# Corrective or Backfire: Characterizing and Predicting User Response to Social Correction

Bing He, Yingchen Ma, Mustaque Ahamad, Srijan Kumar
Georgia Institute of Technology
Atlanta, Georgia, USA
bhe46@gatech.edu,yma473@gatech.edu,mustaq@cc.gatech.edu,srijan@gatech.edu

## ABSTRACT

Online misinformation poses a global risk with harmful implications for society. Ordinary social media users are known to actively *reply* to misinformation posts with counter-misinformation messages, which is shown to be effective in containing the spread of misinformation. Such a practice is defined as "*social correction*". Nevertheless, it remains unknown how users respond to social correction in real-world scenarios, especially, will it have a corrective or backfire effect on users. Investigating this research question is pivotal for developing and refining strategies that maximize the efficacy of social correction initiatives.

To fill this gap, we conduct an in-depth study to characterize and predict the user response to social correction in a data-driven manner through the lens of X (Formerly Twitter), where the user response is instantiated as the reply that is written toward a counter-misinformation message. Particularly, we first create a novel dataset with 55, 549 triples of misinformation tweets, counter-misinformation replies, and responses to counter-misinformation replies, and then curate a taxonomy to illustrate different kinds of user responses. Next, fine-grained statistical analysis of reply linguistic and engagement features as well as repliers' user attributes is conducted to illustrate the characteristics that are significant in determining whether a reply will have a corrective or backfire effect. Finally, we build a user response prediction model to identify whether a social correction will be corrective, neutral, or have a backfire effect, which achieves a promising F1 score of 0.816. Our work enables stakeholders to monitor and predict user responses effectively, thus guiding the use of social correction to maximize their corrective impact and minimize backfire effects. The code and data is accessible on https://github.com/claws-lab/response-to-social-correction.

## CCS CONCEPTS

• **Information systems** → Social networks.

## KEYWORDS

Misinformation, Counter-misinformation, Social Correction

*WEBSCI '24, May 21–24, 2024, Stuttgart, Germany*
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0334-8/24/05
https://doi.org/10.1145/3614419.3644004

**ACM Reference Format:**
Bing He, Yingchen Ma, Mustaque Ahamad, Srijan Kumar. 2024. Corrective or Backfire: Characterizing and Predicting User Response to Social Correction. In *ACM Web Science Conference (WEBSCI '24), May 21–24, 2024, Stuttgart, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3614419.3644004

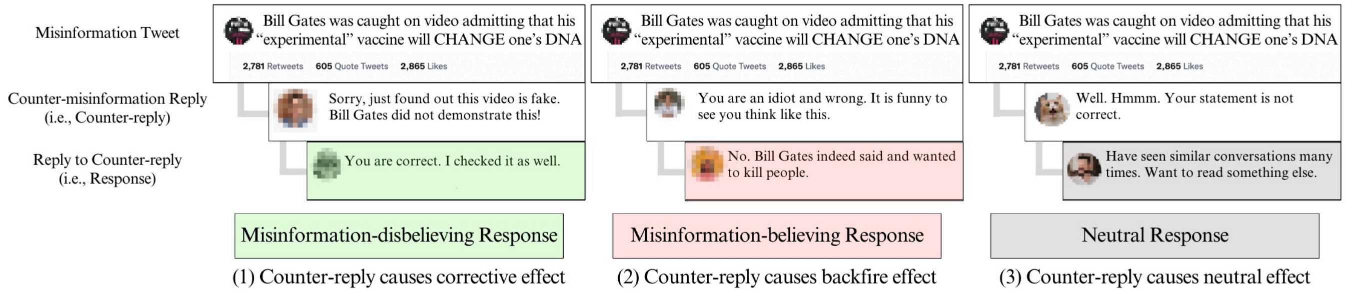
## 1 INTRODUCTION

Online misinformation undermines public health by diminishing trust in vaccines and health policies [4, 33, 46], and has been linked to reduced COVID-19 vaccine uptake [37]. Its impact also extends to inciting violence [3, 51], and negatively affecting well-being [57]. This situation is exacerbated because misinformation typically spreads more rapidly and widely than factual information on online social media platforms [33, 60], making it imperative to curb the spread of misinformation [12, 22, 34, 35, 39, 71, 78].

To combat misinformation, professional fact-checkers and journalists provide valuable objective fact-checks to debunk misinformation [59]. However, their engagement with users is limited [39]. In contrast, ordinary social media users play a proactive role in combating misinformation through their active engagement including their replies, comments, and posts that counter misinformation posted by others [7, 39, 51, 55, 58, 77]. It finally complements the efforts of professionals [2, 26, 31], even accounting for 96% of online counter-misinformation messages [39].

Significantly, recent studies underscore the "*social correction*" [37, 41] - the practice where ordinary users combat misinformation claims in a *conversational* manner by their direct counter misinformation *replies* to misinformation posts - which has shown to be as effective as professional correction, curbing misinformation spread across diverse topics, platforms, and demographics [6–8, 15, 20, 37, 50, 61–64, 64, 65, 72]. One example of social correction is shown in Figure 1.

Nevertheless, little is known about the real-world user response toward social correction. Understanding such responses is beneficial because i) They serve as a critical signal to indicate the impact of social correction in real-world scenarios. If some social corrections are revealed to have corrective effects (e.g., users disbelieve in misinformation) [13], then additional participants can be encouraged to provide reinforcements; ii) Instead, If certain social corrections are found to increase users' beliefs in misinformation (e.g., backfire) [53], targeted efforts can be directed toward improving them. Such instances can be escalated and prioritized for interventions by professionals or social media platforms; iii) Responses can also indicate whether users are entrenched in (counter-)misinformation echo chambers [17], where their beliefs are reinforced by similar viewpoints, or if there is a cross-pollination of ideas. This contributes to understanding polarization around certain topics.


Misinformation Tweet
Counter-misinformation Reply (i.e., Counter-reply)
Reply to Counter-reply (i.e., Response)
Misinformation-disbelieving Response
Misinformation-believing Response
Neutral Response
(1) Counter-reply causes corrective effect
(2) Counter-reply causes backfire effect
(3) Counter-reply causes neutral effect

Bill Gates was caught on video admitting that his "experimental" vaccine will CHANGE one's DNA

2,781 Retweets 605 Quote Tweets 2,865 Likes

Sorry, just found out this video is fake. Bill Gates did not demonstrate this!

You are correct. I checked it as well.

Bill Gates was caught on video admitting that his "experimental" vaccine will CHANGE one's DNA

2,781 Retweets 605 Quote Tweets 2,865 Likes

You are an idiot and wrong. It is funny to see you think like this.

No. Bill Gates indeed said and wanted to kill people.

Bill Gates was caught on video admitting that his "experimental" vaccine will CHANGE one's DNA

2,781 Retweets 605 Quote Tweets 2,865 Likes

Well. Hmmm. Your statement is not correct.

Have seen similar conversations many times. Want to read something else.

**Figure 1: Examples of user responses to social correction. Here, the social correction is the counter-misinformation *reply* posted by ordinary users (the second row), and the user response is the *reply* to the counter-misinformation reply (the third row).**

Despite its advantages, characterizing and predicting social correction is challenging because of multiple reasons. First, current research predominantly utilizes simulated experiments or user studies [9, 32, 55], approaches that may not accurately mirror real-world scenarios. Second, relevant research works and datasets [70, 76] do not consist of conversational-style narratives with tripled misinformation posts, counter-replies, and responses, as shown in Figure 1. The analysis of these conversations can reveal the complex interactions among misinformation spreaders, those who counter-reply, and the responders to these counter-replies. It plays a vital role in demonstrating the organic processes of both correcting and exacerbating misinformation. Third, related research works do not conduct fine-grained investigation of user responses. The traditional four-class stance [70] or two-class sentiment [76] categorization of user responses only provides a shallow classification of user responses. A more comprehensive taxonomy of user actions behind their responses is needed to provide a better understanding of how users respond differently to social correction.

To address these challenges, we first curate conversational-style user response datasets to social correction on Twitter[1] and create the taxonomy of these user responses. Additionally, we conduct a statistical analysis of linguistic-, engagement-, and poster-level characteristics of counter-replies to examine user responses to social correction at a fine-grained level. Finally, we create a prediction model to forecast whether a counter-reply will have a corrective, backfire, or neutral effect on users. In sum, the contributions of the paper are summarized as follows:

- We curate a novel large-scale dataset that contains 1,523,849 misinformation tweets, 254,779 counter-misinformation replies, and 55,549 responses, along with a hand-annotated dataset of misinformation tweets, counter-replies, and counter-replies. Concurrently, we build a taxonomy of user responses to demonstrate different kinds of responses to social correction.
- We perform a fine-grained analysis of the linguistic, engagement, and poster-level characteristics of counter-replies that have a corrective or backfire effect. Our analysis reveals several salient features of counter-replies that are more common in corrective replies (e.g., politeness (4.678%), evidence (8.137%), and positiveness (5.409%)) than in backfire replies.
- We create a user response prediction model to identify whether a counter-reply will be a corrective, backfire, or neutral reply. The model achieves a promising predictive performance with an F1 score of 0.816.

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

## 2 RELATED WORKS

### 2.1 Social Correction on Social Media Platforms

Misinformation causes harms to the society [3, 4, 33, 37, 46, 51]. Social correction by ordinary users (i.e., the crowds) is a crucial approach to combating misinformation [37, 41] due to its accessibility and scalability to complement professional fact-checkers [39]. It has also been shown to be effective by conducted interviews [9, 32, 55], surveys [32, 56], and in-lab experiments [55] in curbing misinformation spread [15, 20, 72] across topics [6–8, 62–64], platforms, and demographics [61, 64, 65]. To facilitate broad social correction among users, Twitter launched the Community Note system [2] [48], where users can report and flag potential misinformation tweets. These works and practices demonstrate that social correction works. However, user response to social correction is less studied in the existing research works despite its benefits in understanding the efficacy of social correction. More critically, current works rely on simulated experiments and user studies [55] and do not study datasets that represent real-world scenarios. To bridge this research gap, we conduct a large-scale data-driven analysis. We specifically focus on examining user replies to counter-misinformation replies on Twitter, thereby providing insights grounded in real-world response behavior.

### 2.2 User Response to Misinformation Correction

To correct misinformation, ordinary users can publish standalone counter-misinformation posts on social media platforms [39]. User responses to this kind of correction have been investigated [69, 70, 76]. For instance, Wang et al. [70] analyze the comments on fake news rebuttal posts through the expressed stance in them. They find that information readability and argument quality improve the acceptance of misinformation rebuttal. They also uncover

[1] Twitter was renamed as "X" in July 2023. We continue to refer to the platform as "Twitter" for illustration.

[2] Community Note is previously known as Birdwatch.

that citing evidence helps [69]. Zhang et al. [76] similarly investigate the sentiment in comments that respond to fact-checking posts. But, all these posts are from official fact-checking organization accounts [76], which is different from our setting of ordinary users. Additionally, none of these corrections occur in a conversational manner like our focus of social correction that has more engagement and visibility between misinformation spreaders and those who counter-reply [37]. These existing conventional four-class stance [70] or two-class sentiment [76] studies only provide coarse-grained analysis of user responses.

Some researchers examine another type of misinformation correction - the warning labels posted around the misinformation posts. For example, Chuai et al. [14] focus on labels as well as the associated fact-check text provided voluntarily by users within Twitter's Community Note system [48]. Different from our response analysis, they only focus on the volume of retweets and likes of the fact-checked tweet. However, retweets and likes are all non-negative signals and are unable to comprehensively capture the user response, especially, the negative responses. In addition, users provide inputs within the Community Note system only, which is restricted (e.g., users cannot write responses to the fact-checking text and labels on the Twitter platform) and does not reflect the larger dynamics of information flow on Twitter.

## 2.3 Backfire and Corrective Effects of Misinformation Correction

When misinformation is debunked, it may have a backfire effect, i.e., users viewing the counter-misinformation post or misinformation spreaders potentially increase their belief in the misinformation due to observing the correction. This has been debated for a long time [34, 43]. Even if some researchers find the backfire effect among particular groups [49] and within certain time frames [45], many studies have failed to replicate the backfire effect [52, 68]. On the other hand, corrective effects, i.e., the audience or the misinformation spreaders instead decrease their belief in misinformation after viewing the counter-misinformation, have been identified by existing research works [13, 20, 47, 50, 66, 67]. Nevertheless, the existing studies of backfire and corrective effects usually leverage simulated experiments to examine their hypothesis about backfire and corrective effects while neglecting real-world scenarios, especially the situations where misinformation is corrected by real-world ordinary users rather than professionals or bot accounts. To fill this gap, we examine these effects through real-world user replies to counter-misinformation posts in a data-driven manner. Since this user response information can indicate the effects of certain textual properties in counter-replies, our work can lead to a better understanding of the impacts of social correction behavior, especially, comprehending the counter-replies that are corrective or backfire.

# 3 DATASET

## 3.1 Definition

### 3.1.1 *Misinformation Tweet*.

We deploy a broad definition of misinformation which includes inaccuracies, falsehoods, rumors, or misleading leaps of logic [74]. Based on the existing work [23, 37], we focus on misinformation related to the COVID-19 vaccine due to its broad impact around the world during the COVID-19 pandemic. Particularly, the misinformative claims include "the vaccine changes genes", "the vaccine leads to infertility", "the vaccine is created by Bill Gates to kill people", and "the vaccine consists of microchips to control people"; these misinformation topics are widely studied by existing research works due to their popularity [1, 23, 25]. The misinformation tweet is represented as $m$.

### 3.1.2 *Counter-misinformation Reply (i.e., Counter-reply)*.

Inspired by existing research works on social correction [37, 41], a direct reply to a misinformation tweet $m$ is considered as a counter-misinformation reply (i.e., counter-reply as shown in Figure 1 and denoted as $c$), if it attempts to counter the misinformation tweet. Particularly, building on existing research works that identify and analyze the text that is countering, debunking, disbelieving, or disagreeing with misinformation [28, 30, 37, 39, 41], a counter-reply is a reply that explicitly or implicitly refutes the misinformation post ("the tweet is wrong. it is misinformation"), targets the tweet poster ("you are born to speak nothing but lies"), or highlights the falsehood ("the COVID-19 vaccine does not change DNA").

### 3.1.3 *Reply to Counter-reply (i.e., Response)*.

On Twitter, users can respond to a counter-reply via a direct reply to it, as shown in Figure 1. These responses denote the responder's stance toward misinformation, serving as a crucial signal to study the impact of counter-reply. Following existing work on similar stance analysis [30, 69, 70], we can group responses into three categories, as shown in Figure 1:

- Misinformation-disbelieving responses: responses disbelieve in misinformation or believe in counter-misinformation (e.g., "You are correct. I checked it as well.");
- Misinformation-believing responses: responses believe in misinformation or disbelieve in counter-misinformation (e.g., "No, Bill Gates indeed said and wanted to kill people");
- Neutral responses: Responses neither believe nor disbelieve in misinformation, lacking sufficient information for judgment (e.g., " Have seen similar conversation many times. Want to read something else.").

## 3.2 Task Objective

Given the set $\mathcal{M}$ of misinformation posts regarding the COVID-19 vaccine, each misinformation post $m \in \mathcal{M}$ has a set of $n$ counter-replies $c = [c_1, c_2, ..., c_n]$ posted in direct reply to $m$. Our goal is to build a classifier $\mathcal{F}$ such that it can output a label $\mathcal{F}(c_i), i \in \{1, 2, ..., n\}$, which indicates whether the counter-reply will have a corrective, backfire, or neutral effect, i.e., the counter-reply will have at least one misinformation-disbelieving response but no misinformation-believing responses (corrective), at least one misinformation-believing response but no misinformation-disbelieving response (backfire), or only neutral responses (neutral)? [3]

## 3.3 Dataset Curation

### 3.3.1 *Misinformation Tweet Collection and Classification*.

In our study, we followed an existing approach [37] and used the

[3] Note that we do not emphasize the case where a counter-reply has both misinformation-believing and misinformation-disbelieving responses, which can be worth exploring in future studies, because (1) it has the lowest volume, accounting for only 0.93% of all 254, 779 counter-replies in our dataset, as shown in Section 4.2; and (2) similar existing research works also do not emphasize this kind of dual labels [27].

Anti-Vax dataset from Hayawi et al. [23], containing around 15.4 million English tweets about COVID-19 vaccines, collected between December 1, 2020, and July 31, 2021. These tweets, which exclude retweets, replies, and quotes, were filtered to include key vaccine-related terms (e.g., “vaccine”, “pfizer”, and “moderna”). From the original set, 14,123,209 tweets are retrievable via the Twitter API while the remaining 1.3 million tweets are unavailable due to deletion by users or Twitter.

To identify misinformation tweets, we followed the definition outlined in Section 3.1.1 and the current approach by Hayawi et al. [23]. Initially, 13, 432 annotated tweets (4,836 misinformation, and 8,596 non-misinformation) were extracted from Hayawi et al. [23]. Using these tweets, we trained a BERT-based text classifier [18], achieving precision, recall, and F-1 score of 0.972, 0.979, and 0.975, respectively, denoting a satisfactory performance for the misinformation classification task.

Applying this classifier to the full dataset, we identified 1, 523, 849 misinformation and 12, 599, 360 non-misinformation tweets. However, since we focus on replies to misinformation tweets and responses to these replies, we only keep misinformation tweets that contain this information, resulting in 44, 557 misinformation tweets.

*3.3.2* ***Counter-reply Collection and Classification.*** For each misinformation tweet, we use the Twitter API to crawl all direct replies to the original tweet. In total, we collect a total of 707, 529 replies to the 44, 557 tweets. One misinformation tweet has an average of approximately 16 replies.

To identify counter-replies, we follow the definition of counter-reply in Section 3.1.2 and build on existing works of counter-reply classification [25, 37]. Particularly, we first crawl a combined 2, 479 annotated replies (1, 425 counter-replies, and 1, 054 non-counter-replies) from [25, 37]. Next, we train a RoBERTa-based lower-case counter-reply classifier [37] attaining precision, recall, and F1 score of 0.801, 0.913, and 0.858, respectively, which is sufficient for counter-reply classification. Finally, we identify 254, 855 replies as counter-replies, and the remaining as non-counter-replies.

*Counter-reply Poster Attribute Collection.* For each counter-reply, we also collect information of the user who posted the counter-misinformation reply, which contains the date and time of account creation, the number of tweets posted, account verification, follower count, and following count. In total, information for 251, 017 unique users was retrieved.

*3.3.3* ***Response Collection and Classification.*** For each counter-reply, we use the Twitter API to crawl all direct replies to counter-replies. In total, we collected a total of 55, 549 replies to 34, 765 counter-replies that have responses. Because it is labor-intensive to manually annotate all 55, 549 responses, we instead aim to train a text-based classifier for annotation. Following the existing works of building tweet text classifiers [30], we first annotate responses and then train the classifier. Particularly, two students annotated 601 randomly selected responses into “misinformation-believing”, “misinformation-disbelieving”, and “neutral” as per the definition in Section 3.1.3. Such annotation results in an inter-rater agreement score of 0.7970. After two students discuss the data points that are labeled differently and reach a consensus, we finally have 213 misinformation-disbelieving responses, 218 misinformation-believing responses, and 170 neutral responses. We then fine-tune a RoBERTa-based classifier, which unfortunately has an under-satisfactory performance in precision, recall, and F1 score of 0.545, 0.526, and 0.511. The potential reason is that one data point consists of three entries (i.e., misinformation tweet, counter-reply, and response) and there are complex inter-relationships between them. This requires a profound understanding of one data point, thus making the RoBERTa-based classification task extremely challenging.

***System:*** Assume you can help people to label the reply-to-reply text. Particularly, in a JSON object, you will have "tweet", "reply", and "reply-to-reply" information where "tweet" is misinformation or false information, "reply" is countering, correcting, or debunking "tweet", and "reply-to-reply" is replying towards "reply". After understanding the content in the JSON object, you provide "label" for "reply-to-reply" where "-1" indicates the "reply-to-reply" disbelieves "reply" or believes in "tweet"; "0" means "reply-to-reply" does not believe or disbelieve "reply", lacking sufficient information for judgment; "1" means "reply-to-reply" believes "reply" or disbelieves "tweet".
***User:*** { "tweet": "Bill Gates was caught on video admitting that his experimental vaccine will CHANGE one's DNA", "reply": "Sorry, just found out this video is fake. Bill Gates did not demonstrate this!", "reply-to-reply": "You are correct. I checked it as well" }
***GPT-4:*** { "label": "1" }
***User:*** { ... }
***GPT-4:*** { ... }

**Figure 2: Illustration of prompts used in GPT-4 annotation.**

On the other hand, Large Language Models have progressed and shown the potential of accurately annotating text due to their human-level understanding of text [21, 79], especially the GPT-4 model [38]. Building on the existing research works regarding ChatGPT-based text annotation in computation social science domains [79], we adopt the well-performed few-shot in-context-learning diagram for GPT-4 annotation [21, 79]. First, to justify the capability of GPT-4 in annotating responses, we randomly sample four annotated data points in each category as the guidance in our carefully crafted prompt, which is presented in Figure 2. Then, we use this prompt to label our remaining annotated responses using a suggested moderate temperature of 0.5 in GPT-4 [21]. After comparing the predicted labels by GPT-4 with the ground truth labels, we find that GPT-4 has a reasonable performance in terms of precision, recall, and F1 score of 0.861, 0.859, and 0.857, respectively. Such results confirm the superior capability of GPT-4 in our annotation task and we then use it to label all responses, resulting in 23, 920 misinformation-believing, 20, 296 misinformation-disbelieving, and 11, 333 neutral responses out of 55, 549 responses. The distribution of the response count per counter-reply is shown in Figure 3.

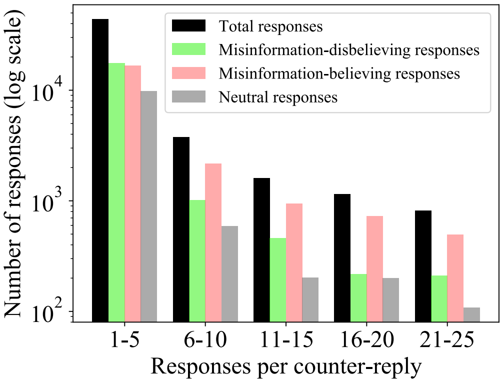


**Figure 3: Distributions of the total number of responses (black), number of misinformation-disbelieving responses (green), number of misinformation-believing responses (red), and number of neutral responses (gray) per counter-reply, each presented on a log scale.**

## 4 USER RESPONSE CHARACTERIZATION

### 4.1 Taxonomy of User Response

Different users respond differently to misinformation correction messages [70, 76]. Analyzing these fine-grained behavioral differences in social correction benefits understanding the impact of social correction, so as to promote the social correction that has corrective effects and demote the ones that have backfire effects. To this end, we first taxonomize user responses to social correction. Given that responses believing or disbelieving in misinformation primarily serve as crucial signals to indicate user reactions, we omit the remaining neutral responses – responses that neither believe nor disbelieve in misinformation – due to their minimal effects. After following similar works on reply analysis [11], we exhaustively analyze manually annotated responses and finally create the taxonomy of user responses in Table 1, presenting user actions employed to demonstrate the corresponding response type. We also assign each user response to its salient action type on a random sample of 220 and finally compute the ratio of each action to the number of total actions within the same response type.

As we can see in Table 1, there are more kinds of user actions in misinformation-believing responses than in misinformation-disbelieving responses. The potential reason is that when people are backfired by social correction, they will explore various ways to express their anger toward the counter-replies. Regarding the ratio of different user actions, we notice that providing additional information to back up misinformation or counter-misinformation is the primary action. This phenomenon may be explained that when users debate or discuss conflicting things on the Internet, they are more likely to include "evidence" to convince others. Altogether, users act differently toward social correction, yet primarily by adding "supporting information" to justify their beliefs.

### 4.2 Types of Counter-reply That Gets User Responses

Different counter-replies can lead to different responses ( i.e., responses showing belief, disbelief, or neither belief nor disbelief in misinformation). To investigate these counter-replies, we group them into four categories based on the response information and then compare them across these four categories. Practically, we first categorize replies based on the number of misinformation-believing, misinformation-disbelieving, and neutral responses a counter-reply has. We follow similar research works [37, 41] and neglect replies that have more than 25 responses since they only account for 0.218% of all counter-replies and these "super-replies" may skew the analysis [37]. Finally, we have the following categories for counter-replies:

- Corrective counter-reply: Counter-replies contain at least one misinformation-disbelieving response but no misinformation-believing responses.
- Backfire counter-reply: Similarly, counter-replies contain at least one misinformation-believing response but no misinformation-disbelieving responses.
- Neutral counter-reply: Counter-replies that only contain neutral responses.
- Dual counter-reply: Counter-replies contain both misinformation-believing and misinformation-disbelieving responses.

Finally, we identify 13, 482 corrective replies, 11, 893 backfire replies, 7, 005 neutral replies, and 2, 385 dual replies in our dataset. Due to the lowest volume of dual replies, we follow the similar research works regarding dual labels [27] and do not emphasize them, which can be worth exploring in future studies. The distribution of the reply count regarding responses per counter-reply is shown in Figure 4.

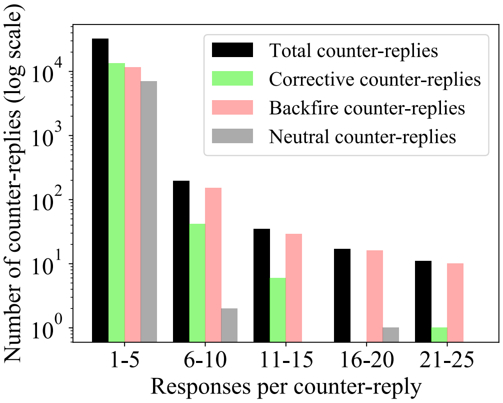


**Figure 4: Distributions of the total number of counter-replies (black), number of corrective counter-replies (green), number of backfire counter-replies (red), and number of neutral replies (gray) based on response per counter-reply, each presented on a log scale.**

| Response type | User actions employed to demonstrate the corresponding response type | Examples | Ratio |
|---|---|---|---|
| Misinformation-disbelieving response | Endorse those who counter-reply | "You are right" | 7.692% |
| | Confirm the counter-misinformation | "I checked the information as well and it is correct" | 15.385% |
| | Debunk or counter the misinformation again | "Again, the first tweet is misinformative and a bait!" | 23.077% |
| | Provide additional evidence or supporting information to back up the counter-misinformation | "Additionally, COVID-19 vaccine only generates spike protein in the cell to protect our body" | 53.846% |
| Misinformation-believing response | Refute or insult those who counter-reply | "You are completely wrong" or "You are such a fool to think in this way" | 19.791% |
| | Reject the counter-reply | "What you said does not make sense to me. The reasoning in your reply is faulty." | 7.291% |
| | Repeat, rephrase, or reconfirm the original misinformation tweet | "No. The vaccine actually is the gene therapy. It aims to change our DNA." and "The first tweet about the vaccine is correct" | 19.792% |
| | Provide additional "evidence", anecdotal experience, or supporting information to back up the misinformation | "I knew my grandfather took the vaccine and died later. So, please do not take it" | 50.000% |
| | Add new types of related misinformation | "Besides changing our DNA, the vaccine is actually developed by Bill Gates to depopulate the people. Take caution!" | 3.125% |

**Table 1: Taxonomy of user responses based on employed user actions within each response type.**

## 4.3 Analysis of Counter-reply

Analyzing and comparing counter-replies contributes to identifying salient features that are correlated with corrective or backfire effects. Given the prominent impacts of corrective and backfire counter-replies on users [13, 73], we focus on these two kinds of counter-replies for the comparison analysis. Particularly, we build on related tweet analysis works [37] and analyze the linguistic, engagement, and poster attribute features [37] as well as the counter-misinformation property [25] features of replies, as follows:

(1) **Reply linguistic attributes**, to analyze the degree to which the reply falls into meaningful personal, psychological, topical, emotional, and other content-related categories.
(2) **Reply engagement attributes**, to analyze how much the reply interacts with online users.
(3) **Reply poster attributes**, to analyze the behavior, popularity, and status of the user behind the counter-reply.
(4) **Counter-misinformation property attributes**, to analyze the extent to which the reply demonstrates the desirable property required for successful debunking backed up by the communication theory [25].

Table 2 displays the full list of attributes we statistically study[4] within each of these four categories.

*4.3.1 Linguistic Attribute Analysis.* First, we find that corrective replies are 5.409% more positive ($p < 0.005$)[5] and 8.581% less negative ($p < 0.0001$) than backfire replies. We find similar results for the "negative emotion" dimension of the LIWC lexicon ($p < 0.05$). This implies that positive tones of counter-replies convey optimistic attitudes to convince users to believe in counter-misinformation, while negative tones attract more attention and friction, and therefore, have more backfiring. Regarding the number of words in the tweet, both corrective and backfire replies have a similar length of text containing around 23 words. No statistical significance is found between these two groups. After LIWC lexicon analysis [44], we identify that backfire replies contain higher usage of affective language (words and phrases that appeal more to emotions) than corrective replies ($p < 0.05$). This indicates that those who continue to believe in misinformation when encountering counter-misinformation posts tend to gravitate more towards replying to counter-replies that induce a stronger emotional reaction. Some research works find a similar role of emotional content affecting users' resistance to correction [19]. Additionally, corrective replies mention more words related to family while backfire replies say more death-related emotions ($p < 0.05$).

*4.3.2 Engagement Attribute Analysis.* In this section, we examine the impact of engagement attributes on whether counter-replies have corrective or backfire effects. We compare the number of total likes, retweets, quotes, and replies that counter-replies receive. Because these engagements serve different purposes and have different functionalities on the platform, it is worth analyzing these metrics separately. Particularly, we find that corrective replies have more retweets (0.875 vs. 0.565 Avg. retweets per reply; $p < 0.001$) and likes (8.629 vs. 6.218 Avg. likes per reply; $p < 0.001$) but fewer replies (1.357 vs. 1.753 Avg replies per reply; $p < 0.001$) than backfire replies while they share a similar number of quotes (0.064 vs 0.065 Avg. quotes per reply) with no statistical difference. These findings may imply that the endorsement through more retweets and likes increases the believability of counter-replies [42], thus having corrective effects. In turn, we can also interpret that the

[4] This statistical test was performed using Welch's unequal variances $t$-test between corrective and backfire counter-replies.
[5] All p-values are calculated using the Welch's unequal variances t-tests.

| Attribute type | List of attributes |
|---|---|
| Reply linguistic | • Number of words in the reply.<br>• VADER [29] positive sentiment, negative sentiment, and compound sentiment of the reply.<br>• For each of the 65 dimensions of the LIWC [44] 2007 lexicon, the number of words for that dimension. |
| Reply engagement | • Number of replies, likes, retweets, and quote of a reply. |
| Reply poster | • Number of followers, and number of users following.<br>• Whether the replier is verified (1) or not (0).<br>• Total number of tweets the replier has posted since account creation. |
| Counter-misinformation property | • Politeness score of the reply, computed as the total number of politeness-related linguistic strategy instances in the reply as proposed by [16].<br>• Refutation score of the reply, obtained by the existing off-the-shelf classifier to indicate to what extent the reply is refuting the misinformation tweet [25].<br>• Evidence score of the reply, derived by checking the existence of high-credibility and fact-checking URLs in the reply [39]. |

**Table 2: List of linguistic, engagement, poster, and counter-misinformation property attributes for the counter-reply analysis.**

misinformation-disbelieving responses make the corrective counter-reply more convincing, finally having more likes and retweets [10]. This mutually-reinforced effect demonstrates the importance of engagement attributes in the analysis.

*4.3.3 Poster Attribute Analysis.* We first examine the impact of the user being verified on the counter-reply having corrective or backfire effects. We find that the proportion of accounts sending corrective counter-replies that are verified is higher than those sending backfire counter-replies (0.021 vs. 0.009 the proportion of verified accounts, $p < 0.001$). Once the account is verified, the audience will be more likely to think it is credible and believe in the counter-misinformation, demonstrating corrective effects. Likewise, unverified accounts may decrease the credibility of the counter-reply, having backfire effects. Similar findings are also identified on another poster feature – the total number of tweets since account creation. Particularly, on average, those having corrective counter-replies have more total tweets since account creation than those having backfire counter-replies ($p < 0.001$). The potential explanation can be that more tweets indicate more active and engaged repliers, thus enhancing their credibility and having corrective effects. Fewer or no tweets make the audience question the validity of the accounts. Regarding the number of followers and followings, even though we do not find a statistical difference in followers, interestingly, we find that those having corrective counter-replies have more followings ($p < 0.001$).

*4.3.4 Counter-misinformation Property Analysis.* Considering the context of counter-misinformation in our analysis, we also examine the three properties that have been shown to be crucial in effective counter-misinformation messages [13, 25]: politeness, evidence, and refutation.

Following the existing work [16, 37], we compute the politeness score of each reply and then compare the average politeness scores between the two groups. Our results find that corrective replies are 4.678% more polite than backfire replies ($p < 0.01$). This result is consistent with the existing theory that polite debunking works better than impolite debunking [13, 25]. Regarding evidence, we check the existence of high-credibility and fact-checking URLs in counter-replies, as suggested by Micallef et al. [39]. The result shows that the proportion of counter-replies that have highly credible or fact-checking URLs is 8.137% higher in corrective replies than in backfire replies. The reason may be that these URLs increase the believability of the counter-reply, finally having corrective effects.

Interestingly, we notice that results in refutation scores are different from the existing theory. Particularly, the refutation score reveals the degree to which the reply refutes the misinformation tweet. The higher the score is, the more explicitly the reply refutes the misinformation tweet, which is needed for effective countering [13]. Note that, the refutation score - where we measure the relationship between the misinformation tweet and counter-reply - is not the same as the previously examined negative sentiment. In practice, after computing the refutation score of each reply using the existing classifier [25] and comparing the average scores between the two categories, we find that corrective replies have lower refutation scores than backfire replies ($p < 0.0001$, and Cohen's $d = 0.202^6$). Even if higher refutation scores are expected in corrective replies [13], our result is still explainable considering when we add more refutation statements in replies, the emotions of some audience can be triggered [5]. This implies that when countering misinformation in real-world scenarios, we need to attend to the degree of refutation to which we reject the false information and avoid the backfire simultaneously.

## 5 USER RESPONSE PREDICTION

In this section, our primary objective is to address the research question: "Given a counter-reply, can we predict whether it will have a corrective, backfire, or neutral effect", as described in Section 3.2.

Being able to accurately predict future interactions following a counter-reply, we can identify sets of online misinformation posts where the counter-reply is organically working, as well as those

[6] Cohen's d here refers to the unweighted Cohen's d values.

| Method | Precision | Recall | F1 score |
|---|---|---|---|
| Logistic Regression | 0.753 | 0.787 | 0.769 |
| XGBoost | 0.814 | 0.764 | 0.788 |
| Neural Network | 0.832 | 0.801 | 0.816 |

**Table 3: Classification performance of whether a counter-reply will have a corrective, backfire, or natural effect.**

requiring additional countermeasures. Finally, we can pinpoint instances of the counter-replies that do not work such that the associated misinformation tweets can be proactively and carefully countered by other users to curb the spread of misinformation.

## 5.1 Dataset

To answer the above research question, we use the aforementioned dataset in Section 4. Particularly, we divide the counter-replies into three sets: (1) corrective counter-replies; (2) backfire counter-replies; and (3) neutral counter-replies, as defined in Section 4.2. The sizes of the three sets are $13,482$, $11,893$, and $7,005$.

## 5.2 Experiment Setup

After choosing the dataset, we follow similar approaches in tweet prediction tasks [39, 75] by building a multi-class classifier. We utilize the set of attributes described in Section 4.3 as features. As shown in the existing tweet prediction work [39], the semantic information from textual embedding benefits the prediction task. Thus, we also generate the embedding vector for each reply using RoBERTa [36]. Finally, we concatenate the above feature vectors to form a reply feature vector to comprehensively represent the reply and use it for classification.

## 5.3 Classifier Creation and Evaluation

Following similar tweet or general text classification tasks [24, 39], we deploy widely-used machine learning classifiers including Logistic Regression, XGBoost, and a Feed-forward Neural Network containing 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.

The classification result is shown in Table 3. As we can see, the model performance is reasonably acceptable. Especially, the neural network achieves the best performance regarding precision, recall, and F-1 score; this finding is also found in other similar tweet classification tasks [39]. This high performance offers the ability to effectively predict whether a counter-reply will have a corrective, backfire, or neutral effect, enabling fact-checkers and social media platforms to organically prioritize counter-replies identified as more likely to backfire.

# 6 DISCUSSION AND CONCLUSION

In this paper, we curate a large-scale conversation-style dataset containing user responses to social correction and build a taxonomy to present different types of these user responses. We also study the text- and user-level properties of counter-replies that have corrective or backfire effects. The in-depth analysis shows that counter-replies expressing positive sentiments and politeness and having evidence are more likely to have corrective effects. Our result also shows that counter-replies that have corrective effects have a higher amount of retweet and like engagement that expresses endorsement. Moreover, we develop a well-performed classifier to predict whether a counter-reply will have a corrective, backfire, or neutral effect. In sum, our work comprehensively demonstrates that the user response to social corrections has implications regarding what kinds of social corrections work better, and sheds light on how to combat misinformation by social correction.

There are still some limitations in our work. First, we only utilize user responses to determine the impact of counter-reply, which is acceptable because user responses can provide both positive and negative feedback through the expressed disbelief and belief in misinformation respectively. However, the number of responses is usually small for one counter-reply, and the signals from the user engagement (e.g., retweets and likes) are not considered together to form a comprehensive evaluation of counter-replies in our work. Another notable limitation is its exclusive focus on Twitter. The dynamics of post engagement and information exchange can vary significantly across different online platforms [40], potentially influencing the nature of social correction and user response to it. Besides, even if GPT-4 demonstrates commendable performance in annotation tasks [38], our study utilizes it exclusively for annotating responses, rather than extending its use to tweets and replies, which does not perform a uniform annotation process for all data points. The reason is that the employment of GPT-4 is constrained by its high API costs. In contrast, traditional low-cost BERT-based classifiers for tweets and replies yield satisfactory results, with both F1 scores exceeding 0.8, which aligns well with the requirements of our research. We also admit that our results will depend on the reliability of classifiers to accurately identify misinformation tweets and counter-replies for downstream analysis. Additionally, our study is limited to English language text since we filter out misinformation tweets in other languages. The dynamics in other languages could present different patterns in misinformation spread and correction. Furthermore, our analysis is confined to discussions around COVID-19 vaccines, a topic that has garnered widespread attention due to the global impact of the COVID-19 pandemic. This focus may not fully represent the dynamics of other prevalent misinformation topics [54], such as climate change misinformation, where the specific countering text and demographics of posters could influence interaction patterns in distinct ways. Finally, we only examine the text information while other modalities like images and videos can manifest various patterns and actions in social correction.

For future work, we can first consider combining user engagements (e.g., likes and retweets of counter-replies) with user responses together to comprehensively determine the impact of counter-replies. Second, we could extend our analysis to the user networks of misinformation posters, those who counter-reply, and responders to the counter-replies to investigate the potential phenomenon of networked "echo chamber" [17]. This would involve examining the followers and followees of these users, as well as the prevalence of misinformation and counter-misinformation within these networks, to identify network attributes that might influence the effect of counter-replies: backfire or corrective effects. In addition, accurately predicting whether a counter-reply can have a corrective, backfire, or neutral effect opens up opportunities for field studies to investigate how specific characteristics of counter-replies might affect a user's belief in misinformation.

**ACKNOWLEDGMENTS** This research/material is based upon work supported by NSF grant CNS-2154118. 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.

## REFERENCES

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

[2] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7 (9 2021). Issue 36. https://doi.org/10.1126/sciadv.abf4393

[3] 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

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

[5] Mareike Bayer, Werner Sommer, and Annekathrin Schacht. 2010. Reading emotional words within sentences: the impact of arousal and valence on event-related potentials. *International Journal of Psychophysiology* 78, 3 (2010), 299–307.

[6] Leticia Bode and Emily K. Vraga. 2015. In Related News, That was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication* 65, 4 (06 2015), 619–638. https://doi.org/10.1111/jcom.12166 arXiv:https://academic.oup.com/joc/article-pdf/65/4/619/22320531/jjnlcom0619.pdf

[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. Issue 9. https://doi.org/10.1080/10410236.2017.1331312

[8] Leticia Bode, Emily K Vraga, and Melissa Tully. 2020. Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review* (2020).

[9] Porismita Borah, Bimbisar Irom, and Ying Chia Hsu. 2021. 'It infuriates me': examining young adults' reactions to and recommendations to fight misinformation about COVID-19. *Journal of Youth Studies* (8 2021), 1–21. https://doi.org/10.1080/13676261.2021.1965108

[10] Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii international conference on system sciences*. IEEE, 1–10.

[11] George Buchanan, Ryan Kelly, Stephann Makri, and Dana McKay. 2022. Reading between the lies: A classification scheme of types of reply to misinformation in public discussion threads. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 243–253.

[12] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. *Proceedings of the 20th international conference on World wide web - WWW '11*, 665. https://doi.org/10.1145/1963405.1963499

[13] 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.

[14] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 2023. The Roll-Out of Community Notes Did Not Reduce Engagement With Misinformation on Twitter. *arXiv preprint arXiv:2307.07960* (2023).

[15] 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.

[16] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *51st Annual Meeting of the Association for Computational Linguistics*. ACL, 250–259.

[17] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2015. Echo chambers in the age of misinformation. *arXiv preprint arXiv:1509.00189* (2015).

[18] 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.

[19] Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* 1, 1 (2022), 13–29.

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

[21] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056* (2023).

[22] Mahak Goindani and Jennifer Neville. 2020. Social Reinforcement Learning to Combat Fake News Spread, Ryan P Adams and Vibhav Gogate (Eds.). *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference* 115, 1006–1016. https://proceedings.mlr.press/v115/goindani20a.html

[23] 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.

[24] Bing He, Mustaque Ahamad, and Srijan Kumar. 2021. Petgen: Personalized text generation attack on deep sequence embedding-based classification models. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 575–584.

[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, Yibo Hu, Yeon-Chang Lee, Soyoung Oh, Gaurav Verma, and Srijan Kumar. 2023. A survey on the role of crowds in combating online misinformation: Annotators, evaluators, and creators. *arXiv preprint arXiv:2310.02095* (2023).

[27] 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.

[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] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, Vol. 8.

[30] 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.

[31] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 324–332.

[32] Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction* 4 (10 2020), 1–27. Issue CSCW2. https://doi.org/10.1145/3415211

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

[34] 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.

[35] Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. 2017. Efficient and timely misinformation blocking under varying cost constraints. *Online Social Networks and Media* 2 (8 2017), 19–31. https://doi.org/10.1016/j.osnem.2017.07.001

[36] 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).

[37] Yingchen Ma, Bing He, Nathan Subrahmanian, and Srijan Kumar. 2023. Characterizing and Predicting Social Correction on Twitter. In *Proceedings of the 15th ACM Web Science Conference 2023*. 86–95.

[38] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. GPTEval: A survey on assessments of ChatGPT and GPT-4. *arXiv preprint arXiv:2308.12488* (2023).

[39] 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.

[40] 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.

[41] Kunihiro Miyazaki, Takayuki Uchiba, Kenji Tanaka, Jisun An, Haewoon Kwak, and Kazutoshi Sasahara. 2023. " This is Fake News": Characterizing the Spontaneous Debunking from Twitter Users to COVID-19 False Information. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 650–661.

[42] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 441–450.

[43] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[44] 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.

[45] Christina Peter and Thomas Koch. 2016. When debunking scientific myths fails (and when it does not) The backfire effect in the context of journalistic coverage and immediate judgments as prevention strategy. *Science Communication* 38, 1 (2016), 3–25.

[46] 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.

[47] Ethan Porter and Thomas J Wood. 2021. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences* 118, 37 (2021), e2104235118.

[48] Nicolas Pröllochs. 2022. Community-based fact-checking on Twitter's Birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 794–805.

[49] Philipp Schmid and Cornelia Betsch. 2022. Benefits and pitfalls of debunking interventions to counter mRNA vaccination misinformation during the COVID-19 pandemic. *Science Communication* 44, 5 (2022), 531–558.

[50] 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.

[51] 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).

[52] 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.

[53] Briony Swire-Thompson, Nicholas Miklaucic, John P Wihbey, David Lazer, and Joseph DeGutis. 2022. The backfire effect after correcting misinformation is strongly associated with reliability. *Journal of Experimental Psychology: General* 151, 7 (2022), 1655.

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

[55] Melissa Tully, Leticia Bode, and Emily K. Vraga. 2020. Mobilizing Users: Does Exposure to Misinformation and Its Correction Affect Users' Responses to a Health Misinformation Post? *Social Media + Society* 6 (10 2020), 205630512097837. Issue 4. https://doi.org/10.1177/2056305120978377

[56] Jeyasushma Veeriah. 2021. YOUNG ADULTS'ABILITY TO DETECT FAKE NEWS AND THEIR NEW MEDIA LITERACY LEVEL IN THE WAKE OF THE COVID-19 PANDEMIC. *Journal of Content, Community and Communication* 13 (2021), 372–383. Issue 7.

[57] 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.

[58] Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 275–284.

[59] 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.

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

[61] Emily Vraga, Melissa Tully, and Leticia Bode. 2021. Assessing the relative merits of news literacy and corrections in responding to misinformation on Twitter. *New Media & Society* (2021), 1461444821998691.

[62] Emily K Vraga and Leticia Bode. 2018. I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society* 21, 10 (2018), 1337–1353.

[63] Emily K Vraga and Leticia Bode. 2020. Correction as a solution for health misinformation on social media. *American Journal of Public Health* 110, Suppl 3 (2020), S278.

[64] Emily K Vraga and Leticia Bode. 2021. Addressing COVID-19 misinformation on social media preemptively and responsively. *Emerging infectious diseases* 27, 2 (2021), 396.

[65] Emily K Vraga, Leticia Bode, and Melissa Tully. 2021. The effects of a news literacy video and real-time corrections to video misinformation related to sunscreen and skin cancer. *Health communication* (2021), 1–9.

[66] Nathan Walter, John J Brooks, Camille J Saucier, and Sapna Suresh. 2021. Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health Communication* 36, 13 (2021), 1776–1784.

[67] Nathan Walter and Sheila T Murphy. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs* 85, 3 (2018), 423–441.

[68] Bairong Wang and Jun Zhuang. 2018. Rumor response, debunking response, and decision makings of misinformed Twitter users during disasters. *Natural Hazards* 93 (2018), 1145–1162.

[69] Xin Wang, Fan Chao, and Guang Yu. 2021. Evaluating rumor debunking effectiveness during the COVID-19 pandemic crisis: utilizing user stance in comments on Sina Weibo. *Frontiers in Public Health* 9 (2021), 770111.

[70] Xin Wang, Fan Chao, Guang Yu, and Kaihang Zhang. 2022. Factors influencing fake news rebuttal acceptance during the COVID-19 pandemic and the moderating effect of cognitive ability. *Computers in human behavior* 130 (2022), 107174.

[71] Zhihong Wang and Yi Guo. 2020. Rumor events detection enhanced by encoding sentimental information into time series division and word representations. *Neurocomputing* 397 (7 2020), 224–243. https://doi.org/10.1016/j.neucom.2020.01.095

[72] Senuri Wijenayake, Danula Hettiachchi, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2021. Effect of Conformity on Perceived Trustworthiness of News in Social Media. *IEEE Internet Computing* 25 (1 2021), 12–19. Issue 1. https://doi.org/10.1109/MIC.2020.3032410

[73] Chloe Wittenberg and Adam J Berinsky. 2020. Misinformation and its correction. *Social media and democracy: The state of the field, prospects for reform* 163 (2020).

[74] 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.

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

[76] Yuqi Zhang, Bin Guo, Yasan Ding, Jiaqi Liu, Chen Qiu, Sicong Liu, and Zhiwen Yu. 2022. Investigation of the determinants for misinformation correction effectiveness on social media during COVID-19 pandemic. *Information Processing & Management* 59, 3 (2022), 102935.

[77] Xinyi Zhou, Kai Shu, Vir V Phoha, Huan Liu, and Reza Zafarani. 2022. "This is Fake! Shared it by Mistake": Assessing the Intent of Fake News Spreaders. In *Proceedings of the ACM Web Conference 2022*. 3685–3694.

[78] Jianming Zhu, Smita Ghosh, and Weili Wu. 2021. Robust rumor blocking problem with uncertain rumor sources in social networks. *World Wide Web* 24 (1 2021), 229–247. Issue 1. https://doi.org/10.1007/s11280-020-00841-8

[79] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? *arXiv preprint arXiv:2305.03514* (2023).