Uncovering Media Bias in Tweets

Silke Husse, 729354

MA Seminar Web Data Collection with R

Department of Politics and Public Administration

University of Konstanz

15.04.2021

Abstract

Media bias analysis for news and information distributed via social media is an upcoming research field as its importance increases. This report uncovers the extent of media bias in the online appearance of German politicians. In detail, URLs cited in tweets from members of the Bundestag are analyzed and captured into an individual bias scoring. Overall, media bias is found to be present in the external citation usage of tweets and matches the political affiliation of the examined politicians. Twitter experiences an increased use as a communication canal. Hence, this work raises awareness of media bias presence on social media platforms and contributes to the growing research.

1 Introduction

The problem of media bias in traditional news is well studied but remains untouched for the rising communication via social media. More information becomes available online, yet readers rely on a small subset of news mediums due to an overwhelming number of available sources or solely habit [13]. Simultaneously, news mediums adapt their content to their primary users' political preference [3]. Paired with the inherent structural problem of media bias in the news production process, skewed or incomplete perception of information is widespread [18, 13].

Potential approaches to solving the media bias problem are located at different points in the news production process [18]. In the production stage itself, efforts

are made to avoid media bias directly. The post-production stage deals with the created media bias via bias diagnosis, measurement, correction, or mitigation. This report follows a media bias diagnosis approach and aims to raise awareness of media bias presence on social media platforms. For this purpose, tweets from members of the German Bundestag are collected and analyzed regarding their external citation usage in the form of Uniform Resource Locators (URLs). Within this work, I uncover the extent of media bias in the online appearance of German politicians. Thereby, I focus on Twitter as an actively used social media platform and communication canal. Identifying biased news coverage on Twitter via external citation usage is novel. The respective GitHub repository can be found at `https://github.com/SilkeHusse/WebScraping`.

2 Background

Media plays an important role in the individual and social decision-making process [10, 18]. Thus, news mediums have a substantial influence on society. On the downside, media is possibly manipulated and hence biased, meaning that news disguisedly incorporates political positions, characterizations, or terminology [8]. Although media bias lacks an universal definition, it often comprises three components: 1) selection, 2) coverage, and 3) statement bias [3, 10]. Selection bias, also known as gate-keeping, is the selective coverage of topics, whereas coverage bias is the disproportionate devotion of space and time to each side. Lastly, statement bias is the preferential treatment of one side in a dispute, and prevalent in the news [5]. More general, media bias is the deviation in the news' content and presentation [13], exemplary demonstrated on a sentence level in an article from `https://www.allsides.com` on Donald Trump's question "if disinfectant or sunlight could be used as a direct treatment for the COVID-19 coronavirus" [1].

The ordinary reader encounters difficulties building a balanced comprehension of a news issue as an event's perception varies significantly depending on the consumed news article [13, 19]. Thus, biased media distorts a reader's understanding of critical public affairs and eventually leads to decreased news use overall [3, 9, 18]. Growing distrust towards conventional media adds to the apparent crisis in traditional media use and gives rise to citizen journalists [3]. They are actively engaged in the information production process as well as part of collaborations by, e.g., commenting on posts of others, linking, or retweeting. Both professional and citizen journalists act and cooperate primarily on social media platforms, with the resulting personal interactions leading to lower levels of perceived media bias [11]. Here, Twitter is the

leading social media platform and the subject of several current studies on political communication. Politicians and citizens are adopting this network to broadcast political messages and engage in public debates [4].

Citations in Twitter can either be internal or external [21]. The former includes retweets within the social media platform, whereas the latter involves links contained in tweets. In this report, the focus is on external citations in the form of URLs referencing external websites in tweets. Politicians often state their perspectives on an issue in tweets and are the top group of actors journalists cite and turn to as sources [4]. Further, the number of followers certainly influences the number of tweets, and in turn, the number of references to external websites [17]. The subsequent section describes the acquisition of such data in detail.

3 Web Data Collection

The whole project is divided into three parts, whereby each component requires a different strategy of web scraping. First, media bias placements of various news mediums are collected from a static website using XPath queries. Further, information about all current German Bundestag members is gathered, including their corresponding political affiliations and Twitter display names. For this second part, RSelenium is used to extract data from a dynamic webpage as well as the R package legislatoR [12]. Lastly, relevant tweets are pulled via the Twitter API. In general, all web data collection components are implemented in R and can be run independently.

3.1 Media Bias Placement

Several websites issue media bias placements of various news mediums [2, 14, 16]. Given the limitation to German politicians, solely data from the website `https://mediabiasfactcheck.com` is considered as it includes media bias placements of German news mediums. They claim to be the most comprehensive online resource of media bias and provide positionings of over 3,600 media sources [16]. These bias placements determine the respective statement bias, also known as editorial bias, and are based on the news medium's perspective on general philosophy, abortion, economic policy, education policy, environmental policy, gay rights, gun rights, health care, immigration, military, personal responsibility, regulation, social views, taxes, voter ID, and worker's or business rights [16].

The overall goal is to collect the news medium's name, website URL, and respective bias placement. The latter is categorized into left, left-center, least, right-center,

and right biased. The main website displays different tabs for each bias which subsequently leads to a sub website with the bias placement appended as a path to the base URL, e.g., `https://mediabiasfactcheck.com/leftcenter/`. Each sub website exhibits a list containing the news medium's name and sometimes its corresponding website URL within brackets. As all five websites considered for data collection are static, each sub website is downloaded, parsed, and queried using XPath. Manual inspection shows that `https://mediabiasfactcheck.com/right-center/` differs in its HTML structure and thus requires a different XPath query to extract names and corresponding website URLs. For news mediums with missing website URL, the respective hyper reference to another sub website, e.g., `https://mediabiasfactcheck.com/radio-com/`, is saved, downloaded, parsed, and queried for the website URL using XPath. Exception handling is included to catch HTTP errors, and various XPath queries account for different HTML structures. After data cleaning, a data frame with columns "name", "url", and "bias" is saved for further processing. In general, `https://mediabiasfactcheck.com` and its sub websites incorporate many inconsistencies, for which reason web data collection has to be robust and account for irregularities.

3.2 Information about Politicians

The second project component comprises two parts and covers specific information about politicians. First, the political affiliation of each current member of the German Bundestag is collected. Second, their corresponding Twitter display names are fetched. Overall, data gathered by the procedures outlined in the following subsections connects the beforehand described media bias data with the subsequently presented Twitter data.

Political Affiliation The current composition of the German Bundestag can be found on the website `https://www.bundestag.de/abgeordnete/biografien` exhibiting all members, including their political affiliation. At the first call of the webpage, only the first 12 politicians, ordered alphabetically by their last name, are displayed. On the right side, a small arrowhead button indicates a sliding window in which the successive 12 politicians are presented as soon as it is clicked on. Inspecting the HTML structure directly online confirms the website's dynamic form and thus the suitability of using RSelenium for this data collection. In total, there are 709 seats in the Bundestag [7], with 12 politicians introduced per page resulting in at least 60 button clicks. Additionally, multiple politicians who either resigned, deceased, or declined the mandate are still presented on the webpage [6]. As mentioned beforehand, the main goal is to collect the politician's name and corresponding polit-

ical party. It is accomplished by extracting the corresponding web element for each sliding window excerpt, followed by an automated click on the arrowhead button. All data is saved in a data frame with columns "name" and "political_party".

Twitter Display Name To proceed with the last project component, a mapping of the politician's name to the corresponding Twitter display name is required. The legislatoR package provides a "social" table containing social media handles and IDs of numerous politicians, particularly respective Twitter display names [12]. The "social" table is joined with the "core" table using "wikidataid" as the key. The latter table includes basic information about politicians, e.g., full name. In turn, a politician's name acts as the key in the join of all the social media information provided by the legislatoR package with the data collected from the German Bundestag webpage. 97 Twitter display names are added manually. After data cleaning, a data frame with columns "name", "political_party", "pageid", "wikidataid", "sex", "ethnicity", and "twitter" is saved for further processing.

3.3 Tweets

Tweets are easily accessible via the official Twitter API [15]. Using OAuth1.0 authentication and the GET statuses/user_timeline endpoint, the 1,000 most recent tweets from each current member of the German Bundestag, who uses Twitter as a communication canal, are retrieved. Exception handling is included to catch errors thrown by not authorized or deleted accounts as well as accounts with no posts. In line with the crawl delay and rate limit restrictions, all requests per politician are combined into a single data frame with columns "id_str", "full_text", "url", and "retweet" and in turn merged into a list with the politicians' names as titles. Besides, a column "url", comprising in each cell a list of cited URLs, is attached to the data frame described in the former subsection. The subsequent section about data exploration and analysis focuses on the external citation use in tweets via URLs.

4 Data Exploration and Analysis

The collected data is explored and analyzed in the following. Non-affiliated politicians are left out in the analysis, leaving the political parties Die Linke, SPD, Bündnis 90/Die Grünen, CDU/CSU, FDP, and AfD. As a convention, referencing the German Bundestag specifies the total number of politicians presented on the official website instead of the actual number of 709 current members.

4.1 Twitter Data Analysis

At first, all relevant information about the Twitter data is presented, e.g., Twitter account usage, Twitter account activity, and URL citation usage.

Twitter Account Usage Only 72.67% of the German Bundestag members are using Twitter as a communication canal. Broken down onto a political party level, 56.69% of CDU/CSU, 68.90% of SPD, 82.95% of AfD, 84.29% of Die Linke, 88.73% Bündnis 90/Die Grünen, and 94.12% of FDP members own a Twitter account. The number of AfD politicians having a Twitter account is higher, but multiple accounts are blocked and thus have been excluded from this analysis. Fig. 1 displays the distribution of the Bundestag taking into account all politicians (l) or only members owning a Twitter account (r). Compared to the regular composition shown on the left, the proportions across political parties are more balanced in the right pie chart. CDU/CSU and SPD are underrepresented in the Twitter data. The most significant loss and gain in representation are evident for CDU/CSU and FDP, respectively.

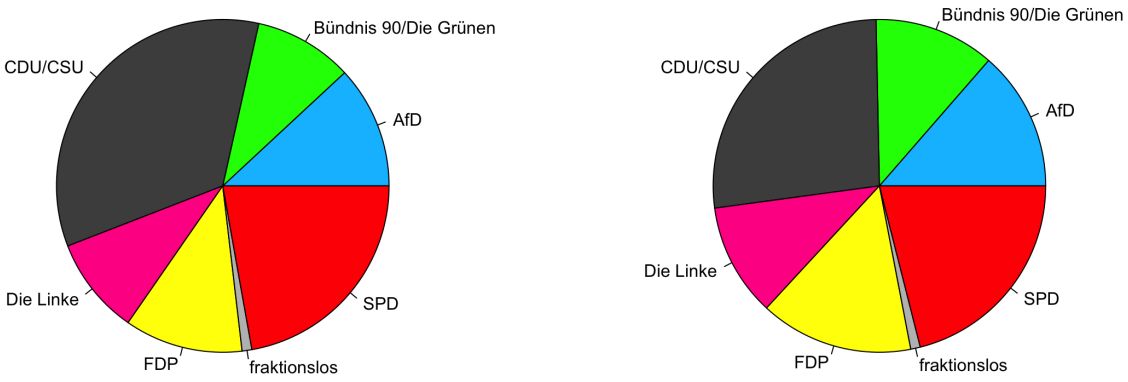

Figure 1: Bundestag composition according all data (l) and Twitter data (r).

Activity on Twitter On average, 751 tweets are gathered per politician. Ideally, this number should be 996, considering a duplicate removal step in the data cleaning of all 1,000 saved tweets. It indicates little activity on Twitter of some members of the Bundestag. Fig. 2 (l) exhibits the average number of gathered tweets per politician, separated by political affiliation. Bündnis 90/Die Grünen (932) and Die Linke (920) demonstrate apparent activity indicating that almost all respective party members tweet regularly. The least activity is shown by CDU/CSU politicians (625).

External Citation Usage The number of followers certainly influences the number of tweets [17]. The resulting number of references to external websites in the form of URLs follows this supposition irregularly. On a political party level, the

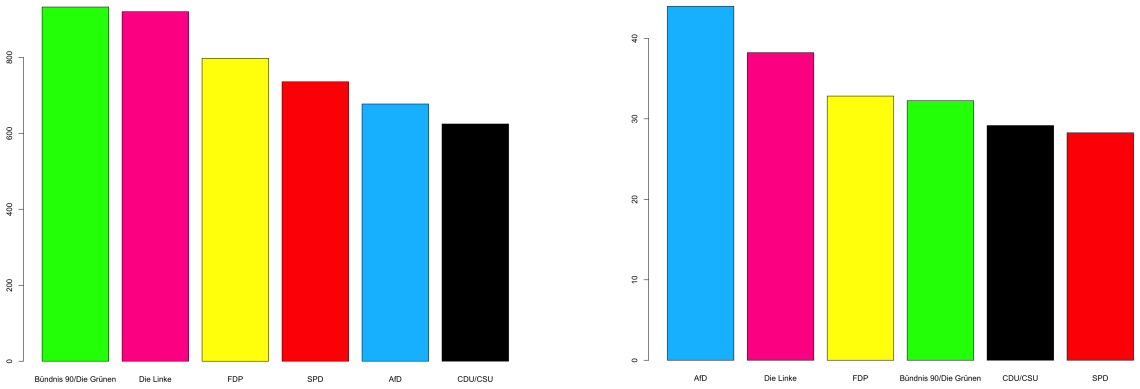

Figure 2: Average number of tweets collected per politician (l) and average external citation usage per politician (r), both separated by political affiliation.

average AfD politician incorporates at least one URL citation in 43.99% of the corresponding tweets. CDU/CSU (29.17%) and SPD (28.27%) members rely on external website referencing the least. Fig. 2 (r) displays the use of URL citations in tweets, separated by political party, in percentage. Mainly, it follows the order demonstrated in Fig. 2 (l), only significantly differing for Bündnis 90/Die Grünen and AfD. Generally, the average politician includes 248 URL citations in 996 tweets.

4.2 Media Bias Analysis

Given the collected data, an individual bias scoring for each politician is computed by depicting media bias placements of news mediums to a numerical scale and then averaging over the correspondingly cited URLs. In detail, left, left-center, center, right-center, and right positionings are mapped to -1, -0.5, 0, 0.5, and 1, respectively. Fig. 3 displays the distribution of media bias scores for the whole Bundestag. It is approximately distributed equally across the spectrum, although extremes are absent. The lack of media bias scores greater than 0.5 indicates that no politician cited a right biased news medium. The variances of the individual media bias scores are small, suggesting the reliability of the media bias score and neither always 0, ensuring diversity within the URL citation sources.

Fig. 4 demonstrates the same concept broken down onto a political affiliation level. Generally, the distribution of media bias scores corresponds to the political direction of a party. The contribution is twofold: 1) media bias placements gathered from `https://mediabiasfactcheck.com` are verified, and 2) media bias presence in tweets is confirmed. Bündnis 90/Die Grünen are comprehensively located on the left side of the numerical scale compared to SPD, contrasting the order presented on the website of the Bundestag. SPD politicians are more broadly positioned, even

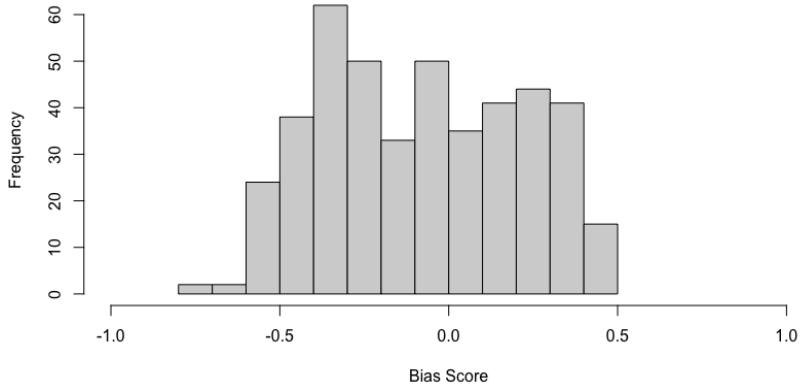

Figure 3: Media bias scores given all politicians.

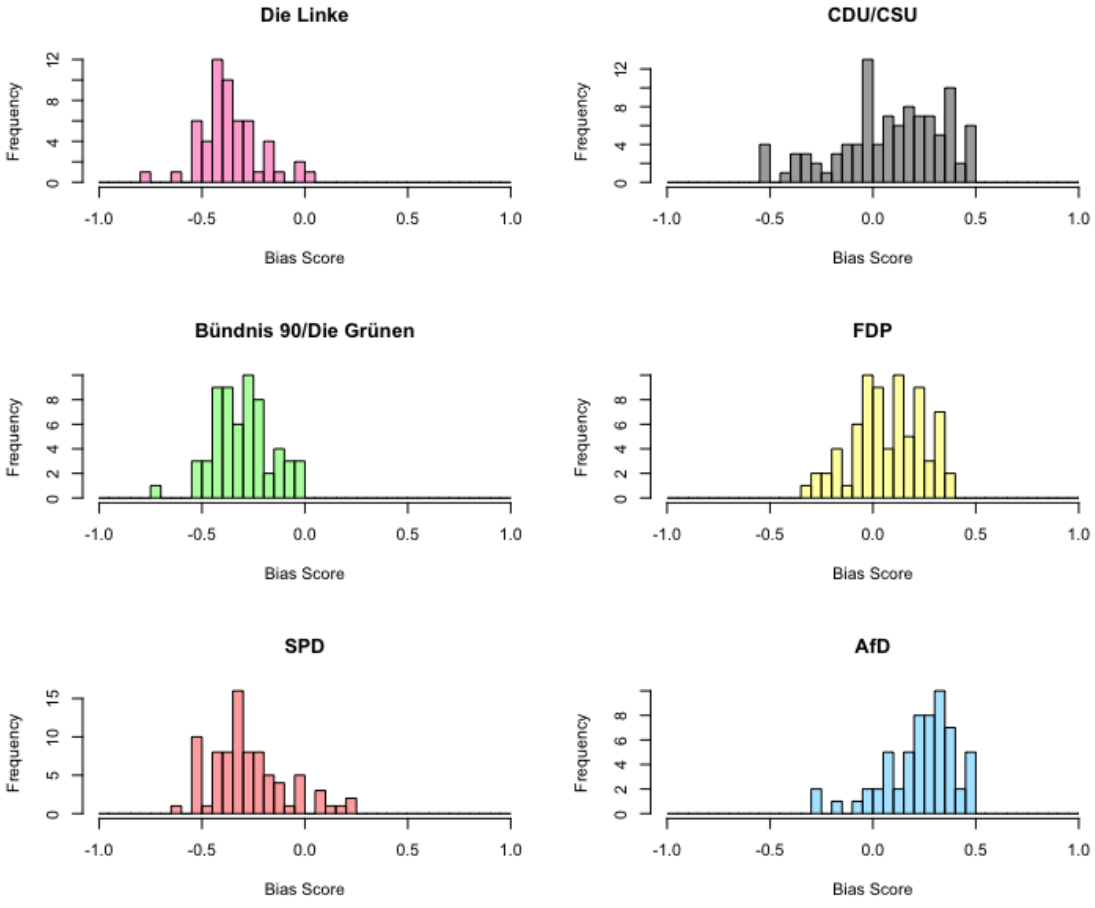

Figure 4: Media bias scores per political party.

having positive media bias scores. CDU/CSU displays a wide range of media bias scores, approximately ranging from -0.5 to 0.5. The same holds for FDP, with a slight tendency towards positive media bias scores as well. AfD approximately resembles a mirrored distribution of SPD. Fig. 5 demonstrates the results further by comparing Die Linke with AfD directly. Both party members primarily do not

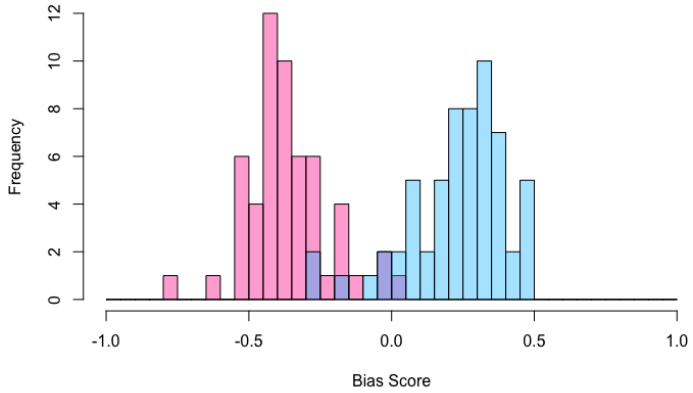

Figure 5: Comparison between Die Linke and AfD.

cite center biased news mediums resulting in a limited overlap between the political parties. This corresponds to the actual political directions and opinions. Fig. 6 exhibits the average media bias score for each political party, endorsing the beforehand described results. Interestingly, FDP has a lower average media bias score than CDU/CSU, again contrasting the political party order presented on the website of the Bundestag. Additionally, it stands out that CDU/CSU and SPD demonstrate an extensive range of media bias scores compared to all other political parties. Besides, outliers are displayed and analyzed in the following. Jörg Cezanne, belonging to Die Linke and having the minimal media bias score, cited only two ULRs and thus is not examined further. Contrarily, Katrin Werner is a member of Die Linke but receives the maximal media bias score within her political party. She mostly cites `https://www.welt.de` which is right-center biased. On par with Jörg Cezanne, Julia Verlinden from Bündnis 90/Die Grünen mainly cites `https://www.change.org`, a left biased website. Marcus Bühl from AfD cites URLs from a diverse range of primarily left-center biased websites, resulting in a media bias score near -0.25.

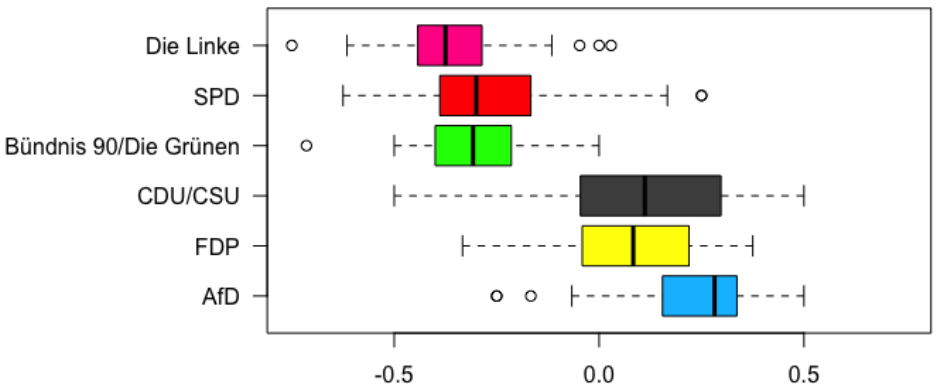

Figure 6: Media bias scores per political party including outliers.

5 Conclusion

The increased use of Twitter as a social media platform and communication canal raises novel concerns, e.g., hidden media bias in tweets. This report raises awareness of the online presence of media bias via addressing external citation usage in tweets of members of the German Bundestag. Data required for the analysis is collected in three steps. First, media bias placements of various news mediums are gathered from `https://mediabiasfactcheck.com`. As its sub websites are static, XPath queries are used to extract the name, website URL, and bias placement of each news medium. Generally, web data collection in this step has to be robust and account for irregularities. Next, information about politicians is gathered. `https://www.bundestag.de/abgeordnete/biografien` exhibits all current members of the Bundestag, including their respective political affiliations. This website is dynamic and thus requires data collection using RSelenium. After saving each politician's name and corresponding political party, data is enriched with Twitter display names via the legislatoR package. Lastly, the $1,000$ most recent tweets from each current member of the Bundestag are retrieved via the Twitter API and examined for URL citations. Overall, media bias is found to be present in the external citation usage of tweets and matches the examined politicians' political directions.

The available amount of media bias placements of German news mediums is not extensive. Most examined politicians cite German news articles, and thus the resulting individual media bias scores may not be fully representative. As media bias is an upcoming research field, I believe that more positionings of German news mediums will become available and hence alleviate the constraint. Further, media bias placements undergo a narrow categorization into five biases, and the method of categorizing news mediums differs depending on the website. Besides, natural language processing is not included in the analysis of URL citations, leaving the possibility of, e.g., sarcasm or the demonstration of an opposite opinion in a tweet. Overall, citation behavior on Twitter differs significantly from traditional standards and has to be examined further [20]. A possible direction would be to explore the effect of retweets on URL citation usage.

References

[1] AllSides. Trump asks if disinfectant, sunlight can treat coronavirus. `https://www.allsides.com/story/trump-questions-if-disinfectant-sunlight-can-treat-coronavirus`, 2020. Accessed: March 31, 2021.

[2] AllSides. Media bias ratings. https://www.allsides.com/media-bias/media-bias-ratings, 2021. Accessed: March 30, 2021.

[3] Alberto Ardèvol-Abreu and Homero Gil de Zúñiga. Effects of editorial media bias perception and media trust on the use of traditional, citizen, and social media news. *Journal. Mass Commun. Q.*, 94(3):703–724, 2017.

[4] Bert Jan Brands, Todd Graham, and Marcel Broersma. Social media sourcing practices: How dutch newspapers use tweets in political news coverage. In *Managing democracy in the digital age*, pages 159–178. Springer, 2018.

[5] Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opin. Q.*, 80(S1):250–271, 2016.

[6] Deutscher Bundestag. Bundestag - biografien. https://www.bundestag.de/abgeordnete/biografien. Accessed: March 30, 2021.

[7] Deutscher Bundestag. Bundestag - fraktionen. https://www.bundestag.de/parlament/fraktionen. Accessed: March 30, 2021.

[8] Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. Analyzing political bias and unfairness in news articles at different levels of granularity. *CoRR*, abs/2010.10652, 2020.

[9] Dave D'Alessio. An experimental examination of readers' perceptions of media bias. *Journal. Mass Commun. Q.*, 80(2):282–294, 2003.

[10] Gabriel Domingos de Arruda, Norton Trevisan Roman, and Ana María Monteiro. Analysing bias in political news. *J. Univers. Comput. Sci.*, 26(2):173–199, 2020.

[11] Homero Gil de Zúñiga, Trevor Diehl, and Alberto Ardèvol-Abreu. When citizens and journalists interact on twitter. *Journal. Stud*, 19(2):227–246, 2018.

[12] Sascha Göbel and Simon Munzert. The comparative legislators database. https://github.com/saschagobel/legislatoR, 2020. Accessed: March 31, 2021.

[13] Felix Hamborg, Norman Meuschke, and Bela Gipp. Matrix-based news aggregation: Exploring different news perspectives. In *JCDL*, pages 69–78, 2017.

[14] Ad Fontes Media Inc. Media bias chart. https://www.adfontesmedia.com/static-mbc/, 2021. Accessed: March 30, 2021.

[15] Twitter Inc. Twitter api. https://developer.twitter.com/en/docs/twitter-api, 2021. Accessed: March 30, 2021.

[16] Media Bias Fact Check LLC. Media bias fact check. https://mediabiasfactcheck.com, 2021. Accessed: March 24, 2021.

[17] José Luis Ortega. To be or not to be on twitter, and its relationship with the tweeting and citation of research papers. *Scientometrics*, 109(2):1353–1364, 2016.

[18] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proc. of CHI*, pages 443–452, 2009.

[19] Souneil Park, Sang Jeong Lee, and Junehwa Song. Aspect-level news browsing: understanding news events from multiple viewpoints. In *Proc. of Conf. on IUI*, pages 41–50, 2010.

[20] Katrin Weller, Evelyn Dröge, and Cornelius Puschmann. Citation analysis in twitter: Approaches for defining and measuring information flows within tweets during scientific conferences. In *Proc. of ESWC Workshop*, volume 718, pages 1–12, 2011.

[21] Katrin Weller and Cornelius Puschmann. Twitter for scientific communication: How can citations/references be identified and measured? 2011.