Toward Fairness in Misinformation Detection Algorithms

Jinkyung Park, Rahul Ellezhuthil, Ramanathan Arunachalam, Lauren Feldman, and Vivek Singh
School of Communication & Information, Rutgers University
{jp1676, re263, ra831, lauren.feldman, v.singh}@rutgers.edu

Abstract

Misinformation in online spaces can stoke mistrust of established media, misinform the public and lead to radicalization. Hence, multiple automated algorithms for misinformation detection have been proposed in the recent past. However, the fairness (e.g., performance across left- and right- leaning news articles) of these algorithms has been repeatedly questioned, leading to decreased trust in such systems. This work motivates and grounds the need for an audit of machine learning based misinformation detection algorithms and possible ways to mitigate bias (if found). Using a large ($N > 100K$) corpus of news articles, we report that multiple standard machine learning based misinformation detection approaches are susceptible to bias. Further, we find that an intuitive post-processing approach (Reject Option Classifier) can reduce bias while maintaining high accuracy in the above setting. The results pave the way for accurate yet fair misinformation detection algorithms.

Introduction

The growth of social media has resulted in an increasing number of users consuming, sharing, and even producing online news. On the positive side, this has democratized news and information by reducing the agenda-setting control of large professional news outlets. On the negative side, this has left the public unprotected from the spread of misinformation by stripping out the gate-keeping role of traditional media (Shoemaker and Vos 2009). The lower barriers to entry, combined with the monetized network structure of social media, have contributed to the rapid proliferation of misinformation online. Indeed, false information on Twitter is typically retweeted more often, and far more rapidly, than true information, especially when the topic is related to politics (Vosoughi, Roy, and Aral 2018). Decision-making based on misinformation entails potential social costs. It poses serious threats to democratic institutions by misinforming the public (Allcott and Gentzkow 2017), deepening political divisions (Faris et al. 2017), fueling mistrust of legitimate media (Guess et al. 2021), reducing demand for accuracy (Allcott and Gentzkow 2017), and even leading to radicalization and violence (Greenhill and Oppenheim 2017). Therefore, verifying online news and preventing the spread of misinformation is critical for enabling trustworthy online environments and protecting democracy.

The traditional way to verify online news via manual fact-checking has become difficult (or almost impossible) due to the enormous volume of information that is generated and disseminated online. This problem has led researchers and platform developers to devise automated algorithms to detect misinformation based on the content and the patterns of the news (Conroy, Rubin, and Chen 2015). However, the very algorithms that are intended to fight off one threat (misinformation) may inadvertently be falling prey to another critical threat (bias of the automatic detection algorithms).

To some extent, the problems of algorithmic bias are the same as those of human-based decision-making. For example, multiple politically right-leaning groups have accused Facebook and YouTube's content moderators to be in favor of the political left (Cummings 2018; Koebler and Cox 2018; Jiang, Robertson, and Wilson 2020; Darcy 2021). However, the impact gets amplified with the application of automatic algorithms that scale the process dramatically. In recent years, machine learning algorithms have been found to systematically discriminate and favor one group over another based on demographic characteristics in multiple domains (e.g., automated decisions on parole or college admission) (Calmon et al. 2017; Buolamwini and Gebru 2018). It is possible that misinformation detection algorithms also may exhibit bias, but in this case, based on the political leaning of the news in question. If such discrimination exists, it can lead to the erosion of public trust online and exacerbate political polarization. Hence, it is important to audit misinformation detection algorithms for political bias and redress problems, if found.

The main research questions in this work are:

RQ1: Are misinformation detection algorithms susceptible to bias in terms of political leaning?

RQ2: Can the level of bias in misinformation detection algorithms be reduced while maintaining accuracy?

Related Work

Misinformation Detection Scholars have used various terms to describe the phenomenon of misinformation. "Misinformation" is an umbrella term used to represent false or misleading information, whereas "disinformation" repre-

sents false information that is "purposely spread to deceive people" (Lazer et al. 2018). Other scholars see misinformation and disinformation as symptomatic of a broader "information disorder" plaguing the media environment (e.g., (Wardle and Derakhshan 2017)). The term "fake news" has been used to describe news articles that are "intentionally and verifiably false and could mislead readers" (Allcott and Gentzkow 2017) or "fabricated information that mimics news media content in form but not in organizational process or intent" (Lazer et al. 2018). Media analyses have found strong structural similarity between "fake news" and traditional journalism (Mourão and Robertson 2019; Tandoc Jr, Thomas, and Bishop 2021); "fake news" is thus understood as a form of "genre blending" that combines "elements of news ideals with features exogenous to the normative model of professional journalism: misinformation, sensationalism, clickbait, and bias" (Mourão and Robertson 2019).

Although some scholars have moved away from the term "fake news" given its use by politicians and others to undermine news that they perceive to be disagreeable, we follow others (Mourão and Robertson 2019; Grinberg et al. 2019; Tandoc Jr, Thomas, and Bishop 2021) in retaining the "fake news" label to refer to media sources that publish false, misleading, hyperpartisan, and sensational content. We also use the term "misinformation" to refer to media content that is false or misleading, and use labels provided by NewsGuard (https://www.newsguardtech.com/) to operationalize it. NewsGuard is a company that employs expert journalists and follows journalistic norms to rate the credibility of news and information websites.

Scholars, policymakers, and media professionals continue to make efforts to stem the flow of misinformation (Lazer et al. 2018). One potential technological approach entails the use of algorithms to detect misinformation online. Misinformation detection has been studied in two broad ways: by analyzing the content of misinformation and by analyzing the spread of misinformation. The former examines content-related features (e.g., textual and images properties), and the latter focuses on network features of the spreading phenomena in assessing the authenticity of the news articles (Zhou and Zafarani 2020; Singh, Ghosh, and Sonagara 2021; Potthast et al. 2017).

In previous literature, the credibility of news has typically been estimated based on the general reputation of the source concerning known fact-checked claims. Most works on misinformation detection which relied on source-level labels assumed that all articles from the source are equally reliable or unreliable depending on the reputation of the medium (Horne et al. 2018; Grinberg et al. 2019). For example, (Horne et al. 2018) trained machine-learning algorithms to predict whether the news article is coming from a factual or unreliable source. In their work, the assumption was that all news articles from a given source share the same credibility level. We also follow this approach by evaluating misinformation at the level of news sources rather than at the level of individual news articles. We do this for two reasons. First, the source of the news article often determines the presence of misinformation, based on the processes of the publisher (Grinberg et al. 2019). According to (Lazer et al. 2018), media outlets that publish false news tend to lack strong journalistic "norms and processes for ensuring the accuracy and credibility of information." Secondly, there is a dearth of fine-grained labels which have been defined at the news article level. State of the art creation of machine learning algorithms for misinformation detection focuses on source level labels for misinformation (Nørregaard, Horne, and Adalı 2019; Grinberg et al. 2019; Sitaula et al. 2020); hence, it makes sense for the bias analysis to be undertaken at the same resolution.

Political Asymmetry in Misinformation Dissemination of and susceptibility to misinformation occurs asymmetrically across the political spectrum. Research by (Faris et al. 2017) points to a broader asymmetry in the U.S. political media ecosystem that bears on the flow and uptake of misinformation on the political right . Their analysis finds that conservative media are more partisan and more insular than left-leaning media. This relative insularity allows for misinformation and misleading claims from politically extreme sites to more easily receive amplification and legitimation within the right-wing media sphere.

Political psychology research points to additional asymmetries in the media use and preferences of liberals and conservatives. Studies have shown that how physically threatened or fearful an individual feels is one of the key factors that predicts whether an individual holds conservative political attitudes (Napier et al. 2018; Jost et al. 2017). In turn, research has found that liberals and conservatives are drawn to different types of political media based on their psychological characteristics; for example, conservatives are attracted to information that monitors for threats and is aggressive in tone (Young 2019). Other research has demonstrated that liberals and conservatives are swayed by different features of persuasive message appeals (Jost and Krochik 2014). Therefore, it is plausible that misinformation targeting conservative and liberal audiences may use different mechanisms (e.g., news frames, emotional appeals, linguistic attributes, etc.) based on assumptions about what might appeal to and influence the intended audience.

Given the ideological asymmetries in the production, proliferation, and interpretation of political misinformation, it is possible that algorithms could be biased in their detection of misinformation in left- versus right-leaning news. In other words, potential differences between left- and right-leaning news may have implications for the fairness of algorithms to detect misinformation. Thus, any solutions for combating misinformation must take ideological asymmetry into account (Lazer et al. 2018).

Algorithmic Fairness The potential issues of the (un)fairness of machine learning algorithms are not limited to misinformation detection algorithms. Various algorithms are applied to make important decisions that were made by humans. However, even with the best intentions, data-driven machine learning algorithms can inherently reflect existing social biases or introduce new ones. The emerging literature on fair machine learning algorithms has identified multiple ways that the algorithms can make a discriminatory

decision. Some of the common scenarios for algorithmic bias include 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 (deliberately or inadvertently) designed to yield unequal decisions (Kamishima et al. 2012; Lepri et al. 2018). Accordingly, techniques to mitigate algorithmic bias attempt to modify the process of the training data (pre-processing), the learning algorithms (in-processing), and the prediction (post-processing) (Lepri et al. 2018).

More specifically, pre-processing techniques focus on optimizing the data before it goes into any algorithms. For instance, disparate impact remover tweaks feature values to increase fairness while preserving rank-order within the group. In this way, it allows each of the considered groups (e.g., political-left and political right) to have equal opportunities to score high on the considered features (Feldman et al. 2015).

To counter the bias stemming from the algorithm processes themselves, an in-processing technique such as adding a "regularizer" can be used. Regularizer acts as a prejudice remover by penalizing discriminatory outcomes generated by an algorithm (Kamishima et al. 2012). Similarly, with the adversarial debiasing technique, classifiers learn to minimize an adversary's ability to predict sensitive features (Alasadi, Al Hilli, and Singh 2019).

Post-processing techniques are often used when it is impossible or undesirable to change the incoming data or the (potentially proprietary) algorithms. Reject Option Classification approach, one of the post-processing techniques, applies rejection options and labels instances to reduce discrimination (Kamiran et al. 2018). More specifically, it tries to balance the outcomes of algorithms by giving desirable outcomes to the unprivileged group and giving undesirable outcomes to the privileged group in the "critical region" i.e., the area near the decision boundary.

Regardless of how algorithmic bias is mitigated, the above techniques share a common idea: responsibility for the social impact generated by algorithmic decision-making. As algorithms are products that involve both human and machine learning, redressing the potential bias inherent in the algorithms is a step forward toward accountable machine learning systems (O'neil 2016).

The issue of human bias in identifying misinformation has been addressed in some of the previous literature. For example, Babaei et al. (Babaei et al. 2021) have investigated the biases in the human process of identifying fake news. Raza et al., (Raza, Reji, and Ding 2022) have identified ways to identify bias (e.g., use of gendered language) within articles and ways to reduce them.

However, *no research has been done on political asymmetry in the algorithms automatically predicting misinformation*. In this work, we address the problem of identifying and reducing discriminatory decisions made by misinformation detection algorithms based on the political identity of the news source. Particularly, we focus on the scenario where there is an unequal representation of politically left- and right-aligned sources in the training data and hence potentially impact fairness in misinformation detection algorithms.

Materials and Methods

We used the NELA-2018 dataset (Nørregaard, Horne, and Adalı 2019) for analyzing bias in misinformation classification. The original dataset contained 713k news articles with source-level labels for credibility and political leaning, compiled from several data sources including NewsGuard, Pew Research Center, Wikipedia, BuzzFeed, and others.

We relied on credibility labels from NewsGuard for multiple reasons. First, NewsGuard's labels cover the highest percentage (35%) of the entire sample. Second, it follows a rigorous labeling process. NewsGuard (https://www.newsguardtech.com/) utilizes trained journalists rather than algorithms to assess the credibility and transparency of news websites. Their analysis creates a points system across 9 dimensions to derive an overall label for credibility. In addition, they allow respective news outlets to comment on the assigned ratings before making them public. Finally, NewsGuard's methodology has been used in recent misinformation detection literature (Nørregaard, Horne, and Adalı 2019; Singh, Ghosh, and Sonagara 2021).

Similarly, we relied on BuzzFeed's labels for political leaning (i.e., left vs. right), as the labels cover 36.3% of the sample in the dataset. We excluded news articles that did not have labels from both BuzzFeed and NewsGuard, leaving 102k articles for further analysis.

The dataset includes political leaning as a sensitive feature (i.e., the dimension to be considered for fairness), having two categorical values (left-aligned and right-aligned). Out of 102k data points, 37.5k points (approximately 36.7%) belonged to left-aligned sources. The remaining 64.5k points belong to right-aligned sources. Table 1 presents a list of news sources in the dataset, their respective number of articles, political leanings, and credibility (e.g., real/fake).

Feature Design

Machine learning literature suggests two major approaches to feature extraction: deep learning and hand-crafted approaches. With large computational resources such as a large dataset, recent deep learning approaches can yield high accuracy. However, they often work as black boxes and do not provide interpretability for the feature extraction (Zihni et al. 2020). Meanwhile, hand-crafted machine learning approaches are often designed by domain experts. These approaches are more interpretable because the role of individual features is more obvious compared to the deep learning approaches. Furthermore, they tend to work well even with the modest size of data and computational resources available. We consider interpretability as an important aspect of our work on fairness in machine learning algorithms, and hence follow the route of theory-driven and hand-crafted feature extraction approaches.

The features were identified based on a combination of 1) the concepts of journalistic news values, 2) relevant theories, and 3) the array of recent empirical studies on misinformation detection.

News Source	Number of Articles	Political Alignment	Real/Fake
Shareblue	2134	Left	Fake
MotherJones	1128	Left	Real
Alternet	4816	Left	Real
Politicus USA	4018	Left	Real
Palmer Report	3539	Left	Fake
Crooks and Liars	2465	Left	Real
Salon	1702	Left	Real
MediaMattersforAmerica	2316	Left	Real
Bipartisan Report	4060	Left	Fake
MSNBC	6604	Left	Real
Raw Story	3719	Left	Real
Daily Kos	994	Left	Fake
Drudge Report	18884	Right	Fake
FrontPage Magazine	892	Right	Fake
Instapundit	15479	Right	Fake
Breitbart	1877	Right	Fake
Fox News	3106	Right	Real
CNS News	5263	Right	Real
News Busters	3240	Right	Real
Infowars	2518	Right	Fake
Bearing Arms	1193	Right	Real
National Review	5129	Right	Real
Real Clear Politics	7206	Right	Real
Daily Signal	308	Right	Real

Table 1: News sources, number of articles, political alignment, and credibility.

Journalistic news values refer to journalists' "shared operational understanding that informs the mediated world that is presented to news audiences" (Tandoc Jr, Thomas, and Bishop 2021). We focused on deviations from journalistic news values (e.g., objectivity, balance) (Tandoc Jr, Thomas, and Bishop 2021; Lazer et al. 2018) when identifying relevant features in misinformation detection algorithms.

In addition, theories in psychology and social science (e.g., social identity theory, four-factor theory, Undeutsch hypothesis, information manipulation theory) grounded our selection of the features. Undeutsch hypothesis suggests that the factual statement differs from a fabricated or fictitious statement in content style and in quality (Amado, Arce, and Fariña 2015). Information manipulation theory states that extreme information quantity exists in deceptive statements (McCornack 1992). According to the four-factor theory (Zuckerman, DePaulo, and Rosenthal 1981), lies are expressed differently in terms of arousal, emotion, and thinking from the truth. Social identity theory suggested that awareness of one's group membership justifies maintaining social distance from the out-group, and this social distance is explained by the feeling of less acceptance, trust, or liking of the out-group members (Ashforth and Mael 1989). This out-group animosity is a powerful predictor of the sharing of political misinformation. Similarly, in-group favoritism is also used in misinformation because individuals tend to see what is favorable to their partisan orientation (Rathje, Van Bavel, and van der Linden 2021).

These theories helped us consider the difference between reliable contents and deceptive contents in terms of linguistic style of the texts (Amado, Arce, and Fariña 2015; McCornack 1992), subjectivity (Amado, Arce, and Fariña 2015), emotional expressions (Zuckerman, DePaulo, and Rosenthal 1981), and social identity manifestations (Ashforth and Mael 1989).

Finally, after reviewing the relevant empirical studies, we identified four broad categories of features: structure, subjectivity, sentiment, and social identity. The detailed explanations of the four categories of features are present below:

1. **Structure**: this category consists of the features describing the organization of the content into different stylistic structures, such as the syntax, text style, and grammatical elements of news content and title. Following the theories of the Undeutsch hypothesis and information manipulation theory and an empirical study (Zhou et al. 2020), we used features such as "number of words," "average words per sentence," and "number of question marks" and complexity measures (e.g., Flesch-Kincaid readability index). Complexity measures were computed using the Textstat Python library, and the other features were computed using LIWC (Pennebaker et al. 2015).
2. **Subjectivity**: this category consists of the features that provide evidence of an effort to convey a certain opinion or viewpoint rather than facts. The category of subjectivity was considered as the deviation from the journalistic news value of "objectivity," impartially pursuing the evidence and demonstrating faith in "facts" (Schudson 1981). Also, Undeutsch hypothesis theoretically grounds that a fictitious statement differs in quality from a true statement. A recent study added empirical evidence to the theory that more than half of the articles published by unreliable news sites contained the personal opinion of the author(s) (Tandoc Jr, Thomas, and Bishop 2021). Therefore, we used "cognitive processes" (e.g. cause, know, ought), "perceptual processes" (e.g., look, heard, feeling), and "informal language" (e.g., swear words) categories and their subcategories from LIWC 2015 (Pennebaker et al. 2015) as the features in this category.
3. **Sentiment**: this category refers to the emotion-arousing aspects of the news stories that contain misinformation. Theoretically, lies are expressed differently in terms of arousal, emotion, and thinking from the truth (Zuckerman, DePaulo, and Rosenthal 1981). Empirically, sentiment features such as positive words, negative words, exclamation marks, and sentiment polarity were used in machine learning algorithms to detect misinformation (Bond et al. 2017; Zhou et al. 2020). Based on the four-factor theory and the empirical studies, we included the "emotional tone" (i.e., positive, neutral, and negative emotions) (Hutto and Gilbert 2014) and "affective processes" category (e.g., happy, cried) and its subcategories (e.g., anxiety, sadness) from LIWC 2015 (Pennebaker et al. 2015)

as the sentiment feature to detect misinformation.

4. **Social Identity**: this category consists of the features that reveal the qualities or beliefs that make a particular group different from others. Social identity theory (Ashforth and Mael 1989) and prior literature confirmed that readers are more easily persuaded by the use of social identity words, such as second-person pronouns (e.g., you), that are unlikely to appear in reliable news articles (Horne and Adali 2017; Singh, Ghosh, and Sonagara 2021). Following the theory and the previous literature, we selected "personal pronouns" category and its subcategories from LIWC 2015 (Pennebaker et al. 2015), "liberal identity words" (e.g., left-wing, democrat), "conservative identity words" (e.g., right-wing, republican), and "moral words" (e.g., guilt, innocent, blame) from the dictionaries created by (Osmundsen et al. 2021).

Table 2 shows the summary of the features that were used in this study. Note that some of the features did not fall exclusively under one category. In this case, we organized the features under the most relevant category in the context of the current study. In addition, we computed separate values of each feature for the title and the body of the news articles. This is because not only the content of the news articles but also the news titles are a strong factor in distinguishing misinformation from reliable news (Horne and Adali 2017). Given that the news titles are short in length, some structural features (e.g., number of paragraphs) were computed only for the body of the news article.

Pre-processing and Model Training

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.

The obtained features were passed to several classification algorithms. Based on a survey of recent literature on machine learning algorithms for misinformation detection (Singh, Ghosh, and Sonagara 2021; Fang et al. 2019; Shu et al. 2020), we considered five frequently used algorithms: Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Multilayer Perceptron Neural Networks. Each of these was implemented using the Sci-kit Learn library (Pedregosa et al. 2011).

We split the dataset in a random manner where 80% of the dataset was set aside for training the machine learning algorithms and the remaining 20% was utilized as a test-set. The split is performed in a stratified manner using labels. This ensures that the proportion of real and fake news articles in train and test sets are same. We ran all the algorithms for 100 iterations and the Random Forest approach yielded the highest accuracy (see Table 3; details follow). Since there is often a trade-off between accuracy and fairness (i.e., as fairness increases, accuracy decreases), the model with the highest accuracy (Random Forest) was picked as the baseline model for further inspection of its fairness towards the sensitive feature (i.e., political leaning). The Random Forest model has 100 estimators/number of trees in the forest with no limits on the maximum depth of the forest.

Auditing for Bias

We selected 'political leaning' as the sensitive feature in this work. Given multiple recent claims that media platforms favor articles from politically left-leaning sources in moderation (Cummings 2018; Koebler and Cox 2018), we consider articles from right-aligned sources to be the unprivileged group when empirically auditing the algorithms for bias. Note that this sensitive feature is not part of the training data but is used at test time to compute various fairness metrics.

There are multiple interpretations of algorithmic fairness such as maximizing utility for groups or respecting various rules such as individual rights and freedoms (Duster 1996). In the current study, we follow John Rawl's interpretation of distributive justice which equates fairness and justice, arguing broadly that fairness is "a demand for impartiality" (Rawls 1999). In other words, the algorithm should yield similar outcomes for different groups irrespective of their demographic description. Focusing on the notion of distributive justice, we consider an algorithm to be fair if its performance does not vary for news articles from the politically left- and right-aligned news sources.

There exist at least two different interpretations within the above mentioned distributive justice paradigm to quantify bias. One approach focuses on equal predictive performance, i.e., an equal ability to identify the "ground truth" labels for the two classes. We consider one such metric: delta accuracy, in this work. It is easy to interpret and follows naturally from the traditional accuracy metric, which remains an unquestionable goal for any classification system.

The other interpretation focuses on the concept of "disparate impact," i.e., when a facially neutral practice has an unjustified adverse impact on members of a protected class (Civil Rights Act 1964). This approach questions the validity of past "ground truth" data that is used to train algorithms. For instance, while using SAT scores to decide on college admissions may "appear" to be objectively fair, it is near impossible to tease apart the impact of systemic injustices which yield poorer SAT scores for under-represented minorities. Hence, *irrespective* of the learning data and the learning process, this interpretation would require that different demographic groups have equal probability of positive outcomes. In fact, in the US legal system, a process is considered biased, irrespective of the intent of the designers, if there is less than 0.8 under-representation in the probability of positive outcome for a demographic group (Civil Rights Act 1964; Feldman et al. 2015). We consider two related metrics based on this line of reasoning: disparate impact and statistical parity difference, in this work.

Delta accuracy Delta accuracy indicates the difference in the accuracy of samples belonging to the privileged and unprivileged groups. If delta accuracy is not zero, it means that the algorithm is classifying more accurately on one group of samples than on the other.

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

Types of Features	Features	Theoretical Support	Empirical Support
Structure	**Title and Body**: Number of words, Number of nouns, Number of verbs, Number of Adverbs, Number of Exclamation marks, Number of question marks, Number of quotation marks, complexity **Body only**: Number of sentences, Number of paragraphs, Average sentences per paragraph, Average words per sentence, Average punctuation per sentence, Average characters per word	Undeutsch Hypothesis (Amado et al., 2015), Information Manipulation Theory (McCornack, 1992)	Zhou et al., 2020, Horne & Adali, 2017
Subjectivity	Cognitive processes, Perceptual processes, Informal language	Undeutsch Hypothesis (Amado et al., 2015)	Tandoc et al., 2021
Sentiment	Affective processes, Positive/Neutral/Negative sentiment	Four-factor theory (Zuckerman, 1981)	Bond et al., 2017, Zhou et al., 2020
Social Identity	Personal pronouns, moral words, liberal identity words, conservative identity words	Social Identity Theory (Ashforth, 1989)	Singh et al., 2020, Rathje et al., 2021

Table 2: Summary of features that were used to build machine learning models

where acc is accuracy and S is the sensitive feature.

Disparate impact Disparate impact (DI) captures the ratio of the probability of favorable outcomes being assigned by the algorithm for the unprivileged group compared to that of the privileged group. Ideally, the value of the disparate impact needs to be 1.0.

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

Statistical parity difference Statistical parity difference (SPD) calculates the difference in the probability of favorable outcomes obtained by the unprivileged group to that of the privileged group. A favorable outcome in the considered setting would be to get assigned a "real" label (as opposed to a "fake" label) for the news article. For an ideal fair model, the statistical parity difference is expected to be zero.

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

Following recent literature on fairness in machine learning, bias audit was undertaken via a statistical t-test on the means of the abovementioned fairness metrics for the two (left-aligned and right-aligned) groups (Alasadi, Al Hilli, and Singh 2019; Singh and Hofenbitzer 2019).

Bias Reduction Approach

Reject Option Classification (ROC) proposed by (Kamiran et al. 2018), was used as a bias reduction approach in this study. ROC is a post-processing algorithm that makes the pre-trained classifier discrimination-aware at the time of prediction. It is useful for a broad range of applications (Kamiran et al. 2018; Iqbal, Karim, and Kamiran 2019), as it does not require any changes in the classification algorithm, nor does it amend or pre-process the dataset before applying the classification algorithm.

ROC labels the instances from the unprivileged groups that lie in the *critical region* (i.e., near the decision boundary in which labels are difficult to identify) as desirable labels. Similarly, the instances belonging to privileged groups that lie in the critical region are assigned an undesirable label. A classifier that predicts the posterior probability p(Y|X) for an instance X closer to 1 or 0, assigns the label with confidence. However, if the same classifier predicts the posterior probability closer to 0.5, it gets into the dilemma of deciding the appropriate label. If an instance belonging to the unprivileged group lies in the *critical region*, then that label is assigned a positive label (Y^+), otherwise, it is assigned a negative label (Y^-). The rest of the instances belonging to the unprivileged group that lie outside the critical region are classified as per usual; meaning that if the posterior probability of P($Y^+|X$) is greater than P($Y^-|X$), then the instance is classified as positive, otherwise, it is classified as negative.

Using the IBM AIF360 library (Bellamy et al. 2018), we implemented the ROC algorithm (optimization metric = statistical parity difference) and ran the classification algorithm 100 times with each iteration having a shuffled version of the dataset.

Note that a classification algorithm is considered to have become less biased if there are changes in the metrics for bias defined above. Specifically, a less biased algorithm will yield *reduced* delta accuracy and statistical parity difference, and the disparate impact score will get closer to 1.0.

Results

Misinformation Classification Results with Different Algorithms

Table 3 summarizes the average results obtained after 100 rounds of experiments for the different algorithms considered. As can be seen, the best performing automated machine learning algorithm (Random Forest) achieved 87.85% accuracy at automatically classifying the credibility of the news article. Random Forests outperforming other algorithms is consonant with trends reported in the past literature, and the obtained accuracy results are also within the range reported in the current state of the art in misinformation detection (Singh, Ghosh, and Sonagara 2021; Zhou and Zafarani 2020).

The three most predictive features for the classification algorithm remained consistent over the 100 iterations. They were the neutrality (sentiment) of the article, the number of words in the title, and the number of words in the article.

The mean values for these features for the four groups: left+true, right+true, left+false, and right+false are shown in Figure 1. The differences in the values for these features across true and false news appear to be different depending on the political leaning of the article. For instance, while right+true articles were more neutral toned than left+true articles, right+false articles were less neutral toned than the left+false articles. Similar transposition occurs in the case of "number of words in news articles," while the effect is less pronounced in the case of "number of words in title." These results suggest that political leaning can play an important role in mediating the production of misinformation.

Auditing Misinformation Detection Algorithms for Political Bias

The accuracy levels achieved with different algorithms when focusing only on the left-aligned and right-aligned sources, and the delta between them, are also shown in Table 3. As can be seen, the level of accuracy varied across algorithms but there was no clear trend of the accuracy being better for the left-leaning or right leaning articles. The accuracy was higher for right-leaning articles when using two algorithms while it was lower in other three algorithms. Importantly, all such differences were below 4%.

On the other hand, we found a strong deviation from the ideal values of 1.0 and 0% respectively in terms of Disparate Impact and Statistical Parity Difference metrics. The DI scores were consistently below the legally required level of 0.8, indicating that the unprivileged group (political right) had significantly lower odds of getting a positive label from the algorithm. A statistical t-test comparing the obtained DI values with the ideal value of 1.0 showed that the results were statistically significantly different from the ideal values (p-values $< .001$) for all five algorithms.

The trends were similar in terms of Statistical Parity Difference. All five algorithms provided noticeably lower probability of positive outcomes for the political right (values ranged from -16% to -25%). The observed SPD values were statistically significantly different from the ideal value of 0 (p-values $< .001$) for all five algorithms.

In summary, while there were minor differences in terms of accuracy across left and right, there were significant differences between the odds of a positive outcome being assigned to a news article based on its political leaning. These results will not be able to meet the legal standards of parity expected (Feldman et al. 2015) and hence, it is important to reduce this bias.

Reducing Algorithmic Bias in Misinformation Detection

We utilized the ROC post-processing method for bias reduction and the various fairness metrics *after* that process are reported in Table 4. For comparison, the results *before* bias reduction are also presented.

The most noticeable differences were those in Disparate Impact and Statistical Parity Difference fairness metrics. For Disparate Impact (ideal value = 1.000), the value moved from 0.6038 (before) to 0.9340 (after). For Statistical Parity Difference (ideal value = 0%), the value changed from -25.29% to -3.86%. Both these changes were large improvements in terms of fairness and were found to be statistically significant based on a t-test (p-values $< .001$).

At the same time, there was a modest dip in overall accuracy from 87.85% to 82.67%. There was also a small increase in the absolute value of delta accuracy from 1.36% to 1.98%. However, given that accuracy stays above 80% and delta accuracy stays below 2%, we consider these changes to be reasonable trade-offs for the much bigger improvements obtained in terms of the DI and SPD fairness metrics.

Discussion

This study reported on the performance of the misinformation detection algorithm in terms of the political leaning of the news source and the potential to reduce the discriminatory performance of the algorithm using the ROC technique.

The first research question in this work was: *(RQ1) Are misinformation detection algorithms susceptible to bias in terms of political leaning?*

As shown in Table 3, the results indicated that the multiple misinformation classification algorithms performed differently based on political leaning of the news source. While the differences in terms of delta accuracy were small (<4%), there were noticeable differences in terms of Disparate Impact and Statistical Parity Difference.

While prior literature motivated an audit on the relative performance of misinformation detection algorithms across political left and right, the precise nature of bias found was unexpected. The bias was found to be lot more noticeable in terms of some metrics (DI, SPD) and not as noticeable or consistent in terms of others (delta accuracy). This motivates future work with more fine-tuned hypothesis development in the space of algorithmic bias in political information.

A potential reason for the observed difference in the probabilities of the positive outcome for left and right might lie in the skew present in the dataset used. This work used one of the largest datasets available for this analysis ($> 100k$ articles with labels for political leaning and true/false news), but the "ground truth" labels had a nearly 80% higher probability of true labels for the left leaning articles.

This could partially be a function of the reported higher prevalence of fake news on right leaning channels (Faris et al. 2017). At the same time, conservatives already are concerned about bias in political fact-checks (Shin and Thorson 2017), content moderation (Usher 2018), and platform regulation (Darcy 2021). For instance, the suspension of the (then) US President Donald Trump from Facebook was considered unfair by multiple stakeholders, and underscores the need to build approaches that are fair and auditable by third-parties (Darcy 2021). Hence, it is unlikely that the conservative population will be accepting of a fake news detector that has a noticeable higher probability of assigning positive outcomes for left-leaning articles. In fact, the value of well-created algorithms lies in being able to create equitable algorithms despite having to work with skewed datasets. This is a challenge, and multiple bias reduction approaches have been proposed in recent literature (Bellamy et al. 2018).

The second research question in this work was: *(RQ2)*

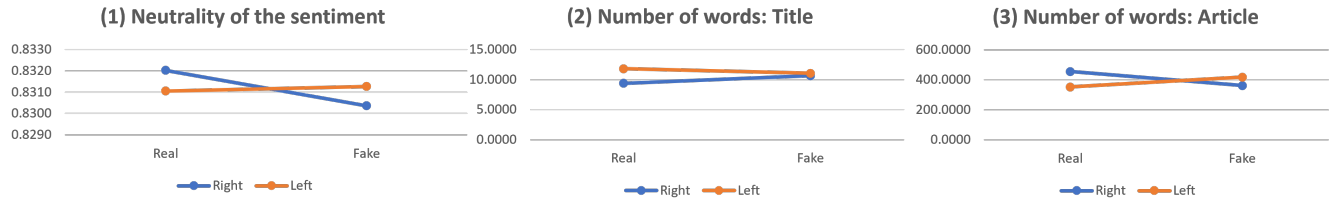

Figure 1: Average values for the four groups (true+right, true+left, false+right, false+left) for the three most predictive features in the misinformation classifier.

Method	Accuracy	Left Accuracy	Right Accuracy	Delta Accuracy	Disparate Impact	Stat. Parity Difference
Random Forest	**87.85%**	86.98%	88.34%	-1.36%	0.6038	-25.29%
Decision Tree	81.05%	82.19%	80.40%	1.79%	0.6485	-22.86%
Logistic Regression	63.89%	61.49%	65.27%	3.78%	0.7319	-16.06%
Multi Layer Perceptron	78.30%	79.47%	77.62%	1.86%	0.6709	-20.93%
Support Vector Machine	64.04%	62.05%	65.19%	-3.14%	0.7437	-15.56%

Table 3: The average accuracy and fairness levels for various models used for misinformation detection after 100 iterations.

Metric	Ideal Value	Before Bias Reduction	After Bias Reduction
Accuracy	100.00%	87.85%	82.67%
Delta Accuracy	0.00%	-1.36%	-1.98%
Disparate Impact	1.0000	0.6038	0.9340
Statistical Parity Difference	0.00%	-25.29%	-3.86%

Table 4: Comparison of delta accuracy, statistical parity difference and disparate impact *before* and *after* bias reduction processing. Average values after 100 iterations using Random Forest classifier.

Can the level of bias in misinformation detection algorithms be reduced while maintaining accuracy?

Based on the observation in the considered dataset, we found that the ROC bias-reduction approach is effective in reducing the disparity in the performance of misinformation detection algorithms across the political leaning of the news source. As shown in Table 4, in terms of DI and SPD (the fairness metrics with noticeable issues in the *before* condition), there was a noticeable improvement in fairness upon applying the ROC bias-reduction approach, and the level of accuracy was still above 80%. Given that the trade-offs between fairness and accuracy are common in similar studies, a modest decrease in accuracy with significant improvements in bias reduction was considered reasonable (Pessach and Shmueli 2020).

To increase public confidence in misinformation detection practices and subsequent corrections, it is critical to establish fairness in misinformation detection algorithms, toward which this study makes an important first contribution. Moreover, improving the fairness of misinformation detection algorithms helps ensure that the practices used to deter the spread of misinformation are not inadvertently exacerbating existing asymmetries in the political media environment or introducing new ones.

The current work has some limitations. It focused on source-level labels for misinformation and political leaning. Yet not all articles published by a particular source are uniformly reliable vs. unreliable (Mourão and Robertson 2019) or conservative-leaning vs. liberal-leaning. In future research, it will be important to explore and extend these results by developing article-level labels. Also, we acknowledge that there are other approaches to build misinformation algorithms and reduce bias than discussed in this work. For instance, going beyond textual features, the misinformation detection algorithms can also use image features or networked propagation features for improving accuracy. Hence, the work undertaken cannot be considered a final word in this space. Rather, its contribution lies in motivating and grounding a new research direction: political bias audit for misinformation detection algorithms and identifying ways to reduce such bias.

Conclusion

This paper grounds the use of political-leaning as a sensitive feature to study fairness in misinformation classification algorithms. The audit of the existing misinformation classification algorithm revealed that the probability of obtaining a positive outcome from the algorithm varied significantly depending on the political leaning. This disparity in the performance was found to be reduced noticeably after the application of a bias reduction algorithm (ROC) without modifying the discriminatory data or tweaking a specific classification algorithm. The results significantly move forward the literature on misinformation classification, particularly with po-

litical leaning as a sensitive feature. Future work could consider reducing bias in misinformation detection algorithms at the article-level, and a newer approach to creating fair and accurate misinformation news classification algorithms to develop and maintain trustworthy online environments.

The results also have implications for our understanding of political media. The disparity in the algorithm's performance points to potential differences in the structure and features of conservative versus liberal fake news articles and/or in their overlap with traditional news. This offers an important area for future research in media and journalism studies.

References

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

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

[Amado, Arce, and Fariña 2015] Amado, B. G.; Arce, R.; and Fariña, F. 2015. Undeutsch hypothesis and criteria based content analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context* 7(1):3–12.

[Ashforth and Mael 1989] Ashforth, B. E., and Mael, F. 1989. Social identity theory and the organization. *Acad. of management review* 14(1):20–39.

[Babaei et al. 2021] Babaei, M.; Kulshrestha, J.; Chakraborty, A.; Redmiles, E. M.; Cha, M.; and Gummadi, K. P. 2021. Analyzing biases in perception of truth in news stories and their implications for fact checking. *IEEE Transactions on Computational Social Systems*.

[Bellamy et al. 2018] Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

[Bond et al. 2017] Bond, G. D.; Holman, R. D.; Eggert, J.-A. L.; Speller, L. F.; Garcia, O. N.; Mejia, S. C.; Mcinnes, K. W.; Ceniceros, E. C.; and Rustige, R. 2017. 'lyin'ted','crooked hillary', and 'deceptive donald': Language of lies in the 2016 us presidential debates. *Applied Cognitive Psychology* 31(6):668–677.

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

[Calmon et al. 2017] Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, 3992–4001.

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

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

[Cummings 2018] Cummings, W. 2018. Diamond and silk tell congress, 'facebook censored our free speech!,'. *USA Today*. Available online: `https://bit.ly/3r6FsJp`.

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

[Duster 1996] Duster, T. 1996. Individual fairness, group preferences, and the california strategy. *Representations* 55:41–58.

[Fang et al. 2019] Fang, Y.; Gao, J.; Huang, C.; Peng, H.; and Wu, R. 2019. Self multi-head attention-based convolutional neural networks for fake news detection. *PloS one* 14(9):e0222713.

[Faris et al. 2017] Faris, R.; Roberts, H.; Etling, B.; Bourassa, N.; Zuckerman, E.; and Benkler, Y. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication* 6.

[Feldman et al. 2015] Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proc. ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

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

[Grinberg et al. 2019] Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on twitter during the 2016 us presidential election. *Science* 363(6425):374–378.

[Guess et al. 2021] Guess, A. M.; Barberá, P.; Munzert, S.; and Yang, J. 2021. The consequences of online partisan media. *Proceedings of the National Academy of Sciences* 118(14).

[Horne and Adali 2017] Horne, B., and Adali, S. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the AAAI Conference on Web and Social Media*, volume 11.

[Horne et al. 2018] Horne, B. D.; Dron, W.; Khedr, S.; and Adali, S. 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.

[Hutto and Gilbert 2014] Hutto, C., and Gilbert, E. 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*, volume 8.

[Iqbal, Karim, and Kamiran 2019] Iqbal, M.; Karim, A.; and Kamiran, F. 2019. Balancing prediction errors for robust sentiment classification. *ACM Trans. on Knowledge Discovery from Data (TKDD)* 13(3):1–21.

[Jiang, Robertson, and Wilson 2020] Jiang, S.; Robertson, R. E.; and Wilson, C. 2020. Reasoning about political bias in content moderation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13669–13672.

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

[Jost et al. 2017] Jost, J. T.; Stern, C.; Rule, N. O.; and Sterling, J. 2017. The politics of fear: Is there an ideological asymmetry in existential motivation? *Social cognition* 35(4):324–353.

[Kamiran et al. 2018] Kamiran, F.; Mansha, S.; Karim, A.; and Zhang, X. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences* 425:18–33.

[Kamishima et al. 2012] Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50. Springer.

[Koebler and Cox 2018] Koebler, J., and Cox, J. 2018. The impossible job: Inside facebook's struggle to moderate two billion people - motherboard. *Motherboard*. Available online.`https://motherboard.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works`.

[Lazer et al. 2018] Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science* 359(6380):1094–1096.

[Lepri et al. 2018] Lepri, B.; Oliver, N.; Letouzé, E.; Pentland, A.; and Vinck, P. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31(4):611–627.

[McCornack 1992] McCornack, S. A. 1992. Information manipulation theory. *Communications Monographs* 59(1):1–16.

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

[Napier et al. 2018] Napier, J. L.; Huang, J.; Vonasch, A. J.; and Bargh, J. A. 2018. Superheroes for change: Physical safety promotes socially (but not economically) progressive attitudes among conservatives. *European Journal of Social Psychology* 48(2):187–195.

[Nørregaard, Horne, and Adalı 2019] Nørregaard, J.; Horne, B. D.; and Adalı, S. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 630–638.

[O'neil 2016] O'neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[Osmundsen et al. 2021] Osmundsen, M.; Bor, A.; Vahlstrup, P. B.; Bechmann, A.; and Petersen, M. B. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *Am. Political Science Review* 115(3):999–1015.

[Pedregosa et al. 2011] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

[Pennebaker et al. 2015] Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of liwc2015. Technical report.

[Pessach and Shmueli 2020] Pessach, D., and Shmueli, E. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

[Potthast et al. 2017] Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

[Rathje, Van Bavel, and van der Linden 2021] Rathje, S.; Van Bavel, J. J.; and van der Linden, S. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* 118(26).

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

[Raza, Reji, and Ding 2022] Raza, S.; Reji, D. J.; and Ding, C. 2022. Dbias: Detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics* 1–23.

[Schudson 1981] Schudson, M. 1981. *Discovering the news: A social history of American newspapers*. Basic books.

[Shin and Thorson 2017] Shin, J., and Thorson, K. 2017. Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication* 67(2):233–255.

[Shoemaker and Vos 2009] Shoemaker, P. J., and Vos, T. 2009. *Gatekeeping theory*. Routledge.

[Shu et al. 2020] Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8(3):171–188.

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

[Singh, Ghosh, and Sonagara 2021] Singh, V. K.; Ghosh, I.; and Sonagara, D. 2021. Detecting fake news stories via multimodal analysis. *Journal of the Assoc. for Information Science and Technology* 72(1):3–17.

[Sitaula et al. 2020] Sitaula, N.; Mohan, C. K.; Grygiel, J.; Zhou, X.; and Zafarani, R. 2020. Credibility-based fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*. Springer. 163–182.

[Tandoc Jr, Thomas, and Bishop 2021] Tandoc Jr, E. C.; Thomas, R. J.; and Bishop, L. 2021. What is (fake) news? analyzing news values (and more) in fake stories. *Media and Communication* 9(1):110–119.

[Usher 2018] Usher, N. 2018. How republicans trick facebook and twitter with claims of bias. *The Washington Post*. Available online: `https://wapo.st/3Jk7NSU`.

[Vosoughi, Roy, and Aral 2018] Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

[Wardle and Derakhshan 2017] Wardle, C., and Derakhshan, H. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report* 27.

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

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

[Zhou et al. 2020] Zhou, X.; Jain, A.; Phoha, V. V.; and Zafarani, R. 2020. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice* 1(2):1–25.

[Zihni et al. 2020] Zihni, E.; Madai, V. I.; Livne, M.; Galinovic, I.; Khalil, A. A.; Fiebach, J. B.; and Frey, D. 2020. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *Plos one* 15(4):e0231166.

[Zuckerman, DePaulo, and Rosenthal 1981] Zuckerman, M.; DePaulo, B. M.; and Rosenthal, R. 1981. Verbal and nonverbal communication of deception. In *Adv. in experimental social psychology*, volume 14. Elsevier. 1–59.