Characterizing and Predicting Social Correction on Twitter

Yingchen Ma
Georgia Institute of Technology
Atlanta, Georgia, USA
yma473@gatech.edu

Bing He
Georgia Institute of Technology
Atlanta, Georgia, USA
bhe46@gatech.edu

Nathan Subrahmanian
Brandeis University
Waltham, Massachusetts, USA
nsubrahmanian@brandeis.edu

Srijan Kumar
Georgia Institute of Technology
Atlanta, Georgia, USA
srijan@gatech.edu

ABSTRACT

Online misinformation has been a serious threat to public health and society. Social media users are known to reply to misinformation posts with counter-misinformation messages, which have been shown to be effective in curbing the spread of misinformation. This is called social correction. However, the characteristics of tweets that attract social correction versus those that do not remain unknown. To close the gap, we focus on answering the following two research questions: (1) "Given a tweet, will it be countered by other users?", and (2) "If yes, what will be the magnitude of countering it?". This exploration will help develop mechanisms to guide users' misinformation correction efforts and to measure disparity across users who get corrected. In this work, we first create a novel dataset with 690,047 pairs of misinformation tweets and counter-misinformation replies. Then, stratified analysis of tweet linguistic and engagement features as well as tweet posters' user attributes are conducted to illustrate the factors that are significant in determining whether a tweet will get countered. Finally, predictive classifiers are created to predict the likelihood of a misinformation tweet to get countered and the degree to which that tweet will be countered. The code and data is accessible on https://github.com/claws-lab/social-correction-twitter.

CCS CONCEPTS

• **Information systems** → Social networks.

KEYWORDS

Misinformation, Counter-misinformation, Social Correction, Twitter, COVID-19 vaccines

ACM Reference Format:
Yingchen Ma, Bing He, Nathan Subrahmanian, and Srijan Kumar. 2023. Characterizing and Predicting Social Correction on Twitter. In *15th ACM Web Science Conference 2023 (WebSci '23), April 30–May 01, 2023, Evanston, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3578503.3583610

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

ACM ISBN 979-8-4007-0089-7/23/04.
https://doi.org/10.1145/3578503.3583610

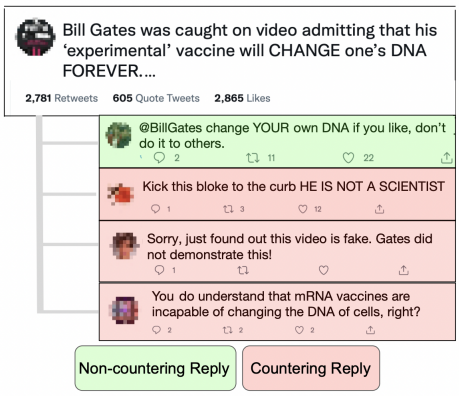


Figure 1: Examples of misinformation tweets and counter-misinformation replies.

1 INTRODUCTION

Online misinformation leads to societal harm including diminishing trust in vaccines and health policies [6, 49], damaging the well-being of users consuming misinformation [35, 63], encouraging violence and harassment [5, 60], and posing a danger to democratic processes and elections [57–59]. The problem has been exacerbated during the COVID-19 pandemic [40, 56]; particularly, COVID-19 vaccine misinformation including false claims that the vaccine causes infertility, contains microchips, and even changes one's DNA and genes has fueled vaccine hesitancy and reduced vaccine uptake [56]. Therefore, it is crucial to restrain the spread of online misinformation [36, 40]. In this work, we use a broad definition of misinformation which contains rumors, falsehoods, inaccuracies, decontextualized truths, or misleading leaps of logic [35, 68].

To combat misinformation, various countermeasures have been developed [40, 42, 65]. Recent work has shown that ordinary users of online platforms play a crucial role in countering misinformation. According to the research study by Micallef et al. [40], the vast majority (96%) of online counter-misinformation responses are made by ordinary users, with the remainder being made by professionals such as fact-checkers and journalists. While fact-checking from these professionals has been widely used due to its prominent and measurable impact [40, 65], this process typically does not involve engaging with the actors spreading misinformation. Instead, the ordinary users' counter-misinformation efforts complement those from professional fact-checkers by directly engaging in countering

conversations through making independent posts or direct replies to misinformation posts made by others [25].

Countering of misinformation messages via direct replies from ordinary users is called *social correction* [7, 34]. One real example is shown in Figure 1. Notably, social correction has been shown to be effective in curbing the spread of misinformation [13, 66], as well as doing so without causing increases in misperception [20, 61, 67]. While certainly not a panacea for convincing people to reconsider potentially misinformative beliefs, they are most effective at reducing the misperceptions of those who may consume it [7, 8, 13, 54].

However, little is known about the characteristics of misinformation tweets that attract social correction. Developing this understanding has several advantages: (1) first, it can help identify inequities in misinformation correction. For example, comparison of correction across users or communities (e.g., political ideologies) can reveal whether certain user types/communities are less likely to be self-correcting, e.g., communities where users correct misinformation when they see it. Identifying these disparities is the first step towards addressing them by redirecting resources towards entities that require external attention to curb misinformation; (2) second, if certain misinformation content is less likely to be socially corrected, targeted efforts can be directed toward countering them. Such instances can be escalated and prioritized for interventions by professionals or social media platforms; (3) third, if certain misinformation content is likely to be socially corrected, then additional participants can be encouraged to provide reinforcements.

Despite these promising benefits, characterizing and predicting social correction is non-trivial due to several challenges. First, existing datasets do not contain conversation-style narratives with paired misinformation posts and counter-replies. Second, existing works (including Miyazaki et al. [42]) do not analyze counter-replies to misinformation in a stratified manner where tweets with different numbers of replies are considered separately. This fine-grained analysis is necessary since comparing across or aggregating statistics across tweets that have drastically different numbers of (counter-)replies can skew the findings [27, 31].

In this work, we seek to characterize and predict counter-replies to misinformation. The contributions can be summarized as follows:

- We curate a novel large-scale dataset that contains 1,523,849 misinformation tweets and 690,047 counter-misinformation replies, along with a hand-annotated dataset of misinformation tweets and counter-replies.
- We perform a stratified, fine-grained analysis of the linguistic, engagement, and poster-level characteristics of misinformation tweets that get countered versus those that do not. Our analysis reveals several features of tweets that attract social correction, such as anger and impoliteness.
- We create two counter-reply prediction models to identify whether a misinformation tweet will be countered or not, and if so, to what degree (i.e. low or high), based on its linguistic, engagement, and poster features. We achieve promising predictive performance with both of these models, with best F-1 scores of 0.860 and 0.801, respectively.

The code and data is accessible on https://github.com/claws-lab/social-correction-twitter.

2 RELATED WORKS

2.1 Social Correction on Social Media Platforms

Misinformation widely spreads on social media platforms, which has caused detrimental effects on society [5, 13, 60], including harassment and personal attacks [40]. To combat misinformation, users actively employ various strategies [43], including replying to and commenting on misinformation [34, 42, 65]. This debunking behavior can broadly reduce the misinformed beliefs of the author and the audience who see the misinformation [7, 12]. Notably, current research works have shown the promising impact of debunking [7, 12] in both curbing the perception of misinformation and reducing the belief of false information [12]. In this work, we deep-dive into this misinformation-countering behavior by looking at both the misinformation posts and the counter-misinformation replies to these posts. Since user response information can indicate the textual properties of misinformation posts that are highly likely to get countered, our work sheds light on better understanding of misinformation-countering behavior, especially, understanding the misinformation tweets that get countered.

2.2 Analysis of Counter-misinformation

Due to the significance of counter-misinformation messages in curbing misinformation, much research has been focused on analyzing and understanding counter-misinformation [40, 65].

One type of work is to analyze and compare misinformation and counter-misinformation messages [40, 65]. For instance, Micallef et al. [40] first created a textual classifier to classify tweets into misinformation, counter-misinformation, and irrelevant groups, and then analyzed the tweets in each group. Interestingly, they find that a surge in misinformation tweets results in a corresponding increase in tweets that reject such misinformation. Vo and Lee [65] first identified fact-checking replies by checking whether a reply contains a fact-checking URL toward two trustworthy fact-checking websites (i.e., Snopes.com and Politifact.com). Then, they retrieved the corresponding misinformation tweet toward which the fact-checking post replies to, and use them to construct pairs of misinformation posts and fact-checking replies for fact-checking content analysis and reply generation.

Meanwhile, Miyazaki et al. [42] curated a large-scale dataset containing pairs of misinformation tweets and debunking replies, by first crawling COVID-19 related misinformation tweets from existing research [14, 28, 33, 39, 55] and then recruiting crowd-sourcing workers via Amazon Mechanical Turk to annotate responses to these tweets as being debunking or not. They then perform analysis to illustrate who counters misinformation and how they do so. However, contrary to this work, we conduct an in-depth *stratified* analysis of the replies to examine which features matter during the countering. Stratification helps to compare similar tweets by controlling for the number of replies it receives. Furthermore, we also conduct analysis of whether tweets get a high or low proportion of countering replies. Importantly, we also build two new tasks of predicting which misinformation posts will get countered and to what degree they get countered. Our work complements the existing counter-misinformation studies.

2.3 Birdwatch (a.k.a. Community Note)

Twitter launched Birdwatch (recently renamed to Community Note) to facilitate misinformation detection by ordinary users. On the platform, users can report suspicious and/or misleading tweets, as well as annotate tweets reported by others. Many have investigated this kind of countering [3, 44] and derived different patterns among this collective countering. For instance, Allen et al. [3] looks at the impact of partisanship during the crowds' annotation by analyzing existing data from the Birdwatch/Community Note platform; they find its users are more likely to (1) give negative annotations of tweets from counter-partisans, and (2) rate annotations from counter-partisans as unhelpful. Though Birdwatch/Community Note enables community-based detection of misinformation, it does not provide a way for users to counter misinformation. Notably, users provide inputs within the Birdwatch ecosystem only, which is restricted and does not reflect the larger dynamics of information flow on Twitter. Recent research has also shown that Birdwatch can be manipulated by motivated bad actors [44]. Therefore, we focus on the misinformation that spreads on Twitter and is countered by ordinary users for a more complete and comprehensive study.

3 DATASET

In this section, we describe the definition of the problem, as well as the corresponding dataset curation.

3.1 Definitions

Misinformation: We employ a broad definition of misinformation which includes falsehoods, inaccuracies, rumors, or misleading leaps of logic [68]. Building on the existing work [24], we focus on misinformation related to the COVID-19 vaccine due to its broad impact around the world during the COVID-19 pandemic. Practically, the misinformative claims include "the vaccine alters DNA", "the vaccine causes infertility", "the vaccine contains dangerous toxins", and "the vaccine contains tracking devices"; these topics are popular and widely studied by existing research works [1, 24].

Counter-reply: Motivated by existing research works on analyzing replies that show disbelief toward misinformation [32] or fact-check misinformation [48], a direct response to a misinformation post m is considered a "countering" reply if it makes an attempt to explicitly or implicitly debunk or counter the misinformation tweet m. Otherwise, the reply is considered as non-countering. Practically, given a reply r, it is a:

- **Countering reply:** Motivated by existing research works on identifying and analyzing text that is countering, debunking, disbelieving, or disagreeing with misinformation [28, 32, 40], a countering reply is a reply that explicitly or implicitly refutes the misinformation post ("this is misinformation"), points out the falsehood ("the COVID-19 vaccine does not change DNA"), insults the tweet poster ("you are born to lie"), or questions the misinformation ("Is there any reference I can check?").
- **Non-countering reply:** Instead of countering, a non-countering reply supports, is in favor of, comments, repeats misinformation, etc., such as "This is not the vaccine but the gene therapy", "Yes, I agree with you", or "It makes sense".

A post m is considered to be countered if it receives at least one counter-reply. Meanwhile, given that different misinformation tweets have various numbers of replies, to have a normalized measure of the magnitude of which a misinformation tweet gets countered, we define the proportion of counter-replies to total replies, denoted as $p(m)$.

3.2 Task Objective

We consider the set $\mathcal{M}$ of misinformation posts about the COVID-19 vaccine. Each misinformation post $m \in \mathcal{M}$ has a set of n replies $r = [r_1, ..., r_n]$ posted in direct response to m. Our final goal is to build a classifier $\mathcal{F}$ such that it can output a binary label $\mathcal{F}(m)$, which indicates whether the misinformation post will be countered or not, i.e., whether it will receive at least one counter-reply.

3.3 Dataset Curation

3.3.1 Misinformation Tweet Collection and Classification. We utilize the Anti-Vax dataset from Hayawi et al. [23], a large-scale dataset of tweets related to the topic of COVID-19 vaccines, in order to identify misinformation tweets for our study. These tweets range eight months from December 1, 2020 to July 31, 2021, which was the relevant period covering a substantial part of the time from when the vaccines were approved by the FDA in December 2020 [56]. Also during this period, many uncertainties and misinformation about COVID-19 vaccines were spreading on social media [23, 45, 56]. The original dataset consists of approximately 15.4 million tweets collected from the Twitter API [23], each containing at least one of the following COVID-19 vaccine relevant keywords: {'vaccine', 'pfizer', 'moderna', 'astrazeneca', 'sputnik', 'sinopharm'}. Only original tweets were considered, i.e., retweet, reply, or quote tweets were removed. We utilized the Twitter API to retrieve the tweet text, user ID of the tweet author, datetime, conversation ID, reply settings, and tweet engagement metrics (like, retweet, quote, and reply counts). In total, we were able to retrieve 14,123,209 tweets from the original dataset while the remaining 1.3 million tweets were unavailable due to the deletion by the users or the Twitter platform.

Following the definition of misinformation in Section 3.1 and the current approach of identifying COVID-19 vaccine related misinformation tweets [23], we first get the annotated misinformation tweets from Hayawi et al. [23], train a text classifier to determine if a tweet is misinformation or not, and classify all non-annotated tweets. Specifically, we first crawl and get 4,836 annotated misinformation and 8,596 annotated non-misinformation tweets from Hayawi et al. [23]. Next, we build a text classifier using BERT [16]. This classifier has a promising performance in precision, recall, and F-1 scores of 0.972, 0.979, and 0.975, respectively. This performance is comparable to the reported one in the original paper by Hayawi et al. [23] (i.e., the precision, recall, and F-1 scores of 0.97, 0.98, and 0.98). The classifier has high performance as per the metrics and thus can be used for downstream classification tasks.

Finally, we use this misinformation classifier to identify misinformation tweets in the entire dataset, resulting in 1,523,849 misinformation tweets and 12,599,360 non-misinformation tweets. Since we only focus on replies to misinformation in this work, we only use misinformation tweets for downstream analyses.

Next, we perform filtering of the dataset. Since our work focuses on categorizing misinformation by the composition of their replies, we further discard misinformation tweets that have zero replies. In addition, we discard tweets where the poster has limited the set of users who can reply to their tweet, to ensure that all tweets in our dataset have an equal opportunity to be replied to. This information is obtained from the Twitter API.

Finally, our COVID-19 vaccine misinformation tweet dataset consists of 268,990 tweets where each tweet has at least one reply. This is the final set of misinformation tweets that we use.

3.3.2 Counter-misinformation Reply Collection and Classification. For each tweet in our misinformation dataset, we use the Twitter API to crawl all direct replies to the original tweet. In total, we collected a total of 1,991,611 replies to the 268,990 tweets. One misinformation tweet has an average of approximately 7.4 replies. The distribution of the reply count per tweet is shown in Figure 2 in blue.

Building a Counter-reply Classifier: Since it is of high cost to manually annotate all replies, in order to identify all the counter-replies (and non-counter-replies) from this set of 1.9 million replies, we train another text-based classifier to determine if a reply counters the tweet or not. Here, we call this a "counter-reply classifier".

Building on the existing works of the reply classification task [32], we first annotated replies and then built the classifier. Specifically, two students each first annotated 500 randomly-selected pairs of tweets and replies based on the textual contents into 'Countering' or 'Non-Countering' as per the definition provided in Section 3.1. This annotation resulted in an inter-rater agreement score of 0.7033 measured by percent agreement, resulting in 244 responses expressing countering while the remainder were non-countering. Then, after discussing the disagreements and creating the same annotation standard, each annotator labeled another 545 randomly selected pairs of tweets and replies. In total, we get 802 countering replies and 788 non-countering replies in our final annotated counter-reply dataset.

After getting the annotated replies, we utilize the Roberta-base lower-case architecture [37] as the classifier to which the input is the pair of tweets and replies. After the hyperparameter search across batch size and learning rate, the classifier achieves a decent performance with a precision of 0.834, a recall of 0.819, and an F1-score of 0.822, which is sufficient for counter-reply classification on unlabeled replies.

Finally, we classify 690,047 (34.65%) replies as counter-replies, and the remaining 1,301,564 (65.35%) as non-counter-replies. The distribution of the counter-reply count per tweet is shown in Figure 2 in orange. The average number of counter-replies that a misinformation tweet has is 2.57, and the average proportion of all replies of a misinformation tweet that are counter-replies is 0.271.

3.3.3 Misinformation Poster Attribute Collection. For each misinformation tweet, we also collect information of the user who posted the misinformation tweet, which includes date and time of account creation, number of tweets posted, account verification, follower count, and following count. In total, information for 137,929 unique users was retrieved.

Additionally, we collected all the tweets that the user posted in the 7 days leading up to them posting the misinformation tweet; we refer to these tweets as "pre-misinformation" tweets. Only original and quote tweets were retrieved; replies and other retweets were excluded. We pull the same set of attributes as in the misinformation tweet crawling. In total, we retrieved a total of 31,450,114 "pre-misinformation" tweets, with an average of 116.9 "pre-misinformation" tweets per misinformation tweet. Note that these numbers include duplicate tweets if the user had posted two misinformation tweets within 7 days of each other.

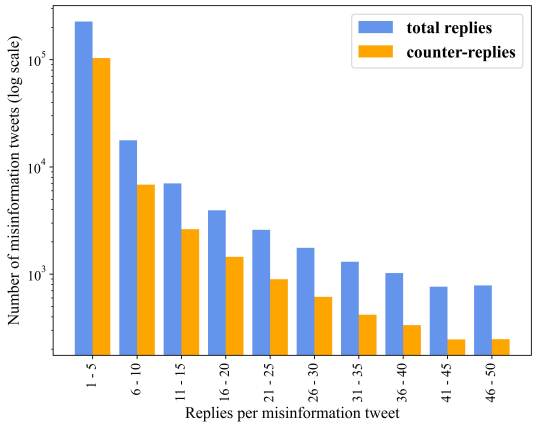


Figure 2: Distributions of the total number of replies (blue) and number of counter-replies (orange) per misinformation tweet, each presented on a log scale.

As a final step, we identify the subset of "pre-misinformation" tweets that are related to the topic of COVID-19 vaccines, as well as those that are also misinformative. We define a "pre-misinformation" tweet belonging to that subset if it contains at least one of the aforementioned six keywords that were used to collect the original Anti-Vax dataset, namely {'vaccine', 'pfizer', 'moderna', 'astrazeneca', 'sputnik', 'sinopharm'}. In total, 1,781,161 (5.71%) of the "pre-misinformation" tweets are labeled as being about COVID-19 vaccines. We then utilize the aforementioned misinformation classifier to identify COVID-19 vaccine misinformation within this subset of "pre- misinformation" tweets, of which 335,458 (18.83%) were classified as misinformative.

4 CHARACTERIZATION OF COUNTER-REPLY

In this section, we analyze the properties of misinformation tweets with respect to the degree to which their misinformation gets countered. In order to do so, we identify the tweets that see a high proportion of their replies being counter-replies, and compare it to the group that see a low proportion of their replies being counter-replies.

To avoid skewing the results due to extreme data points, for this analysis, we do not consider tweets at the two extremes of the "reply count" distribution – specifically, we remove tweets with fewer than three replies, as well as the top 2% of tweets that have the greatest number of replies, following similar tweet filtering procedures in existing research works [2, 4, 69]. This is done to remove dataset noise related to low-engagement tweets, along with outliers associated with the highest engagement tweets. After this

process, we are left with 74,663 misinformation tweets, with reply counts ranging from 3 to 52 (both inclusive).

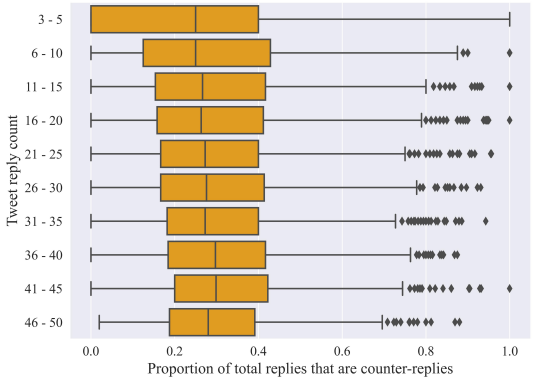


Figure 3: Distribution of proportion of counter-replies for each stratum. Each boxplot represents a stratum, displaying the minimum, maximum, quartiles, and (any) outliers.

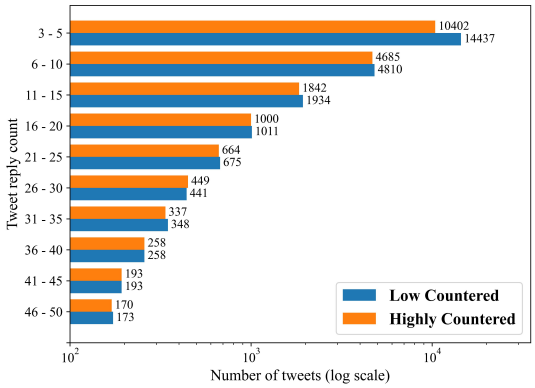


Figure 4: Number of tweets in each of the "Low Countered" (yellow) and "Highly Countered" (red) groups for each stratum, presented on a log scale.

4.1 Stratified Dataset Creation

The linguistic, engagement, and user-level properties of tweets that get a low number of replies are different from those of tweets that receive many replies [9, 10, 38]. Thus, to avoid conflating the factors that lead to receiving a high number of replies with the factors to receive counter-replies, we define and create several strata based on the number of replies that a misinformation tweet receives. Specifically, the strata are defined as follows: *[3, 5], [6, 10], [11, 15], ..., [46, 50]*. Each stratum contains similar misinformation tweets that receive a similar number of replies, with some tweets that get countered and others that do not. We then compare these two groups within each stratum. Figure 3 shows the distribution of counter-reply proportion within each stratum. We observe that, with the exception of tweets with a lower number of replies (that have more tweets with relatively fewer counter-replies), the distribution is similar across reply counts.

Within each stratum, we assign tweets to a "Highly Countered" group if its counter-reply proportion is in the top quartile (also within that stratum), a "Low Countered" group if its counter-reply proportion is in the bottom quartile (within that stratum), or discard it if it does not fall into either of the two groups. Figure 4 shows the distribution of the tweets in the two relevant categories.

Within each stratum, we compare misinformation tweets between the two groups. We identify three types of attributes to perform this comparison along:

(1) **Tweet linguistic attributes**, to analyze the degree to which the tweet falls into meaningful personal, psychological, topical, emotion, and other content-related categories.
(2) **Tweet engagement attributes**, to analyze how and how much the tweet is interacted with among online users.
(3) **Tweet poster attributes**, to analyze the behavior, popularity, and status of the user behind the tweet.

Table 1 displays the full list of attributes we study within each of these categories. We present results in the following subsections.

4.2 Linguistic Attributes of Tweets that are Countered

First, we observe from Figure 5a that on average, "Highly Countered" tweets contain 32.1% higher usage of affective language (words and phrases that appeal more to emotions) than "Low Countered" tweets ($p < 0.05$ for all strata[2]; average Cohen's $d = 0.277$)[3]. This indicates that those who post counter-replies tend to gravitate more towards replying to misinformation that induces a stronger emotional reaction in them. This is consistent with the finding that emotional content gets more attention on social media in existing research works [52]. Further, we find that "Highly Countered" tweets express significantly higher negative sentiment than "Low Countered" tweets across all strata. Figure 5b shows this result for VADER negative sentiment ($p < 0.05$ for all strata; average Cohen's $d = 0.304$); we find similar results for the "negative emotion" dimension of the LIWC lexicon ($p < 0.05$ for all strata; average Cohen's $d = 0.279$). In particular, we find that on average, "Highly Countered" tweets contain 104% more anger-related words than "Low Countered" tweets (see Figure 5c) ($p < 0.01$ for all strata; average Cohen's $d = 0.347$). This implies that the negative tone of misinformation tweets attract more attention [22, 29], and therefore, more counter-replies.

In addition, we measure differences in the degree to which the misinformation tweet expresses politeness and impoliteness. We do this by identifying the sets of linguistic strategies associated with each as presented in [15], and compute the total number of linguistic instances associated with each set to derive the "politeness" and "impoliteness" score, respectively. As shown in Figure 5d, on average, "Highly Countered" tweets utilize 23.1% more strategies

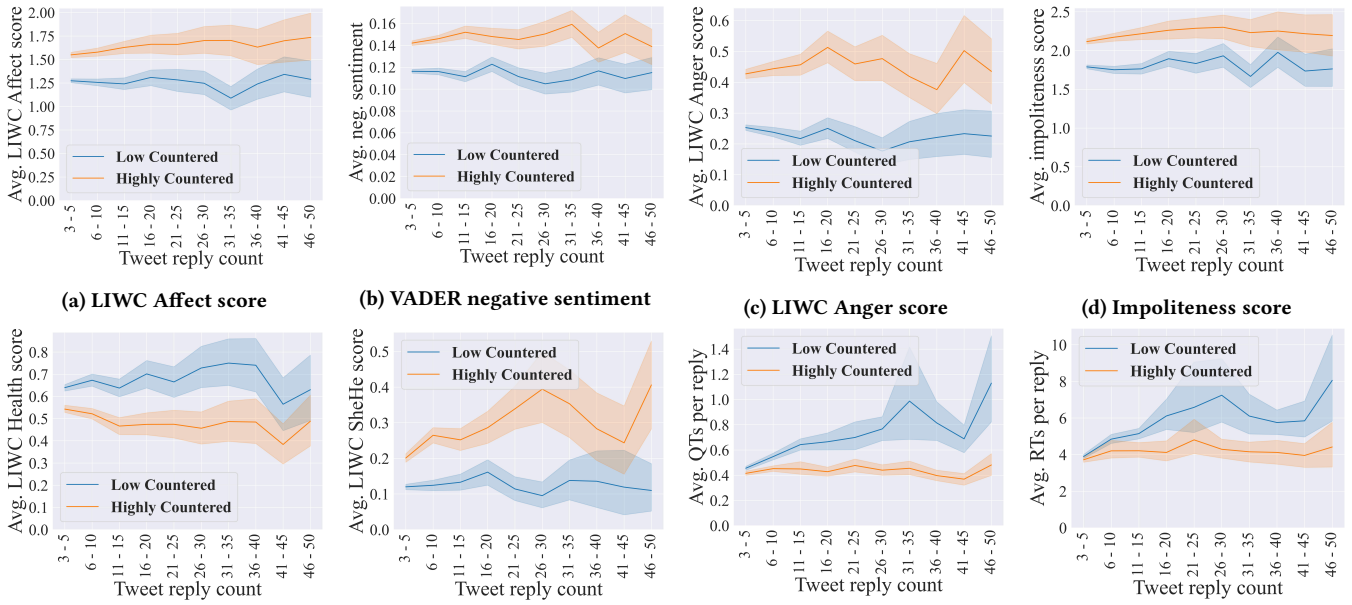
[1]This statistical test was performed using Welch's unequal variances t-test between the upper and lower quartiles (with respect to the proportion of counter-replies) of the data visualized in Figure 3.

[2]all p-values in Sections 4.2, 4.3, and 4.4 are calculated using Welch's unequal variances t-tests.

[3]"average Cohen's d" here (and elsewhere in this paper) refers to the unweighted average of Cohen's d values of each stratum.

Attribute type	List of attributes
Tweet linguistic	• number of words in the tweet*** • VADER [30] positive sentiment, negative sentiment***, and compound sentiment*** of the tweet • Politeness*** and impoliteness*** scores of the tweet, computed as the total number of linguistic strategy instances in the tweet positively and negatively correlated (respectively) with politeness as proposed by [15]. • For each of the 65 (47*** + 18) dimensions of the LIWC [46] 2007 lexicon, the number of words for that dimension.
Tweet engagement	• number of replies***, likes***, retweets*** (RTs), and quote tweets (QTs)*** • number of likes, retweets (RTs), and quote tweets (QTs)***, each divided by the number of replies
Tweet poster	• number of followers, number of users following***, whether the user is verified (1) or not (0)*** • Total number of tweets the user has posted since account creation*** • In the week (7 days) leading up to the misinformation tweet: the average # of tweets posted per day***, the median count of likes*** and retweets*** received on their tweets, the number of tweets the user posted about COVID-19 vaccines***, and the proportion of COVID-19 vaccine tweets that are misinformation.

Table 1: List of linguistic, engagement, and poster attributes considered for the analysis in Section 4. A set of three asterisks(***) next to the attribute indicates a statistical test result of $p < 0.001$.[1]This subset of statistically significant attributes are considered for the predictive tasks in Section 6.



(a) LIWC Affect score **(b) VADER negative sentiment** **(c) LIWC Anger score** **(d) Impoliteness score**

(e) LIWC Health score **(f) LIWC SheHe score** **(g) Fraction of quote tweets (QTs) among all the replies** **(h) Fraction of retweets (RTs) among all the replies**

Figure 5: Means and 95% confidence intervals of the linguistic and engagement attributes of misinformation tweets that get highly countered versus those that do not.

associated with impoliteness than "Low Countered" tweets ($p < 0.05$ for all but one stratum; average Cohen's $d = 0.248$); this finding is consistent with the previous findings involving negative sentiment. Meanwhile, we do not find a significant difference between the groups for strategies associated with politeness, implying that trying to be polite in presenting a misinformation tweet does not significantly impact the chance of being countered.

Next, we find that there exist differences in topical presence between "Highly Countered" and "Low Countered" tweets. Figure 5e shows that on average, "Highly Countered" tweets utilize 28.7% fewer health-related terms than "Low Countered" tweets ($p < 0.05$ for all but one stratum; average Cohen's $d = 0.220$). This suggests that for the average counter-reply poster, the inclusion of more technical medical terminology might pose a barrier for

their willingness or ability to post an effective debunking response. One possible reason is that the inclusion of technical health-related terms can signal authority over the topic and be more convincing to the reader [11, 21].

We also find that "Highly Countered" tweets use 2.5 times more third-person pronouns (e.g., 'he', 'she', 'they', 'them', etc.) than "Low Countered" tweets ($p < 0.05$ for all but one stratum; average Cohen's $d = 0.259$; see Figure 5f).

4.3 Engagement Attributes of Tweets that are Countered

In this subsection, we study the impact of engagement attributes (e.g., likes, retweets, etc.) on whether misinformation gets countered. There are two possibilities: (1) first, misinformation tweets with higher engagement get countered more often because the misinformation gets more attention and therefore, have a higher likelihood of becoming accessible to someone who would counter it; (2) second, misinformation tweets that get countered are less likely to be liked or retweeted by others. We investigate which of the two possibilities hold as per the data.

In addition to the reply count, we compare tweets using the number of likes, retweets (RT), and quotes (QT) they receive. As these methods of engagement on the platform serve a different purpose and have different functionality than the "reply" method, it is worth using these metrics in our cross-group comparison. In order to effectively capture these differences with respect to reply count, we first perform a scaling of these attributes by dividing by the reply count, then performing comparisons of this quotient across the two groups.

Figure 5g shows that on average, "Highly Countered" tweets receive 37.6% fewer QTs relative to replies on average ($p < 0.05$ for all strata). This difference is very small at the lowest stratum (8.9% fewer; Cohen's $d = 0.05$), but is much higher on the highest stratum (57.4% fewer; Cohen's $d = 0.37$). We receive similar results for retweets and likes; on average, "Highly Countered" tweets receive 27.4% fewer retweets relative to replies ($p < 0.05$ for all but one stratum; see Figure 5h) and 25.6% fewer likes relative to replies ($p < 0.05$ for all but 3 strata).

These findings show that the presence of counter-replies on a tweet organically decreases engagement by average users, suggesting that the practice of countering is potentially effective at reducing the spread of misinformation [13, 18, 66].

4.4 User Attributes of Tweet Posters that are Countered

First, we study the impact of the user being verified on Twitter on the tweet getting countered. We find that, on average, the proportion of "Highly Countered" misinformation posters that are verified is 16.8% higher than that for "Low Countered" misinformation posters ($p < 0.05$ in all but 3 strata; average Cohen's $d = 0.143$).

Since the majority of the posters on Twitter are non-verified, we study that set of users next. We compare the attributes of non-verified users in the "Highly Countered" group versus the "Low Countered" group. For the remainder of the attributes, we found none of them to be statistically different across the two groups. Thus, together with the linguistic results presented in Section 4.2, we find that the content of the misinformation tweet is more important in attracting countering than the user who posts the misinformation.

5 INEQUALITY IN SOCIAL CORRECTION

We further investigate the potential inequality in social correction. This can help identify whether certain types of users are less likely to be countered, leading to an increase in disparity. Motivated by existing work [63], we use education level as a key demographic variable to illustrate the potential inequality between different users. Since lack of education and literacy play a crucial role in believing in misinformation [19, 51, 62], it is important to study whether it also impacts correction.

We derived the education level of users by quantifying the readability of their posts using the Automated Readability Index (ARI), which is known to produce an approximate representation of education level in prior works [17, 50, 53]. A higher ARI corresponds to a higher education level. We use the "pre-misinformation" posts of each user (i.e., posts made within the 7 days prior to posting the misinformation tweet) to calculate that user's ARI. Then, for each post, we compute the ARI score [17, 50, 53]. Finally, we compute the average of these scores, and use it as the final ARI value to present the education level of the user. Thus, it should be noted that the ARI score is *not* the education level portrayed in the misinformation tweet, but instead, the education level derived across the *historical* posts of the user who spread misinformation tweets. We randomly sampled 10,000 users who spread misinformation in our dataset to illustrate the inequality phenomenon.

As shown in Figure 6, we find that misinformation posts made by users with lower education levels have a higher likelihood of getting corrected. There is a systematically negative trend with an increase in the user's (perceived) education level. This highlights a need to pay attention to misinformation spread by users who portray a higher education level, since ordinary users are less likely to correct them.

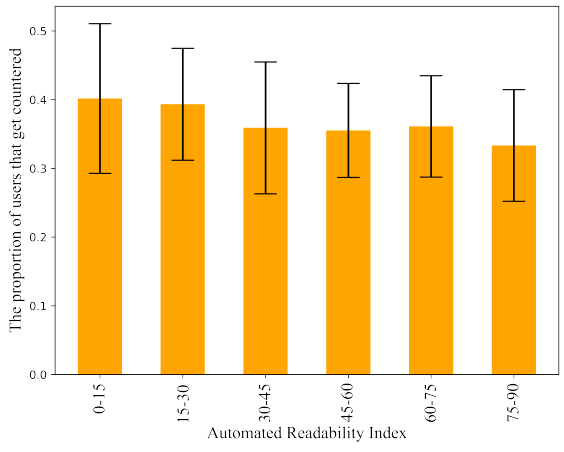


Figure 6: Comparison of user communities with different education levels. As shown, users with lower education levels will have higher possibilities of getting countered when sending misinformation tweets.

6 SOCIAL CORRECTION PREDICTION

In this section, we aim to answer two research questions:

- RQ1: Given a misinformation tweet, can we predict whether it will be countered or not in the future?
- RQ2: Given a misinformation tweet that will be countered in the future, can we predict whether it will be countered with fewer or more counter-replies?

Both RQs are important to address for the combating of future misinformation. By being able to effectively predict future interactions surrounding misinformation tweets, we can better identify sets of online interactions where misinformation is being organically countered, along with those where additional countering needs to be performed. Answering RQ1 can identify sets of misinformation posts where other users may take the initiative in posting a counter-reply, while answering RQ2 can predict the intensity or magnitude of countering.

6.1 Dataset

For both the research questions, we use the aforementioned dataset as we described in Section 4.

For RQ1, we divide the dataset into two sets of misinformation tweets: (1) misinformation tweets that have replies but none of them are counter-replies; (2) misinformation tweets that have at least one counter-reply. The sizes of these sets are 17,787 and 55,136, respectively.

For RQ2, we divide the misinformation tweets into two groups: one with a low proportion of counter-replies, and another group with a high proportion of counter-replies. Similar to the stratified setup in Section 4.1, we use the proportion of counter-replies as an indicator of membership for the two groups. The bottom 25% of posts with respect to their countering proportion are assigned to the low countered group. On the other hand, the top 25% posts with the highest proportion of countering replies are assigned to the highly countered group. The sizes of these sets are 14,274 and 15,224, respectively.

6.2 Experimental setup

Using each of these datasets, we follow similar approaches in tweet prediction tasks [40, 70] to address both RQ1 and RQ2. We aim to build a binary classifier for each of RQ1 and RQ2, using the label definitions described above. For both RQs, we use the same set of features. We begin with the set of attributes listed in Table 1 with $p < 0.001$ and have non-null values for all datapoints; there are 63 such attributes (53 linguistic, 5 engagement, 5 poster). As shown in the existing tweet prediction task [40], the semantic information from textual embedding benefits the prediction task. Thus, we also generate the embedding vector for each tweet using RoBERTa [37], which results in a 768-dimensional feature vector. Finally, we concatenate the above feature vectors to form a tweet feature vector to comprehensively represent the tweet and use it for classification.

Classifier: Following similar tweet classification tasks [26, 40], we deploy widely-used conventional machine learning classifiers including Logistic Regression, XGBoost, and a Feed-forward Neural Network with a single hidden layer, using the feature vector as input. During the experiment, 10-fold cross-validation is deployed, and we report precision, recall, and F-1 score as the performance metrics.

6.3 Classifier Performance

Method	Precision	Recall	F-1 score
Logistic Regression	0.801	0.929	0.860
XGBoost	0.803	0.908	0.852
Neural Network	0.804	0.914	0.855

Table 2: RQ1: Classifier performance of whether tweets will get countered or not.

In Table 2, we report the classification result for RQ1. As we can see, all three models are able to achieve good performance on the task. The logistic regression achieves the best performance in terms of precision, recall, and F-1 score; this result is also found in other similar tweet classification tasks [40]. This high performance grants the ability to effectively predict whether a tweet will be countered or not, enabling fact-checkers and social media platforms to prioritize countering tweets identified as less likely to be countered organically.

Method	Precision	Recall	F1 score
Logistic Regression	0.731	0.742	0.737
XGBoost	0.841	0.756	0.796
Neural Network	0.848	0.759	0.801

Table 3: RQ2: Classification performance of whether tweets will be highly countered versus that will be low countered.

For RQ2, the classification result is shown in Table 3. As we can see, the model performance is still reasonably acceptable, but is worse compared to RQ1. This decrease in performance may imply that the task to identify the *intensity* of countering tweets is not only more difficult, but also distinct from the task to identify *whether* a tweet will be countered. In other words, the phenomenon of posting of the first counter-reply is easier to forecast than that of the posting of additional counter-replies given that at least one has already been posted.

7 DISCUSSION AND CONCLUSION

In this paper, we studied the tweet and user-level properties of misinformation tweets that get countered versus those that do not. The in-depth analysis shows that misinformation tweets expressing negative emotion, strong emotion, third-person pronouns, and strategies associated with impoliteness are more likely to result in more countering replies from users. Our result also shows that tweets that get countered have a higher amount of reply engagement in proportion to like, retweet, and quote tweet engagement. Moreover, we develop well-performing classifiers to predict whether a misinformation tweet will be countered or not, and if so, to what degree they will be countered (i.e. the proportion of its replies that end up being counter-replies).

Given the statistical significance of our analysis and the high performance of our classifiers, we demonstrate that it is possible

to identify tweets that are more or less likely to get countered. In particular, nearly all of these attributes (tweet linguistic attributes and user attributes) are readily available as soon as the tweet is posted, allowing for the quantity of future counter-misinformation (or the lack thereof) to be reasonably forecast. This can have major implications in times of breaking news or other such events in which large quantities of (mis-)information are posted to online platforms at a rapid rate; in conjunction with state-of-the-art misinformation detection approaches, the counter-reply prediction approach presented in this paper can be used to identify tweets that are less likely to be countered, possibly necessitating additional platform-level approaches to control the spread of misinformation for these tweets. One of these approaches may be adding or increasing interventions to draw attention towards accuracy, an approach that has been shown to be effective in discouraging users from spreading misinformation [47].

A limitation of this work that it focuses on only one platform: Twitter. On other online platforms, different mechanisms of post and user engagement, as well as information exchange, may be present [41], possibly influencing the types of misinformation tweets and posters users will choose to counter. Another limitation is that it studies only one topic (COVID-19 vaccines), which has become one of the most widely discussed topics in our society due to the universal effects of the COVID-19 pandemic. On misinformation-related topics that might be more obscure or less widely discussed (e.g. flat earth theories), it could be possible that the more specific demographics of misinformation and/or counter-reply posters may affect the ways in which they interact. In addition, we only study text in the English language; the dynamics and discussion in other languages and other modalities (images, videos) may differ [64].

For future work, similar analysis can be performed on the user network surrounding the misinformation poster and counter-reply poster (e.g. their followers and those they follow, how much misinformation these accounts spread, etc.) in order to assess if there are any network-related attributes that may increase the likelihood of counter-replies. In addition, given that we can reasonably determine which tweets will and will not be countered, it would also be valuable to perform user studies or field studies to evaluate if certain characteristics about online encounters with misinformation can increase (or decrease) the likelihood of a user posting a counter-reply. Also, while we explore it in Section 5, further studies can be done to understand the inequities surrounding counter-reply targets along additional demographic, social, political, and/or geographic dimensions; this can allow further exploration of the greater societal implications surrounding counter-misinformation.

ACKNOWLEDGMENTS This research/material is based upon work supported in part by NSF grants CNS-2154118, IIS-2027689, ITE-2137724, ITE-2230692, CNS-2239879, and funding from Microsoft, Google, and Adobe Inc. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the position or policy of NSF and no official endorsement should be inferred. We thank the CLAWS research group members for their help on the project.

REFERENCES

[1] Jennifer Abbasi. 2022. Widespread misinformation about infertility continues to create COVID-19 vaccine hesitancy. *JAMA* 327, 11 (2022), 1013–1015.

[2] Ana Aleksandric, Sayak Saha Roy, and Shirin Nilizadeh. 2022. Twitter Users' Behavioral Response to Toxic Replies. *arXiv preprint arXiv:2210.13420* (2022).

[3] Jennifer Allen, Cameron Martel, and David G Rand. 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 Conference on Human Factors in Computing Systems*. 1–19.

[4] David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. In *Tenth international AAAI conference on web and social media*.

[5] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 20 (Nov. 2018), 27 pages. https://doi.org/10.1145/3274289

[6] Philip Ball and Amy Maxmen. 2020. The epic battle against coronavirus misinformation and conspiracy theories. https://www.nature.com/articles/d41586-020-01452-z.

[7] Leticia Bode and Emily K Vraga. 2018. See something, say something: Correction of global health misinformation on social media. *Health communication* 33, 9 (2018), 1131–1140.

[8] Leticia Bode and Emily K. Vraga. 2021. Correction Experiences on Social Media During COVID-19. *Social Media + Society* 7 (4 2021), 205630512110088. Issue 2. https://doi.org/10.1177/20563051211008829

[9] Axel Bruns and Stefan Stieglitz. 2012. Quantitative approaches to comparing communication patterns on Twitter. *Journal of technology in human services* 30, 3-4 (2012), 160–185.

[10] Axel Bruns and Stefan Stieglitz. 2013. Towards more systematic Twitter analysis: metrics for tweeting activities. *International journal of social research methodology* 16, 2 (2013), 91–108.

[11] Michelle M Buehl, Patricia A Alexander, P Karen Murphy, and Christopher T Sperl. 2001. Profiling persuasion: The role of beliefs, knowledge, and interest in the processing of persuasive texts that vary by argument structure. *Journal of Literacy Research* 33, 2 (2001), 269–301.

[12] Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science* 28, 11 (2017), 1531–1546.

[13] Jonas Colliander. 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 (2019), 202–215.

[14] Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885* (2020).

[15] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Daniel Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Annual Meeting of the Association for Computational Linguistics*.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[17] Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 313–319.

[18] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[19] Neophytos Georgiou, Paul Delfabbro, and Ryan Balzan. 2020. COVID-19-related conspiracy beliefs and their relationship with perceived stress and pre-existing conspiracy beliefs. *Personality and individual differences* 166 (2020), 110201.

[20] Andrew Guess and Alexander Coppock. 2020. Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science* 50, 4 (2020), 1497–1515.

[21] Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 1214–1223.

[22] Martin Haselmayer, Thomas M Meyer, and Markus Wagner. 2019. Fighting for attention: Media coverage of negative campaign messages. *Party Politics* 25, 3 (2019), 412–423.

[23] K. Hayawi, S. Shahriar, M.A. Serhani, I. Taleb, and S.S. Mathew. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health* 203 (2022), 23–30. https://doi.org/10.1016/j.puhe.2021.11.022

[24] Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public health* 203 (2022), 23–30.

[25] Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19

Vaccine Misinformation. In *Proceedings of the ACM Web Conference 2023*.

[26] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 90–94.

[27] Julian PT Higgins, Ian R White, and Judith Anzures-Cabrera. 2008. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. *Statistics in medicine* 27, 29 (2008), 6072–6092.

[28] Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. (2020).

[29] Yu-Xia Huang and Yue-Jia Luo. 2007. Attention shortage resistance of negative stimuli in an implicit emotional task. *Neuroscience Letters* 412, 2 (2007), 134–138.

[30] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

[31] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 245–251.

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

[33] Jisu Kim, Jihwan Aum, SangEun Lee, Yeonju Jang, Eunil Park, and Daejin Choi. 2021. FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics and Informatics* 64 (2021), 101688.

[34] Neta Kligler-Vilenchik. 2022. Collective social correction: addressing misinformation through group practices of information verification on WhatsApp. *Digital Journalism* 10, 2 (2022), 300–318.

[35] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559* (2018).

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

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[38] Kazuyuki Matsumoto, Yuta Hada, Minoru Yoshida, and Kenji Kita. 2019. Analysis of Reply-Tweets for Buzz Tweet Detection. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), Hakodate, Japan*. 13–15.

[39] Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791* (2020).

[40] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The role of the crowd in countering misinformation: A case study of the COVID-19 infodemic. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 748–757.

[41] Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. 2022. Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 651–662.

[42] Kunihiro Miyazaki, Takayuki Uchiba, Kenji Tanaka, Jisun An, Haewoon Kwak, and Kazutoshi Sasahara. 2022. " This is Fake News": Characterizing the Spontaneous Debunking from Twitter Users to COVID-19 False Information. *arXiv preprint arXiv:2203.14242* (2022).

[43] Yida Mu, Pu Niu, and Nikolaos Aletras. 2022. Identifying and Characterizing Active Citizens who Refute Misinformation in Social Media. *14th ACM Web Science Conference 2022* (2022).

[44] Rohit Mujumdar and Srijan Kumar. 2021. HawkEye: a robust reputation system for community-based counter-misinformation. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 188–192.

[45] Goran Muric, Yusong Wu, Emilio Ferrara, et al. 2021. COVID-19 vaccine hesitancy on social media: building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR public health and surveillance* 7, 11 (2021), e30642.

[46] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[47] Gordon Pennycook, Jonathon Mcphetres, Zhang Yunhao, and Jackson Lu. 2020. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science* 31 (06 2020), 770–780. https://doi.org/10.1177/0956797620939054

[48] Cefas Garcia Pereira and Humberto Torres Marques-Neto. 2022. Characterizing the impact of fact-checking on the COVID-19 misinformation combat. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 1789–1796.

[49] Francesco Pierri, Brea L Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. 2022. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific reports* 12, 1 (2022), 1–7.

[50] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on Twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 97–106.

[51] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* 7, 10 (2020), 201199.

[52] Melanie Schreiner, Thomas Fischer, and Rene Riedl. 2021. Impact of content characteristics and emotion on behavioral engagement in social media: literature review and research agenda. *Electronic Commerce Research* 21, 2 (2021), 329–345.

[53] RJ Senter and Edgar A Smith. 1967. *Automated readability index*. Technical Report. Cincinnati Univ OH.

[54] Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. 2022. 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*, Vol. 16. 896–907.

[55] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. *Online social networks and media* 22 (2021), 100104.

[56] Karishma Sharma, Yizhou Zhang, and Yan Liu. 2022. COVID-19 Vaccine Misinformation Campaigns and Social Media Narratives. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 920–931.

[57] Jieun Shin and Kjerstin Thorson. 2017. Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media. *Journal of Communication* 67, 2 (02 2017), 233–255. https://doi.org/10.1111/jcom.12284 arXiv:https://academic.oup.com/joc/article-pdf/67/2/233/22321279/jjnlcom0233.pdf

[58] Craig Silverman. 2015. Lies, damn lies, and viral content: How news websites spread (and Debunk) online rumors, unverified claims and misinformation. *Tow Center for Digital Journalism* 168, 4 (2015), 134–140.

[59] Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed news* 16 (2016).

[60] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. *IConference 2014 Proceedings* (2014).

[61] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of applied research in memory and cognition* 9, 3 (2020), 286–299.

[62] Jan-Willem van Prooijen. 2017. Why education predicts decreased belief in conspiracy theories. *Applied cognitive psychology* 31, 1 (2017), 50–58.

[63] Gaurav Verma, Ankur Bhardwaj, Talayeh Aledavood, Munmun De Choudhury, and Srijan Kumar. 2022. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports* 12, 1 (2022), 1–9.

[64] Gaurav Verma, Rohit Mujumdar, Zijie J Wang, Munmun De Choudhury, and Srijan Kumar. 2022. Overcoming Language Disparity in Online Content Classification with Multimodal Learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1040–1051.

[65] Nguyen Vo and Kyumin Lee. 2019. Learning from fact-checkers: analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–344.

[66] Senuri Wijenayake, Danula Hettiachchi, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2020. Effect of conformity on perceived trustworthiness of news in social media. *IEEE Internet Computing* 25, 1 (2020), 12–19.

[67] Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41, 1 (2019), 135–163.

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

[69] Nicholas Jing Yuan, Yuan Zhong, Fuzheng Zhang, Xing Xie, Chin-Yew Lin, and Yong Rui. 2016. Who will reply to/retweet this tweet? The dynamics of intimacy from online social interactions. In *Proceedings of the ninth ACM international conference on web search and data mining*. 3–12.

[70] Hamada M Zahera, Ibrahim A Elgendy, Richa Jalota, and Mohamed Ahmed Sherif. 2019. Fine-tuned BERT Model for Multi-Label Tweets Classification.. In *TREC*. 1–7.