Misinformation Detection Algorithms and Fairness across Political Ideologies: The Impact of Article Level Labeling

Jinkyung Park
Vanderbilt University
Nashville, USA
jinkyung.park@vanderbilt.edu

Rahul Dev Ellezhuthil
Rutgers University
New Brunswick, USA
re263@rutgers.edu

Joseph Isaac
Rutgers University
New Brunswick, USA
joseph.isaac@fulbrightmail.org

Christoph Mergerson
University of Maryland
College Park, USA
cm495@umd.edu

Lauren Feldman
Rutgers University
New Brunswick, USA
lauren.feldman@rutgers.edu

Vivek K. Singh
Rutgers University
New Brunswick, USA
v.singh@rutgers.edu

ABSTRACT

Multiple recent efforts have used large-scale data and computational models to automatically detect misinformation in online news articles. Given the potential impact of misinformation on democracy, many of these efforts have also used the political ideology of these articles to better model misinformation and study political bias in such algorithms. However, almost all such efforts have used source level labels for credibility and political alignment, thereby assigning the same credibility and political alignment label to *all* articles from the same source (e.g., the New York Times or Breitbart). Here, we report on the impact of journalistic best practices to label *individual* news articles for their credibility and political alignment. We found that while source level labels are decent proxies for political alignment labeling, they are very poor proxies – almost the same as flipping a coin – for credibility ratings. Next, we study the implications of such source level labeling on downstream processes such as the development of automated misinformation detection algorithms and political fairness audits therein. We find that the automated misinformation detection and fairness algorithms can be suitably revised to support their intended goals but might require different assumptions and methods than those which are appropriate using source level labeling. The results suggest caution in generalizing recent results on misinformation detection and political bias therein. On a positive note, this work shares a new dataset of journalistic quality individually labeled articles and an approach for misinformation detection and fairness audits.

CCS CONCEPTS

• **Human-centered computing**; • **Applied computing** → **Publishing**;

KEYWORDS

misinformation detection, algorithmic fairness, political bias, article level labeling

ACM Reference Format:
Jinkyung Park, Rahul Dev Ellezhuthil, Joseph Isaac, Christoph Mergerson, Lauren Feldman, and Vivek K. Singh. 2023. Misinformation Detection Algorithms and Fairness across Political Ideologies: The Impact of Article Level Labeling. In *15th ACM Web Science Conference 2023 (WebSci '23), April 30–May 01, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3578503.3583617

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebSci '23, April 30–May 01, 2023, Austin, TX, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0089-7/23/04…$15.00
https://doi.org/10.1145/3578503.3583617

1 INTRODUCTION

Misinformation is false or inaccurate information that is deliberately created and intentionally or unintentionally propagated [56]. The growth in social media has led to unprecedented growth in web-based dissemination, consumption, and propagation of news. While this reduces barriers to entry in the production of news articles, it also allows vested entities to abuse the web to rapidly disseminate misinformation. False or inaccurate information is known to be propagated more often, and far more rapidly than true information, especially when the topic is related to politics [54]. The rapid propagation of misinformation to a large group of people leads to serious threats to democracy by misleading the public [2], intensifying the political divide [13], increasing mistrust of legitimate media [20], and even leading to radicalization and violence [17]. In fact, recent research identifies misinformation as one of the biggest challenges to democracy [39].

Hence, validating online news and preventing the spread of misinformation is critical for ensuring trustworthy online environments and protecting democracy. One way to counter the spread of misinformation is the use of algorithms to detect misinformation based on the content and the propagation patterns of online news [9]. In the previous literature, the credibility of news articles and their political leanings have typically been evaluated based on the general reputation of the source [18, 22, 44]. That is, all news articles from the source are deemed equally reliable or unreliable (respectively right leaning or left leaning) depending on the reputation of the source. This is mainly because there is a dearth of fine-grained labels defined at the news article level.

We acknowledge that labeling each news article may not be feasible given the massive volume of news articles that are published and disseminated on the web. At the same time, there are

reasons to question the validity of datasets labeled at the source level [4, 40, 49]. For instance, a limited match between source-level and content-based labels was documented in recent studies [4, 16]. As highlighted by the previous research [15], the validity of computational science heavily depends on the integrity of the data, and it is an academic responsibility to set strict methodological expectations when using large-scale datasets. Motivated by prior studies, in this work, we rigorously assessed the credibility and political leaning of 1,000 news articles and used these article-level labels to build misinformation detection algorithms. Then, we evaluated how the labeling methodology (source-level vs article-level) impacts the performance of misinformation detection algorithms.

To understand the downstream impact of such article level labeling on political fairness in misinformation detection, we compare the quality of results as obtained for articles with different sensitive attributes (e.g., left/right leaning). Recently multiple politically right-leaning groups have claimed that the practices of the content moderators on Facebook and YouTube favor the political left [10, 11, 27, 31]. Similarly, multiple recent studies have reported on political bias in the algorithms used by YouTube recommendation [37] and Twitter search [33].

This follows a line of recent results where machine learning algorithms have been found to systematically make discriminatory decisions based on protected characteristics (e.g., race/ethnicity) in multiple domains [6, 7]. A recent article [44] reported on the presence of a political ideology-based bias in misinformation detection algorithms using large scale source-level labeling for credibility and political leaning of news articles. Here, we study the impact of article-level labeling on the process and study if similar bias exists when applying the same machine-learning approach at article level and if bias reduction approaches can be effectively applied when dealing with individually labeled articles.

The main contributions of this work are as follows:
C1: Introducing and sharing a journalistic quality dataset [1] with article level labels for true/fake news and political leaning ($N > 700$)
C2: Comparing and contrasting the impact of labeling resolution (source vs article level) for credibility and political labeling.

The above methodological contributions allow us to answer the following questions at *article level*:
RQ1: Are misinformation detection algorithms based on article-level labels susceptible to bias in terms of political leaning?
RQ2: Can the level of bias in the aforementioned misinformation detection algorithms be reduced while maintaining high accuracy?

1.1 Related Work

1.1.1 Misinformation Detection. "Misinformation" is an umbrella term used to represent false or misleading information. For instance, the term "disinformation" is used to describe information that is inaccurate and is usually distinguished from misinformation by the purpose to deceive [35, 56]. The term "fake news" refers to news articles that intentionally spread false information to mislead the audience [2]. Fake news is understood as a form of "genre blending" [40] due to its structural similarities with traditional journalism [40, 53]. It mimics news media content in a way that it combines "elements of news ideals with features exogenous to the normative model of professional journalism" [35, 40]. While some scholars refrain from the use of the term "fake news" due to its political misuse, we follow others [18, 40, 53] who utilize the "fake news" label to describe false, misleading, hyper-partisan, and sensationalized content. Also, we use the term "misinformation" to refer to false or misleading media content. One technological approach to curtail the flow of misinformation on the web is the use of automated algorithms to detect misinformation. Misinformation detection has been applied in various ways. Broadly, misinformation is classified by analyzing the content of misinformation (e.g., textual or image component) and by analyzing the propagation of misinformation (e.g., how misinformation circulates among users/networks) [47, 51, 58]. In this work, we study misinformation classification algorithms built upon the textual features of news on the web.

1.1.2 Resolution of Analysis: Source vs. Article level. In previous literature, there are two primary approaches to labeling misinformation datasets: 1) using the *credibility of the sources* publishing the content or 2) verifying the *credibility of news articles* with fact-checking parties. With the former approach, the credibility of news is determined based on the reputation of the source. Existing literature on misinformation detection that relied on source-level labels assumes that all new articles from a given source share the same credibility level (reliable vs. unreliable) depending on the reputation of the source [18, 22, 44]. The second approach is to have news articles verified by fact-checking agencies (e.g., Snopes, PolitiFact, BuzzFeed). This approach is frequently used to construct datasets with a few hundred or thousand labeled misinformation contents (e.g., rumors, short statements) [34, 38, 50, 55]. Although labor-intensive and less scalable, this approach could provide more accurate labels than the source-level credibility label [47, 49].

Some recent efforts have started focusing on studying the differences between source and article level labeling. Sharma et al., [49] discuss the issue of disparity between source-level and article-level credibility labels. Focusing their analysis on tweets (some of which may refer to news sources), they found a high degree of consistency between source and article level labels and was able to use source level labels as "weak labels" for model refinement. However, we focus on full-length articles (rather than short texts on Twitter) and have different findings (as detailed later).

Asr et al., [4] also study the interplay between source reputation and content credibility. Based on a dataset of 312 articles sourced from Snopes, they study the correlations between source-level labels and human-labeled content credibility. They found a limited match between source-level and content-based labels. Based on a smaller dataset of 145 articles, they also study the impact of source-level vs. content-based labels on misinformation detection algorithms. Some of the relevant issues with this work are potential bias in data selection (based on those selected by Snopes, Buzzfeed, etc. for detailed fact-checking), missing details on the expertise of the human labelers, diversity of the news article topics, and the smaller sample size.

In another work [16], researchers also explored how the political leaning of news articles (i.e., Liberal, Neutral, Conservative) can be different from those of their source outlets. They compared the political leaning labels of 460 news articles regarding gun policy and immigration and found that more than 50% of the article-level

[1] DOI 10.17605/OSF.IO/QWNSF

political leanings do not match their source-level political leanings. However, the study relied on political learning labels provided by non-experts (i.e., crowd-sourced participants), rather than experts' assessment of the political leaning of the news articles.

We try to counter some of these issues in our work by starting with a larger dataset that has been balanced in the context of misinformation in political news. Unlike the existing research that focused on either credibility or political learning labels, we labeled both the credibility and the political learning of the news articles at the individual article level and compared those labels with source-level labels. Using source level labels as the starting point, we consider 1,000 articles, which are drawn equally from four groups based on two axes of credibility and political leaning. The article level labels (credibility and political leaning) are based on rigorous assessments by a team that includes journalism and political communication experts. This allows us to understand the discrepancy levels in political leaning labels in consonance with credibility labels and study algorithmic fairness across political ideologies.

1.1.3 Ideological Asymmetry in Political Misinformation. According to previous literature, there is a broad ideological asymmetry in the propagation and consumption of political misinformation, and conservative (right-aligned) media are more partisan and more insular than liberal (left-aligned) media [13]. For instance, misinformation consumption during the 2016 election was disproportionately concentrated among Trump supporters, particularly those with the heaviest conservative information diets [19].

The ideological asymmetry in political misinformation stems in part from the psychological differences between liberals and conservatives. That is, liberals and conservatives find different features of messages persuasive and appealing [28]. For instance, compared to liberals, conservatives are more attracted to information that is aggressive in tone and deals with threats [57]. Thus, misinformation targeting conservative and liberal audiences may use different strategies to appeal to its audience. Due to the ideological asymmetry in the production and consumption of political misinformation, misinformation detection algorithms may be biased for left- versus right-leaning news. If algorithms are more effective at weeding out misinformation in conservative news versus liberal news (or vice versa), it risks selectively suppressing certain views and manipulating the information environment to benefit one political side, thereby threatening democratic discourse. Hence, any solutions to counter misinformation must take political asymmetry into consideration [35].

1.1.4 Algorithmic Fairness. Various algorithms are applied to make important decisions that were made by humans in the past. A plethora of research has suggested that machine learning algorithms are susceptible to discriminatory decision-making. If the performance of algorithms varies depending on protective classes (e.g., age, race/ethnicity, gender, socio-economic class, etc.), the algorithms are considered biased or unfair. Recently, significant effort has been made to mitigate bias and promote fairness in machine learning algorithms [3, 32, 43, 52]. The existing literature points to multiple scenarios for algorithmic bias including when (a) input data has unequal representation from different groups, (b) historically there is not enough positive outcome for the unprivileged group, and when (c) the algorithm processes are (intentionally and unintentionally) designed to yield unequal decisions [23, 30, 36]. Accordingly, techniques to mitigate algorithmic bias attempt to modify (a) the process of the training data (pre-processing), (b) the learning algorithms (in-processing), and (c) the prediction (post-processing) [36]. For instance, Park et al. [43] audited phone-based mental health prediction algorithms for gender bias (female vs male) and applied one of the pre-processing techniques (e.g., Disparate Impact Remover) to mitigate such bias. Another work by Singh and Hofenbitzer [52] tackled the problem of biased decisions made by cyberbullying detection algorithms. They applied one of the post-processing approaches (e.g., Equalized Odds) to lower the discrepancies in the performance of cyberbullying detection algorithms depending on one's network position.

Recently, Park et al., [44] addressed the problem of identifying and reducing discriminatory decisions made by misinformation detection algorithms based on source level labels for credibility and political leaning. Focusing on two fairness metrics (i.e., Disparate Impact and Statistical Parity Difference), they applied the Reject Option Classification (ROC), one of the post-processing techniques, to mitigate algorithmic bias in misinformation detection. In this work, we address the same problem of discriminatory decisions made by misinformation detection algorithms with credibility and political-leaning labels at the article level. To this end, we apply another bias reduction approach called Disparate Impact Remover (DIR), one of the pre-processing techniques, and focus on different fairness metrics (i.e., Delta Accuracy).

2 MATERIALS AND METHODS

2.1 Labeling of Articles for political leaning and credibility

2.1.1 Dataset and source level labeling. To analyze bias in misinformation detection algorithms, we used a subset of the NELA-2018 dataset [41]. The original dataset contained 713K news articles from 194 media sources. We follow the approach adopted by [44] for obtaining source-level labels for credibility and political leaning. They used source-based credibility labels (fake vs. real) from NewsGuard as it works with trained journalists to evaluate the credibility of news sources on the web. They used BuzzFeed for source-level political leaning (i.e., left vs. right) labels, as its sources covered a large proportion (36.3%) of the articles in the dataset. With this process, a total of (N = 102k) news articles received labels for both credibility and political leanings. Out of 102k source-level labeled articles, 37.5k articles (36.7%) belonged to left-aligned sources and the remaining 64.5k articles (63,7%) belonged to right-aligned sources.

2.1.2 Article level labeling. Instead of relying on source-level labels for credibility and political leaning, we wanted to create content-based article-level labels for the news articles. This necessitated a smaller sample from the abovementioned dataset of 102k articles.

To maintain contextual integrity, we decided to focus only on news stories about U.S. politics. Hence, the articles were passed through the IPTC newscode API to restrict the focus to only political articles [25]. We also eliminated sources from the dataset that had an explicit non-U.S. focus (e.g., The Irish Times, The Moscow Times,

France24). Further, we eliminated satirical news sources (e.g., The Onion, The Spoof) from the dataset.

Next, we binned the set of 102K articles into four groups based on source-level political learning and credibility labels: (1) true + right-leaning; (2) false + right-leaning; (3) true + left-leaning; (4) false + right-leaning. From each bin, an equal number of articles (i.e., 250) were selected for detailed human coding. Articles were sampled from between Feb. 1, 2018 – Feb. 10, 2018. Article-level coding was conducted by a journalism faculty member who is also an expert on political communication and two journalism Ph.D. students, one of whom previously worked as a professional fact-checker. A codebook was developed with coding rules (described in more detail below) to determine the news article's focus (on U.S. politics or not), political leaning (liberal vs. conservative), and credibility (mostly true, a mix of true and false, mostly false).

The three coders participated in four rounds of practice coding to fine-tune the codebook and improve inter-coder reliability. Practice coding and reliability coding were conducted on articles randomly sampled from the larger NELA dataset. During practice coding, coders labeled the articles independently and then discussed any disagreements. A final test of inter-coder reliability, calculated using Krippendorff's alpha [21], was conducted on a sample of 110 news articles. For the final coding (N = 1,000), articles were divided among the two Ph.D. students to code independently. The two coders manually coded the credibility as well as the political leaning of each individual article based on the following guidelines:

- Does the article read like a valid U.S. political news article, as opposed to other topics (e.g., sports, entertainment)? This criterion filtered down the subset to 907 articles
- Identify the political leaning of the article from a scale of 1 to 5 (1: strong liberal, 5: strong conservative). Articles labeled as Neutral (totaling 104) were not coded further.
- Identify the credibility of the contents of the article on a scale from 1 to 3 (1: mostly true, 3: mostly false). The coders could not decide on the credibility rating for 85 articles. Hence a total of 706 articles received valid scores after this step.

2.1.3 Political News. As noted above, in constructing our sample, we took steps to limit the sample to only articles focused on U.S. politics. However, to account for the possibility that some non-U.S. political articles would nonetheless end up in the sample, we coded each article for whether it focused on U.S. political news or not (Krippendorff's alpha = .80). Politics included topics related to elections, parties, politicians and elected officials, government institutions and processes, policies and policymaking (including U.S. foreign policy), and foundational political concepts like justice, freedom, etc. As a result, 93 articles were labeled as unrelated to U.S. politics and were eliminated from further coding.

2.1.4 Political Leaning. After eliminating articles that were not focused on U.S. politics, the remaining articles (N = 907) were coded for political leaning (strongly liberal, mildly liberal, neutral, mildly conservative, strongly conservative; Krippendorff's alpha = .91). Source information was blinded when coding for political leaning. Articles were coded as strongly liberal or conservative if they clearly promoted liberal/Democratic [conservative/Republican] politicians and issue positions and/or attacked conservative/Republican [liberal/Democratic] politicians and issue positions. Strongly biased articles typically included inflammatory, loaded language. Articles were coded as mildly biased if they used neutral language but the tone of the article was somewhat favorable or considered "good news" for liberals/Democrats [or conservatives/Republicans], or if the balance of the article favored liberal/Democratic [or conservative/Republican] politicians and positions but included some information that acknowledged the validity of the other side. Neutral articles did not clearly favor one side or the other or were otherwise ambiguous in tone. Final coding results indicated strongly liberal n = 334, mildly liberal n = 111, neutral n = 104, mildly conservative n = 125, and strongly conservative n = 233. For political leaning, we combined the counts of "strongly liberal" and "mildly liberal" into "liberal" and "mildly conservative" and "strongly conservative" into "conservative." Neutral articles were eliminated from further coding.

2.1.5 Credibility. After eliminating neutral articles, the remaining articles (N = 803) were coded for credibility using the categories similar to those used by fact-checking organizations such as BuzzFeed (mostly true, mix of true and false, mostly false, unverifiable; Krippendorff's alpha = .73). For each article, coders fact-checked the claims in the article using existing fact-checking resources (e.g., PolitiFact, FactCheck.org, Snopes), relevant primary source information (e.g., official statements from government officials, video of speeches/interviews, legislative documents, etc.), and/or coverage in multiple reputable news outlets (e.g., The Washington Post, The Wall Street Journal, etc.). For articles coded as mostly true, all significant details, including relevant numbers, quotes, and event descriptions, were represented accurately, and any opinion or evaluative statement was clearly presented as such. Articles coded as a mix of true and false contained some accurate information but also included at least one major claim that was taken out of context, exaggerated, speculative, or otherwise misleading, or included a misleading headline. In mostly false articles, many of the article's claims were inaccurate or missing essential context, and/or the central claim or event described in the article was false. Finally, the unverifiable label was used for articles that did not include any substantive facts that could be fact-checked (i.e., pure opinion) or for articles that were otherwise unverifiable using typical fact-checking strategies. This resulted in 516 articles that were mostly true, 169 that were a mix of true and false, 21 that were mostly false, and 97 unverifiable. For purposes of analysis, we collapsed two groups "mix of true and false" and "mostly false" to the label "fake news," as a mixture of true and false is a very important category of misinformation that this work aims to tackle.

Table 1 shows the final distribution of coded articles across political leaning and credibility (N = 706). Table 2 gives the breakdown of the number of articles in the final dataset across different news sources and their political alignment and credibility assessed at the source level compared to the article level.

2.2 Misinformation Detection using Machine Learning and Fairness Audits

2.2.1 Misinformation Detection. There are two major approaches in machine learning literature to extract features: deep learning and

Table 1: The final distribution of coded articles across political leaning and credibility

Political Leaning	True news	False news
Liberal	317	78
Conservative	199	112
Total	516	190

hand-crafted features. This work follows the approach proposed by a recent effort [44] that worked with source level labels for the same task (misinformation detection) using the same initial dataset (NELA-GT). In fact, the comparison of performance across source-level and article-level processes is one of the goals of this work. The features were designed in a hand-crafted manner by domain experts. Based on a survey of previous misinformation literature, we focus on four broad categories of features: structure, subjectivity, sentiment, and social identity. [2]

(1) **Structure**: the organization of the content into different stylistic structures, such as the syntax, text style, and grammatical elements of news content and title (e.g."number of words," "number of question marks").
(2) **Subjectivity**: features that present an effort to convey a certain opinion or viewpoint rather than facts. For instance, we used "cognitive processes" (e.g. cause, know, ought) and "perceptual processes" (e.g., look, heard, feeling) categories and their subcategories from LIWC 2015 [45].
(3) **Sentiment**: the emotion-arousing aspects of the news articles that contain misinformation such as "emotional tone" (i.e., positive and negative emotions) [24] and "affective processes" category (e.g., happy, cried) from LIWC 2015 [45].
(4) **Social Identity**: the features that show the qualities or beliefs that make a particular group distinct from other groups including "conservative identity words" (e.g., republican, right-wing), "liberal identity words" (e.g., democrat, left-wing), and "moral words" (e.g., blame, innocent, guilt) [42].

Before applying any of the machine learning algorithms, the missing values were filled with the median values of corresponding features. To reduce the impact of features with high variance, features were standardized by centering their mean to zero and by scaling them to unit variance. We split the dataset in a random manner, setting aside 80% of the dataset for training the machine learning algorithms and using the remaining 20% for testing. The split was performed in a stratified manner using both credibility labels as well as political orientation. This was to ensure that the proportion of real and fake news articles in the training and test sets are the same, reducing the possibility of bias issues arising from the imbalanced dataset.

Similar to previous work [44] that uses the same features, we found that the Random Forest approach yields the highest accuracy for misinformation detection. Hence, we chose the Random Forest for machine learning model development. The number of estimators/ number of trees for Random Forest was set to 100 and there was no limit set regarding the maximum depth of the forest.

[2] refer to [44] for a complete set of features

2.2.2 Bias Audit and Bias Reduction. Given several recent claims that media platforms favor politically left [10, 31], we consider right-aligned articles as an 'unprivileged group' when auditing the algorithms for bias. Note that the sensitive feature (political alignment) was not included in the training phase, but was used to compute various fairness metrics during the testing phase.

The notion of fairness has multiple interpretations in the extant literature which include maximizing the communal good, calibration of outcomes for similar input, restorative justice, etc. Here we adopt the notion of *distributional justice* as has been suggested by John Rawl [48]. This notion interprets fairness as the *demand for impartiality*. Again, there exist multiple interpretations of impartiality, but here we focus on group-level fairness for two political ideologies (left vs. right in the US political context). Group-level fairness can be operationalized as impartiality in the accuracy levels observed for articles with different political leaning. Hence, "delta accuracy" can be used as a fairness metric [46]. The ideal value for delta accuracy is 0.

$$\Delta acc = acc(S = unprivileged) - acc(S = privileged) \quad (1)$$

Another interpretation of impartiality focuses on "disparate impact." This interpretation underscores that any algorithmic decision is based on the input data, which cannot always be assumed to be fair. For instance, if past data in the criminal justice system is biased and/or an outcome of unequal policing, it would be unreasonable to call for equal accuracy for different groups based on such input data as being fair. Hence, this interpretation argues a case for the equal statistical representation of positive outcomes out of an algorithm irrespective of the input data. In many legal cases, if the representation of a group (e.g., women) falls below 80% of what it should be expected on a statistical basis, the process is considered to be biased, irrespective of the intention of the system designers [8, 14]. Hence, we consider two related metrics: Disparate Impact (DI) [14] and Statistical Parity Difference (SPD) [12].

Disparate impact quantifies the ratio of participation of the two groups in the positive class (credible news). The ideal value of the disparate impact equals 1.0.

$$\frac{p(\hat{Y} = 1|S = unprivileged)}{p(\hat{Y} = 1|S = privileged)} \quad (2)$$

Similarly, the statistical parity difference quantifies the gap in the ratio of participation of the two groups in the positive class (credible news). The ideal value for SPD is 0.

$$SPD = p(\hat{Y} = 1|S = unprivileged) - p(\hat{Y} = 1|S = privileged) \quad (3)$$

Following recent literature [1, 52], we used a statistical t-test on the means of the above-mentioned fairness metrics for the two groups (left-aligned and right-aligned) when auditing misinformation detection algorithm for bias. We consider algorithms to be less biased if the algorithms yield *reduced* delta accuracy and statistical parity difference, and move the disparate impact score closer to 1.0.

We considered two different techniques for bias reduction. The first is Reject Option Classification (ROC) [29] which was also used by [44] in a similar setting in past work. ROC is a post-processing algorithm that works by modifying the "as-is" results from the

Table 2: News sources, number of articles, political alignment, and credibility at source level and article level

News Source	Number of Articles	Political Alignment Source	Political Alignment Article	Credibility Rating Source	Credibility Rating Article
Bearing Arms	2	Right	L:0, R:2	Real	R:1, F:1
Bipartisan Report	76	Left	L:75, R:1	Fake	R:64, F:12
Breitbart	80	Right	L:6, R:74	Fake	R:62, F:18
CNN	39	Center	L:34, R:5	Real	R:39, F:0
CNS News	31	Right	L:0, R:31	Real	R:21, F:10
Daily Kos	9	Left	L:9, R:0	Fake	R:7, F:2
Daily Signal	6	Right	L:0, R:6	Real	R:2, F:4
Drudge Report	12	Right	L:5, R:7	Fake	R:11, F:1
FrontPage Magazine	23	Right	L:0, R:23	Fake	R:5, F:18
Infowars	43	Right	L:5, R:38	Fake	R:30, F:13
Media Matters for America	16	Left	L:11, R:5	Real	R:12, F:4
NPR	5	Center	L:5, R:0	Real	R:5, F:0
National Review	37	Right	L:3, R:34	Real	R:28, F:9
News Busters	79	Right	L:0, R:79	Real	R:47, F:32
Palmer Report	63	Left	L:63, R:0	Fake	R:16, F:47
Politicus USA	47	Left	L:46, R:1	Real	R:39, F:8
Salon	76	Left	L:71, R:5	Real	R:72, F:4
Shareblue	62	Left	L:62, R:0	Fake	R:55, F:7

classifier to make them fairer. Specifically, it focuses on the instances that lie near the decision boundary in which labels are difficult to identify. For such instances, it flips the outcomes in a probabilistic manner and gives preferential treatment (i.e., higher odds of a positive outcome) to the unprivileged class. Similarly, the instances belonging to the privileged class that lie near the decision boundary are assigned an undesirable label.

The second bias reduction approach considered is Disparate Impact Remover (DIR) [14]. DIR is a pre-processing technique designed to minimize the effects of an attribute acting as a strong signal for privileged/unprivileged class membership. For an unprivileged class, the disparity can materialize as a significant shift of attribute values when compared to the privileged class (e.g., the distribution of heights of malnourished children might be different from their well-fed counterparts). When training a model, these differences can end up acting as a proxy for privileged/unprivileged class membership and can influence the model's decision-making process even when the model has no information about privileged/unprivileged classes. Disparate Impact Remover (DIR) works by editing the distribution of values within a feature such that the distribution of the feature for privileged and unprivileged classes is made similar.

We considered the above two techniques as they do not require any changes in the classification algorithm and hence are used for a broad range of applications [26, 29]. We implemented ROC and DIR using the IBM AIF360 library [5]. The classification algorithms were run 100 times and each iteration had a shuffled version of the dataset. The average results across the 100 iterations are reported. Note that classification algorithms are considered less biased if they yield *reduced* delta accuracy, DI value closer to 1.0, and SPD value closer to 0.

Table 3: Comparison of Source level and Article level labeling along Political Leaning and Credibility axes

Political Orientation	Credibility Label	Source level	Article level
Left	Real	139	278
Left	Fake	210	78
Right	Real	155	112
Right	Fake	158	194

3 RESULTS

3.1 Comparing Source-level and Article-level Labeling

Table 2 describes the political alignment and credibility labels for the articles from different sources using the two different processes (source vs. article level labeling) (N = 706). Table 3 summarizes the effect of labeling at the article level as opposed to source-level annotations in a 2 x 2 matrix. Please note that 44 out of 706 articles in Table 2 (from CNN and NPR) were categorized as neutral at the source level and hence are not part of the 2 x 2 analysis in Table 3.

When using source-level labeling, 139 articles were categorized as left-leaning and real as opposed to 278 articles being categorized in the same category when using article-level labeling. Similarly, 210 articles were categorized as left-leaning and fake when using source-level labeling, only 78 were categorized as such when article-level labeling was used. In contrast, the difference between articles categorized as right-leaning and real when using source-level and article-level annotations was smaller (155 to 112). The difference between Right/Fake articles was again smaller (158 to 194). To

Table 4: Comparison of Source type and Article type labeling along the Political Leaning axis (Left/Right)

		Article Level	
		Left	Right
Source Level	Left	337	12
	Right	19	294

Table 5: Comparison of Source type and Article type labeling along the Credibility axis (Fake/Real)

		Article Level	
		Fake	Real
Source Level	Fake	118	250
	Real	72	222

compare the overall differences, we utilized a Chi-squared analysis of the differences between source-level annotation and article-level annotations. The results indicated that this difference is statistically significant ($p < 0.05$).

To further elucidate the impact of using article-level labels as opposed to source-level labels, we zoom into the differences considering only the political leaning (Table 4) and credibility (Table 5). From Table 4, we see that source-level labeling is a reasonable proxy for inferring political leaning ($accuracy > 95\%$). On the other hand, from Table 5, we see that source-level labels for credibility cannot be used as a proxy for inferring the actual credibility of the article ($accuracy = 51\%$), which is almost the same as flipping a coin to assign a credibility rating.

This can be interpreted in more detail by consulting Table 2. For example, Bipartisan Report was identified as a source of fake news as per the source level labeling, and hence 100% of the news articles from that source were considered as fake news in source level analysis. However, article-level labeling yielded only 12 out of 76 articles (15.78%) of the articles to be fake. Hence, 85% of the articles from this source will be unjustifiably labeled as false with a source-level analysis. A reverse example was News Busters. While source-level labeling considered it to be real i.e., 0% fake news, 32 out of its 70 articles (i.e., 45.71%) were deemed to be fake news as per our article-level analysis. Given the widespread prevalence of research articles that consider source-level labels for identifying true and fake news, we believe that these results are reflection-worthy.

3.2 Misinformation Detection using Machine Learning and Fairness Audits

3.2.1 Prediction Model and Bias Quantification. Table 6 summarizes the performance of the developed misinformation detection algorithm with article-level labels. The classifier achieves an overall accuracy of 74.5%. We note that this is noticeably higher than a baseline majority class classifier (one that classifies all instances into the majority class, 58.9%); however, it is lower than that reported in [44] (87.85%) which utilized source-level labels for credibility using the same set of features and the same machine learning algorithm (i.e., Random Forest). When considering the algorithm's accuracy on the subset of left and right-leaning articles, the issue of political bias becomes obvious. Here, the classifier is able to classify left-leaning articles with an accuracy of 81.5% while right-leaning articles only have 65.9% accuracy. This marks a "delta accuracy" gap of 15.37% between the left-aligned and right-aligned news articles, indicating issues of bias within the model. While this value is high, statistical parity difference (SPD) and disparate impact (DI) are near optimal levels (3.04%, 1.03, respectively). As SPD and DI focus on the favorable outcome assignment rather than the unfavorable outcome assignment, we interpret the results to mean that while the model is fairly assigning a favorable outcome to both privileged and unprivileged classes, it does not do so when predicting an unfavorable outcome.

Table 6: Performance of the misinformation detection in terms of accuracy and fairness

Metric	Ideal Value	Observed Value
Overall Accuracy	100.00%	74.54%
Left Accuracy	100.00%	81.50%
Right Accuracy	100.00%	65.94%
Delta Accuracy	0.00%	15.56%
Disparate Impact	1.00	1.03
Statistical Parity Difference	0.00%	3.04%

3.2.2 Bias Reduction. Table 7 delineates the impact of two bias reduction mechanisms: Reject Option Classification (ROC) and Disparate Impact Removal (DIR) on the developed misinformation detection model. Performance is measured using accuracy as well as the three bias metrics described in the previous section.

While both the bias reduction approaches help to reduce bias in terms of delta accuracy (which was the primary metric of concern), the DIR approach is more effective in reducing this metric. Further, the DIR approach yields better overall accuracy. The ROC approach performed yields better results in terms of DI and SPD, but given that the primary metric of concern was delta accuracy and the associated accuracy gain, we consider DIR to be a better choice for bias reduction in this case. With this bias reduction process, delta accuracy drops to 2.39% signifying a reduction in bias and accuracy increases to 75.20%. This result and the corresponding accuracy for left-aligned and right-aligned articles are visualized in Figure 1.

4 DISCUSSION

The main contributions of this work are to introduce a new journalistic quality dataset with article level labels for credibility and political leaning ($N > 700$) and to study the impact of labeling resolution (source vs article level) on misinformation detection algorithms and fairness audits.

In Table 4, we showed that source-level labeling is a reasonable proxy for inferring political leaning (> 90% alignment). Given, the high cost of article level labeling, it might be reasonable to use source level labeling for political leaning. Interestingly, the results of our study are confronted with the results from the previous study that confirmed discrepancies (more than 50%) between source level and article level political leaning labels. This could be due to the difference in how political leaning was evaluated. In

Table 7: Performance of the misinformation detection in terms of accuracy and fairness with different bias reduction processing approaches

Metric	Ideal Value	Without Bias Reduction	With ROC Bias Reduction	With DIR Bias Reduction
Overall Accuracy	100.00%	74.54%	70.69%	75.20%
Delta Accuracy	0.00%	15.56%	9.71%	2.39%
Disparate Impact	1.00	1.03	1.01	1.06
Statistical Parity Difference	0.00%	3.04%	1.11%	5.25%

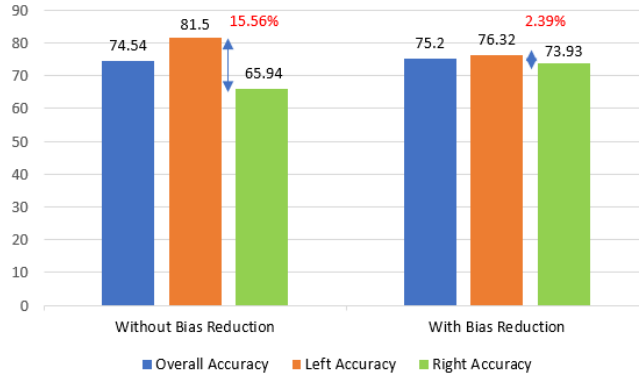

Figure 1: Performance of the misinformation detection algorithm based on political leaning, before and after the bias reduction process

previous research, political leaning scores were averaged from five crowd-sourced participants; whereas, a team of journalism and political communication experts independently evaluated the political leaning of the news article in our work. In addition, we focused on news articles that addressed U.S. politics as a whole, while gun policy and immigration were chosen as specific topics to explore in the previous literature. Examining how annotation approaches and news topics affect the difference between source level and article level political leaning labels warrants further inspection.

As can be seen in Tables 2-5, we found that the credibility labeling for news articles can be quite different based on whether it was undertaken at the source level vs. the individual level. Once again, given the high cost of article level labeling, it might be reasonable to use source level labeling for political alignment. However, even with the high-cost argument, it seems unreasonable to continue using source level labels for the credibility of news articles. A 51% alignment between source and article level labels indicates the very poor quality of the match. The results were not surprising given similar findings of recent works that examined the differences between source and article level credibility labeling [4, 49].

With a further comparison of source and article level labeling, we find that this mismatch is not in a single direction (e.g., always overestimating or underestimating the number of fake news articles). In Table 2, we see multiple examples of source level true labeled venues to have fake news articles and vice versa. Similarly, as shown in Table 3, this mismatch is not constrained to only Left-leaning or only Right-leaning sources. However, what is noteworthy in our work is that the degree of mismatch was much more pronounced for left-aligned articles than the right-aligned articles. This could mean that the use of source-level labeling (rather than article level labeling) results in much poorer quality of data for left-aligned articles than right-aligned articles. This interplay between political alignment and the quality of credibility labels at different granularity is worth further exploration in future work.

Here, we interpret these results to mean that the issue of misalignment between source level and article level labels for credibility is frequent and widespread. At the same time, the number of research efforts that use source level labels for misinformation is rapidly growing and impacting hundreds of other efforts that cite and build upon them. Hence, the results of the current work suggest caution and introspection in identifying aspects that can be generalized from source level labeling and those which cannot.

The second major goal of this work is to study the impact of using article level labels on the workflow for detecting misinformation, bias therein, and ways to reduce such bias. We note that we used the same set of features and machine learning method as reported in previous work [44] that used source level labels for credibility. Our results confirmed that the accuracy of the algorithm was lower (around 75%) with article level labeling than that obtained with source level labeling in previous work (around 88%). One way to interpret this result is that source level labeling is an easier problem to tackle due to multiple repeated credibility labels for articles from the same source than the individual article labeling problem. Using linguistic structure and syntax as input variables, it might be easier for machine learning models to build shortcuts to identify news sources than to identify the finer variations between true and fake news, potentially coming from the same source.

Next, the results in Table 6 indicate that multiple misinformation classification algorithms performed differently based on the political leaning of the news article. While the difference in Disparate Impact was small, there were noticeable differences in terms of Delta Accuracy. Again, we note that the trends of these differences diverge from those reported with source level analysis in [44], where the biggest disparities were found in terms of the other two metrics (i.e. SPD and DI). With different issues coming up as prominent, it is not surprising that we found that the Reject Option Classification bias reduction method adopted in [44] was no longer the most effective way to reduce the bias levels. Instead, we found the Disparate Impact Remover method to be useful to reduce bias in this work. The DIR approach was able to significantly reduce the delta accuracy gap from 15.37% to 2.39%. Despite the clear differences from the source level labeling scenario, our results indicate that there is clear evidence of value in using theory-driven feature engineering machine learning to predict misinformation, the need for auditing such results for bias along a political leaning

dimension, and for bias reduction to ensure equitable performance across different groups.

To increase public confidence in misinformation detection practices and subsequent corrections, there is a need to ensure the validity and fairness of results. And once again, the validity of computational science heavily depends on the integrity of the data, and it is our responsibility to set strict methodological expectations when using large-scale datasets [15]. This work marks an important early step in that direction. It underscores the need for article level credibility labels for ensuring validity and reliability for misinformation detection. It also adds evidence for the need to audit such misinformation detection algorithms for bias across political leaning and ways to increase fairness in algorithmic decision-making.

Zooming out of the immediate findings, we recognize the inherent human labor needed for article level labeling. In fact, our research questions were partially motivated by the desire to use source level labels as proxies for article level labels. The authors would like to explicitly acknowledge their own use of source level labels in their past work. This work is not intended to critique any specific works that have used source level labels. Rather, it is intended to be a call-to-action to produce article level labeled datasets (to which this work aims to contribute) and devise new approaches that can combine source level and article level information to build more accurate and fair misinformation detection algorithms. This way, misinformation detection models are most likely to benefit all regardless of asymmetry in the political media environment, which may further undermine media trust and perpetuate misinformation on one political side.

5 CONCLUSION

This work aims to increase public confidence in misinformation detection practices and subsequent corrections by ensuring the validity and fairness of results. It reports the consequential impact of article level labeling, as opposed to source level labeling, on credibility and political leaning labels. The results indicate that while source level labels might be a decent proxy for political leaning, they are poor proxies for the credibility of news articles. The downstream impact of these changes in labels is studied in terms of their impact on misinformation detection algorithms, audits for fairness, and bias reduction procedures. The outcomes for each of these steps differed based on the granularity of labels. At the same time, the results indicate the feasibility of a machine-learning approach to obtain reasonable accuracy and fairness in practical settings.

To help with the development of newer approaches with article level labeling, this work introduces a new journalistic quality dataset with labels for true/fake news and political leaning. The dataset and the conceptual results aim to pave the way for more reliable and fair misinformation detection algorithms.

ACKNOWLEDGMENTS

This material is in part based upon work supported by the National Science Foundation under Grant No. SES-1915790.

REFERENCES

[1] Jamal Alasadi, Ahmed Al Hilli, and Vivek K Singh. 2019. Toward fairness in face matching algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*. 19–25.

[2] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[3] Abdulaziz A Almuzaini, Chidansh A Bhatt, David M Pennock, and Vivek K Singh. 2022. ABCinML: Anticipatory Bias Correction in Machine Learning Applications. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1552–1560.

[4] Fatemeh Torabi Asr and Maite Taboada. 2018. The data challenge in misinformation detection: Source reputation vs. content veracity. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*. 10–15.

[5] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[7] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[8] Civil Rights Act. 1964. Civil Rights Act of 1964. *Title VII, Equal Employment Opportunities* (1964).

[9] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of Association for Information Science and Technology* 52, 1 (2015), 1–4.

[10] W Cummings. 2018. Diamond and Silk tell Congress, 'Facebook censored our free speech!,'. *USA Today* (2018). Available online: https://bit.ly/3r6FsJp.

[11] O Darcy. 2021. Republicans and right-wing media use Facebook Oversight Board's Trump decision to claim bias. *CNN* (2021). Available online: https://www.cnn.com/2021/05/05/media/facebook-oversight-board-trump-right-wing-reaction/index.html.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[13] Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center Research Publication* 6 (2017).

[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proc. ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[15] Devin Gaffney and J Nathan Matias. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS one* 13, 7 (2018), e0200162.

[16] 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 International AAAI Conference on Web and Social Media*, Vol. 14. 939–943.

[17] Kelly M Greenhill and Ben Oppenheim. 2017. Rumor has it: The adoption of unverified information in conflict zones. *International Studies Quarterly* 61, 3 (2017), 660–676.

[18] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.

[19] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council* 9, 3 (2018), 4.

[20] Andrew M Guess, Pablo Barberá, Simon Munzert, and JungHwan Yang. 2021. The consequences of online partisan media. *Proceedings of the National Academy of Sciences* 118, 14 (2021).

[21] Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1, 1 (2007), 77–89.

[22] Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018*. 235–238.

[23] Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass* 15, 8 (2021), e12432.

[24] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[25] IPTC. n.d.. IPTC API newscodes. https://show.newscodes.org/index.html

[26] Mohsin Iqbal, Asim Karim, and Faisal Kamiran. 2019. Balancing prediction errors for robust sentiment classification. *ACM Trans. on Knowledge Discovery from Data (TKDD)* 13, 3 (2019), 1–21.

[27] Shan Jiang, Ronald E Robertson, and Christo Wilson. 2020. Reasoning about political bias in content moderation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13669–13672.

[28] John T Jost and Margarita Krochik. 2014. Ideological differences in epistemic motivation: Implications for attitude structure, depth of information processing, susceptibility to persuasion, and stereotyping. In *Advances in motivation science*. Vol. 1. Elsevier, 181–231.

[29] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences* 425 (2018), 18–33.

[30] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[31] J Koebler and J Cox. 2018. The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People - Motherboard. *Motherboard* (2018). Available online.https://motherboard.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works.

[32] Omkar N Kulkarni, Vikram Patil, Vivek K Singh, and Pradeep K Atrey. 2021. Accuracy and Fairness in Pupil Detection Algorithm. In *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*. IEEE, 17–24.

[33] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 417–432.

[34] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PloS one* 12, 1 (2017), e0168344.

[35] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[36] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.

[37] Michael Lutz, Sanjana Gadaginmath, Natraj Vairavan, and Phil Mui. 2021. Examining Political Bias within YouTube Search and Recommendation Algorithms. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–7.

[38] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. (2016).

[39] Susan Morgan. 2018. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy* 3, 1 (2018), 39–43.

[40] Rachel R Mourão and Craig T Robertson. 2019. Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism Studies* 20, 14 (2019), 2077–2095.

[41] Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 630–638.

[42] Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *Am. Political Science Review* 115, 3 (2021), 999–1015.

[43] Jinkyung Park, Ramanathan Arunachalam, Vincent Silenzio, Vivek K Singh, et al. 2022. Fairness in Mobile Phone–Based Mental Health Assessment Algorithms: Exploratory Study. *JMIR formative research* 6, 6 (2022), e34366.

[44] Jinkyung Park, Rahul Ellezhuthil, Ramanathan Arunachalam, Lauren Feldman, and Vivek Singh. 2022. Toward Fairness in Misinformation Detection Algorithms. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media. Retrieved from https://doi. org/10.36190*.

[45] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[46] Dana Pessach and Erez Shmueli. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784* (2020).

[47] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).

[48] John Rawls. 1999. *A theory of justice: Revised edition*. Harvard university press.

[49] Karishma Sharma, Emilio Ferrara, and Yan Liu. 2022. Construction of Large-Scale Misinformation Labeled Datasets from Social Media Discourse using Label Refinement. In *Proceedings of the ACM Web Conference 2022*. 3755–3764.

[50] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[51] Vivek K Singh, Isha Ghosh, and Darshan Sonagara. 2021. Detecting fake news stories via multimodal analysis. *Journal of the Assoc. for Information Science and Technology* 72, 1 (2021), 3–17.

[52] Vivek K Singh and Connor Hofenbitzer. 2019. Fairness across network positions in cyberbullying detection algorithms. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 557–559.

[53] Edson C Tandoc Jr, Ryan J Thomas, and Lauren Bishop. 2021. What is (fake) news? Analyzing news values (and more) in fake stories. *Media and Communication* 9, 1 (2021), 110–119.

[54] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[55] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[56] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 80–90.

[57] Dannagal Goldthwaite Young. 2019. *Irony and outrage: The polarized landscape of rage, fear, and laughter in the United States*. Oxford University Press, USA.

[58] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.