"This is Fake News": Characterizing the Spontaneous Debunking from Twitter Users to COVID-19 False Information

Kunihiro Miyazaki[1], Takayuki Uchiba[2], Kenji Tanaka[3], Jisun An[1], Haewoon Kwak[1], Kazutoshi Sasahara[4]

[1] Indiana University Bloomington
[2] Sugakubunka
[3] The University of Tokyo
[4] Tokyo Institute of Technology
{kunihirom,jisun.an,haewoon}@acm.org, takayuki.uchiba@sugakubunka.com,
tanaka@tmi.t.u-tokyo.ac.jp, sasahara.k.aa@m.titech.ac.jp

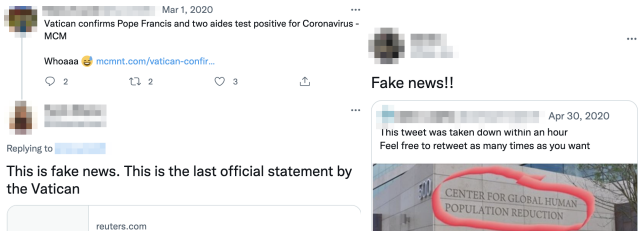
Abstract

False information spreads on social media, and fact-checking is a potential countermeasure. However, there is a severe shortage of fact-checkers; an efficient way to scale fact-checking is desperately needed, especially in pandemics like COVID-19. In this study, we focus on spontaneous debunking by social media users, which has been missed in existing research despite its indicated usefulness for fact-checking and countering false information. Specifically, we characterize the tweets with false information, or fake tweets, that tend to be debunked and Twitter users who often debunk fake tweets. For this analysis, we create a comprehensive dataset of responses to fake tweets, annotate a subset of them, and build a classification model for detecting debunking behaviors. We find that most fake tweets are left undebunked, spontaneous debunking is slower than other forms of responses, and spontaneous debunking exhibits partisanship in political topics. These results provide actionable insights into utilizing spontaneous debunking to scale conventional fact-checking, thereby supplementing existing research from a new perspective.

1 Introduction

The spread of false information has been a severe problem in our society (Lazer et al. 2018). In the recent COVID-19 pandemic, false information and its spread has been considered as dangerous as the virus (Naughton 2020). For example, the information about wrong treatment has led people to die (Coleman 2020), and a conspiracy theory about vaccines has made people less likely to get vaccinated, which unnecessarily undermines the utility of society as a whole (Burki 2019). The World Health Organization (WHO) called the prevalence of such false information an "infodemic" (WHO 2020) and has set it as an important global issue.

Various countermeasures have been implemented in order to combat this infodemic, among which fact-checking is a prominent approach. Fact-checking is an activity to verify the correctness of the information, news, and discourse spread (Walter et al. 2020), which is conducted by an individual, a group, or an organization, e.g., Snopes[1] and Vaccination Demand Observatory[2]. However, it takes time and resources to train fact-checkers (Graves 2017). Given the volume of false information generated and circulated on social media (Stewart 2020), it is almost impossible to have enough fact-checkers to verify all the suspicious information (Rodrigo 2020).



(a) Reply. (b) Quote Tweet (QT).

Figure 1: Examples of spontaneous debunking.

Academics and the industry have been exploring ways to scale fact-checking. The efforts can be divided into largely two directions: automated fact-checking (e.g., Liu et al. 2015a) and fact-checking by crowdsourcing (e.g., Allen et al. 2021). Automated fact-checking mainly relies on algorithms such as machine learning to detect false information and has already contributed to removing fake news from some platforms (Facebook 2021). Fact-checking by crowdsourcing exploits the wisdom of crowds. It is reported that their judgment has a high correlation with professional judgment in fact-checking (Epstein, Pennycook, and Rand 2020), although not yet as accurate as professionals (Godel et al. 2021). While these methods are promising to combat the infodemic, machine-learning-based models are known to be highly context-dependent (Bang et al. 2021); they tend to perform poorly for newly circulating false information, and crowdsourcing requires a significant amount of time and cost if the crowd needs to verify a large amount of information.

Spontaneous debunking, one type of debunking by crowds, is a user's act of pointing out false information in other users' posts on social media (i.e., social correction (Bode and Vraga 2018)) without any particular incen-

[1] https://www.snopes.com/

[2] https://vaccinationdemandobservatory.org/

tive, which we can often see in the wild. Figure 1 shows examples of spontaneous debunking on Twitter: Twitter users voluntarily pointed out that the original tweet contains false information by reply and retweet with comments (i.e., Quote tweet, QT). Understanding the characteristics of these spontaneous debunking actions allows various ways to leverage them. First, it may be possible to prioritize potential false information that needs verification by observing the amount and tendency of spontaneous debunking toward them. Second, by understanding their motivation, it may be possible to encourage more participation in debunking. Third, spontaneous debunkers can work as social sensors (Liu et al. 2015b) to detect false information quickly. Despite having great potential, spontaneous debunking has been less explored.

With this in mind, we aim to characterize spontaneous debunking on social media. We exhaustively collect tweets containing false information and responses (i.e., reply and QT) from other users toward these tweets. Then, we identify *debunking* tweets among the responses by using the language model fine-tuned by our annotated dataset. After a first glance at our dataset, we pose the following research questions (RQs) to examine tweets with false information and debunking:

- RQ1: What are the characteristics of tweets with false information (fake tweets) that tend to be debunked?
- RQ2: What are the characteristics of spontaneous debunkers?

Our analysis confirms that much of the fake news is left undebunked and debunking behavior is generally slower than other responses. Fake tweets related to the factual tweets, such as the status of infection around the world and the countermeasures against COVID-19, are especially less debunked than other topics. We also find that the most frequent debunkers are highly partisan and well connected in each group.

Our contributions are as follows. We create a spontaneous debunking dataset consisting of tweets with false information and the responses to them. We annotate 10,000 responses, including debunking and non-debunking behavior. We build the sentence classification model by the annotated samples and classify all the responses. The codes and data are available[3]. Our dataset will be a valuable resource for upcoming studies. We shed light on debunking behavior by characterizing fake tweets likely to be debunked, debunking tweets, and debunkers, which has been underexplored by previous research.

2 Related Works

2.1 Social Correction

Debunking is an action to inform authors of posts as inaccurate, which is reported to generally reduce false beliefs of the authors (Wood and Porter 2019). To debunk as much false information as possible is important because leaving false information undebunked is said to raise the "implied truth effect," (Pennycook et al. 2020) where undebunked false information gives its witnesses the impression that it is true. On the other hand, a potential "backfire effect," where being debunked would ironically counterstrengthen its originator's beliefs, has been debated for a long time (e.g., Lewandowsky et al. 2012). Nonetheless, recent research suggests that the support for this effect is small (e.g., Swire-Thompson, DeGutis, and Lazer 2020).

Debunking by social media users is referred to as social correction (Kligler-Vilenchik 2021). It is known that social correction is effective not only in curbing false perceptions of authors of fake posts (Shu et al. 2020) but also in reducing the false belief of those who witness false information being debunked on social media (Bode and Vraga 2018; Colliander 2019). In addition, debunking by crowdsourcing showed that the crowd's assessment of the credibility of the news publisher (e.g., publisher's website domain) was highly correlated with the assessment of professional fact-checkers, which showed the effectiveness of the wisdom of crowds, however unstable (e.g., Pennycook and Rand 2019). Moreover, in the literature on automated fake news detection, social contexts, such as responses from other users, are known to be essential signals to significantly improve the accuracy of the model (e.g., Cui, Wang, and Lee 2019). Nevertheless, the characteristics of debunking behavior from social media users have not been fully explored.

In this work, we call the debunking behavior "spontaneous debunking" as we focus particularly on the social correction without any incentives, which is different from artificially-generated debunking in lab/field experiments (Mosleh et al. 2021). The study by Vo and Lee (2018), which analyzed how social media users use fact-checked information, e.g., by Snopes, especially in replies, is the closest to our work. We also analyze replies, but the difference is that we ensure the original posts targeted by replies contain false information. This difference comes from the different purposes of the analysis. They ultimately focused on how the fact-checked information is used and shared, whereas our primary focus is on which fake tweets are most likely to attract debunking, which necessitates that the targeted tweets always contain false information. Additionally, we analyze "other responses" that are not debunking. By basing our analysis on comparisons between debunking behavior and other responses, we can more clearly characterize debunking behavior (e.g., in the analysis of "speed of debunking"). Also, Vo and Lee (2019) examined the linguistic characteristics of tweets containing fact-checked information and attempted to create a model for automatically generating fact-checked tweets, which is different from our primary objective of analyzing the behavioral characteristics of debunking.

2.2 Twitter Birdwatch

Twitter has recently started a new initiative called "Birdwatch" (Coleman 2021). It allows Twitter users to report suspicious tweets they encounter. Thus, it can be called fact-checking that leverages the collective intelligence of social media. The data of reported suspicious tweets are publicly available and have already started being analyzed (Pröllochs

[3]https://github.com/Mmichio/spontaneous_debunking_public

2021; Allen, Martel, and Rand 2022).

The difference between this study and Birdwatch is that Birdwatch collects the reporting from users, not publicly debunking the tweets that have false information. In addition, the Birdwatch data contains tweets reported as false by Twitter users only. This study analyzes both tweets with false information debunked by Twitter users *and* those undebunked.

3 Building a Spontaneous Debunking Dataset

We aim to analyze false information during COVID-19 and how it is debunked (and not debunked). In doing so, we first collect tweets that contain false information from various sources (§3.1), and then we collect 'responses,' i.e., replies and QTs (§3.2). We use the term *fake tweet* to refer to a tweet containing false information, such as false claims and fake news.

3.1 Collection of Fake Tweets

The widely used methods to collect fake tweets are machine-learning-based inference (Patwa et al. 2021), hashtag search (Al-Rawi, Groshek, and Zhang 2018), and getting neighbors of fake tweets using edit distance or distance in embedding space (Shaar et al. 2020). However, none of these methods can guarantee high veracity (e.g., Bang et al. 2021), and cannot avoid containing non-fake tweets. We need to avoid false positives as much as possible since the primary focus of this work is *debunking behavior* to fake tweets. To this end, we take two approaches: domain-based and claim-based approaches, which can achieve relatively high veracity in collecting fake tweets. In both, we leverage already-fact-checked tweets and domains from previous research.

Claim-based approach. We utilize tweet datasets that were manually fact-checked and confirmed as fake tweets in existing studies. We adopt the following conditions to select datasets: (1) fake tweets, (2) in English, (3) about COVID-19, (4) labeled by the fact-checking organizations or experts, (5) whose labels are publicly available, and (6) whose IDs are also publicly available (e.g., the dataset of Patwa et al. (2021) provides text, but we need tweet IDs for the search of responses). We examine publicly available datasets (see the survey of Murayama 2021) to see if they satisfy the above conditions. This results in the six datasets shown in Table 1. Some datasets have fewer tweets than the original because we only use tweets labeled as fake tweets.

Datasets	Count
CoAID (Cui and Lee 2020)	7,267
FibVID (Kim et al. 2021)	743
COVIDLies (Hossain et al. 2020)	114
COVID-Alam (Alam et al. 2021)	6
CMU-MisCov19 (Memon and Carley 2020)	855
Misinformation_COVID19 (Shahi, Dirkson, and Majchrzak 2021)	1,221

Table 1: Datasets that were manually fact-checked in existing studies and the number of tweets labeled as fake tweets.

These tweets total 18,929. We retrieve these tweets using their IDs and get 10,190 tweets (Table 2). The rest of the 8,738 tweets have already been deleted or their accounts are no longer public as of November 2021.

Domain-based approach. We also collect fake tweets by searching for tweets containing URLs of suspicious domains in our COVID-19-related tweet data set. Our COVID-19-related tweet dataset consists of 86,357,693 English-language tweets collected from February 2020 to February 2021 excluding retweets. The query words for building this dataset are "corona virus," "coronavirus," "COVID19," "2019-nCoV," "SARS-CoV-2," and "wuhanpneumonia." The suspicious domains are from the list from CoVaxxy (DeVerna et al. 2021), which is the dashboard of COVID-19 misinformation on social media. This list is publicly available and consists of 674 domains. We obtain the fake tweets by searching all domains in this list. As a result, we retrieved 333,470 tweets with URLs (Table 2).

Approach	Type of tweets		Count	Ratio (%)
Claim-based	All		10,190	100
	Responded	Reply	2,200	21.59
		QT	1,885	18.50
	Debunked	Reply	1,521	14.93
		QT	1,184	11.62
Domain-based	All		333,470	100
	Responded via	Reply	30,296	9.09
		QT	17,166	5.15
	Debunked via	Reply	10,504	3.15
		QT	3,429	1.03

Table 2: Statistics of fake tweets. The displayed numbers are the amount (and ratio) of all the fake tweets, those that get responses at least once and that is debunked at least once.

3.2 Collection of Responses to Fake Tweets

We collect replies and QTs as responses to fake tweets.

Replies. We search fake tweet IDs and obtain replies using conversation_id[4]. We get 2,639,104 replies. We then remove non-English replies (25.7% of the total). Also, as the search with conversation_id returns all reply trees, including replies to a reply, we only keep the direct replies to the fake tweets and eliminate further replies. Finally, we get 1,190,643 replies in total (1,011,179 for claim-based and 179,464 for domain-based fake tweets, see Table 3).

Quote tweets. We obtain QTs by searching with the condition that the URL included in the tweet contains the ID of the targeted tweet because Twitter's API treats the quoted tweets as URLs, just like news media. We obtain 884,060 QTs, and retain only those in English (34.8% of total). As a result, we get 527,007 QTs in total (470,349 for claim-based and 56,658 for domain-based fake tweets, see Table 3).

[4]https://developer.twitter.com/en/docs/twitter-api/conversation-id

Response	Original fake tweets	Total	Debunking	Ratio of debunking
Reply	Claim-based	1,011,179	481,538	47.6%
	Domain-based	179,464	58,475	32.6%
	Total	1,190,643	540,013	45.4%
QT	Claim-based	470,349	117,777	25.0%
	Domain-based	56,658	9,444	16.7%
	Total	527,007	127,235	24.1%

Table 3: Statistics of responses to fake tweets.

4 Detection of Debunking Behavior

We first identify debunking behaviors among the responses by using the sentence classification model. In this study, we build a machine learning model to classify whether or not the responses to fake tweets are debunking behavior, where the input is the text of the response and the target variable is the type of response.

4.1 Annotating Debunking Tweets

Since no ground truth labels for debunking tweets of COVID-19 false information exist, we manually label a subset of the responses to fake tweets using Amazon Mechanical Turk (MTurk). We aim to annotate 5,000 replies and QTs each. As the ratio of debunking in replies and QTs is unknown, we carefully design a multistep annotation process to avoid class imbalance, which often impacts the model performance (Tayyar Madabushi, Kochkina, and Castelle 2019). We first randomly sample 3,000 replies and 3,000 QTs and annotate them. Then, we find that 47.5% of replies and 25.2% of the QTs are labeled as debunking. We randomly sample 2,000 more replies because replies do not show a significant class imbalance. As for QTs, we randomly sample 2,000 more from quasi-debunking QTs, which is the preliminary model's output fine-tuned by the first 3,000 QTs. The preliminary model is trained with the undersampling technique to address the imbalances in the first 3,000 QTs.

We assign three annotators for each tweet in MTurk. We choose workers who 1) report their location as the U.S., 2) mark their approval rate as greater than 98%, and 3) have a number of approved tasks (HIT) greater than 10,000, referring to the official guideline of MTurk (MTurk 2019). Furthermore, we choose "Master Workers" whose quality of work is officially verified by MTruk (MTurk 2015). Since *debunking* may be an ambiguous word, we offer concrete guidelines and examples. We ask the annotators "Do you think the displayed reply/quote tweet (retweet with comment) is critical of the original tweet and debunks it?". We show examples to annotators as follows:

- Debunking: Point out fakes, Insult to tweet author, Refute with logic, Order to retract: e.g., "This is fake news," "You are mad, liar," "Ginger does not cure COVID-19," "Retract."
- Others (Support, Comment, Queries, etc.): e.g., "True," "How did it happen?" "I heard this news on yesterday."

If the class is ambiguous, we ask the annotators to choose Others. The categories and examples of debunking are iteratively refined through pilot tests. The examples of Others as Support, Comment, and Queries come from the previous study on rumor detection (Cheng, Nazarian, and Bogdan 2020; Cheng et al. 2021). As a result, we get the labels with the Fleiss Kappa score at 0.542 for replies and 0.541 for QTs, which are moderate agreement (Landis and Koch 1977). The majority voting of the three annotators decides the final label of each response. Table 4 shows the summary of the annotation results.

	Reply			QT		
	Agreement			Agreement		
Labels	2	3	Total	2	3	Total
Debunking	785	1,629	2,414	884	1,544	2,428
Others	934	1,652	2,586	837	1,735	2,572
Total	1,719	3,281	5,000	1,721	3,279	5,000

Table 4: Result of annotation by three annotators for Replies and QTs. The agreement indicates the number of annotators whose responses matched, i.e., 3 indicates the complete agreement. Labels are determined by majority vote, e.g., among the 5,000 replies, all three annotators agree 1,629 replies are debunking, while 785 replies are decided as debunking from voting with 2 Debunking vs. 1 Others.

4.2 Building a Debunking Classifier

We evaluate several choices to build the classification model. First, we adopt the basic machine learning models, i.e., Logistic Regression (LR), Random Forest (RF), and Linear SVM (SVM). As the input, we use embedding features by SentenceBERT (Reimers and Gurevych 2019), which has been popularly used to represent a short text (An et al. 2021). We also use pre-trained BERT models for prediction by fine-tuning them with our annotated dataset. In particular, we use: regular BERT (Devlin et al. 2019), trained on Wikipedia data (BERT-Wiki); BERTweet (Nguyen, Vu, and Nguyen 2020), trained on the Twitter corpus of general topics (BERT–Tweet); and COVID-Twitter-BERT, trained on the Twitter corpus on COVID-19 topics (BERT–Tweet–COVID19) (Müller, Salathé, and Kummervold 2020). For evaluation, we obtain the average F1 score by the 10-fold cross-validation.

	Reply		QT	
Model	Mean	SD	Mean	SD
SBERT + LR	0.703	±0.023	0.735	±0.019
SBERT + RF	0.673	±0.021	0.699	±0.024
SBERT + SVM	0.716	±0.017	0.736	±0.016
BERT–Wiki	0.739	±0.023	0.752	±0.023
BERT–Tweet	**0.792**	±0.018	0.783	±0.029
BERT–Tweet–COVID19	**0.792**	±0.018	**0.820**	±0.030

Table 5: F1 scores for prediction of debunking behaviors by the 10-fold cross-validation. The highest scores in Reply and QT are shown in bold.

Table 5 shows the performance of various models. We find that the BERT–Tweet–COVID19 model shows the best performance for both reply and QT—we obtain the F1 score of

0.792 (SD = ±0.018) for replies and 0.820 (SD = ±0.030) for QTs. To further demonstrate the generality of our model in the full dataset, we conduct a manual evaluation. We randomly extract 100 replies and 100 QTs that are not used in MTurk annotation, and then manually annotate them whether they are actual debunking or not. Compared to the predicted results, we get F1 scores of 0.773 for replies and 0.813 for QTs. These values are almost the same as the result of the original 10-fold cross-validation, and thus the result demonstrates the generality of our model. Also, as the classification model with an F1 score of about 0.8 has been employed in previous studies (He et al. 2021), in this study, we use our BERT–Tweet–COVID19 model to analyze the overall tendency of a large amount of data (330k fake tweets and 1.8m responses) in the subsequent analyses.

Finally, we fine-tune the model using all annotated samples and infer whether or not the collected replies and QTs are debunking. As a result, we get 559,702 debunkings from replies (44.8% of total) and 127,235 debunkings from QTs (24.1% of total), which are summarized in Table 3.

5 A First Look at Debunking

In this section, we conduct an exploratory analysis to provide an overview of debunking behavior.

Do fake tweets receive (debunking) responses? The majority of fake tweets do not get any responses. As Table 2 in §3.1 shows, 21.59% and 18.50% of claim-based fake tweets get replies and QTs, respectively, and 9.09% and 5.15% of domain-based fake tweets get replies and QTs, respectively. Among them, the tweets debunked by replies and QTs are even lower in number. In the same table, we can see the proportion dramatically drops to 1.03% (domain-based tweets debunked by QTs).

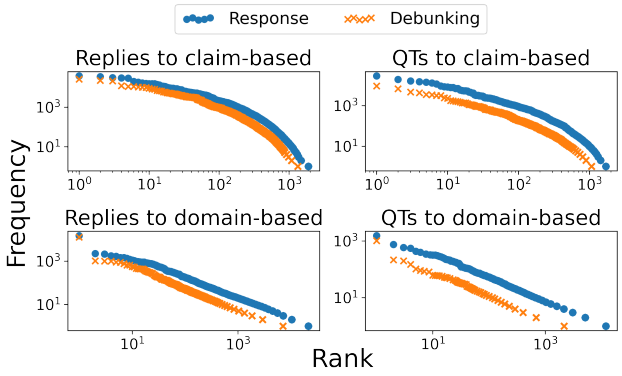


Figure 2: Log-log plots for fake tweets in terms of frequency of responses/debunking to fake tweets (x-axis) and their rank of frequency (y-axis).

Figure 2 shows its skewed nature; only a small number of fake tweets receive a large number of responses. In the figure, the relationship between the frequency of responses and debunking a fake tweet gets and its rank shows linearity in log-log space. Figure 2 also indicates the selection bias of claim-based fake tweets, i.e., manually collected fake tweets. The response to claim-based fake tweets is skewed slightly away from the linear shape (Figure 4a, 4b), where the frequencies of low-ranked fake tweets are less than the expected values. This may imply a potential selection bias that more prominent fake tweets are more likely to be collected in previous works. Alternatively, since the domain-based fake tweets are obtained exhaustively, there is no deviation from the linear shape (Figure 4c, 4d).

Correlation of responses and debunking. The more responses a fake tweet receives, the more debunking it receives. As shown in Figure 3, the number of other responses and debunking received by fake tweets have a linear relationship. In fact, their correlations are 0.527, 0.586, 0.486, and 0.543 for reply-claim-based, reply-domain-based, QT-claim-based, and QT-domain-based, respectively ($p<0.001$ for all values).

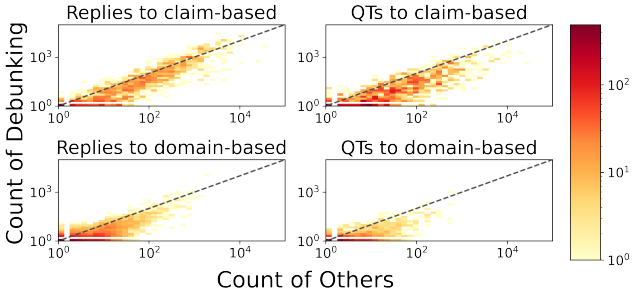


Figure 3: Heatmap for fake tweets in terms of the amounts of debunking they get (x-axis) and the amounts of other responses they get.

What kind of words in debunking? We identify the most representative words of debunking and other responses by using the log-odds ratio (Monroe, Colaresi, and Quinn 2008), which is widely used for comparing multiple corpora (An et al. 2021). We aggregate the corpus by aggregating all tweets in a group and comparing them with the tweets of the other group. We remove terms that appear less than ten times to avoid overemphasis on rare jargon and compute a log-odds ratio of all unigrams. As the prior, we compute the background word frequency using all tweets in our data collection. The unigrams are then ranked by their estimated z-scores. The result is shown in Figure 4.

Debunking contains many words that can be directly used for debunking (i.e., "lie," "liar," "fake"), regardless of replies and QTs. Conversely, other responses indicate encouraging words such as "thank" and "love" in replies and QTs, as well as words related to the conspicuous fake news, such as "Tom Hanks" (O'Rourke 2021), in QTs.

Since we see many offensive words in debunking tweets, we quantify the toxicity of the responses in each group. We use the Jigsaw's Perspective API[5] to quantify the toxicity. This API measures the toxicity of text on a scale of 0 to 1. As a result, the median scores are 0.411 for debunking and 0.116 for other responses with $p < 0.0001$ for the Mann-Whitney U test.

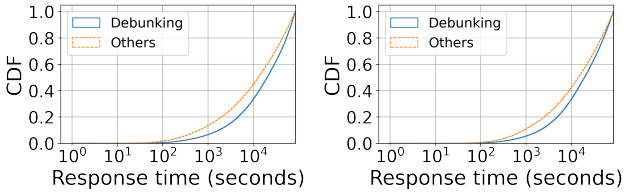
[5]https://www.perspectiveapi.com/

(a) Reply-debunking. (b) Reply-others.

(c) QT-debunking. (d) QT-others.

Figure 4: Outstanding words in responses to fake tweets. The size of a word in each wordcloud reflects the z-scores by log-odds ratio.

How quickly are fake tweets debunked? We also examine the timing of debunking. Figure 5 shows the CDF of the interval time between the responses and the targeted fake tweets. We can see that debunking occurs later than other responses in Reply and QT. For example, the median intervals in other responses were 13,024 seconds in reply and 14,429 seconds in QT, while for debunking, they were 20,438 seconds in reply and 19,820 seconds in QT. Other responses are faster than debunking at $p < 0.0001$ with the Mann-Whitney U test. Debunking is slower than other responses probably because debunking is an action that generally requires a lot of thought and research, and therefore takes more time on average than other responses.



(a) Reply. (b) QT.

Figure 5: CDF of interval time between the fake tweets and their responses.

6 RQ1: What Are the Characteristics of Fake Tweets That Tend to Be Debunked?

We investigate the potential reasons and motivations for users to debunk fake tweets. In particular, we focus on three different perspectives of fake tweets: who is sharing the tweet (*User features*), what the tweet is about (*Content features*), and how people react to it (*Engagement features*). We then conduct regression analysis to examine which features have the strongest predictive power to detect which fake tweets are likely to be debunked. Specifically, we predict whether or not fake tweets get at least one debunking response by using the logistic regression model, and analyze the regression coefficiency of the features. We conduct the prediction and analysis for reply and QT separately.

6.1 Regression Features

We select the following features that are expected to be relevant to the conditions of occurrence of spontaneous debunking based on related works.

User features. We use (1) followers count, i.e., how conspicuous the users are, (2) length of bio, i.e., user's willingness to appeal, and (3) verification by Twitter, i.e., user's publicity. Verification by Twitter is a binary measure of whether or not Twitter has authorized an account[6].

Content features. We consider topics and styles of fake tweets as content features.

Topic features. For topic extraction, we use a biterm topic model (Yan et al. 2013), which is known to work better for short sentences than other widely-used topic models such as LDA. We choose seven as the number of topics by comparing the perplexity scores (Zhao et al. 2015). We assign each tweet to one of the seven topics as dummy variables once we build the topic model. We determined that assigning one topic with the highest probability of belonging to one tweet is sufficient. The seven topics, their representative words, and the proportions of tweets for each topic are summarized in Table 6. Early emergence and spread of COVID-19 in other countries (e.g., China, Russia, and Iran (Andrew Osborn 2020)) and politics are the two most prevalent topics among fake tweets during COVID-19. We note that we use "Cases/Deaths" as a reference when creating dummy variables of topics for building the regression model to avoid the "dummy variable trap" (Gujarati 1970).

In addition, to account for potential differences between the two types of datasets, we also add a binary feature, whether an original fake tweet is claim-based or domain-based (Type of fake tweet: 1 for claim-based; 0 for domain-based), to the model.

Linguistic features. We use features that capture the linguistic style of fake tweets, such as (1) length of the text, i.e., the volume of information of tweets; (2) the existence of a URL, i.e., the existence of evidence; and (3) sentiment, i.e., impressions to readers. For the sentiment, we use a pre-trained model of Barbieri et al. (2020) and use the positive and negative values.

Engagement features. We use the retweet count as the feature of attention to fake tweets. The count of favorites is alternative, but they are highly correlated and produce multicollinearity; thus, we only choose retweet count. In addition, we consider the numbers of replies and QTs as features to account for the general likeliness of receiving replies and QTs.

6.2 Regression Results

Table 7 shows the results of the logistic regression. The p-values are computed using two-tailed z-tests. All variance

[6]https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Topic labels	Ratio	Representative words
Control measures	2.6%	ban, governor, lockdown, order, michigan, whitmer, house, que, force, task
Political figures	21.6%	biden, watch, video, mask, cuomo, joe, die, american, lockdown, donald
Status around the world	25.5%	lockdown, china, outbreak, due, test, case, crisis, russia, spread, iran
Cases/Deaths	14.3%	death, case, number, cdc, rate, toll, day, patient, state, study
Vaccine	12.4%	test, vaccine, positive, gate, bill, fight, covid, corona, treatment, study
Wuhan lab	12.9%	china, wuhan, lab, chinese, claim, time, hoax, scientist, doctor, video
Players against COVID-19	10.7%	bill, pelosi, democrat, china, relief, house, illegal, american, stimulus, crisis

Table 6: Topics of fake tweets by biterm topic model. The ratio is the percentage of tweet count assigned in each topic per total. Representative words are the top 10 words with the highest probability of occurring in each topic.

	Reply	QT	Mean	Std
(Intercept)	-8.51***	-10.84***	-	-
Followers count (log)	0.08***	0.23***	7.75	2.68
Length of bio (log)	0.03*	0.13***	4.03	1.58
Verification	0.03	-0.28***	0.11	0.32
Control measures	-2.02***	-3.57***	0.03	0.16
Political figures	0.02	0.44***	0.22	0.41
Status around the world	-0.60***	-0.39***	0.26	0.44
Vaccine	-0.07	0.18*	0.12	0.33
Wuhan lab	0.26***	0.36***	0.13	0.34
Players against COVID-19	0.00	0.56***	0.11	0.31
Type of fake tweet	-1.67***	-1.88***	0.03	0.17
Length of text (log)	0.04	0.20**	5.20	0.35
URL	2.43***	1.09***	1.00	0.07
Positive	0.12	-0.39**	0.09	0.17
Negative	0.59***	0.68***	0.43	0.29
Retweet count (log)	-0.15***	0.03	0.44	0.93
Reply count (log)	3.95***	0.10**	0.13	0.54
QT count (log)	-1.21***	2.53***	0.07	0.42
Pseudo R-squ.	0.57	0.55	-	-

Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7: Results of the regression analysis predicting whether or not a fake tweet gets debunked (number of fake tweets is N = 343,549). Each number indicates the regression coefficients.

inflation factors (VIFs) (O'brien 2007) are less than three, indicating that multicollinearity is negligible. We conduct log transform to some features when needed, indicated as (log) in Table 7.

Among the user features, followers count has positive associations in replies and QTs, which means tweets from conspicuous accounts are more likely to get debunks. Among the topic features, the topic of Wuhan lab has positive and significant coefficients both in reply and QT. The topic of political figures and players against COVID-19 has positive and significant coefficients in QT. In contrast, fake tweets about status around the world and control measures against COVID-19 are less likely to get debunked both in reply and QT. For the other topics, the coefficients tend to be greater than 0, although some of them are not significant. Also, claim-based fake tweets are less likely to get debunked compared to domain-based fake tweets (the type of fake tweet has significant negative coefficients). Among the linguistic features, fake tweets with URLs and negative fake tweets are more likely to get debunked in both reply and QT. Lastly, reply count shows a positive association for both reply and QT, and QT count has a negative value in reply and positive value in QT.

In summary, as expected, conspicuous accounts are more likely to receive debunking (follower count), and tweets that get more replies tend to get debunked in reply, as does QT. Also, tweets expressing negative emotions and containing URLs tend to get debunked. After controlling for these factors and focusing on topics, we find that debunking is more common for Wuhan lab in both reply and QT. Politics and players against COVID-19 get more debunking in QT. In contrast, there is less debunking to the topics such as the status of infection, measures around the world, and the countermeasures against COVID-19 itself.

7 RQ2: What Are the Characteristics of Spontaneous Debunkers?

In this section, we characterize debunkers by profile descriptions and networks.

We first examine how active users are in debunking fake tweets. Among all respondents, almost half of the users have debunked at least once (Debunking ≥ 1) by replies (340k (48.5%)), and a quarter of the users have debunked by QTs (97k (25.9%)). Those who have debunked three or more times (Debunking ≥ 3) are 46k (6.6%) with replies and 5.8k (1.6%) with QTs, indicating that it is not so common to debunk multiple times, but there is a considerable number of users who frequently debunk fake tweets.

7.1 Who Are Debunkers?

Difference between debunkers and non-debunkers. To better understand debunkers, we conduct a comparative analysis between debunkers and non-debunkers. For comparison, we define debunkers as those who have debunked more than 3 times and non-debunkers as those who have responded to fake tweets more than 3 times but have never debunked them. This criterion is to reduce the impact of the error of our debunking detection model on the analysis.

We examine the difference in bio descriptions between debunkers and non-debunkers. We calculate the log-odds ra-

tio of words in bios of debunkers and non-debunkers (Monroe, Colaresi, and Quinn 2008). We use our COVID-19-related tweets as a background corpus. Figure 6 shows the representative words of debunkers' and non-debunkers' bios. Interestingly, we see many political words in debunkers' bios. For example, words relating to conservatives, such as "maga," "conservative," and "trump," and those relating to liberals, such as "blue," "liberal," and "democrat" are observed. In other words, debunkers tend to be highly partisan accounts. Conversely, non-debunkers' representative words are far less political and include words like includes "health," "news," and "endorsement" (e.g., "RTs are not endorsements").

(a) Debunkers. (b) Non debunkers

Figure 6: Outstanding words in bios in debunkers and non-debunkers. The size of words in each wordcloud is along with the z-scores by log-odds ratio.

Types of debunkers. Next, we identify types of debunkers by applying the biterm topic model to their bios. We then examine the average debunk ratio for each group obtained by the topic model. Table 8 shows that the highest debunk ratio is found in the conservative group, followed by the liberal group, and then those without bios. In contrast, the users who have information about Business/Politics/Health in their bio have the lowest debunk ratio. The difference between these groups is significant at $p < 0.001$ for all pairs of Mann-Whitney U tests with Bonferroni correction. Business/Politics/Health is the only group that has the word "science" in the top 20, although it is not listed in the table, so it is likely that these users are relatively interested in science, but their debunk ratio is low.

7.2 How Are Debunkers Connected?

Finally, we examine social connections between debunkers to understand how far or close they locate on social media. We take the top 1,000 users with the most debunkings in replies and QTs each (1,644 unique users), and get all of their followers. We use Louvain clustering (Blondel et al. 2008) to identify clusters.

Figure 7 shows the follower network with identified clusters. We observe two large, separated clusters, which are slightly connected with each other. This indicates a clear echo-chamber/polarization phenomena (Barberá et al. 2015). By looking into the bios of each cluster, we found that purple is conservative (50.06%), and green is liberal (41.24%). Other than these two large clusters, the rest of the users in the center of Figure 7 are all isolated, not even connected with the two large clusters. Overall, the partisan

Topic labels	User ratio	Representative words	Debunk ratio
Conservative	16.3%	trump, proud, god, conservative, maga, husband, father, american, family, country	0.267
Business /Politics /Health	22.4%	view, writer, former, business, politics, news, health, opinion, director, social	0.166
Liberal	24.0%	mom, resist, blm, wife, dog, animal, life, proud, liberal, mother	0.230
Spiritual	7.7%	life, people, good, world, live, truth, work, try, god, take	0.205
Hobby	15.8%	fan, sport, music, game, husband, football, dad, father, movie, enthusiast	0.192
No bio	13.8%	-	0.223

Table 8: Topics in the bios of respondents. The debunk ratio is the average ratio of debunking tweets among their responses to fake tweets.

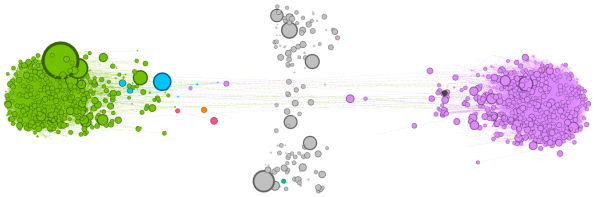


Figure 7: Follower network of top debunkers. Nodes indicate the users who conduct the debunking the most frequently (top 1,000 in reply and QT, 1,664 in total without duplicates). The size of nodes corresponds to the frequency of debunking. Edges indicate the follower relationship.

debunkers are highly connected, and they can see the debunking behavior of each other, which may lower the hurdle of debunking behavior.

8 Discussion and Conclusion

8.1 Main Findings

We have shown several remarkable findings in spontaneous debunking on Twitter. First, we confirm that many fake tweets are left undebunked and debunking behavior is generally slower than other responses. If we expect spontaneous debunkers to act as social sensors (Liu et al. 2015b), we need to design an environment that incentivizes them to debunk more and faster. Second, the topics of debunked fake tweets are unevenly distributed. Debunkings tend to be directed at fake tweets about Wuhan lab, seeming to be the most apparent conspiracy theory. Also, fake tweets about political figures and players against COVID-19 tend to get debunked partially. Instead, there was relatively little debunk-

ing to fake tweets about the global infection/control situation and infection control measurement itself (e.g., lockdown, masks). These may be topics that fact-checkers should focus on, as understanding the status of infections and countermeasures is vital information to encourage people to act rationally. Third, we found that the most frequent debunkers are partisan and connected well in each group. It seems political motivation drives debunking behaviors. In other words, we may consider enlisting the help of their debunking behavior for politically relevant topics. However, according to Allen et al. (2021), politically balanced people are more accurate in debunking; thus, the debunking of partisan users has to be used with caution. Then again, it may be necessary to encourage users with different interests to do the debunking for other topics. For example, researchers were requested to cooperate in countering fake news even before COVID-19 (Lazer et al. 2018).

8.2 Limitations and Future Works

Credibility of social correction. Even though a "wisdom of the crowds" is highly correlated with professional fact-checking, as Yasseri and Menczer pointed out, there is a risk in relying entirely on the social correction. In particular, (Allen et al. 2021) reported that when the partisanship of debunkers becomes high, the correctness of debunking lessens. It is challenging to identify meaningful information from social correction for fact-checking.

Alternatively, it may be possible to examine the differences in debunking among different types of debunkers. For example, this study showed that many of the top debunkers are partisan, and it could be possible that what is true for one side may be seen as a fake by the other side. In such a situation, it would be interesting to find out which posts were judged as false by both parties.

Definition of debunking. Our working definition of debunking is an action to inform authors of posts as inaccurate. We use this broad definition in order to capture the overall tendency of possible debunking behaviors after separating all responses into debunking and other responses. Our analysis then found multiple styles of debunking, including "pointing out" and "insulting." We have to note that insulting is not necessarily considered as debunking in some cases, as such behavior is often driven by partisan motivation in political information. A detailed categorization of debunking and characterization of different categories can be considered in a future study. Also, while this study focuses on debunking, focusing on supporting responses may help to find fake tweets that people tend to accept. In this case, we would find blind spots to improve digital media literacy (The Lancet 2022).

Classification accuracy This study is highly dependent on the quality of the classifier. Therefore, we built a model with high accuracy to conduct the subsequent analyses and the performance of the model is comparable with an existing study (He et al. 2021). To demonstrate the generality of our fine-tuned model, we also conducted a manual evaluation of the prediction results (in §4.2).

Moreover, we further conducted an error analysis to understand how to improve our model. We hypothesized that it would also be difficult for the annotators to make an accurate decision for the tweets the model failed to predict. We analyzed the relationship between the disagreement of the annotators' judgments and the incorrectness of the model prediction. Specifically, we calculated the F1 score with respect to a binary variable of whether the three annotators agreed or disagreed and a binary variable of whether the predicted result was correct or not, and found a certain positive correlation: 0.63 for replies and 0.62 for QTs. The contents that are difficult for people to judge are also difficult for the classifier. However, there are various patterns in missed tweets, so further research may be needed.

Lastly, it would be better to achieve even greater accuracy in the classification of debunking behavior. Therefore, future research includes further improvement of accuracy. When it comes to the balance of recall and accuracy, both the certainty of the predicted result and the target coverage are important, and it is not easy to prioritize one over the other. However, we note that precision would be more important considering the confidence in the subsequent analyses.

Bias in datasets The domain-based approach uses keyword-based tweet collection to gather COVID-19-related tweets, but not all COVID19-related tweets may have been collected, such as the name of emerging variants. However, since this is a trend analysis using large-scale data, some lack of keywords can be considered to have a minor impact on the analysis.

In addition, in this study, we assumed all information from suspicious domains as false information. While this is a common assumption used to study fake-news detection and spread (Baly et al. 2020), not all information is completely false, even if the domain is suspicious. In particular, factual information, such as infection status, is less likely to be false, which may have affected the regression analysis results. A deeper investigation may be required in the future.

In terms of the result of our regression analysis, we found that people tend to debunk domain-based fake news more than claim-based fake news, which is consistent between RT and QT. This indicates that the debunking of claim-based fake tweets is less likely to occur than domain-based fake tweets. This may imply that "suspicious domains" can be an easy indicator of fake information, and thus, domain-based fake tweets are more easily identified and debunked. This result highlights that online space needs better support for recognizing claim-based fake tweets.

Application to other topics This work introduces a framework to analyze debunking behaviors. Since the analysis procedure of this study is not topic-dependent (i.e., we can collect and analyze fake tweets and responses on an arbitrary topic), it can also be applied to other topics. In particular, since some debunking behavior is not dependent on the content of the fake tweet (e.g., debunking such as "This is fake" can exist in any topic), it is possible that our debunking classifier can be applied to other fake news in different domains. However, some debunking behaviors are indeed content-dependent, and we cannot make accurate speculation at this time. Transfer learning of debunking classification sounds like a very interesting topic and we will leave it as a future research topic.

Ethical Statement

We pay the utmost attention to the privacy of individuals in this study. We did not include personal names or account names in our analysis. Moreover, in the example figures, we blur user identity-related features (name, photo, and user id) to maintain anonymity. Lastly, for sharing our tweet data, we will publish only a list of tweet IDs, without any text or information, according to Twitter's guidelines.

References

Al-Rawi, A.; Groshek, J.; and Zhang, L. 2018. What the fake? Assessing the extent of networked political spamming and bots in the propagation of# fakenews on Twitter. *Online Information Review*.

Alam, F.; et al. 2021. Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to Arms. In *ICWSM*, volume 15, 913–922.

Allen, J.; Arechar, A. A.; Pennycook, G.; and Rand, D. G. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36).

Allen, J.; Martel, C.; and Rand, D. G. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *CHI*, 1–19.

An, J.; Kwak, H.; Lee, C. S.; Jun, B.; and Ahn, Y.-Y. 2021. Predicting Anti-Asian Hateful Users on Twitter during COVID-19. *EMNLP*.

Andrew Osborn, P. N. 2020. Russia's coronavirus cases surge past 100,000 after record daily rise — Reuters. https://www.reuters.com/article/idUSKBN22C12L. Accessed: 2021-12-17.

Baly, R.; Karadzhov, G.; An, J.; Kwak, H.; Dinkov, Y.; Ali, A.; Glass, J.; and Nakov, P. 2020. What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context. In *ACL*.

Bang, Y.; Ishii, E.; Cahyawijaya, S.; Ji, Z.; and Fung, P. 2021. Model Generalization on COVID-19 Fake News Detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, 128–140. Cham: Springer International Publishing.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10): 1531–1542.

Barbieri, F.; Camacho-Collados, J.; Anke, L. E.; and Neves, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *EMNLP*, 1644–1650.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Bode, L.; and Vraga, E. K. 2018. See something, say something: correction of global health misinformation on social media. *Health communication*, 33(9): 1131–1140.

Burki, T. 2019. Vaccine misinformation and social media. *The Lancet Digital Health*, 1(6): e258–e259.

Cheng, M.; Nazarian, S.; and Bogdan, P. 2020. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In *TheWebConf*, 2892–2898.

Cheng, M.; Wang, S.; Yan, X.; Yang, T.; Wang, W.; Huang, Z.; Xiao, X.; Nazarian, S.; and Bogdan, P. 2021. A COVID-19 Rumor Dataset. *Frontiers in Psychology*, 12.

Coleman, A. 2020. 'Hundreds dead' because of Covid-19 misinformation - BBC News. https://www.bbc.com/news/world-53755067. Accessed: 2022-01-03.

Coleman, K. 2021. Introducing Birdwatch, a community-based approach to misinformation. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation. Accessed: 2021-12-20.

Colliander, J. 2019. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, 97: 202–215.

Cui, L.; and Lee, D. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Cui, L.; Wang, S.; and Lee, D. 2019. Same: sentiment-aware multi-modal embedding for detecting fake news. In *ASONAM*, 41–48.

DeVerna, M. R.; et al. 2021. CoVaxxy: A Collection of English-Language Twitter Posts About COVID-19 Vaccines. In *ICWSM*, volume 15, 992–999.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Epstein, Z.; Pennycook, G.; and Rand, D. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *CHI*, 1–11.

Facebook. 2021. Facebook Removed 20 Million Pieces of Covid-19 Misinformation - Bloomberg. https://www.bloomberg.com/news/articles/2021-08-18/facebook-removed-20-million-pieces-of-covid-19-misinformation. Accessed: 2021-12-17.

Godel, W.; Sanderson, Z.; Aslett, K.; Nagler, J.; Bonneau, R.; Persily, N.; and Tucker, J. A. 2021. Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*, 1(1).

Graves, L. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture & Critique*, 10(3): 518–537.

Gujarati, D. 1970. Use of dummy variables in testing for equality between sets of coefficients in linear regressions: A generalization. *The American Statistician*, 24(5): 18–22.

He, B.; Ziems, C.; Soni, S.; Ramakrishnan, N.; Yang, D.; and Kumar, S. 2021. Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis. In *ASONAM*, 90–94.

Hossain, T.; Logan IV, R. L.; Ugarte, A.; Matsubara, Y.; Young, S.; and Singh, S. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Kim, J.; Aum, J.; Lee, S.; Jang, Y.; Park, E.; and Choi, D. 2021. FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics and Informatics*, 64: 101688.

Kligler-Vilenchik, N. 2021. Collective Social Correction: Addressing Misinformation through Group Practices of Information Verification on WhatsApp. *Digital Journalism*, 1–19.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Lazer, D. M.; et al. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.

Lewandowsky, S.; Ecker, U. K.; Seifert, C. M.; Schwarz, N.; and Cook, J. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3): 106–131.

Liu, X.; Nourbakhsh, A.; Li, Q.; Fang, R.; and Shah, S. 2015a. Real-time rumor debunking on twitter. In *CIKM*, 1867–1870.

Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; and Shi, L. 2015b. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105(3): 512–530.

Memon, S. A.; and Carley, K. M. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4): 372–403.

Mosleh, M.; Martel, C.; Eckles, D.; and Rand, D. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *CHI*, 1–13.

MTurk. 2015. Simplified Masters Qualifications — by Amazon Mechanical Turk — Happenings at MTurk. https://blog.mturk.com/simplified-masters-qualifications-137d77647d1c. Accessed: 2021-12-17.

MTurk. 2019. Qualifications and Worker Task Quality. https://blog.mturk.com/qualifications-and-worker-task-quality-best-practices-886f1f4e03fc. Accessed: 2021-12-17.

Müller, M.; Salathé, M.; and Kummervold, P. E. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv preprint arXiv:2005.07503*.

Murayama, T. 2021. Dataset of Fake News Detection and Fact Verification: A Survey. *arXiv:2111.03299*.

Naughton, J. 2020. Fake news about Covid-19 can be as dangerous as the virus — The Guardian. https://www.theguardian.com/commentisfree/2020/mar/14/fake-news-about-covid-19-can-be-as-dangerous-as-the-virus. Accessed: 2021-12-17.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. In *System Demonstrations at EMNLP*, 9–14.

O'Rourke, C. 2021. PolitiFact — No, Tom Hanks didn't die. https://www.politifact.com/factchecks/2021/mar/17/viral-image/no-tom-hanks-didnt-die/. Accessed: 2022-01-10.

O'brien, R. M. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5): 673–690.

Patwa, P.; et al. 2021. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, 42–53. Springer.

Pennycook, G.; Bear, A.; Collins, E. T.; and Rand, D. G. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11): 4944–4957.

Pennycook, G.; and Rand, D. G. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7): 2521–2526.

Pröllochs, N. 2021. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *arXiv:2104.07175*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP)*, 3982–3992.

Rodrigo, C. M. 2020. Critics fear Facebook fact-checkers losing misinformation fight — TheHill. https://thehill.com/policy/technology/478896-critics-fear-facebook-fact-checkers-losing-misinformation-fight?rl=1. Accessed: 2021-12-17.

Shaar, S.; Babulkov, N.; Da San Martino, G.; and Nakov, P. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *ACL*, 3607–3618.

Shahi, G. K.; Dirkson, A.; and Majchrzak, T. A. 2021. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22: 100104.

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

Stewart, E. 2020. America's growing fake news problem on social media, in one chart - Vox. https://www.vox.com/policy-and-politics/2020/12/22/22195488/fake-news-social-media-2020. Accessed: 2022-01-04.

Swire-Thompson, B.; DeGutis, J.; and Lazer, D. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*.

Tayyar Madabushi, H.; Kochkina, E.; and Castelle, M. 2019. Cost-Sensitive BERT for Generalisable Sentence Classification on Imbalanced Data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 125–134. Hong Kong, China.

The Lancet. 2022. The state of science and society in 2022. *The Lancet*, 399(10319): 1.

Vo, N.; and Lee, K. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *SIGIR*, 275–284.

Vo, N.; and Lee, K. 2019. Learning from fact-checkers: analysis and generation of fact-checking language. In *SIGIR*, 335–344.

Walter, N.; Cohen, J.; Holbert, R. L.; and Morag, Y. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3): 350–375.

WHO. 2020. Managing the COVID-19 infodemic. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation. Accessed: 2022-01-04.

Wood, T.; and Porter, E. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1): 135–163.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *TheWebConf*, 1445–1456.

Yasseri, T.; and Menczer, F. 2021. Can crowdsourcing rescue the social marketplace of ideas? *arXiv:2104.13754*.

Zhao, W.; Chen, J. J.; Perkins, R.; Liu, Z.; Ge, W.; Ding, Y.; and Zou, W. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics*, volume 16, 1–10. Springer.