The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias

TIMO SPINDE, University of Göttingen, Germany
SMI HINTERREITER*, University of Würzburg, Germany
FABIAN HAAK*, TH Köln - University of Applied Sciences, Germany
TERRY RUAS, University of Göttingen, Germany
HELGE GIESE, Charité – Universitätsmedizin Berlin, Germany
NORMAN MEUSCHKE, University of Göttingen, Germany
BELA GIPP, University of Göttingen, Germany

The way the media presents events can significantly affect public perception, which in turn can alter people's beliefs and views. Media bias describes a one-sided or polarizing perspective on a topic. This article summarizes the research on computational methods to detect media bias by systematically reviewing 3140 research papers published between 2019 and 2022. To structure our review and support a mutual understanding of bias across research domains, we introduce the Media Bias Taxonomy, which provides a coherent overview of the current state of research on media bias from different perspectives. We show that media bias detection is a highly active research field, in which transformer-based classification approaches have led to significant improvements in recent years. These improvements include higher classification accuracy and the ability to detect more fine-granular types of bias. However, we have identified a lack of interdisciplinarity in existing projects, and a need for more awareness of the various types of media bias to support methodologically thorough performance evaluations of media bias detection systems. Concluding from our analysis, we see the integration of recent machine learning advancements with reliable and diverse bias assessment strategies from other research areas as the most promising area for future research contributions in the field.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → **Retrieval tasks and goals**.

Additional Key Words and Phrases: media bias, gender bias, racial bias, hate speech, text retrieval, news slant

ACM Reference Format:
Timo Spinde, Smi Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2023. The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias. *ACM Comput. Surv.* 1, 1 (December 2023), 41 pages. https://doi.org/10.1145/1122445.1122456

*Both authors contributed equally to this research.

Authors' addresses: Timo Spinde, University of Göttingen, Germany, Göttingen, Timo.Spinde@uni-konstanz.de; Smi Hinterreiter, University of Würzburg, Würzburg, Germany; Fabian Haak, TH Köln - University of Applied Sciences, Köln, Germany, fabian.haak@th-koeln.de; Terry Ruas, University of Göttingen, Germany, Göttingen, ruas@uni-goettingen.de; Helge Giese, Charité – Universitätsmedizin Berlin, Berlin, Germany, helge.giese@uni-konstanz.de; Norman Meuschke, University of Göttingen, Germany, Göttingen, Meuschke@uni-goettingen.de; Bela Gipp, University of Göttingen, Germany, Göttingen, gipp@uni-goettingen.de.

1 INTRODUCTION

Online news articles have become a crucial source of information, replacing traditional media like television, radio broadcasts, and print media (e.g., newspapers, magazines) [48]. However, news outlets often are biased [245]. The primary reason for this bias is that opinionated, entertaining, and sensationalist content is more likely to attract a larger audience while being less expensive to produce [11].

Media bias is widely recognized as having a strong impact on the public's perception of reported topics [48, 98, 145]. Media bias aggravates the problem known as filter bubbles or echo chambers [216], where readers consume only news corresponding to their beliefs, views, or personal liking [145]. The behavior likely leads to poor awareness of particular issues, a narrow and one-sided perspective [214], and can influence voting behavior [53, 60].

Highlighting media bias instances has positive implications and can mitigate the effects of such biases [21]. While completely eliminating bias may be an unrealistic goal, drawing attention to its existence by informing readers that content is biased allows them to compare content easily. It can also enable journalists and publishers to assess their work objectively [48]. In the following, we list systems designed to help readers mitigate the effects of media bias on their decision-making. Most of these systems focus on aggregating articles about the same event from various news sources to provide different perspectives [145]. For example, news aggregators like Allsides[1] and Ground News[2] allow readers to compare articles on the same topic from media outlets known to have different political views. Media bias charts, such as the AllSides media bias chart[3] or the Ad Fontes media bias chart[4] provide up-to-date information on media outlets' political slants. However, it is uncertain whether readers have the possibility and, more importantly, the desire to read several articles on the same topic and compare them.

Media bias has become the subject of increasing interdisciplinary research, particularly in automated methods to identify bias. However, the concept of media bias remains loosely defined in the literature [98]. Existing work uses different subcategories and types of bias [190, 221], but authors tend to focus on only one media bias subcategory while disregarding similar kinds of bias concepts. publications on media bias often work on similar concepts but assign different names to them, leading to confusion and imprecise use of terms. For example, some authors refer to word-based bias as linguistic bias [190], while others call it bias by word choice [222], but the exact difference or overlap between these terms is undefined. The lack of clarity surrounding media bias can have negative effects on measuring media bias perception [217]. Additionally, recent advances in Deep Learning have shown how awareness of tasks within complex domains, such as media bias, could potentially lead to large performance increases [9]. However, these advancements have yet to be incorporated into media bias research [218].

Our literature review seeks to create awareness of media bias detection as a task and to provide a summary of existing conceptual work on media bias and automated systems to detect it. To achieve this, we compare and contrast computer science research while also incorporating media bias-related concepts from non-technical disciplines such as framing effects [66], hate speech [51], and racial bias [234].

We propose a unified taxonomy for the media bias domain to mitigate ambiguity around its various concepts and names in prior work. In addition, we classify and summarize computer science contributions to media bias detection in six categories[5]: (1) traditional natural language processing (tNLP) methods [171], (2) simple non-neural ML techniques [207], (3) transformer-based (tbML) [210] and (4) non-transformer-based (ntbML) [69] machine learning. We also include

[1] https://www.allsides.com
[2] https://ground.news
[3] https://www.allsides.com/media-bias/media-bias-chart
[4] https://www.adfontesmedia.com/
[5] We reason and detail our categories in Section 5.

(5) non-neural network (nNN)-based (Section 5.2.3) [186] as well as graph-based [92] approaches. Lastly, we provide an overview of available datasets. Our aim is to provide an overview of the current state-of-the-art in media bias and increase awareness of promising methods. We show how computer science methods can benefit from incorporating user and perception-related variables in different datasets to improve accuracy. To facilitate the usage of such variables, we give an overview of recent findings about cognitive processes behind media bias. We believe that a systematic overview of the media bias domain is overdue given the numerous papers covering related issues. Such an overview can benefit future work in computer science and other areas, such as Psychology, Social Science, or Linguistics, which all cover media bias. As we show in detail in Section 3, existing literature reviews on media bias [98, 167, 185] do not cover crucial aspects. They do not give a systematic overview of related concepts, instead presenting how media bias can develop. Aside from the major developments within the media bias domain since 2021, they lack details on computer science methods and psychological and social science research.

In summary, our literature review answers the following research questions:

(RQ1) What are the relationships among the various forms of bias covered in the literature?
(RQ2) What are the major developments in the research on automated methods to identify media bias?
(RQ3) What are the most promising computer science methods to automatically identify media bias?
(RQ4) How does social science research approach media bias, and how can social science and computer science research benefit each other?

All resources for our review are publicly available at https://github.com/Media-Bias-Group/Media-Bias-Taxonomy.

2 METHODOLOGY

The core contribution of this article is a systematic literature review that provides a structured and comprehensive overview of the application of computer science methods for detecting media bias. This review also clarifies and establishes connections among the various concepts employed in the field of media bias. Reviews are susceptible to incomplete data and deficiencies in the selection, structure, and presentation of the content [68], especially when aiming for extensive coverage. To overcome these challenges, we designed our collection and selection processes carefully, with a focus on mitigating common risks associated with literature reviews.

We used automated, keyword-based literature retrieval (described in Section 2.1), followed by a manual selection (Section 2.2), and adhered to established best practices for systematic literature reviews [76, 113, 177].

The number of concepts (and keywords) relevant to media bias is high but hard to define.[6] Reviewing all papers for all related concepts is unfeasible[7]. Therefore, we applied filter criteria to select candidate documents. Moreover, we excluded references from the selected papers as additional candidates since determining an unbiased stopping criterion would be challenging. Our review covers the literature published between January 2019 and May 2022, thus providing a comprehensive overview of the state-of-the-art in the field.

To ensure diversity in the computer science publications included in our review, we retrieved literature from two sources: DBLP (DataBase systems and Logic Programming)[8] and Semantic Scholar[9]. Both sources are reliable and diverse and therefore meet the criteria for suitable sources for literature reviews [30, 124]. DBLP is the most extensive database for computer science publications to date, containing documents from major peer-reviewed computer science

[6] For example, the term bias also yields many health-related papers that are irrelevant to our review.
[7] Based on the keywords we searched for, which we detail in Section 2.1, we found over 100.000 publications.
[8] https://dblp.org/
[9] https://www.semanticscholar.org/

journals and proceedings. It is a primary literature platform used in other reviews [57, 168, 250]. Semantic Scholar draws on a considerably larger database than DBLP, going beyond computer science into other research areas. It is also frequently used in literature reviews [100, 238, 247] and allows for applying more filter criteria to searches, particularly filtering by scientific field.

Both platforms are accessible through an API and facilitate the use of an automated retrieval pipeline, which we require to filter our search results efficiently. We retrieved results for a selection of search terms (see Section 2.1). While Semantic Scholar is an extensive general knowledge archive, DBLP focuses on in-depth coverage of computer science. By including both major archives, we aim to retrieve an exhaustive set of candidate documents in computer science.

2.1 Retrieving Candidate Documents

We used media bias terms encountered during our initial manual retrieval step (depicted in Figure 1) as search queries to create candidate lists for our literature review.[10] These terms also served as the basis of the media bias categories we consolidated in our Media Bias Taxonomy in Section 4.2. In step 2 (Figure 1), we employed a Python pipeline to retrieve computer science documents from both DBLP and Semantic Scholar, merge and unify the search results, and export them as tabular data.[11] We scraped a list of 1496 publications from DBLP and 1274 publications from Semantic Scholar for the given time frame. We present the complete list and search keywords in our repository. As shown in Figure 1, we obtained a list of 3140 candidates for the literature review. After removing 531 duplicates between the Semantic Scholar and DBLP results, the final list contained 2609 publications. All search results were tagged with the relevant search queries and exported as a CSV file for the selection step.

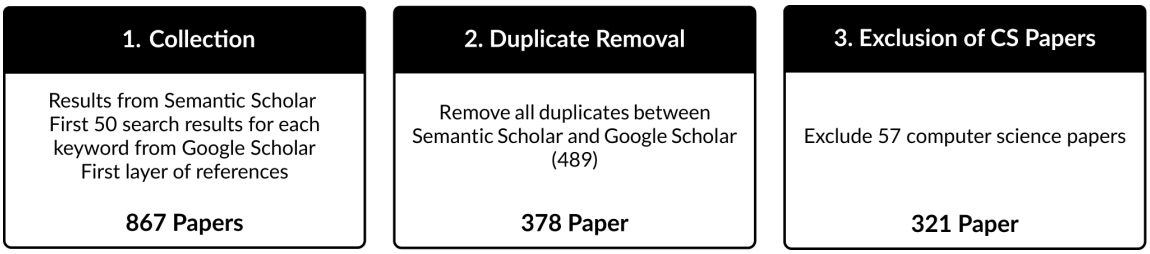

Fig. 1. Number of publications at each step of the literature retrieval and review of computer science publications.

2.2 Candidate Selection

We followed a multi-stage process to select relevant publications, as shown in Figure 1. The figure also shows the number of publications in each step. Three reviewers (Ph.D. students in computer science) filtered the results after the automatic scrape (step 2) and duplicate removal (step 3). In step 4, they filtered for documents that cover media bias, based on the title, abstract, and text, which resulted in 299 documents. In step 5, one reviewer per paper thoroughly inspected every publication to investigate whether computer science methods were used to detect media bias. For each publication, we exported the used methods and datasets (see Section 5). In step 6, a second reviewer verified the choice of the first reviewer for each publication. In case of disagreement or uncertainty, the third reviewer was consulted. For

[10] Initially, we used more general terms such as media bias", hate speech", linguistic bias", and racial bias" which are widely known. We manually identified additional bias concepts in the retrieved publications during our searches depicted in Figure 1 and Figure 2 and added them to our list of search queries. Subsequently, we searched for these newly identified keywords, creating the media bias keyword list presented in Figure 3.
[11] We have made the crawler publicly available for use in other projects. The code and instructions can be found in our repository.

each publication, at least two of the three reviewers must deem the publication suitable for our review. The detailed selection criteria for each step are available in our repository. In the end, we selected 96 relevant documents. We assigned each paper to its computer science methods category according to Figure 4.

2.3 Finding Additional Conceptual Literature for the Media Bias Taxonomy

One goal of our systematic literature review is to develop a taxonomy that organizes the various definitions of media bias into distinct types. However, while conducting our search, we recognized that most computer science publications focus on methodology rather than defining bias types. Therefore, we expanded our search to other research areas that may have different perspectives on media bias. For this purpose, we conducted a second search, as shown in Figure 2, replacing DBLP with Google Scholar to identify more non-computer science research[12]. We manually selected papers from the first 50 search results for each keyword on Google Scholar and Semantic Scholar[13] and checked the first layer of their references for additional relevant literature.

Overall, the additional search step for non-computer-science publications yielded 867 results, of which 489 were duplicates between Google Scholar and Semantic Scholar. Of the 378 non-duplicate publications, 57 were included in the search for computer science publications. We present the results of our searches in Section 4.2[14].

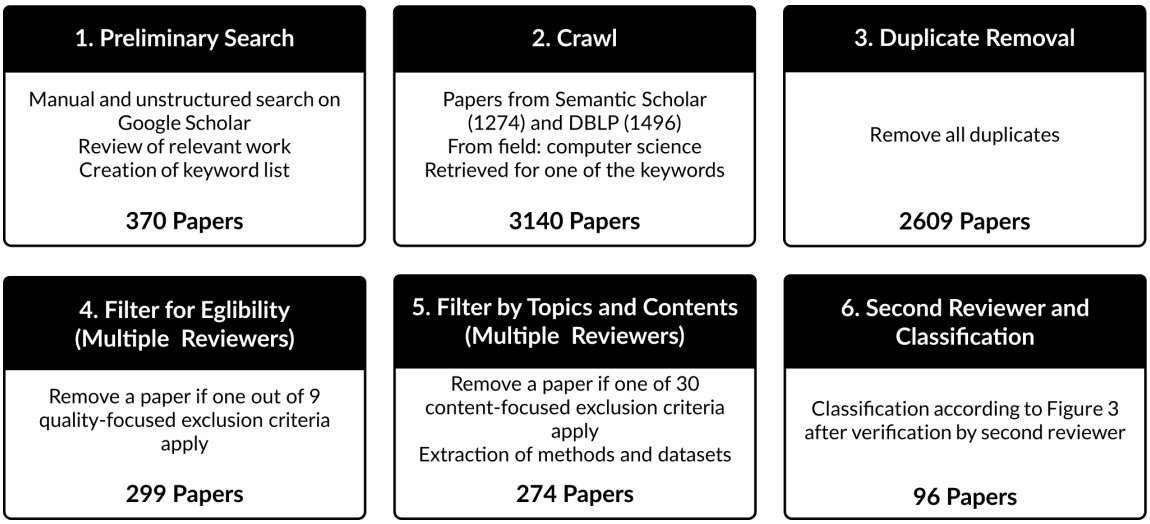

Fig. 2. Number of publications at each step of the literature retrieval and review for the Media Bias Taxonomy.

3 RELATED LITERATURE REVIEWS

Related literature reviews[15] on media bias are scarce. Our literature crawl and search (Section 2) yielded only three such results [98, 167, 185]. An additional search for the terms "media bias" and "news bias"[16] on Google Scholar did not yield more findings. In their literature review, Hamborg et al. [98] defined sub-categories of media bias from a social science

[12] Google Scholar is also a reliable and diverse database, meeting the criteria recommended in systematic literature review guidelines [30, 124].
[13] In this step, we excluded computer science publications in the Semantic Scholar results.
[14] Due to space restrictions we do not cite all of the filtered works in this article but omit publications focusing on highly similar concepts.
[15] We considered a publication a literature review if its main focus is a critical summary and evaluation of research about a topic related to media bias.
[16] We manually examined the first 50 results on Google Scholar.

perspective and showed how they emerge during journalistic work. Further, the authors described the advancements in computer science and indicated that frame analysis exists in both social sciences and computer science.

In the second work, Nakov et al. [167] surveyed media profiling approaches. They summarized computer science methods to analyze factuality (i.e., stance and reliability) and various forms of media bias (selection bias, presentation bias, framing bias, and news slant). The authors separated four prediction bases for media bias: 1) textual content and linguistic features, 2) multimedia content, 3) audience homophily, and 4) infrastructure characteristics.

Lastly, Puglisi and Snyder [185] surveyed the literature on media bias from a sociological perspective and offered an overview of possible bias measurements. They grouped biases into three kinds of measurement: comparing media outlets with other actors, the intensity of media coverage, and tone.

The earlier literature reviews exhibit three major shortcomings. First, both computer science-focused reviews [98, 167] lack a systematic literature search. They only covered selected computer science approaches and datasets. Second, Hamborg et al. [98] and Nakov et al. [167] did not cover the psychological perspective on bias, which we argue is essential to create and evaluate detection methods and datasets [217]. Third, no work thus far has provided a detailed overview of the various concepts and subcategories that fall under the umbrella term media bias. Current literature on media bias often addresses related concepts like hate speech, gender bias, and cognitive bias, but uses the umbrella term of media bias without clearly differentiating between overlapping categories and their relationships.

To our knowledge, we are the first to offer a large-scale, systematic analysis of the media bias domain. As a result, we provide our Media Bias Taxonomy, which connects the various definitions and concepts in the area. In addition, we briefly summarize the state-of-the-art psychological research on media bias and provide an in-depth overview of all computer science methods currently used to tackle media bias-related issues.

Our review focuses exclusively on media bias and does not include publications on related topics such as fake news. For details on fake news and its detection, we recommend referring to the two literature reviews [63, 227].

4 RELATED WORK AND THEORETICAL EMBEDDING

This section will provide an overview of media bias, followed by a presentation and organization of related concepts in our novel Media Bias Taxonomy.

4.1 Media Bias

Media bias is a complex concept [213, 217] that has been researched at least since the 1950s [242]. It describes slanted news coverage or other biased media content [98], which can be intentional, i.e., purposefully express a tendency towards a perspective, ideology, or result [243], or unintentional [21, 243]. Different stages of the news production process can introduce various forms of media bias [98].

The lack of a precise and unified definition for media bias, sometimes referred to as editorial slant [60], has contributed to the conceptual fragmentation in the field [217]. For instance, D'Alessio and Allen [47] categorized media bias into three primary groups [47]: gatekeeping bias, coverage bias, and statement bias. In contrast, Mullainathan and Shleifer [164] proposed two types of media bias: ideology bias and spin bias [164]. Some scholars referred to media bias as lexical or linguistic bias [23]. Others have proposed less specific definitions. For instance, Spinde et al. [222, p. 2] described media bias as "slanted news coverage or internal bias reflected in news articles." Lazaridou et al. [134, p. 1268] defined it as news reporting that "leans towards or against a certain person or opinion by making one-sided misleading or unfair judgments," and Lee et al. [138, p. 1] defined it as reporting "in a prejudiced manner or with a slanted viewpoint." None of these definitions is based on a comprehensive literature review. Therefore, we provide a comprehensive and

well-organized description of media bias in Section 4.2, which includes its sub-fields and related computer science methods and discuss the common ground of all media bias concepts in our review in Section 7.

It is worth mentioning that media bias does not only manifest via text but also via pictures or text/news layout [158, 178]. Moreover, biased reporting in one outlet can also cause biased reporting in other outlets by direct citations [91]. Our literature review focuses on text-based media bias and methods only.

4.2 The Media Bias Taxonomy

As media bias definitions often overlap, a clear distinction between its types is challenging. We propose the Media Bias Taxonomy, depicted in Figure 3 to give a comprehensive overview of the media bias domain. Based on a manual selection after the literature search process, described in Section 2.3, we split media bias into four major bias categories: linguistic, cognitive, text-level context, reporting-level, as well as related concepts, which are detailed in the following subsections. We show detailed examples in Appendix A.2 for all subtypes of bias[17].

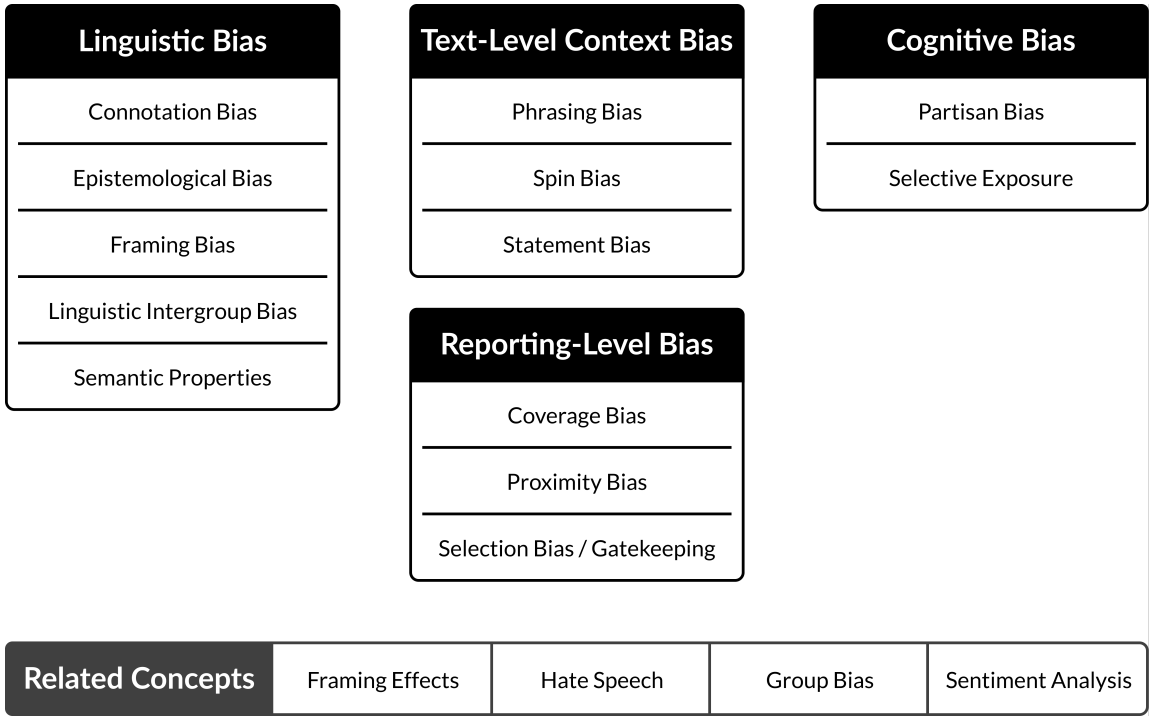

Fig. 3. The Media Bias Taxonomy. The four subcategories of media bias consist of different bias types.

4.2.1 Linguistic Bias. Linguistic bias, sometimes called lexical bias [70], refers to a pattern of using certain words that reflects a particular way of thinking about a group or an individual based on their social category. This bias involves a systematic preference for certain words or phrases that may reflect stereotypes or preconceived notions about the

[17] Other, overarching concepts exist, such as persuasiveness [89], which we do not cover or organize within this work. In future work, we will address concepts containing mmultiple forms of bias

group or individual being described [23]. In simpler terms, linguistic bias means using language that reflects a particular attitude or viewpoint towards a particular group or individual.

We identified five bias types within this category: linguistic intergroup bias [206], framing bias [190], epistemological bias [190], bias by semantic properties [90], and connotation bias [187]. Table 3 lists examples for each subcategory.

Linguistic Intergroup Bias describes which group members use specific language [206]. The concept is based on the linguistic category model (LCM), which categorizes words into different levels of abstraction (action words, interpretive action words, state verbs, and adjectives) according to their purpose [58, 206]. The use of biased language is often subtle and reinforces stereotypes [23, 151]. Maass et al. [151] illustrated linguistic intergroup bias with the following example:

- They considered the hypothetical scenario where "Person A is hitting Person B's arm with his fist" [151, p. 982].
- Describing the scenario using the least abstract form of language, one could say, "A is punching B" [151, p. 982]. This entails no kind of valuation or implication and only describes what happened.
- In contrast, using the most abstract form of language, one could say "A is aggressive" [151, p. 982]. This might or might not be accurate and cannot be judged from the fact that A hit B.

Framing Bias is defined as the use of "subjective words or phrases linked with a particular point of view" [190, p. 1650] to sway the meaning of a statement. The subjective words are often either one-sided terms or subjective intensifiers [190]. One-sided terms are words that "reflect only one of the sides of a contentious issue" [190, p. 1653], while subjective intensifiers are adjectives or adverbs that reinforce the meaning of a sentence.

Epistemological Bias describes the use of linguistic features that subtly focus on the credibility of a statement [190]. Word classes associated with epistemological bias are factive verbs, entailments, assertive verbs, and hedges, see examples in Table 3. Factive verbs indicate truthfulness; entailments are relations where one word implies the truth of another word. Assertive verbs state clearly and definitely that something is true. Hedges are words used to introduce vagueness to a statement. In contrast to framing bias, epistemological bias is rather subtle and implicit [190].

Bias by Semantic Properties describes how word choice affects the framing of content and triggers bias, similar to framing bias and epistemological bias. The difference, however, is that framing and epistemological bias refer to the individual words used, whereas bias by semantic properties refers to how the sentence is structured [90].

Connotation Bias refers to using connotations to introduce bias to a statement [187]. While the denotation of a word expresses its literal meaning, the connotation refers to a secondary meaning besides the denotation. The connotation is usually linked to certain feelings or emotions associated with a point of view [187].

4.2.2 Text-level Context Bias. Similar to linguistic bias, text-level context bias refers to the way the context of a text is expressed. Words and statements have the power to alter the article's context, influencing the reader's opinion [110]. The types of bias belonging to this category are statement bias [47], phrasing bias [110], and spin bias [155], which consists of omission bias and informational bias [155]. Table 4 lists examples for each subcategory.

Statement Bias refers to "members of the media interjecting their own opinions into the text" [47, p. 136], which leads to certain news being reported in a way that is more or less favorable towards a particular position [47]. These opinions can be very faint and are expressed "by disproportionately criticizing one side" [31, p. 250] rather than "directly advocating for a preferred [side]" [31, p. 250].

Phrasing Bias is characterized by inflammatory words, i.e., non-neutral language [110]. Depending on the context, a word can change from neutral to inflammatory. Therefore, when analyzing bias, the inter-dependencies between words and phrases must be considered [110].

Spin Bias describes a form of bias introduced either by leaving out necessary information [155, 164] or by adding unnecessary information [70]. The underlying motivation is to tell a simple and memorable story [164]. Spin bias can be divided into omission, and informational bias [155]. Omission bias, also known as simplification, is the act of omitting words from a sentence [155, 164]. Informational bias, or exaggeration, is defined as adding speculative, tangential, or irrelevant information to a news story [70].

4.2.3 Reporting-level Context Bias. Reporting-level context bias subsumes all bias types on the reporting level. While text-level context bias observes bias within an article, reporting-level bias observes the general attention for specific topics [31, 47, 77, 199]. Bias types in this category are selection bias, proximity bias, and coverage bias, which are all closely connected. Table 5 lists examples for each subcategory.

Selection Bias (or gatekeeping bias) refers to the selection of content from the body of potential stories by writers and editors [47]. Obviously, not all news events can be reported due to the limited resources of newspapers. However, this decision-making process is prone to bias from personal preferences [47, 164, 199, 242].

Coverage Bias describes situations in which two or more sides of an issue receive imbalanced amounts of attention, such as pro-life vs. pro-choice statements [47].[18] The level of attention can be measured either in absolute numbers (e.g., there are more articles discussing pro-life than pro-choice topics), how much space the topics get in a newspaper (e.g., printed on the front page), or as the length of the article (e.g., pro-life articles are longer and receive more in-depth coverage than pro-choice articles) [47, 199].

Proximity Bias focuses on cultural similarity and geographic proximity as decisive factors. Newspapers tend to report more frequently and more in-depth on events that happened nearby [199]. For instance, the more two countries are culturally similar, the more likely it is that events from one region or country will be reported in the other, and the coverage will be more in-depth [77, 199].

4.2.4 Cognitive Bias. The processing of media information may also be biased by the reader of an article and the state the reader is in during reading. In this review, we use the term cognitive bias, defined as "a systematic deviation from rationality in judgment or decision-making" [27, p. 1], to summarize how this processing may be negatively affected. While a failure to detect biased media in a given set of articles may be explained by a lack of ability or motivation (e.g., being inattentive/ disinterested, focusing on identity instead of accuracy motives), biased processing of news by the reader is often attributed to a need for a consistent world view and for overcoming dissonances evoked by discordant information [169]. In this line of reasoning, repeated exposure and increased familiarity with an argument as well as source cues for a reputable, world-view-consistent source, may increase the trust in information quality.

Selective Exposure. Similar to the selection bias of editors and authors, readers also actively select which articles they read [125]. Given this choice, they tend to favor reading information consistent with their views, exacerbating already existent biases through selective exposure to one-sided news reports [55, 170]. Additionally, such selective exposure tends to extend to social tie formation. Topic information is solely exchanged among like-minded individuals, a phenomenon often dubbed echo chamber or filter bubble [141][19], hampering unbiased information processing.

Partisan Bias. Selective attention to world-view-consistent news has led to research on the effects of political identity. There, the evaluation of veracity seems dependent on the fit to the reader's party affiliation, a phenomenon dubbed partisan bias [18, 82]. Similarly, the hostile media phenomenon (HMP) describes the general observation that members of opposing groups rate a news article as biased against their point of view [231].

[18] Coverage bias refers to a particular event, whereas reporting-level context bias refers to the general attention a topic receives.
[19] In case an algorithm was trained to this preference.

4.2.5 Related Concepts. The last category contains definitions that cannot be exclusively assigned to any other media bias category. Concepts belonging to this category are framing effects [53], hate speech [51], sentiment analysis, and group bias [39], which consists of gender bias [43], and religion bias [153]. Much research focuses on these concepts, so we introduce them only briefly and refer to other sources for more information.

Framing Effects refer to how media discourse is structured into interpretive packages that give meaning to an issue, so-called frames. Frames promote a specific interpretation of the content or highlight certain aspects while overlooking others. In other words, this type subsumes biases resulting from how events and entities are framed in a text [53, 65].

Hate Speech is defined as any language expressing hatred towards a targeted group or intended to be derogatory, humiliate, or insult [51]. Often, hateful language is biased [163]. The consequences of hate speech in media content are severe, as it reinforces tension between all actors involved [3, 163].

Group Bias. We categorize gender bias, racial bias, and religion bias under the umbrella term "group bias," as they all refer to biased views toward certain groups.

Gender Bias is characterized by the dominance of one gender over others in any medium [43], resulting in the under-representation of the less dominant gender and the formation of stereotypes [43, 183]. It is associated with selection bias [10, 116], coverage bias [32, 135], and context bias at the text level. For instance, women are quoted more frequently than men for "Lifestyle" or "Healthcare" topics, while men are quoted more frequently in "Business" or "Politics" [186]. Linguistic research on gender bias aims to identify gender-specific and gender-neutral words [46] and create lexicons of verbs and adjectives based on gender stereotypes [71].

Racial Bias and **Religion Bias** are other types of group bias. Racial bias refers to the systematic disproportionate representation of ethical groups, often minorities [39], in a specific context [39, 161].

Religion, racial, and gender biases can be observed in word embeddings. For example, "Muslim" is spatially close to "terrorist" in some embeddings [153], which may result from biased texts in the data used to derive these embeddings (as word embeddings depend on their input).

Group biases can manifest in other forms, such as hate speech, which is a subgroup of biases. Although the distinction between racial and gender biases is not always evident, they can exist independently [85, 161].

Sentiment Analysis involves examining text for its emotional content or polarity [64]. In the context of media bias, sentiment analysis can detect biases in statements or articles [97, 109] and help identify other concepts like hate speech, political ideology, or linguistic bias [3, 193].

5 COMPUTER SCIENCE RESEARCH ON MEDIA BIAS

Computer science research on media bias primarily focuses on methods used to analyze, mitigate, and eliminate bias in texts. Detecting bias is a prerequisite for other applications [189]. Bias detection systems could also be employed to check computer-generated texts for bias. Hereafter, we provide a comprehensive overview of computer science methods used in media bias research in recent years based on a systematic literature review. The methodology of the review is described in Section 2. A systematic overview of computer science methods is essential for capturing the state of media bias research and identifying research trends and gaps. To the best of our knowledge, this is the most comprehensive survey on media bias detection methods so far, as discussed in Section 3.

Table 1 organizes the findings of our literature review by the year of publication and category of employed computer science method.[20] We chose the employed methods as the main categorical property to structure the publications since

[20] We do not report performance measures for most models, as most approaches work on different datasets and tasks, causing the scores to be incomparable. Instead, we summarize our findings on the most promising approaches at the end of this section.

the methods are typically described in more detail than the type of investigated bias. Our analysis shows that media bias detection methods use approaches ranging from traditional natural language processing (tNLP) methods (e.g., [171]) and simple ML techniques (e.g., [207]) to complex computer science frameworks that combine different advanced classification approaches (e.g., [97]), and graph-learning-based approaches (e.g., [105]). Therefore, we introduce the classification depicted in Figure 4.

Approaches we classify as tNLP (Section 5.1) do not use complex ML techniques and are commonly employed in social sciences (e.g., [130, 160]). We categorize the tNLP publications into two groups: first, count-based techniques supported by lexical resources, and second, more sophisticated embedding-based techniques.

ML-based approaches (Section 5.2) are organized into transformer-based machine learning (tbML), non-transformer-based (ntbML), and non-neural network (nNN)-based (Section 5.2.3) approaches, ordered by the frequency of application in the reviewed literature. Graph-based models represent the third major category presented in Section 5.3.

Appendix A.1 shows the number of publications per year and category according to our search criteria (cf. Section 2). An increasing majority of publications use tbML approaches, while the numbers of nNN- and ntbML-based approaches decrease. Although our review does not fully cover 2022, the numbers suggest that these trends continue.

5.1 Traditional Natural Language Processing Techniques

The tNLP category encompasses all publications that identify media bias using techniques not based on ML or graph-based approaches. We include the term "traditional" in the category name to differentiate it from ML and similar techniques. Moreover, techniques similar to what we label as tNLP have already been employed in computational linguistics as early as the sixties and seventies [1]. Frequently, tNLP methods are used as a baseline when introducing new datasets due to their explainability and proven effectiveness (e.g. [46, 136, 202]). Furthermore, social sciences are increasingly adopting them because of their accessibility and ease of use [130]. Although some approaches leverage ML techniques (e.g., [49]), we classify them as tNLP if the main contribution is a non-ML approach. The tNLP methods can be divided into count-based and embedding-based approaches. Count-based approaches quantify words and n-grams in the text to analyze bias, while embedding-based approaches are more sophisticated and serve to represent texts for either facilitating comparisons (e.g., [136, 175]) or analyzing text associations and inherent biases (e.g., [33, 72]).

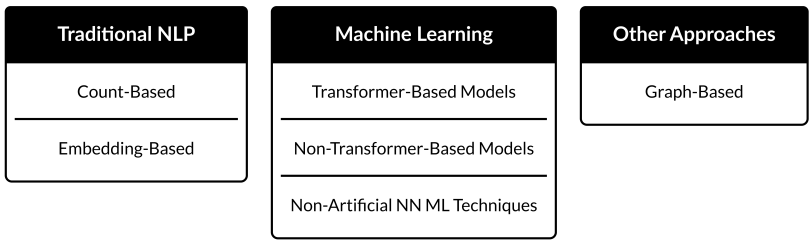

Fig. 4. Classification of computer science methods for media bias detection we use in our analysis.

	2019	2020	2021	2022	total
tNLP	[12, 41, 241],[203]*	[2, 52, 72, 160, 171, 175, 215, 248], [136]*	[46, 49, 130, 133, 202, 214, 220],[45]*		21
tbML	[56, 70, 80, 99, 184]	[14, 24, 132, 144, 163, 174, 208, 228, 232],[16, 162]*	[26, 97, 103, 111, 117, 147, 152, 209, 212, 219],[210, 251]*	[128, 139, 150, 189, 218]	33
ntbML	[7, 28, 78, 110, 114]	[38, 44, 74, 108, 115, 156, 239, 246]	[19, 84, 129, 196, 197]	[69, 149]	20
nNN	[8, 15]	[6, 29, 79, 118, 134, 207] ,[37]*	[186, 221, 222], [173]*		13
graph-based	[143]	[36, 40, 223]	[92, 105, 237, 252]	[95]	9
total	17	38	33	8	96

Table 1. Results of the literature study on computer science methods used for media bias detection.

* We refer to this paper in multiple sections. If a publication covers multiple categories, we assign the most used category. If two categories apply equally, we assign one based on the method performing best.

5.1.1 Count-Based Approaches. While recent applications of tNLP techniques primarily employ embedding-based methods, simpler count-based approaches are still in use. Count-based approaches most commonly use word counts and a lexicon as a reference to quantify linguistic characteristics and compare texts.

Niven and Kao [171] measured the alignment of texts to authoritarian state media using a count-based methodology that leveraged the LIWC lexicon [179] for topical categorization. Similarly, Spinde et al. [215] applied various count-based techniques to a custom dataset of German news articles and assessed their effectiveness for media bias detection. They reported precision, recall, and F_1 scores for bias and sentiment lexicons, word embeddings, and general TF-IDF measures, evaluating the identification of human-annotated bias in their dataset. A custom bias lexicon yielded the best performance with a low F_1 score of 0.31.

Sapiro-Gheiler [203] employed Naive Bayes (NB) decision tree, support vector machine (SVM), and lasso-penalty regression models based on bag-of-word representations to classify politicians' ideological positions and trustworthiness. De Arruda et al. [52] used a count-based approach within an outlier detection framework to identify selection, statement, and coverage bias in political news. D'Alonzo and Tegmark [49] presented a singular value decomposition (SVD) approach that predicts the newspaper that published an article based on word and n-gram frequencies. Discriminative words and n-grams were derived from a multi-stage (automatic and manual) purging process. The system generates a conditional probability distribution that enables the projection of newspapers and phrases into a left-right bias space.

Zahid et al. [248] used a contingency table showing mention counts and polarity rates for sources (S) and entities (E) within news-related content on Twitter to calculate media bias measures based on definitions for absolute and relative media bias [200]. They investigated coverage, selection, and statement bias towards specific topics and entities, and further quantified and compared the number of positive and negative reports from media outlets on Twitter.

Cuéllar-Hidalgo et al. [45] presented their contribution to the ICON2021 Shared Task on Multilingual Gender Biased and Communal Language Identification [131], where the goal is to classify texts as aggressive, gender biased, or

communally charged. They used k-nearest neighbors (KNN) and a mixed approach consisting of NB, SVM, random forest (RF), GBM, Adaboost, and a multi-layer perceptron, for classifying texts.[21]

Dacon and Liu [46] presented a study on gender bias in news abstracts using centering resonance analysis based on specifically filtered attribute words. This technique employs rich linguistic features and graph-based techniques.

5.1.2 Word Embedding-Based Techniques. A second group of tNLP techniques detects media bias by deriving word associations through word embeddings. We exclude publications that investigate bias in pre-trained word embeddings, e.g., to understand potential biases in systems that use the embeddings, as this analysis does not represent a media bias investigation. However, we include work that uses word embeddings as proxies to help understand biases in texts used for training the embeddings. This is typically done by constructing word embeddings based on a collection of texts and investigating associations in these embeddings (e.g., [72, 136, 175]). We differentiate between sparse and dense embedding-based techniques. Sparse embeddings, primarily based on TF-IDFs, are mostly used to survey the occurrence of certain words [136]. Dense embeddings are employed to examine associations with specific terms [72, 241].

Sparse Word Embeddings. Leavy [136] investigated gender bias in Irish newspapers, examining various discriminative features such as TF-IDFs. Alongside ML techniques, she used count-based tNLP approaches to detect coverage bias towards female politicians. Employing a bag-of-words approach, TF-IDFs, and linguistic labels on word forms, she provided data for classification models and directly detected bias. For instance, she found articles mentioning spouses of female politicians four times more often than male politicians.

Dense Word Embeddings. Most word embedding-based techniques in this section use methods similar to the word embedding association test (WEAT) introduced by Caliskan et al. [33]. WEAT investigates bias in the resulting word embeddings trained on a specific text corpus by measuring the cosine similarity between two sets of tokens (e.g., male and female pronouns) and another two sets of tokens, typically topic or stereotype-based words.

Ferrer et al. [72] explored various aspects of linguistic and gender bias on Reddit using a technique akin to WEAT, while also examining biases through count-based approaches and sentiment analysis. Badjatiya et al. [12] proposed a debiasing strategy using bias-sensitive words as reference, primarily focusing on replacing bias-sensitive words with less sensitive synonyms to debias text datasets. They identified replacement words using word embeddings with different algorithms such as KNN or a centroid function.

Mendelsohn et al. [160] primarily employed embedding-based tNLP techniques to investigate the development of dehumanization towards the LGBTQ community in New York Times articles from 1986 to 2015. Wevers [241] conducted a study on gender bias in Dutch newspapers between 1950 and 1990, measuring the distance of "three sets of target words" [241, p. 3] to two gender-representative vectors. These vectors were constructed from the average of lists of "gender words"[241], such as "man," "his," "father," and similar terms for the male vector.

Similarly, Papakyriakopoulos et al. [175] used word embedding associations to compare gender bias in Wikipedia and social media texts. Kroon et al. [130] analyzed implicit associations with word embeddings to detect racial bias, using the term "ethnically stereotyped bias" in their work. Spinde et al. [220] trained two word embedding models on slanted news corpora: one using left-wing news from HuffPost and another based on right-wing Breitbart news. They employed the Word2Vec Continuous Skip-gram architecture for training and subsequently applied a distance-based technique with their word embeddings to identify strongly biased words, beginning with biased seed words.

Kwak et al. [133] presented a distinct approach to bias detection based on word embeddings. They introduced a method for characterizing documents by identifying the most relevant semantic framing axes ("microframes") that

[21] This work employed both tNLP and NN based methods. However, since the majority of the techniques fall into the tNLP category, we discuss it here.

are overrepresented in the text. They then assessed the extent of bias and activity of a given microframe, ultimately providing a more detailed description of the documents. For instance, they might identify that the axis of "depressing" and "cheerful" is central to an article and then analyze the wording that led to this classification [133].

Sales et al. [202] employed a mix of tNLP techniques based on word embeddings to detect subjectivity bias, utilizing methods such as lexicon translation and document similarity measures.

5.2 Machine Learning

The following section includes publications that used ML for bias detection. We start by presenting transformer-based models (tbML), which were most frequently applied in the reviewed literature, followed by non-transformer-based models (ntbML), and non-neural network models (nNN). TbML increased in popularity after their introduction in 2017 [235], as shown in Table 1 and Appendix A.1. Transformers use self-attention to weigh the importance of input data and can be fine-tuned with specific datasets, saving time and resources [235]. Their universal architecture captures dependencies across domains but can over-fit in case of limited training data [146].

5.2.1 Transformer-Based Models. Researchers frequently used tbML to detect linguistic bias or political stance with an encoder-only architecture and bias-specific pre-training. Most often they used BERT or models derived from it, e.g, RoBERTa [111, 128, 152, 189, 221, 232, 251], DistilBERT [26, 189, 218], or ALBERT [117]. Several papers compare the performance of BERT-based models with other transformer models, e.g. T5 [128], BART [103, 139], ELECTRA [219] or XLNet [147]. BERT-based models were also applied to detect media bias in languages other than English, such as Korean ((Kor)BERT) [99], Indian (IndicBERT) [117] or fine-tuning BERT on African American [163]. When researchers used an encoder-decoder architecture model like BART, they used the encoder only for the detection task, while the decoder performed the debiasing task [103, 139]. BERT-based models often outperformed other transformers for most of the tasks and groups we defined for linguistic bias [111, 189, 209, 219], and for political stance detection [56, 210], which typically associates linguistic bias with specific political stances [210].

The prevalent approach in tbML is to create or select bias-specific datasets, fine-tune the most popular models on them, and test the performance of the encoder-only architecture by comparing F_1-scores to baselines of tNLP methods (e.g., [80, 97, 128, 174, 219]). To facilitate the evaluation of using different transformers for identifying various media bias types, we structure our review of tbML by the type of bias used in fine-tuning.

Linguistic Bias. Most tbML applications focus on detecting linguistic bias. Spinde et al. [219] detected bias by word choice following a distant supervision approach with BERT. Based on the BABE dataset, BERT outperformed RoBERTa and other ML classifiers in their application. In contrast, Huguet Cabot et al. [111] achieved the best performance on their Us vs. Them dataset with RoBERTa. Sinha and Dasgupta [209] also fine-tuned BERT with a custom dataset and contextual embeddings. In addition, they parsed sentences using a GCN model with an additional layer of bidirectional long short-term memory (LSTM) to exploit structural information. Raza et al. [189] proposed a four-phase pipeline consisting of detection (DistilBERT), recognition (RoBERTa), bias masking, and debiasing. The system, fine-tuned on the MBIC dataset [221], detected biased words, masked them, and suggested a set of sentences with new words that are bias-free or less biased. Pryzant et al. [184] detected and automatically transformed inappropriate subjective texts into a more neutral version. Using a corpus of sentence pairs from Wikipedia edits, their system used BERT as an encoder to identify subjective words as part of the generation process.

Political Stance Detection. The second most researched classification problem is political stance detection, an umbrella term closely related to partisan bias (cf. Section 4.2) that identifies linguistic biases to identify the political biases

of authors. Sinno et al. [210] studied the ideology of specific policies under discussion and presented the first diachronic dataset of news articles annotated at the paragraph level by trained political scientists and linguists. Their fine-tuned BERT model performed best. Dinkov et al. [56] integrated audio, video, metadata, and subtitles in their multimodal dataset. In addition to the text analysis with BERT, their application included metadata and audio data through open SMILE[22], resulting in the highest accuracy. Sinno et al. [210] presented a manually annotated dataset focusing on linguistic bias in news articles. Based on their dataset, in addition to several BERT-based classification approaches, they used a 2-layer bidirectional LSTM for ideology prediction, which was outperformed by all transformer-based systems.

Framing Bias. Mokhberian et al. [162] used BERT with tweet embeddings, fine-tuned on the All The News dataset[23], and an intensity score for moral frames classification based on the moral foundation theory[24]. Kwak et al. [132] proposed a similar BERT-based method for conducting sociological frame analysis to detect framing bias. Lee et al. [139] proposed a system for framing bias detection and neutral summary generation from multiple news headlines of varying political leanings to facilitate balanced and unbiased news reading. They performed multi-document summarization, multi-task learning with two tasks, and based their work on BART.

Spin/Informational Bias. Fan et al. [70] investigated lexical and informational bias with BERT on their BASIL dataset, which others also used in their research [150, 209, 232]. van den Berg and Markert [232] fine-tuned RoBERTa as a context-inclusive model, exploring neighboring sentences, the full article, articles on the same event from other news publishers, and articles from the same domain. Their model is domain-and-task-adapted for informational bias detection on the BASIL corpus. They reported that integrating event context improved classification performance.

Racial/Group Bias. For group bias detection, He et al. [103] presented DEPEN, which employs a fine-tuned BERT model to detect biased writing styles. Subsequently, they used BART to debias and rewrite these detected sentences.

Sentiment Analysis. We exclude general sentiment analysis but include publications that leveraged sentiment analysis for linguistic bias detection as a stand-in for political stance detection (cf. Section 5.2.1). Huguet Cabot et al. [111] investigated populist mindsets, social groups, and related typical emotions using RoBERTa fine-tuned on their populist attitude dataset *Us vs. Them*. Gao et al. [80] utilized BERT in aspect-level sentiment classification, achieving promising performances on three public sentiment datasets[25]. They showed that incorporating target information is crucial for BERT's performance improvement. Hamborg and Donnay [97] applied target-dependent sentiment classification (TSC) with BERT, RoBERTa, XLNET, and a BiGRU. They proposed a classifier, GRU-TSC, that incorporated contextual embeddings of the sentences and representations of external knowledge sources.

Unreliable News Detection. Zhou et al. [251] used RoBERTa to detect unreliable news—a task that overlaps with media bias detection. Further, they proposed ways to minimize selection bias when creating datasets by including a simple model as a difficulty/bias probe. They also suggested that future model development uses a clean non-overlapping site and date split [251].

5.2.2 Non-Transformer-Based Models. This section presents publications that use non-transformer-based machine learning for media bias detection, categorized by the type of detected bias. Most commonly, ntbML methods are used to detect media bias at the document level, e.g., hyperpartisanship and political stance. Despite the homogeneity of detected biases, publications using ntbML evaluate numerous aspects of the identification methodology, including training data [156, 210], word embeddings [44, 114, 239], and pseudo-labeling [197].

[22] https://www.audeering.com/de/research/opensmile/
[23] https://www.kaggle.com/datasets/snapcrack/all-the-news
[24] Moral foundation theory explains moral differences across cultures. For more information, see the original work by Haidt and Joseph [96].
[25] The datasets include restaurant and laptop reviews, and tweets [80].

Linguistic/Text-Level Bias. The detection of hyperpartisanship[26] is the most common application of ntbML. The task's popularity is partly due to the SemEval 2019 hyperpartisan news detection task [122] and the associated dataset, which inspired many publications. Hyperpartisanship is defined as non-neutral news reporting [122], which can be described as a combination of linguistic and text-level biases on a document level. The approach of Jiang et al. [114] performed best in the task. It leveraged a convolutional neural network (CNN) along with batch normalization and ELMo embeddings. In a follow-up study, Jiang et al. [115] incorporated Latent Dirichlet Allocation (LDA) distributions with different approaches to hyperpartisan news detection. They implemented multiple methods, such as a CNN, a recurrent neural network (RNN), a transformer encoder approach, and a hierarchical attention network (HAN) with and without LDA topic modeling. Their results suggested that, in most cases, LDA topic modeling improves the effectiveness of the methods, and hierarchical models outperform non-hierarchical models. Webson et al. [239] presented another study based on the SemEval 2019 hyperpartisan news detection task. They focused on decomposing pre-trained embeddings into separate denotation and connotation spaces to identify biased words descriptively. Although their primary goal was to improve the embeddings' reflection of the implied meaning of words, they showed how the discrepancy between the denotation space and the pre-trained embeddings reflects partisanship [239]. Cruz et al. [44] used different ML approaches (e.g., RNN, CNN, bidirectional LSTM/GRU, and the attention-based approaches AttnBL, HAN) trained on the SemEval 2019 dataset. They evaluated the effects of attention mechanisms and embeddings based on different granularities, tokens, and sentences on the effectiveness of the models. Ruan et al. [197] focused on introducing methods for generating additional data. They presented two approaches for pseudo-labeling (overlap-checking and meta-learning) and introduced a system detecting media bias using sentence representations from averaged word embeddings generated from a pre-trained ELMo model and batch normalization. The same authors also employed an ELMo-based classifier and a data augmentation method using pseudo-labeling [196].

Political Stance Detection. Baly et al. [16] trained two models based on LSTM and BERT for classifying news texts as left-wing, center, or right-wing. Their main contribution is the evaluation of techniques for eliminating the effects of outlet-specific language characteristics (here: political ideology expressed by linguistic bias) from the training process. They used adversarial adaptation and triplet loss pre-training for removing linguistic characteristics from the training data. Further, they incorporated news outlets' Wikipedia articles and the bio of their Twitter followers in the training processes to reduce the effects of outlet-specific language characteristics. While a transformer-based classification outperformed the LSTM model, the techniques for improving training effectiveness improved both models' classification results.[27] As part of their political stance detection approach, Gangula et al. [78] proposed a headline attention network approach to bias detection in Telugu news articles. It leveraged a bidirectional LSTM attention mechanism to identify key parts of the articles based on their headlines, which were then used to detect bias toward political stances. They compared the results of their approach with NB, SVM, and CNN approaches, all of which the headline attention network outperformed. To depolarize political news articles, Fagni and Cresci [69] mapped Italian social media users into a 2D space. Their solution initially leveraged a NN for learning latent user representations. Then, they forwarded these representations to a UMAP [159] model to project and position users in a latent political ideology space, allowing them to leverage properties of the ideology space to infer the political leaning of every user, via clustering.

Gender/Group Bias. Field and Tsvetkov [74] presented an unsupervised approach for identifying gender bias in Facebook comments. They used a bidirectional LSTM to predict the gender of the addressee of Facebook comments and,

[26] Hyperpartisanship is not to be confused with partisan bias as described in Section 4.2. It describes one-sidedness that can manifest in a range of biases [122].
[27] Since transformers are not the paper's focus, we discuss it here.

in doing so, identify gender biases in these comments. Mathew et al. [156] introduced HateXplain, a dataset on hate speech and gender bias that includes expert labels on the target community towards which the hate speech is aimed. They further included labels of words annotators identified as bias-inducing. They evaluated the effects of including the rationale labels in the training process of a BiRNN and a BERT model on the models' bias detection capabilities. Including the rationale labels increased the bias classification performance for both models.

5.2.3 Non-Neural Network Machine Learning Techniques. Besides state-of-the-art approaches using tbML or deep learning techniques for bias detection, other (nNN) ML approaches are still widely used for bias detection. Many employ LDA, SVM, or regression models, but a wide range of models is usually used and compared. These models are particularly common in papers presenting new datasets, as they can be seen as a solid and widely known baseline for the quality of labels within a dataset.

Based on the MBIC dataset, Spinde et al. [221, 222] presented a traditional feature-based bias classifier. They evaluated various models (e.g., LDA, logistic regression (LR), XGBoost, and others), trained with features such as a bias lexicon, sentiment values, and linguistic word characteristics (such as boosters or attitude markers [222]). Alzhrani [6] contributed a dataset of personalized news. Furthermore, she used a range of classifiers (Ridge classifier, nearest centroid, SVM with SDG, NB) for political affiliation detection. Rao and Taboada [186] investigated coverage and gender bias in their dataset of Canadian news articles. They employed LDA topic modeling to detect biased topic distributions for articles that contain predominantly male or female sources. Kameswari et al. [118] presented a dataset of 200 unbiased and 850 biased articles written in Telugu. They used NB (Bernoulli and multinomial), LR, SVM, RF, and MLP classifiers to evaluate the effectiveness of adding presuppositions as model input. Shahid et al. [207] researched framing effects in news articles using their proposed dataset. They trained an SVM classifier to detect and classify moral framing and compared it to a baseline lexicon-based natural language processing approach, investigating moral framing aspects such as authority, betrayal, care, cheating, etc. Ganguly et al. [79] explored various biases that can occur while constructing a media bias dataset. Part of their work examined the correlation between the political stance of news articles and the political stances of their media outlets. To evaluate this correlation, they compared multinominal NB, SVM, LR, and RF models using ground-truth labels. Several other publications described the application of nNN ML approaches in addition to other ML techniques for data evaluation [134, 136, 162, 173, 251]. We have already mentioned these in Section 5.1 and Section 5.2.2.

Baly et al. [15] presented a multi-task ordinal regression framework for simultaneously classifying political stance and trustworthiness at different Likert scales. This approach is based on the assumption that the two phenomena are intertwined. They employed a copula ordinal regression along with a range of features derived from their previous work, including complexity and morality labels, linguistic features, and sentiment scores. Anthonio and Kloppenburg [8] presented an additional[28] model for the SemEval 2019 hyperpartisan news detection task [122]. They used a linear SVM with VADER sentiment scores as a feature, relying exclusively on the intensity of negative sentiment in texts to derive political stances expressed in texts. With a F_1 score of 0.694, their approach failed to match the other competitors in the task. In addition to a FastText classifier, the approach presented by Lazaridou et al. [134] included a manual selection of training data containing examples of media bias. Aside from contributing to a new media bias dataset and evaluating the effect of expert and non-expert annotators, they presented a curriculum learning approach for media bias detection. They concluded that high-quality expert-labeled data improves the performance of the model.

[28] We mention multiple models for the task within Section 5.2.2.

5.3 Graph-Based

The research described in this section leverages graph data structures to analyze online social networks through their users and text interactions, which requires a distinctive set of methods for bias analysis. Although most publications used ML, we treat them separately due to the unique characteristics of the analyzed data representations. Graph-based approaches are primarily used to investigate framing bias, echo chambers, and political stances. Therefore, we structure our overview of corresponding publications by the type of bias they investigate.

Framing Bias. The SLAP4SLIP framework [105] detects how concepts are discussed in different parts of a social network with predefined linguistic features, graph NN, and structured sparsity. The authors exploit the network structure of discussion forums on Reddit without explicitly labeled data and minimally supervised features representing ideologically driven agenda setting and framing. Training graph auto-encoders, Hofmann et al. [105] modeled agenda setting, and framing for identifying ideological polarization within network structures of online discussion forums. They modeled polarization along the dimensions of salience and framing. Further, they proposed MultiCTX (Multi-level ConTeXt), a model consisting of contrastive learning and sentence graph attention networks to encode different levels of context, i.e., neighborhood context of adjacent sentences, article context, and event context.

Guo and Zhu [95] built on the SLAP4SLIP framework [105] to detect informational bias and ideological radicalization by combining contrastive learning and sentential graph networks. Similarly, Tran [228] proposed a framework for identifying bias in news sources. The authors used BERT Base for aspect-based sentiment analysis and assigned a bias score to each source with a graph-based algorithm.

Echo Chambers. Villa et al. [237] applied community detection strategies and modeled a COVID-19-related conversation graph to detect echo chambers. Their method considered the relationship between individuals and the semantic aspects of their shared content on Twitter. By partitioning four different representations of a graph (i.e., topology-based, sentiment-based, topic-based, and hybrid) with the METIS algorithm[29], followed up by qualitative methods, they assessed both the relationships connecting individuals and semantic aspects related to the content they share over Twitter. They also analyzed the controversy and homogeneity among the different polarized groups obtained.

Political Stance Detection. Stance detection[30] is a typical application of graph-based classification techniques. Zhou et al. [252] combined network structure learning analysis and NN to predict the political stance of news media outlets. With their semi-supervised network embedding approach, the authors built a training corpus on network information, including macro- and micro-network views. They primarily employed network embedding learning and graph-based label propagation to overcome label sparsity. By integrating graph embeddings as a feature, Stefanov et al. [223] detected the stance and political stance of Twitter users and online media by leveraging their retweet behavior. They used a user-to-hashtag graph and a user-to-mention graph and then ran node2vec. They achieved the best result for combining BERT with valence scores[31]. Guimarães et al. [92] analyzed news stories and political opinions shared on Brazilian Facebook. They proposed a graph-based semi-supervised learning approach to classify Facebook pages as politically left or right. Utilizing audience interaction information by inferring self-reported political leaning from Facebook pages, Guimarães et al. [92] built an interest graph to determine the stance of media outlets and public figures. The authors achieved the best results for label propagation with a spectral graph transducer. Li and Goldwasser [143] captured social context with a neural architecture for representing relational information with

[29] As proposed by Karypis and Kumar [120].
[30] We defined stance detection as political bias detection via the identification of linguistic biases, compare Section 5.2.1.
[31] A valence score [223] close to zero reflects that an influencer is cited evenly among different groups in a network. Conversely, a score close to −1 or 1 indicates that one group disproportionately cites an influencer compared to another group. In their paper, Stefanov et al. [223] indicated that valence scores are essential in identifying media bias in social networks.

graph-based representations and a graph convolutional network. They showed that using social information, such as Twitter users who have shared the article, can significantly improve performance with distant and direct supervision.

5.4 Bias in Language Models

Detecting bias inherent to language models is an important research area due to the models' popularity for many NLP tasks. Researchers have investigated bias in texts and other media generated by language models as well as in classification performed with language models. We did not include publications that address these forms of bias.[32] However, we would like to give some examples to raise awareness of biased language models. Nadeem et al. [166] analyzed stereotypical bias with the crowdsourced dataset StereoSet in BERT, GPT-2, ROBERTA, and XLNET, concluding that all models exhibit strong stereotypical bias. Vig et al. [236] used causal mediation analysis to analyze gender bias in language models. Their results showed that gender bias effects exist in specific components of language models. Bhardwaj et al. [24] also analyzed gender bias within BERT-layers and concluded that the layers are generally biased. In Liu et al. [148], the authors detected bias in texts generated by GPT-2 and discussed means of mitigating gender bias in language models by using a reinforcement learning framework.

5.5 Datasets

During our review, we collected both methods and datasets from the publications we selected for inclusion. In total, we found 123 datasets. We categorize the datasets according to the concepts proposed in our Media Bias Taxonomy, similar to the discussion of methodologies as shown in Table 2. We added the category General Linguistic Bias as several datasets do not define the subcategory of bias they contain. We did not evaluate the quality of the datasets as they address distinct tasks and objectives but leave this assessment for future work (cf. Section 7).

Only two of the 123 datasets include information on the background of annotators. Moreover, dataset sizes are generally small; only 21 of the 123 datasets contain more than 30,000 annotations. We believe that the use of multiple datasets is promising for future work as we discuss in Section 7. As part of this review, we present the datasets, their statistics, and tasks merely as a starting point for future work, without further assessment. We give a detailed overview of publications, sizes, availability, tasks, type of label, link, and publication summary for each dataset in our repository.

[32] We focus exclusively on detection methods; the field of bias in language models is extensive enough for a dedicated literature review.

Media Bias Category	Media Bias Type	Amount
Linguistic Bias		45
	General Linguistic Bias	26
	Framing Bias	15
	Epistemological Bias	3
	Bias by Semantic Properties	1
Text-level Context Bias		5
	Statement Bias	2
	Phrasing Bias	3
Reporting-level Context Bias		6
	General Reporting-level Context Bias	2
	Selection Bias	1
	Coverage Bias	2
	Proximity Bias	1
Cognitive Bias.		28
	Partisan Bias	28
Related Concepts		
	Hate Speech	14
	Group Bias	20
	Sentiment Analysis	10

Table 2. Overview of datasets found during our literature review

6 HUMAN-CENTERED RESEARCH ON MEDIA BIAS

Human-centered research on media bias aims to understand why people perceive media as biased, explore the societal and digital consequences, and develop strategies to overcome biased perception and detect media bias. Debates on all these factors are ongoing and experimental effects tend to be minor. Hereafter, we highlight some of these debates.

6.1 Reasons for biased media perception

One explanation for the emergence of cognitive biases in media perception is that information is processed in light of prior expectations, which may be distorted [169]. The veracity of claims is often judged based on familiarity, potentially resulting in illusory truths [75, 81, 127, 211]. Cognitive dissonance theory posits that people experience discomfort when confronted with information inconsistent with their convictions, motivating them to discount it [73].

Extending this notion to groups, Tajfel et al. [226] suggested in their social identity and categorization theory that basic self-esteem is derived from personal affiliation with positively-connotated groups. This results in in-group favoritism, out-group derogation [121, 198], and behavior and information processing in line with group identity. People easily regard reports that negatively affect groups they strongly identify with as a personal threat to their self-esteem and devalue these reports [86, 102]. Furthermore, Turner [230] posited that when people self-categorize with a specific group, they evaluate the validity of arguments by congruence to in-group norms and in-group consensus. This pattern

aligns with empirical findings showing that news acceptance depends on group identification and congruent group membership cues of the news source [191, 205].

Generally, prior works expect selective exposure to media to be consistent with previous viewpoints [125], further strengthening prior convictions. Such behaviour can be referred to as confirmation bias [169] through repeated exposure [54]. In the age of social media and the abundance of information available, these cognitive biases may further allow for confrontation only with attitude-consistent information and like-minded individuals in echo chambers [165, 170, 224]. Moreover, algorithms trained on these biases may further limit the available media spectrum in filter bubbles [176].

Consequently, limited exposure to alternative viewpoints may also impact the perception of social norms and the prevalence of opinions. The overestimation of the frequency of one's own position, known as the false consensus effect [195], has been widely documented even before the introduction of social media and may be partially due to identity motivations explained earlier [154]. However, when echo chambers are used to gauge the frequency of opinions and social norms, even larger shifts between groups are expected [141]. This feeds into a vicious circle of polarizing group norms, discounting information inconsistent with these shifted norms, and feeling encouraged to voice even more extreme positions (e.g., [59, 157, 205, 230]). These mechanisms lead to expectations that media perception is polarized based on social categories and prior beliefs and that the introduction of social media has exacerbated this phenomenon.

6.2 Consequences of biased media perception

Partisan individuals tend to select media that aligns with their prior beliefs and political attitudes, a phenomenon known as the Friendly Media Phenomenon (FMP) [17, 88, 127]. This tendency may be partially due to interpersonal communication among like-minded individuals [104]. People also tend to assess the veracity of information based on its fit with their political convictions, exhibiting partisan bias [82].

Biased media perception can lead to the Hostile Media Phenomenon (HMP), where people perceive media coverage as biased against their side, regardless of the actual political position of the article [101, 181, 231]. This effect increases with the extremity of party affiliation and is primarily due to the derogation of dissenting media [17, 83, 216], making it a cognitive bias rather than a characteristic of the media landscape. Discussions and feedback from like-minded individuals can further amplify the HMP, leading to the perception of general media bias even when primarily exposed to self-selected, like-minded media [34, 126].

Methodologically, the HMP, FMP, and partisan bias complicate the assessment of media bias, as raters' perceptions of bias may reflect more on individual affiliations and idiosyncrasies than the objective properties of the rated article [217]. Subjective bias ratings are relative to their social context; their quality as a scientific measure of media bias depends on the representativeness of raters. Therefore, such ratings should be supplemented by objective bipartisan bias criteria (e.g., language biases).

Socially, the HMP can lead to the mobilization of more extreme positions, distrust in the social system, and, in cases of low efficacy beliefs, political withdrawal [181]. Both the HMP and FMP can contribute to increased political segmentation and polarization, which can negatively impact political communication and interaction, essential for a peaceful and democratic society [83]. Exposure to certain media can also have social consequences, such as altered political participation [126]. For example, Dvir-Gvirsman et al. [62] found that exposure to congruent media is tied to biased perceptions of the opinion climate, influencing how participants communicate their political beliefs and engage in politically meaningful acts, while incongruent exposure has little effect.

The role of the social media environment in this process is somewhat disputed: While selective exposure in social media is widely documented [55, 170], some authors argue that social media is not the main contributor to the variety

of media diets globally. For example, Zuiderveen Borgesius et al. [253] deem its general impact negligible and suggest it may expose users to more diverse information compared to traditional media. According to Dubois and Blank [61], people may even cope with this high-choice media environment by developing strategies like verifying news in different outlets, and—even though social networks are polarized—only a subset of the population regards itself as susceptible to echo chambers. After all, the phenomena and underlying cognitive processes were known before the advent of social media. The effects observed in social media may just be more visible to researchers than they were before [253]. In addition, exposure to biased media may not be sufficient to significantly affect attitudes [233]. As such, it is challenging to determine the overall effect of social media on biased media perception and social consequences today, though some feedback loops can be expected [59]. This problem is even more pressing for algorithmic filtering than for personal selections, as the algorithms involved are not transparently disclosed, their application is in flux, and they are not accessible to the user [253]. This fact illustrates that parts of the conclusion on the impact of social media on media bias phenomena are also driven by the selection of media and the assessment method of the effects.

6.3 Recipient-oriented approaches to reduce media bias

Given that selective media exposure partially explains cognitive media bias phenomena, one intervention approach is to encourage and facilitate a diverse media diet to reduce media bias [61]. This can be achieved by plug-ins that actively diversify the media displayed in a search by identifying the topic and sampling other articles or information related to it [172], or by providing media based on another individual's platform history [25]. In a similar approach, Munson et al. [165] used a browser widget to provide feedback on the balance of a user's media diet, successfully encouraging these users to explore more media from centrist and opposing viewpoints.

Other experiments and observations of counter-attitudinal exposure illustrate that the mere presentation and reception of opposing viewpoints do not always decrease the HMP and may even exacerbate the problem. For instance, Weeks et al. [240] found that people who were incidentally exposed to counter-attitudinal information are more likely to subsequently select information that aligns with their attitudes. Other studies found that exposure to incongruent comments increases the perception of bias and decreases the perception of the credibility of a later, neutral news report [83], and that exposure to opposing tweets may backfire and intensify political polarization, particularly for Republicans [13]. These findings are consistent with the notion of motivated reasoning, as the potential threat of backfiring from inconsistent exposure—though rather dependent on the specific materials to which readers are exposed [225]—may be explained by the threat of the presented material to the reader's identity. As a result, diverse exposure with well-crafted materials may help but is not a comprehensive solution for the HMP, FMP, and biased media perception.

As an alternative, some studies have attempted to alter the user's mindset during news processing and shift the attentional focus to aspects of a user's self-identity that are not challenged by the news report. For example, inducing self-affirming thoughts aimed at mitigating the potentially self-threatening aspect of belief-inconsistent arguments has been shown to successfully evoke more unbiased processing of such information [42]. Similarly, focusing readers' attention on a value that may be threatened by information increases their perception of media bias in that article [123]. Likewise, people seem more open to sharing and are better at judging news headlines based on their veracity when nudged to think about their own accuracy instead of their identity motives [180]. Opening the mindset may thus be an effective, albeit situational, approach when tackling phenomena such as the HMP and media bias detection during exposure to attitude-inconsistent materials.

As an additional step, forewarning messages that draw attention to biased media and potential influencing attempts can help "inoculate" against this media by provoking reactance towards manipulations [194, 201, 216]. Exposing

individuals to examples of media bias through such messages may teach them to detect and cope with it. In this vein, various forms of training have been tested and generally increase a reader's ability to identify biased media and distinguish it from congruency with one's political stance [213, 216, 233]. This detailed training is necessary, as mere awareness of media bias as part of general news media literacy may not be sufficient for a balanced media diet [229].

Overall, all approaches have yielded relatively small effects on improving media bias detection, and more research on effective interventions is necessary. Regarding partisan bias, there is some indication that interventions are not equally effective in reducing the bias for liberals and conservatives—potentially inadvertently biasing the overall discourse on media towards the less open-minded faction [188, 216]. Thus, further testing of the effectiveness of approaches in reducing partisan media perception and the HMP is warranted.

7 DISCUSSION

To address RQ1, we have established a Media Bias Taxonomy that allows to precisely categorize the various sub-concepts related to media bias [217, 221]. We emphasize the complexity of media bias and note that researchers often fail to clearly define the type of media bias they investigate, which leads to confusion when comparing different studies. Furthermore, existing literature reviews on the topic do not address the various media bias concepts [98], making it difficult to understand problems and solutions across different approaches.

Our Media Bias Taxonomy is a crucial first step in establishing a common ground for more clearly defined media bias research. We divide media bias into five major categories: linguistic bias, cognitive bias, text-level context bias, reporting-level bias, and related concepts. We provide subgroups for each of these categories. Throughout the creation of our taxonomy, we engaged in frequent discussions and revised our definitions and structure multiple times, revealing the numerous options available for defining media bias.

While our taxonomy provides a practical foundation and effective starting point for research in the domain, future research should critically re-examine the discussed concepts. We believe that the main common ground among the various types of media bias we identified is smaller than that of existing universal definitions (see Section 4.1) and primarily refers to one-sided media content

To answer RQ2 and RQ3, we provided an extensive overview of recently published literature on computer science methods and datasets for media bias detection. We manually inspected over 1,528 computer science research papers on the topic published between 2019 to May 2022 after automatically filtering over 100,000 keyword-related publications. Our review reveals valuable insights into best practices and trends in the research field.

In recent years, transformers have quickly become the most frequently used and most reliable method for media bias detection and debiasing [221]. Platforms like Hugging Face facilitate the implementation of the models and their adaption to various tasks [56]. However, as we show in Section 5, the new models have not yet made their way into all subtypes of bias, leaving room for future experiments. Additionally, available media bias classifiers are largely based on small in-domain datasets. Recent advancements in natural language processing, especially transformer-based models, demonstrate how accurate results can be achieved by unsupervised or supervised training on massive text corpora [9] and by model pre-training using inter and cross-domain datasets [9].

Although graph-based methods are not as popular as transformers, their application to media bias detection is increasing but mostly limited to analyzing social network content, activities, and structures, and identifying structural political stances within these entities [92, 143, 223, 252]. Transformer-based approaches cannot accomplish such an analysis due to the network properties of the explored data.

Established methods still play a role in media bias detection. Traditional natural language processing approaches, as well as non-transformer-based (deep NN) machine learning models, are simpler and more explainable compared to language-model-based approaches, making them advantageous in applications where transparency of classification decisions is critical (e.g., [222]). Since traditional approaches have been used in many media bias identification tasks, they often serve as a baseline to compare new (transformer-based) approaches. Given their higher explainability and long-term testing, we don't expect language models to completely replace other approaches soon.

Apart from these major trends, including information on spreading behavior, social information [40, 143, 223], metadata [56], and examining the vector spaces of word embeddings [56] also show promise in improving classifier performance to detect media bias.

We addressed RQ4 by reviewing social science research on media bias. One significant takeaway is that media bias datasets largely ignore insights from social science research on the topic, leading to low annotator agreement and less accurate annotations [221]. The perception of bias depends on factors beyond content, such as the reader's background and understanding of the text. Moreover, limited exposure to alternative viewpoints can impact how social norms and opinions are perceived. These insights have never been fully integrated into automated detection methods or datasets. Integrating bias perception research in language models is a promising way to improve annotation-based detection systems [24], which can potentially be achieved by further developing standardized questions within the domain [221].

We see a need to develop further methods to increase news consumers' bias awareness and believe that computer science methods, as described in this review, can be a powerful tool to build such awareness-increasing tools. While some tools already exist, none have been applied on a larger scale in a real-world scenario, which is a promising direction for future research.

Our literature review also exhibits limitations. First, we excluded work from areas other than media bias due to the high number of publications involved, potentially leaving out valuable contributions. Investigating promising concepts from other areas will be necessary for future work. Second, for all computer science methods, we only included literature from 2019 to 2022, excluding valuable earlier research. Analyzing a longer period could yield an even more complete picture of the research domain. Lastly, although we distinguish several categories within our Media Bias Taxonomy, the concepts related to media bias still overlap and appear concurrently. We believe that future work should further discuss and adapt the taxonomy. Although the taxonomy we present is merely a starting point to connect works in the area, we believe it can benefit future approaches by raising awareness of concepts, methods, and datasets in the research domain. During the writing of this literature review, the taxonomy's outline frequently changed in permanent discussions among the authors.

8 CONCLUSION

In 2018, Hamborg et al. [98] concluded that (1) powerful computer science methods (such as word embeddings and deep learning) had not yet made their way into the automated detection of media bias and that (2) the interdisciplinarity of media bias research should be improved in the future. The authors suggested (3) that approaches in computer science did not account for bias having many different forms and usually only focused on narrow bias definitions [98].

Our literature review reveals that two of these propositions (1 and 3) have been addressed to some extent, but there is still considerable room for improvement. Transformer and graph-based methods have led to significant increases in the performance of automated methods for detecting media bias, and numerous types of bias have received research attention. However, these concepts are primarily used and analyzed individually, with knowledge overlaps between

them remaining unexplored [212]. Recent modeling techniques, such as multi-task learning, enable the use of related datasets to improve classification performance [9].

Regarding (2), datasets and systems still exhibit limited conceptual work, with the cognitive dimension of media bias rarely mentioned in computer science research. Our literature review aims to provide a foundation for increased awareness of bias in media bias datasets (through standardized annotator background assessments), enhanced interdisciplinarity in the research domain (which we believe is particularly relevant since reasonable classifications cannot exist without clear conceptualizations), and future computer science methods.

We are confident that this review will facilitate entry into media bias research and help experienced researchers identify related works. We hope that our findings will contribute to the development of more effective and efficient media bias detection methods and systems to increase media bias awareness. Finally, we plan to repeat our workflow in three years to reassess the state of the research domain.

ACKNOWLEDGMENTS

We thank Elisabeth Richter, Felix Blochwitz, Jerome Wassmuth, Sudharsana Kannan, and Jelena Mitrović for supporting this project through fruitful discussions. We are grateful for the financial support of this project provided by the Hanns-Seidel Foundation, the DAAD (German Academic Exchange Service), the Lower Saxony Ministry of Science and Culture, and the VW Foundation.

REFERENCES

[1] 1967. *Second Conference Internationale Sur Le Traitement Automatique Des Langues, COLING 1967, Grenoble, France, August 1967.* https://aclanthology.org/volumes/C67-1/

[2] Victoria Patricia Aires, Juliana Freire, and Fabiola G. Nakamura. 2020. An Information Theory Approach to Detect Media Bias in News Websites. (2020), 9.

[3] Muhammad Z. Ali, Ehsan-Ul-Haq, Sahar Rauf, Kashif Javed, and Sarmad Hussain. 2021. Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis. *IEEE Access* 9 (2021), 84296–84305. https://doi.org/10.1109/ACCESS.2021.3087827

[4] Emily Allaway and Kathleen McKeown. 2021. A Unified Feature Representation for Lexical Connotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 2145–2163. https://doi.org/10.18653/v1/2021.eacl-main.184

[5] Karel Jan Alsem, Steven Brakman, Lex Hoogduin, and Gerard Kuper. 2008. The impact of newspapers on consumer confidence: does spin bias exist? *Applied Economics* 40, 5 (2008), 531–539. https://doi.org/10.1080/00036840600707100

[6] Khudran Alzhrani. 2020. Ideology Detection of Personalized Political News Coverage: A New Dataset. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis* (Silicon Valley CA USA). ACM, 10–15. https://doi.org/10.1145/3388142.3388149

[7] Ananya, Nitya Parthasarthi, and Sameer Singh. 2019. GenderQuant: Quantifying Mention-Level Genderedness. In *Proceedings of the 2019 Conference of the North* (Minneapolis, Minnesota). Association for Computational Linguistics, 2959–2969. https://doi.org/10.18653/v1/N19-1303

[8] Talita Anthonio and Lennart Kloppenburg. 2019. Team Kermit-the-frog at SemEval-2019 Task 4: Bias Detection Through Sentiment Analysis and Simple Linguistic Features. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (Minneapolis, Minnesota, USA). Association for Computational Linguistics, 1016–1020. https://doi.org/10.18653/v1/S19-2177

[9] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Prakash Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. *CoRR* abs/2111.10952 (2021). https://doi.org/10.48550/arXiv.2111.10952

[10] Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. *PLOS ONE* 16, 1 (01 2021), 1–28. https://doi.org/10.1371/journal.pone.0245533

[11] Larry Atkins. 2016. *Skewed: A Critical Thinker's Guide to Media Bias.* Prometheus Books, Amherst, New York.

[12] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. In *The World Wide Web Conference on - WWW '19* (San Francisco, CA, USA). ACM Press, 49–59. https://doi.org/10.1145/3308558.3313504

[13] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the*

National Academy of Sciences 115, 37 (2018), 9216–9221. https://doi.org/10.1073/pnas.1804840115

[14] Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 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* (Online). Association for Computational Linguistics, 3364–3374. https://doi.org/10.18653/v1/2020.acl-main.308

[15] Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-Task Ordinal Regression for Jointly Predicting the Trustworthiness and the Leading Political Ideology of News Media. arXiv:1904.00542 [cs, stat] http://arxiv.org/abs/1904.00542

[16] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. arXiv:2010.05338 [cs] http://arxiv.org/abs/2010.05338

[17] Matthew Barnidge, Albert C. Gunther, Jinha Kim, Yangsun Hong, Mallory Perryman, Swee Kiat Tay, and Sandra Knisely. 2020. Politically Motivated Selective Exposure and Perceived Media Bias. *Communication Research* 47, 1 (2020), 82–103. https://doi.org/10.1177/0093650217713066

[18] Cédric Batailler, Skylar M. Brannon, Paul E. Teas, and Bertram Gawronski. 2022. A Signal Detection Approach to Understanding the Identification of Fake News. *Perspectives on Psychological Science* 17, 1 (2022), 78–98. https://doi.org/10.1177/1745691620986135 PMID: 34264150.

[19] Lisa Bauer, Karthik Gopalakrishnan, Spandana Gella, Yang Liu, Mohit Bansal, and Dilek Hakkani-Tur. 2022. Analyzing the Limits of Self-Supervision in Handling Bias in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7372–7386. https://aclanthology.org/2022.findings-emnlp.545

[20] Matthew A. Baum and Phil Gussin. 2008. In the Eye of the Beholder: How Information Shortcuts Shape Individual Perceptions of Bias in the Media. *Quarterly Journal of Political Science* 3, 1 (2008), 1–31. https://doi.org/10.1561/100.00007010

[21] Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and Comparing Computational Approaches for Identifying the Language of Framing in Political News. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. Denver, Colorado, 1472–1482,. https://doi.org/10.3115/v1/N15-1171

[22] Jonathan Berant, Ido Dagan, Meni Adler, and Jacob Goldberger. 2012. Efficient Tree-based Approximation for Entailment Graph Learning. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, 117–125. https://aclanthology.org/P12-1013

[23] Camiel J. Beukeboom and Christian Burgers. 2017. Linguistic Bias. https://doi.org/10.1093/acrefore/9780190228613.013.439

[24] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating Gender Bias in BERT. *Cognitive Computation* 13, 4 (2021), 1008–1018. https://doi.org/10.1007/s12559-021-09881-2

[25] Md Momen Bhuiyan, Carlos Augusto Bautista Isaza, Tanushree Mitra, and Sang Won Lee. 2022. OtherTube: Facilitating Content Discovery and Reflection by Exchanging YouTube Recommendations with Strangers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 204, 17 pages. https://doi.org/10.1145/3491102.3502028

[26] Ciara Blackledge and Amir Atapour-Abarghouei. 2021. Transforming Fake News: Robust Generalisable News Classification Using Transformers. (2021). https://doi.org/10.48550/ARXIV.2109.09796

[27] Fernando Blanco. 2017. *Cognitive Bias*. Springer International Publishing, Cham, 1–7. https://doi.org/10.1007/978-3-319-47829-6_1244-1

[28] Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North* (Minneapolis, Minnesota). Association for Computational Linguistics, 7–15. https://doi.org/10.18653/v1/N19-3002

[29] Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. 2018. Selection Bias in News Coverage: Learning it, Fighting it. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (Lyon, France). ACM Press, 535–543. https://doi.org/10.1145/3184558.3188724

[30] Pearl Brereton, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil. 2007. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software* 80, 4 (2007), 571–583. https://doi.org/10.1016/j.jss.2006.07.009

[31] Ceren Budak, Sharad Goel, and Justin M. Rao. 2016. Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. *Public Opinion Quarterly* 80, S1 (04 2016), 250–271. https://doi.org/10.1093/poq/nfw007

[32] Dianne G. Bystrom, Terry A. Robertson, and Mary Christine Banwart. 2001. Framing the Fight: An Analysis of Media Coverage of Female and Male Candidates in Primary Races for Governor and U.S. Senate in 2000. *American Behavioral Scientist* 44, 12 (2001), 1999–2013. https://doi.org/10.1177/00027640121958456

[33] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230

[34] Laia Castro, David Nicolas Hopmann, and Lilach Nir. 2020. Whose media are hostile? The spillover effect of interpersonal discussions on media bias perceptions. *Communications* (2020), 000010151520190140. https://doi.org/10.1515/commun-2019-0140

[35] Tuhin Chakrabarty, Christopher Hidey, and Smaranda Muresan. 2021. ENTRUST: Argument Reframing with Language Models and Entailment. arXiv:2103.06758 [cs.CL]

[36] Wen Chen, Diogo Pacheco, Kai-Cheng Yang, and Filippo Menczer. 2021. Neutral bots probe political bias on social media. 12, 1 (2021), 5580. https://doi.org/10.1038/s41467-021-25738-6

[37] Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting Media Bias in News Articles using Gaussian Bias Distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (Online). Association for Computational Linguistics, 4290–4300. https://doi.org/10.18653/v1/2020.findings-emnlp.383

[38] Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science* (Online). Association for Computational Linguistics, 149–154. https://doi.org/10.18653/v1/2020.nlpcss-1.16

[39] Lamogha Chiazor, Geeth de Mel, Graham White, Gwilym Newton, Joe Pavitt, and Richard Tomsett. [n.d.]. An Automated Framework to Identify and Eliminate Systemic Racial Bias in the Media. *CEUR Workshop Proceedings* 2812 (02 [n. d.]), 32–36. http://ceur-ws.org/Vol-2812/

[40] Uthsav Chitra and Christopher Musco. 2020. Analyzing the Impact of Filter Bubbles on Social Network Polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston TX USA). ACM, 115–123. https://doi.org/10.1145/3336191.3371825

[41] Laurenz A. Cornelissen, Lucia I. Daly, Qhama Sinandile, Heinrich de Lange, and Richard J. Barnett. 2019. A Computational Analysis of News Media Bias: A South African Case Study. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019 on ZZZ - SAICSIT '19* (Skukuza, South Africa). ACM Press, 1–10. https://doi.org/10.1145/3351108.3351134

[42] Joshua Correll, Steven J. Spencer, and Mark P. Zanna. 2004. An affirmed self and an open mind: Self-affirmation and sensitivity to argument strength. *Journal of Experimental Social Psychology* 40, 3 (2004), 350–356. https://doi.org/10.1016/j.jesp.2003.07.001

[43] Marta Costa-jussa. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence* 1 (11 2019), 495–496. https://doi.org/10.1038/s42256-019-0105-5

[44] André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. 2020. On document representations for detection of biased news articles. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (Brno Czech Republic). Association for Computing Machinery, New York, NY, USA, 892–899. https://doi.org/10.1145/3341105.3374025

[45] Rodrigo Cuéllar-Hidalgo, Julio de Jesús Guerrero-Zambrano, Dominic Forest, Gerardo Reyes-Salgado, and Juan-Manuel Torres-Moreno. 2021. LUC at ComMA-2021 Shared Task: Multilingual Gender Biased and Communal Language Identification without using linguistic features. arXiv:2112.10189 [cs] http://arxiv.org/abs/2112.10189

[46] Jamell Dacon and Haochen Liu. 2021. Does Gender Matter in the News? Detecting and Examining Gender Bias in News Articles. In *Companion Proceedings of the Web Conference 2021* (Ljubljana Slovenia). Association for Computing Machinery, New York, NY, USA, 385–392. https://doi.org/10.1145/3442442.3452325

[47] Dave D'Alessio and Mike Allen. 2000. Media Bias in Presidential Elections: A Meta-Analysis. *Journal of Communication* 50, 4 (01 2000), 133–156. https://doi.org/10.1111/j.1460-2466.2000.tb02866.x

[48] Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. 2015. Media Bias in German Online Newspapers. In *HT '15*.

[49] Samantha D'Alonzo and Max Tegmark. 2021. Machine-Learning media bias. arXiv:2109.00024 [cs] http://arxiv.org/abs/2109.00024

[50] Russell J. Dalton, Paul A. Beck, and Robert Huckfeldt. 1998. Partisan Cues and the Media: Information Flows in the 1992 Presidential Election. *The American Political Science Review* 92, 1 (1998), 111–126. https://doi.org/10.2307/2585932

[51] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (05 2017), 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

[52] Gabriel De Arruda, Norton Roman, and Ana Monteiro. 2020. Analysing Bias in Political News. 26, 2 (2020), 173–199. https://doi.org/10.3897/jucs.2020.011

[53] Claes H. de Vreese. 2005. News framing: Theory and typology. *Information Design Journal* 13, 1 (2005), 51–62. https://doi.org/10.1075/idjdd.13.1.06vre

[54] Alice Dechêne, Christoph Stahl, Jochim Hansen, and Michaela Wänke. 2010. The truth about the truth: A meta-analytic review of the truth effect. *Personality 10.1073/pnas.1517441113 Social Psychology Review* 14 (2010), 238–257. https://doi.org/10.1177/1088868309352251

[55] Michela Del Vicario, Fabiana Zollo, Guido Caldarelli, Antonio Scala, and Walter Quattrociocchi. 2017. Mapping social dynamics on Facebook: The Brexit debate. *Social Networks* 50 (2017), 6–16. https://doi.org/10.1016/j.socnet.2017.02.002

[56] Yoan Dinkov, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Predicting the Leading Political Ideology of YouTube Channels Using Acoustic, Textual, and Metadata Information. In *Interspeech 2019*. ISCA, 501–505. https://doi.org/10.21437/Interspeech.2019-2965

[57] Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A Survey of Natural Language Generation. *ACM Comput. Surv.* (jul 2022). https://doi.org/10.1145/3554727

[58] Marko Dragojevic, Alexander Sink, and Dana Mastro. 2017. Evidence of Linguistic Intergroup Bias in U.S. Print News Coverage of Immigration. *Journal of Language and Social Psychology* 36, 4 (2017), 462–472. https://doi.org/10.1177/0261927X16666884

[59] James N. Druckman, Matthew S. Levendusky, and Audrey McLain. 2018. No Need to Watch: How the Effects of Partisan Media Can Spread via Interpersonal Discussions. *American Journal of Political Science* 62, 1 (2018), 99–112. https://doi.org/10.7910/DVN/TJKIWN

[60] James N. Druckman and Michael Parkin. 2005. The Impact of Media Bias: How Editorial Slant Affects Voters. *The Journal of Politics* 67, 4 (2005), 1030–1049. https://doi.org/10.1111/j.1468-2508.2005.00349.x

[61] Elizabeth Dubois and Grant Blank. 2018. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society* 21, 5 (2018), 729–745. https://doi.org/10.1080/1369118X.2018.1428656

[62] Shira Dvir-Gvirsman, R. Kelly Garrett, and Yariv Tsfati. 2018. Why Do Partisan Audiences Participate? Perceived Public Opinion as the Mediating Mechanism. *Communication Research* 45, 1 (2018), 112–136. https://doi.org/10.1177/0093650215593145

[63] Jana Laura Egelhofer and Sophie Lecheler. 2019. Fake news as a two-dimensional phenomenon: a framework and research agenda. *Annals of the International Communication Association* 43, 2 (2019), 97–116. https://doi.org/10.1080/23808985.2019.1602782

[64] Kenneth C. Enevoldsen and Lasse Hansen. 2017. Analysing Political Biases in Danish Newspapers Using Sentiment Analysis. *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift* 2, 2 (07 2017), 87–98. https://tidsskrift.dk/lwo/article/view/96014

[65] Robert M. Entman. 2007. Framing Bias: Media in the Distribution of Power. *Journal of Communication* 57, 1 (02 2007), 163–173. https://doi.org/10.1111/j.1460-2466.2006.00336.x

[66] Robert M. Entman. 2010. Media framing biases and political power: Explaining slant in news of Campaign 2008. *Journalism* 11, 4 (2010), 389–408.

[67] William P. Eveland Jr. and Dhavan V. Shah. 2003. The Impact of Individual and Interpersonal Factors on Perceived News Media Bias. *Political Psychology* 24, 1 (2003), 101–117. https://doi.org/10.1111/0162-895X.00318

[68] Jody Condit Fagan. 2017. An Evidence-Based Review of Academic Web Search Engines, 2014-2016: Implications for Librarians' Practice and Research Agenda. *Information Technology and Libraries* 36, 2 (Jun. 2017), 7–47. https://doi.org/10.6017/ital.v36i2.9718

[69] Tiziano Fagni and Stefano Cresci. 2022. Fine-grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning. 73 (2022), 633–672. https://doi.org/10.1613/jair.1.13112

[70] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In Plain Sight: Media Bias Through the Lens of Factual Reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China). Association for Computational Linguistics, 6342–6348. https://doi.org/10.18653/v1/D19-1664

[71] Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Tenth International AAAI Conference on Web and Social Media*.

[72] Xavier Ferrer, Tom van Nuenen, Jose M. Such, and Natalia Criado. 2020. Discovering and Categorising Language Biases in Reddit. (2020). https://doi.org/10.48550/ARXIV.2008.02754

[73] Leon Festinger. 1957. *A theory of cognitive dissonance*. Vol. 2. Stanford university press.

[74] Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised Discovery of Implicit Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online). Association for Computational Linguistics, 596–608. https://doi.org/10.18653/v1/2020.emnlp-main.44

[75] Peter Fischer, Eva Jonas, Dieter Frey, and Stefan Schulz-Hardt. 2005. Selective exposure to information: the impact of information limits. *European Journal of Social Psychology* 35, 4 (2005), 469–492. https://doi.org/10.1002/ejsp.264

[76] Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2019. Academic Plagiarism Detection: A Systematic Literature Review. *ACM Comput. Surv.* 52, 6, Article 112 (Oct. 2019), 42 pages. https://doi.org/10.1145/3345317

[77] Johan Galtung and Mari Holmboe Ruge. 1965. The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers. *Journal of Peace Research* 2, 1 (1965), 64–90. https://doi.org/10.1177/002234336500200104

[78] Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting Political Bias in News Articles Using Headline Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy, 77–84. https://doi.org/10.18653/v1/W19-4809

[79] Soumen Ganguly, Juhi Kulshrestha, Jisun An, and Haewoon Kwak. 2020. Empirical Evaluation of Three Common Assumptions in Building Political Media Bias Datasets. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles (Eds.). AAAI Press, 939–943. https://doi.org/10.1609/icwsm.v14i1.7362

[80] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-Dependent Sentiment Classification With BERT. *IEEE Access* 7 (2019), 154290–154299. https://doi.org/10.1109/ACCESS.2019.2946594

[81] R. Kelly Garrett. 2009. Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. *Journal of Communication* 59, 4 (2009), 676–699. https://doi.org/10.1111/j.1460-2466.2009.01452.x

[82] Bertram Gawronski. 2021. Partisan bias in the identification of fake news. *Trends in Cognitive Sciences* 25, 9 (2021), 723–724. https://doi.org/10.1016/j.tics.2021.05.001

[83] Sherice Gearhart, Alexander Moe, and Bingbing Zhang. 2020. Hostile media bias on social media: Testing the effect of user comments on perceptions of news bias and credibility. *Human Behavior and Emerging Technologies* 2, 2 (2020), 140–148. https://doi.org/10.1002/hbe2.185

[84] Gerald Ki Wei Huang and Jun Choi Lee. 2021. Hyperpartisan News Classification with ELMo and Bias Feature. 37, 5 (2021). https://doi.org/10.6688/JISE.202109_37(5).0013

[85] Sarah Gershon. 2012. When Race, Gender, and the Media Intersect: Campaign News Coverage of Minority Congresswomen. *Journal of Women, Politics & Policy* 33, 2 (2012), 105–125. https://doi.org/10.1080/1554477X.2012.667743

[86] Bryan T. Gervais. 2015. Incivility Online: Affective and Behavioral Reactions to Uncivil Political Posts in a Web-based Experiment. *Journal of Information Technology & Politics* 12, 2 (2015), 167–185. https://doi.org/10.1080/19331681.2014.997416

[87] Carroll J. Glynn and Michael E. Huge. 2014. How Pervasive Are Perceptions of Bias? Exploring Judgments of Media Bias in Financial News. *International Journal of Public Opinion Research* 26, 4 (02 2014), 543–553. https://doi.org/10.1093/ijpor/edu004

[88] Seth K. Goldman and Diana C. Mutz. 2011. The Friendly Media Phenomenon: A Cross-National Analysis of Cross-Cutting Exposure. *Political Communication* 28, 1 (2011), 42–66. https://doi.org/10.1080/10584609.2010.544280

[89] Melanie C. Green and Timothy C. Brock. 2005. Persuasiveness of Narratives. In *Persuasion: Psychological insights and perspectives, 2nd ed.* Sage Publications, Inc, Thousand Oaks, CA, US, 117–142.

[90] Stephan Greene and Philip Resnik. 2009. More than Words: Syntactic Packaging and Implicit Sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Boulder, Colorado, 503–511. https://aclanthology.org/N09-1057

[91] Tim Groseclose and Jeffrey Milyo. 2005. A social-science perspective on media bias. *Critical Review* 17, 3-4 (2005), 305–314. https://doi.org/10.1080/08913810508443641

[92] Samuel S. Guimarães, Julio C. S. Reis, Marisa Vasconcelos, and Fabrício Benevenuto. 2021. Characterizing political bias and comments associated with news on Brazilian Facebook. 11, 1 (2021), 94. https://doi.org/10.1007/s13278-021-00806-3

[93] Albert C. Gunther and Janice L. Liebhart. 2006. Broad Reach or Biased Source? Decomposing the Hostile Media Effect. *Journal of Communication* 56, 3 (2006), 449–466. https://doi.org/10.1111/j.1460-2466.2006.00295.x

[94] Albert C. Gunther and Kathleen Schmitt. 2004. Mapping Boundaries of the Hostile Media Effect. *Journal of Communication* 54, 1 (2004), 55–70. https://doi.org/10.1111/j.1460-2466.2004.tb02613.x

[95] Shijia Guo and Kenny Q. Zhu. 2022. Modeling Multi-level Context for Informational Bias Detection by Contrastive Learning and Sentential Graph Network. arXiv:2201.10376 [cs] http://arxiv.org/abs/2201.10376

[96] Jonathan Haidt and Craig Joseph. 2004. Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus* 133, 4 (2004), 55–66. https://doi.org/10.1162/0011526042365555

[97] Felix Hamborg and Karsten Donnay. 2021. NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1663–1675. https://doi.org/10.18653/v1/2021.eacl-main.142

[98] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20, 4 (01 Dec 2019), 391–415. https://doi.org/10.1007/s00799-018-0261-y

[99] Jiyoung Han, Youngin Lee, Junbum Lee, and Meeyoung Cha. 2019. The Fallacy of Echo Chambers: Analyzing the Political Slants of User-Generated News Comments in Korean Media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)* (Hong Kong, China). Association for Computational Linguistics, 370–374. https://doi.org/10.18653/v1/D19-5548

[100] Abdelhakim Hannousse. 2021. Searching relevant papers for software engineering secondary studies: Semantic Scholar coverage and identification role. *IET Software* 15, 1 (2021), 126–146. https://doi.org/10.1049/sfw2.12011

[101] Glenn J. Hansen and Hyunjung Kim. 2011. Is the Media Biased Against Me? A Meta-Analysis of the Hostile Media Effect Research. *Communication Research Reports* 28, 2 (2011), 169–179. https://doi.org/10.1080/08824096.2011.565280

[102] Tilo Hartmann and Martin Tanis. 2013. Examining the hostile media effect as an intergroup phenomenon: The role of ingroup identification and status. *Journal of Communication* 63, 3 (2013), 535–555. https://doi.org/10.1111/jcom.12031

[103] Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Detect and Perturb: Neutral Rewriting of Biased and Sensitive Text via Gradient-based Decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (Punta Cana, Dominican Republic). Association for Computational Linguistics, 4173–4181. https://doi.org/10.18653/v1/2021.findings-emnlp.352

[104] Jay D. Hmielowski, Sarah Staggs, Myiah J. Hutchens, and Michael A. Beam. 2022. Talking Politics: The Relationship Between Supportive and Opposing Discussion With Partisan Media Credibility and Use. *Communication Research* 49, 2 (2022), 221–244. https://doi.org/10.1177/0093650220915041

[105] Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2021. Modeling Ideological Agenda Setting and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity. (2021). https://doi.org/10.48550/ARXIV.2104.08829

[106] Joan B. Hooper. 1975. *On Assertive Predicates*. Brill, Leiden, Niederlande, 91 – 124. https://doi.org/10.1163/9789004368828_005

[107] J. Brian Houston, Glenn J. Hansen, and Gwendelyn S. Nisbett. 2011. Influence of User Comments on Perceptions of Media Bias and Third-Person Effect in Online News. *Electronic News* 5, 2 (2011), 79–92. https://doi.org/10.1177/1931243111407618

[108] Christoph Hube. 2020. Methods for detecting and mitigating linguistic bias in text corpora. (2020). https://doi.org/10.15488/9873 Publisher: Hannover : Institutionelles Repositorium der Leibniz Universität Hannover.

[109] Christoph Hube and Besnik Fetahu. 2018. Detecting Biased Statements in Wikipedia. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1779–1786. https://doi.org/10.1145/3184558.3191640

[110] Christoph Hube and Besnik Fetahu. 2019. Neural Based Statement Classification for Biased Language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC Australia). ACM, 195–203. https://doi.org/10.1145/3289600.3291018

[111] Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 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* (Online). Association for Computational Linguistics, 1921–1945. https://doi.org/10.18653/v1/2021.eacl-main.165

[112] Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. A&C Black.

[113] Rami Ibrahim and M. Omair Shafiq. 2022. Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. *ACM Comput. Surv.* (aug 2022). https://doi.org/10.1145/3563691

[114] Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 840–844. https://doi.org/10.18653/v1/S19-2146

[115] Ye Jiang, Yimin Wang, Xingyi Song, and Diana Maynard. 2020. Comparing Topic-Aware Neural Networks for Bias Detection of News. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020) (Frontiers in Artificial Intelligence and Applications, Vol. 325)*, Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang (Eds.). IOS Press, 2054–2061. https://doi.org/10.3233/FAIA200327

[116] Valgerður Jóhannsdóttir and Þorgerður Einarsdóttir. 2015. Gender Bias in the Media: The Case of Iceland. *Stjórnmál og Stjórnsýsla* 11, 2 (Autumn 2015), 207–230. https://doi.org/10.13177/irpa.a.2015.11.2.5

[117] Lalitha Kameswari and Radhika Mamidi. 2021. Towards Quantifying Magnitude of Political Bias in News Articles Using a Novel Annotation Schema. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021*, Galia Angelova, Maria Kunilovskaya, Ruslan Mitkov, and Ivelina Nikolova-Koleva (Eds.). INCOMA Ltd., 671–678. https://aclanthology.org/2021.ranlp-1.76

[118] Lalitha Kameswari, Dama Sravani, and Radhika Mamidi. 2020. Enhancing Bias Detection in Political News Using Pragmatic Presupposition. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media* (Online). Association for Computational Linguistics, 1–6. https://doi.org/10.18653/v1/2020.socialnlp-1.1

[119] Lauri Karttunen. 1971. Implicative Verbs. *Language* 47, 2 (1971), 340–358. https://doi.org/10.2307/412084

[120] George Karypis and Vipin Kumar. 1995. METIS – Unstructured Graph Partitioning and Sparse Matrix Ordering System, Version 2.0. Technical report. *University of Minnesota, Department of Computer Science, Minneapolis, MN 55455* (01 1995).

[121] Caroline Kelly. 1989. Political identity and perceived intragroup homogeneity. *British Journal of Social Psychology* 28, 3 (1989), 239–250. https://doi.org/10.1111/j.2044-8309.1989.tb00866.x

[122] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 829–839. https://doi.org/10.18653/v1/S19-2145

[123] Kenneth Kim. 2019. The Hostile Media Phenomenon: Testing the Effect of News Framing on Perceptions of Media Bias. *Communication Research Reports* 36, 1 (2019), 35–44. https://doi.org/10.1080/08824096.2018.1555659

[124] Barbara Kitchenham. 2004. *Procedures for Performing Systematic Reviews*. Keele University. Technical Report TR/SE-0401. Department of Computer Science, Keele University, UK.

[125] Joseph T. Klapper. 1960. The effects of mass communication. (1960).

[126] Jan Kleinnijenhuis, Tilo Hartmann, Martin Tanis, and Anita M. J. van Hoof. 2020. Hostile Media Perceptions of Friendly Media Do Reinforce Partisanship. *Communication Research* 47, 2 (2020), 276–298. https://doi.org/10.1177/0093650219836059

[127] Silvia Knobloch-Westerwick and Jingbo Meng. 2009. Looking the Other Way: Selective Exposure to Attitude-Consistent and Counterattitudinal Political Information. *Communication Research* 36, 3 (2009), 426–448. https://doi.org/10.1177/0093650209333030

[128] Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 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* (Cologne Germany). ACM, 1–7. https://doi.org/10.1145/3529372.3530932

[129] Maria Krommyda, Anastasios Rigos, Kostas Bouklas, and Angelos Amditis. 2021. An Experimental Analysis of Data Annotation Methodologies for Emotion Detection in Short Text Posted on Social Media. 8, 1 (2021), 19. https://doi.org/10.3390/informatics8010019

[130] Anne C. Kroon, Damian Trilling, and Tamara Raats. 2021. Guilty by Association: Using Word Embeddings to Measure Ethnic Stereotypes in News Coverage. 98, 2 (2021), 451–477. https://doi.org/10.1177/1077699020932304

[131] Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Akanksha Bansal. 2021. ComMA@ICON: Multilingual Gender Biased and Communal Language Identification Task at ICON-2021. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*. NLP Association of India (NLPAI), NIT Silchar, 1–12. https://aclanthology.org/2021.icon-multigen.1

[132] Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A Systematic Media Frame Analysis of 1.5 Million New York Times Articles from 2000 to 2017. In *12th ACM Conference on Web Science*. ACM, Southampton United Kingdom, 305–314. https://doi.org/10.1145/3394231.3397921

[133] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science* 7 (2021), e644. https://doi.org/10.7717/peerj-cs.644

[134] Konstantina Lazaridou, Alexander Löser, Maria Mestre, and Felix Naumann. 2020. Discovering Biased News Articles Leveraging Multiple Human Annotations. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1268–1277. https://www.aclweb.org/anthology/2020.lrec-1.159

[135] Susan Leavy. 2018. Uncovering gender bias in newspaper coverage of Irish politicians using machine learning. *Digital Scholarship in the Humanities* 34, 1 (06 2018), 48–63. https://doi.org/10.1093/llc/fqy005

[136] Susan Leavy. 2020. Uncovering Gender Bias in Media Coverage of Politicians with Machine Learning. arXiv:2005.07734 [cs] http://arxiv.org/abs/2005.07734

[137] Eun-Ju Lee. 2012. That's Not the Way It Is: How User-Generated Comments on the News Affect Perceived Media Bias. *Journal of Computer-Mediated Communication* 18, 1 (2012), 32–45. https://doi.org/10.1111/j.1083-6101.2012.01597.x

[138] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Mitigating Media Bias through Neutral Article Generation. *CoRR* abs/2104.00336 (2021). https://doi.org/10.48550/arXiv.2104.00336

[139] Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral Multi-News Summarization for Mitigating Framing Bias. https://doi.org/10.48550/arXiv.2204.04902

[140] Tien-Tsung Lee. 2005. The Liberal Media Myth Revisited: An Examination of Factors Influencing Perceptions of Media Bias. *Journal of Broadcasting & Electronic Media* 49, 1 (2005), 43–64. https://doi.org/10.1207/s15506878jobem4901_4

[141] Kristina Lerman, Xiaoran Yan, and Xin Zeng Wu. 2016. The "majority illusion" in social networks. *PLoS ONE* 11, 2 (2016), 1–13. https://doi.org/10.1371/journal.pone.0147617

[142] Beth Levin. 1993. *English verb classes and alternations: a preliminary investigation.* University of Chicago Press, London; Chicago.

[143] Chang Li and Dan Goldwasser. 2019. Encoding Social Information with Graph Convolutional Networks forPolitical Perspective Detection in News Media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Florence, Italy, 2594–2604. https://doi.org/10.18653/v1/P19-1247

[144] Chang Li and Dan Goldwasser. 2021. Using Social and Linguistic Information to Adapt Pretrained Representations for Political Perspective Identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Online). Association for Computational Linguistics, 4569–4579. https://doi.org/10.18653/v1/2021.findings-acl.401

[145] Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. 2018. Understanding Characteristics of Biased Sentences in News Articles. *CIKM Workshops* (2018).

[146] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. A Survey of Transformers. *CoRR* abs/2106.04554 (2021). https://doi.org/10.48550/arXiv.2106.04554

[147] Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021. A Transformer-based Framework for Neutralizing and Reversing the Political Polarity of News Articles. *Proceedings of the ACM on Human-Computer Interaction* 5, 65 (2021), 1–26. Issue CSCW1. https://doi.org/10.1145/3449139

[148] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating Political Bias in Language Models Through Reinforced Calibration. *ArXiv* abs/2104.14795 (2021).

[149] Ruibo Liu, Lili Wang, Chenyan Jia, and Soroush Vosoughi. 2021. Political Depolarization of News Articles Using Attribute-aware Word Embeddings. (2021). https://doi.org/10.48550/ARXIV.2101.01391

[150] Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. (2022). https://doi.org/10.48550/ARXIV.2205.00619

[151] Anne Maass, Daniela Salvi, Luciano Arcuri, and Gun Semin. 1989. Language use in intergroup contexts: The linguistic intergroup bias. *Journal of Personality and Social Psychology* 67, 6 (1989), 981–993. https://doi.org/10.1037/0022-3514.57.6.981

[152] Karthic Madanagopal and James Caverlee. 2021. Towards Ongoing Detection of Linguistic Bias on Wikipedia. In *Companion Proceedings of the Web Conference 2021* (Ljubljana Slovenia). ACM, 629–631. https://doi.org/10.1145/3442442.3452353

[153] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 615–621. https://doi.org/10.18653/v1/N19-1062

[154] Gary Marks and Norman Miller. 1987. Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin* 102, 1 (1987), 72–90. https://doi.org/10.1037//0033-2909.102.1.72

[155] Héctor Martínez Alonso, Amaury Delamaire, and Benoît Sagot. 2017. Annotating omission in statement pairs. In *Proceedings of the 11th Linguistic Annotation Workshop.* Association for Computational Linguistics, Valencia, Spain, 41–45. https://doi.org/10.18653/v1/W17-0805

[156] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. (2020). https://doi.org/10.48550/ARXIV.2012.10289 Publisher: arXiv Version Number: 2.

[157] Jörg Matthes, Johannes Knoll, and Christian von Sikorski. 2018. The "Spiral of Silence" Revisited: A Meta-Analysis on the Relationship Between Perceptions of Opinion Support and Political Opinion Expression. *Communication Research* 45, 1 (2018), 3–33. https://doi.org/10.1177/0093650217745429

[158] Jörg Matthes, Desirée Schmuck, and Christian von Sikorski. 2021. In the Eye of the Beholder: A Case for the Visual Hostile Media Phenomenon. *Communication Research* (2021), 1–25. https://doi.org/10.1177/00936502211018596

[159] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (2018), 861. https://doi.org/10.21105/joss.00861

[160] Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence* 3 (aug 2020), 55. https://doi.org/10.3389/frai.2020.00055

[161] Seong-Jae Min and John C. Feaster. 2010. Missing Children in National News Coverage: Racial and Gender Representations of Missing Children Cases. *Communication Research Reports* 27, 3 (2010), 207–216. https://doi.org/10.1080/08824091003776289

[162] Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral Framing and Ideological Bias of News. Vol. 12467. 206–219. https://doi.org/10.1007/978-3-030-60975-7_16

[163] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE* 15, 8 (08 2020), 1–26. https://doi.org/10.1371/journal.pone.0237861

[164] Sendhil Mullainathan and Andrei Shleifer. 2002. *Media Bias.* Working Paper 9295. National Bureau of Economic Research. https://doi.org/10.3386/w9295

[165] Sean Munson, Stephanie Lee, and Paul Resnick. 2021. Encouraging Reading of Diverse Political Viewpoints with a Browser Widget. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (Aug. 2021), 419–428. https://doi.org/10.1609/icwsm.v7i1.14429

[166] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Online). Association for Computational Linguistics, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

[167] Preslav Nakov, Husrev Taha Sencar, Jisun An, and Haewoon Kwak. 2021. A Survey on Predicting the Factuality and the Bias of News Media. *ArXiv* abs/2103.12506 (2021).

[168] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of Post-OCR Processing Approaches. *ACM Comput. Surv.* 54, 6 (jul 2021), 1–37. https://doi.org/10.1145/3453476

[169] Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

[170] Dimitar Nikolov, Mounia Lalmas, Alessandro Flammini, and Filippo Menczer. 2019. Quantifying biases in online information exposure. *Journal of the Association for Information Science and Technology* 70, 3 (2019), 218–229. https://doi.org/10.1002/asi.24121

[171] Timothy Niven and Hung-Yu Kao. 2020. Measuring Alignment to Authoritarian State Media as Framing Bias. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda* (Barcelona, Spain (Online)). International Committee on Computational Linguistics (ICCL), 11–21. https://aclanthology.org/2020.nlp4if-1.2

[172] Ruchi Ookalkar, Kolli Vishal Reddy, and Eric Gilbert. 2019. Pop: Bursting News Filter Bubbles on Twitter Through Diverse Exposure. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 18–22. https://doi.org/10.1145/3311957.3359513

[173] Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2020. Out of the Echo Chamber: Detecting Countering Debate Speeches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online). Association for Computational Linguistics, 7073–7086. https://doi.org/10.18653/v1/2020.acl-main.633

[174] Kartikey Pant, Tanvi Dadu, and Radhika Mamidi. 2020. Towards Detection of Subjective Bias using Contextualized Word Embeddings. In *Companion Proceedings of the Web Conference 2020* (Taipei Taiwan). ACM, 75–76. https://doi.org/10.1145/3366424.3382704

[175] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona Spain). ACM, 446–457. https://doi.org/10.1145/3351095.3372843

[176] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin, UK.

[177] Eun Hee Park and Veda C. Storey. 2022. Emotion Ontology Studies: A Framework for Expressing Feelings Digitally and Its Application to Sentiment Analysis. *ACM Comput. Surv.* (aug 2022). https://doi.org/10.1145/3555719

[178] Yilang Peng. 2018. Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision. *Journal of Communication* 68, 5 (10 2018), 920–941. https://doi.org/10.1093/joc/jqy041

[179] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count.* Lawerence Erlbaum Associates, Mahwah, NJ.

[180] Gordon Pennycook and David G. Rand. 2021. The Psychology of Fake News. *Trends in Cognitive Sciences* 25, 5 (2021), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

[181] Richard M. Perloff. 2015. A Three-Decade Retrospective on the Hostile Media Effect. *Mass Communication and Society* 18, 6 (2015), 701–729. https://doi.org/10.1080/15205436.2015.1051234

[182] Christopher Piñón. 2001. A Finer Look at the Causative-Inchoative Alternation. *Semantics and Linguistic Theory* 11 (2001), 346–364. https://doi.org/10.3765/salt.v11i0.2858

[183] Kate Power, Lucy Rak, and Marianne Kim. 2019. Women in business media: A Critical Discourse Analysis of Representations of Women in Forbes, Fortune and Bloomberg BusinessWeek, 2015-2017. *Critical Approaches to Discourse Analysis Across Disciplines* 11, 2 (2019).

[184] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically Neutralizing Subjective Bias in Text. (2019). https://doi.org/10.48550/ARXIV.1911.09709

[185] Riccardo Puglisi and James M. Snyder. 2015. Chapter 15 - Empirical Studies of Media Bias. In *Handbook of Media Economics*, Simon P. Anderson, Joel Waldfogel, and David Strömberg (Eds.). Handbook of Media Economics, Vol. 1. North-Holland, 647–667. https://doi.org/10.1016/B978-0-444-63685-0.00015-2 ISSN: 2213-6630.

[186] Prashanth Rao and Maite Taboada. 2021. Gender bias in the news: A scalable topic modelling and visualization framework. *Frontiers in Artificial Intelligence* 4 (2021). https://doi.org/10.3389/frai.2021.664737

[187] Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation Frames: A Data-Driven Investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 311–321. https://doi.org/10.18653/v1/P16-1030

[188] Steve Rathje, Jon Roozenbeek, Cecilie Traberg, Jay Van Bavel, and Sander van der Linden. 2022. Letter to the Editors of Psychological Science: Meta-Analysis Reveals that Accuracy Nudges Have Little to No Effect for U.S. Conservatives: Regarding Pennycook et al. (2020). *Psychological Science* (01 2022). https://doi.org/10.25384/SAGE.12594110.v2

[189] Shaina Raza, Deepak John Reji, and Chen Ding. 2022. Dbias: Detecting Biases and Ensuring Fairness in News Articles. *International Journal of Data Science and Analytics* (Sept. 2022). https://doi.org/10.1007/s41060-022-00359-4

[190] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 1650–1659. https://aclanthology.org/P13-1162

[191] Scott A. Reid. 2012. A Self-Categorization Explanation for the Hostile Media Effect. *Journal of Communication* 62, 3 (2012), 381–399. https://doi.org/10.1111/j.1460-2466.2012.01647.x

[192] Ellen Riloff and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. 105–112. https://aclanthology.org/W03-1014

[193] Axel Rodríguez, Carlos Argueta, and Yi-Ling Chen. 2019. Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 169–174. https://doi.org/10.1109/ICAIIC.2019.8669073

[194] Jon Roozenbeek, Cecilie S. Traberg, and Sander van der Linden. 2022. Technique-based inoculation against real-world misinformation. *Royal Society Open Science* 9, 5 (2022), 211719. https://doi.org/10.1098/rsos.211719

[195] Lee Ross, David Greene, and Pamela House. 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology* 13, 3 (may 1977), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

[196] Qin Ruan, Brian Mac Namee, and Ruihai Dong. 2021. Pseudo-labelling Enhanced Media Bias Detection. (2021). https://doi.org/10.48550/ARXIV.2107.07705

[197] Qin Ruan, Brian Mac Namee, and Ruihai Dong. 2021. Bias Bubbles: Using Semi-Supervised Learning to Measure How Many Biased News Articles Are Around Us. In *The 29th Irish Conference on Artificial Intelligence and Cognitive Science 2021, Dublin, Republic of Ireland, December 9-10, 2021 (CEUR Workshop Proceedings, Vol. 3105)*, Arjun Pakrashi, Ellen Rushe, Mehran Hossein Zadeh Bazargani, and Brian Mac Namee (Eds.). CEUR-WS.org, 153–164. http://ceur-ws.org/Vol-3105/paper40.pdf

[198] Mark Rubin and Constantina Badea. 2007. Why Do People Perceive Ingroup Homogeneity on Ingroup Traits and Outgroup Homogeneity on Outgroup Traits? *Personality and Social Psychology Bulletin* 33, 1 (2007), 31–42. https://doi.org/10.1177/0146167206293190 PMID: 17178928.

[199] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social Media News Communities: Gatekeeping, Coverage, and Statement Bias. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 1679–1684. https://doi.org/10.1145/2505515.2505623

[200] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social Media News Communities: Gatekeeping, Coverage, and Statement Bias. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 1679–1684. https://doi.org/10.1145/2505515.2505623

[201] Nabil F. Saleh, Jon Roozenbeek, FadiADI A. MakkiAKKI, William P. MCClanahan, and Sander van der Linden. 2021. Active inoculation boosts attitudinal resistance against extremist persuasion techniques: a novel approach towards the prevention of violent extremism. *Behavioural Public Policy* (2021), 1–24. https://doi.org/10.1017/bpp.2020.60

[202] Allan Sales, Albin Zehe, Leandro Balby Marinho, Adriano Veloso, Andreas Hotho, and Janna Omeliyanenko. 2021. Assessing Media Bias in Cross-Linguistic and Cross-National Populations. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, Ceren Budak, Meeyoung Cha, Daniele Quercia, and Lexing Xie (Eds.). AAAI Press, 561–572. https://doi.org/10.1609/icwsm.v15i1.18084

[203] Eitan Sapiro-Gheiler. 2019. Examining Political Trustworthiness through Text-Based Measures of Ideology. 33 (2019), 10029–10030. https://doi.org/10.1609/aaai.v33i01.330110029

[204] Kathleen M. Schmitt, Albert C. Gunther, and Janice L. Liebhart. 2004. Why Partisans See Mass Media as Biased. *Communication Research* 31, 6 (2004), 623–641. https://doi.org/10.1177/0093650204269390

[205] Anne Schulz, Werner Wirth, and Philipp Müller. 2020. We Are the People and You Are Fake News: A Social Identity Approach to Populist Citizens' False Consensus and Hostile Media Perceptions. *Communication Research* 47, 2 (2020), 201–226. https://doi.org/10.1177/0093650218794854

[206] Gün R. Semin and Klaus Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology* 54, 4 (1988), 558–568. https://doi.org/10.1037/0022-3514.54.4.558

[207] Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events* (Online). Association for Computational Linguistics, 120–125. https://doi.org/10.18653/v1/2020.nuse-1.15

[208] Jay Kachhadia Ania Korsunska Shloak Gupta, Sarah Bolden and Jennifer Stromer-Galley. 2020. PoliBERT: Classifying political social media messages with BERT. *2020 International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation Working Papers* 13 (2020). https://doi.org/10.1007/978-3-030-61255-9

[209] Manjira Sinha and Tirthankar Dasgupta. 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* (Virtual Event Queensland Australia). ACM, 3418–3422. https://doi.org/10.1145/3459637.3482084

[210] Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022-05-03. Political Ideology and Polarization of Policy Positions: A Multi-dimensional Approach. https://doi.org/10.48550/arXiv.2106.14387

[211] Steven M. Smith, Leandre R. Fabrigar, and Meghan E. Norris. 2008. Reflecting on Six Decades of Selective Exposure Research: Progress, Challenges, and Opportunities. *Social and Personality Psychology Compass* 2, 1 (2008), 464–493. https://doi.org/10.1111/j.1751-9004.2007.00060.x

[212] Timo Spinde. 2021. An Interdisciplinary Approach for the Automated Detection and Visualization of Media Bias in News Articles. In *2021 International Conference on Data Mining Workshops (ICDMW)*. 1096–1103. https://doi.org/10.1109/ICDMW53433.2021.00144

[213] Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. 2020. Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (2020-01-01) *(JCDL '20)*. Association for Computing Machinery, Virtual Event, China, 389–392. https://doi.org/10.1145/3383583.3398619

[214] Timo Spinde, Felix Hamborg, and Bela Gipp. 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* (2020-01-01) *(JCDL '20)*. Association for Computing Machinery, Virtual Event, China, 505–506. https://doi.org/10.1145/3383583.3398585

[215] Timo Spinde, Felix Hamborg, and Bela Gipp. 2020. Media Bias in German News Articles : A Combined Approach. In *Proceedings of the 8th International Workshop on News Recommendation and Analytics (INRA 2020)*. Virtual event. https://doi.org/10.1007/978-3-030-65965-3_41

[216] Timo Spinde, Christin Jeggle, Magdalena Haupt, Wolfgang Gaissmaier, and Helge Giese. 2022. How do we raise media bias awareness effectively? Effects of visualizations to communicate bias. *PLOS ONE* 17, 4 (2022), 1–14. https://doi.org/10.1371/journal.pone.0266204

[217] Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021. Do You Think It's Biased? How To Ask For The Perception Of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2021-09-01). https://doi.org/10.1109/JCDL52503.2021.00018

[218] Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. 2022. Exploiting Transformer-Based Multitask Learning for the Detection of Media Bias in News Articles. In *Information for a Better World: Shaping the Global Future. iConference 2022.*, Malte Smits (Ed.). Vol. 13192. Springer International Publishing, 225–235. https://doi.org/10.1007/978-3-030-96957-8_20 Series Title: Lecture Notes in Computer Science.

[219] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 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. https://doi.org/10.18653/v1/2021.findings-emnlp.101

[220] Timo Spinde, Lada Rudnitckaia, Felix Hamborg, and Bela Gipp. [n.d.]. Identification of Biased Terms in News Articles by Comparison of Outlet-specific Word Embeddings. In *Proceedings of the 16th International Conference (iConference 2021)* (Beijing, China (Virtual Event), 2021-03-01). https://doi.org/10.1007/978-3-030-71305-8_17

[221] Timo Spinde, Lada Rudnitckaia, Sinha Kanishka, Felix Hamborg, Bela Gipp, and Karsten Donnay. 2021. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics. In *Proceedings of the iConference 2021*. Beijing, China (Virtual Event). https://doi.org/10.6084/m9.figshare.17192924

[222] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management* 58, 3 (2021), 102505. https://doi.org/10.1016/j.ipm.2021.102505

[223] Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the Topical Stance and Political Leaning of Media using Tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online). Association for Computational Linguistics, 527–537. https://doi.org/10.18653/v1/2020.acl-main.50

[224] Cass R. Sunstein. 2009. *Going to extremes: How like minds unite and divide.* Oxford University Press.

[225] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the Backfire Effect: Measurement and Design Considerations. *Journal of Applied Research in Memory and Cognition* 9, 3 (2020), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

[226] Henri Tajfel, John C. Turner, William G. Austin, and Stephen Worchel. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader* 56, 65 (1979), 9780203505984–16.

[227] Edson C. Tandoc Jr. 2019. The facts of fake news: A research review. *Sociology Compass* 13, 9 (2019), e12724. https://doi.org/10.1111/soc4.12724

[228] Minh Tran. 2020. How biased are American media outlets? A framework for presentation bias regression. In *2020 IEEE International Conference on Big Data (Big Data)* (Atlanta, GA, USA). IEEE, 4359–4364. https://doi.org/10.1109/BigData50022.2020.9377987

[229] Melissa Tully, Emily K Vraga, and Anne-Bennett Smithson. 2020. News media literacy, perceptions of bias, and interpretation of news. *Journalism* 21, 2 (2020), 209–226. https://doi.org/10.1177/1464884918805262

[230] John C. Turner. 1991. *Social influence.* Thomson Brooks/Cole Publishing Co.

[231] Robert P. Vallone, Lee Ross, and Mark R. Lepper. 1985. The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology* 49, 3 (1985), 557–585. https://doi.org/10.1037/0022-3514.49.3.577

[232] Esther van den Berg and Katja Markert. 2020. Context in Informational Bias Detection. In *Proceedings of the 28th International Conference on Computational Linguistics* (Barcelona, Spain (Online)). International Committee on Computational Linguistics, 6315–6326. https://doi.org/10.18653/v1/2020.coling-main.556

[233] Sander van der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine* 28, 3 (2022), 460–467. https://doi.org/10.1038/s41591-022-01713-6

[234] Teun A. van Dijk. 2007. Chapter 12 Discourse and Racism.

[235] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

[236] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12388–12401. https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf

[237] Giacomo Villa, Gabriella Pasi, and Marco Viviani. 2021. Echo chamber detection and analysis: A topology- and content-based approach in the COVID-19 scenario. 11, 1 (2021), 78. https://doi.org/10.1007/s13278-021-00779-3

[238] Jun Wang and Bei Yu. 2021. News2PubMed: A Browser Extension for Linking Health News to Medical Literature. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2605–2609. https://doi.org/10.1145/3404835.3462788

[239] Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. 2020. Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling Denotation and Connotation in Vector Spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online). Association for Computational Linguistics, 4090–4105. https://doi.org/10.18653/v1/2020.emnlp-main.335

[240] Brian E. Weeks, Daniel S. Lane, Dam Hee Kim, Slgi S. Lee, and Nojin Kwak. 2017. Incidental Exposure, Selective Exposure, and Political Information Sharing: Integrating Online Exposure Patterns and Expression on Social Media. *Journal of Computer-Mediated Communication* 22, 6 (11 2017), 363–379. https://doi.org/10.1111/jcc4.12199

[241] Melvin Wevers. 2019. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (Florence, Italy). Association for Computational Linguistics, 92–97. https://doi.org/10.18653/v1/W19-4712

[242] David Manning White. 1950. The "Gate Keeper": A Case Study in the Selection of News. *Journalism Quarterly* 27, 4 (1950), 383–390. https://doi.org/10.1177/107769905002700403

[243] Alden Williams. 1975. Unbiased Study of Television News Bias.

[244] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (Vancouver, British Columbia, Canada) *(HLT '05)*. Association for Computational Linguistics, USA, 347–354. https://doi.org/10.3115/1220575.1220619

[245] Stephane Wolton. 2017. *Are Biased Media Bad for Democracy?* Asymmetric & Private Information eJournal, Microeconomics.

[246] Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. 2020. TIMME: Twitter Ideology-detection via Multi-task Multi-relational Embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event CA USA). ACM, 2258–2268. https://doi.org/10.1145/3394486.3403275

[247] Chenyan Xiong, Zhengzhong Liu, Jamie Callan, and Tie-Yan Liu. 2018. Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling. In *The 41st International ACM SIGIR Conference on Research &; Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 575–584. https://doi.org/10.1145/3209978.3209982

[248] Anam Zahid, Maham Nasir Khan, Ahmer Latif Khan, Faisal Kamiran, and Bilal Nasir. 2020. Modeling, Quantifying and Visualizing Media Bias on Twitter. 8 (2020), 81812–81821. https://doi.org/10.1109/ACCESS.2020.2990800

[249] Xueying Zhang and Mei-Chen Lin. 2021. The Effects of Social Identities and Issue Involvement on Perceptions of Media Bias Against Gun Owners and Intention to Participate in Discursive Activities: In the Context of the Media Coverage of Mass Shootings. *Mass Communication and Society* 0, 0 (2021), 1–22. https://doi.org/10.1080/15205436.2021.1916036

[250] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2021. A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances. *ACM Comput. Surv.* 54, 2, Article 27 (mar 2021), 36 pages. https://doi.org/10.1145/3433000

[251] Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden Biases in Unreliable News Detection Datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Online). Association for Computational Linguistics, 2482–2492. https://doi.org/10.18653/v1/2021.eacl-main.211

[252] Yiyi Zhou, Rongrong Ji, Jinsong Su, and Jiaquan Yao. 2021. Uncovering Media Bias via Social Network Learning. 12, 1 (2021), 1–12. https://doi.org/10.1145/3422181

[253] Frederik J. Zuiderveen Borgesius, Damian Trilling, Judith Möller, Bodó Balázs, Claes H. De Vreese, and Natali Helberger. 2016. Should we worry about filter bubbles? *Internet Policy Review* 5, 1 (2016), 1–16. https://doi.org/10.14763/2016.1.401

GLOSSARY

AttnBL attention-based bidirectional long short-term memory. 16

BiRNN bidirectional recurrent neural network. 17

CNN convolutional neural network. 16

FMP Friendly Media Phenomenon. 21, 22

GCN graph convolutional network. 14
GRU gated recurrent unit. 16

HAN hierarchical attention network. 16
HMP hostile media phenomenon. 21–23

KNN k-nearest neighbors. 13

LDA Latent Dirichlet Allocation. 16, 17
LR logistic regression. 17
LSTM long short-term memory. 14–16

ML Machine Learning. 2, 11, 13, 14, 16–18
MLP multilayer perceptron. 17

NB Naive Bayes. 12, 13, 16, 17
NN neural network. 13, 16, 18, 24
nNN non-neural networks. 3, 11, 12, 14, 17
ntbML non-transformer-based machine learning. 2, 11, 12, 14–16

RF random forest. 13, 17
RNN recurrent neural network. 16

SVM support vector machine. 12, 13, 16, 17

tbML transformer-based machine learning. 2, 11, 12, 14, 17
tNLP traditional natural language processing. 2, 11–14

UMAP Uniform manifold approximation and projection for dimension reduction. 16

VADER valence aware dictionary for sentiment reasoning. 17

XGBoost extreme gradient boosting. 17

A APPENDICES

A.1 Publications per Bias Category

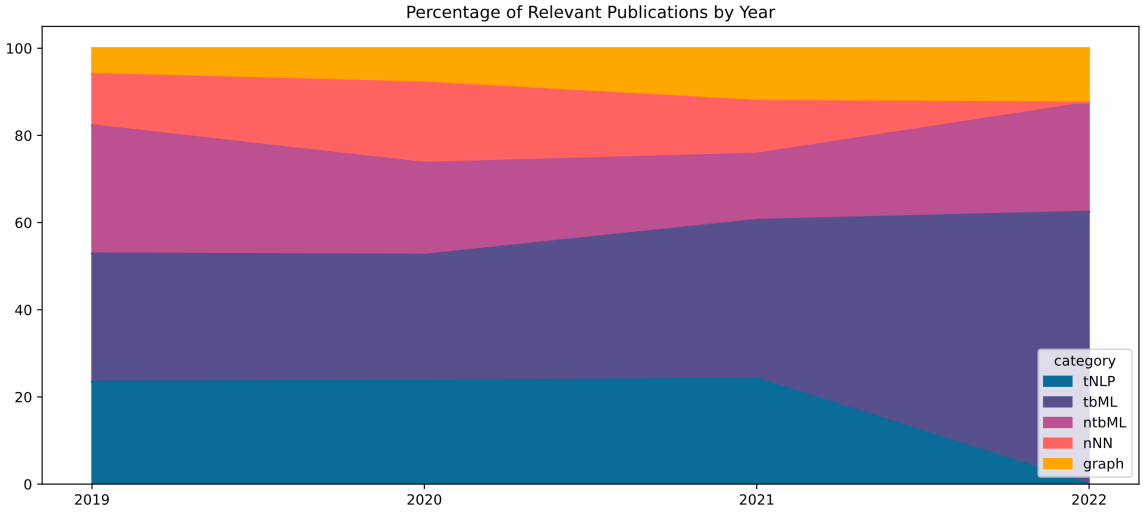

Fig. 5. Overview of relevant publications per category and year. Publications can appear in multiple categories.

A.2 Definitions and Examples of Media Bias

Linguistic Bias Overview

Subcategory	Features	Definition	Source	Example
Framing Bias		Subjective words or phrases linked with a certain point of view.	[190]	Usually, smaller cottage-style houses have been demolished to make way for these ***McMansions***.
	Subjectivity Clues	Subjectivity clues are words that, in most (*strongsubj*) or certain (*weaksubj*) contexts, are considered to add subjective meaning to a statement.	[192, 244]	Subjective adjectives (e.g., preposterous, unseemly) and metaphorical or idiomatic phrases (e.g., dealt a blow, swept off one's feet).
	One-sided Terms	One-sided terms only reflect one side of a controversial subject where the same event can be seen from two or more opposing perspectives.	[190, p. 4]	"***Pro-life***" vs. "anti-abortion".
Epistemological Bias		Linguistic features that subtly focus on the believably of a position.	[190]	Kuypers ***claimed*** that the mainstream press in America tends to favor liberal viewpoints.

Subcategory	Features	Definition	Source	Example
	Factive Verbs	Factive verbs presuppose the truth of a complement clause.	[106], [119, p. 5]	"John ***realized*** that he had no money" implies "John had no money".
	Entailments	Entailments are directional relations where one statement implies the truth of another.	[22, p. 1], [119]	"Cyprus ***was invaded by*** the Ottoman Empire in 1571" implies the hypothesis "The ***Ottomans attacked*** Cyprus".
	Assertive Verbs	Assertives imply that the speaker or subject of a sentence has an affirmative opinion on the truth value of a complements proposition.	[106]	E.g., regret, resent, forget, amuse, suffice, bother.
	Hedge Words	Hedge words are used to lessen the probability and the writer's guarantee of the truthfulness of a statement.	[112]	E.g., perhaps, might, may.
Bias by Semantic Properties		A set of semantic properties which are grammatically relevant and serve as an interface between syntax and semantics.	[90, p. 3]	"The gunmen shot the opposition leader" vs. "The shooting killed the opposition leader".
	Causative-Inchoative Alternations	Causative-inchoative alternations are verbs that can either emphasize a change of state or the bringing about of the change of state.	[90, 142], [182, p. 1]	"Rebecca broke the pencil" vs. "The pencil broke".
Connotation Bias		Bias through the use of different words for the same subject with identical denotations but differing (often biased) connotations.	[35, 239]	E.g. "immigrants" vs. "***aliens***", "estate tax" vs. "***death tax***".
	Verb Connotations	Verbs with identical denotation but differing connotations.	[187]	The story begins in Illinois in 1987 when a 17- year-old girl ***suffered*** a botched abortion.
	Noun and Adjective Connotations	Nouns and adjectives with identical denotation but differing connotations.	[4]	relentless (-) vs. persistent (+) gentleman (+) vs. man (=) protection (-) vs. security (=)

Subcategory	Features	Definition	Source	Example
Linguistic Intergroup Bias		Variations in language lead to the creation or maintenance of the reputations of certain social groups.	[58, p. 2]	"Mary hit Mike" vs. "Mary hurt Mike" vs. "Mary hates Mike" vs. "Mary is aggressive".

Table 3. Linguistic Bias Overview

Context Bias (Text Level)

Subcategory	Definition	Source	Example
Spin Bias	Emerges through a newspaper's attempt to create a memorable story by means of simplification or exaggeration. In the case of simplification, newspapers discard some information, whereas exaggeration shapes the story in a certain direction. Both methods try to present the news in a way to attract potential readers.	[5, 164]	President Donald Trump ***gloated*** over mass layoffs at multiple news outlets on Saturday.
Statement Bias	Refers to the process of how facts are reported. This can be by means of expressing own opinions, criticizing the counterpart, or advocating the own viewpoints in a way that news reporting is favorable or unfavorable towards a certain standpoint.	[47]	A political protest in which people sat in the middle of a street blocking traffic can be described as "peaceful" and "productive," or others may describe it as "aggressive" and "disruptive".
Omission Bias / Commission Bias	Omission or Commission Bias results from the journalists deciding what (substantial) information to exclude and what information to include in the news report.	[98, 155]	CNN previously reported on the FBI's hate crime statistics released last November, which showed the number of hate crimes reported to the bureau rose by about 17% in 2017 compared to 2016. 2017 is the latest year for which those statistics are available. It was the third-straight year that hate crime incidents rose.[33]

[33]Example from https://www.allsides.com/media-bias/how-to-spot-types-of-media-bias

Subcategory	Definition	Source	Example
Informational Bias	A statement containing Informational Bias influences the readers' opinion by adding information that is not necessarily relevant to the event or is rather speculative.	[70]	Looking at two articles that report on the same event, the Huffington Post and Fox News each frame entities of opposing stances negatively. HPO states an assumed future action of Donald Trump as a fact, and FOX implies Democrats are taking advantage of political turmoil.
Phrasing Bias	Phrasing Bias results from using words in a sentence or statement which are inflammatory or partial.	[110, p. 2]	"Aborting the fetus" vs. "***Killing the baby***".

Table 4. Context Bias (Text Level)

Context Bias (Reporting Level)

Subcategory	Definition	Source	Example
Coverage Bias	Coverage bias is concerned with the different amounts of coverage diverse sides of the same issue receive.	[47]	A candidate of a political party receiving more coverage than the candidate of the other party.
Selection / Gatekeeping Bias	Selection bias occurs because individuals in the news media are biased in their selection of events to report on.	[47]	During the Roosevelt administration, the entire press was hostile to it in some parts of the country, and it was impossible to get the other side of the story.

Table 5. Context Bias (Reporting Level)

Cognitive / Perception Bias

Subcategory	Definition	Source	Example
Political Ideology	Political Ideology determines the likelihood of a news consumer perceiving media as biased and influences what news sources are viewed as biased.	[87, 140]	Viewers of Fox News held a distinct set of attitudes towards President George W. Bush (i.e., more favorable) as to his political opponents (i.e., less favorable).
Hostile Media Phenomenon	The tendency of partisans to view media coverage as negative towards their own positions. Partisans also fear that nonpartisans could be swayed into a hostile direction by the media coverage.	[50, 93, 94, 191, 204, 222, 231]	Both Pro-Arab and pro-Israeli subjects who saw US news programs on the Israeli move into West Beirut perceived it as biased in favor of the other side.

Subcategory	Definition	Source	Example
Interpersonal Factors	Interpersonal Factors (e.g., ideological similarity) are related to the perception of media bias.	[34, 67]	Conversations with ideologically like-minded others increase the readers' likelihood of perceiving media as biased.
Article Comments	A news consumer being exposed to other users' comments may influence the readers' perception of this article being biased. Research shows that, specifically, those individuals who are confronted with incongruent opinions are likely to perceive stronger levels of media bias.	[83, 107, 137]	Readers who saw a neutral Facebook post of an article on abortion with comments in line with their opinion perceived the story, writer, and outlet as less biased against their opinion.
Exposure to Opposing Views	Constant exposure to messages with opposing views from politicians and opinion leaders leads media consumers to perceive more bias in news media.	[13]	Republicans who followed Twitter bots that retweeted liberal content exhibited more conservative views after one month of exposure.
Topic Identification / Involvedness	High involvement or identification with a topic or an issue is a reason consumers perceive media to be biased.	[87, 249]	Persons who strongly identified with the republican party and were pro-gun ownership perceived media bias in mass shooting news against gun owners to a higher degree.
Outlet Reputation	The reputation of media outlets' ideological orientation can lead consumers to perceive bias in balanced news.	[20]	The perception of bias in the same news article depended on the brand or outlet name (e.g., CNN or FOX) study participants saw.

Table 6. Cognitive / Perception Bias