10.48550/ARXIV.2102.04567>, arXiv. URL <https://arxiv.org/abs/2102.04567>.
- Gruppi, M., Horne, B. D., & Adali, S. (2022). NELA-GT-2021: A large multi-labelled news dataset for the study of misinformation in news articles. <http://dx.doi.org/10.48550/ARXIV.2203.05659>, arXiv. URL <https://arxiv.org/abs/2203.05659>.
- Gupta, V., Jolly, B. L. K., Kaur, R., & Chakraborty, T. (2019). Clark kent at SemEval-2019 task 4: Stylometric insights into hyperpartisan news detection. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 934–938).
- Hajare, P., Kamal, S., Krishnan, S., & Bagavathi, A. (2021). A machine learning pipeline to examine political bias with congressional speeches. In *2021 20th IEEE international conference on machine learning and applications* (pp. 239–243). IEEE.
- Hamborg, F. (2020). 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* (pp. 79–87).
- Hamborg, F., Donnay, K., & Gipp, B. (2019). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4), 391–415.
- Hamborg, F., Zhukova, A., & Gipp, B. (2019). Automated identification of media bias by word choice and labeling in news articles. In *2019 ACM/IEEE joint conference on digital libraries* (pp. 196–205). <http://dx.doi.org/10.1109/JCDL.2019.00036>.
- Harzing, A.-W. (2010). *The publish or perish book*. Australia: Tarma Software Research Pty Limited Melbourne.
- Holsanova, J., Rahm, H., & Holmqvist, K. (2006). Entry points and reading paths on newspaper spreads: Comparing a semiotic analysis with eye-tracking measurements. *Visual Communication*, 5(1), 65–93.
- Horne, B. D., Gruppi, M., Joseph, K., Green, J., Wihbey, J. P., & Adali, S. (2022). NELA-local: A dataset of US local news articles for the study of county-level news ecosystems. In *Proceedings of the international AAAI conference on web and social media*, Vol. 16 (pp. 1275–1284).
- Horne, B. D., Khedr, S., & Adali, S. (2018). 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* (pp. 518–527).

- Hube, C., & Fetahu, B. (2018). Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018* (pp. 1779–1786).
- Hube, C., & Fetahu, B. (2019). Neural based statement classification for biased language. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 195–203).
- Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014). Political ideology detection using recursive neural networks. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1113–1122).
- Jiang, T., Guo, Q., Chen, S., & Yang, J. (2019). What prompts users to click on news headlines? Evidence from unobtrusive data analysis. *Aslib Journal of Information Management*.
- Jiang, Y., Wang, Y., Song, X., & Maynard, D. (2020). Comparing topic-aware neural networks for bias detection of news. In *ECAI 2020* (pp. 2054–2061). IOS Press.
- Kameswari, L., Sravani, D., & Mamidi, R. (2020). Enhancing bias detection in political news using pragmatic presupposition. In *Proceedings of the eighth international workshop on natural language processing for social media* (pp. 1–6).
- Kang, H., & Yang, J. (2022). Quantifying perceived political bias of newspapers through a document classification technique. *Journal of Quantitative Linguistics*, 29(2), 127–150.
- Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*.
- Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., et al. (2019). Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 829–839).
- Kim, M. Y., & Johnson, K. (2022). CloSE: Contrastive learning of subframe embeddings for political bias classification of news media. In *Proceedings of the 29th international conference on computational linguistics* (pp. 2780–2793).
- Kohlmeier, M. (2018). Overblown claims. *BMJ Nutrition, Prevention & Health*, 1(1), 5.
- Krestel, R., Wall, A., & Nejdl, W. (2012). Treehugger or petrolhead? Identifying bias by comparing online news articles with political speeches. In *Proceedings of the 21st international conference on world wide web* (pp. 547–548).
- Krieger, J.-D., Spinde, T., Ruas, T., Kulshrestha, J., & Gipp, B. (2022). A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the 22nd ACM/IEEE joint conference on digital libraries* (pp. 1–7).
- Kuculo, T., Gottschalk, S., & Demidova, E. (2022). : A multilingual knowledge graph of quotes. In *European semantic web conference* (pp. 353–369). Springer.
- Law, J. (2020). Looking for media bias in coverage of Trump's Covid diagnosis. *JLaw's R Blog*, URL <https://jlaw.netlify.app/2020/10/07/looking-for-media-bias-in-coverage-of-trump-s-covid-diagnosis/>.
- Lazaridou, K., & Krestel, R. (2016). Identifying political bias in news articles. *Bulletin of the IEEE TCDD*, 12.
- Lazaridou, K., Krestel, R., & Naumann, F. (2017). Identifying media bias by analyzing reported speech. In *2017 IEEE international conference on data mining* (pp. 943–948). IEEE.
- Lei, Y., Huang, R., Wang, L., & Beauchamp, N. (2022). Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10040–10050).
- Lim, S., Jatowt, A., Färber, M., & Yoshikawa, M. (2020). Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the 12th language resources and evaluation conference* (pp. 1478–1484).
- Lim, S., Jatowt, A., & Yoshikawa, M. (2018a). Towards bias inducing word detection by linguistic cue analysis in news. In *DEIM forum* (pp. C1–3).
- Lim, S., Jatowt, A., & Yoshikawa, M. (2018b). Understanding characteristics of biased sentences in news articles. In *CIKM workshops*.
- Lin, Y.-R., Bagrow, J., & Lazer, D. (2011). More voices than ever? quantifying media bias in networks. In *Proceedings of the international AAAI conference on web and social media*, Vol. 5, no. 1 (pp. 193–200).
- Madhu Kumar, S., Chacko, A. M., et al. (2022). Towards smart fake news detection through explainable AI. arXiv e-prints, arXiv:2207.
- Mastrine, J., Sowers, K., Alhariri, S., & Nilsson, J. (2022). How to spot 16 types of media bias. *AllSides*, URL <https://www.allsides.com/media-bias/how-to-spot-types-of-media-bias>.
- Moher, D., Altman, D. G., Liberati, A., & Tetzlaff, J. (2011). PRISMA statement. *Epidemiology*, 22(1), 128.
- Mullainathan, S., & Shleifer, A. (2002). *Media bias: Working paper series no. 9295*, National Bureau of Economic Research, <http://dx.doi.org/10.3386/w9295>, URL <http://www.nber.org/papers/w9295>.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.
- Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015). Quotos: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th international conference on world wide web* (pp. 798–808).
- Nimmo, B. (2015). *Anatomy of an info-war: how Russia's propaganda machine works, and how to counter it*, Vol. 15. Central European Policy Institute.
- Nørregaard, J., Horne, B. D., & Adali, S. (2019). NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13 (pp. 630–638).
- Ogawa, T., Ma, Q., & Yoshikawa, M. (2011). News bias analysis based on stakeholder mining. *IEICE Transactions on Information and Systems*, 94(3), 578–586.
- Özge, C., & Ercan, G. S. (2020). Discursive functions of reported speech in turkish op-ed articles. *Dilbilim Araştırmaları Dergisi*, 31(2), 265–288.
- Palić, N., Vladika, J., Čubelić, D., Lovrenčić, I., Buljan, M., & Šnajder, J. (2019). Takelab at SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 995–998).
- Pant, K., Dadu, T., & Mamidi, R. (2020). Towards detection of subjective bias using contextualized word embeddings. In *Companion proceedings of the web conference 2020* (pp. 75–76). Taipei Taiwan: ACM, <http://dx.doi.org/10.1145/3366424.3382704>, URL <https://dl.acm.org/doi/10.1145/3366424.3382704>.
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. penguin UK.
- Park, S., Lee, K.-S., & Song, J. (2011). 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* (pp. 340–349).
- Parker-Bass, B., Fette, I., Mans, P., Seth, M., Sullivan, J., & Washburn, P. (2022). News bias explored. <http://websites.umich.edu/~newsbias/>. (Accessed on 27 Dec 2022).
- Patricia Aires, V., G. Nakamura, F., & F. Nakamura, E. (2019). A link-based approach to detect media bias in news websites. In *Companion proceedings of the 2019 world wide web conference* (pp. 742–745).
- Pothast, M., Köpsel, S., Stein, B., & Hagen, M. (2016). Clickbait detection. In *European conference on information retrieval* (pp. 810–817). Springer.
- Preoțiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 729–740).
- Quijote, T., Zamoras, A., & Ceniza, A. (2019). Bias detection in Philippine political news articles using SentiWordNet and inverse reinforcement model. *IOP Conference Series: Materials Science and Engineering*, 482(1), Article 012036.
- Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). 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* (pp. 2931–2937).
- Rawat, S., & Vadivu, G. (2022). Media bias detection using sentimental analysis and clustering algorithms. In *Proceedings of international conference on deep learning, computing and intelligence: ICDICI 2021* (pp. 485–494). Springer.
- Rebele, T., Suchanek, F., Hoffart, J., Biega, J., Kuzey, E., & Weikum, G. (2016). YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *International semantic web conference* (pp. 177–185). Springer.
- Rodrigo-Ginés, F.-J., Carrillo-de Albornoz, J., & Plaza, L. (2021). UNEDBiasTeam at IberLEF 2021's EXIST task: Detecting sexism using bias techniques. In *IberLEF@SEPLN* (pp. 522–532).
- Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and Knowledge*, 103–135.
- Ruiz, J. B., & Bell, R. A. (2014). Understanding vaccination resistance: Vaccine search term selection bias and the valence of retrieved information. *Vaccine*, 32(44), 5776–5780.
- Saez-Trumper, D., Castillo, C., & Lalmas, M. (2013). Social media news communities: Gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on information & knowledge management* (pp. 1679–1684).
- Samory, M., Cappelleri, V.-M., & Peserico, E. (2017). Quotes reveal community structure and interaction dynamics. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 322–335).
- Shannon, C. E., & Weaver, W. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Shelke, S., & Attar, V. (2019). Source detection of rumor in social network—a review. *Online Social Networks and Media*, 9, 30–42.
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, Article 132306.
- Sims, M., & Bamman, D. (2020). Measuring information propagation in literary social networks. In *Proceedings of the 2020 conference on empirical methods in natural language processing*.
- Sinha, M., & Dasgupta, T. (2021). 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* (pp. 3418–3422).
- Spinde, T., Hamborg, F., & Gipp, B. (2020). An integrated approach to detect media bias in german news articles. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020* (pp. 505–506).
- Spinde, T., Krieger, J.-D., Ruas, T., Mitrović, J., Götz-Hahn, F., Aizawa, A., et al. (2022). Exploiting transformer-based multitask learning for the detection of media bias in news articles. In *International conference on information* (pp. 225–235). Springer.
- Spinde, T., Plank, M., Krieger, J.-D., Ruas, T., Gipp, B., & Aizawa, A. (2021). Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the association for computational linguistics: EMNLP 2021*. Dominican Republic: <http://dx.doi.org/10.18653/v1/2021.findings-emnlp.101>, URL https://media-bias-research.org/wp-content/uploads/2022/01/Neural_Media_Bias_Detection_Using_Distant_Supervision_With_BABE_Bias_Annotations_By_Experts_MBG.pdf.
- Spinde, T., Rudnitskaia, L., Mitrović, J., Hamborg, F., Granitzer, M., Gipp, B., et al. (2021). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3), Article 102505.

- Spinde, T., Rudnitskaia, L., Sinha, K., Hamborg, F., Gipp, B., & Donnay, K. (2021). MBIC—a media bias annotation dataset including annotator characteristics. *arXiv preprint arXiv:2105.11910*.
- Stafford, T. (2014). Psychology: Why bad news dominates the headlines. *BBC Future. BBC*, 28.
- Strömbäck, J. (2005). In search of a standard: Four models of democracy and their normative implications for journalism. *Journalism Studies*, 6(3), 331–345.
- Sunstein, C. R. (2009). *Going to extremes: How like minds unite and divide*. Oxford University Press.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27.
- Sutter, D. (2000). Can the media be so liberal—the economics of media bias. *Cato Journal*, 20, 431.
- Tangri, K. (2021). *Using natural language to predict bias and factuality in media with a study on rationalization* (Ph.D. thesis), Massachusetts Institute of Technology.
- Taulé, M., Martí, M. A., Rangel, F. M., Rosso, P., Bosco, C., & Patti, V. (2017). 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 (pp. 157–177). CEUR-WS.
- Van Vleet, J. E. (2021). *Informal logical fallacies: A brief guide*. Hamilton Books.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 347–354).
- Yap, A. (2013). Ad hominem fallacies, bias, and testimony. *Argumentation*, 27(2), 97–109.
- Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), 1–40.