Pitfalls in Popular Misinformation Detection Methods and How to Avoid Them

by

Lia Bozarth

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2022

Doctoral Committee:

Professor Ceren Budak, Chair
Professor Mark Ackerman
Professor Kelly Garrett
Doctor Alexandra Olteanu
Professor Paul Resnick

Lia Bozarth
lbozarth@umich.edu
ORCID iD: 0000-0002-0879-4234

DEDICATION

This dissertation is dedicated to my parents, Jen and Tom. To my mom, who brought us to the United States so I can have a better life, and to my dad, who raised me as his own.

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge and express my deepest appreciation for my advisor Ceren Budak. This dissertation would not have come to completion without the six years of steady guidance and mentorship from her. I want to particularly thank Ceren for giving me the freedom and the encouragement to pursue topics and research questions that are of interest to me, and for her consistent reassurance and support when my graduate studies were stalling. I would also like to acknowledge my committee members for providing valuable insights along the way to the completion of this dissertation.

Next, I want to acknowledge and thank my cohort for their continuous support throughout my graduate studies. This includes giving valuable feedback on my research ideas, work-in-progress manuscripts, and academic/job presentations; answering the many questions I had about specific methods or research approaches; and, finally, contributing to a supportive and empathetic research environment. In alphabetic order, I thank Heeryung Choi, Jane Im, Harmanpreet Kaur, Brad Lott, Danaja Maldeniya, Jiaqi Ma, Megh Marathe, Christopher Quarles, Ashwin Rajadesingan, Allison Tyler, and Jeremy York. Similarly, I want to thank Professor Joyojeet Pal and Professor Kelly Garrett for their mentorship in the research projects we conducted together and for taking the time to assist me with obtaining gainful employment. Additionally, I want to acknowledge all of my smart and hardworking coauthors. Their significant contributions have made our publications possible.

Special thanks to my dearest friends, my basketball group, my Dungeons and Dragons group, and my running buddies. Their presence, friendship, and dependable support kept me balanced outside my academic work. In particular, I would like to acknowledge Alexis Alvarez, Helen Chao, Heeryung Choi, Carolina Chung, Natacha Comandante, Sangmi Kim, Danaja Maldeniya, Nirmala Maldeniya, Megh Marathe, Jessie Marie, Christopher Quarles, Ellen Quarles, Cindy Tung, Sherry Great, Jeremy York, Jin Yu, and Yulin Yu.

Finally, I would like to thank my partner Alexis Alvarez and our pets, Daj, Courage, and Ginsberg, for being my anchor. You are the best things that life can offer, and it's my greatest happiness that I get to share my life with you.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

Misinformation is a major challenge today, and much academic work has been done to study misinformation, including evaluating its prevalence, trend, behavior, and impact. Many scholars have also sought effective mitigation strategies to curtail the spread and influence of misinformation.

In this dissertation, we focus on misinformation detection (MID). We evaluate three popular MID approaches: i) expert labeling, ii) automated methods, and iii) crowd wisdom. For each approach, we first review its theorized strength and weaknesses, in addition to its existing and potential real-world applications. We then empirically evaluate the extent of its strength and weaknesses. Our studies identify shared caveats and potential improvements for some or all of the three approaches. Specifically, we demonstrate the need to include domain and task-specific performance evaluation and bias assessment procedures. Similarly, we show that the performance of some of these approaches can change significantly over time and amid external shocks. Additionally, we also reveal that the lack of transparency can have a direct impact on the actual and perceived usability of an approach. Moreover, we demonstrate that user preference for MID approaches is not always about performance and is also not fully rational. Finally, we synthesize our results and formulate concrete strategies to mitigate the observed caveats.

CHAPTER 1

Introduction

"Don't believe in everything you read on the internet just because there is a picture of a famous dude next to it—-George Washington, 1792"

Following the 2016 U.S. presidential election, misinformation swiftly becomes a major phenomenon and concern to the general public [233, 154], a critical obstacle to social media platforms [73, 1], and a rapidly growing research field for academia [145, 187]. Thus far, misinformation-related studies have extended to a wide range of domains, including civic engagement, health and healthcare, financial markets, and disaster/crisis management [208, 162, 191, 232, 159, 242, 141]. These ongoing research efforts include the conceptualization of misinformation [246, 264, 87, 78, 119, 285], detection [198, 230, 231, 215, 152, 229, 110, 39, 276], behavior and impact analysis [11, 260, 256, 100, 227], and finally, mitigation and remedy [147, 249, 257, 140, 34, 50, 252].

For this dissertation, we will focus on *misinformation detection*, which we will refer to as MID. The ability to reliably detect misinformation is a precursor to many subsequent studies, including behavior and impact evaluation [32, 31]. As such, a sizable amount of resources had gone into MID-related work. This includes expanding third-party fact-checking organizations [197, 200], designing advanced automated MID systems [198, 230, 229, 110, 39, 276], and utilizing the knowledge and opinion of laypersons [261, 196, 84]. Yet, the latest research suggests much more is needed to improve existing MID approaches. Notably, related work has highlighted several inadequacies in existing processes, including detection biases [30], speed of detection [153], scalability [196], and reliability over time [111]. These limitations likely hinder the real-world applicability of existing approaches. For example, despite social media giant Facebook's repeated attempt to find and curtail fake news on its site, recent work found that misinformation was widely present on the platform and generated billions of views during the 2020 U.S. presidential election [69]. This, at least in part, is likely due to Facebook's current MID approach, which is significantly reliant on external fact-checkers [266, 73]. Given these considerations, our goal is to conduct in-depth analyses of existing MID approaches, identify their pitfalls and associated consequences, and then

explore ways to improve existing practices such that these pitfalls can be mitigated.

Broadly, there are three popular MID approaches: 1) expert labeling, 2) automated methods, and 3) crowd wisdom-based detection. Each process has its unique advantages and drawbacks. Third-party fact-checkers have high expertise, but they differ in defining and conceptualizing misinformation [32]. Further, fact-checkers may preferentially fact-check certain content over others [32]. Similarly, fact-checker evaluation procedures are sometimes surprisingly opaque to outsiders, and may even change over time. Additionally, as mentioned previously, this approach also lacks scalability. Automated systems enable early detection and are highly scalable [110]. But, many existing models lack comprehensive evaluations of performance, fairness, accountability, and transparency [30]. Finally, the crowd wisdom-based approach may absolve platforms of some editorial duties. However, the availability and reliability of crowd wisdom remain an open question [3]. Thus far, work that rigorously examines the drawbacks and pitfalls in the existing MID approaches is limited. In practice, social media giants like Facebook have used a combination of crowd signals, AI models, and third-party fact-checkers for misinformation detection [73]. Yet, without in-depth analysis, unexplored and ill-understood limitations of these MID methods could lead to costly consequences. These include misinformation not being detected, biased detection (e.g., never-before-encountered fake news publishers being less likely to be detected), or the suppression of factual content. Additionally, public trust in platforms that utilize these approaches likely will weaken as a result [186]. Given these considerations, this dissertation aims to answer the following question: *"What are the meaningful pitfalls in the three popular MID approaches? More importantly, how can we improve upon existing practices to avoid these pitfalls?"*

1.1 Three Popular Misinformation Detection Approaches

Here, we provide a brief background on each of the three popular misinformation detection approaches (Figure 1.1). For each approach, we first describe its expected strength and possible weaknesses in addition to its existing and potential use cases. We then formulate targeted research questions to examine whether or not the theorized key strength and weaknesses of an approach are supported empirically.

1.1.1 Expert Labeling

Expert labeling refers to the labor-intensive approach of using domain experts (e.g., third-party fact-checking organizations) to manually review and verify content. Despite this approach's lack of scalability and timeliness, its proponents underscored the high accuracy of expert labels and their capacity to correct viewer misperceptions [184]. Expert labels are also expected to be highly

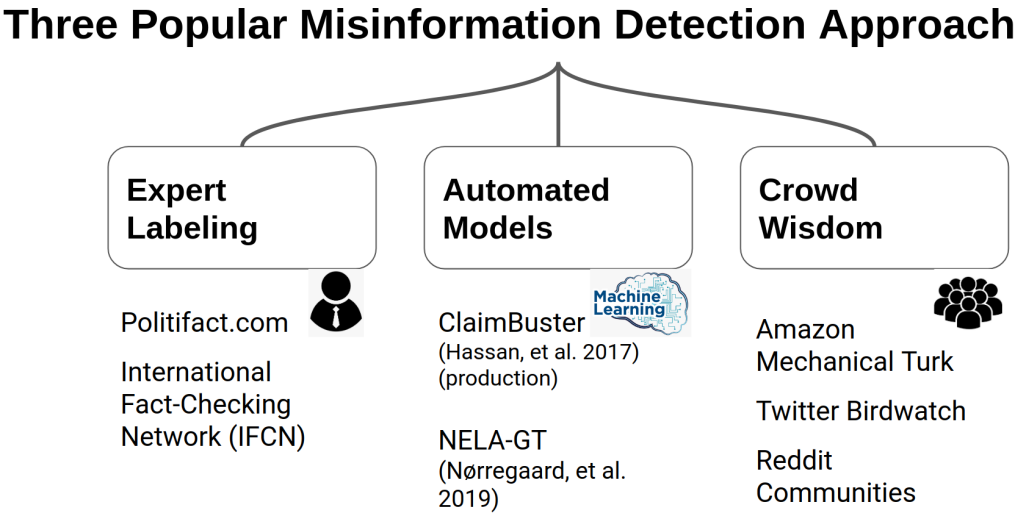


Figure 1.1: Three Popular Misinformation Detection Approaches.

correlated with each other. That is, high accuracy and high agreement are commonly considered the key strength of expert labeling. As such, labels from experts are commonly used as groundtruth in various downstream academic work [226, 107, 230, 92, 18, 262, 10, 11]. Yet, in practice, we note that even among the experts, there is currently no single consistent definition of misinformation [148, 264, 275]. Further, we also observe that academic scholars and third-party fact-checkers have their own annotation procedures and metrics to identify and label misinformation [285, 100]. Critics of expert labeling have similarly argued that there is lack of consistency in expert labels [184]. For example, past work argues that there is a lack of systematic content selection method (i.e., experts' tendency to pick and choose the entities/content they check). Additionally, critics have also expressed concerns about bias and transparency issues in expert labeling [184, 82]. For example, a Poynter study has shown that half of all Americans (and 70% of Republicans) thought that expert fact-checkers are ideologically biased and distrusted labels from fact-checking organizations [82]. Given these considerations, we ask the following questions.

1. To what extent do labels provided by experts differ? Further, what are some possible explanations for their differences?

2. Does groundtruth choice (i.e., choosing which expert labels to use) impact the results of downstream analyses? If so, to what extent?

The first question addresses concerns surrounding the lack of agreement between experts, and factors (e.g., biases) that potentially explain such differences. In other words, it speaks to the proposed strength and weaknesses of expert labels. The second question explores the impact of disagreement between the experts on existing academic use cases of expert labels.

1.1.2 Automated Methods

Machine learning-based misinformation detection is likely one of the most popular topics in misinformation research [245, 160, 71, 8, 110, 208, 162, 191, 134, 277]. Scholars have emphasized the scalability [101], responsiveness [152], and potentially generalizable performance [134, 277] of automated methods. For instance, researchers have proposed various automated MID models for early detection [152, 142]. These models can potentially identify misinformation before it's even visible on social media. Similarly, computer scientists have also introduced domain-agnostic models, such that the performance of these models is comparable across different information domains (political vs. financial content) [134, 277, 39]. Further, other scholars [143, 228] have also built models targeting the more harmful types of misinformation (e.g., manipulated and fabricated content) [264, 275].

These are all noteworthy contributions. Yet, historically, the adoption of automated methods in real-world applications is plagued by unexpected model biases and subpar performance [190, 171]. Further, prior literature has also demonstrated that many machine learning models' performance lacks consistency across different datasets and evaluation metrics [281, 149]. The machine learning field has since moved toward more explainable, transparent, fair, and robust model-building processes [171, 137]. Nevertheless, existing work focused on evaluating the robustness of automated MID models is limited [111, 21, 138]. We aim to contribute to this literature. We ask:

3. To what extent is the performance of MID models consistent across different datasets and contexts (e.g., does model performance vary over time)?

4. What types of biases, if any, do the models present?

These two questions speak to the general weaknesses observed in many automated approaches across various research domains. Results here can shed light on whether existing MID models also have performance consistency issues and biases.

1.1.3 Crowd Wisdom

Many recent studies have also explored ways to leverage crowd wisdom for misinformation detection [261, 196, 84, 72, 22, 183, 12, 247]. Some scholars have demonstrated that labels provided by crowdsourced workers can be highly correlated with those given by experts [196, 72, 12, 213]. In other words, crowds of ordinary individuals have the potential to be a scalable alternative to professional fact-checkers.

Moreover, platforms can also rely on crowdsourced flagging and crowdsourced fact-checking for misinformation detection [51, 90]. Here, crowdsourced flagging refers to crowds using the

report button on the platform to flag posts as misinformation [51], and crowdsourced fact-checking is defined as users calling out other users' content as misinformation via commenting [125]. For instance, a user replies to another user's submission and says, "This post is misinformation. Please see [link] from Snopes.com". Both crowdsourced flagging and crowdsourced fact-checking are crowd signals readily available on platforms.

Scholars have posited that crowd wisdom provides a practical mechanism that scales with the magnitude of online content, and it also affords stakeholders (e.g., platform admins, managers of online communities) certain legitimacy to remove problematic content [51]. Nevertheless, crowd signals can be noisy and biased [90, 136]. Further, coordinated bad actors can even "game the system" and mislead platforms [90]. Though, platforms may employ various means to distinguish useful signals from misleading signals [219].

Thus far, most studies were from the perspective of platforms [90, 49]. Unlike platforms, managers of self-governing online communities (e.g., Facebook groups, and Reddit communities) only have limited access to the wisdom of their own communities. Further, these communities differ significantly in their sizes, rules, norms, and user bases. Such differences suggest that the quality of crowd wisdom varies across different communities, given that past work by Allen et al. [12] has demonstrated crowd composition has a direct impact on the accuracy of crowd wisdom. Related work that explores how community managers utilize crowd wisdom for misinformation detection is limited. Further, it's also unclear how community managers have experienced and dealt with bias and misleading crowd signals. We ask the following:

5. How are online community managers dealing with misinformation? Further, what's the role of crowd wisdom in the moderation process?

6. More importantly, how can we help platform stakeholders (e.g., platform admins, and community managers) better leverage crowd wisdom for misinformation detection?

We highlight that the questions we ask of expert labeling and automated methods are centered on their expected strength and weaknesses. Here, we assess the real-world characteristics and value of crowd wisdom in misinformation detection from the perspectives of one type of platform stakeholders: the moderators of online communities.

1.2 Research Scope

1.2.1 Domain Selection

While misinformation is prevalent in a wide range of domains (e.g., climate change [250], financial markets [103], and politics [193]) and circumstances (e.g., unexpected natural disasters such as

hurricanes [207], and scheduled occasions such as elections [232]), we focus on MID associated with two key events: *the 2016 U.S. presidential election*, and *the COVID-19 pandemic*, which is also formally known as the SARS-CoV-2 outbreak [274]. Specifically, the first two projects studying MID using expert labels and automated systems are focused on the 2016 U.S. presidential election. The third project explores crowd wisdom-based MID and is focused on COVID-19.

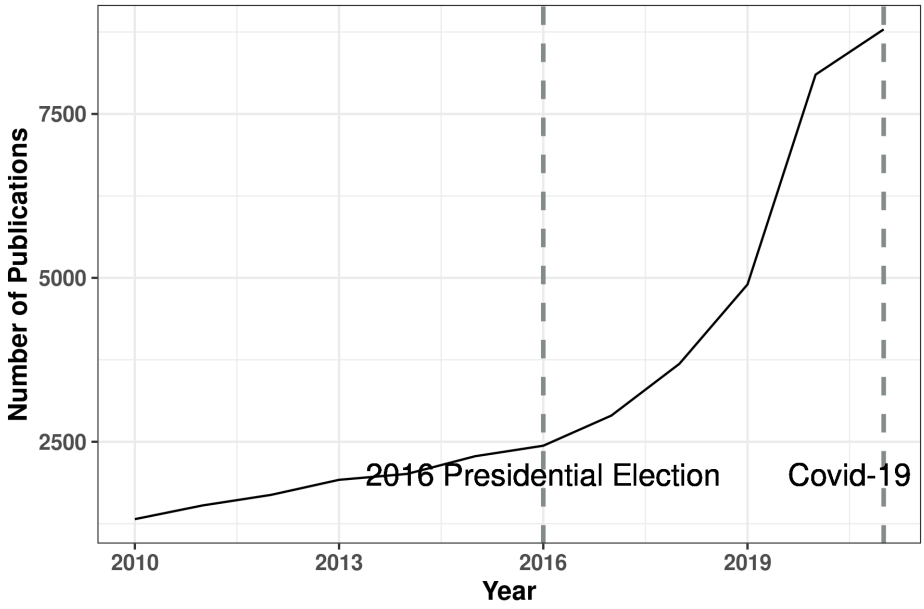


Figure 1.2: Number of Manuscripts on Google Scholars Published between 2010 to 2021 Containing the Terms “Misinformation Detection".

We choose these two events for several reasons. First, both events are extraordinarily consequential. The 2016 U.S. presidential election contributed to large-scale social unrest [81, 117], and threatened the legitimacy of established political institutions [164]. The COVID-19 pandemic has, thus far, resulted in 47 million positive cases and 0.77 million deaths in the U.S. alone [1]. It is also linked to rising violent crime rates [25], broad mental health problems among vulnerable populations such as healthcare workers [2], and persistent economic tolls [170]. More importantly, misinformation is a significant factor in both events [232, 100, 156, 37]. For instance, COVID-19 vaccine hesitation among the unvaccinated U.S. population is significantly associated with their misinformation consumption [156, 37, 124].

Second, both events have also rapidly amassed considerable research interest in misinformation detection from academia. As shown in Figure 1.2, the 2016 U.S. presidential election was followed by an exponential growth of MID-related research. For instance, only 1.3K papers on Google Scholar (https://scholar.google.com/) published in 2012 contained the terms

[1]Data was collected on Nov. 22nd 2021 from https://covid.cdc.gov/COVID-data-tracker

“misinformation" + “detection". In comparison, 2.9K papers published in 2017 did. In addition, research interest in misinformation detection continued to grow exponentially during the COVID-19 pandemic. In 2021, the total number of manuscripts that contain the terms “misinformation" + “detection" + “covid" was 5.1K, approximately three-fifths of all the papers that had the terms “misinformation" + “detection". While this metric is coarse, it's clearly indicative of the research community's heightened attention to MID-related work, likely as a product of these two events. Thus far, interest in detecting misinformation about the two events has led to a significant increase in third-party fact-checkers [74, 200, 197], and novel MID models [229, 110, 39]. However, this rapid growth could also potentially lead to problems such as label inconsistencies (e.g., divergent conceptualization of misinformation) or a lack of benchmarks across different research communities. This dissertation serves to identify and address such problems.

Third and lastly, the most popular social media platforms are particularly invested in detecting and combating misinformation related to the U.S. elections (likely due to the role of false news in the 2016 presidential election) and COVID-19. For example, both Twitter and Facebook have actively created new platform features or updated existing ones (e.g., user reporting) with the aim of reducing misinformation related to the 2020 presidential election (e.g., voting locations) and COVID-19 (e.g., vaccination) [45, 86]. Likewise, Reddit has also rolled out policies and tools specifically targeting COVID-19 misinformation [5]. In other words, MID-related work about the two aforementioned events is particularly in demand by the industry.

1.2.2 A Bounded Definition of Misinformation

The term “misinformation" has various definitions [246, 275]. As such, a piece of content could be misinformation under one conceptual framework, but may not be under another. This section serves to clarify which content is considered to be “misinformation" in this dissertation.

For studies 1 and 2 (Section 1.3), we focus on “fake news", a subcategory (or an instance) of misinformation [246, 275]. Fake news is defined as misinformation in the format of news articles that's intentionally distributed as if it's genuine news, with the aim of misleading readers [145, 246, 275]. Leading scholars have advocated to focus on the fake news publishers rather than individual stories because “we view the defining element of fake news to be the intent and processes of the publisher." [145]. There is currently no agreement on which news producers are fake news producers [246]. As such, studies 1 and 2 rely on the various lists of fake news sites as defined and provided by different experts [285, 100, 10, 255, 197, 226].

Next, study 3 (Section 1.3) is focused on how social media platforms' stakeholders (e.g., platform admins, moderators of online communities) use crowd wisdom for misinformation detection. Instead of providing an existing academic definition of misinformation, this project relies on semi-

structured interviews to understand how these stakeholders think about misinformation, what they consider to be misinformation, and how they deal with misinformation. We use open coding to synthesize their comments. Through our qualitative analysis, we observe that stakeholders' conceptualization of misinformation has three key dimensions: content facticity, potential harm, and intent. This is highly correlated with existing scholarly frameworks [246, 275].

1.3 A High-level Overview of the Three Studies

In this section, we provide a high-level overview of each study. We briefly describe our data, methods, and key analyses. The full studies are available in Chapters 2, 3, and 4.

1.3.1 Study I (Expert Labeling): Higher Ground? How Does Groundtruth Selection Impact Our Understanding of Fake News About the 2016 U.S. Presidential Nominees

In the first study, we focus on existing expert-labeled lists of news publishers. For instance, the *New York Times*, `nytimes.com`, is labeled by Vargo et al. [256] as a traditional mainstream publisher; whereas the *Boston Tribune*, `thebostontribune.com`, is labeled by Zimdars et al. [285] as a fake news publisher. We identify five lists of fake news publishers and three lists of traditional news publishers provided by different experts, including misinformation researchers and professional fact-checking organizations [285, 262, 269, 146, 255].

We first review these lists' annotation procedures, sizes, temporal changes, and real-world applications. We then evaluate the similarities between the lists by measuring their overlap. Moreover, we also assess whether domain-level explanatory variables (e.g., a website's popularity as measured using monthly viewership) are indicative of a given domain's likelihood of being included in a list. Results here can shed light on whether some lists are preferentially including or excluding particular domains than other lists. For example, compared to other lists, is a given list more likely to include conservative-leaning fake news sites but not the liberal-leaning ones (or vice versa)?

Next, we also conduct downstream analyses using tweets of the two 2016 presidential nominees (Hillary Clinton and Donald Trump) that linked to news articles from these publishers. We examine whether groundtruth choice (i.e., which fake and traditional news lists are used) leads to diverging results. We focus on three distinct downstream analyses: i) the prevalence of fake news (e.g., what fraction of news articles are from fake news sites?), ii) trend (i.e., whether the prevalence of fake news is increasing, decreasing, or staying stationary over time), and iii) agenda-setting priorities. These three types of analyses are frequently the focus of related work [232, 10, 28, 11, 100]. Study 1 focuses on answering the following: i) to what extent expert labels differ and what are

some possible explanations, and ii) whether groundtruth choice significantly changes observations in downstream analyses. Results from this study reveal that expert-labeled lists of fake news sites differ significantly, and their low agreement is possibly due to the experts having different selection pools of potential fake news sites. Additionally, while groundtruth choice has a direct impact on prevalence analysis, it has a limited to modest impact on analyses focused on the behaviors of fake news sites.

1.3.2 Study II (Automated Models): Toward a Better Performance Evaluation Framework for Fake News Classification

In study 2, we focus on automated MID models. We first review distinct classifiers from MID-related work published in a wide range of venues. We explore the distribution of these models based on their i) data sources, ii) data types, iii) feature engineering techniques, and iv) machine learning paradigms. We also determine the out-of-sample testing techniques, and evaluation metrics/steps most commonly adopted by these models. Results here can inform us whether the performance of existing models was evaluated sufficiently.

After acquiring code repositories of existing models, we evaluate these models using the following facets: i) datasets collected through distinct processes, ii) different metrics (e.g., roc auc, F1, precision, recall), iii) varied out-of-sample testing techniques (e.g., k-fold cross validation vs. leaving-p-out), iv) domain-specific biases (i.e., does performance vary for different types of news publishers?), v) temporality (i.e., does performance vary over time?), and finally vi) topic (i.e., does performance vary across different topics?). Finally, we also examine if these models exhibit shared performance issues.

Results from this study reveal meaningful performance issues and biases in existing models. That is, the performance of many MID models varies significantly across different datasets, evaluation metrics, and out-of-sample evaluation techniques. Further, we also observe important model biases with respect to publisher ideology and article topic. Future machine learning researchers can use our evaluation facets to build more robust MID models.

1.3.3 Study III (Crowd Wisdom):

The final project is focused on understanding how online community moderators use crowd wisdom for misinformation detection. We choose to focus on the Reddit platform because it relies heavily on subreddit communities themselves to self-regulate misinformation. Many prior studies had also used Reddit to study community practices [120, 206, 268].

Here, we rely on semi-structured interviews with Reddit moderators. First, we determine the general moderation workflow used by the participants to regulate COVID-19 misinformation. We

explore how moderators think about misinformation (e.g., what content they perceive to be misinformation and what false content they prioritize), how they identify potential misinformation, and how they make the final judgment (e.g., deciding whether or not content is actually misinformation). Next, we focus on the value of crowd wisdom in the moderation workflow. Particularly, we assess the value of crowdsourced flagging (i.e., user report). We determine whether and to what extent crowdsourced flagging is used by moderators to identify potential misinformation, and to make the final judgment. Moreover, we also examine whether moderators have experienced significant issues with crowdsourced flagging, and potential strategies to address these problems. Additionally, we also survey the interviewees to possibly identify other types of crowd wisdom used in their moderation workflow. Lastly, we introduce an alternative modqueue design that contains additional types of crowd signals (e.g., crowdsourced fact-checking comments) and labels from experts. We explore the strength, weaknesses, and potential use of these signals by soliciting participant feedback. Finally, we synthesize our observations and provide concrete strategies such that platform stakeholders can better leverage existing crowd wisdom for misinformation detection.

Among the key findings, our analysis reveals that moderators rely on the wisdom of two different crowds— ordinary users and other moderators. The wisdom of ordinary users is used to identify potential misinformation, whereas the wisdom of other moderators is used to assist with deciding difficult, ambiguous cases. Further, we also show that platform stakeholders can better leverage crowd wisdom by making the current report system more informative and transparent.

CHAPTER 2

Study I (Expert Labeling): An In-depth Meta-review of Expert Labels and How Groundtruth Selection Impacts Our Understanding of Fake News about the 2016 U.S. Presidential Nominees

2.1 Introduction

Following the 2016 U.S. presidential election, fake news swiftly became a topic of interest and scrutiny for political pundits, media scholars, and the general public [232, 100]—driving increased research efforts on fake news. The research community has been struggling to define fake news. While there is currently no consensus on the topic, leading scholars advocate "... focusing on the original sources—the publishers—rather than individual stories, because we view the defining element of fake news to be the intent and processes of the publisher." [145]. Yet, there is currently no agreement on which news producers are fake news producers either [246].

Consequently, there are a number of lists with opaque generation processes [285, 100, 10, 255, 197, 226] being used by studies with important implications such as examining fake news cascading behavior [10, 11], assessing agenda-setting powers of fake and traditional news sites [256, 100, 178] or characterizing changes in fake news trends [11]. How robust are these studies, particularly the ones focused on the 2016 presidential elections, with respect to the choice of groundtruth lists that define which publishers are producers of fake or traditional news? We set out to answer this question through meta-analysis—a methodology used to overcome the limitations of any single study by consolidating multiple data sources or studies that aim to address the same research questions, and determining their similarities and differences [26].

Here, we aggregate 5 lists of fake and 3 of mainstream news sites contributed by both the academia and other reputable sources [285, 269, 255, 197, 11]. We first review the labeling processes of these lists, assess their similarities and temporal changes. We then determine how selec-

tion impacts prevalence, temporal trends, and agenda-setting analysis of fake news about the 2016 presidential nominees.

We first examine prevalence given the divergent findings in recent work [232, 10, 28] [1]. A careful analysis of prevalence can also help lawmakers/platforms in better prioritizing anti-misinformation actions [145]. Next, we investigate the robustness of trend analysis since having an accurate assessment of temporal patterns of fake news can assist lawmakers/platforms in evaluating whether their efforts to curtail fake news is successful [11]. Finally, we turn to topic analysis. Agenda-setting theory [169] postulates that the most frequently covered topics are what the general public considers the most important. Relatedly, fake news sites could have led voters to re-evaluate issue importance and nominee viability by prioritizing certain topics over others [100]. Determining the robustness of fake news agenda-setting effects is consequential to media effects research.

Our paper makes the following contributions:

- We demonstrate that existing fake news lists share very few domains in common. Additionally, popular fake news sites are more likely to be included (and included earlier) than unpopular ones. Further, domains in *hate*, *junksci*, *clickbait* subcategories are less likely to be included by lists compared to domains in the *fake* subcategory.

- Based on the groundtruth choice, the prevalence of fake news varies considerably (2%-to-40%). This discrepancy is mostly due to the inclusion or exclusion of domains with mixed factualness.

- We show that the time-series correlation between most lists is high, especially for the general election period where we observe an increase in fake news prevalence regardless of groundtruth choice. Further, we also show that scheduled events contribute to a temporary drop in fake news prevalence. Observations for scandals are not as robust and are dependent on selection.

- Studying the agenda-setting priority difference between fake and traditional news sites, we observe that whether a topic (e.g., immigration) was more central to the coverage from fake news outlets compared to the traditional news sites is robust to the choice of groundtruth.

- Finally, groundtruth selection of mainstream news lists has a very limited impact on all downstream analyses.

To summarize, through meta-analysis, we characterize what makes a domain a fake news producer to some but to not others. We show that the use of different groundtruth sets can account for

[1] For instance, Silverman ((**year?**)) suggests that fake news articles garnered "...sometimes more than twice as many as legitimate news scoops in major outlets". Whereas, Allcott and Gentzkow ((**year?**)) suggests that an average adult only saw and remembered 1.14 fake news articles during the 2016 presidential election.

diverging fake news prevalence findings. Further, despite the varied labeling and validation procedures used and domains listed by fake news annotators, the groundtruth selection has a limited to modest impact on studies reporting on the behaviors of fake news sites (e.g., agenda-setting).

2.2 Related Work

Researchers have extensively documented the negative impact fake news has on the quality of civic engagement, healthcare, markets, and disaster management [208, 162, 191], both within the United States [232, 159, 242] and internationally [141, 9].

Many studies aim to distinguish false content from credible news articles at scale. Prior studies have identified differences in i) linguistic patterns such as punctuation and word choices [198], ii) auxiliary data [230, 231], iii) network cascading attributes such as depth, breadth, and speed [227, 11, 260], and iv) agenda-setting priorities [256]. These differences are then used to build automated fake news detection platforms [109] in an effort to curtail fake news.

However, efforts to study fake news and to diminish its spread are difficult [36], partly because scholars do not have a consistent definition for fake news [246, 264]. For instance, Tandoc et al. identify 2 primary dimensions of fake news: levels of facticity and deception. Wardle, on the other hand, conceptualizes fake news using 3 distinct dimensions: type of content, motivation, and dissemination method. Moreover, existing fake news labelsets [197, 285, 255, 269, 146] have considerably different annotation and categorization procedures.

We first consolidate existing groundtruth labelsets of fake and mainstream news sites that have been generated by various groups. We then assess whether and to what extent differences in groundtruth selection affect downstream studies.

2.3 Data

We use 3 types of data: i) lists of fake and traditional news sites, ii) tweets about the two nominees during the 2016 U.S. presidential election, and iii) webpages, or news articles, corresponding to the URLs shared in those tweets.

Fake and Traditional News Site Lists: We collect 5 distinct fake news lists and 3 traditional news lists from both the academia and the press [285, 100, 10, 255, 197, 226], resulting in 1884 aggregated fake news sites and 8238 traditional news sites. We describe and evaluate these lists in Section 2.4.

Twitter Data: The social media dataset is described in detail in [23]. The data collection was performed using Sysomos MAP - a social media search engine that includes access to all tweets (Twitter firehose) going back one year. For any given day between May 23, 2014, and January

1, 2017, our dataset includes i.) 5,000 tweets randomly sampled from all tweets that included the keyword “Trump", and ii) 5,000 tweets similarly sampled from all that mentioned “Clinton". The resulting dataset includes approximately 4.8 million tweets each about Donald Trump and Hillary Clinton respectively.

Webpages (News Articles): The webpages dataset [35] includes the content of the webpages shared in the Twitter dataset described above. For each tweet with an external URL, the dataset includes a record with: i) the shortened URL, ii) the original URL, iii) domain name, iv) title of the document, v) body of the document, (vi) the date of the tweet, vii) Twitter account id of the user sharing the URL, and viii) a binary categorization that indicates whether this tweet is about Clinton or Trump. We remove the records with domains not listed in the aforementioned 10K+ news sites and filter out the tweets posted before 12/01/2015 or after 01/01/2017. We derive approximately 244K unique articles shared by 1M Tweets on Twitter.

2.4 Meta-review

In this section, we first examine the characteristics and applications of the available lists of fake and traditional news websites. Then, focusing on fake news lists, we assess their commonalities and differences and explore the characteristics of websites that are correlated with them being included in or excluded from any given list. Finally, we explore fake news domains' likelihood of becoming defunct.

Lists of Fake News Sites:

We collect 5 fake news lists.

1. `ZDR`: We refer to the set of fake news websites annotated by Zimdars et al. ([285]) as `ZDR`. `ZDR` tags each website with at most 3 of the following 10 subcategories: *fake, satire, bias, conspiracy, rumor, state, junksci, hate, clickbait*, and *unreliable* [2]. Among these subcategories, *unreliable* and *clickbait* are noted to have “mixed" factualness.

2. `MBFC`: The set of sites labeled by *Media Bias/Fact Check*–an independent online media outlet maintained by a small team of researchers and journalists [255]–will be referred to as `MBFC`. Similar to `ZDR`, `MBFC` assigns domains to subcategories: *fake, conspiracy, satire*. Moreover, it also labels websites with political ideology (*extreme left, left, center, right, extreme right, unlabeled*) and rates websites by their factualness (*low, mixed, high*).

[2] Zimdars et al. also list a small subset of domains as *political, reliable* and *unidentified* which are not fake news sites and therefore removed from subsequent analyses.

3. POLIT: The staff of PolitiFact, in collaboration with Facebook, identified the list of most-shared fake news sites on Facebook during the 2016 election [197]. This list–referred to as POLIT–labels sites to *fake, imposter, some fake*, or *parody*.

4. DDOT: This list is shared by *the Daily Dot*, a mainstream online news site [269]. This list is largely created by referencing other pre-existing fake news lists and does not contain subcategories.

5. AGZ: [11] aggregated the following five lists: POLIT, Grinberg et al. ([95]), Silverman ([232]), Schaedel ([217]), and Guess et al.([99]). This list is referred to as AGZ. The subcategorization process in AGZ is somewhat complex. For instance, POLIT subcategories were ignored and all the domains were relabeled as *fake*. However, the subcategories *black, red, orange* (black: completely false, red/orange: has unreliable claims) of [95] were maintained. Finally, all domains from other referenced lists were labeled as *fake*.

A synthesis of these lists reveals that 4 out of the 5 lists share 2 common subcategories: i) a subcategory containing domains with *mixed* factualness, and ii) a *fake* subcategory (entirely fabricated). This consistency suggests that *mixed* or *fake* domains are conceptually distinct from others. Thus, studies should take this distinction into consideration.

Lists of Traditional News Sites:

We consider the following 3 traditional news lists.

1. ALEXA: Alexa is an online domain directory owned by Amazon [270]. We crawl for all the websites listed under Alexa's *News* category.

2. MBFC(T): *Media Bias/Fact Check* also lists a large set of traditional news sites. We refer to this list as MBFC(T).

3. VARGO: This list contains fact-based news websites compiled through manual content analysis of the top news media websites found in GDELT's global knowledge graph [256].

Considering fake news domain list quantities, DDOT has the fewest with 175 domains, followed by POLIT (327) and AGZ (673). ZDR (786) and MBFC (1183) are the largest lists. Traditional news site list quantities are MBFC(T) (1685), VARGO (2649) and ALEXA (5497). Table 2.1 provides a summary of the annotation processes and the uses of these lists. As is evident from the second column (*Annotation and Quality*), most lists do not have a transparent annotation and quality evaluation procedure. Perhaps due to the absence of such robust procedures, there is no consensus on which of these lists should be treated as the ultimate groundtruth. This is clear from

List	Annotation and Quality	Applications
DDOT	no information	build automated fake news trackers [226, 107], assess agenda-setting powers of fake and traditional news sites [256, 100, 178]
AGZ	authors aggregate lists generated by others, and then use various combinations of these list for result robustness check	assess impact on election, examine fake news cascading behavior [10]; examining fake news trend [11]
MBFC	annotated by staff; authors examine wording, source, story selection, and political affiliation	studies of the Alt-right [159], globalism [242], the virality of fake news [56], information literacy [76], polarization [52], and information quality [181]
POLIT	no information	study the diffusion of fake news on social media [11], information literacy [178], automate fake news detection [92]
ZDR	annotated by scholars and librarians; domain name, about us page, writing style, aesthetics, and social media accounts are among the examined characteristics	examine network cascading behavior difference between fake and real news articles during the 2016 Election [10, 11], build fake news classifiers [226, 109, 110], assess agenda-setting powers of fake and real news sites [256, 100], impact assessment [212, 80, 65], ethics and policy [77, 139]
MBFC-T	see MBFC	see MBFC
ALEXA	no information	examine cascading behavior differences between fake and traditional news articles [10, 11], news sharing behavior in right-leaning echo chambers [150]
VARGO	annotated by authors; intercoder reliability of 0.988 Krippendorff's alpha.	assess agenda-setting power of fake and real news sites [100, 256]

Table 2.1: Traditional and Fake News Lists and Their Applications. Some studies below use multiple sources.

the third column (*Applications*). More than 20 studies have used these lists of fake and traditional news sites. The lists are used for various important purposes such as building automated fake news classifiers or assessing the impact of fake news on the 2016 election. This highlights the importance of identifying similarities and differences between the lists.

Thus, we conduct downstream analysis using different groundtruth pairs (f,t) where $f \in \{\texttt{ZDR}, \texttt{MBFC}, \texttt{POLIT}, \texttt{DDOT}, \texttt{AGZ}\}$, and $t \in \{\texttt{ALEXA}, \texttt{MBFC(T)}, \texttt{VARGO}\}$.

List Overlap:

Here, we identify the overlap among the 5 fake news lists using 2 metrics. We first calculate the fraction of websites being present in at least 2 of the 5 lists, then 3, then 4. We observe that close to 50% of all domains are only included in a single list. In fact, only 5.7% of the domains are included by all fake lists. Second, we also calculate the Jaccard similarity score [91] of each pair of lists. We observe that more than half of the 15 pairs of fake news lists have a similarity of $<= 0.1$. We note that MBFC and DDOT have the lowest Jaccard similarity score of 0.08, and AGZ and POLIT

have the highest score of 0.48.

The extent of dissimilarity between the lists is surprising, and we identify four potential measures: i) *popularity*, defined as the number of times a URL from a given domain is shared in the Twitter dataset, ii) *age* (we collect data using whois.com, an online domain registration service), iii) *subcategory*, as defined by Zimdars et al. ([285])[3], and finally vi) *ideology*, as defined by *Media Bias/Fact Check* [255][4]. The details of the regression model and analysis are provided in the Appendix. We observe that the popularity of a website is positively correlated with being included in lists (though the variable is not significant for DDOT and POLIT). Further, ideology is not predictive of whether a domain will be included by lists except for AGZ (conservative-leaning domains are more likely to be listed). Finally, we observe that compared to domains subcategorized as *fake* by ZDR, domains that belong to other subcategories are uniformly less likely to be present in other lists.

Domain Addition and Removal through Time:

We further examined how the lists changed over time and found the types of changes to be largely consistent. For the lists we have temporal information for (MBFC, ZDR, and DDOT), we observe the following: i) they include more popular domains earlier on–adding the less popular ones later, ii.) they include the sites that publish fake news earlier compared to sites that publish less problematic categories such as *clickbait* and *bias*, and iii.) interestingly, sites labeled as *satirical* are added early on to the lists, perhaps due to the ease of identification. For the regression model for temporal analysis, we refer the reader to the Appendix.

Besides the addition of domains through time, we also looked into i) domain removals and ii) domains with changed subcategories. We observe very few to no removals [5]; same for changes of subcategories.

Active and Defunct Domains:

Once flagged as fake news websites, these publishers may aim to bypass fact-checking systems by using simple tricks such as abandoning their domains and migrating to new ones [85]. We observe that 68.9% of all websites listed under POLIT are no longer active—the highest defunct

[3]Zimdars et al. ([285]) have the most comprehensive subcategories and a coherent labeling guideline. Subcategory is *unknown* if a domain is not listed by Zimdars et al. ([285]).

[4]Ideology is *unknown* if the domain is not listed by *Media Bias/Fact Check* [255] or if *Media Bias/Fact Check* didn't mark it with an ideological label (approximately 18.6% domains). Here we collapse MBFC's *extreme left* and *left* categories into single *liberal* class. Same for *conservative*.

[5]The exception being DDOT: in late 2016, DDOT contained 98 websites; it then removed a substantial number of sites and reduced its size to 25 in mid-2017; its latest version has a size of 175. No explanation was given for each change.

rate among all lists [6]. Further, AGZ and DDOT have comparable defunct rates of 64% and 62% respectively. In comparison, ZDR and MBFC have considerably lower rates of 40.6% and 30.9%. Similar to the previous section, we assess a domain's likelihood of being defunct as a function of its *popularity, age, subcategory*, and *ideology* (see Appendix).

We show that older, more popular, and ideologically conservative or ambiguous domains are less likely to be defunct. Further, compared to domains subcategorized by ZDR as *fake*, domains with other subcategories (e.g., *junksci*, *satire*) are less likely to be defunct. Thus, one possible explanation that ZDR and MBFC have lower defunct rates is that both sources include more domains that do not belong to the subcategory *fake* (e.g., *unreliable* and *conspiracy* websites), and these types of domains are targeted less frequently by fact-checking platforms and thus have less incentive to migrate.

2.5 Analysis & Results

A meta-review of the fake news lists in the previous section demonstrates marked differences between these lists. How do these differences affect the downstream analysis? We aim to answer this question in this section. To that end, we first assess how groundtruth selection impacts the perceived prevalence of fake news during the 2016 election. Next, we measure the similarities or dissimilarities of fake news time-series generated using different groundtruth pairs (f,t). Finally, we determine whether there are any marked differences in agenda-setting priorities of fake and real news sites due to choice in groundtruth.

2.5.1 Prevalence

Here, we define *prevalence* as the fraction of tweets containing URLs that are from fake news sites. We examine to what extent groundtruth difference impacts perceived pervasiveness of fake news using 3 distinct boundary conditions (strictness in definition) for each fake news list: *all*, *all-except-mixed*, and *fake*. More specifically, given a groundtruth pair (f,t), we write f_{all} as the entire set of domains in f, f_{mixed} and f_{fake} as the set of domains in f_{all} that belong to subcategories with mixed factualness and the subcategory *fake* respectively. We then calculate *prevalence* as $\frac{|f_{all}|^s}{|f_{all}|^s+|t_{all}|^s}$, $\frac{|f_{all}|^s-|f_{mixed}|^s}{|f_{all}|^s+|t_{all}|^s}$, and $\frac{|f_{fake|1}|^s}{|f_{all}|^s+|t_{all}|^s}$ where $|f_{all}|^s$ is the number of tweets, or shares, contributed by f_{all}.

Results are shown in Figure 2.1. For the *all* condition, based on (f,t), fake news could amount to be more than 40% of total news shares or as low as less than 3%. Further, for robustness check

[6] We use *scrapy* [174], a Python crawler library, to scrape website homepages. Domains timed-out during scraping, or returned 404 errors (Not Found), 502 (Bad Gateway), 503 (Service Unavailable), et cetera are labeled as defunct.

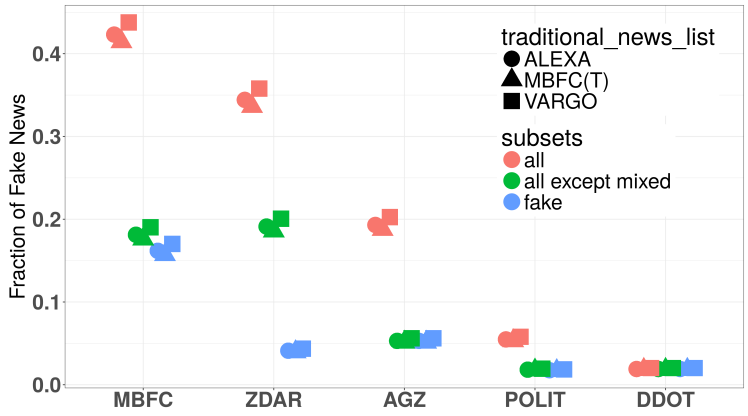


Figure 2.1: Fraction of Fake News. The x-axis indicates fake news lists. Each list is divided into subsets (marked by color) of *all*, *all-except-mixed* (not including domains in *mixed* subcategory), and *fake* (only domains in *fake* subcategory). The shape of each point denotes mainstream news lists, and y-axis is the fraction of tweets contain fake news.

(details in Section 2.5.4), we also redefine prevalence as the fraction of unique accounts that posted at least 1 fake news tweet and observe comparable results.

Additionally, if we discard all domains with mixed factualness, prevalence drops substantially to between 1.3% and 20.1%. Further, the fraction of fake news are comparable for the conditions *all-except-mixed* and *fake* except for ZDR. In other words, domains that are low in quality but not necessarily fake, (i.e. *mixed*), contribute to a large fraction of total articles shared, and domains that are neither *fake* nor *mixed* are not as popular on Twitter. To further illustrate this point, we calculate the average number of tweet shares per domain for each type of subcategories. We observe that *mixed* domains have an average of 0.7K to 2.4K tweet shares, 4 to 5 times that of the average of all subcategories for each list; in fact, *mixed* domains, on average, are considerably more popular than traditional news outlets which had an average tweet share of 0.15K to 0.66K.

Our analysis helps explain the divergent findings in the literature. While some studies raise significant concerns about the prevalence of fake news [232], others claimed limited prevalence [11] [7]. Here, similar to work by Grinberg et al. ([95]) which showed that the analysis on fake news exposure is significantly dependent on whether domains of mixed factualness were included, we see that drastically divergent conclusions can be reached even with the same Twitter data as a function of the fake and traditional news lists and fake-ness definitions (e.g., fake, mixed) one chooses to use. In sum, the more comprehensive a fake news list is, the higher the fake news prevalence.

[7] More specifically, Silverman ([232]) selected the top 20 highest performing fake news stories from hundreds of known fake news sites and demonstrated, on aggregate, they had a larger number of tweet shares compared to the top 20 news stories selected from the top 13 traditional news sites. In comparison, Allcott et al. [11] aggregated 673 fake news sites and showed that an average adult saw and remembered a single fake news story.

2.5.2 Time-Series Analysis

In this section, we first construct a time-series representing the fraction of fake news over all available news per day for each (f,t) from 3 different time periods (primary, general election, and after election) accounting for only Clinton tweets, only Trump tweets, and all tweets (for both nominee). Specifically, for each election phase i where $i \in \{primary, general\ \ election, after\ election\}$, given a groundtruth pair (f,t) and nominee n (where $n \in \{clinton, trump, both\}$), we write $|f|^s_{0,n}$, and $|t|^s_{0,n}$ as the total number of tweets, or shares, that mention n and contain URLs from f or t at day 0 [8]. We then derive the time-series $P^i(f,t,n) = \{\frac{|f|^s_{0,n}}{|f|^s_{0,n}+|t|^s_{0,n}}, \frac{|f|^s_{1,n}}{|f|^s_{1,n}+|t|^s_{1,n}}...\}$.

Next, we then compare these time-series from 3 distinct dimensions: i) correlation, ii) trend, and iii) effects of external events. For example, one might be interested to know how consistent the fake news trend is over time for discussions about Clinton (n) during the primary (i) when using $(\texttt{MBFC}, \texttt{ALEXA})$ or $(\texttt{AGZ}, \texttt{ALEXA})$ as the groundtruth pair. For that, we can use $P^{primary}(\texttt{MBFC}, \texttt{ALEXA}, Clinton)$ and $P^{primary}(\texttt{AGZ}, \texttt{ALEXA}, Clinton)$. Furthermore, instead of comparing only 2 pairs, we can compute and contrast the findings for all 15 pairs to examine overall consistency of Clinton conversations during the primary season.

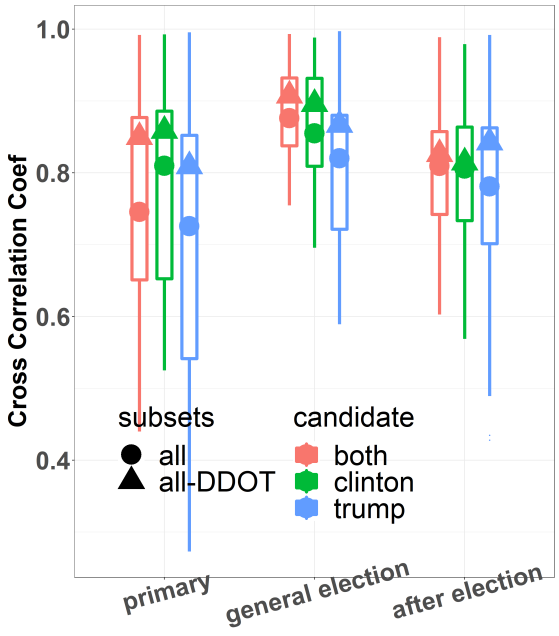
Time-series Correlation:

We calculate correlation separately for each time period and nominee. For each n and i, given 2 groundtruth pairs $(f1,t1)$ and $(f2,t2)$ where $f_1 \neq f_2$ or $t_1 \neq t_2$, we compute the maximum normalized cross correlation coefficient and the corresponding time lag [105] of $P^i(f_1,t_1,n)$ and $P^i(f_2,t_2,n)$.

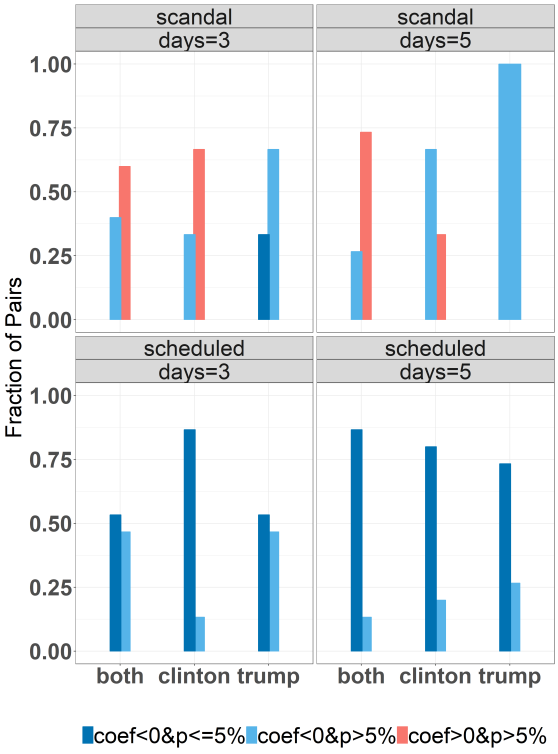
We observe that the highest correlation scores of all pair-wise comparisons occur at 0 lag, indicating that no single time-series is "ahead" or "behind" others. Correlation scores are plotted in Figure 2.2a. Normalized coefficients have a range between $\{-1,1\}$. As shown, correlation for $P(f_1,t_1,n)$ and $P(f_2,t_2,n)$ is the highest when $f_1 \equiv f_2$ but $t_1 \neq t_2$, indicating traditional news list selection (choosing `ALEXA`, `MBFC(T)`, or `VARGO`) has little impact here. Further, we also note that certain fake news lists have considerably high correlation (e.g., `ZDR` and `MBFC` have correlation consistently higher than 0.9). Yet, `DDOT` diverges significantly from others.

We further observe that the correlation is highest for the *general election* season (median correlation between the pairs for each nominee n are all above 0.8). Most efforts in fake news detection were motivated by the spread of fake news during the 2016 presidential election. This provides one potential explanation—fact-checkers and scholars could have had a stronger emphasis on the publishers that were active in this time frame, resulting in higher agreement.

[8] Here, we pick 2015-12-01, 2016-06-15, and 2016-11-09 as day 0 for primary, general election, and after election; and 2016-06-21, 2016-11-15, and 2017-01-01 as the last day.



(a) Cross correlation coefficient between pairs of lists for each time period. Circle and triangle points are the median coef scores when including or excluding DDOT.



(b) Level effect of scandals and scheduled events.

Figure 2.2: Time-series analysis results for *correlation* and *effects of external events*.

Trend:

Similar to prior work [11, 145], we are also interested in assessing whether there was an increase in fake news prevalence and to what degree findings would depend on the choice of groundtruth pairs. Here, we first employ seasonal decomposition using moving averages [20] to deconstruct each time-series $P^i(f,t,n)$ into its components *trend*, *seasonal*, and *residual*. This is to remove the seasonality and residuals from the original time-series. Next, we apply both Augmented Dickey-Fuller (ADF) and Kwiatkowski Phillips Schmidt Shin (KPSS), 2 commonly used methods to test for stationarity [42] on the *trend* component of $P^i(f,t,n)$. If any one of the tests show that unit root is non-stationary, we run the linear regression model $y^i(f,t,n) = \beta_0 + \beta_1 * T + \varepsilon$ (where $y^i(f,t,n)$ contains the values from *trend*, and T is the time elapsed since the start of the time-series). Here, a positive β_1 suggests a rise of fake news. Finally, we assess whether trend analysis results for each nominee n and time period i using all pairs (f,t) are consistent.

Results are on Table 2.2. Column “majority trend" shows the trend result shared by the largest fraction of groundtruth pairs and column “majority frac" is the size of that fraction. Additional, median β_1 scores indicate the median estimated percentage increase, over all congruent pairs, of

period	nominee	majority trend	majority frac	median β_1	least.congruent
primary	trump	positive	0.67	0.04%	DDOT, POLTI
	clinton	stationary	0.60	NA	
	both	stationary, positive	0.47	NA, 0.02%	
general election	trump	positive	**1.00**	0.03%	NA
	clinton	positive	**1.00**	0.09%	
	both	positive	**1.00**	0.07%	
after election	trump	negative	0.67	-0.01%	AGZ, POLTI
	clinton	positive	0.80	0.07%	
	both	positive	0.67	0.05%	

Table 2.2: Fraction of Fake News Per Day Time-series Trend Using Different GroundTruth. Column *majority trend* denotes the trend observed by a majority of pairs, column *majority frac* is the fraction of pairs in majority.

fake news per day. As shown, conclusions for the *general election* are remarkably consistent: all lists pairs indicate an increase in fake news. In other words, regardless of whether groundtruth choice is (MBFC, ALEXA), (AGZ, VARGO), or any of the other combinations, we repeatedly see a positive trend for fake news in the general elections. Results for other election phases are less congruent (e.g., 80% pairs show a positive trend for Clinton-related fake news after the election, but the other 20% show stationarity or a negative trend). We observe that DDOT and POLIT disagree the most with other fake news lists (measured by the number of times a list diverges from "majority vote") in *primary*, whereas it's AGZ and POLIT in *after election*.

In sum, similar to time-series correlation analysis, we see a higher consistency for the *general election* period compared to *primary* and *after election*. Accurate trend analysis is vital given that it impacts platform owners and policymakers' decision-making. Facebook's fact-checking system targets domains listed in POLIT and AGZ, and consequently [11] shows significantly reduced content from these domains on Facebook over time. We do not have access to Facebook data and therefore cannot check the robustness of their curtailing efforts. Yet, we do demonstrate that caution must be taken when examining fake news spread outside of the general election period.

Effects of External Events:

Many prior studies examined media coverage of i) unexpected political events such as scandals [203] as well as ii) scheduled high-profile events such as the presidential debates [218]. Such events are shown to have important effects on campaign news coverage. Here, we examine whether these 2 distinct categories of events have a temporary effect on the prevalence of fake news, specifically in the *general election* period.

We first obtain a list of scandals and planned key events of Trump, Clinton, or both that occurred in the general election from *ABC News* and *The Guardian*. The list, ordered chronically,

includes: Republican nomination (07/18), Democrat nomination (07/28), Clinton “deplorable" and “pneumonia" scandals (09/09), first debate (09/26), Clinton email involving Wikileaks and Trump Hollywood tape scandals (10/07), second debate (10/09), Clinton email scandals involving the FBI (10/28, 11/06), and finally election day (11/08). Here, nominations, debates, and election day are assigned to *scheduled* and others to *scandal*.

Next, we use the extended version of autoregressive integrated moving average (ARIMA) time-series model [243] to run interrupted time-series analysis and identify whether *scandals* and *scheduled* events are associated with level changes in the fraction of fake news per day for x days where $x \in \{3,5,7\}$. In our paper, we use *auto.arima*, a common ARIMA model selection function [161] from R's forecast library. Given a time-series, $P^i(f,t,n)$, and a set of external regressors (i.e., events), *auto.arima* selects the best ARIMA model based on the corrected Akaike information criterion (AIC). Here, we have 2 external regressors for each n. We denote $xreg^1_{n,T} = \{0,0,...1,1,1...\}$ where $xreg^1_{n,t} = 1$ if day t is within x days of the nearest *scandal* (after it has occurred) involving n. Similarly, we write $xreg^2_{n,T}$ for *scheduled* [9].

A positive coefficient returned by *auto.arima* for $xreg^1_{n,T}$ would mean that *scandals* temporarily increase the fraction of fake news per day. As shown in Figure 2.2b [10], regardless of the groundtruth selection, *scheduled* events generally contribute to a reduction of fake news. This does not mean planned events reduced the absolute volume of fake news. One possible explanation is mainstream media simply covered scheduled events much more, thus $\frac{|f|^s}{|f|^s+|t|^s}$ is smaller. Results for *scandal* are, however, more varied, suggesting that groundtruth pair selection has an impact on perceived effects of scandals. For instance, we see that *scandals* contributed to a short-term *increase* in the fraction of fake news shared per day when given groundtruth pair $(\texttt{ZDR}, \texttt{ALEXA})$, but a *decrease* if pair is $(\texttt{POLIT}, \texttt{ALEXA})$. This discrepancy is particularly important to studies that examine how scandals and negative media coverage diminish voter turnout in the 2016 election, particularly for Clinton [75].

2.5.3 Agenda-setting Priorities

In this section, we first use an iterative topic modeling process to extract issues, or topics, being covered by both fake and traditional news sites and assign each news article to its corresponding topic. Next, we examine whether the choice of groundtruth pairs impacts agenda-setting conclusions.

[9] If n is *both*, we only use events that involve both nominees.

[10] Trend results for when $x = 7$ is omitted due to space.

topic	doc frac	most weighted tokens	f1
abortion	0.96%	woman abort life plan_parenthood issu punish femal	0.87
benghazi	0.60%	attack benghazi libya committe report secretari secur	0.75
c-health	0.86%	medic doctor releas report mental suffer pneumonia	0.75
climate	1.40%	climat coal environment industri land administr regul	0.89
wst	0.30%	speech wall_street talk ask issu transcript releas	0.82
d&i	0.75%	commun lgbt issu equal woman discrimin anti marriag	0.78
economy	4.4%	trade job china deal compani manufactur econom	0.79
election	20.3%	sander berni primari voter percent poll voter cruz	0.77
email	5.76%	email depart investig server classifi comey secretari	0.84
border	2.28%	immigr border mexico wall illeg deport mexican build	0.85
mid-east	3.86%	muslim islam israel isi terror terrorist attack unit syria	0.76
religion	1.14%	christian evangel church faith religi leader pastor pope	0.78
russia	1.81%	russia russian putin intellig hack offici govern	0.76
security	1.70%	iran china nuclear polici foreign deal nato secur	0.78
sexual	1.93%	woman accus alleg rape husband sexual claim sexual_assault	0.82

Table 2.3: List of Topics, Fraction of Total Documents Accounted for, Most Weighted Keywords, and F1

Topic Modeling of News Articles Using Guided LDA:

We use Guided LDA for topic modeling. It is an extension of the base LDA that allows sets of keywords to guide document topic assignment by increasing their "confidence" or weights [118].

First, we use base LDA and manual labeling to extract seed words from news articles[11]. More specifically, we use *gensim* [209] to generate several base LDA models[12]. We then select the model which has the optimal coherence score[13]. From it, we obtain the top 30 most representative words for each topic. Next, we manually inspect words and categorize them into coherent sets (i.e., topics). Using this approach, we obtain 409 unique seed words divided into 33 different sets. Next, we run the guided using the derived seed word sets[14]. We filter out the subset of topics that lacked coherent themes and collapse topics that share the same human-interpretable theme into a single topic. This process results in 19 distinct topics. Finally, we assign each document into a single topic according to the maximum probability of its topic distribution. This topic is later referred to as the document's *predicted* topic label.

[11] We remove stop words, lemmatize and perform stemming. Finally, we remove all articles that have <100 or >800 word tokens.

[12] The number of topics are $\{50, 75, 100, 125, 150\}$ respectively for the models. In addition, we set all models to ignore words and bigrams that have a frequency of less than 100 or occur in more than 50% of total documents.

[13] Coherence score for a topic is the average of the pairwise word-similarity scores of its words [182]. A model's coherence score is the sum over its topic coherence scores.

[14] We adjust model's seed confidence to 0.25 and set the number of total topics to 125. We use perplexity score [173] to determine the optimal number of topics given that *gensim* does not support coherence calculation for guided LDA.

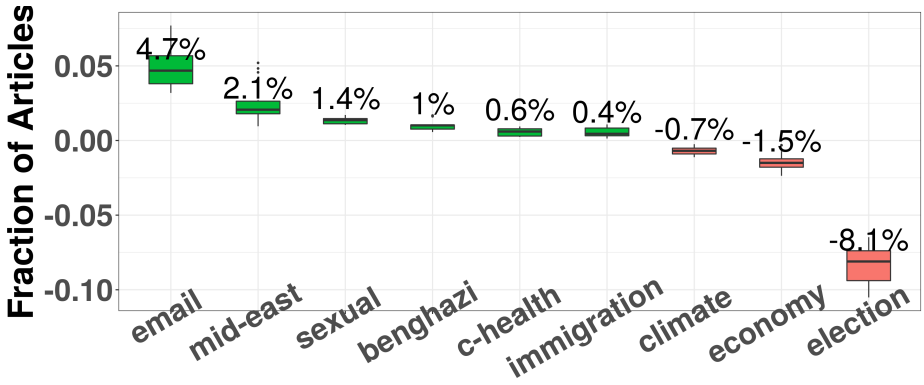


Figure 2.3: Relative agenda priority difference between fake and traditional news. Y-axis is fraction of fake news articles on topic i subtracted by the fraction of traditional news articles on i. Topics colored in green indicate a higher priority by fake news.

Topic Modeling Quality Assessment and Selection:

For each topic, we randomly sample 0.2% of its documents (or 10 if the size of a topic is small). This gives us 434 unique documents. We also sample 0.2% documents from the articles not included in the 19 topics. This results in 525 documents. Finally, we shuffle and publish the 1K $(434 + 525)$ documents on MTurk for crowdsourced labeling[15].

We assign 3 independent workers to categorize each document[16] and mark the *manual* topic of each article according to the majority vote [17]. Next, for each topic, we calculate its precision, recall, and f1 scores using the *manual* and *predicted* topic labels. We filter out the topics that have an f1 score of < 0.75. This process produces 15 distinct topics accounting for 49% of total articles. Table 2.3 provides this list of topics, their names, prevalence across domains that are listed by at least one fake or traditional news list, most weighted keywords, and f1 score. As shown, *election* is the most prevalent topic accounting for 20.3% of total news articles, followed by Clinton's email scandal, and the economy.

Agenda-setting Priorities:

Next, we assess whether groundtruth choice affects the perceived agenda-setting difference between fake and mainstream news.

[15]The success of a crowdsourcing task relies heavily on the right mechanisms to ensure worker qualifications. We require that workers: 1) reside in the U.S. 2) have successfully completed at least 1,000 HITs; and 3) have an approval rate of at least 98%.

[16]Workers are given a list of categories (19 topics listed in Table 2.3 + 1 *none of the above* option) to choose from and are instructed to select a single *primary* category of a given article. We use Krippendorff's alpha [106] to measure interrater reliability. It is 0.62, which means a moderate agreement.

[17]Articles that do not have a majority is labeled as *unknown*. We observe 46, or 8.6% *unknown* documents. Note, *unknown* documents differ from *none of the above*.

For any given groundtruth pair (f,t), we derive the following topic distributions $K_{(f,t)} = \{k^1_{(f,t)}, k^2_{(f,t)} ... k^{16}_{(f,t)}\}$ and $L_{(f,t)} = \{l^1_{(f,t)}, l^2_{(f,t)} ... l^{16}_{(f,t)}\}$ where $k^i_{(f,t)}$ and $l^i_{(f,t)}$ are the fractions of fake and traditional news articles on topic i respectively, and $\sum_{i=1}^{16} k^i_{(f,t)} = 1$, $\sum_{i=1}^{16} l^i_{(f,t)} = 1$. Then, for each topic i in $\mathscr{I}$ (where $\mathscr{I}$ is the entire set of topics), and all groundtruth pairs (F,T), we apply Student's T-test on $K^i(F,T)$ and $L^i(F,T)$ to determine whether the difference in mean is statistically significant between these 2 distributions (here, $K^i(F,T) = \{K^i(f1,t1), K^i(f2,t1) ... K^i(f5,t3)\}$). In other words, we assess whether fake news sites have published significantly more or fewer articles (measured using normalized fractions) on certain topics than traditional news sites and vice versa. We observe a significant difference in 9 topics. For instance, the average fraction of traditional news articles focusing on *election* is 22.5%, while the average is less than 15% for fake news articles. Traditional news sites are also more concentrated on topics including *economy* and *climate*. Fake news sites, on the other hand, spend a considerable fraction, approximately 10%, of all articles on Clinton's *email* scandal alone, twice that of traditional news sites. Fake news sites also place a stronger emphasis on topics such as *sexual* scandals (mostly related to Bill Clinton), and Hillary's pneumonia and claims of early onset dementia.

For each pair (f,t), we calculate the difference distribution $D_{(f,t)} = \{d^1_{(f,t)}, d^2_{(f,t)} ... d^9_{(f,t)}\}$ where $d^1_{(f,t)} = k^1_{(f,t)} - l^1_{(f,t)}$. We plot $D^{\mathscr{I}}_{(F,T)}$ in Figure 2.3. Notably, the data points of $D^{\mathscr{I}}_{(f,t)}$ consistently stay above or below the horizontal $y = 0$ line. For instance, given groundtruth (`AGZ`, `VARGO`), 13.1% and 23.6% of fake and traditional news articles covered the election. The negative difference is statistically significant, suggesting that fake news places less priority on the horse-race coverage compared to traditional news. Further, the negative difference persists for all pairs (f,t). Similarly, for all pairs (f,t), we consistently see a higher fraction coverage of Clinton's email scandal by fake news outlets. In other words, the assessment as to whether a topic was more central to the coverage of fake vs. traditional news outlets is robust to the choice of groundtruth pairs. This is good news for studies that are focused on misinformation publishers' agenda-setting functions [256]: fake news domains commonly prioritize hyperpolarizing and hyperpartisan issues, and including more or fewer domains in a study is unlikely to change the overall results.

Groundtruth Difference Using Factor Analysis:

Here, we provide an analysis of how topics contribute to the variance in agenda-setting across groundtruth pairs. We apply PCA [273] on $D^{\mathscr{I}}_{F,T}$ and extract the first 2 principal components (the first and second component explains 68% and 23% of the total variance). The resulting biplot is shown in Figure 2.4. We see that `MBFC`, `ZDR`, and `AGZ` are more similar in their topic distributions. In comparison, fake sites in `POLIT` have a higher fraction of articles on *election*. One possible explanation is that this list is specifically created to reduce election-related fake news [197]. Additionally, we also see that fake news sites in `DDOT` have a higher priority for scandals and

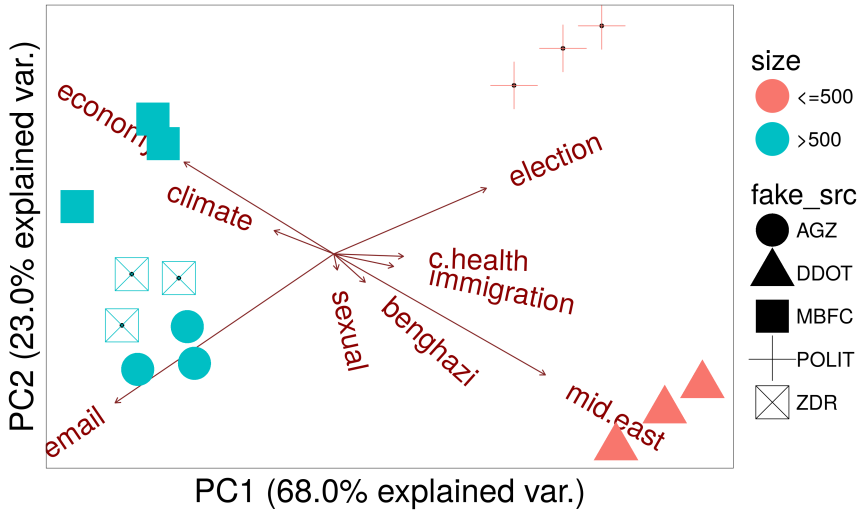


Figure 2.4: PCA plot for topic fractional difference distribution between fake and traditional news described in Section. 2.5.3 Fake news lists are marked by shape.

controversial issues including *benghazi* and *sexual*, perhaps due to *Daily Dot* being a social news site focused on fake news sites that wrote entertaining false content.

2.5.4 Robustness Checks

In this section, we conduct additional analysis to ensure that our results on the prevalence, temporal attributes, and agenda-setting priorities of fake news with respect to groundtruth choice are robust.

Fake News Characteristics Measured Using User Participation:

Thus far, we have approximated prevalence using the number of tweets. Yet, it's possible to have a few exceedingly active and concentrated accounts post a large amount of tweets containing fake news without gaining traction in the general population [95]. Here, we reexamine fake news characteristics using the number of users. We observe comparable results.

First, we redefine prevalence as the fraction of accounts that posted at least 1 tweet containing fake news and observe that, depending on the groundtruth choice, the prevalence of fake news ranges from 3.9% to 55.7% (compared to from 1.3% to 43.7% when measured using tweets).

Focusing on temporal patterns, we again see a consistent positive trend on the fraction of users who shared fake news during the general election period. That is, regardless of groundtruth choice, we observe that the closer the time was to the general election date, the higher the fraction of users who shared fake news. Further, *scheduled* events are consistently associated with a short-term decrease in the fraction of users who shared fake news, whereas results for *scandals* are dependent on groundtruth choice (e.g., scandals are correlated with a short-term *decrease* in fake news when

the groundtruth pair is $(\texttt{AGZ}, \texttt{ALEXA})$, but a short-term *increase* if pair is $(\texttt{MBFC}, \texttt{ALEXA})$). These observations are comparable to prior results obtained using tweets.

Finally, agenda-setting priority differences between fake and traditional news media measured by user participation (i.e., defining priority as how many unique accounts posted about a given topic versus how many tweets were posted about that topic) result in comparable conclusions. We observe that, for all combinations of f and t, topics including *email*, *mid-east*, and *sexual* have the highest priority in fake news, whereas *climate*, *economy*, and *election* have the highest priority in traditional news. In sum, we arrive at similar results when conducting analysis using user participation compared to when using tweets.

Addressing Potential Biases in Keywords-based Data Collection

Another concern lies with data incompleteness leading to biased observations. Thus far, we only use keywords “trump" and “clinton" to collect tweets concerning each of the two presidential nominees respectively. Therefore, a tweet about Hillary Clinton that only includes the first name “hillary" is absent from our original data. Here, we expand our dataset to include the 2 additional random sample of tweets that contain the keywords “hillary" and “hillary clinton" respectively–collected using the Sysomos MAP pipeline (see “Data" section). We then repeat our prior analysis. While the additional data increases the total number of tweets for Clinton to 13.3M, 2.8 times the size of the original dataset, downstream results generally remain the same. For instance, fake news prevalence ranges from 2.2% to 47.7% when using the expanded dataset—similar to the range of 1.3% to 43.7% when using the original dataset. Further, time-series generated using the 2 datasets are also highly correlated (e.g., the median normalized cross-correlation for the time-series on Clinton is 0.94). In fact, the expanded dataset only resulted in 306 additional number of unique articles (a mere 0.13% increase from the total 244K).

Overall, the results suggest that our analysis are robust. However, we note that our dataset and assessments remain only focused on the two 2016 presidential nominees. Our data do not include other related subjects, or personalities, such as political parties and congressional candidates, and the study of these subjects is outside the scope of this paper.

2.6 Discussion

In this paper, we first provided a comprehensive overview of the publicly available lists of fake and traditional news sites. We showed that these lists have divergent labeling processes and very few domains in common. In addition, we illustrated that the perceived prevalence of fake news varies substantially based on groundtruth choice. Despite these initially discouraging results, we

were able to reach several important robust conclusions. We noted an increase in fake news during the general election season regardless of the groundtruth selection and a temporary reduction of fake news due to scheduled events (conclusions for scandals were more mixed). Finally, after an iterative topic modeling process, we showed that agenda-setting priority differences between fake and mainstream news sites are relatively robust to the groundtruth pair choice. Overall, our results suggest groundtruth selection has a sizable impact on prevalence analysis and limited impact on downstream analysis in i) temporal characteristics, and ii) agenda-setting priorities.

There are several caveats to our study. First, our analysis of groundtruth difference and its impact is limited to domain-level labels. There are more granular datasets that annotate content at article–or even sentence–level. Second, while the focus of our meta-analysis—prevalence, temporal characteristics, and agenda-setting priorities—asks important research questions, future work should also review existing literature on similarly significant issues, such as fake news exposure in different demographics [95] or supervised fake news detection [226], identify similarities and potentially contradictory results, and determine whether groundtruth choice contributes to the observed differences (e.g., how groundtruth affect the performance of automated fake news classifiers).

Third, our dataset and analysis are only focused on the subset of fake news surroundings the two presidential nominees in the 2016 presidential election. Future work should address how the study of fake news in other fields (e.g., misinformation concerning vaccination) could also be potentially impacted by groundtruth choice.

Where do we go from here? How can we make progress as a research community despite the lack of agreement between fake news lists and domains with potential to be considered fake? Our findings can be leveraged to provide guidance.

Guidance on List Expansion and Maintenance for List Creators: Both fake news websites and groundtruth labels are indeed changing through time. List creators should include methods that track and evaluate these changes.

For efficient and timely list expansion, one key roadblock is the amount of manual labor required[18]. List creators can reduce workload by using supervised machine learning models to classify unlabeled news domains into fake or mainstream provided that potential model biases are examined and understood [19].

[18]For instance, for `MBFC`, unaffiliated individuals first submit questionable websites which automatically go into *pending* status, then the staff will review each pending domain and reach a decision using existing annotation procedure. Given that `mediabiasfactcheck.com` currently has a backlog of 500+ domains in pending, it suffices to say that the process is painstakingly slow.

[19]For instance, creators can assess whether a model is biased against domains' i) ideology-leaning, ii) popularity, iii) age and iv) subcategory. Biases in a model may not automatically disqualify it from being employed, but documenting these biases can help future scholars using these lists and models better conceptualize how potential limitations may impact the validity of their studies.

Second, for list maintenance, we urge researchers to undertake the following tasks. First, it's valuable to i) document the exact timestamp when a domain is added, removed, updated (e.g., change of subcategory), or defunct in the list. Further, if a change is unusual (e.g., subcategory modification), creators should also ii) underline reasons for the change. Next, if the annotation process is updated (e.g., ZDR introduced more subcategories as the list expands), it's also integral to iii) keep both the initial and updated procedures separate, highlight the differences, and note the time of change. These tasks not only generate useful metadata that is required by various studies, they also make the maintenance process much more transparent, which can enhance the list's credibility and help researchers identify potential discrepancies or errors early on.

Lastly, given that top fake news stories in 2016 ostensibly target white, older conservative men and favor Trump over Clinton [95, 10], we posit that the ideological-leaning of fake news sites will be undoubtedly valuable to future work in this field, and propose that creators also include the meta-data and the relevant annotation process in the lists.

Guidance on Groundtruth Selection for List Users: First, researchers need to consider whether an analysis is directly affected by list size, as in the case of prevalence. Other types of analysis that depend on the nature of the fake news domain (as opposed to counts) are more robust to the choice (e.g., temporal and topical analysis).

The second consideration relates to which lists one should use for evaluation. We first observe that the choice of traditional news lists seems to not matter, thus reducing the effort to carry out research. Second, we also see consistent clustering of fake news lists across different analyses and we recommend selecting a list from each cluster. MBFC, AGZ, and ZDR are commonly clustered together (e.g., topic analysis latent space and prevalence). POLIT and DDOT are rather distinct from the rest. By selecting a list from each (e.g., MBFC and POLIT), researchers can determine informative bounds on their analyses. Finally, if the findings diverge, expanding the set of lists used as a function of (i) annotation and quality measure described in the meta-analysis and (ii) list clustering, i.e., considering the next most distinct list, can help explore this data space systematically.

2.7 Appendix

Regression Model for Domain Inclusion

For a given domain i that's listed by at least one f where $f \in \{\texttt{ZDR}, \texttt{MBFC}, \texttt{AGZ}, \texttt{DDOT}, \texttt{POLIT}\}$, let the binary variable $y_{i,s} = \{0,1\}$ denote whether domain i exists in the list of fake news sites f. We fit model for each f using *ideology*, *subcategory*, *popularity*, and *age* as the explanatory variables.

$$y_f = \beta_0 + \beta_1 ideology + \beta_2 subtype + \beta_3 popularity + \beta_4 age + \varepsilon_i$$

	Model 1					Model 2		Model 3
	(DDOT)	(POLIT)	(AGZ)	(MBFC)	(ZDR)	(ZDR)	(MBFC)	(defunct)
independent variables								
ideology_conservative	−0.007	0.049	0.125*		0.006	0.124**	0.033	−0.130*
ideology_unknown	−0.006	0.027	0.170**		0.126	0.064	−0.066	−0.142**
subtype_bias	−0.664***	−0.480***	−0.375***	0.008		0.094**		−0.229***
subtype_clickbait	−0.677***	−0.462***	−0.373***	−0.128*		0.124*		−0.150*
subtype_conspiracy	−0.686***	−0.442***	−0.370***	0.021		−0.047	−0.001	−0.219***
subtype_hate	−0.703***	−0.452***	−0.348***	−0.251***		0.172**		−0.034
subtype_junksci	−0.705***	−0.460***	−0.388***	0.055		−0.032		−0.297***
subtype_rumor	−0.704***	−0.420***	−0.408***	−0.148		0.012		−0.058
subtype_satire	−0.704***	−0.345***	−0.284***	0.047		−0.136***	−0.115*	−0.271***
subtype_unknown	−0.703***	−0.412***	−0.268***	0.295***				−0.186***
subtype_unreliable	−0.703***	−0.512***	−0.511***	−0.178***		0.068		−0.205***
popularity	−0.0004	−0.002	0.023***	0.048***	0.042***	−0.018***	−0.046***	−0.021***
age_in_year	−0.0001	−0.009***	−0.024***	0.008***	−0.003	0.001	0.005*	−0.021***
Observations	1,644	1,644	1,644	1,644	1,644	695	724	1,644
p-value	0.62	0.21	0.18	0.17	0.053	0.057	0.067	0.173
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 2.4: All three models have different dependent variables. Model 1 assesses a domain's likelihood of being listed by a source (DDOT, POLIT, AGZ, MBFC, ZDR) given its i) ideology, ii) subcategory, iii) age, and iv) popularity. Model 2 examines characteristics that contribute to a domain's time of inclusion in sources ZDR and MBFC. Model 3 analyzes attributes correlated with the likelihood of a domain being defunct. Please note that the referencing group for ideology is *liberal*, and for subtype is *fake*.

Results are summarized on Table 2.4 (Model 1).

Regression Model for the Time of Addition

We first use web.archive.org and authors' websites to obtain 3 timestamped snapshots [20] of ZDR, MBFC, and DDOT. Let i be a website that was added to ZDR in one of its 3 snapshots and remained on the list thereafter, we determine i's preferred *ideology*, *subcategory*, *popularity*, and *age*. Let the variable $y_{i,zdr} = \{0,1,2\}$ denote whether domain i was added in the 1st, 2nd, or 3rd version of ZDR, we fit the following:

$$y_{zdr} = \beta_0 + \beta_1 ideology + \beta_2 subtype + \beta_3 popularity + \beta_4 age + \varepsilon_i$$

We repeat the same procedure for DDOT and MBFC. Regression results are summarized on Table 2.4 (Model 2) [21].

Regression Model for Active and Defunct Domains

For a given domain i that's listed by at least one f where $f \in \{\text{ZDR}, \text{MBFC}, \text{AGZ}, \text{DDOT}, \text{POLIT}\}$, let the binary variable $y_{i,s} = \{0,1\}$ denote whether domain i is defunct (i.e. $y = 1$ when i is no longer

[20] December 2016, June 2017, and December 2017 with each separated by 6 months

[21] Results for DDOT are removed given none of the variables are significant.

active). We fit model:

$$y_f = \beta_0 + \beta_1 ideology + \beta_2 subtype + \beta_3 popularity + \beta_4 age + \varepsilon_i$$

Results are summarized on Table 2.4 (Model 3).

CHAPTER 3

Study II (Automated Models): Toward a Better Performance Evaluation Framework for Fake News Classification

3.1 Introduction

In the United States, many political pundits and media scholars alike have cautioned against the rising influence of fake news [233, 15], stressing that the spread of false information weakens the legitimacy and public trust in the established political and media institutions. Outside of the U.S., fake news has been tied to Brexit in Europe [141], and the rising hate, violence, and nationalism in Indonesia [144]. It has also been linked to the endangerment of election integrity of European and Latin American nations [83, 9]. Indeed, fake news, backed by armies of social bots, disseminates significantly faster and deeper than mainstream news [227]. Additionally, subsequent research also suggests that it is difficult for the general public to distinguish fake news from credible content. Equally alarming is that repeated exposure causes readers to perceive false content as more accurate [15]. Past work also shows the importance of detecting and combating misinformation in its early phases of spread [36]. Thus, timely, scalable, and high-performing fake news detection automatons become a vital component in combating fake news.

Thus far, researchers have leveraged linguistic attributes, user network characteristics, temporal propagation patterns of news articles, and various machine learning paradigms to build effective models that separate fake news from traditional news content [215, 152, 229, 110, 39, 276]. These are all valuable contributions. Some of these novel approaches led to high performing classifiers with exceptionally high accuracy and F1 scores. Yet, our review also reveals a key gap: many papers lack comprehensive model performance evaluation and error analysis steps. Here, we first review 23 distinct classifiers from related work and consolidate each according to 4 major components: data source, data types, feature engineering techniques, and machine learning paradigms. We then reached out to the authors of 17 papers and acquired a total of 5 classifiers. Next, we

evaluate these models using a few very simple procedures. Our paper makes the following contributions:

- We show that model performance may vary drastically based on the choice of dataset. As such, results from individual papers, especially those that use a single dataset for evaluation, should be taken with a grain of salt.
- Additionally, classifiers generally have significantly higher performance when trained and validated using the common 5-fold 80/20 data split evaluation archetype compared to validation using a set of domains never before encountered in training. This suggests that classifiers might be learning domain-specific features as opposed to actual distinctions between fake and traditional news.
- We show that all classifiers studied here demonstrate significantly higher false-positive rates for right-leaning mainstream news sites. This bias raises an important concern for trust in fake news detection systems. Similarly, articles from small credible news sites are also classified as fake news more often than those from large sites.
- Next, the performance of classifiers can be worse following external shocks such as scandals. This indicates that temporal variations in classifier performance need to be taken into consideration when taking actions based on these predictions.
- Finally, models generally have a higher false-positive rate when classifying news articles involving scandals, and a higher false-negative rate for articles focused on the 2016 election (e.g., polling results).

In sum, our simple evaluation approach reveals potential biases and significant weaknesses in existing classifiers. It also provides a cautionary tale for real-world applications that use these models. Our work sufficiently demonstrates that using simple metrics such as accuracy, AUC, or F1 to evaluate and compare models is insufficient. As a community, we need to collectively investigate and construct a more comprehensive performance evaluation framework for fake news classification.

3.2 Fake News Classifiers—Review & Selection

We first review a total of 23 existing fake news classifiers in Section 3.2.1. We observe that text, relational and temporality data are the most commonly used data to construct featuresets: within the 23 classifiers reviewed, 17 (or 74%) exclusively use 1 or more of the 3 data types. The remaining 6 use at least 1 additional data type (e.g., images). Next, we reach out to authors of all 17 papers and obtained 5 code repositories. We describe this subset of models in detail in Section 3.2.2.

3.2.1 Meta-review

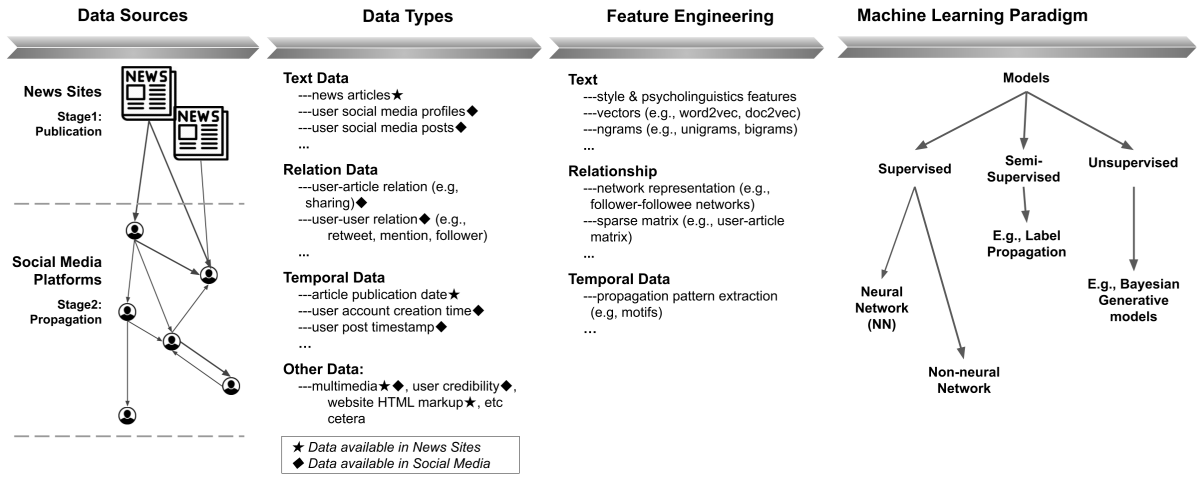


Figure 3.1: An Overview of the Fake News Detection Process. This process consists of 4 major choices in: i) datasources, ii) data types, iii) feature engineering techniques, and iv) machine learning paradigm.

As shown in Figure 3.1, the process of building a fake news classifier consists of 4 major decisions: i) choosing datasources, ii) selecting a subset of data from all available data types, iii) deciding on the techniques that transform raw data into features (i.e., feature engineering). These features are bundled closely with iv) the specific machine learning algorithm(s) one adopts.

Data Sources:

News sites and social media platforms are the two primary data sources. Our review shows that 12 (or 52.1%), and 15 (or 65.2%) out of the 23 classifiers use data from news sites and social media platforms respectively, with four (or 17.4%) using both sources.

Data Types:

We show *text* data are by far the most common with 21 (or 91.3%) classifiers using at least some text-based data. It's followed by *relational* data (e.g., follower, friend) at 47.8% and *temporality* data at 26.0% usage. A small number of classifiers also use additional data types including multimedia images and videos [102, 24]. Other data types include author age, gender, and credibility [155, 231]; website DNS records [158]; web markup and advertising [39]; and geo location [59].

Feature Engineering:

A large arsenal of techniques are available to transform raw data of varied types into usable features.

Text Data: First, features can be extracted from text using existing theories and domain knowledge such as psycholinguistic theories and frameworks [110, 39, 283]. Focusing on news articles, these features include i) quality, complexity and style (e.g. word count, lexicon diversity, readability), and ii) psychological attributes (e.g., sentiment, subjectivity, biases). See [282] for a detailed literature review. Intuitively, fake news articles are likely to include more “clickbaity" elements—capitalization of all words in the title, use of many exclamation marks, or adoption of sharp and sentimental words (e.g., “poisoning") in the text. Within a social context, user profile descriptions or user posts can be used to derive implicit features such as a user’s personality, gender, and age [231]. These features are also used to detect fake news. As such, this feature extraction approach often leads to highly explainable and transparent classifiers.

Additionally, text can be transformed into i) *ngrams*, commonly combined with tfidf weighting [6, 204]; ii) *vectors*, i.e., converting content to numeric vectors using variations of GloVE, Skipgram, CBOW, and then word2vector, sent2vec, or doc2vec [211, 93]; and iii) *tensors* [97, 192], which are 3-dimensional vector representations of words or documents. Extracted features can also undergo additional transformation steps including feature reduction.

Finally, researchers also use text to build networks (e.g., hashtag-hashtag or text similarity-based networks) and derive graph-based features [214].

Relational Data: Relational data are generally used to construct networks (e.g., follower-followee network, retweet network, user-article bipartite network). Researchers then use these networks to derive usable features including communities, clustering coefficient, and network motifs. [46, 259]. Networks can also be represented as matrices which can then be reduced into a low-dimensional representation and adopted as features [215]. Finally, these networks can be used by semi-supervised label propagation classifiers [244].

Temporality Data: Many existing studies use temporality data to build classifiers [127, 152, 215, 64]. For instance, Do et al. ([64]) and Ruchansky et al. ([215]) partition user interactions (posts) about news based on the timestamps. Posts within the same partition are treated as a single text document. Liu and Wu ([152]) model the propagation path of each news story as a multivariate time series which are then used as features. Additionally, similar to relational data, propagation networks can also be used to extract useful graph-based features such as motifs.

Machine Learning Paradigms:

Reviewed classifiers can be categorized as *supervised*, *semi-supervised* or *unsupervised*. Work by Katsaros et al. ([132]) provides an overview of existing classifiers categorized with respect to paradigms. Further, supervised models can be subcategorized as neural network [259, 211, 215, 263, 157, 279] or non-neural network based approaches [6, 110, 39, 93]. *Semi-supervised* models use label propagation techniques [127, 244, 214, 97]. Last, the *unsupervised* paradigm is very rare–we identified only one related prior work [276].

3.2.2 Representative Fake News Classifiers

From the 17 models, we see that 3 include a code repository link in their original publication. We then emailed authors of the remaining 14 papers, and 5 of them responded. Finally, we include 2 of the 5 models in our subsequent analysis [1] In sum, we collect a total of 5 distinct classifiers: `BTC`, `CSI`, `HOAX`, `NELA`, and `RDEL`. The code repositories (including our code) are available at https://github.com/lbozarth/fakenewsMPE.

CLF	Source	Text	Relational	Temporal	Other	Number of Features	Machine Learning Paradigm
BTC	News Sites	stylistic; psycholinguistics; co-ntent complexity; word2vec	X	X	geotext features (from text)	70 (e.g., stylistic); 300 (word2vec); N (geotext)	supervised; non-NN; AdaBoost
CSI	Social Media	doc2vec	user-user	YES	X	122	supervised; NN; LSTM
HOAX	Social Media	X	user-article	X	X	X	semi-supervised; propagation
NELA	News Sites	stylistic; psycholinguistics; co-ntent complexity	X	X	X	122	supervised; non-NN; RandomForest
RDEL	News Sites	ngram (tfidf); cosine similarity between title and text	X	X	X	4001	supervised; NN; Multi-layer Perceptron

Table 3.1: Overview of Classifiers based on i) data source, ii) data types, iii) feature engineering, and iv) machine learning paradigm

[1] Two of the classifiers [27, 97] are omitted due to our lack of experience with the programming language (e.g., MatLab). For the 3rd, our performance analysis reveals that it likely has an over-fitting issue due to its HTML-based features. To elaborate, we use the pre-generated HTML-based features and model provided by the paper. We train the model on 90% of the domains and validate on the remaining 10% (see the *bydomains* archetype in Section 3.4.1). The accuracy score provided in the paper using 5-fold training and validation is 0.86. In comparison, here, accuracy scores are 0.84 and 0.75 for training and validation respectively. Further, the AUC scores are 0.92 and 0.69, suggesting overfitting. We also used the leave-one-out training and validation approach. Observations are similar. Given the overfitting and that collecting HTML features for 0.7M webpages in our dataset (see Section 3.3) is also a costly task, we choose to omit this model from our analysis.

BTC: This classifier [93] uses only news article text data. The authors extract 70 stylistic and psycholinguistic features (e.g., number of unique words, sentence readability, and sentiment) and geo-location features. Additionally, they transform each word into a vector using GLOVE and then sum the vectors to generate a vector representation of each article. The authors use Adaboost [195] and concatenation of listed features to model fake news.

CSI: For each news article, this paper [215] first partitions user engagements (e.g., tweets) with an article based on timestamps of the posts. All engagements within a partition are treated as a single document. They then use LSTM to capture the temporal patterns of the documents. Additionally, the authors also build a user-user network with the edge weight being the number of shared articles between pairs of users. This network's corresponding adjacency matrix is then used to generate lower dimensional features that capture the similarity of users' article sharing behavior. Finally, both sets of features are integrated together using another neural network layer.

HOAX: The authors [244] construct a user-article bipartite graph based on whether a user liked or shared an article or a post. They then use semi-supervised harmonic label propagation to classify unlabeled articles. This approach is based on the hypothesis that users who frequently like or share fake or low-quality content can be used to identify the quality of unlabeled content.

NELA: This classifier [110] uses the following 3 distinct dimensions of text-based features to predict fake news: i) style features (e.g. exclamation marks, verb tense, pronoun usage), ii) psycholinguistic features such as sentiment scores using LIWC, SentiStrength [248] and iii) content complexity features including readability (Mc Laughlin, 1969), dictionary size, and average word length. We refer readers to the original paper for the complete list of 100+ features. The authors use Linear Support Vector Machine (SVM) and Random Forest as their classification algorithms.

RDEL: This model [211] first tokenizes text from news articles and extracts the most frequent ngrams (unigram, bigram). Then, for each news article, it constructs the corresponding term frequency-inverse document frequency (TF-IDF) vectors for article title and body separately, and computes the cosine similarity between the 2 vectors. Finally, the authors concatenated the features together and use Multilayer Perceptron [195] to classify fake and real news articles.

The data source, data types, feature engineering process, and machine learning paradigm for each of the 5 classifiers are summarized on Table 3.1. As shown, this set of classifiers encompasses both neural network and non-neural network based supervised learning paradigms, as well as the semi-supervised paradigm. Additionally, it includes all 3 most common data types: text, relational, and temporality. Focusing on feature engineering, the classifiers collectively cover the most common text feature engineering approaches: theory-driven, ngrams, word2vec, and doc2vec. Similarly, they also cover some common relational data feature engineering techniques: user-article network, user-user network. Notably, these classifiers do not include popular approaches such as leveraging user profile descriptions, or follower-followee networks. Overall, however, we argue

that this set of classifiers is sufficiently representative.

3.3 Data

We use two datasets in this study. The summary statistics are available on Table 3.2. Both datasets use *Media Bias Fact Check* [255] as ground truth to label whether a domain is a fake news site. *Media Bias Fact Check* contains one of the most comprehensive list of fake and mainstream news sites and is also used by many related work [242, 159]. Prior research shows that the choice of ground truth can have a significant effect on downstream analysis [32]. Therefore, we note that our study is only the first step towards building a comprehensive fake news model evaluation framework. Future work should consider multiple ground truth labels.

	Election-2016	NELA-GT
Time Period	12/2015-01/2017	02/2018-10/2018
# Domains	1390	130
# Fake News Domains	335 (24%)	38 (29%)
# Articles	231.6K	304.1K
# Fake News Articles	31.3K (16%)	37.4 (14%)
# Tweets	1.16M	
# Fake News tweets	141.9K (12.2%)	
# Users	215.2K	
# Fake News Users	37.9K (18%)	

Table 3.2: Basic Statistics for Datasets

Election-2016:

This dataset is primarily focused on the 2016 U.S. presidential candidates and consists of both social media data and news articles. Social media data collection is described in detail in Bode et al. ([23]). The data collection was performed using Sysomos MAP. For any given day between December, 2015, and January 1, 2017, this dataset includes i.) 5,000 tweets randomly sampled from all tweets that included the keyword “Trump", and ii) 5,000 tweets similarly sampled from all that mentioned “Clinton". The webpages dataset [35] includes the content of the webpages shared in the Twitter dataset described above. For each tweet with an external URL, the dataset includes a record with: i) the shortened URL, ii) the original URL, iii) domain name, iv) title of the document, v) body of the document, (vi) the date of the tweet, and vii) Twitter account id of the user sharing the URL. Here, we use *Media Bias Fact Check* to identify the list of fake and mainstream news sites present in *Election-2016*, and filter out non news-related sites. As shown on Table 3.2, *Election-2016* contains 231.6K unique articles (16% of which are fake news), and

1.16M tweets (12.2% of which contain links to fake news articles) shared by 215.2K unique users (18% users shared at least 1 fake news article).

NELA-GT:

This dataset [185] contains articles scraped from 194 news sites between 02/2018 and 11/2018. The list of domains is collected by aggregating existing lists of fake and mainstream news sites provided by other researchers and organizations. News content is scraped via the RSS feed of these sites. Each domain has source-level veracity labels from 1 or more independent assessments (e.g., *Media Bias Fact Check, News Guard*). We note that 130 out of 194 domains have labels from *Media Bias Fact Check*.

As shown on Table 3.2, *NELA-GT* contains 304.1K unique articles (14% of which are fake news).

There are several key distinctions between *Election-2016* and *NELA-GT*. First and foremost, news articles in *Election-2016* are collected through tweets mentioning Trump or Clinton. As such, these articles are almost exclusively about one or both candidates. In comparison, creators of *NELA-GT* directly access and scrape the websites which likely resulted in more diverse news topics. Next, *Election-2016* contains 1.4K distinct news sites, 10 times that of *NELA-GT*, yet the latter has 72.5K more news articles. Indeed, the median number of articles per domain is 13 for *Election-2016* and 1.12K for *NELA*. Finally, over 30% of all fake news domains active in 2016 have since become defunct [32], thus they are in *Election-2016* but not *NELA-GT*.

Additional Auxiliary Data:

We also obtain the following data for each news site: i) *average monthly traffic* using similarweb.com, a popular web analytics platform [234, 236]; ii) *age* using whois.com, a domain name registrar database [177]; and finally, iii) *ideology* using *Media Bias Fact Check*. Additionally, for each article, we also include its publication date (For *Election-2016*, an article's date is approximated using the timestamp of the earliest tweet that included it).

Data Preprocessing:

For each dataset, we first filter out all news articles i) without a title, or ii) with fewer than 10 words in the article body. Next, we aggregate 4.1K words and phrases representing news sites names [2] (e.g., "Daily Dot", "CNNPolitics"); for each news article, we remove matching words and phrases from the article.

[2]Here, we use the list of news site names provided by *Media Bias Fact Check*.

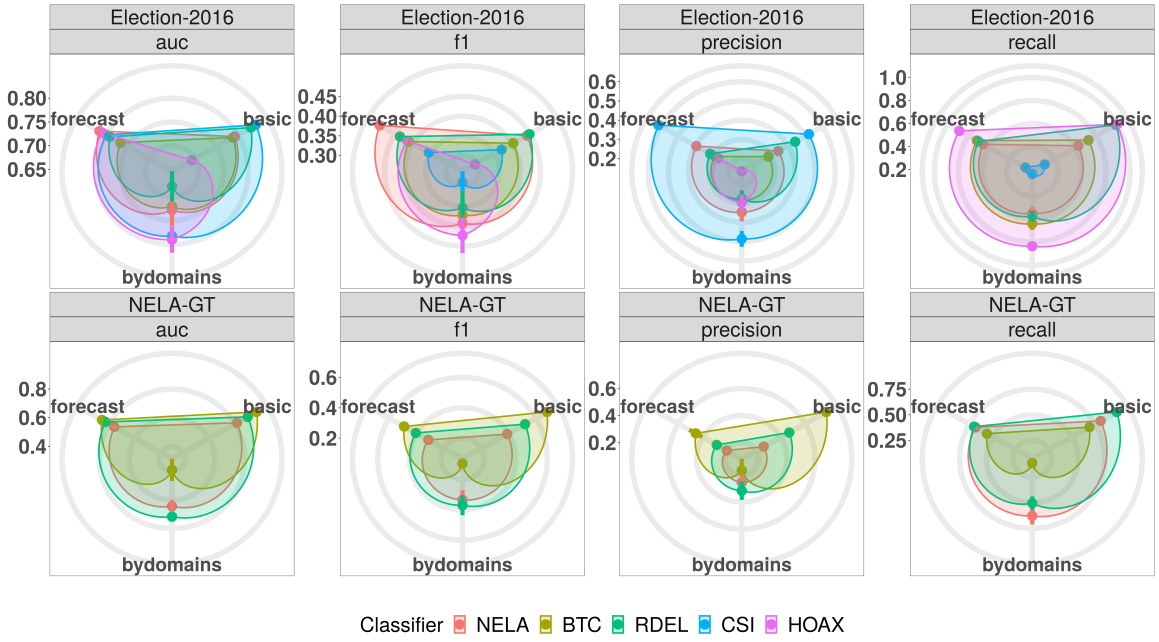


Figure 3.2: Performance Overview For All Classifiers. As shown, we separate the classifiers by colors. Additionally, each grid represents a distinct dataset and performance metric combination. Within a given grid, data-points that lie in the outer-rings represent higher performance. For instance, the upper left corner grid contains each classifier's AUC scores for the *Election-2016* dataset. We see that `RDEL`, colored in green, has a significantly higher AUC (data point lies on the outer-ring) when validated against the `basic` archetype compared to `forecast` and `bydomains`. Please note that both CSI and HOAX are excluded in the bottom row because these two models require social media data, which is not available in the NELA-GT dataset.

3.4 Analysis & Results

In this section, we first compare and contrast classifier performance using different evaluation archetypes. Then, we conduct an in-depth error analysis focusing on domains and articles of varied attributes. Finally, we examine model performance and bias tradeoffs.

3.4.1 Performance Overview

We first underline the 3 most common training and performance evaluation archetypes `basic`, `forecast`, and `bydomains`. We then present our evaluation metrics and results.

(1) Basic N-folds (`basic`): Using this common approach, data are split into N-folds for training and cross-validation. Here, for each dataset, we use 5-fold (i.e., 80/20) training and validation data split.

(2) Forecasting into the Future (`forecast`): Here, given a time t, classifiers are trained on fake and mainstream news articles that were written before t, and tested against those that were written after t. For each dataset, we randomly sample 10 dates within its data collection period and split data into training and validation accordingly.

(3) Predicting Never Before Encountered Domains (`bydomains`): In this archetype, classifiers are trained on articles from 90% randomly sampled domains and tested against articles from the remaining 10% that are not present in training. We repeat this sampling process 10 times for each dataset.

Evaluation Metrics: We use *AUC*, *F1* (fake news articles as the positive label), *precision*, and *recall* to evaluate model overall performance. We omit *accuracy* due to label imbalance in the datasets [112].

Results: Results are summarized in Figure 3.2. As shown, we denote classifiers using different colors. Additionally, each grid represents a distinct dataset and performance metric combination. Within a given grid, outer-rings represent higher performance. Note that *NELA-GT* doesn't provide social media data, so analyses for `CSI` and `HOAX` are not available for this dataset. Overall, we see that classifier performance varies considerably based on the i) dataset, ii) evaluation archetype, and iii) metric.

(i) Dataset Effects: The average *AUC* scores for `BTC` under the `basic` evaluation archetype are 0.78 and 0.96 when the datasets are *Election-2016* and *NELA-GT* respectively, a considerable difference. We also observe a similar but less significant effect for `RDEL` and `NELA`. One possible explanation is that *Election-2016* is specifically focused on Donald Trump and Hillary Clinton. Thus, news content from fake and mainstream publishers in this dataset is presumably much more similar compared to *NELA-GT*, which scrapes the entire RSS feed of news sites daily. The higher article similarity likely contributes to a performance drop in content-based models.

(ii) Archetype Effects: `RDEL` and `BTC` both perform considerably better under `basic` evaluation compared to predicting new domains (i.e., `bydomains`). For instance, when dataset is *NELA-GT*, the *AUC* score for `RDEL` is 0.97 using `basic` archetype but 0.72 using `bydomains` (similar patterns for `BTC`). One possible explanation is that, despite removing news site name-related tokens (e.g., "daily beast", "nytimes") from data, certain remaining word tokens may still be indicative of a domain and its practices (e.g., New York Times has the custom of using the word "Mr." when addressing the President of the United States [47]). Thus, if word2vec or ngram-based classifiers such as `BTC` and `RDEL` rely heavily on site-specific features for training, they may perform poorly when validating on newly encountered domains. Similarly, we also see that `BTC` and `RDEL` perform worse in `forecast` than `basic`. Findings here are complementary to prior research [111] which demonstrates that the performance of text-based models decreases over time due to changes in news content. Interestingly, `NELA`, the only other exclusively text-based classifier, has more comparable *AUC* scores across the archetypes—the *AUC* scores are 0.80, 0.73, and 0.82 for `basic`, `bydomains`, and `forecast` respectively when dataset is *Election-2016*. We note that `NELA` only uses the stylistic and psycholinguistic features of articles. As such, it may be more robust to site-specific linguistic eccentricities. Compared to text feature-only classifiers,

both CSI and HOAX—the classifiers that only partially or not-at-all rely on text content— have more comparable *AUC* scores across all archetypes.

Out of the 3 archetypes, basic is the most common. That is, model performance is typically examined using 5-fold cross-validation with 80/20 training and validation data split. Yet, if researchers only assess classifiers' performance based on *basic*, they may not discover the potential weakness their models have against never before encountered domains. In fact, our results demonstrate that the word2vec and ngram text feature-based models maybe especially need to be evaluated against bydomains.

Evaluation Metric Effects: The ranking of classifiers change based on whether we use *AUC* or *F1*. For instance, when the dataset is *Election-2016* and the archetype is bydomains, we show that CSI has *rank* $= 2$ using *AUC* and *rank* $= 4$ using *F1*. Additionally, all classifiers excluding CSI generally have significantly higher *recall* than *precision*. In fact, *CSI* has the highest *precision* across all classifiers and the lowest recall. As such, the definition of "best-performing" is dependent on whether one prefers precision over recall. Generally, high precision is preferred in high stake circumstances. For instance, Google and Facebook both have banned hundreds of fake news sites from using their Advertising services [194]. In this context, a credible mainstream news site would suffer considerable economic drawbacks if it's falsely labeled as fake. As such, CSI should be the preferred model despite having a low recall score. Alternatively, classifiers with a high recall can serve as a useful filter. As an example, researchers and organizations interested in identifying new fake news domains can first apply a high recall model to obtain a list of presumptive fake news domains, and then manually review each site to label true fakes.

In sum, we demonstrate that the performance of classifiers varies considerably from case to case as a result of the difference in i) dataset, ii) evaluation archetype, and iii) metric.

3.4.2 Domain and Context-specific Error Analysis

In this section, we conduct in-depth error analysis using false-negative rate **(FNR)** and false-positive rate **(FPR)**. Here, for each $c \in \{$BTC, CSI, HOAX, NELA, RDEL$\}$, we first i) assess whether models perform better or worse on classifying domains of particular subcategories. Next, we then explore ii) errors that are of significant interest in the context of the 2016 election.

Domain-level Error Analysis:

Here, we examine error rates based on domain i) ideological-leaning, ii) age, and iii) popularity. To elaborate, conservative elites have long criticized both the academia and tech firms for having a liberal bias [96]. Thus, we aim to determine whether classifiers indeed contain biases such as having a higher *fnr* for liberal-leaning fake news sites, and/or a higher *fpr* for conservative-leaning

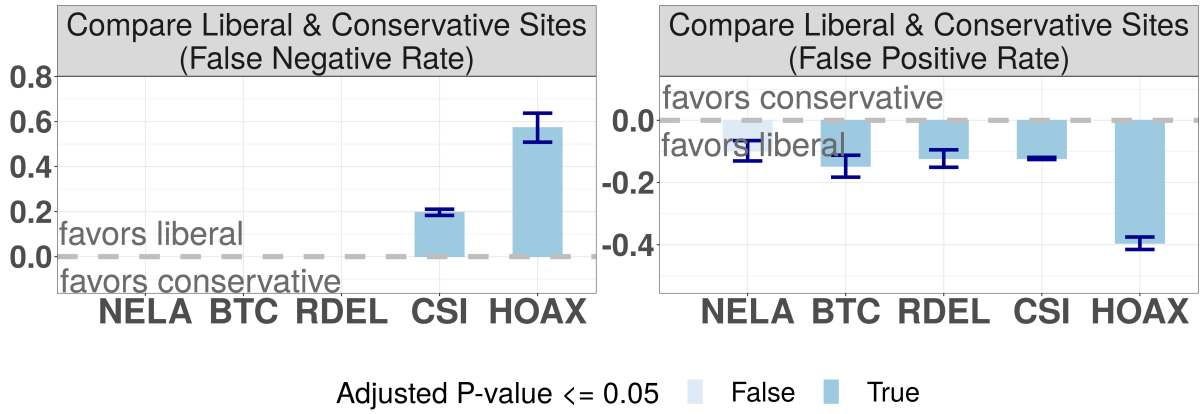
mainstream news sites. Similarly, a model's performance may also vary based on a domain's age and popularity (e.g., incorrectly classifying recently created websites with small viewerships as fake news at a higher rate compared to mature domains with heavy traffic). We note that analysis here is focused on the *Election-2016* [3] dataset and `bydomains` archetype. We focus on this archetype for potential impact. While tech giants and online platforms have blacklisted hundreds of fake news sites, reports show that owners of these domains are ramping up for the 2020 election by creating new sites [271, 239].

Ideological Biases: For each article i in a given validation set, we first assign i to a bin using i's corresponding domain's ideology $\{unknown, conservative, center, liberal\}$. Then, for each classifier c, we calculate c's *fpr* and *fnr* for each bin separately. Here, we denote *fpr* for the liberal(left)-leaning and conservative(right)-leanings bins as $fpr(l)$ and $fpr(r)$ respectively. Next, to evaluate "liberal bias", we examine i) whether liberal-leaning fake news sites on average are significantly more often classified by c as credible news (i.e., $fnr(l) > fnr(r)$), and ii) whether articles by conservative-leaning mainstream news sites on average are significantly more often classified by c as fake news (i.e., $fpr(l) < fpr(r)$).

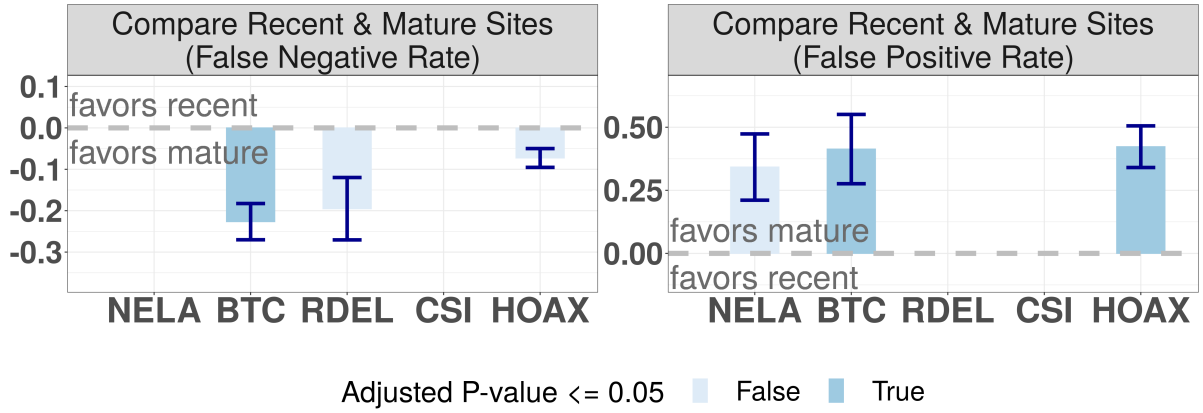
To elaborate, we first apply Student's T-test [88] on the distributions $fpr(V,l,c)$ and $fpr(V,r,c)$. Here, $fpr(V,l,c)$ is classifier c's false-positive rates for the liberal-leaning bin l across all the validation sets V of the `bydomains` archetype. For results that are statistically significant ($p-value <= 0.05$), we compute the mean differences between the distributions and then plot the values. We repeat this process for false-negative rates. As shown in Figure 3.3a, all classifiers have a higher false-positive rate for articles from conservative-leaning credible news sites. Furthermore, all results remain significant, except for `NELA`, even after adjusting the p-values using the Holm-Bonferroni method [108] to account for multiple hypothesis testing. One possible explanation for this bias is that there are significantly fewer liberal-leaning fake news sites available for training. Another explanation is provided by Benkler, et al. ([19]). They argue in a recent book that some traditional right-leaning news outlets participated in the dissemination of fake news by echoing and giving platform to false claims initially produced and campaigned by fake news sites. This might bring into question the ground truth definition as opposed to the classifier results. As a whole, results in this section demonstrate that researchers should actively consider potential ideological biases when building and evaluating fake news classifiers.

Domain Age: Similar to the previous section, we first partition each domain i based on its age into 3 bins: i) unknown (i.e., DNS record is not available), ii) recent ($<= 3 years$), and iii) mature ($> 3 year$). We then compare *fnr* and *fpr* between the bins and assess whether the mean differences are statistically significant. As shown in Figure 3.3b, both `BTC` and `HOAX` have a

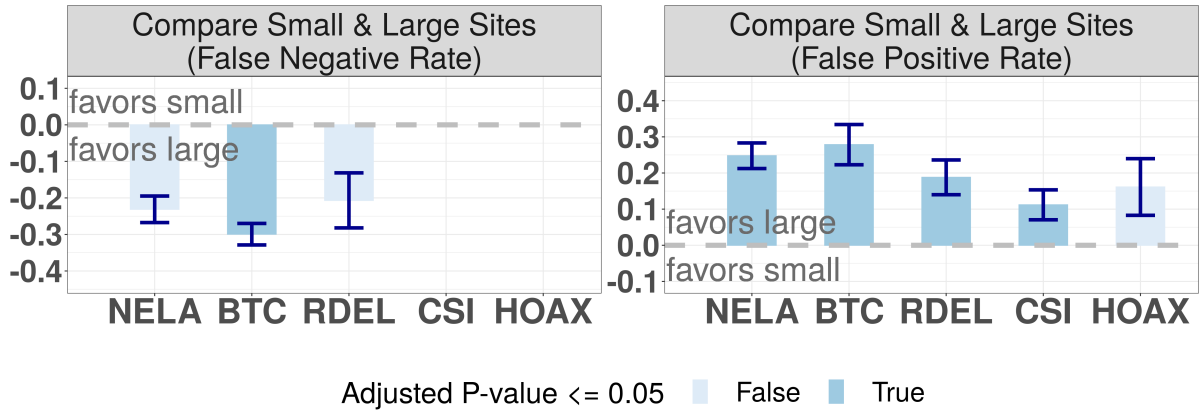
[3] We also repeat the analysis using the *NELA-GT* dataset. Results are largely insignificant given that *NELA-GT* only contains 130 domains. In comparison, *Election-2016* has 1.4K.



(a) Domain Ideology.



(b) Domain Age.



(c) Domain Popularity

Figure 3.3: Mean Differences in Error Rates.We first assign each prediction from the validation dataset into groups using its corresponding domain i) ideology, ii) age, or iii) popularity. We then compare the classifiers' false-positive and false-negative error rates for each group of predictions. Here, the x-axis denotes the classifiers, and the y-axis denotes each classifier's mean differences in error rates between a selected group and the baseline group (e.g., subtract a classifier's average false-positive rate for liberal-leaning sites by the average for conservative-leaning sites). The baseline groups are {*conservative*, *mature site*, *large site*} for domain ideology, age, and popularity respectively. Note*, results that are not statistically significant are removed. Further, results that are insignificant after adjusting p-values using the Holm–Bonferroni method [108] are colored in lightblue; finally, results that remain significant after adjustment are in darkblue.

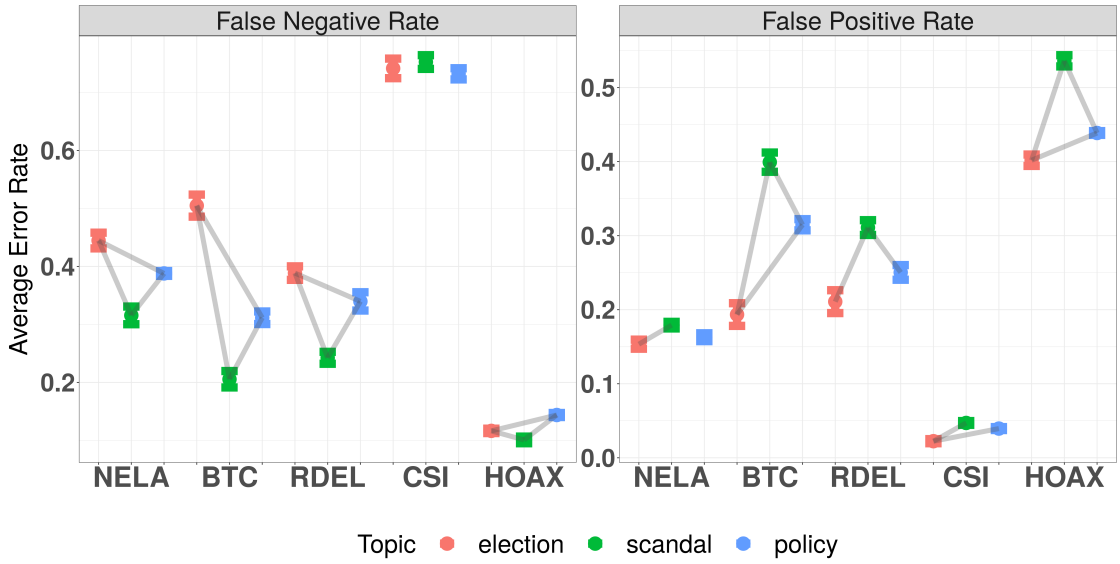


Figure 3.4: Article Topic-level Error Rates. The x-axis denotes the classifiers, and the y-axis denotes a topic's average *fpr* (or, *fnr*). Topic are differentiated by color. Further, if a pairwise comparison between 2 topics is statistically significant even after adjustment for multiple hypothesis testing, the pair is linked by a gray line. For instance, the mean differences in false-positive rate between the pairs (election, scandal) and (scandal, policy) are significant for *RDEL*.

significantly higher *fpr* for recent domains. In other words, the 2 classifiers more often label articles by recently created credible news domains as fake news. A potential explanation is that newly created mainstream news domains have published fewer articles and thus models have less data to train on. Alternatively, newer mainstream sites might have language and consumers that are better aligned with fake news outlets. Finally, for robustness check, we also repeat the evaluation and set the partitions into recent ($<= 5 years$) and mature ($> 5 years$). We observe similar patterns.

Domain Popularity (web-traffic): For each dataset, we divide domains based on web-traffic into: i) unknown (no web-traffic data on `similarweb.com`), ii) small (web-traffic percentile calculated to be between 0%-33% percentile), iii) median (33%-66% percentile), and iv) large. We again compare *fnr* and *fpr* across the bins. As shown in Figure 3.3c, all classifiers, except for `HOAX` have a significantly higher false-positive rate for small websites. This bias potentially causes small but legitimate news sites to be more often incorrectly labeled as fake news domains. Finally, for robustness check, we also repeat the evaluation and set the partitions into small (0%-50% percentile) and large (50%-100% percentile). We observe similar patterns.

Context-specific Error Analysis *(2016 election)*:

As previously stated in Section 3.3, the *Election-2016* dataset is collected specifically to study the 2016 U.S. presidential candidates. Here we identity 2 additional error analyses that are of

significance to the election. First, prior research [32] has demonstrated that the prevalence of fake news temporarily decreases after scheduled high-profile events (e.g. presidential debates). Though, results are inconclusive for scandals (e.g., Trump Hollywood tape). We expect the fake news articles produced shortly after these shocks to differ from those published in other time periods. Accordingly, we examine model performance for such articles. Furthermore, political communications studies show that news coverage of different topics have varied impact on a voter's knowledge, decision-making, and trust in the government [201, 253]. As such, we also aim to determine whether model performance differs across article topics. In sum, we conduct error analysis based on i) the types of external events, and ii) article topics. We note that analysis here is focused on the `forecast` archetype given shocks are temporal events and news coverage of different topics varies over time.

External Events: We first obtain a list of scandals and planned key events of Trump, Clinton, or both that occurred in the general election from *ABC News* and *The Guardian*. The list, ordered chronically, includes: Republican nomination (07/18), Democrat nomination (07/28), Clinton "deplorable" and "pneumonia" scandals (09/09), first debate (09/26), Clinton email involving Wikileaks and Trump Hollywood tape scandals (10/07), second debate (10/09), Clinton email scandals involving the FBI (10/28, 11/06), and finally, the election day (11/08). Here, nominations, debates, and election day are assigned to *scheduled* and others to *scandal*. We also randomly select 10 dates and assign them as *baseline* for comparison purposes. Next, given classifier c, for each day $t \in \{07/18/2016, 07/28/2016...\}$, we train c using articles before t, and validate c on the articles that were published and shared within x days after t where $x \in \{3,5,7\}$. We again compute *fpr* and *fnr* for articles published right after *scheduled, scandal* events and compare these error rates to that of *baseline*. Surprisingly, we see that error rates are generally comparable across different types of external shocks for all classifiers except for `HOAX`, which has a considerably higher false-positive rate for predicting articles published shortly after scandals (the mean difference between *fpr* for *scandal* and *baseline* is 10.0%). In our paper, `HOAX` is the only model that exclusively adopts a user-article network-based classification approach. It's possible that users may have temporarily altered their news-sharing behavior right after scandals. In comparison, results from text-based models suggest that the linguistics features of articles published after shocks are not significantly different from those in *baseline*.

Article Topics: We obtain the topic for each article in the *Election-2016* dataset from related work [32]. We refer readers to the original paper to review the detailed topic-modeling process which assigns 49% of total articles into 15 unique topics (the remaining 51% is labeled as *other*). Here, we further cluster documents into broader topics $\{election, scandal, policy, other\}$ [4]. We

[4]We merge {'email', 'clinton-health', 'sexual', 'clinton-wst', 'benghazi'} as *scandal*, and {'russia', 'economy', 'abortion', 'climate', 'mid-east', 'd&i', 'security', 'religion'} as *policy*.

observe that 20.3%, 8.9% and 19.2% of all articles belong in *election* (e.g., news involving polling results), *scandal* (e.g., news involving Clinton's pneumonia incident), and *policy* (e.g., the economy) respectively.

Next, we calculate and compare the average *fpr* and *fnr* rates for the topics. Results are summarized in Figure 3.4. The x-axis denotes the classifiers, and the y-axis denotes a topic's average *fpr* (or, *fnr*). Topics are differentiated by color. Further, if a given pair of topics, e.g. $(scandal, election)$, has a significant difference in mean error rates even after p-value adjustment, the pair are then linked together by a gray line. As shown, all text-based models and `HOAX` have a significantly higher *fpr* for *scandal* compared to both *policy* and *election*. That is, articles written by credible sites about scandals are significantly more often mislabeled as fake. We also see that, for 4 out of the 5 models, fake news articles focused on the election are more often labeled as credible compared to those about scandals or policy.

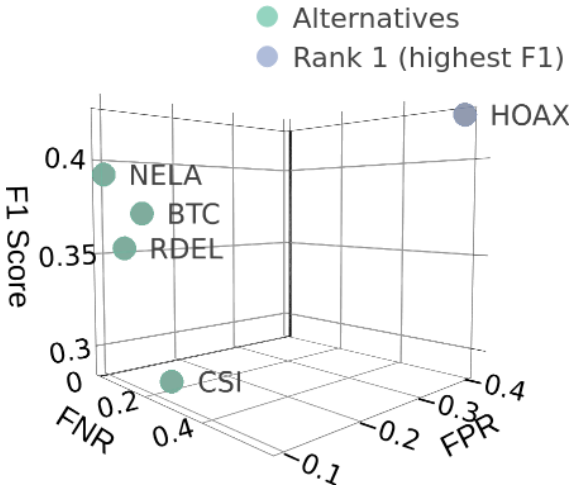
What does this mean? Past studies in voter decision-making demonstrate that scandals have a strong and long lingering impact on voter choice [201]. As such, perhaps misclassifying articles involving scandals would have a more detrimental downstream impact. On the other hand, erroneous election-related coverage such as fake polling results and endorsements also affect voter behavior [253]. It is important to note one caveat. The ground truth labels are provided at the outlet level. It is entirely possible that fake news sites generally publish accurate election-related articles. Regardless, our analysis shows that one might need to adjust prediction w.r.t. the proportions of topics covered to determine the quality of news outlets. Further, we highlight that researchers invested in building effective models should consider the impact/weight of different types of misclassifications (e.g., misclassifying a credible celebrity gossip piece as fake news may be harmless, but incorrectly labeling vaccine-related misinformation as credible is harmful).

3.4.3 Performance and Bias Trade-off

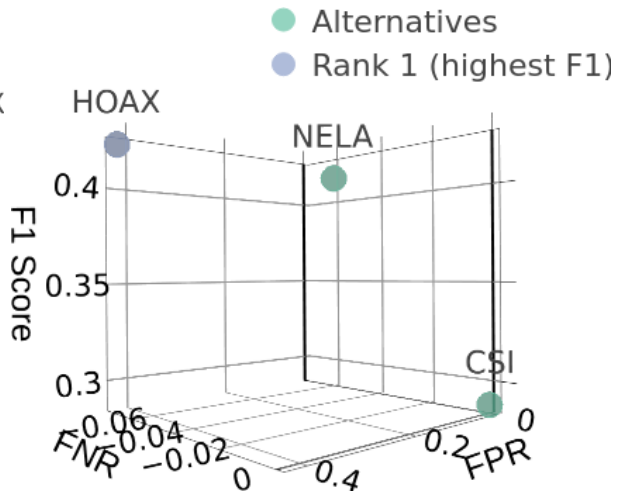
In this section, we examine error bias and performance trade-off with respect to i) domain ideology, ii) domain age, iii) domain popularity, iv) external events and v) article topic [5]. In other words, we ask if a classifier has the highest $F1$ (or, AUC) score, but significantly favors liberal-leaning news sites, does there exist another classifier which has a slightly worse $F1$(or, AUC) score but lower or no significant bias with respect to domain ideology?

Here, we first rank all classifiers based on their average $F1$ scores for the validation datasets, and denote the one with the highest $F1$ as $c\star$. We then add models with worse $F1$ scores but lower bias (with respect to domain ideology, domain age, etc.) than $c\star$ into the set C_{alt} (these are the alternative options). Models that have worse $F1$ scores in addition to higher bias when compared

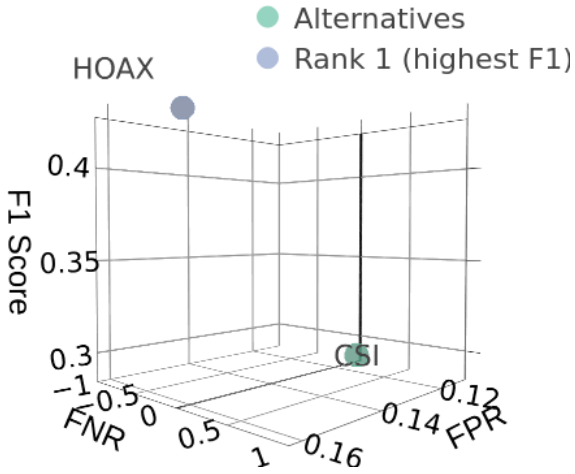
[5] We focus on the topics *scandal* and *election* given that mean differences in error rates between the pair are the highest.



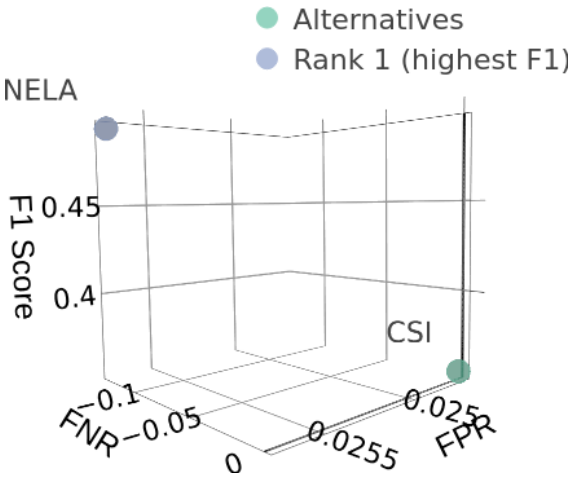
(a) Domain Ideology. Trade performance to reduce bias that favors *liberal-leaning* sites compared to *conservative* sites.



(b) Domain Age. Trade performance to reduce bias that favors *mature* sites compared to *recently created* sites.



(c) Domain Popularity. Trade performance to reduce bias that favors *large* sites compared to *small* sites.



(d) Article Topic. Trade performance to reduce bias that favors articles about *election* compared to those about *scandal.*

Figure 3.5: Error Bias and Performance Trade-off. Here, the z-axis denotes F1 scores, the x-axis and y-axis denote false-positive and false-negative error-based bias respectively. Further, the model $c\star$ with highest $F1$ score is colored in blue, and alternative models with worse $F1$ scores but lower bias than $c\star$ are in green.

to $c\star$ are excluded from analysis. Next, we plot the $F1$ scores and error bias measurements of $c\star$ and C_{alt} in Figure 3.5. Here, the z-axis denotes $F1$ scores, the x-axis and y-axis denotes fnr and fpr based bias respectively. That is, the x-axis and y-axis values are equivalent to the mean differences calculated in Section 3.4.2 and summarized in Figure 3.3. A higher absolute value of x and/or y implies a higher bias. We note that results for event shocks are omitted given that none of the alternative models provides a notable reduction in bias.

As shown in Figure 3.5a which focuses on ideological bias, *HOAX* has the highest $F1 = 0.42$, yet it has $fpr = -0.40$ and $fnr = 0.57$. In other words, *hoax* erroneously classifies articles published by mainstream conservative-leaning news sites as fake news 40% more often compared to articles by mainstream liberal-leaning publishers. Further, it also misclassifies articles by liberal-leaning fake news sites as credible 57% more often compared to articles by conservative-leaning fake sites. We can reduce this bias by choosing the alternative `NELA`, which has $F1 = 0.39$, $fpr = -0.10$, and $fnr = 0.0$. In other words, by trading a small reduction of $F1$, we can significantly reduce ideological bias. Focusing on Figure 3.5b, we see that trade a small drop in $F1$ from 0.42 to 0.39 can lead to a modest reduction in domain age-based bias. For domain popularity (Figure 3.5c) and article topic (Figure 3.5d), however, any reduction in bias requires a substantial drop in performance.

3.5 Discussion

We reviewed 23 existing fake news classification models and provided a comprehensive overview of the current state of this research field. Furthermore, by reaching out to the authors of 17 papers, we collected a representative set of 5 classifiers that we used for additional performance evaluation and error analysis. The results reveal important concerns about generalizability. Performance of fake news classifiers varies significantly from one dataset to another, from one evaluation archetype to another, and from one evaluation metric to another. We also observed important bias: articles from small and/or conservative-leaning mainstream sites, for example, were more often labeled incorrectly as fake news. Furthermore, we also showed that model error rate varies across different topics. Finally, we illustrated that, in some cases, we can trade the model that has the best overall performance but high-bias with another that has a slightly worse performance but a substantially lower bias.

There are several limitations to our work. First, we used the list of fake and mainstream news sites provided by *Media Bias Fact Check*. Yet, many other sources such as *News Guard* provide domain-level veracity labels. Related work [32] has highlighted that the choice in ground truth labels affects downstream observations. As such, future work should evaluate models using different ground truth of fake and mainstream news sites to ensure robustness. Similarly, certain

sources [285] also partition fake news domains into more fine-grained subcategories (e.g. junk science, state-sponsored misinformation sites, clickbait). Understanding whether existing models perform better on some subcategories compared to others can also provide valuable insights into potential model bias and weaknesses.

Despite these limitations, we believe this work takes an important step towards reproducibility and replicability in fake news classification. We invite other scholars to build on this effort and help collectively build towards a well-formulated and comprehensive evaluation framework for fake news detection. As a very first step, we argue that our community needs to make datasets and code repositories available to others. In our case, we were able to acquire only 5 out of the 17 published classifiers. The code repositories for the 5 classifiers and our own code is available at https://github.com/lbozarth/fakenewsMPE. Easier access would allow our community to compare and contrast different datasets and algorithms. Such access would also enable our community to develop a robust evaluation framework more quickly.

A simple guideline on model evaluation: Here, we provide a simple checklist for researchers interested in building robust and effective models. (1) Models should be evaluated using multiple datasets and ground truth labels provided by various parties and may be of varying granularity. (2) Individuals interested in creating a model with the intent of long-term use should adopt variations of the *forecast* archetype to train and validate the model. Similarly, individuals should use the *bydomains* archetype if the aim is to label never-before-encountered websites with unknown factualness. (3) Focusing on metrics, researchers can also consider using the precision-recall curve [29] to optimize precision over recall (or vise versa) if a model's particular usage warrants it (e.g., favoring precision over recall if the impact of a false-positive is high). (4) Ingrained prejudices are shown to be present in machine learning models used in real-world applications [190]. When such biases come to light, they can significantly raise the public's distrust in automatons. As such, using domain-expertise to identify potential high-cost biases and evaluate the model for these biases are crucial. Finally, (5) if the highest-performing model is indeed biased, researchers should consider a trade-off between maximizing overall performance and reducing potential biases.

We note that this guide is far from being a comprehensive framework. Indeed, it is commonplace machine learning best-practice guidance. Yet, our analysis revealed that these simple guides have yet to be fully followed in this important area of research so far.

CHAPTER 4

Study III (Crowdsourced Wisdom): Leverage the Crowd for Covid-19 Misinformation Detection on Reddit

4.1 Introduction

The wisdom of the crowd, or the collective knowledge of a group of people, was found to match or even exceed the knowledge of experts in a wide range of domains, including financial markets, education, and health services [70, 4, 238, 7]. With the rising apprehension about the widespread misinformation and its costly consequences [145], many recent studies have also explored ways to leverage crowd wisdom for misinformation detection [261, 196, 72, 22, 12, 247]. Broadly, researchers have evaluated whether crowdsourced workers can substitute professional fact-checkers and identify misinformation at scale [196, 72, 22, 12], and whether crowd signals available on social media platforms can be aggregated to detect misinformation [151, 251, 136]. These crowd signals include *crowdsourced flagging* which refers to users using platform-wide report systems to flag (i.e., report) content as misinformation to platform stakeholders (e.g., platform admins), and *crowdsourced fact-checking* which is defined as users fact-checking other users' posts via commenting (e.g., a user replies to another user's content with "this is fake"). Crowdsourced flagging signals are only available to platform stakeholders (e.g., platform admins). In contrast, crowdsourced fact-checking signals are public but not always easily accessible to platform admins unless they discover the comments. In practice, Facebook has relied on crowdsourced flagging to identify potential misinformation on its platform [199], and Twitter has recently introduced the BirdWatch program to facilitate crowdsourced fact-checking [202].

Thus far, most existing [199, 202] and theorized applications [196, 72, 22, 12, 247, 136] of crowd wisdom for misinformation detection are from the perspective of platforms. Yet, platforms like Facebook and Reddit often offload misinformation moderation to their communities [114]. Currently, however, there is a lack of knowledge on the role and value of crowd wisdom in misin-

formation moderation for self-governing online communities, particularly from the perspective of community moderators.

Indeed, online communities vary significantly in their sizes, purposes, rules, norms, and user bases [79, 66, 41]. This suggests that the characteristics of crowd wisdom for each community vary as the quality and availability of crowd wisdom are directly impacted by crowd composition [12]. Likewise, the decision-makers (i.e., community managers, moderators) of these communities also differ in their personal beliefs, moderation philosophy, and expertise [129, 222]. These differences may affect how they perceive and moderate misinformation. Given the extent of variance in online communities, *can we find shared practices of misinformation moderation across these diverse online communities? Further, what is the role of crowd wisdom in the moderation workflow? More importantly, how can we help platform stakeholders (e.g., platform admins, and community managers) improve the existing moderation process based on such understanding?*

To tackle these research questions, we focus on Reddit—a platform that is uniquely fitted for the study of crowd wisdom, given the heavy reliance of the platform on communities (i.e., crowds) themselves to self-moderate. Further, we also focus on COVID-19 misinformation, given that many Reddit moderators had publicly expressed that COVID-19 misinformation was rampant on Reddit, platform-wide actions against it weren't sufficient, and that they had significant difficulty moderating COVID-19 misinformation [5, 54].

Here, we conduct semi-structured interviews with 18 Reddit moderators to understand i) how moderators conceptualize misinformation (e.g., how they decide whether a post is misinformation), ii) their moderation workflow and practices, and iii) the role of crowd wisdom in the workflow. Further, we also seek moderators' thoughts on novel moderation queue designs that aggregate various public crowd signals to improve the moderation process. These crowd signals include crowdsourced fact-checking and moderation actions of other moderators, which are not easily accessible in the current Reddit moderation queue's interface (see detail in Section 4.3.4). Using design probes, we iv) explore these crowd signals' potential use and caveats. Finally, we v) compare the values of these crowd signals to labels from professional fact-checkers.

Our analysis reveals a general workflow model that encapsulates the moderation processes adopted by most participants to moderate COVID-19 misinformation. Further, we show that the moderation process revolves around three elements: content facticity, user intent, and perceived harm. The earlier part of the moderation process is centered on identifying misinformation content (i.e., content facticity) and users who are intentionally spreading false narratives (i.e., user intent), And, moderation actions taken in the latter part of the process are affected by all three elements. Next, we show that Reddit moderators broadly rely on two types of crowd wisdom for misinformation detection: the wisdom of ordinary users, and the wisdom of other moderators. Specifically, we observe that almost all Reddit moderators are heavily reliant on crowdsourced flagging by or-

dinary users to come upon potential COVID-19 misinformation (i.e., the “maybe pile”). When encountering difficult-to-identify, ambiguous cases, moderators often seek support from their own modteam or from expert moderators of other communities. Nevertheless, we also observe significant issues with crowd wisdom in the current moderation workflow (e.g., user abuse of the report system). When asked their thoughts about new moderation queue designs, close to half of the participants preferred crowd signals over labels from professional fact-checkers. This is largely because crowd signals can assist participants with evaluating user intent. Additionally, a quarter of the participants distrust professional fact-checkers, raising important concerns about misinformation moderation. Finally, we synthesize the findings to conclude with concrete strategies such that platform stakeholders can better leverage crowd wisdom for misinformation moderation.

4.2 Related Work

Online governance is a complex, multi-faceted process that involves many different actors and competing interests [90]. Past work has demonstrated that the success of self-governance is critically dependent on the contributions of both moderators and users in these communities. Here, we first review existing studies focused on the moderation practices of community moderators. We then discuss possible implications for misinformation detection based on the findings from these past studies. Next, we discuss the wisdom of laypeople and its uses in misinformation detection. We observe that most of the related studies were from the perspective of platforms and users, and there is minimal work exploring how moderators use crowd wisdom for misinformation detection. Nevertheless, we hypothesize the extent to which past findings may be applicable to our work.

4.2.1 The Moderation Practices of Community Moderators

Moderators’ Roles and Moderation Practices

Community moderators have many roles and responsibilities. These include making rules to facilitate positive community norms, collaborating with other moderators, regulating problematic content and user behavior, and even undertaking collective action to promote platform-wide changes [223, 222, 165, 166]. Past work by Seering et al. [223] suggested that community moderators spend most of their time regulating misbehavior, and facilitating open/constructive discussions. Additionally, the occurrence and strictness of moderation are dependent on the community’s goals and individual moderator’s moderation philosophy and personal values [79, 223, 222]. Indeed, what’s considered rule-breaking in one community may be viewed as normal user behavior in another [79, 41]. Further, work by Seering et al. [222] showed that some moderators viewed themselves as the “referee" or “juror", emphasizing their role in overseeing and regulating com-

munities. Others referred to themselves as a “neutral representative" who only stepped in when absolutely necessary, while a small group of “anti-censorship" moderators preferred minimal or no moderation.

These studies drive us to posit that subreddit goals and moderators' moderation philosophies likely affect misinformation moderation. For instance, “juror" moderators are likely to be more actively taking actions against misinformation posts and the corresponding posters, whereas “anti-censorship" moderators are likely to be avoidant of punitive moderation actions. Similarly, some communities may be specifically created to host high-quality content and conversations, and, therefore, require moderators to be more strict against misinformation [278].

Facilitating the Moderation of Problematic Content and User Behavior

Past work [135, 166, 278] demonstrated that moderators employ various approaches and mechanics for moderation. For instance, most moderators introduce community-specific rules (e.g., no off-topic or low-quality content), and then promote user behavior aligned with community values through the enforcement of these rules [114, 79, 41, 278]. Moderators also rely on various technical systems—native to the platform or third-party supported—for routine tasks [272, 121]. As an example, both Reddit and Twitch have auto moderators to automatically flag or remove problematic content. Third, community moderators often rely on good-faith members of the community to set positive examples, promote community values, and assist with moderation [135, 128]. Though, controversial communities often attract bad-faith users who actively hinder moderation [57]. Finally, past work showed that moderators' reliance on these methods is dependent on the community-level characteristics [114, 40]. For instance, larger communities rely on more algorithmic methods, such as bots, for moderation [221], while smaller communities are moderated chiefly through manual efforts [113]. And, niche, highly-specialized subreddits such as r/science are reliant on expert community members to assist with moderation, such as identifying misinformation [128].

We explore the extent to which moderators of different communities rely on crowd wisdom for misinformation detection as opposed to algorithmic approaches or other means. We posit, based on past work described above, that larger subreddits with high volumes of new content are likely to be more reliant on algorithm approaches to detection misinformation than crowd wisdom [221]. Likewise, we speculate that small subreddits are largely dependent on manual efforts for moderation. Further, past work by Gillespie [90] showed that bad-faith users often purposefully generate erroneous crowd signals to mislead platform stakeholders. As such, we hypothesize that controversial communities (e.g., r/political_revolution) that attract norm-violating users are less able to rely on crowd wisdom for misinformation detection [57].

Next, community moderators face many moderation-related challenges, and past literature has

explored various ways to assist community moderators with regulating problematic content and user behavior [40, 122]. This includes studies that demonstrated making rules more transparent and salient can reduce problematic user behavior [122, 135, 167]. Other researchers have also explored ways to improve moderation through expanding existing algorithmic tools [40, 121]. Notably, Chandrasekharan et al. [40] introduced an advanced auto moderator that leverages cross-community learning. That is, it aggregates moderation decisions made by moderators of many different communities to learn and predict the moderation decisions of a given community. Unlike these related studies, we aim to improve moderators' misinformation moderation process through better leveraging crowd wisdom. The potential advantages and caveats of using crowd wisdom for misinformation detection are discussed in detail in Section 4.2.2.

4.2.2 Crowd Wisdom in Online Communities

Existing work focused on leveraging crowd wisdom for misinformation detection can be broadly assigned into two categories. The first [261, 196, 72, 22, 12] focused on the wisdom of *synthetic crowds*. These studies employed laypeople from crowdsourcing platforms (e.g., Amazon Mechanical Turk) to create crowds for content labeling tasks. Thus far, past work demonstrated that labels provided by synthetic crowds, when aggregated, are highly correlated with those provided by professional fact-checkers [196, 72, 12, 213], particularly when a crowd is balanced, and/or has received constructive treatments [196, 12, 210]. The second category of studies examined the value of *organic crowds*: communities of ordinary users present on social media platforms. These studies demonstrated that organic crowds provide two types of signals to assist platforms with misinformation detection. The first is crowds using the platform-wide report button to flag (i.e., report) posts as misinformation, which we refer to as *crowdsourced flagging*. The second is users commenting on other users' posts and calling out these posts as misinformation (e.g., a user replies to another user's submission and says, "this post is fake"). We refer to the second type as *crowdsourced fact-checking*. Crowdsourced flags serve to directly alert platform stakeholders and are only available to such stakeholders. In comparison, crowdsourced fact-checking posts are public, but platform stakeholders like community moderators may not be aware of them unless they stumble upon the related threads. Below, we describe these two types of crowd signals in detail, and discuss how they might assist community moderators with misinformation moderation.

Crowdsourced Flagging

Crowdsourced flagging is one of the most common ways for platform stakeholders to solicit help from ordinary users to identify problematic content [90]. Past work argued that crowdsourced flagging provides a practical mechanism that scales with the magnitude of online content [51,

90]. Moreover, it also affords stakeholders such as moderators the legitimacy to remove flagged content [51]. In recent years, social media giants, including Twitter, Facebook, and Reddit, all updated their reporting systems to enable users to flag content as misinformation. Though, there is limited transparency in how these platforms use crowdsourced flags. Notably, Facebook stated that it uses flags as one of the signals to determine whether content needs to be reviewed by third-party fact-checkers [199].

Prior work also identified meaningful caveats with crowdsourced flagging [90, 49, 220]. This includes findings about flags providing "little room to express the degree of concern, contextualize the complaint, or take issue with the rules" [90]. Further, flags may be biased due to user ideology and the users who regularly flag may not be representative of the user base [90, 49]. Additionally, flagging can be gamed or abused by bad actors [220]. These caveats likely affect the accuracy of flags. As an example, a recent Twitter internal study suggested that only 10% of COVID-19 content flagged as misinformation were in violation of Twitter policies. Researchers suggested that flag abuse can be mitigated through various means, such as distinguishing high-quality flaggers from low-quality ones [219].

Thus far, existing studies largely focused on the values and caveats of crowdsourced flagging from the perspective of platforms. Unlike platform admins, moderators of communities embedded in these platforms have fewer resources to address flag abuse due to more limited administrative privileges [90], highlighting an important challenge. Yet, work that explores how moderators use crowdsourced flagging in their moderation practices is sparse [62]. Notably, Diakopoulos and Naaman [62] interviewed moderators of online newsrooms, and their study suggested that communities with diverse user bases who do not share the same interests and values tend to experience flag abuse. For instance, a user may flag content containing viewpoints that they disagree with but does not necessarily violate community rules. This result is related to past work on trolling and group conflicts [57, 67] which demonstrated that moderators of controversial or minority communities often experienced disruptions from non-community members. Given the contentiousness of politics [13], we hypothesize that crowdsourced flagging is less accurate and more frequently abused in partisan communities, especially if these communities attract users of opposing ideologies. In contrast, apolitical, mainstream communities are likely to have higher quality flags.

Crowdsourced Fact-checking

Ordinary users also fact-check misinformation posted by other users via commenting. Thus far, the prevalence of crowdsourced fact-checking posts is not well understood. For instance, results from Micallef et al. [172] demonstrated that ordinary users generated the vast majority (96%) of all fact-checking tweets on Twitter. This is primarily because the number of ordinary users far exceeds the number of professional fact-checkers. Another study by Jiang et al. [125] showed that ap-

proximately 15% of the posts containing false claims made on Reddit had at least 1 crowdsourced fact-checking comment. Further, results from Achimescu and Chachev [3] suggested that 5% of all posts on Reddit had at least 1 crowdsourced fact-checking comment. In [258], the authors Vo and Lee identified Twitter users who have been actively engaging in fact-checking, and explored ways to increase these users' fact-checking activity levels through recommendation systems.

Scholars suggested that crowdsourced fact-checking posts can serve to correct viewer misconceptions and curtail the spread of misinformation [224, 258, 179], particularly when the refuting sources are reputable, or when the user who wrote the fact-checking post is perceived to be credible. Though, in practice, crowdsourced fact-checking is not always appreciated. For instance, Parekh et al. [193] explored the reception of crowdsourced fact-checking across three different subreddits: r/politics, r/the_donald, r/hillaryclinton. They observed that it was more prevalent and better received by users in r/politics than r/the_donald and r/hillaryclinton. Based on the results, we posit that crowdsourced fact-checking is less available and potentially less effective for partisan online communities.

Here, we observe that related work primarily focused on the value of crowdsourced fact-checking from the perspective of ordinary users. Similar to crowdsourced flagging, there is limited work that explored whether moderators consider crowdsourced fact-checking when moderating misinformation. We argue that this is largely because, unlike crowdsourced flagging, crowdsourced fact-checking is not readily available to moderators through platform mechanisms. Indeed, platform report systems are well integrated into the existing moderation pipelines, whereas additional, non-trivial algorithmic means are needed to identify fact-checking posts [125, 3, 258]. In this paper, we discuss the feasibility of moderators leveraging crowdsourced fact-checking, and also introduce related design probes (see Section 4.3.4.2). We use these design probes to better understand whether crowdsourced fact-checking can assist moderators with misinformation detection. Additionally, we explore whether, similar to ordinary users, moderators also value fact-checking posts containing reputable sources and are written by credible users.

4.3 Method

4.3.1 Research Context

This project uses semi-structured interviews to understand how Reddit moderators regulate misinformation. The Reddit platform is uniquely fitted for studies focused on crowd wisdom. Unlike other social media giants such as Twitter and Facebook, Reddit adopts a more decentralized approach to content regulation [114]. While it had, in rare circumstances, quarantined and banned problematic subreddits and users en mass in the past, it is generally hesitant and slow to take

content moderation actions [165, 114, 222]. Instead, it heavily relies on communities called subreddits to self-govern. Though, it provides some tools (e.g., spam filter settings and crowd control functionalities) to facilitate community-specific moderation [166, 222].

Ethical Considerations: Our study has received an IRB exemption for posing minimal risks of criminal or civil liabilities to participants. Nonetheless, research with human subjects often risks exposing participants to unintended consequences. To mitigate harm, we pseudonymized the subreddits that our participants moderate. Furthermore, we also provide aggregated participants' demographic data instead of providing the exact information.

In the following sections, we first illustrate the process we adopted to generate the list of subreddits that we used for purposeful sampling, and our recruitment procedure. We then summarize the subset of moderators (and the subreddits they moderate) that participated in our study. Next, we describe our data collection process which includes both our interview protocol and design probe. Finally, we discuss our qualitative data analysis approach.

4.3.2 Subreddit Selection and Recruitment

4.3.2.1 Subreddit Selection Pool

Reddit has over 3 million subreddits [1]. Our goal was to focus on subreddits of reasonable size where COVID content has been actively moderated and where COVID misinformation could have been an issue. To do so, we narrowed down the sample pool to subreddits that: 1) had ≥ 1000 subscribers (approximately 98th-percentile), 2) frequently hosted COVID-related posts, and 3) contained COVID-related posts that subreddit moderators took moderation actions on.

We used two processes to identify subreddits that contained frequent discussions around COVID. First, we identified the subset of subreddits where COVID was the primary discussion topic (e.g. r/coronavirus, r/china_flu) by extracting the subreddits that contain at least 1 of the COVID-related keywords in their subreddit description field [2]. This list of keywords was obtained through related work [44, 133, 237]. Then, we manually reviewed all matched subreddits and filtered out subreddits that were not specifically about COVID. We also filtered out subreddits that were private, restricted, quarantined, or banned (e.g., r/nonewnormal). Our second process aimed to find subreddits that were not specifically about COVID but where COVID-related content nevertheless frequently occurred (e.g., r/health, r/massachusetts). To do so, we first parsed all submissions posted in all subreddits between April 2020 and June 2021 using `pushshift.io`.

[1] See the full list of subreddits: `https://frontpagemetrics.com/list-all-subreddits`

[2] Full keyword list: {covid, covid-19, covid19, covid_19, corona, coronavirus, corona virus, wuhan virus, wuhan, wuhanvirus, china virus, chinavirus, 2019ncov, ncov19, ncov2019, sars-cov-2, ncov, kungflu, pandemic, lockdown, outbreak, quarantine}

We then identified subreddits where at least 1000 submissions and at least 5% of all posted submissions in the time period were related to COVID. A submission was considered COVID-related if it contained at least one of the COVID keywords. We also filtered out adult 18+ subreddits (e.g., r/bostonr4r). These two processes collectively resulted in 572 candidate subreddits.

Next, we removed all subreddits without significant COVID moderation. Any subreddit that had fewer than 10 COVID submissions removed by moderators was taken out of the sample. This resulted in a final selection pool of 424 subreddits.

4.3.2.2 Recruitment Process

We used stratified sampling across three distinct subreddit attributes of interest to select participants for recruitment. These attributes are described below:

Subreddit Type: We assigned subreddits into two groups: COVID specific subreddits (e.g. r/coronavirus, r/china_flu) or non COVID specific subreddits (e.g., r/health, r/stimuluscheck, r/massachusetts). We observe that 67 or 15.8% of the 424 subreddits are COVID-specific.

Subreddit Size: The largest subreddit in our selection pool had 2.4 million Redditors, the smallest had 1.3K. Prior literature showed that large subreddits commonly use automated approaches to address content that contains obvious violations; moderators of small subreddits voiced less need for it [121]. As such, we expected the subreddit size to play a role in the degree to which moderators rely on automated, computer-aided or manual approaches to identify misinformation. To capture this nuance, we assigned subreddits into small ($\leq$33.3-percentile), medium (33.3 to 66.7-percentile), and large ($>$66.7-percentile) based on each subreddit's percentile ranking of number of subscribers. Specifically, we classified subreddits that had fewer than 45K users as small, subreddits with 45K to 105K users as medium, and subreddits with more than 105K users as large.

Subreddit Ideological Leaning: Prior studies identified ideology as a significant factor in how individuals view COVID-related misinformation [38]. Additionally, past work showed that users' ideological biases could impact flag quality [49]. Here, we used the method proposed by [240, 205] to compute the ideological leaning of subreddits. Briefly, given a subreddit *s*, we defined liberal-leaning users in *s* as users that i) had posted more frequently in liberal subreddits (r/politics, r/liberal, r/progressive, r/democrats) than conservative subreddits (r/the_donald, r/conservative, r/republican), ii) had a higher average karma score in liberal subreddits than conservative subreddits, and iii) their average karma score in liberal subreddits was greater than 1 [205]. We also defined conservative-leaning users using a comparable definition. We then calculated the ideological-leaning of *s* using the difference between the numbers of liberal and conservative-leaning users, divided by the total number of subscribers of *s*. This metric was then standardized to a scale of [-1, 1] across the 424 subreddits in our selection pool. We refer readers to the orig-

inal papers [240, 205] for a detailed description of this approach. Here, we labeled a subreddit as conservative-leaning if its userbase was more conservative than 95% of all subreddits in our selection pool. Similarly, a subreddit was labeled as liberal-leaning if its userbase was more liberal than 95% of all subreddits in our pool. We observe that subreddits including r/republican, r/askthe_donald, r/conservatives, r/lockdownskepticism, r/ccp_virus are conservative-leaning, and subreddits such as r/political_revolution, r/coronavirusmn, r/joebiden are liberal-leaning. Additionally, subreddits such as r/vancouver, r/COVID19, r/miami are not ideologically aligned. Finally, we note that a subreddit's moderators' ideology is not strictly aligned with the majority of its userbase.

We stratified the 424 subreddits according to the three dimensions listed above. This resulted in 18 (type $\times$ size $\times$ ideology) distinct strata. Subreddits were then assigned into waves; each wave containing one subreddit from each stratum. While each wave is not a representative subset of all the subreddits, sampling from each stratum allows us to examine the moderation practices of more diverse communities. Note that wave size varies because certain strata had fewer subreddits than others (e.g., we had zero large, conservative-leaning, COVID subreddit, and only one large, liberal-leaning, COVID subreddit).

Finally, we invited moderators to participate in the study using five waves. For each subreddit, the moderation team was contacted through *modmail*, Reddit's system for messaging moderators. Responses from the four participants in the first wave were used to fine-tune the interview protocol prior to subsequent waves. Altogether, 78 subreddits were contacted between March and May 2022, which led to a total of 18 interviews (response rate of 23%).

4.3.3 Participants Description

Tables 4.1 and 4.2 contain the descriptions of the mods and subreddits that participated in our interview. As shown, we interviewed a total of 18 moderators accounting for 20 subreddits (note that during the interview, a couple of the interviewees discussed 2 distinct subreddits that they are moderating, both of which experienced COVID misinformation). As shown in Table 4.1, half of the subreddits are COVID-specific subreddits. Further, there are 8 large subreddits, 8 medium subreddits, and 4 small subreddits. We also observe that half of the subreddits are liberal-leaning, 9 have no particular leaning (not-aligned) or are apolitical, and only 1 is conservative-leaning. Moderators from conservative-leaning subreddits overwhelmingly declined to be interviewed. Additionally, the vast majority of COVID-specific subreddits are not ideologically aligned.

Next, Table 4.2 contains a summary of the participants we interviewed (See 4.3.4.1). To ensure participant privacy, we only provide aggregated demographic information. The average participant age is 39, and only 1 person identifies as female. Most of the moderators reside in the United

Table 4.1: Subreddit Summary.

subreddit code	subreddit type	subreddit size	subreddit user political leaning	description
SR1	covid	large	liberal	U.S. regional covid-specific subreddit
SR2	covid	large	not-aligned	A subreddit for users who have covid-19
SR3	covid	large	not-aligned	A subreddit dedicated to covid-19 monitoring and discussions with a focus on high-quality posts
SR4	covid	large	not-aligned	A subreddit with highly strict rules that's dedicated to scientific news and discussions of covid-19
SR5	covid	medium	not-aligned	U.S. regional covid-specific subreddit
SR6	covid	medium	not-aligned	U.S. regional covid-specific subreddit
SR7	covid	medium	not-aligned	A satirical covid-specific subreddit
SR8	covid	medium	not-aligned	U.K. regional covid-specific subreddit
SR9	covid	small	conservative	A subreddit focused on discussing lockdowns & other pandemic policies
SR10	covid	small	liberal	U.S. regional covid-specific subreddit
SR11	not covid	large	liberal	Liberal political subreddit
SR12	not covid	large	liberal	U.S. regional subreddit
SR13	not covid	large	liberal	One of the largest political analysis and discussion subreddits
SR14	not covid	large	not-aligned	U.S. regional subreddit
SR15	not covid	medium	liberal	Political question answering subreddit focused on user opinion
SR16	not covid	medium	liberal	Political analysis and discussion subreddit with a focus on recent events
SR17	not covid	medium	liberal	A subreddit dedicated to documenting and debunking political conspiracies
SR18	not covid	medium	liberal	A political subreddit focused on a particular candidate
SR19	not covid	small	liberal	A progressive news and talk show subreddit
SR20	not covid	small	not-aligned	U.S. regional subreddit

States, and two participants are located outside of the U.S.. Further, two-thirds of the participants moderate 1 or 2 subreddits, and only 2 out of the 18 moderate 5 or more subreddits. In addition, moderators on average have 2.6 years of moderation experience on Reddit. The least experienced moderator has 2 months of experience, and the most experienced mod has over 6 years. Further, an average participant spends 1 hour moderating per day. Focusing on moderators' COVID expertise, we observe that moderators' self-assessed COVID-19 knowledge score on average is 7 (response to the question, *"How would you rate your level of COVID-19 knowledge? On a scale of 1 to 10 where 1 is no knowledge, and 10 is expert level"*). Surprisingly, self-assessed scores are comparable for moderators of COVID subreddits and non-COVID subreddits. Moderators also state that they obtain COVID information primarily from academic sources, reputable news sites, and government websites. Several mods also obtain information from their social and professional connections. Finally, five moderators (or 27.8%) are working or had worked in the health and healthcare industry.

4.3.4 Data Collection

The research team conducted semi-structured interviews from early March to mid May, 2022. All participants were interviewed via Zoom. We obtained the audio recording of each interview after participants gave affirmative consent. The average interview duration is 1 hour and 10 minutes, the minimum duration is 56 minutes, and the maximum is 1 hour and 50 minutes.

Table 4.2: Moderator (Interviewee) Summary

Mod code	subreddit code	In Health industry	average mod time (hours per day)	mod exp(subs)	mod exp(years)	covid knowledge (1-10)	covid information sources
P1	SR1	No	1.0	5+	3.0	8.0	Mod experience; journal articles
P2	SR2	Yes	1.5	1	<0.5	8.0	Mayo clinic; John Hopkins
P3	SR3,SR4	Yes	<0.5	4	1.5	8.0	Professional training; cohort
P4	SR5	No	0.5-1.0	2	2.0	6.0	State government sources; local news media; Reuters and AP
P5	SR6	Past	<0.5	1	2.5	8.0	Medical journals; New England Journal of Medicine; the Lancet; Google search
P6	SR7	No	2.0	3	1.5	6.0	CDC; institutional research
P7	SR8	No	1.5	5+	2.0	7.0	Prepublishing bio archive
P8	SR9	No	<0.5	2	2.0	7.0	Medics; people in biomedicine; news and preprints
P9	SR10	No	<0.5	1	2.0	8.0	Medical journals
P10	SR12	Yes	1.5	2	4.5	8.0	The CDC or the DHS, state government sources
P11	SR11	No	1.0	1	4.0	6.0	Contact who work for the FDA
P12	SR14	No	1.0	2	1.0	8.0	CDC
P13	SR15	No	0.5-1.0	1	<0.5	5.0	Word of mouth; articles on the internet and Reddit
P14	SR17	Past	2.0	2	3.5	8.5	News sources, PBS, NPR, BBC, and other big news outlets
P15	SR18	No	1.0	5+	6.0	7.5	Own experience; own doctor
P16	SR16,SR13	No	2.5	5+	4.5	4.5	News on reddit; another moderator who's a covid-19 researcher
P17	SR19	No	1.0	1	2.0	6.0	Progressive media
P18	SR20	No	0.5-1.0	2	5.0	7.5	CDC and NIH

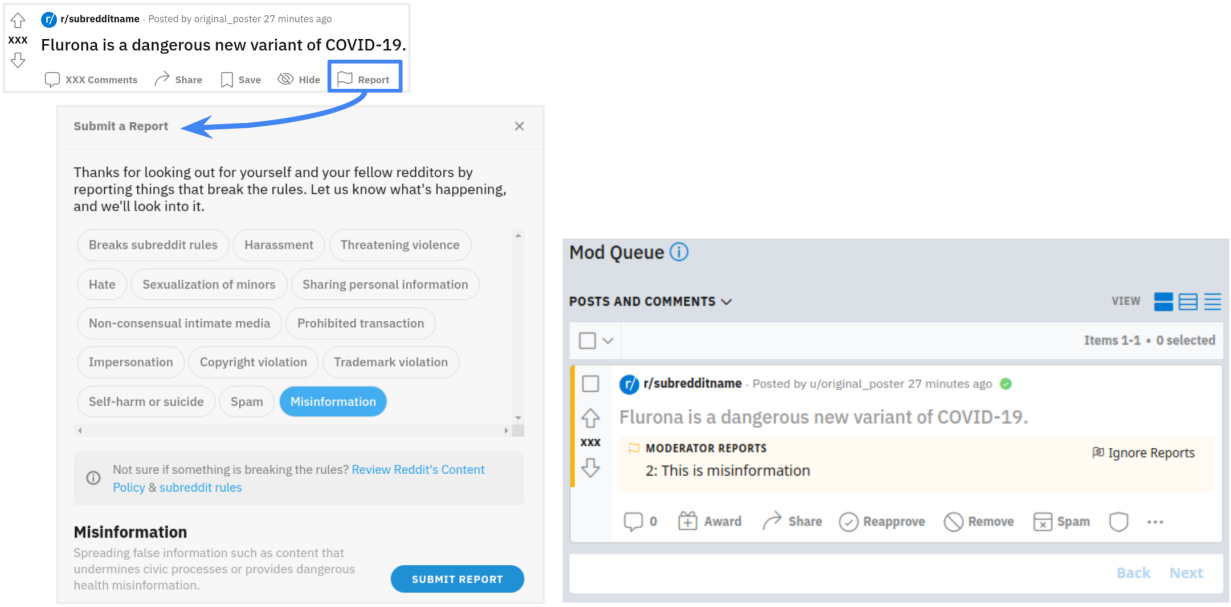
4.3.4.1 Interview Protocol

The initial protocol consisted of seven sections, which explored: (1) the participant's moderation experience and self-assessed COVID knowledge, (2) the extent to which COVID misinformation has been a problem for the subreddit and Reddit as a whole, (3) how potential misinformation come to the moderator's attention, (4) how moderators decide whether content is misinformation, (5) the moderation actions taken by the moderator, (6) whether moderation of COVID misinformation is different from that of other problematic content, and (7) the usefulness of additional crowd-based wisdom, as shown in Section 4.3.4.2. We also collected demographic information from each moderator.

Protocol Update: After the first wave of four interviews, the research team reviewed the high-level observations to discuss whether and how the protocol should be revised. We found that the initial protocol didn't reveal the extent to which moderators relied on manual moderation as opposed to automated moderation. We concluded additional insights were needed to better understand the relative importance of moderators' manual efforts to detect misinformation as opposed to relying on other means, such as automated tools and crowd wisdom, to detect misinformation. Similarly, we observed that moderators relied on content characteristics to identify misinformation and user attributes to identify misinformation spreaders. However, the relative values of the two

(user vs. content) were unclear. As such, we updated the protocol to ask about (8) the relative importance of automated moderation vs. human moderation, and (9) whether the moderators rely more on user account info than submission/comment content in making decisions concerning misinformation moderation. Further, we included additional design updates, which will be discussed in Section 4.3.4.2.

4.3.4.2 Design Probe



(a) Default User Report System

(b) Default Moderation Queue (Modqueue)

Figure 4.1: The current Reddit native implementations of the user report system and the moderation queue. When users report a submission (or a comment) as misinformation (see Figure 4.1a on the left), the reports show up on the moderation queue (see Figure 4.1b on the right). Here, the modqueue indicates that two users reported the submission by *original_user* as misinformation.

In this work, we are interested in whether and how moderators might use additional types of crowd wisdom to detect and moderate misinformation. We use a simple example of misinformation throughout our design probes: "Flurona is a dangerous new variant of COVID-19". We pick an easy and banal example because we want moderators to focus their attention on understanding our widget instead of getting distracted. Here, we first describe Reddit's current report system and modqueue (see Figure 4.1). We then introduce an alternative modqueue design that contains a widget with additional types of information.

Currently, when a user reports a post as misinformation on Reddit (see Figure 4.1a), the reported

content and the number of reports show up on the modqueue as depicted in Figure 4.1b [3].

For our design probe, we introduce a mockup widget that contains three components and additional types of information (see Figure 4.2) that may be helpful to moderators in deciding whether the initial misinformation reports are accurate. The three components are *crowdsourced fact-checking, similar posts*, and *expert labels*. We chose these three categories based on potential usefulness and technical feasibility discussed in prior literature, which we describe below. Our interview process probes 1) to what extent moderators find the widget useful, 2) what additional information can make the widget more useful, and 3) the potential drawbacks of the widget.

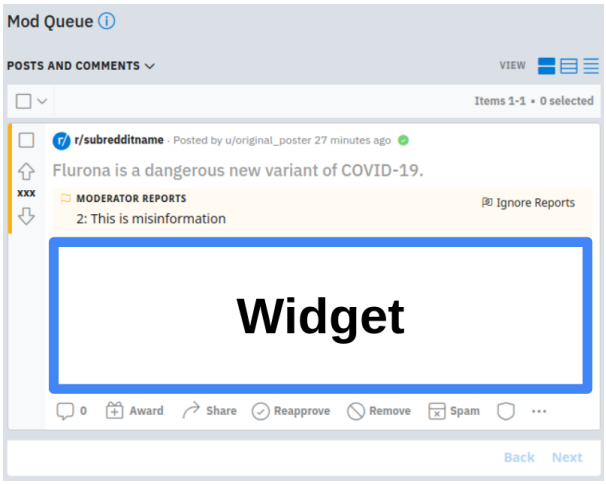
Crowdsourced Fact-checking: Similar to prior work [3, 125], we define *crowdsourced fact-checking* as the comments that reply to the original post and claim that the post is misinformation. For example, as shown in Figure 4.2b, the comments *"This is misinformation. The term refers to simultaneously getting the flu and covid. See [link]*" and *"Fake!*" are both crowdsourced fact-checking comments (Figure 4.2b). Unlike Reddit's native modqueue, with this widget element, moderators can learn which users wrote these crowdsourced fact-checking comments. Further, we posit that the additional text/evidence can inform moderators whether the original post is indeed misinformation [224, 258, 179].

Design Update: After the first wave of interviews, we updated our widget to include a popup that contains user account detail. This was done because in our first wave of interviews, moderators disclosed that they often rely on user account age, total karma score, post history, and other account characteristics to determine whether a user is purposefully spreading misinformation (i.e., bad-faith) as opposed to being genuinely confused (i.e., good-faith). As shown in Figure 4.2e, moderators can click on any username and learn its account age, total karma, most frequented subreddits, most frequently posted websites, and moderator notes [4] on the account.

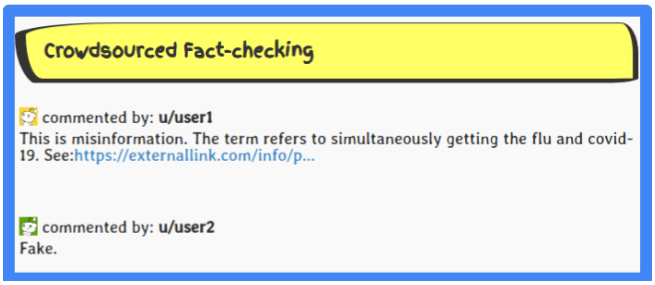
Similar Posts: Misinformation can be shared across many different subreddits by the same user or by different users. As depicted in Figure 4.2c, the *similar posts* component contains posts comparable to the original content that is reported as misinformation. Additionally, it also contains crowdsourced fact-checking associated with these similar posts. If a similar post was removed by a moderator, the moderation action is also shown. Unlike the previous widget component, which relies on the users of a single community to provide fact-checking comments, the *similar posts*

[3]Moderators have access to the number of users who reported the original post as misinformation through the modqueue. However, the reporters' usernames and account information are unavailable. The modqueue also doesn't include any explanation from the reporters as to why they marked the original post as misinformation. This is because Reddit's existing report UI doesn't include any input field (see Figure 4.1a). Moderators can click on the original post to go into the thread to find out more about the context of the reports. Moderators can also click on the reported user to examine their history (e.g., account age, total karma, and previous posts).

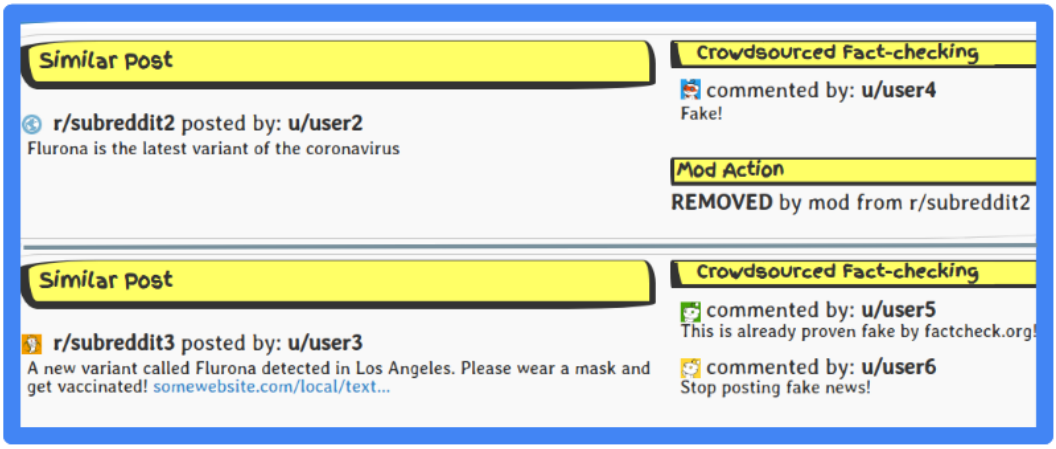
[4]Moderator notes allows moderators of the same community to keep track of the participation histories of members of their communities. See https://mods.reddithelp.com/hc/en-us/articles/4635680764557-Mod-Notes-and-User-Mod-Log



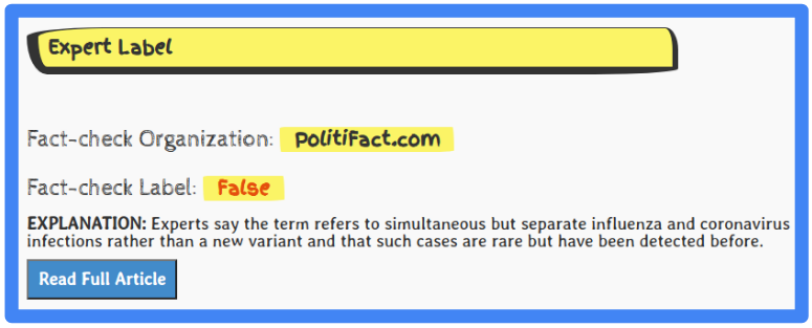
(a) Alternative Modqueue Design



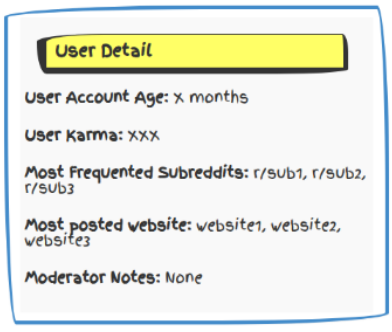
(b) Widget Component: Crowdsourced Fact-checking.



(c) Widget Component: Similar Posts



(d) Widget Component: Expert Labels



(e) User Popup

Figure 4.2: Widget Mockup. The placement of the widget in the modqueue is given in Figure (a). The widget has 3 different components given in Figures (b)-to-(d). Each component includes additional information that may help moderators decide whether a user-reported post is misinformation and whether the poster is intentionally pushing a narrative. Note, moderators can hover over any of the usernames in the components *crowdsourced fact-checking* and *similar posts*, and the user popup element shown in Figure (e) will appear. The elements *user popup* and *mod action* were added in the second wave of interviews.

element gathers relevant crowdsourced fact-checking comments across the entire Reddit platform. In other words, moderators of one community can have access to the crowd wisdom of other communities. This component is motivated by past work [151, 251, 136], which suggests that aggregated crowdsourced fact-checking across entire social media platforms can be used to detect misinformation.

Design Update: For the second wave of interviews, we also included the user detail popup. We also added the *mod action* element based on participants' feedback. If similar content is removed by moderators of other subreddits, our widget will make this action known.

Expert Labels: Thus far both widget components *crowdsourced fact-checking* and *similar posts* contain information from Reddit. Here, as shown in Figure 4.2d, the *expert labels* component includes the label (e.g., "false", "missing context", "true") and explanation from third-party fact-checking organizations (e.g., Politifact, Snopes).

Technical Feasibility of the Widget: Several studies have proposed promising techniques to extract crowdsourced fact-checking using part-of-speech based rule parsing [3, 193], or pretrained deep-learning models [125]. Next, the information needed for the user popup is available via Reddit's native API. Several existing third-party tools, such as Context Bot [5] and Safer Bot [6] already include functionalities comparable to the user detail popup.

For the similar posts component, prior literature has proposed many deep-learning based paraphrasing models to identify similar content [225, 17, 280, 126]. For instance, many past studies relied on the pretrained paraphrase models (e.g., *paraphrase-MiniLM-L3-v2*) from the SentenceTransformer library[7]. Variations of these models have already been adopted by existing work focused on building automated fact-checking systems [225, 17]. Further, efforts like `pushshift.io` also makes it easier to identify moderator removals [8].

Finally, the expert labels component requires a dataset of claims that were already verified by fact-checkers, and a method to determine if a user-reported Reddit post contains a claim that matches one of the verified claims. Currently, there are two large datasets of verified COVID-19 claims extensively used in related work: *IFCN* [172, 14, 241] and *Google Fact Check Tools* [71, 43]. Once we obtain these verified claims, we can again use existing paraphrase models to determine if a user-reported Reddit post contains a claim that's already fact-checked.

[5] https://github.com/FoxxMD/context-mod

[6] Saferbot is a third-party tool that allows Reddit moderators to define a list of problematic subreddits, and then ban users in their own communities that participated in those problematic subreddits https://www.reddit.com/r/Saferbot/wiki

[7] These models can be fine-tuned using sentence pairs $(s1, s2)$ where $s1$ and $s2$ are paraphrases of each other (e.g., "this is a happy person" and "this person is very happy"). See https://www.sbert.net/index.html.

[8] Pushshift fetches and stores the submissions and comments soon after they are submitted. If a post is later removed by moderators, the removed post has the text "[removed]" instead of the original text. Though, it's worth noting that auto-removed posts are not captured by Pushshift.

4.3.5 Qualitative Analysis

The lead researcher (lead coder) transcribed and corrected all interviews via rev.com. We used inductive coding [216] to analyze the transcripts. The lead and second coders independently reviewed and conducted line-by-line open coding of 8 semi-structured interviews (4 interviews from the first wave, and then another 4 from the remaining waves). Then, the two coders met multiple times to discuss and resolve disagreements in order to create one consistent codebook. This process consisted of merging codes into one, creating new subcodes, refining the original code's description, and removing codes that were found to be unnecessary by both coders. The entire research team had several group meetings to review the codes and discuss initial themes. During the discussions, we decided to organize the codes according to moderation workflow activities (see work activity affinity diagram described in work by Hartson and Pyla [104]). For example, top-level codes concerning the automoderator [121] were sorted to be next to each other. The workflow activities were informed by the interview transcriptions, prior studies on Reddit moderators' moderation practices [135, 121, 40], and the lead coder's own experience creating a new subreddit and moderating it.

With the merged codebook, the lead and second coders each independently coded half of the 18 interview transcriptions (note that the 8 initial transcriptions were recoded). During the independent coding process, the two coders regularly checked in with each other to discuss potential updates to the codebook, including adding new codes, updating descriptions of existing ones, or breaking an existing code into several ones. This allowed the codebook to stay consistent. The two coders reviewed all codings to ensure they reflected the final codebook. Next, the lead coder assigned the codes into final themes using affinity diagramming, and the second coder then reviewed each assignment. Finally, the two coders resolved all disagreements to reach a census.

In sum, we had a total of 137 codes. We identified various themes related to i) participants' moderation practices and workflow, ii) the role of crowd wisdom in the workflow, iii) the challenges of relying on crowd wisdom for misinformation moderation, and iv) moderators' perceived potential use and issues of each widget component. We describe these themes in the next section.

4.4 Findings

4.4.1 Moderation Workflow and Practices

In this section, we first describe the general moderation workflow model, which encompasses the moderation practices of almost all subreddits. Further, we also show that moderation workflows commonly revolve around three elements: content facticity, user intent, and harm. Finally, we discuss variations in moderation practices as a result of moderator and subreddit characteristics.

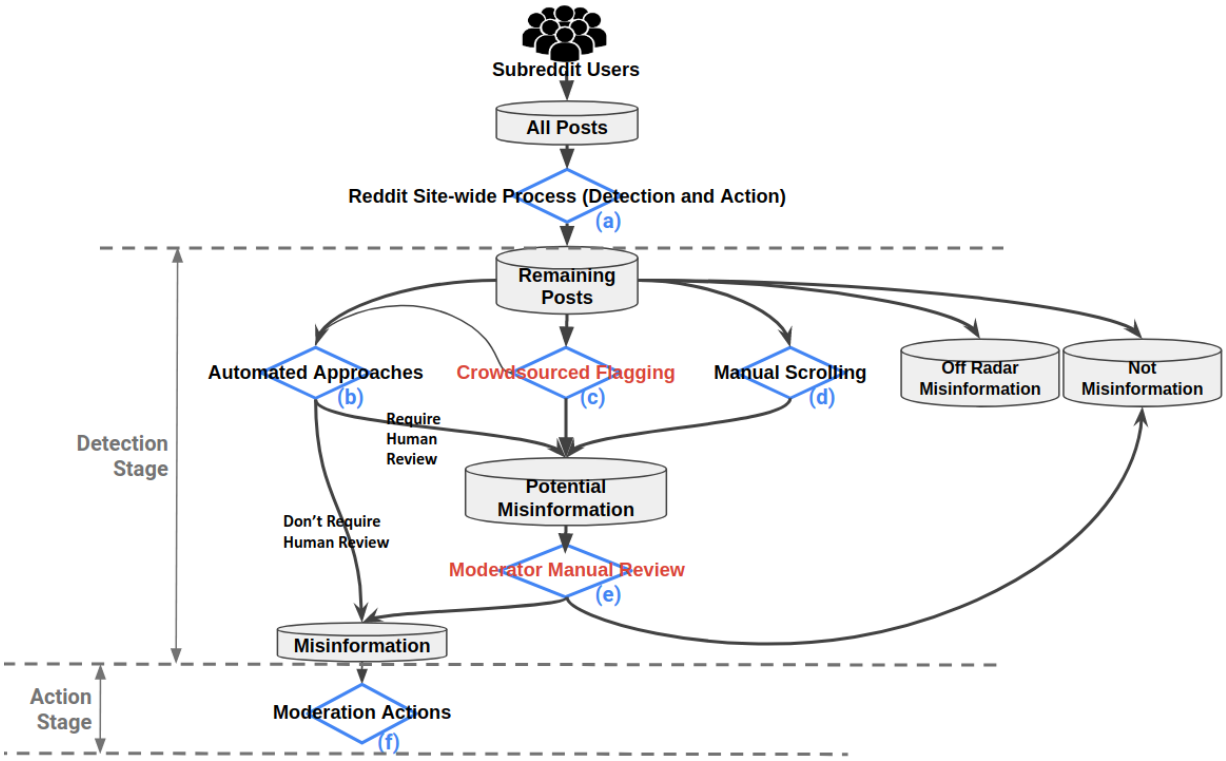


Figure 4.3: Synthesized Moderation Workflow Model. Blue diamonds labeled *a*-through-*f* correspond to subprocesses used by the moderators in misinformation moderation. In the *detection stage*, moderators rely on content and user characteristics to evaluate/classify content facticity and user intent. In the *action stage*, the occurrence and severity of moderation actions are dependent on content facticity, user intent, and harm. The subprocesses with red labels incorporate crowd wisdom.

4.4.1.1 Brief Overview of COVID-19 Misinformation Moderation Workflow

Our interviews reveal a general misinformation moderation workflow model that encapsulates all subreddits' moderation practices except one subreddit, SR15, which does not moderate misinformation. This model is summarized in Figure 4.3. It broadly consists of 6 subprocesses (blue diamonds). Each subprocess is employed by Reddit admins or subreddit moderators. Here, we provide a high-level overview of this model.

As shown, when a user makes a post in a given subreddit, it is immediately analyzed by Reddit's site-wide automated procedures (subprocess (a) in Figure 4.3). For instance, P14 disclosed that, "Certain websites are banned from Reddit from linking ... I know that Great Awakening. That one's banned. You can't link." If this post doesn't invoke any moderation actions from the site-wide algorithms, it can still be identified as *potentially* containing misinformation or its poster can be labeled as a potential misinformation spreader by three distinct subreddit processes (subprocesses (b), (c), and (d) in Figure 4.3). These three processes are i) automated approaches at the subreddit level like the automod, ii) crowdsourced flagging, and iii) manual scrolling (i.e., moderators manually scrolling through their own subreddit and reading the submissions and comments). We discuss these three approaches in detail in Section 4.4.2.1.

Next, moderators predominately rely on manual review (subprocess (e) in Figure 4.3) to determine whether or not the post indeed contains misinformation, and whether or not the poster is indeed purposefully spreading misinformation (we described this more in detail in Section 4.4.1.2). Though, a handful of participants also noted that a few of the automated approaches they used (e.g., Saferbot) can bypass the manual review process.

Lastly, if the post is evaluated to be misinformation or if the poster is perceived to be a misinformation spreader, actions (subprocess (f) in Figure 4.3) are taken against the post and its poster.

Our analysis also suggests that this moderation model can be generalized to other types of misinformation and problematic content. In fact, a majority of the participants noted that their moderation for COVID-19 misinformation is comparable to other types of problematic content. The remaining participants noted that their moderation priority and difficulty differed across different types of problematic content.

4.4.1.2 Moderation Revolves Around Content-level Facticity, User Intent, and Harm

When describing their workflows, two-fifth of the interviewees mentioned that moderation of COVID-19 misinformation is an informal process, or that their moderation practice likely varies from other subreddits due to Reddit not providing any official guidelines or policies on COVID-19 misinformation. Several also wished to have official policies on misinformation and additional technical support from Reddit.

> "Because obviously moderators across the site all have different rules in their subreddits and there are lots of different levels of what moderators feel is misinformation. So I feel like if Reddit stepped up and set specific policies about what people should remove, I think that would be very helpful." -[P12]

Nevertheless, our qualitative analysis reveals that misinformation moderation generally revolves around three elements: *content facticity, user intent*, and *harm*. Interestingly, the three elements emerged from our analysis are largely aligned with the elements of misinformation ecosystem described in [265]. Though, interviewees were focused more on *harm* than the *dissemination mechanisms* of misinformation. We observe that, in the *detection stage* (Figure 4.3), moderators rely on content-level and user-level attributes to evaluate content facticity and user intent. Moreover, in the *action stage* (Figure 4.3), the occurrence and severity of moderation actions are dependent on all three elements. Below, we describe each element separately. However, in moderation, these three elements are interconnected. For instance, a few participants observed that users who post obvious misinformation tend to operate with deceitful intentions.

Content Facticity: When asked "Are there specific types of Covid-19 misinformation that you're particularly vigilant against (very important to identify and moderate)?", one-third of the participants responded that they prioritize obvious, blatantly false claims that contradict facts with an established scientific consensus or are supported by a lot of existing data, such as misinformation about vaccines being dangerous. Some moderators described obvious misinformation as "you will know it when you see it"-[P3]. A couple of interviewees also said that *it* is posts that contain wild claims without providing any evidence to substantiate them.

In the *detection stage* (Figure 4.3), moderators relied on content-level characteristics to evaluate content facticity. A third of the subreddits relied on the automod and keyword regex matching to identify potential COVID-19 misinformation (subprocess (b) in Figure 4.3). For instance, P11 remarked that they "remove all posts that are from Fox News and also have COVID numbers because they've been a major source of misinformation early on." Some participants also considered content-level characteristics during manually scrolling (subprocess (d) in Figure 4.3). For example, a couple of participants asserted that they were more attentive with threads concerning topics (e.g., COVID-19) that were more likely to attract misinformation. Finally, half of the participants mentioned that, during manual review (subprocess (e) in Figure 4.3), if a post contained an URL, they relied on the reputation of the website as a signal to determine whether or not the post is misinformation.

In *action stage* (Figure 4.3), some moderators also took more severe actions against obvious misinformation. That is, the moderators would remove the post and also ban the poster, as opposed to just removing the post. Several participants noted that this type of misinformation is easy to

spot and easy to ban. P18 asserted that their subreddit is more strict with obvious misinformation because the correct information is readily available and, as such, there is little value to having posts or discussions on this type of misinformation.

> "Vaccines were found to be safe, specifically Pfizer, Moderna, and Johnson & Johnson. They were found to overall be safe. To be completely candid, we just don't have time for that anymore, and there's enough of a knowledge base that is freely available." -[P18]

User Intent: Our analysis suggests that user intent is equally if not more important than content facticity. We observe that moderators depended on the following strategies to assess user intent in the *detection stage* (Figure 4.3). First, as a general moderation practice, almost all subreddits only allow users who meet specific account age and karma score requirements to post. Next, a handful of participants, most of whom moderate COVID-19 subreddits, remarked that they relied on several algorithmic tools (subprocess (b) in Figure 4.3) like Saferbot to scan users' post histories to determine their frequented subreddits as a way to identify potential misinformation spreaders. During manual review (subprocess (e) in Figure 4.3), moderators also relied on a variety of user-level characteristics to determine whether or not a user was purposefully posting COVID-19 misinformation. These characteristics, ordered from the most frequently mentioned by the participants to the least, include: account age, total karma, whether the user had been making similar claims in the past in different subreddits, most/recently frequented subreddits, user activity gaps, most frequently posted websites, username, and user ideology. Finally, a third of the participants mentioned that they adopted several existing third-party tools (e.g, Mod Toolkit, Reddit Enhancement Suite) or created custom tools to synthesize important user account and post history metrics to make manual review easier. Particularly, many moderators relied on moderator notes to keep track of users' problematic behaviors.

Next, most of the participants said that, in *action stage* (Figure 4.3), the severity of their moderation actions and their willingness to engage with the poster are dependent on the perceived user intent. Some participants emphasized the educational aspects for good-faith users. These moderators contended that removing a user's post and banning them will never enable this user to obtain the correct information. As a result, the user would remain misinformed.

> "They're genuinely misunderstanding the topic. They are repeating a fact that they heard, thinking that it's true. We haven't tried to do anything, but guide them to the truth through conversation. We don't ban them from the subreddit. We don't remove their messages. We interact with them and try to educate." -[P4]

> "Typically, we'll look at the user's history. And more often than not, it is somebody bouncing subreddit to subreddit and we'll just delete their comment, ban them and just move on; [if] they seem like a legitimate person, we will remove their comment and just let them know, 'Hey, we don't do that here.' " -[P10]

Still, many moderators believed that education efforts are futile for bad-faith users whose goal is to push false narratives. A handful of moderators also claimed that, in their experience, good-faith users who also posted misinformation were relatively rare in their communities.

Perceived Harm: The majority of participants also highlighted that they prioritize misinformation that they perceive would lead to harm at individual-level (e.g., causing a person physical harm) or at societal-level (e.g., causing harm to a collective). Examples of this type of misinformation include those that downplay the effects of COVID-19 (e.g., COVID-19 is just the flu or COVID-19 hasn't killed that many people), discourage others to get vaccinated, or encourage people to pursue "alternative cures" such as hydroxychloroquine or ivermectin.

> "But it is dangerous that you think that [alternative cures] will cure Coronavirus. Because I really felt like this could really lead to someone getting hurt. And so, that's something I was really particular about." -[P17]

It's unclear whether or not harm evaluation is a part of the *detection stage* (Figure 4.3). Though, our analysis suggests that the topic of misinformation and volume are indicative of perceived harm. We also observe that perceived harm varies across the participants. Additionally, what moderators consider to be harmful changes over time. For instance, a few moderators viewed misinformation about masks don't work against COVID-19 to be harmful early on, but less so now. Some of the moderators mentioned that this is because the volume of mask-related misinformation has significantly dropped, and that the official guidelines on masks have also changed over time.

Finally, some moderators stated that, in *action stage* (Figure 4.3), their moderation actions are more severe against instances they perceived to be harmful.

4.4.1.3 Moderator and Community Characteristics, and Variations in Moderation

We also observed noteworthy differences in moderation practices as a result of moderator and community characteristics. Indeed, moderators vary in their moderation philosophies [222]. Most notably, we observed that a few of the participants are strongly averse to punitive actions including removals and bans (subprocess (f) in Figure 4.3). These participants emphasized the need for open discussions and limited censorship. For instance, one moderator referred to themselves as a caretaker rather than an authoritative figure. In contrast, another moderator stated that they had been very strict with COVID-19 misinformation because they believed that they were brought

onto the moderation team to ban people. Further, moderation process is also dependent on the moderators' expertise. In particular, we observed that most of the moderators who are or were employed in the health industry said that, during manual review (subprocess (e) in Figure 4.3), they often relied on their own knowledge to determine content facticity.

Moderation practices are also affected by community characteristics [79]. For instance, P13 emphasized that the goal of their subreddit is to understand why Redditors believe what they believe, and it made little sense to moderate users' opinions even when they contain misinformation. A few of the non-COVID subreddits noted that they have specific rules against off-topic content, and that these rules allowed them to take more aggressive actions with COVID-19 misinformation (subprocess (f) in Figure 4.3) because such content is unrelated to their communities.

Overall, observations here are aligned with many prior studies [135, 79, 223, 222] which found that moderator and community characteristics directly impact moderation practices.

4.4.2 The Wisdom of Two Crowds

In this section, we explore the role of two crowds in the generalized moderation workflow model: ordinary users and moderators. We show that moderators rely on the wisdom of the general crowd to detect potential misinformation (the maybe pile), and the wisdom of other moderators to decide ambiguous cases. Finally, we note that ordinary users and moderators are not two completely separate crowds. In our interviews, several moderators mentioned that expert users who demonstrated good character are often recruited to become moderators.

4.4.2.1 The Role of Ordinary Users

Our analysis shows that ordinary users assist moderators through *crowdsourced flagging* (subprocess (c) in Figure 4.3). Indeed, *crowdsourced flagging* is one of the most important, if not the most important, means used by participants to come upon potential misinformation and potential bad-faith misinformation posters.

> "So there were the users flagging. It is the number one way it [potential misinformation] comes to our attention. Number two is by scanning the background of new users coming into the community, [use Saferbot] to flag people that may be liable to post misinformation. The list of subreddits changes day to day; r/churchofcovid is one of them, r/nonewnormal." -[P1]

> "So things will come to our attention in two ways. Either someone will bring it to our attention by reporting it, because we've got specific report reasons that people can use.

> Or that we'll just be browsing the subreddit because we are still community members at the end of the day." -[P7]

Comparison to Other Methods: Some subreddits also rely on *automated approaches* and *manual scrolling* (subprocesses (b) and (d) in Figure 4.3) to identify potential misinformation. Approximately a third of subreddits relied on the automod to identify posts containing keywords or URLs that are associated with COVID-19 misinformation. Some used the Saferbot and its variations to detect potential misinformation posters. An even smaller number of participants created their own unique algorithmic tools. We observe that large subreddits and COVID-19 subreddits employed more automated methods to identify potential COVID-19 misinformation. This is likely due to them having higher volumes of COVID-19 content. Moderators noted that some of these tools were very useful. Particularly, P3 noted that algorithmic approaches can preemptively capture potential misinformation before other users are exposed to it. Additionally, we also observe that subreddits with limited volumes of new content per day and acceptable moderation workload preferred manual scrolling. Results here are aligned with past work [221, 113] which found that subreddit size is a significant factor in the extent to which moderators relied on algorithmic means for moderation as opposed to manual efforts.

Among the three approaches, crowdsourced flagging is the *most widely used* method for identifying potential misinformation. Almost all subreddits that actively moderate COVID-19 misinformation rely heavily on crowdsourced flagging to find potential misinformation. Indeed, the vast majority of the participants remarked that moderation workload had been a significant challenge in their subreddits. As such, these participants cannot manage manual scrolling. While automated tools are highly scalable, some participants noted that functionalities of many existing automated tools are limited, and that algorithmic tools can be difficult to implement well. For instance, both P7 and P16 noted that algorithmic tools can often result in too many false positives. A few participants also mentioned trying out new tools but finding them too complicated to adopt. In comparison, crowdsourced flagging is well integrated into the moderation pipeline, and in the words of P5, "it's a very straightforward process". These circumstances likely contributed to participants' significant reliance on crowdsourced flagging.

4.4.2.2 Limitations of Crowdsourced Flagging on Reddit

Despite the heavy reliance on crowdsourced flagging, moderators also stressed that it has several significant weaknesses. That is, similar to past work [51, 90], we also observe that crowdsourced flagging is often misused/abused, report rate can be low, and that a flag contains very limited information. However, we also show that flagging-related challenges faced by communities varied significantly. And, our findings provide an additional layer of granularity to past work. We describe

these findings in detail below. Additionally, we also discuss potential improvements derived from the interviews for addressing each limitation.

Prevalent Report Misuse and Abuse: The report system (Figure 4.1a) can be abused (i.e., used fraudulently) by bad-faith users, especially because the system is anonymous. All moderators have reported experiencing *report abuse*. Half of the participants also stated that report abuse is a significant problem, and it takes time and effort to address. Many moderators asserted that the *misinformation* option included in the report system UI is commonly misused as a "big disagreement button"-[p16]. Further, it tends to happen when users are arguing with each other, and one of them decides to report all recent posts made by another user as misinformation.

> "The report button often is used as a, 'Hey, I strongly disagree with this statement button.' So we'll go look in the modqueue and we'll see one user has been reported 20 times and everything's innocuous. Ten reports and it's all reporting the same user going back three days, four different discussions. That's usually what we'll see." -[P10]

Other times, bad-faith users abuse the report system to cause an additional workload for moderators.

> "We started getting a lot of trolling and spamming of moderator reports by people who didn't like the moderation. That's probably the single biggest problem that I've had in the whole process, when you get somebody who's a really dedicated troll who comes in, and it happens periodically." -[P9]

Some moderators also considered that sometimes this can be an honest mistake and the reporter sincerely thought that what they reported was misinformation. For instance, P6 mentioned that people can accidentally report satirical content. Outside of the abuse of the *misinformation* option, some moderators also noted that the report options *harassment*, *self-harm and suicide*, and *sexualization of minors* are also frequently used as a way to harass targeted users, or try to get them banned. One moderator in fact stated that the *misinformation* option had been one of the less abused options in their experience.

Further, as indicated by past work [57, 67], moderators of controversial or partisan communities (e.g., P6) stated that they had experienced rampant report abuse. Interestingly, some moderators of apolitical, mainstream subreddits also mentioned that report abuse is significant in their communities. A possible explanation is that participants have varied perception of *significance* (e.g., a 20% abuse rate could be noteworthy to one moderator but not another).

Finally, some moderators noted that Reddit has been making an effort to address report abuse [9]. However, our analysis reveals that many participants felt existing efforts are insufficient.

Potential Improvements: identifying bad-faith users and limiting their ability to report. Most notably, a third of the moderators desired more transparency in the report system.

> "Let the moderator see who's doing the reporting. Because I strongly suspect that a lot of those types of reports, the frivolous reporting, is the same subset of people. It's the same small group of people doing it." -[P5]

Though, some of the moderators conceded that Reddit admins are probably concerned about retribution, and will likely not implement this transparency. As an alternative, others have also proposed to make available various metrics that can be computed from user account and report history: i) number of reports made recently, ii) whether the reporter is a brand new user, iii) whether the reporter has previously participated in the community. These metrics can be used to inform moderators whether or not a report is made in bad-faith without having to reveal the identity of the reporter. Further, some moderators also proposed to preemptively limit who can report and how many times they can report. For instance, P16 proposed that when a user is banned from a subreddit, they should no longer be able to report on that subreddit. P13 argued that report should not be available for users who have never interacted with a subreddit before, or who fall below a certain amount of activity on the subreddit. Interestingly, one moderator said they would like to have the ability to contact bad-faith reporters to tell them to stop misusing the report system.

Potential Improvements: identifying good-faith users. Some moderators mentioned that they keep notes of regular and expert users in their communities. In addition, some of these users may receive a fact-checker flair [10] or other flairs that are indicative of expertise. Our analysis suggests that moderators give more weight to inputs and reports from these regular and expert users.

> "Some of our longer time users are very keen to report troubling posts because many of them feel very invested in the space too. I talked one on one with a number of them. They are quick to report things." -[P8]

Given that the current report system is anonymous, account characteristics that are indicative of a reporter being a regular or expert user are unavailable. While none of the moderators proposed any specific ideas, we hypothesize that including additional reporter metrics can help moderators determine whether a report is made in good-faith. Some examples include i) whether the reporter

[9] https://www.reddit.com/r/modnews/comments/mlgsw5/safety_updates_on_preventing_harassment_and_more/

[10] User flair is an icon or text that appears next to Reddit usernames. Each subreddit has its own user flairs set up by the community's moderators.

had received a positive flair from the community moderators, and ii) percentage of previous reports made by the reporter that was approved. These metrics can be useful to moderators without revealing the identity of the reporters.

Low Report Rate: Another issue faced by some moderators is low report rate. As noted by P11:

> "A lot of our actions will happen in response to that [crowdsourced flagging]. If a user doesn't report a post, then we don't notice it most of the time. Finding those posts before they get to the front page of the subreddit and get exposed to the wider audience is definitely the hardest part." -[P11]

A couple of the participants speculated that perhaps many users are unaware of the report button and what it does, or unaware that reporting is anonymous. Further, users may also choose to engage with the misinformation poster directly rather than reporting them to the moderators and Reddit admins. That would keep the moderators uninformed of the existence of the misinformation.

Potential Improvements: Moderators have used various strategies to improve crowdsourced flagging rate. For instance, P9 stated that their subreddit has clear rules on the sidebar requiring users to backup their claims using reputable sources, and that users in their community are very good at reporting unsubstantiated posts. Some moderators also adopted (or considered) ways to promote community awareness of the fact that misinformation is taken seriously and encourage users to report more.

> "To get more quality reports, it's possible to have auto moderator sticky a comment on all posts. You could attach a [moderator] comment saying, 'This post is likely to attract a fair amount of COVID-19 misinformation. Please report this kind of thing.' " -[P16]

Results here suggest that transparency and salience of rules and moderation practices can increase report rate [122, 135, 167]. Further, a couple of moderators also mentioned that they thought giving expert users flairs (e.g., a fact-checker flair) may incentivize these users to report more frequently. It's worthwhile to note that past studies [180, 63] have also explored various ways to increase flagging rate by promoting a sense of personal responsibility in bystanders.

Limited Information: Finally, as stated in the previous section, moderation practice is heavily reliant on content-level facticity, perceived harm, and user intent. However, moderators expressed lacking information to efficiently make use of these three aspects, as Reddit's current report system only makes available the number of times a post is reported as misinformation. A few moderators, however, mentioned that the number along is very informative.

> "So if [redacted number] people make a report on a submission or a comment, it gets removed and put into our modqueue for us to look at." -[P4]

Many of these moderators said that if the number of reports has exceeded a certain threshold, the reports are generally accurate. However, some mods also expressed wanting additional information from the reports. For instance, P10 noted that it would be nice to be able to contact the reporter for additional evidence. Result here resonates with prior work that asserts crowdsourced flagging is a thin expression [51, 90].

Potential Improvements: A simple and straightforward way to allow reports to have additional context is to include an input box where reporters can write some text to tag along with their reports. One moderator, P3, noted that their subreddit has an updated report UI with a *custom response* option. P3 said that users had used this option in the past to provide additional information to the moderators. However, not every subreddit has a *custom response* option. More importantly, past work that's focused on using flagging for misinformation detection has been more invested in classifying content facticity [49, 136]. It is unclear whether additional text provided by reporters can help moderators evaluate harm and the intent of the reported user.

4.4.2.3 The Role of Other Moderators

During manual review (subprocess (e) in Figure 4.3), moderators can read the content, view user post history, and use various means to discern content facticity, and user intent. Our qualitative analysis demonstrates that moderators often rely on the wisdom of their modteam for ambiguous cases. Further, some participants also stated that they also rely on moderators from affiliated subreddits, or reputable subreddits to moderate misinformation.

Deciding Ambiguous Cases: Some participants noted that in many cases, it's not difficult to tell whether the content is misinformation, and whether the user is purposefully spreading misinformation. When something is murkier, however, many of the participants would check in with fellow mods for inputs, using Discord or Slack.

> "Generally, we'll talk it over as mods and we try to err on the side of leaving it up. We might even post a moderator comment saying, 'We had a discussion about this and we're really not sure if this is accurate or not, but it's plausible.' " -[P5]

The vast majority of moderators also said that they keep moderator notes on users that could potentially be problematic, such as users who participated in COVID-denial, antivaxxer subreddits. These notes are shared among moderators of the same community.

> "If it's something that's super on the fence where I'm legitimately like,'This person could be doing it in good faith. I don't want to stifle that.' I would usually post it.

> The mod team has a Slack where we can communicate with each other. I would post it there and be like, 'Hey, does anybody have any experience on this? Do you know a little bit more?' If they gave the okay and the user was commenting in good faith and having conversations with people and they didn't have any past history, like have they ever gotten a little slap on the wrist for whatever, then it would probably just stay up. But, I leave notes to the other moderators, 'Oh, I was kind of iffy about this one. If they keep it up in the future, it might be something to look at.' "-[P11]

The notes allow moderators to identify users who repeatedly skirt subreddits' rules. Most of the participants disclosed that their moderation actions escalate for handling these users. For high-stake situations such as *permanent ban*, a few moderators also said that they tend to discuss with fellow mods to reach an agreement.

Cross-community Support: Participants also sought support from moderators of other subreddits. For instance, P7 mentioned that, when evaluating the intent of a user, they would sometimes message moderators of other subreddits that the user participated in. More commonly, there are Discord channels for moderators of various subreddits to "bounce off each other"-[p14]. For example, P1 mentioned that they are a member of a Discord channel for location-based subreddits in the same geographic region. This provided an extra information resource for moderators in that region, and a way to keep track of malicious users who post in those subreddits.

In addition to asking for support, a couple of participants also lent their own expertise to other subreddits. In these cases, the subreddits that sought external help are often non-COVID subreddits (e.g., r/starwars, r/hockey). Moderators of these communities had little experience regulating COVID-19 content, and lacked the knowledge to separate facts from misinformation. The following is an excerpt from the interview with P3, who had helped moderators of other subreddits.

> "What we would do is, we would hop into Discord with them [moderators of other subreddits]. They gave us access to their general Discord that they use for coordinating and planning. And then, if they had questions on whether something that someone was talking about was actually true or not, they could just bring it to us. And, we could take a look at it and answer, 'Yeah, that's accurate to the best of our knowledge, or to the best of the scientific literature.' Or, 'No, that's absolute nonsense.' " -[P3]

Past work observed cross-community interactions between moderators in high-stake scenarios such as collective actions, or when moderators have certain affiliations such as moderating similar communities [165, 166]. Our analysis shows that moderators also seek external support for everyday incidents revolving around misinformation.

Outside of misinformation detection, some participants also noted that they rely on other affiliated moderators to create and implement new subreddit rules, and to tweak their moderation

workflow (e.g., implement auto-ban bots) to target COVID-19 misinformation. Sometimes, moderators (e.g., P1) temporarily volunteered to moderate other subreddits to assist with the influx of user activities. This is consistent with prior work [223], which demonstrates that affiliated moderators rely on each other to develop and update moderation practices.

4.4.3 Moderators' Feedback on Alternative Modqueue Designs

In Section 4.3.4.2, we introduced a design probe of a widget with multiple components and additional types of information to assist moderators with misinformation detection. In this section, we discuss participant feedback in detail. Among our key findings, we show that participants considered *expert labels* to be very useful for identifying content veracity. However, crowd signals are useful for determining the intent of the parties involved. Further, we also show that the perceived usefulness of information is dependent on participants' perceived credibility of the information providers. Moreover, participants also discussed that the data and metrics they currently rely on (or proposed in the design probe) for determining credibility are not always reliable and can be gamed. Finally, we synthesize participants' feedback to highlight some additional improvements to the designs.

4.4.3.1 Comparing the Values of Different Components

Almost all participants found the widget or some components of the widget to be useful for moderating misinformation. The vast majority of participants viewed *crowdsourced fact-checking*, *similar posts*, and *expert labels* favorably, and almost all participants appreciated *user popup* (components given in Figure 4.2). We also observe that *expert labels* were considered to be the most helpful but, surprisingly, only for a mere majority of moderators. Of the remaining participants, one participant thought that the widget is not that useful; the others were evenly divided between *crowdsourced fact-checking* and *similar posts* in terms of which component they considered to be the most useful. More interestingly, several participants also asserted that *crowdsourced fact-checking* and *similar posts* are useful for identifying user intent, something that *expert labels* cannot do.

> "If it's a fact-checking organization you trust, that's a very quick read. Okay. This is a source that I trust. They're deeming it false. I can act on this quickly. 'Is this user problematic?' Then, I would say, *crowdsourced fact-checking* would be a good avenue to go in, to really investigate what's the intent of this user, just based on the sources they are posting and the consistency of their posts. *Similar post* is interesting and helpful for reposts, and just people spamming subreddits." -[P18]

In other words, our analysis demonstrates that the extent to which a type of information is useful is dependent on whether a moderator needs help evaluating content facticity or user intent. Further, we observe that while some participants mentioned that they had more difficulty with subtle misinformation, others had more trouble with skilled trolls. This suggests that all three types of information are needed to cover different use cases.

4.4.3.2 The Role of Perceived Credibility in the Perceived Usefulness of Information

Our analysis shows that the perceived usefulness of signals provided by different entities (e.g., users, professional fact-checkers) is dependent on the perceived credibility of the entities involved. Indeed, we observe that participants preferred crowdsourced fact-checking that i) were posted by regular or expert users, ii) contained links to reputable sources, and iii) explained why the reported content is misinformation.

> "Yeah. That [user reputation] would absolutely add weight to making a decision probably easier. Google does with their reviewers. If I see, like Wikipedia, somebody who's done lots of good articles that I would know, I would rely on their report quicker. Or, if I see somebody has a new account, then I'm gonna look at it a little closer." -[P10]

Results here are aligned with past work [224, 179] which showed that ordinary users also valued crowdsourced fact-checking posts containing reputable sources or were written by users who were perceived to be credible. Additionally, some moderators noted that bad-faith users can also "fact-check". These moderators stated that they are suspicious of fact-checking comments from new user accounts with low karma, and also wouldn't trust crowdsourced fact-checking containing unusual sources or sources masked using URL shorteners. A few participants, however, stated that the number of crowdsourced fact-checking alone is a good indicator. Finally, a couple of interviewees noted that they wouldn't automatically agree with the fact-checking from regular users. However, they would assume that these users contributed in good-faith.

Similarly, the majority of participants thought it is useful to have access to other subreddits' crowdsourced fact-checking and moderator actions on *similar posts*. However, to some, it matters which subreddits they were. A few moderators said that they preferred signals from reputable COVID-19 subreddits such as r/coronavirus and r/covid19. In addition, a couple of interviewees said that signals from subreddits they considered to be unreliable would have an opposite effect.

> "If r/conservative are saying like, "Oh, this is fake." That's going to change how I interpret their comments. Something that they think is fake is what I know is true." -[P11]

Finally, a quarter of the participants had reservations about *expert labels* because they viewed these fact-checking organizations as being too politicized or lacking medical expertise. Many participants also mentioned that it would be difficult to find sources and professional fact-checkers that are broadly viewed as credible. Our own analysis suggests that participants generally trusted academic sources the most, followed by government sources such as the CDC. Though, one moderator remarked that the accuracy of information from government sources is dependent on who wins the elections.

Overall, results here are consistent with the findings in Section 4.4.2, and with prior work [90]. Participants preferred information from entities that they considered to be good-faith and credible. Those signals were also given more weight during the moderation process.

4.4.3.3 Limitations of Crowd Signals: Fabricated Credibility and Over-reliance

Anticipating Bad-faith Users "Gaming the System": Interestingly, participants also anticipated various ways that bad-faith actors can mislead moderators by fabricating credibility. Indeed, some interviewees emphasized that user account attributes, such as account age and karma scores, that moderators have been relying on to evaluate intent and credibility, are not necessarily authentic. For instance, year-old Reddit accounts with moderate karma scores can be purchased cheaply en mass. Users can also sit on their sock accounts until the accounts reach a certain age. Bad-faith users can also obtain more karma from free karma subreddits (e.g., r/freekarma4u). Likewise, with *crowdsourced fact-checking*, a few moderators were concerned about users misrepresenting sources. For instance, bad-faith users can provide links to tangentially related articles from reputable sources that would not necessarily refute the reported content. For *similar posts*, some moderators emphasized that it takes very little to create a subreddit and become a moderator. Incompetent or even malicious moderators may accidentally or purposefully remove factual content, which could potentially lead to a damaging cascade.

> "Because then suddenly, someone could be posting something and with a certain amount of concerted effort, someone could trick a moderator into banning something that they shouldn't have and then they would cascade through. It would be helpful if it was a true positive in terms of misinformation that needs to be banned. But it makes the effect of a false positive heavy." -[P6]

Over-reliance on the Widget can Undermine Moderation: A couple of participants argued that moderators may also become too reliant on user attributes (shown in *user popups*) rather than analyzing each user from their own perspectives. These interviewees expressed concern that good-faith users may be treated unfairly.

> "There are subreddits I could go to and make what I believe are perfectly innocent, truthful, non-controversial comments, but it's against their narrative and I'll be swarmed with downvotes ... Again, it's useful to have quick information [user popup], but just people have to know that it doesn't mean that they're trolling." -[P15]

Similarly, many participants also indicated that they preferred making their own moderation decision rather than relying on the actions of other subreddits' moderators (shown in *similar posts*) because each subreddit has its own goals and rules.

4.4.3.4 Improving Modqueue Designs: Credibility-focus Approaches

Interestingly, our analysis shows that most of the recommendations made by participants to help improve the modqueue designs are centered on intent and credibility. First, participants proposed several additional metrics to be included in the user popup. These include large gaps in user activities, user account country of origin, and user activities in free karma subreddits. A few moderators also expressed interest in knowing whether a user was banned from other subreddits, and having access to notes of other subreddits' moderators. Finally, one moderator asked for more advanced NLP techniques that can detect abrupt changes in an account's writing style and topics of interest. Participants asserted that these metrics can help them separate sock puppet accounts and bad-faith users from organic users. Next, a couple of participants also suggested only including moderation actions from known reputable subreddits such r/coronavirus. Finally, several moderators indicated that the widget can include an additional search bar for academic sources such as PubMed or PMC search. These moderators believed that academic sources are more credible than third-party fact-checkers.

Additional Limitations and Concerns: The most common concern participants had with the widget is its technical limitations and AI trustworthiness. For instance, P3 questioned how well an algorithm can identify crowdsourced fact-checking. P16 mentioned that other types of user comments can also be indicative of misinformation.

> "You're gonna see a lot more people saying like F*** off to the person posting misinformation. They post other things, verbal cues that aren't quite as obvious." -[P16]

In addition to these technical limitations, moderators also expressed concern about the social aspects of the widget. Notably, one moderator asserted that they would prefer users to not engage directly with misinformation posters via commenting as it would likely lead to conflicts. Rather, users should flag misinformation using the report system. Similarly, a few moderators were concerned that Reddit admins preferred to "keep communities from spilling into one another and causing drama", raising concerns about the *similar posts* section of our widget. Finally, a couple of moderators also noted that ordinary users (and Reddit admins) may have user privacy concerns.

4.5 Discussion

4.5.1 Improving the Efficacy of Crowd Wisdom on Reddit

In this chapter, we found crowd wisdom—in the form of crowdsourced flagging and modteam support—to be a powerful tool for detecting COVID-19 misinformation. Most of the participants heavily relied on crowdsourced flagging to identify potential misinformation, and many also relied on inputs from other moderators for ambiguous cases. Furthermore, we also identified concrete strategies to improve the efficacy of crowd wisdom.

First, similar to prior work [90], we observed that the use of crowdsourced flagging is limited beyond identifying potential misinformation. This is likely because the current report system neither allows the reporters to provide evidence along with the reports, nor affords reporters to express their degree of concern. We argue that platforms should enable reporters to explain their reasoning for flagging. Platforms should also allow reporters to express their concern-level along with their reports. This can be accomplished by including additional inputboxes, scales, or checkboxes in the report system. An important caveat is worth mentioning: additional user input requirements may reduce the report rate [130]. Future work should also explore how this trade-off can be managed. Additionally, we observed that controversial and partisan communities indeed had significant issues (e.g., report abuse) with using crowd wisdom for misinformation detection, which is align with related work [57, 67]. More interestingly, we also discovered that similar issues exist in some apolitical, mainstream communities. Future work should explore community characteristics that are predictive of report abuse, and generate insights into how to best reduce such abuse.

Second, while Reddit has various subreddits (e.g., r/modsupport, r/modguide) specifically designed to support moderators, our interviews revealed that moderators are dependent on other platforms including Discord and Slack to seek and provide networked support. This includes check-ins with each other for specific, difficult cases. Further, our analysis also suggested that moderators of smaller subreddits have less access to these networks. Reddit and other social media giants should consider providing additional platforms (e.g., r/mod_discord_finder) such that isolated moderators may find other moderator teams that can assist them in various moderation aspects.

Third, the perceived intent and credibility of ordinary users and moderators alike are important factors that significantly moderate the perceived usefulness of their inputs. As such, incorporating metrics that are indicative of credibility in the moderation workflow can improve the efficacy of crowd wisdom for misinformation detection. Specifically, many participants called for the reporting system to include additional information on the reporter (e.g., account age, karma, whether or not the reporter is a regular member of the community). Moderators can use this information to better evaluate whether the reports are made in good faith. Similarly, moderators also only sought assistance from other moderators that they are affiliated with and trust. Likewise, the perceived use-

fulness of crowd-based signals in the widget is also dependent on whether or not these signals come from reputable users and moderators. For instance, some interviewees considered *crowdsourced flagging* by reputable users in their communities to have more weight in their decision-making. Lastly, platforms may be hesitant to make user information available in the report system or elsewhere due to privacy concerns. However, our analysis showed that platforms could provide various useful metrics without revealing the identity of these users. For example, mods mentioned wanting metrics like number of reports made per user, which lets mods to more easily decide whether they would take a report seriously without disclosing the reporter's username.

Finally, and interestingly, we also observed sophisticated anticipation from participants in that they had also identified ways that credibility-informing metrics can be gamed. Moderators suggested additional user account metrics that they believed to be harder to game. These metrics include sudden gaps in users' activities, or abrupt changes in users' linguistic patterns. Though, how to best implement these account metrics at scale is still an open question.

4.5.2 One Size Doesn't Fit All

Our findings in Section 4.4 encompassed the moderation practices of many of the participants. Here, we reiterate the variations in moderation, and highlight the complexity and the myriad of ways in which moderation happens.

Similar to related work [135, 223], we observed that moderation practices is dependent on moderator characteristics and subreddit attributes. Indeed, moderators' tech-savviness and trust in technology affect the extent to which they employed algorithmic tools to moderate misinformation. Similarly, moderators' expertise influenced how much they relied on their own knowledge to identify misinformation. Moderators' personal beliefs and philosophies also shaped the occurrence and severity of moderation actions they take after identifying misinformation. Likewise, community goals and rules affected whether and how misinformation is moderated. In particular, our analysis suggested that strict rules (e.g., not allowing posts containing nonreputable sources, not allowing off-topic content) and dedicated moderation enforcing the rules helped participants manage COVID-19 misinformation.

Overall, results suggest that the recommendations we made to improve misinformation moderation likely will have varied effects across different communities. For instance, moderators with low COVID knowledge may find labels from professional fact-checkers more useful than moderators with high COVID knowledge. Moderators who overlook repeated problematic user behaviors to preserve free speech will likely find less value in signals that are indicative of intent. Likewise, communities that do not moderate misinformation as that would contradict their goals will benefit minimally–if at all–from our proposals. Finally, moderators of communities with high levels

of report abuse, may benefit significantly more from transparent report systems such as the one proposed in this work.

For communities with moderation practices that significantly diverge from the norm, future work should explore ways to better assist them with misinformation moderation. Nevertheless, while one size doesn't fit all [123], it does fit many. The high-level findings and concrete strategies from our study can help many online communities with detecting misinformation.

4.5.3 Distrust in Fact-checking Services and the Limitations of Knowledge-based Misinformation Detection

One of the findings that we found surprising is the level of participant distrust in fact-checking services. That is, a quarter of all participants, explicitly expressed their distrust in third-party fact-checkers. Among them, there were no noticeable differences between those moderating political subreddits versus non-political ones. Some of them considered fact-checking organizations to be polarized and biased. Others were concerned about fact-checkers lacking domain expertise, and thus misinterpreting or mislabeling some subtle but important medical information. While the broader public's distrust in third-party fact-checkers was observed in past literature [184], our study revealed that such distrust also exists in a significant portion of online platform decision-makers. This result has important implications for misinformation detection on Reddit. Most notably, moderators who do not trust labels from professional fact-checkers may have to rely more on their own research and knowledge, the wisdom of their communities, or even the Reddit platform itself to identify misinformation during manual review (subprocess (e) in Figure 4.3). As such, we argue that future work focused on improving crowd wisdom to assist with manual review could be especially valuable to these moderators [53].

Alternatively, while there is likely no universal fact-checking site that all moderators trust, there seemed to be a hierarchy of sources that interviewees rank from the most reputable to the least reputable. Generally, participants tend to consider academic sources as the most authoritative. This is followed by official information from public health institutions. Though, moderators differ in whether they place more trust in federal institutions or local leadership. Given these considerations, Reddit could include facts and fact-checks from various sources, both academic and non-academic, and then let moderators choose which sources they rely on.

Beyond skepticism, our study demonstrated another limitation of professional fact-checking services. That is, while most participants considered labels from professional fact-checkers to be valuable for identifying content facticity, these labels cannot assist moderators with evaluating user intent. Interestingly, academic scholars have contended that intent is highly subjective and difficult to quantify [176]. Yet, platform stakeholders are nevertheless using various user characteristics

to interpret intent and use this inferred intent to choose moderation actions. Given these considerations, we argue that future work in misinformation detection should also explore methods to identify the intent of misinformation posters in addition to content veracity.

4.5.4 Generalizing Findings to Other Misinformation Domains

To what extent do our findings and strategies generalize to other types of misinformation? As mentioned in Section 4.1.1, for a majority of the participants, moderation of COVID-19 misinformation is similar to other types of problematic content. Additionally, the remaining moderators noted that their priority and difficulty varied across different domains (e.g., political vs. COVID-19 content) of misinformation. Though, there is no clear directionality. For instance, some moderators commented that they had more difficulty with COVID-19 misinformation, whereas others said the opposite. Some participants similarly mentioned that they were more strict with COVID-19 misinformation, whereas others said that they were more lenient. Our analysis didn't reveal any clear indicators that can explain this variance. Overall, our results suggest that while the difficulty and priority varied across different types of misinformation, the workflow model in Figure 4.3 is generalizable to other misinformation domains for many subreddits.

More interestingly, we also observed that many participants had gone through unique knowledge acquisition steps in order to moderate COVID-19 misinformation. Particularly, as the pandemic progressed and more information became available, moderators needed to continuously update their knowledge on what's misinformation and what's not. A few others, however, noted that things had settled down more recently. Finally, a couple of moderators also commented that they had more difficulty digesting medical and scientific information than political content. Our analysis suggested that moderators needed a certain level of domain-specific literacy and knowledge in order to moderate misinformation of different domains.

4.6 Limitations and Future Work

We note the following limitations in our study. First, while we covered a range of participants and subreddits, we had limited participation from conservative-leaning subreddits. Despite our best efforts, we were not able to obtain interview opportunities with these subreddits. It's possible that moderators of conservative-leaning subreddits have distinct experiences with crowd wisdom (e.g., they may have encountered significantly more misreports due to inter-community conflicts [57]). Second, our analysis is focused exclusively on COVID-19 misinformation. While we are confident that many of the high-level findings are generalizable to other types of misinformation, additional future work is needed to explore the extent to which moderation practices vary across different

information domains. For instance, we observed some evidence that moderators encountered more misreports on political content. This suggests that crowdsourced flagging is less reliable for detecting political misinformation. Similarly, our study was conducted during a time when there was no significant development (i.e., external shocks) in COVID-19. A few of our participants had noted that their reliance on crowd wisdom had changed over time, and their moderation practices also varied amid external events. For instance, a few moderators considered users to be less knowledgeable and reliable early on than later in the pandemic. Given these considerations, future work should also explore the temporal variations in the reliability of crowd wisdom, and its implications for misinformation detection in self-governing online communities. Next, we also highlight that our results are based on semi-structured interviews. Using quantitative analysis on larger scale datasets would be interesting and important to complement our findings. For instance, if data is available, future work should quantitatively examine the frequency of report abuse across different communities, and determine which community-level attributes are indicative of report abuse. Finally, we note that this work is largely focused on improving misinformation detection from the perspective of online community moderators, and we didn't explore which moderation strategies/actions are the most effective at safeguarding communities against misinformation. We highlight that misinformation detection is only the first step in the broader objective of combating misinformation. Much more work is needed to examine its characteristics, and find the best ways to curtail its prevalence and influence.

CHAPTER 5

Conclusion

5.1 Results and Discussion

In this dissertation, we studied three popular misinformation detection (MID) approaches. We empirically explored the theorized strength and caveats of each approach with respect to their actual and potential use cases. In this section, we first review the main findings, discuss their implications and focus on the real-world applications and concerns of these three popular MID approaches. Finally, we provide recommendations for future developers and users of the three MID approaches.

5.1.1 Summary of Key Findings

In study 1, we evaluated i) to what extent expert labels differ and what are some possible explanations, and ii) whether groundtruth choice significantly changes observations in downstream analyses. We observed that existing lists of fake news publishers provided by different experts vary significantly in size and have very few domains in common. Additionally, some lists remained static while others changed over time (e.g., adding/removing news publishers). We also found that popular fake news publishers tend to be included in the lists earlier than unpopular ones. Furthermore, while certain lists included detailed annotation procedures, others either provided vague descriptions or omitted to provide their labeling process altogether. More importantly, we also demonstrated that the estimated prevalence of fake news, and to some extent the temporal trend, is significantly dependent on which fake news list is used as the groundtruth. However, news sites are similar in their agenda-setting [168] behavior. For example, regardless of groundtruth choice, fake news sites more frequently published articles about Hilary Clinton's email scandal than traditional news media. This suggested that, despite the lists having very few news sites in common, the websites themselves are similar in behavior. As such, the low overlap of domains is likely due to expert fact-checkers having a limited capacity and only manually examining a subset of the tens of thousands of news sites available online.

For study 2, we examined a representative subset of automated MID models. We explored i) whether the performance of these models is consistent across different contexts, and ii) biases present in these models. We showed model performance varied substantially based on the choice of dataset and evaluation metrics. Additionally, models generally performed worse on articles from never-before-encountered news publishers. This raises concern regarding these models' real-world applicability, given that new fake news sites are routinely created [115]. Focusing on bias analysis, we also observed that all classifiers demonstrated significantly higher false-positive rates for right-leaning mainstream news sites. This result could be due to conservative mainstream news publishers more frequently referencing articles from fake news sites [19]. Nevertheless, this bias raises an important consideration for MID systems aimed at detecting ideologically-motivated (mis)information. Further, the performance of classifiers can decrease following external shocks such as political scandals. Similarly, models were worse at classifying news articles about scandals involving the two presidential nominees than those about other topics such as polling results. In sum, our simple evaluation steps revealed potential biases and performance inconsistencies in existing MID classifiers. Finally, these results echoed weaknesses seen in many machine learning models in other research domains.

Finally, in study 3, we explored i) how moderators of online communities regulate misinformation, ii) the role of crowd wisdom in the moderation process, and iii) how to better leverage crowd wisdom to improve this process. We relied on semi-structured interviews with Reddit moderators for this study. Our research revealed a general workflow adopted by most interviewees to moderate COVID-19 misinformation. Interestingly, we also identified patterns of moderation practice centered on content facticity, perceived harm, and user intent. This is very similar to the academic definitions of misinformation [275, 246]. Next, we also showed that Reddit moderators rely on two types of crowd wisdom for misinformation detection: the wisdom of ordinary users, and the wisdom of other moderators. We observed that almost all Reddit moderators are heavily reliant on crowdsourced flagging (i.e., users use the platform-wide reporting system to flag content as misinformation) to come upon potential COVID-19 misinformation. However, the report system is often abused; though, less so for COVID-19 than political content. Further, crowdsourced flagging is currently limited in its ability to help moderators evaluate content facticity, harm, and user intent. Many moderators suggested that the usefulness of the report system can be improved by including additional evidence that substantiates the credibility of the reporter, and the evidence that directly rebukes claims made in the reported posts. Next, when encountering difficult, ambiguous cases, moderators often seek inputs from their moderation team or from expert moderators of other communities to make the final judgment. Though, many moderators are reluctant to rely on the inputs from moderators of questionable subreddits. Finally, we showed interviewees additional types of crowd signals readily available on Reddit. We observed that close to half of the interviewees

preferred these crowd signals over labels provided by professional fact-checkers.

5.1.2 Implications on Real World Applications

What are the important implications of our findings with respect to these misinformation detection approaches' potential use cases and real-world applications?

First, results from studies 1 and 3 suggested that despite a lack of clear consensus on what's misinformation, experts and platform stakeholders both considered content facticity and publisher/user intent as key characteristics of misinformation. The degree of content veracity and the extent of harmful intentions both affected how experts label fake news sites, and also how platform stakeholders prioritize misinformation. For instance, experts often gave fake news sites that were purposefully created to mislead and harm sublabels (e.g., *impostor sites*) to separate them from ones that intended to serve useful purposes like *satire*. Similarly, platform stakeholders were also particularly vigilant against bad-faith users and coordinated efforts to spread misinformation. In practice, however, publisher and user intent are often challenging to determine [176, 89]. For instance, existing knowledge-based approaches can not separate intentionally distributed misinformation from non-intentional false information. Moreover, benchmark misinformation datasets are limited, and even fewer datasets include intent labels [284]. Furthermore, past work [94, 116] also suggested that it's often difficult to separate parody from extreme views, and bad-faith fake news publishers and online trolls can always claim that their content is "satirical" when facing blowback or criticism. Additionally, intent is also a subjective perception of the entity (e.g., a professional fact-checker or an online community moderator) that's evaluating it [89]. Nevertheless, results from studies 1 and 3 suggested that in real-world applications, experts and online community decision-makers often make a considerable effort to determine intent (e.g., is a user genuinely trying to learn about the facts or is the user pushing a particular agenda?). And, platform decision-makers also take different moderation actions based on perceived user intent. For instance, many participants in study 3 preferred to educate a good-faith user but ban a bad-faith one. Overall, we argue that it's crucial for future work in misinformation detection to consider classifying user/publisher intent in addition to evaluating content veracity.

Next, studies 1 and 3 also revealed some noteworthy, shared heuristics among experts, and also among community moderators for evaluating intent. These heuristics gave us some intuitions about how to approach intent detection from the perspective of experts or community moderators. In study 1, experts' annotation procedures suggested that experts relied on domain name and domain aesthetics (e.g., HTML layouts) to identify impostor sites (e.g., abcnews.com.go, times.com.mx) that were purposefully masquerading themselves as well-known sources. Similarly, experts depended on the "About Us" page to evaluate the transparency and goals of a fake news site to

determine its intent (e.g., is a site transparent about being a parody site?). Experts also relied on the writing style, link sources, political affiliations, and social media accounts of these news sites [285, 254]. Future work can explore whether automated and crowd wisdom-based MID approaches can adopt these annotation procedures to identify publisher intent. Next, study 3 showed that platform stakeholders commonly relied on a user's community associations (i.e., which online communities is the user a member of?) and past behavior patterns (e.g., did the user repeatedly post misinformation?) to assess intent. Arguably, automated MID models can mirror the practices of platform stakeholders. For instance, future work can use social network analysis [68] (e.g., community detection) to classify a user's community affiliations. Likewise, future work can also train automated models using features generated from past user behaviors, and then use the models to predict future user behaviors. Lastly, we note that our studies only revealed some of the heuristics used by experts and platform stakeholders for intent detection. Future work should more systematically identify and evaluate these heuristics.

Second, studies 2 and 3 underscored the need for task and domain-specific (e.g., political vs. scientific misinformation) performance evaluation. For example, study 2 showed that a model might be great at detecting newly published articles from known fake news sites. Yet, it could perform poorly at identifying fake news articles from never-before-encountered fake news sites. Similarly, some moderators in study 3 noted that crowdsourced flagging is less reliable and more likely to be abused for political content (e.g., Jan 6 riot) than COVID information. One possible explanation is that crowds are more likely to misreport in order to achieve partisan ends, particularly in cases where ideology is salient [60]. As such, online communities that rely on their users for misinformation detection can potentially have more trouble finding political misinformation than misinformation in other less polarized domains.

Similarly, studies 2 and 3 also revealed that the performance for automated methods and crowdsourced flagging change over time and during external shocks. For example, study 2 suggested models that heavily relied on user behavior-based features have lower performance in detecting misinformation published amid external shocks due to the sudden changes in user behavior during these unexpected events. This finding is aligned with prior work [111], which showed that the performance for models that relied on NLP features degrades over time. Likewise, some interviewees in study 3 noted that they were less able to rely on users in their community for fact-checking early into the pandemic. These moderators emphasized that there was less official information on COVID-19 and more speculations from users in their community. However, the same participants noted that the pandemic has been ongoing for over two years, and they can now rely more on their regular users to do the fact-checking because these users have acquired factual COVID-19 knowledge over time. This is aligned with prior literature, which demonstrated that crowds can gain experience and become better at fact-checking over time [213]. Overall, our findings sug-

gested that performance evaluation procedures centered on temporal variations of misinformation can further inform researchers of the robustness of different existing approaches.

Third, studies 1 and 2 highlighted the need to include domain-specific biases detection steps. Study 2 clearly showed that ideological bias in automated misinformation detection methods is a potential concern. Many automated models were significantly more likely to mislabel news articles from mainstream conservative sites as fake. Moreover, study 2 also demonstrated that even within a single domain, model performance varied across different topics. Within the broader context, scholars have been increasingly invested in understanding meaningful biases in algorithmic methods [189, 55, 267]. These related studies suggested that biases could potentially be presented in i) the data collection and processing stage, ii) the modeling stage, iii) the evaluation stage, and iv) the application stage. Study 2, in particular, revealed that the model training and testing technique used by many misinformation classifiers were not appropriate for many real-world applications. Further, the evaluation procedures used by many studies were minimal. Next, in study 1, we didn't observe a significant difference in ideological biases among experts. In other words, the ideological leaning of a fake news site was not correlated with it being more likely or less likely to be included by one expert than other experts. Though, it's possible that experts were similarly biased. Future work should explore experts' content selection process (i.e., choose which news sites or which claims to fact-check). Regardless, ideology remains a crucial concern for critics of third-party fact-checking organizations. And, appropriate bias detection steps can be used to address and alleviate this concern. Finally, while study 3 didn't include any bias analysis, prior studies demonstrated that systematic bias (e.g., ideological bias) within a crowd could undermine its accuracy [235]. Consequently, we argue that bias assessment procedures are also necessary for researchers interested in leveraging crowd wisdom for misinformation detection.

Fourth, transparency, or rather the lack of it, is an issue in all three approaches. Study 1 showed that some experts such as PolitiFact didn't provide any annotation procedures describing how they identify fake news sites, or had kept their procedures vague. This is surprising because when providing claim-level labels, third-party fact-checking organizations like PolitiFact made their fact-checking procedures transparent, and also consistently included detailed explanations. Given that expert labels are commonly used as groundtruth, and past work [33] demonstrated that transparency is associated with user trust in expert labels, we argue that experts should make transparent additional useful metadata along with their labels and explanations. Similarly, in study 3, many participants emphasized that they want more transparency in the user report system, such as being able to see who the reporter is. In fact, some moderators attributed the lack of transparency in the report system as the key challenge in misinformation detection. While platforms may prefer to keep the report system opaque due to concerns about user privacy, our study identified several strategies to expand transparency in the reporting system while preserving reporter privacy. Fi-

nally, transparency wasn't the focus of study 2. However, the broader machine learning field is moving toward more explainable and transparent models [189, 171, 137].

An important concern regarding expanding transparency in existing MID approaches is that bad-faith users can exploit it to game the system [61, 129]. Indeed, professional fact-checkers' efforts to make their labeling procedures more nuanced and transparent can be used to discredit or undermine them [48]. Similarly, while increasing transparency in automated models has many positive effects, such as preventing overtrust in fake news classifiers [175], this would also arguably makes it easier for bad actors to game the AI (e.g., adversarial attacks) [58]. Studies 1 and 2 didn't explore the trade-offs between transparency and security in expert labeling and automated models. However, study 3 demonstrated that experienced moderators can predict the actions of bad-faith, norm-violating users, and formulate strategies to make the crowd wisdom-based MID system more difficult to game. Further, some interviewees also acknowledged that complete transparency is not optimal for moderation, and that they preferred more opaque moderation practices (e.g., shadow bans and removals) under certain circumstances. This is aligned with past work [129], which showed that moderators tend to keep some moderation processes opaque, especially if they consider these processes easy to game. Our work suggested that individuals who are experienced with using a particular MID approach in real-world applications can provide valuable insights into the transparency and security trade-offs of the approach. Future work should consider leveraging such insights when introducing additional transparency to existing systems.

Finally, an important finding from study 3 also highlighted that the preferred MID method is not always straightforward and fully rational. For instance, scholars commonly view expert labeling as the most accurate among the three popular MID approaches. However, in practice, platform stakeholders may prefer other less accurate methods due to the lack of perceived expert credibility. Specifically, moderators in study 3 considered that many fact-checking organizations are ideologically biased, or that these fact-checking organizations lacked domain/subject expertise. Similarly, some moderators also voiced their distrust of automated approaches. A couple of participants even stated that they lack confidence in the crowdsourced reporting system despite their belief that users in their communities generally participate in good-faith. We argue that the lack of transparency and potential biases in existing approaches likely hindered their adoption in the real world. Another possible explanation is that the increased polarization had led to a reduced perception of credibility in fact-checking services in populations from both the right and the left [33]. Likewise, individuals can also overtrust existing approaches, and view these methods as more accurate than they actually are [175]. Here, we argue that more robust evaluation and bias assessment procedures, and additional transparency can assist in correcting these misperceptions.

5.1.3 Practical Recommendations

Thus far, we have reviewed the key findings from each study. We have also synthesized characteristics shared by some or all of the three MID approaches. In this section, we provide some simple but practical high-level recommendations for both developers and users of these MID approaches. Our recommendations are guided by existing use cases and concerns.

Expert Labeling

Expert Fact-checkers: We provide suggestions for the entity and content selection process, the labeling phase, and the label maintenance stage. First, we observed that experts varied significantly in which publishers they chose to examine and label. For example, PolitiFact noted that they worked with Facebook to generate the list of popular fake news sites active on Facebook during the 2016 U.S. presidential election. In comparison, Media Bias Fact Check relies on unaffiliated individuals to first submit questionable websites to their queue, and then its professional staff will review each pending site. Our observation reflects related work [163, 98], which demonstrated that experts lack a systematic content selection process. Here, we recommend that experts make their selection pools transparent and provide detailed descriptions of how their pools are generated, and what criteria are used to choose which sites or claims to verify. For experts who do not have a systematic way of generating the selection pool, we propose to use automated models to help identify potential fake news sites and potential misinformation content, provided that model biases are examined, understood, and documented. We recommend automated models, given that many models need few resources as opposed to other more costly alternatives (e.g., crowdsourced workers). Additionally, automated models are highly scalable, easily tunable, and stay static unless retrained (e.g., a model will always return the same label for a given set of features).

Next, focusing on the labeling phase, we propose that experts include the ideological slants of their labeled fake news sites and false claims. This is because the propagation and consumption of misinformation across many domains (e.g., elections, COVID-19) are correlated with political ideology [95, 10]. Additionally, many downstream studies relying on expert labels were also focused on ideology [131]. Further, we also encourage experts to include labels that are indicative of intent in addition to content facticity labels. Arguably, it's often difficult to establish intent for popular claims circulated online by various actors. However, when providing publisher-level labels, experts can adopt various frameworks proposed by related work [264, 275] to separate the least harmful publishers (e.g., clickbait sites that have sensational headlines but factual articles) from the most harmful publishers (e.g., impostor sites). Similarly, focusing on the label maintenance stage, we note that labels may change over time with the introduction of new information. Experts can also change their fact-checking practices over time. We also encourage experts to keep track

of all these changes. These metadata can help both experts and users of these expert labels identify potential biases, discrepancies, or errors early on. For instance, as shown in study 2, timestamps can help researchers determine whether experts are more likely to include news sites of particular characteristics earlier than others (i.e., expert bias).

Users of Expert Labels: We propose the following for scholars interested in using expert labels for research. First, similar to past work [16], we showed that the prevalence of misinformation is heavily dependent on what's considered misinformation (e.g., do we include or exclude fake news sites of mixed factualness). Here, we suggest scholars use expert-labeled datasets that have granular labels for both content facticity and intent. Additionally, we also recommend researchers conduct multiple, separate prevalence analyses ranging from the most broad definition of misinformation to the most restrictive. This can provide informative bounds on prevalence analyses. For researchers interested in exploring the behavior of fake news publishers, it matters less which expert labels are used. Next, our study also demonstrated that some expert-labeled datasets are static while others change over time. For instance, PolitiFact never updated its list of fake news sites since early 2017. However, Media Bias Fact Check is continuously updating its list to include new fake news publishers. As such, scholars interested in studying the temporal variation of recent misinformation are likely better off with the "live" datasets. Further, when everything else is comparable, we recommend researchers consider more transparent and better-documented expert labels. These labels are not necessarily more accurate than labels provided by experts with opaque annotation processes. But, they can help users more quickly identify potential issues and biases. Finally, as a general good practice, we recommend that researchers use more than one expert dataset for robustness checks. Researchers can use various clustering techniques to identify expert labels that are significantly distinct from each other, and then use the most distinct expert labels to evaluate whether downstream results are consistent.

Automated Methods

We make the following recommendations for machine learning researchers focusing on task-specific evaluation procedures and bias assessment steps. We note that these recommendations are not novel. They are also simple to adopt. Yet, many related studies surprisingly did not incorporate these practices.

First, we propose that, instead of using generic n-fold cross-validation, researchers should separate the training and validation data using time frames (i.e., misinformation posted before time t is used for training, and misinformation posted after time t is used for evaluation). Researchers can also use the leave-p-out validation technique such that a model is trained using misinformation published by a subset of publishers (or users), and then validated against misinformation published

by another set of publishers (or users). Both of these training/validation methods are more aligned with real-world applications of MID models than n-fold cross-validation. Particularly, the former is better for evaluating a model's performance at detecting recently published misinformation, and the latter for misinformation shared by never-before-encountered publishers. Second, we also encourage researchers to use a wider range of evaluation metrics, including f1, roc auc, precision and recall. We argue that a model does not need to have high scores across all metrics. Rather, these metrics can be used to determine which types of tasks a model is best fitted for. Further, researchers should also train and evaluate a model using different benchmarking datasets.

Finally, we focus on bias assessment steps. Past work has clearly demonstrated that many machine learning models used in real-world applications exhibit harmful biases [190]. Our results from study 2 also identified potential ideological biases in many proposed MID models. While none of the models we examined in study 2 are used in real-world applications, we again encourage domain-experts to identify potential high-cost biases and evaluate MID models for these biases. For high-performing but biased models, researchers can consider various trade-off techniques suggested by literature in the Fairness, Accountability, Transparency, and Ethics in AI (FATE) field [137, 267, 55].

Crowd Wisdom

Finally, for platforms invested in leveraging crowd wisdom to assist online community moderators in regulating misinformation, we make the following recommendations.

First, our study demonstrated that crowdsourced flagging is very useful to moderators for detecting potential misinformation. However, a flag is a "thin expression" [51]. We recommend that platforms update their report UI so that reporters can provide evidence proving the reported content is indeed misinformation or that the reported user is indeed intentionally spreading misinformation. Next, community moderators often experience users misusing or abusing the report system. Thus far, there is little room for moderators to address the abusers due to a lack of transparency. Platforms should consider providing greater transparency for their report systems while maintaining reporter privacy and safeguarding against bad-faith users from abusing the expanded transparency. Some of our suggestions include making visible i) reporter account age, ii) whether the reporter has been a regular member of the community, and iii) the number of reports made by the reporter recently. Next, our study also demonstrated that moderators value crowdsourced fact-checking posts from good-faith users in their communities. As such, we recommend platforms make these crowd signals available to moderators.

Further, our study also revealed that moderators rely on the wisdom of other moderators for misinformation detection. For example, our work showed that moderators of sports subreddits (e.g., r/basketball) sought help from moderators of reputable COVID-19 subreddits with regulat-

ing COVID-19 misinformation. Therefore, we also recommend platforms incorporate additional affordances to facilitate the cooperation between moderators. For instance, platforms can identify communities that may have trouble moderating misinformation and communities with a lot of experience and are successful at moderating misinformation, and then facilitate the interactions between the moderators of these different communities.

5.2 Limitations and Future Work

There are several noteworthy limitations in this dissertation. First, each of our studies was focused on misinformation related to one of the two historical events: the 2016 U.S. presidential election or the COVID-19 pandemic. As such, future work should consider evaluating the characteristics of each MID approach across misinformation of different events and domains.

Next, our studies explored the potential use and significant caveats of each misinformation approach separately. That is, we did not compare and contrast the strength and weaknesses of these MID approaches to determine their best use cases. Related work and our own findings suggested that expert labeling is the preferred approach for small scope detection and high-stake scenarios. Some potential high-stake scenarios include i) detecting misinformation amid significant external shocks, ii) fact-checking public speeches and debates involving elites, and iii) when platforms are using the labels to make long-lasting decisions (e.g., permanently blocking a fake news publisher from using its platform). Further, the automated MID approach is the most scalable and responsive (early detection) of the three. Though, our studies showed that many existing MID models lacked proper evaluation and bias assessment. Perhaps then, the automated MID approach can be used for identifying potential misinformation at its early state in low-stake scenarios. Finally, the crowd wisdom-based MID approach can potentially be as scalable as automated methods, and as accurate as expert labeling. However, crowd wisdom is less responsive than automated methods. Many crowd signals are only available once misinformation is already viewable to many online users. Future work should empirically compare and contrast these three approaches to identify their best use cases.

Similarly, many recent studies also proposed more complex misinformation detection pipelines that incorporated a combination of the three basic approaches [101]. For instance, [258] proposed a hybrid human-machine collaboration model that first uses an automated method to detect suspicious URLs; these URLs are then recommended to be fact-checked by crowds. Likewise, [188] uses outputs from automated models to assist crowds in misinformation detection. Future work should review instances of these more complex pipelines to empirically determine their strength, weaknesses, and potential use.

Finally, we highlight that misinformation detection is an important first step in the broader goal

of curtailing the spread and influence of misinformation. Findings and recommendations distilled from this dissertation can be used to guide future research in misinformation detection.

BIBLIOGRAPHY

[1] New Rules for Twitter & Facebook.

[2] Alison Abbott. Covid's mental-health toll: how scientists are tracking a surge in depression. *Nature*, 590(7845):194–195, 2021.

[3] Vlad Achimescu and Pavel Dimitrov Chachev. Raising the Flag: Monitoring User Perceived Disinformation on Reddit. *Information*, 12(1):4, December 2020.

[4] Samantha A Adams. Sourcing the crowd for health services improvement: The reflexive patient and "share-your-experience" websites. *Social science & medicine*, 72(7):1069–1076, 2011.

[5] Reddit Admin. Misinformation and covid-19: What reddit is doing, 2020.

[6] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In Issa Traore, Isaac Woungang, and Ahmed Awad, editors, *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, Lecture Notes in Computer Science, pages 127–138. Springer International Publishing, 2017.

[7] Hadeel S Alenezi and Maha H Faisal. Utilizing crowdsourcing and machine learning in education: Literature review. *Education and Information Technologies*, pages 1–16, 2020.

[8] Sana Ali. Combatting Against Covid-19 & Misinformation: A Systematic Review. *Hu Arenas*, October 2020.

[9] Katitza Rodriguez Alimonti and Veridiana. "fake news" offers latin american consolidated powers an opportunity to censor opponents, Apr 2018.

[10] Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017.

[11] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *arXiv preprint arXiv:1809.05901*, 2018.

[12] Jennifer Nancy Lee Allen, Antonio Alonso Arechar, Gordon Pennycook, and David Gertler Rand. Scaling Up Fact-Checking Using the Wisdom of Crowds. preprint, PsyArXiv, October 2020.

[13] Monica Anderson and Dennis Quinn. 46% of us social media users say they are 'worn out'by political posts and discussions. 2019.

[14] Philip Ball and Amy Maxmen. The epic battle against coronavirus misinformation and conspiracy theories. *Nature*, 581(7809):371–375, 2020.

[15] Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research*, 41(3):430–454, 2014.

[16] João Pedro Baptista and Anabela Gradim. A working definition of fake news. *Encyclopedia*, 2(1):632–645, 2022.

[17] Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 215–236. Springer, 2020.

[18] Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 2: Factuality. *CLEF (Working Notes)*, 2125, 2018.

[19] Yochai Benkler, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.

[20] Stephen Beveridge and Charles R Nelson. A new approach to decomposition of economic time series into permanent and transitory components. *Journal of Monetary economics*, 7(2):151–174, 1981.

[21] Meghana Moorthy Bhat and Srinivasan Parthasarathy. How effectively can machines defend against machine-generated fake news? an empirical study. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 48–53, 2020.

[22] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2):1–26, October 2020. arXiv: 2008.09533.

[23] Leticia Bode, Ceren Budak, Jonathan M. Ladd, Frank Newport, Josh Pasek, Lisa O. Singh, Stuart N. Soroka, and Michael W. Traugott. *Words That Matter: How the News and Social Media Shaped the 2016 Presidential Campaign*. Brookings Institution Press, Washington, D.C., 2020.

[24] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, March 2018.

[25] John H Boman and Owen Gallupe. Has covid-19 changed crime? crime rates in the united states during the pandemic. *American journal of criminal justice*, 45(4):537–545, 2020.

[26] Shelley Boulianne. Social media use and participation: A meta-analysis of current research. *Information, communication & society*, 18(5):524–538, 2015.

[27] Peter Bourgonje, Julián Moreno Schneider, and Georg Rehm. From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In *NLPmJ@EMNLP*, 2017.

[28] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7, 2019.

[29] Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 451–466. Springer, 2013.

[30] Lia Bozarth and Ceren Budak. Toward a better performance evaluation framework for fake news classification. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 60–71, 2020.

[31] Lia Bozarth and Ceren Budak. Market forces: Quantifying the role of top credible ad servers in the fake news ecosystem. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 83–94, 2021.

[32] Lia Bozarth, Aparajita Saraf, and Ceren Budak. Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 us presidential nominees. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 48–59, 2020.

[33] Petter Bae Brandtzaeg and Asbjørn Følstad. Trust and distrust in online fact-checking services. *Communications of the ACM*, 60(9):65–71, 2017.

[34] Joshua A. Braun and Jessica L. Eklund. Fake news, real money: Ad tech platforms, profit-driven hoaxes, and the business of journalism. *Digital Journalism*, 7(1):1–21, January 2019.

[35] Ceren Budak. What happened? the spread of fake news publisher content during the 2016 u.s. presidential election. WWW '19, 2019.

[36] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 665–674, New York, NY, USA, 2011. ACM.

[37] Ceren Budak, Ashley Muddiman, Yujin Kim, Caroline C Murray, and Natalie J Stroud. Covid-19 coverage by cable and broadcast networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 952–960, 2021.

[38] Dustin P Calvillo, Bryan J Ross, Ryan JB Garcia, Thomas J Smelter, and Abraham M Rutchick. Political ideology predicts perceptions of the threat of covid-19 (and susceptibility to fake news about it). *Social Psychological and Personality Science*, 11(8):1119–1128, 2020.

[39] Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. A Topic-Agnostic Approach for Identifying Fake News Pages. *Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19*, pages 975–980, 2019. arXiv: 1905.00957.

[40] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):1–30, November 2019.

[41] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, November 2018.

[42] Wojciech W Charemza and Ewa M Syczewska. Joint application of the dickey-fuller and kpss tests. *Economics Letters*, 61(1):17–21, 1998.

[43] Marina Charquero-Ballester, Jessica G Walter, Ida A Nissen, and Anja Bechmann. Different types of covid-19 misinformation have different emotional valence on twitter. *Big Data & Society*, 8(2):20539517211041279, 2021.

[44] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020.

[45] Nick Clegg. Combating covid-19 misinformation across our apps, Aug 2021.

[46] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. A Motif-based Approach for Identifying Controversy. *arXiv:1703.05053 [cs]*, March 2017. arXiv: 1703.05053.

[47] Corbett, P. Why the times calls trump 'mr.' (no, we're not being rude), 2017.

[48] David Corn. How to beat the fact checkers. politicians have figured it out: When caught in a lie, attack the truth cops. *Mother Jones*, 2012.

[49] Michele Coscia and Luca Rossi. Distortions of political bias in crowdsourced misinformation flagging. *J. R. Soc. Interface.*, 17(167):20200020, June 2020.

[50] Diane Coyle. Making the most of platforms: a policy research agenda. *Available at SSRN 2857188*, 2016.

[51] Kate Crawford, Tarleton Gillespie, and Nms /new Media. What is a flag for?, 2014.

[52] Michelle Croft and Raeal Moore. Checking what students know about checking the news. 2017.

[53] Ryan Cross. Will public trust in science survive the pandemic. *Chemical and Engineering News*, 99(3), 2021.

[54] Milmo Dan. Reddit bans covid misinformation forum after 'go dark' protest, Sep 2021.

[55] David Danks and Alex John London. Algorithmic bias in autonomous systems. In *IJCAI*, volume 17, pages 4691–4697, 2017.

[56] Kareem Darwish, Walid Magdy, and Tahar Zanouda. Trump vs. hillary: What went viral during the 2016 us presidential election. In *International Conference on Social Informatics*, 2017.

[57] Srayan Datta and Eytan Adar. Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, volume 13, pages 146–157, 2019.

[58] Karl de Fine Licht and Jenny de Fine Licht. Artificial intelligence, transparency, and public decision-making. *AI & society*, 35(4):917–926, 2020.

[59] Nikos Deligiannis, Tien Do Huu, Duc Minh Nguyen, and Xiao Luo. Deep Learning for Geolocating Social Media Users and Detecting Fake News. 2018.

[60] AR Dennis, A Kim, and P Moravec. Facebook's bad idea: Crowdsourced ratings work for toasters, but not news. *Buzzfeed*, 2018.

[61] Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.

[62] Nicholas Diakopoulos and Mor Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 133–142, 2011.

[63] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[64] Tien Huu Do, Xiao Luo, Duc Minh Nguyen, and Nikos Deligiannis. Rumour Detection via News Propagation Dynamics and User Representation Learning. *arXiv:1905.03042 [cs, stat]*, April 2019. arXiv: 1905.03042.

[65] Anil R Doshi, Sharat Raghavan, Rebecca Weiss, and Eric Petitt. The impact of the supply of fake news on consumer behavior during the 2016 us election. 2018.

[66] Bryan Dosono and Bryan Semaan. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk, May 2019. ACM.

[67] Bryan Dosono and Bryan Semaan. Decolonizing tactics as collective resilience: Identity work of aapi communities on reddit. *Proceedings of the ACM on Human-Computer interaction*, 4(CSCW1):1–20, 2020.

[68] David Easley, Jon Kleinberg, et al. Networks, crowds, and markets. *Cambridge Books*, 2012.

[69] Laura Edelson, Minh-Kha Nguyen, Ian Goldstein, Oana Goga, Damon McCoy, and Tobias Lauinger. Understanding engagement with us (mis) information news sources on facebook. In *Proceedings of the 21st ACM Internet Measurement Conference*, pages 444–463, 2021.

[70] Matthias Eickhoff and Jan Muntermann. Stock analysts vs. the crowd: Mutual prediction and the drivers of crowd wisdom. *Information & Management*, 53(7):835–845, 2016.

[71] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. Detecting misleading information on covid-19. *Ieee Access*, 8:165201–165215, 2020.

[72] Ziv Epstein, Gordon Pennycook, and David Gertler Rand. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. preprint, PsyArXiv, April 2019.

[73] Facebook. How people help fight false news, Nov 2019.

[74] FactCheck.org. Our mission. 2018.

[75] Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. 2017.

[76] Lesley SJ Farmer. Don't get faked out by the news: Becoming an informed citizen. In *The Fifth European Conference on Information Literacy (ECIL)*, page 174, 2017.

[77] Gheorghe-Ilie Farte and Daniel-Rares Obada. Reactive public relations strategies for managing fake news in the online environment. 2018.

[78] Caitlin Candice Ferreira, Jeandri Robertson, and Marnell Kirsten. The truth (as i see it): philosophical considerations influencing a typology of fake news. *Journal of Product & Brand Management*, 2019.

[79] Casey Fiesler. Reddit Rules! Characterizing an Ecosystem of Governance. page 10, 2018.

[80] Álvaro Figueira and Luciana Oliveira. The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121:817–825, 2017.

[81] Dana R Fisher, Dawn M Dow, and Rashawn Ray. Intersectionality takes it to the streets: Mobilizing across diverse interests for the women's march. *Science Advances*, 3(9):eaao1390, 2017.

[82] D Flamini. Most republicans don't trust fact-checkers, and most americans don't trust the media., 2019.

[83] Richard Fletcher, Alessio Cornia, Lucas Graves, and Rasmus Kleis Nielsen. Measuring the reach of "fake news" and online disinformation in europe. *Reuters Institute Factsheet*, 2018.

[84] David Mandell Freeman. Can You Spot the Fakes?: On the Limitations of User Feedback in Online Social Networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1093–1102, Perth Australia, April 2017. International World Wide Web Conferences Steering Committee.

[85] Daniel Funke. Want to get away with posting fake news on facebook? just change your website domain., Jan 2019.

[86] Vijaya Gadde and Kayvon Beykpour. Additional steps we're taking ahead of the 2020 us election, Oct 2020.

[87] Axel Gelfert. Fake news: A definition. *Informal Logic*, 38(1):84–117, 2018.

[88] Jean D Gibbons and S Chakraborti. Comparisons of the mann-whitney, student'st, and alternate t tests for means of normal distributions. *The Journal of Experimental Education*, 59(3):258–267, 1991.

[89] Fabio Giglietto, Laura Iannelli, Augusto Valeriani, and Luca Rossi. 'fake news' is the invention of a liar: How false information circulates within the hybrid news system. *Current sociology*, 67(4):625–642, 2019.

[90] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[91] David W Goodall. A new similarity index based on probability. *Biometrics*, pages 882–907, 1966.

[92] Torstein Granskogen. Automatic detection of fake news in social media using contextual information. Master's thesis, NTNU, 2018.

[93] Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213, August 2019.

[94] James Grimmelmann. The platform is the message. *Geo. L. Tech. Rev.*, 2:217, 2017.

[95] N Grinberg, K Joseph, L Friedland, B Swire-Thompson, and D Lazer. Fake news on twitter during the 2016 us presidential election. Technical report, 2018.

[96] Neil Gross and Solon Simmons. Americans' views of political bias in the academy and academic freedom. In *annual meeting of the American Association of University Professors*, 2006.

[97] Gisel Bastidas Guacho, Sara Abdali, Neil Shah, and Evangelos E. Papalexakis. Semi-supervised Content-based Detection of Misinformation via Tensor Embeddings. *arXiv:1804.09088 [cs, stat]*, April 2018. arXiv: 1804.09088.

[98] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586, 2019.

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

[100] Lei Guo and Chris Vargo. "Fake News" and Emerging Online Media Ecosystem. *Communication Research*, page 009365021877717, June 2018.

[101] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking.

[102] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 729–736, New York, NY, USA, 2013. ACM.

[103] Sebastián G Guzmán. "should i trust the bank or the social movement?" motivated reasoning and debtors' work to accept misinformation. In *Sociological Forum*, volume 30, pages 900–924. Wiley Online Library, 2015.

[104] Rex Hartson and Pardha S Pyla. *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier, 2012.

[105] Larry D Haugh. Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach. *Journal of ASA*, 71, 1976.

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

[107] Stefan Helmstetter and Heiko Paulheim. Weakly supervised learning for fake news detection on twitter. pages 274–277. IEEE, 2018.

[108] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[109] Benjamin D. Horne and Sibel Adali. This Just In. *arXiv:1703.09398 [cs]*, March 2017. arXiv: 1703.09398.

[110] Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. Assessing the News Landscape. Lyon, France, 2018. ACM Press.

[111] Benjamin D Horne, Jeppe Nørregaard, and Sibel Adali. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(1):1–23, 2019.

[112] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[113] Sohyeon Hwang and Jeremy D Foote. Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

[114] Waleed Iqbal, Gareth Tyson, and Ignacio Castro. Looking on Efficiency of Content Moderation Systems from the Lens of Reddit's Content Moderation Experience During COVID-19. *SSRN Electronic Journal*, 2022.

[115] Soares Isa. The fake news machine: Inside a town gearing up for 2020, 2017.

[116] Caroline Jack. Lexicon of lies: Terms for problematic information. *Data & Society*, 3(22):1094–1096, 2017.

[117] Gary C Jacobson. Donald trump and the parties: Impeachment, pandemic, protest, and electoral politics in 2020. *Presidential Studies Quarterly*, 50(4):762–795, 2020.

[118] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, 2012.

[119] Amelia Jamison, David A Broniatowski, Michael C Smith, Kajal S Parikh, Adeena Malik, Mark Dredze, and Sandra C Quinn. Adapting and extending a typology to identify vaccine misinformation on twitter. *American Journal of Public Health*, 110(S3):S331–S339, 2020.

[120] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5):1–35, September 2019.

[121] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. 26(5):1–35, 2019.

[122] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):150:1–150:27, November 2019.

[123] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. A trade-off-centered framework of content moderation. *arXiv preprint arXiv:2206.03450*, 2022.

[124] Julie Jiang, Emily Chen, Shen Yan, Kristina Lerman, and Emilio Ferrara. Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies*, 2(3):200–211, 2020.

[125] Shan Jiang, Miriam Metzger, Andrew Flanagin, and Christo Wilson. Modeling and measuring expressed (dis) belief in (mis) information. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 315–326, 2020.

[126] Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36, 2022.

[127] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. News Credibility Evaluation on Microblog with a Hierarchical Propagation Model. In *2014 IEEE International Conference on Data Mining*, pages 230–239, December 2014. ISSN: 2374-8486.

[128] Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. r/science: Challenges and Opportunities in Online Science Communication. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Glasgow Scotland Uk, May 2019. ACM.

[129] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–35, January 2020.

[130] Anja Kalch and Teresa K Naab. Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. 2017.

[131] Mansooreh Karami, Tahora H Nazer, and Huan Liu. Profiling fake news spreaders on social media through psychological and motivational factors. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 225–230, 2021.

[132] Dimitrios Katsaros, George Stavropoulos, and Dimitrios Papakostas. Which machine learning paradigm for fake news detection? In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 383–387, October 2019. ISSN: null.

[133] Simranpreet Kaur, Pallavi Kaul, and Pooya Moradian Zadeh. Monitoring the dynamics of emotions during covid-19 using twitter data. *Procedia Computer Science*, 177:423–430, 2020.

[134] Junaed Younus Khan, Md Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032, 2021.

[135] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. Surviving an" eternal september" how an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1152–1156, 2016.

[136] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schoelkopf, and Manuel Gomez-Rodriguez. Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. *arXiv:1711.09918 [cs, stat]*, November 2017. arXiv: 1711.09918.

[137] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

[138] Camille Koenders, Johannes Filla, Nicolai Schneider, and Vinicius Woloszyn. How vulnerable are automatic fake news detection methods to adversarial attacks? *arXiv preprint arXiv:2107.07970*, 2021.

[139] Vasilis Koulolias, Gideon Mekonnen Jonathan, Miriam Fernandez, and Dimitris Sotirchos. Combating misinformation: An ecosystem in co-creation, 2018.

[140] Nir Kshetri and Jeffrey Voas. The economics of "fake news". *IT Professional*, 19(6):8–12, 2017.

[141] Adam Kucharski. Post-truth: Study epidemiology of fake news. *Nature*, 540(7634):525, 2016.

[142] Srijan Kumar and Neil Shah. False Information on Web and Social Media: A Survey. *arXiv:1804.08559 [cs]*, April 2018. arXiv: 1804.08559.

[143] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 591–602, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[144] Yenni Kwok. Where memes could kill: Indonesia's worsening problem of fake news. *Time, January*, 6, 2017.

[145] 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. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[146] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, pages 1–49. Citeseer, 2013.

[147] Lawrence Lessig. *Code: Version 2.0*. Shoeisha Co., Ltd., 2006.

[148] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.

[149] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[150] Lucas Lima, Julio Reis, Philipe Melo, Fabricio Murai, Leandro Araújo, Pantelis Vikatos, and Fabrício Benevenuto. Inside the right-leaning echo chambers. 2018.

[151] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1867–1870, 2015.

[152] Yang Liu and Yi-Fang Brook Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[153] Yang Liu and Yi-Fang Brook Wu. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–33, 2020.

[154] Steve Lohr. It's true: False news spreads faster and wider. and humans are to blame. *The New York Times*, 8, 2018.

[155] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. Fake News Detection Through Multi-Perspective Speaker Profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[156] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

[157] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[158] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. Beyond blacklists: learning to detect malicious web sites from suspicious urls. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1245–1254. ACM, 2009.

[159] Thomas J Main. *The Rise of the Alt-Right*. Brookings Institution Press, 2018.

[160] Adrija Majumdar and Indranil Bose. Detection of financial rumors using big data analytics: the case of the bombay stock exchange. *Journal of Organizational Computing and Electronic Commerce*, 28(2):79–97, 2018.

[161] Spyros Makridakis, Steven Wheelwright, and Rob Hyndman. *Forecasting methods and applications*. John wiley & sons, 2008.

[162] Alessandro R Marcon, Blake Murdoch, and Timothy Caulfield. Fake news portrayals of stem cells and stem cell research. *Regenerative medicine*, pages 765–775, 2017.

[163] Morgan Marietta, David C Barker, and Todd Bowser. Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? In *The Forum*, volume 13, pages 577–596. De Gruyter, 2015.

[164] Jason L Mast. Legitimacy troubles and the performance of power in the 2016 us presidential election. *American Journal of Cultural Sociology*, 5(3):460–480, 2017.

[165] J. Nathan Matias. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1138–1151, San Jose California USA, May 2016. ACM.

[166] J. Nathan Matias. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 5(2):2056305119836778, April 2019. Publisher: SAGE Publications Ltd.

[167] J. Nathan Matias. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20):9785–9789, May 2019.

[168] Maxwell McCombs. A look at agenda-setting: Past, present and future. *Journalism studies*, 6(4):543–557, 2005.

[169] Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2), 1972.

[170] Warwick McKibbin and Roshen Fernando. The global macroeconomic impacts of covid-19: Seven scenarios. *Asian Economic Papers*, 20(2):1–30, 2021.

[171] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[172] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. *arXiv:2011.05773 [cs]*, November 2020. arXiv: 2011.05773.

[173] Hemant Misra, Olivier Cappé, and François Yvon. Using lda to detect semantically incoherent documents. pages 41–48. Association for Computational Linguistics, 2008.

[174] Ryan Mitchell. *Web Scraping with Python: Collecting More Data from the Modern Web*. " O'Reilly Media, Inc.", 2018.

[175] Sina Mohseni, Fan Yang, Shiva K Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric D Ragan. Machine learning explanations to prevent overtrust in fake news detection. In *ICWSM*, pages 421–431, 2021.

[176] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee. "fake news" is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist*, 65(2):180–212, 2021.

[177] Milton Mueller and Mawaki Chango. Disrupting global governance: the internet whois service, icann, and privacy. *Journal of Information Technology & Politics*, 5(3):303–325, 2008.

[178] Nikil S Mukerji. A conceptual analysis of fake news.

[179] Teresa K Naab, Dominique Heinbach, Marc Ziegele, and Marie-Theres Grasberger. Comments and credibility: how critical user comments decrease perceived news article credibility. *Journalism studies*, 21(6):783–801, 2020.

[180] Teresa K Naab, Anja Kalch, and Tino GK Meitz. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, 20(2):777–795, 2018.

[181] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. Social media is polarized. pages 957–970. ACM, 2018.

[182] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. pages 100–108. Association for Computational Linguistics, 2010.

[183] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *The 31st Annual ACM Symposium on User Interface Software and Technology - UIST '18*, pages 189–199, Berlin, Germany, 2018. ACM Press.

[184] Sakari Nieminen and Lauri Rapeli. Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, 17(3):296–309, 2019.

[185] Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. 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, pages 630–638, 2019.

[186] Michael Nunez. Former facebook workers: we routinely suppressed conservative news. *Gizmodo [Internet]*, 2016.

[187] Katherine Ognyanova, David Lazer, Ronald E Robertson, and Christo Wilson. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, 2020.

[188] Efeosasere Moibi Okoro, Benjamin Abara, Aneyelewa Alan-Ajonye, Zayyad Isa, and Alex Umagba. Effects of human and human-machine fake news detection approaches on user detection performance. *International Journal of Advanced Research in Computer Science*, 10(1), 2019.

[189] Alexandra Olteanu, Emre Kıcıman, and Carlos Castillo. A critical review of online social data: Biases, methodological pitfalls, and ethical boundaries. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 785–786, 2018.

[190] Cathy O'neil. *Weapons of math destruction*. Broadway Books, 2016.

[191] Leysia Palen and Amanda L Hughes. Social media in disaster communication. In *Handbook of Disaster Research*. Springer, 2018.

[192] Frosso Papanastasiou, Georgios Katsimpras, and Georgios Paliouras. Tensor Factorization with Label Information for Fake News Detection. *arXiv:1908.03957 [cs]*, August 2019. arXiv: 1908.03957.

[193] Deven Parekh, Drew Margolin, and Derek Ruths. Comparing audience appreciation to fact-checking across political communities on reddit. In *12th ACM Conference on Web Science*, pages 144–154, 2020.

[194] Dave Paresh. Without these ads, there wouldn't be money in fake news, December 2016.

[195] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[196] Gordon Pennycook and David G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc Natl Acad Sci USA*, 116(7):2521–2526, February 2019.

[197] Politifact staff. Politifact guide to fake news websites and what they peddle, 2018.

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

[199] P. Pourghomi, F. Safieddine, W. Masri, and M. Dordevic. How to stop spread of misinformation on social media: Facebook plans vs. right-click authenticate approach. In *2017 International Conference on Engineering MIS (ICEMIS)*, pages 1–8, May 2017. ISSN: 2575-1328.

[200] Poynter.org. IFCN Covid-19 Misinformation, 2020.

[201] Rodrigo Praino, Daniel Stockemer, and Vincent G Moscardelli. The lingering effect of scandals in congressional elections: Incumbents, challengers, and voters. *Social Science Quarterly*, 94(4):1045–1061, 2013.

[202] Nicolas Pröllochs. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *arXiv:2104.07175 [cs]*, May 2021. arXiv: 2104.07175.

[203] Riccardo Puglisi and James M Snyder Jr. Newspaper coverage of political scandals. *The Journal of Politics*, 73(3):931–950, 2011.

[204] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[205] Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 525–536, 2021.

[206] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568, 2020.

[207] Meet Rajdev and Kyumin Lee. Fake and spam messages: Detecting misinformation during natural disasters on social media. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 17–20. IEEE, 2015.

[208] K Rapoza. Can 'fake news' impact the stock market?, 2017.

[209] R Rehurek and P Sojka. Gensim–python framework for vector space modelling. 2011.

[210] Paul Resnick, Aljohara Alfayez, Jane Im, and Eric Gilbert. Informed crowds can effectively identify misinformation. *arXiv preprint arXiv:2108.07898*, 2021.

[211] Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *arXiv:1707.03264 [cs]*, July 2017. arXiv: 1707.03264.

[212] Regina Rini. Fake news and partisan epistemology. *Kennedy Institute of Ethics Journal*, 27(2):E–43, 2017.

[213] Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19. *arXiv:2107.11755 [cs]*, July 2021. arXiv: 2107.11755.

[214] Victoria L. Rubin. Semi-supervised Content-based Fake News Detection using Tensor Embeddings and Label Propagation. 2018.

[215] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A Hybrid Deep Model for Fake News Detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, pages 797–806, 2017. arXiv: 1703.06959.

[216] Johnny Saldaña. *The coding manual for qualitative researchers*. sage, 2021.

[217] S Schaedel. Websites that post fake and satirical stories. factcheck, 2018.

[218] Dietram A Scheufele, Eunkyung Kim, and Dominique Brossard. My friend's enemy: How split-screen debate coverage influences evaluation of presidential debates. *Communication Research*, 34(1):3–24, 2007.

[219] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Z. Tan, and Amy X. Zhang. Modular Politics: Toward a Governance Layer for Online Communities. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1):1–26, April 2021.

[220] Joseph Seering. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, October 2020.

[221] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–29, 2018.

[222] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. Metaphors in moderation. *New Media & Society*, 24(3):621–640, March 2022.

[223] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, July 2019. Publisher: SAGE Publications.

[224] Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. If you have a reliable source, say something: Effects of correction comments on covid-19 misinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 896–907, 2022.

[225] Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeno, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, et al. Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media. In *CLEF (Working Notes)*, 2020.

[226] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy. *WWW '16 Companion*, 2016. arXiv: 1603.01511.

[227] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. pages 96–104, 2017.

[228] Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1385, 2020.

[229] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19*, pages 395–405, Anchorage, AK, USA, 2019. ACM Press.

[230] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[231] Kai Shu, Suhang Wang, and Huan Liu. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on (MIPR)*, pages 430–435. IEEE, 2018.

[232] Craig Silverman. Here are 50 of the biggest fake news hits on facebook from 2016. 2016.

[233] Craig Silverman. The fake news watchdog, 2017.

[234] SimilarWeb. Website Demographic Data, 2019.

[235] Joseph P Simmons, Leif D Nelson, Jeff Galak, and Shane Frederick. Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, 38(1):1–15, 2011.

[236] Himani Singal and Shruti Kohli. Trust necessitated through metrics: estimating the trustworthiness of websites. *Procedia Computer Science*, 2016.

[237] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*, 2020.

[238] Monika Skaržauskaitė. The application of crowd sourcing in educational activities. *Social technologies*, 2(1):67–76, 2012.

[239] Isa Soares. The fake news machine: Inside a town gearing up for 2020, 2019.

[240] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*, pages 259–263, 2019.

[241] Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. Classification aware neural topic model for covid-19 disinformation categorisation. *PloS one*, 16(2):e0247086, 2021.

[242] Kate Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, pages 230–239, 2017.

[243] James H Stock, Mark W Watson, et al. *Introduction to econometrics*, volume 104. Addison Wesley Boston, 2003.

[244] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. Some Like it Hoax: Automated Fake News Detection in Social Networks. *arXiv:1704.07506 [cs]*, April 2017. arXiv: 1704.07506.

[245] Tetsuro Takahashi and Nobuyuki Igata. Rumor detection on twitter. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 452–457. IEEE, 2012.

[246] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining "fake news" a typology of scholarly definitions. *Digital Journalism*, 6(2):137–153, 2018.

[247] Franklin Tchakounté, Ahmadou Faissal, Marcellin Atemkeng, and Achille Ntyam. A Reliable Weighting Scheme for the Aggregation of Crowd Intelligence to Detect Fake News. *Information*, 11(6):319, June 2020.

[248] Mike Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. Springer, 2017.

[249] Joel Timmer. Fighting falsity: Fake news, facebook, and the first amendment. *Cardozo Arts and Ent. LJ*, 35:669, 2016.

[250] Kathie M d'I Treen, Hywel TP Williams, and Saffron J O'Neill. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):e665, 2020.

[251] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake News Detection in Social Networks via Crowd Signals. *arXiv:1711.09025 [cs]*, March 2018. arXiv: 1711.09025.

[252] Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2):1600008, 2017.

[253] Tom WG Van der Meer, Armen Hakhverdian, and Loes Aaldering. Off the fence, onto the bandwagon? a large-scale survey experiment on effect of real-life poll outcomes on subsequent vote intentions. *International Journal of Public Opinion Research*, 28(1):46–72, 2016.

[254] Dave Van Zandt. Media bias/fact check (mbfc news) about. https://mediabiasfactcheck.com/about/, 2018. Accessed: 2018-09-30.

[255] Van Zandt, Dave. Media bias/fact check (mbfc news), 2018.

[256] Chris J Vargo, Lei Guo, and Michelle A Amazeen. The agenda-setting power of fake news. *new media & society*, 20(5):2028–2049, 2018.

[257] Norman Vasu, Benjamin Ang, Terri-Anne Teo, Shashi Jayakumar, Muhammad Raizal, and Juhi Ahuja. *Fake news: National security in the post-truth era*. RSIS, 2018.

[258] Nguyen Vo and Kyumin Lee. The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 275–284, Ann Arbor MI USA, June 2018. ACM.

[259] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada, July 2017. Association for Computational Linguistics.

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

[261] Gang Wang, Manish Mohanlal, Christo Wilson, Xiao Wang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. Social Turing Tests: Crowdsourcing Sybil Detection. *arXiv:1205.3856 [physics]*, December 2012. arXiv: 1205.3856.

[262] William Wang. "Liar, Liar Pants on Fire". *arXiv:1705.00648*, 2017.

[263] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 849–857, New York, NY, USA, 2018. ACM. event-place: London, United Kingdom.

[264] Claire Wardle. The need for smarter definitions and practical, timely empirical research on information disorder. *Digital journalism*, 6(8):951–963, 2018.

[265] Claire Wardle, Hossein Derakhshan, et al. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. *Ireton, Cherilyn; Posetti, Julie. Journalism,'fake news'& disinformation. Paris: Unesco*, pages 43–54, 2018.

[266] Jen Weedon, William Nuland, and Alex Stamos. Information operations and facebook. *Retrieved from Facebook: https://fbnewsroomus. files. wordpress. com/2017/04/facebook-and-information-operations-v1. pdf*, 2017.

[267] Hilde JP Weerts. An introduction to algorithmic fairness. *arXiv preprint arXiv:2105.05595*, 2021.

[268] Galen Weld, Maria Glenski, and Tim Althoff. Political Bias and Factualness in News Sharing across more than 100,000 Online Communities. *arXiv:2102.08537 [cs]*, February 2021. arXiv: 2102.08537.

[269] Nicholas White. The daily dot, 2018.

[270] Wikipedia contributors. Alexa internet — Wikipedia, the free encyclopedia, 2019. [Online; accessed 4-May-2019].

[271] Nick Wingfield, Mike Isaac, and Katie Benner. Google and facebook take aim at fake news sites. *The New York Times*, Nov 2016.

[272] Donghee Yvette Wohn. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[273] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemom. Intell. Lab. Syst.*, 2(1-3):37–52, 1987.

[274] Di Wu, Tiantian Wu, Qun Liu, and Zhicong Yang. The sars-cov-2 outbreak: what we know. *International Journal of Infectious Diseases*, 94:44–48, 2020.

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

[276] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of 33rd AAAI Conference on Artificial Intelligence*, 2019.

[277] Hua Yuan, Jie Zheng, Qiongwei Ye, Yu Qian, and Yan Zhang. Improving fake news detection with domain-adversarial and graph-attention neural network. *Decision Support Systems*, 151:113633, 2021.

[278] Jason Shuo Zhang, Brian C. Keegan, Qin Lv, and Chenhao Tan. Understanding the Diverging User Trajectories in Highly-related Online Communities during the COVID-19 Pandemic. *arXiv:2006.04816 [cs]*, June 2020. arXiv: 2006.04816 version: 1.

[279] Jiawei Zhang, Bowen Dong, and Philip S. Yu. FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network. *arXiv:1805.08751 [cs, stat]*, August 2019. arXiv: 1805.08751.

[280] Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, 2021.

[281] Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. *arXiv preprint arXiv:2004.13606*, 2020.

[282] Xinyi Zhou and Reza Zafarani. Fake News: A Survey of Research, Detection Methods, and Opportunities. *arXiv:1812.00315 [cs]*, December 2018. arXiv: 1812.00315.

[283] Xinyi Zhou and Reza Zafarani. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter*, 21(2):48–60, 2019.

[284] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 836–837, 2019.

[285] Melissa Zimdars. My "fake news list" went viral. but made-up stories are only part of the problem. *The Washington Post*, 2016.