

**LEVERAGING AI TO COMBAT MISINFORMATION BY EMPOWERING
CROWDS AND EVALUATING DETECTORS**

A Ph.D. Thesis
Presented to
The Academic Faculty

By

Bing He

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing
School of Cybersecurity and Privacy

Georgia Institute of Technology

August 2024

© Bing He 2024

**LEVERAGING AI TO COMBAT MISINFORMATION BY EMPOWERING
CROWDS AND EVALUATING DETECTORS**

Thesis committee:

Dr. Mustaque Ahamad (Advisor)
School of Cybersecurity and Privacy
Georgia Institute of Technology

Dr. Frank Li
School of Cybersecurity and Privacy
Georgia Institute of Technology

Dr. Srijan Kumar (Advisor)
School of Computational Science and En-
gineering
Georgia Institute of Technology

Dr. Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology

Dr. Nasir Memon
Tandon School of Engineering
New York University

Date approved: May 6, 2024

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my thesis committee for their help in the completion of this work.

I am very grateful to my academic advisors, Professor Mustaque Ahamad and Professor Srijan Kumar, for their unlimited support, guidance, and encouragement throughout my Ph.D. journey. Their expertise, constructive feedback, and dedication have been invaluable, shaping both the content and direction of this work. I extend my sincere thanks to the members of my dissertation committee for their time, expertise, and thoughtful insights. Professor Frank Li and Professor Munmun De Choudhury have provided constructive critiques and suggestions since the Ph.D. proposal. Professor Nasir Memon has also collaborated with us for a long time and significantly enhanced the quality of our research papers.

I am also grateful to the Georgia Institute of Technology, Institute for Data Engineering and Science (IDEaS) and Microsoft Azure for providing computational resources, which allowed me to solely focus on my research without any resource constraints.

A big shout-out to my colleagues in the CLAWS lab who have offered intellectual companionship, encouragement, and feedback. Their support has been a source of inspiration and motivation. Finally, I want to express my deepest gratitude to my family and friends. Their unconditional support, understanding, and encouragement have been the main fuel for my academic journey.

Again, thank you to everyone for being part of this significant milestone in my life.

TABLE OF CONTENTS

Acknowledgments	iii
List of Tables	vii
List of Figures	ix
Summary	xi
Chapter 1: Introduction and Background	1
1.1 Introduction	1
1.1.1 Motivation	1
1.1.2 Overview and Contributions	3
1.2 Background and Related Works	6
1.2.1 Analysis of Counter-Misinformation and Misinformation Correction	7
1.2.2 User Response to Counter-misinformation by Crowds	9
1.2.3 Counter-misinformation Reply Generation	10
1.2.4 Deep Sequence Embedding-based Classification Models	11
1.2.5 Attack and Defense on Deep Sequence Embedding-based Classification Models	12
Chapter 2: Characterizing Crowd-Generated Counter-Misinformation	14

2.1	Introduction	14
2.2	Data Curation	17
2.3	Counter-misinformation Characterization	24
2.4	Conclusion	32
Chapter 3: Characterizing and Predicting User Response to Crowd-generated Counter-misinformation Replies		33
3.1	Introduction	33
3.2	Dataset	35
3.3	User Response Characterization	40
3.4	User Response Prediction	47
3.5	Conclusion	49
Chapter 4: Generating Counter-Misinformation to Empower Crowds		50
4.1	Introduction	50
4.2	Problem Definition	53
4.3	Counter-response Datasets: In-the-Wild and Crowdsourced	54
4.4	MisinfoCorrect: A Counter-Response Generation Model	58
4.5	Experimental Evaluation	63
4.6	Conclusion	68
Chapter 5: Evaluating the Robustness of Deep Sequence Embedding-based Detectors		70
5.1	Introduction	70
5.2	Problem Definition	73

5.3	Methodology	74
5.4	Experiments	81
5.5	Conclusion	89
Chapter 6: Improving the Robustness of Deep Sequence Embedding-based Detectors		90
6.1	Introduction	90
6.2	Problem Definition	92
6.3	Methodology	93
6.4	Evaluation	98
6.5	Conclusion	104
Chapter 7: Concluding Remarks		105
7.1	Opportunities for Real-World Impact	105
7.2	Limitations	107
7.3	Conclusions	109
7.4	Future Work	110
References		113
Vita		133

LIST OF TABLES

2.1	Examples of tweets and assigned labels in misinformation and counter-misinformation studies.	21
2.2	Misinformation and counter-misinformation tweet classification performance for 5G topic.	23
2.3	Misinformation and counter-misinformation tweet classification performance for fake cures topic.	23
2.4	Misinformation and counter-misinformation twitter Data statistics	24
3.1	Taxonomy of user responses based on employed user actions within each response type.	41
3.2	List of linguistic, engagement, poster, and counter-misinformation property attributes for the counter-reply analysis.	44
3.3	Classification performance of whether a counter-reply will have a corrective, backfire, or natural effect.	48
4.1	Statistics of 754 social media counter-responses in MisinfoCorrect.	57
4.2	Effectiveness of MisinfoCorrect. Performance comparison of counter-response generators when trained on social media and crowdsourced responses.	65
4.3	Examples of generated counter-responses by MisinfoCorrect and baseline methods.	65
4.4	Effectiveness of MisinfoCorrect. Performance comparison of counter-response generators when trained on social media responses only.	66
4.5	Ablation study of MisinfoCorrect.	67

5.1	Table of major notations used in PETGEN.	74
5.2	Statistics of datasets used in PETGEN	82
5.3	<i>White-box attack performance</i> of PETGEN and existing methods on HRNN and TIES classifiers. PETGEN is the most effective attack (lowest F1 and highest Atk score).	82
5.4	<i>Black-box attack performance</i> of PETGEN and existing methods on HRNN and TIES classifiers. PETGEN is the most effective attack (lowest F1 and highest Atk score).	83
5.5	Comparison the quality of text generated by PETGEN and other attack strategies. PETGEN generates higher quality text in all but one case across all metrics.	85
5.6	Ablation studies of PETGEN showing the contribution of each component in PETGEN.	86
6.1	Table of notations used in the defense model	93
6.2	Dataset statistics used in the defense model	98
6.3	Comparison of classification performance by different defense models on Wikipedia dataset	101
6.4	Comparison of classification performance by different defense models on Yelp dataset	101
6.5	Comparison of robustness performance by different defense models on Wikipedia dataset. Here, we report the F1 score after the attack.	102
6.6	Comparison of robustness performance by different defense models on Wikipedia dataset. Here, we report the F1 score after the attack.	103
6.7	Ablation studies of our defense model on Wikipedia dataset, measured by F1 score.	104
6.8	Ablation studies of our defense model on Yelp dataset, measured by F1 score.	104

LIST OF FIGURES

1.1	Landscape of the PhD Research	3
2.1	Most frequent external sources in misinformation and counter-misinformation tweets. High-credibility sources are more common in counter-misinformation tweets.	25
2.2	Tweet volume and engagement of misinformation and counter-misinformation over time. Both volume and engagement of countering misinformation are comparable to misinformation.	26
2.3	Sentiment and politeness of misinformation and counter-misinformation for fake cures. Misinformation tweets are more positive and professional fact-checks are neutral (left). Evidence-based citizen responses are less polite than misinformation (right).	28
2.4	LIWC results of misinformation and counter-misinformation for fake cures (left) and 5G (right) tweets. Opinion-based citizen responses have similar linguistic properties as misinformation.	30
3.1	Examples of user responses to social correction. Here, the social correction is the counter-misinformation <i>reply</i> posted by ordinary users (the second row), and the user response is the <i>reply</i> to the counter-misinformation reply (the third row).	33
3.2	Illustration of prompts used in GPT-4 annotation.	39
3.3	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.	40

3.4	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.	43
4.1	An overview of counter-misinformation response generation task in MisinfoCorrect.	51
4.2	The overview of the MisinfoCorrect framework.	59
5.1	Application Setting of PETGEN: Deep user sequence embedding-based classification models are used to detect malicious users (top row). However, an evasion attack by an adversary by creating a new fake post can lead the same model to misclassify it as a benign user (bottom row). Our method, PETGEN, generates personalized text posts to adversarially attack the classifier.	71
5.2	Overview of the PETGEN architecture: The sequence-aware text generator utilizes the sequence of post and context to generate text that maintains the contextual post relevance. Then, the multi-stage multi-task learning module fine-tunes the text by different tasks to generate attack text.	75
5.3	The overview of the sequence-aware conditional text generator in PETGEN. We first create the sequence embedding from the post embedding of each post in a sequence. We also compute the attention score between the target context and the user’s historical contexts to capture their pairwise relevance, resulting in a context-aware attention vector. After multiplying the generated sequence embedding and attention vector, we get the context-biased user sequence embedding. We concatenate it with the generated tokens for sequence-aware conditional text generation.	76
6.1	Deep sequence classification models are used to detect malicious users. However, the next post attack by an adversary by creating a new fake post can lead the same model to misclassify it as a benign user. Our proposed solution built on the local-global attended and adversary-aware modules can accurately and robustly identify malicious users.	90
6.2	Overview of the proposed defense model.	93
6.3	Local-global attended module in the defense model	94
6.4	Adversary-aware module in the defense model	95

SUMMARY

Online misinformation poses a global risk, leading to threatening real-world implications. To combat misinformation, existing research works either focus on leveraging the expertise of professionals including journalists and fact-checkers to annotate and debunk misinformation, or develop automatic ML methods to detect misinformation and its spreaders. However, the efficacy of professionals is limited because their manual processes do not scale with the volume of misinformation; ML methods rely on deep sequence embedding-based classifiers for detecting misinformation spreaders, but their vulnerabilities are rarely examined. To complement professionals, non-expert ordinary users (a.k.a. crowds) can act as eyes-on-the-ground who proactively question and counter misinformation, showing promise in overcoming the limitations of solely relying on professionals. However, little is known about how these crowds organically combat misinformation. Concurrently, AI has progressed dramatically, demonstrating the potential to help combat misinformation. In this thesis, we aim to utilize AI to investigate the aforementioned challenges and provide insights and solutions to empower crowds to better counter misinformation.

We first characterize crowds who counter misinformation on social media platforms and how users respond to these counter-misinformation messages, and then assist crowds by generating more effective counter-misinformation replies. We apply advanced AI techniques to characterize the spread and textual properties of counter-misinformation generated by crowds as well as their characteristics during the COVID-19 pandemic. Interestingly, we found 96% counter-misinformation posts are made by crowds, which confirms their prominent role in combating misinformation. We also analyze user responses toward crowd-generated counter-misinformation replies in a conversation to investigate the impact of these counter-misinformation replies. As expected, we discovered that counter-misinformation replies that are polite, positive, and evidenced have a higher possibility of having a corrective effect on users. Our analysis work provides insights into how online

misinformation is organically countered by crowds and how users respond to such counter-misinformation. Alarming, we also noticed that 2 out of 3 crowd messages are rude and lack evidence, and impolite and non-evidence replies may cause backfire. Generating an effective counter-misinformation response is thus crucial but challenging due to the absence of high-quality datasets and communication theory-backed models. To address these challenges, we first create two novel datasets of misinformation and counter-misinformation response pairs from in-the-wild social media and in-lab crowdsourcing, and then propose a reinforcement learning-based AI algorithm, called MisinfoCorrect, that learns to generate high-quality counter-misinformation responses for an input misinformation post. Our work illustrates the promise of AI for empowering crowds in combating misinformation.

On the other hand, deep sequence embedding-based classification methods, which use a sequence of user posts to generate user embeddings and detect malicious users, are also employed to identify misinformation spreaders on social media platforms. Although deep learning models are shown to be vulnerable to adversarial attacks in computer vision and natural language processing domains, the vulnerability of deep sequence embedding-based detectors remains unknown. Thus, we evaluate existing detectors by proposing a novel end-to-end AI algorithm, called PETGEN (PERSONalized Text GENERator), that simultaneously reduces the efficacy of the detection model and generates high-quality personalized posts. Next, to improve the robustness of these detection models against the next post attack, we propose a novel transformer-based detection model. The algorithm first comprehensively encodes the local and global information (i.e., the post and sequence information) by transformer encoder and decoder blocks, and then deploys the contrastive learning-enhanced classification loss to consider the adversarial attack scenario during training. Building on our efforts, we pave the path toward the next generation of adversary-aware deep sequence embedding-based classification models to robustly identify misinformation spreaders.

Our AI-based approaches lead to solutions that can empower crowds and better automated detectors for efficiently and effectively combating misinformation.

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Introduction

1.1.1 Motivation

Most people receive information and news from social media [1], which is often “ground-zero” for misinformation and where online misinformation spreads faster and farther than truth [2, 3]. Online misinformation reduces trust in vaccines and health policies [4, 5, 3], leads to violence and harassment [6, 7], questions democratic processes, increases polarization [8], and harms well-being [9]. For example, COVID-19 misinformation (e.g., 5G causes COVID-19.) and related vaccine misinformation (e.g., COVID-19 vaccine changes DNA.) has fueled vaccine hesitancy, reduced vaccine uptake, and prolonged the pandemic. Besides, misinformation also causes harm to people directly. For instance, misinformation that Bill Gates created vaccines to depopulate people led to distrust and verbal attacks [10]. Thus, it is critical to combat the spread of online misinformation¹ [13, 14, 15, 16, 17, 18, 19].

To combat misinformation, some research works rely on professional fact-checkers and journalists to form the first line of defense by fact-checking popular false claims [20]. These professionals provide objective fact-checks for viral claims and release their determination on their websites, which are useful for debunking misinformation. However, they do not engage directly in conversations with misinformation spreaders to have large-scale impact [13]. Even if non-expert ordinary users (a.k.a. **Crowds**) can act as eyes-on-the-ground who proactively question and counter misinformation, little is known about how these crowds organically combat online misinformation in the real world. Therefore, it is

¹In this work, we use a broad definition of misinformation which includes falsehoods, inaccuracies, rumors, decontextualized truths, or misleading leaps of logic [11, 12]

imperative to characterize crowds who counter misinformation and how to assist them, so as to complement fact checkers who can only verify a handful of stories after they have gone viral [21].

On the other hand, some researchers design and apply automatic ML methods to detect misinformation and its spreaders, aiming to halt its spread and ensure an informed public discourse. They utilize user reports through reinforcement learning [22], leverage textual and visual features from posts by adversarial neural network [23], or employ user features with Bayes Network and k-Nearest Neighbors [24]. Among these solutions, deep sequence embedding-based classification models [25] are widely used to identify malicious users including misinformation spreaders. Especially, Facebook proposed TIES for malicious account detection [25]. However, deep learning models can be vulnerable to adversarial attacks [26], and existing research works have shown the vulnerabilities of graph representation learning [27], natural language processing [28], and computer vision [26] models. Unfortunately, the vulnerability of deep sequence embedding-based classification models remains unknown. Even worse, a malicious user can adaptively write a new post to flip the classification from malicious to benign to bypass the detector and continue spreading misinformation. Thus, identifying vulnerabilities of deep sequence embedding-based classification models and improving their robustness is crucial.

In recent years, Artificial Intelligence (AI) has progressed dramatically with a substantial and multifaceted impact across numerous domains ranging from education to healthcare [29, 30]. The AI techniques have also been widely used in combating misinformation ranging from detecting misinformation [31, 32] and its spreaders [33, 25] to assisting professional fact-checkers [34, 35]. These applications demonstrate the potential of AI in the fight against misinformation. However, few researchers have utilized AI from the dimensions of (i) empowering crowds who proactively counter misinformation and, (2) identifying vulnerabilities of deep sequence embedding-based classifiers. To this end, it is imperative to explore the utilization of AI through these two dimensions to effectively

combat misinformation.

1.1.2 Overview and Contributions

To close the aforementioned research gaps derived from professionals and ML classifiers, the goal of this thesis is to **design and apply AI algorithms to empower crowds (Chapter 2, 3, and 4) and evaluate detectors (Chapter 5 and 6) for combating misinformation**, as illustrated in Figure 1.1.

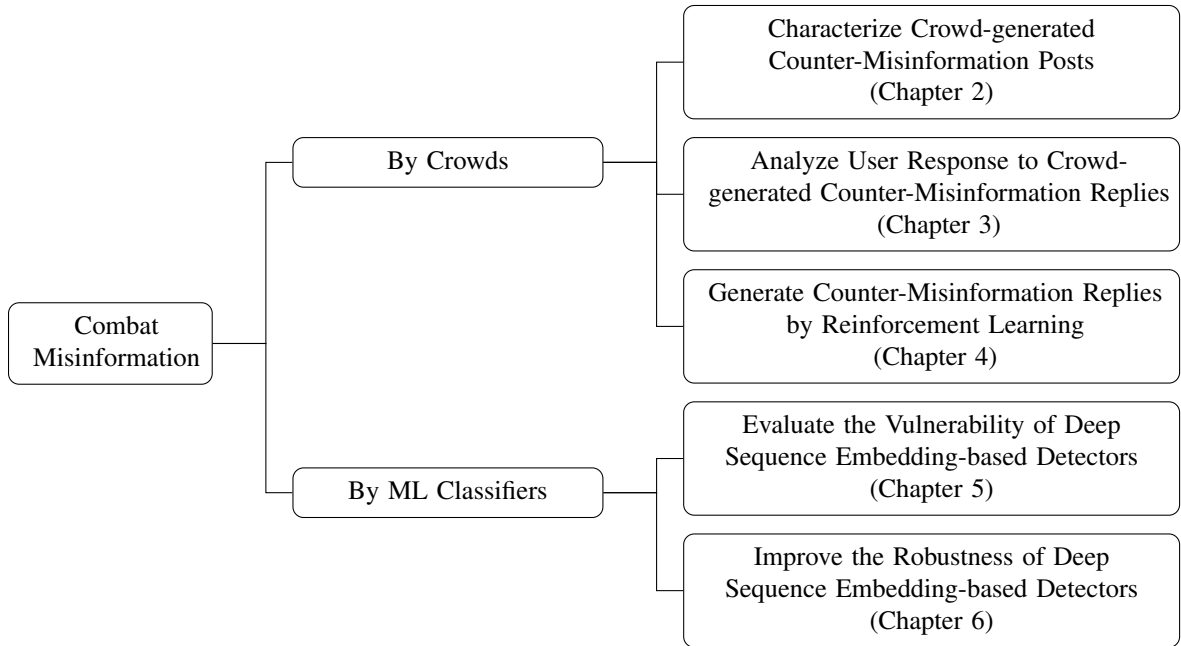


Figure 1.1: Landscape of the PhD Research

We give a brief description of each proposed effort in each thrust.

Characterizing how crowds counter misinformation (in Chapter 2). We conduct a data-driven study of misinformation on Twitter² [13], analyzing the spread of misinformation, professional fact checks, and the crowd responses to popular misleading claims about COVID-19. We curate a dataset of misinformation tweets and counter-misinformation tweets that seek to challenge or refute false information. We train a classifier to create a novel dataset of 155,468 COVID-19-related tweets, containing 33,237 misinformation

²Twitter has been rebranded as X from Aug 2023.

tweets and 33,413 counter-misinformation tweets. Our findings show that professional fact-checking tweets have limited volume and reach. In contrast, 96% counter-misinformation tweets come from crowds, and the surge in misinformation tweets results in a quick crowd response and a corresponding increase in tweets that refute such misinformation. More importantly, we find contrasting differences in the way the crowd refutes tweets, particularly, 67% tweets appear to be opinions, while others contain concrete evidence, such as a link to a reputed source. Our work provides insights into how misinformation is organically countered in social platforms by some of their users and the role they play in amplifying professional fact checks.

Analyzing how users respond to crowd-generated counter-misinformation replies (in Chapter 3). We conduct a data-driven study to characterize and predict the user response to counter-misinformation replies by crowds, where the user response is instantiated as the reply that is written toward a counter-misinformation message. Specifically, we first create a novel dataset with 55,549 triples of misinformation tweets, counter-misinformation replies, and responses to counter-misinformation replies. 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.

Generating counter-response for crowds to correct misinformation (in Chapter 4) When analyzing the counter-misinformation responses by crowds [13], we found that 2/3 times, these responses are rude and lack evidence. Even worse, these responses can have backfire effects. Thus, we created a counter-misinformation response generation model to empower users to effectively correct misinformation using high-quality responses in [36].

This objective is challenging due to the absence of datasets containing ground-truth of high-quality counter-misinformation responses, and the lack of models that can generate responses backed by communication theories. In this work, we create two novel datasets of misinformation and counter-misinformation response pairs from in-the-wild social media and crowdsourcing from college-educated students. We annotate the collected data to distinguish low-quality from high-quality responses that are factual, polite, and refute misinformation. We propose MisinfoCorrect, a reinforcement learning-based framework that learns to generate counter-misinformation responses for an input misinformation post. The model rewards the generator to increase the politeness, factuality, and refutation attitude while retaining text fluency and relevancy. The quantitative and qualitative evaluation shows that our model outperforms several baselines with an average 8.42% improvement in all evaluation metrics (i.e., politeness, refutation, evidence, perplexity, and relevance) by generating high-quality counter-responses.

Evaluating the robustness of existing deep sequence embedding-based detectors (in Chapter 5). We propose a novel adversarial attack model called PETGEN [37] against deep user sequence embedding-based classification models, which use a sequence of user posts to generate user embeddings and detect malicious users. In the attack, PETGEN generates a new post to fool the classifier such that it simultaneously reduces the efficacy of the detection model and generates desirable posts. Specifically, PETGEN generates posts that are personalized to the user’s writing style, have knowledge about a given target context, are aware of the user’s historical posts on the target context, and encapsulate the user’s recent topical interests. We conduct extensive experiments on two real-world datasets (Yelp and Wikipedia, both with ground-truth of malicious users) to show that PETGEN significantly reduces the performance of popular deep sequence embedding-based classification models by an average drop of 17.66% in F1 score. PETGEN outperforms five attack baselines in terms of text quality and attack efficacy in both white-box and black-box classifier settings.

Improving the robustness of existing deep sequence embedding-based detectors

(in Chapter 6). To improve the model robustness, we create a novel transformer-based adversary-aware local-global attended detector. It leverages the transformer encoder block to encode each post bidirectionally, thus building a comprehensive post embedding. Next, the model adopts the transformer decoder block to model the sequential pattern in the post embeddings by using an attention mechanism to generate the sequence embedding. Through this design, we leverage different levels of information in the sequence of user posts (e.g., local and global information) to enhance the model. Finally, sequence embeddings of original sequences and modified sequences of mimicked attackers are fed together into a contrastive-learning-enhanced classification layer to enhance the model knowledge so that it can be stable against adversarial attacks. After evaluating on Yelp and Wikipedia datasets again, we find that our model can outperform representative compared methods by robustly detecting malicious users under state-of-the-art attacks with the lowest reduced F1 score. Overall, this work paves the path toward the next generation of adversary-aware robust sequence classification models.

Each of our explored AI-based endeavors can serve as a “lighthouse” that introduces new research directions for academia. Altogether, we conduct data-driven analysis and develop novel algorithms to empower crowds and evaluate detectors to combat misinformation for a reliable and responsible online social network ecosystem.

1.2 Background and Related Works

Combating misinformation has been widely investigated in academia by exploring various directions (e.g., crowds and ML detectors) so as to mitigate the spread of misinformation. We first cover several key aspects regarding the crowds who counter misinformation: (1) Current counter-misinformation studies and correction of misinformation by crowds in Section 1.2.1; (2) User responses to misinformation correction by crowds in Section 1.2.2; (3) Existing counter-misinformation reply generation methods that can assist crowds in Section 1.2.3. Then, we introduce the existing deep sequence embedding-based classification

models used for detecting misinformation spreaders in Section 1.2.4, as well as attack and defense on these models in Section 1.2.5.

1.2.1 Analysis of Counter-Misinformation and Misinformation Correction

Social media contains many types of misinformation, such as rumors [38], hoaxes [39], false news [2] and false information [11]. Due to its harmful impact on society, various combating efforts appear [38, 23, 40]. Among them, counter-misinformation contents serve as a critical part of the misinformation-combating ecosystem due to their potential to correct misconceptions and reduce the spread of misinformation. Many researchers have explored counter-misinformation contents [22, 23, 40, 41, 42], and these efforts can be grouped into three categories:

(1) Misinformation and Counter-Misinformation Detection: Different data sources and methods have been developed, including user reports through reinforcement learning [22], textual and visual features from posts by adversarial neural network [23], new user and tweets features with Bayes Network, k-Nearest Neighbors and Adaptive Boosting supervised learning framework [24]. Besides, some computational methods are designed to limit the propagation of misinformation by spreading counter-misinformation[43]. However, the existing research works either focus more on misinformation, or emphasize finding the set of users who initiate the propagation of counter-misinformation while neglecting to consider and analyze the textual information of counter-misinformation.

(2) User Survey-based Experimental Studies: In [40, 41, 42], user surveys are conducted to determine how people can combat misinformation by posting reliable rebuttals on social media platforms. However, these studies are small-scale.

(3) Observational Studies: Past research [38] used comments on rumors that included a URL to fact-checking websites to study the spread of misinformation and counter-misinformation. However, very few posts receive such comments, while most countering is done without using URLs. Thus, this approach is limited in scale. Further, [2] used six fact-checking

websites to determine whether a tweet containing news information is true or false by text comparison and analyzed their spread. However, fact-checks were not examined. Additionally, some works [44, 45] studied the assessment differences when different people judge the credibility of articles. Meanwhile, other works [46, 6] studied the spreading patterns of tweets that contain false rumors and confirmed news, using URLs and news content.

Different from previous works, our research investigates the broad information ecosystem which consists of misinformation and its rebuttals. These rebuttals can come from crowds, through opinion-based or evidence-based messages [13]. Recent studies have shown the remarkable effectiveness of correction by crowds by conducting experiments via interviews [47, 48, 49], surveys [50, 48], and in-lab experiments [49]. This correction has been shown to be as effective as professional correction [51], curbs misinformation spread [38, 52, 53], and works across topics [42, 54, 55, 56, 57, 58], platforms and demographics [55, 59, 60, 61]. Notably, users' polite and evidenced responses that refute misinformation are shown to effectively counter misinformation and reduce the belief in misinformation [62, 63, 64, 51, 65, 65, 66, 67]. Users correct others, typically friends [68], owing to a sense of social duty [38, 50, 69, 70, 71], anger, or guilt [72]. These works provide considerable evidence that correction by ordinary users is effective when countering misinformation and mitigating the spread of misinformation. On the other hand, considering the limited capability of professional fact-checkers, the large number of ordinary users and their efforts in social correction show great potential for a scalable solution to countering misinformation. Emerging research has analyzed the role that crowds play in countering misinformation. Twitter's Birdwatch [73] is a platform that allows users to report and flag misinformation. Studies have analyzed the data from Twitter Birdwatch [73, 74, 75, 76], which have shown how users actively engage to identify tweets that they believe are misleading and provide contextual notes to debunk them. Users have different levels of debunking capability. However, Birdwatch only allows users to flag misinformation and does not allow user-to-user communication and countering of misinformation on

Twitter. Thus, user flagging behavior within the Birdwatch ecosystem is not representative of user behavior on the broader Twitter platform or on other social media platforms.

1.2.2 User Response to Counter-misinformation by Crowds

To correct misinformation, ordinary users can publish standalone counter-misinformation posts on social media platforms [13]. User responses to this kind of correction have been investigated [77, 78, 79]. For instance, [77] analyzes 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 that citing evidence helps [79]. [78] similarly investigate the sentiment in comments that respond to fact-checking posts. But, all these posts are from official fact-checking organization accounts [78], 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 [80]. These existing conventional four-class stance [77] or two-class sentiment [78] 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, [81] focus on labels as well as the associated fact-check text provided voluntarily by users within Twitter's Community Note system [73]. 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.

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 [14, 82]. Even if some researchers find the backfire effect among particular groups [83] and within certain time frames [84], many studies have failed to replicate the backfire effect [85, 86]. 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 [67, 87, 88, 89, 51, 38]. 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 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.

1.2.3 Counter-misinformation Reply Generation

To assist crowds in combating misinformation, one approach is to generate counter misinformation responses. This task is related to fact-check generation in the existing literature [90]. The goal of fact-check generation methods [90, 91] is to respond to misinformation with a fact-checking URL. However, we consider a broader task of counter-response generation where the response text has to be generated. Existing works [90, 91] consider any post with a fact-checking URL from fact-checking websites (e.g., Snopes.com) as a fact-checking response, which is an inaccurate assumption – a fact-checking URL can be present to ridicule or oppose the fact-check [92] and can be taken out of context [92]. Importantly, only 1 out of 3 users use URL evidence when correcting misinformation [13]

and YouTube is the most frequently used URL, instead of fact-checking URLs [92]; consequently, studies relying only on fact-checking URLs are limited in their scope and do not learn from the majority of user-generated corrective posts.

Counter-hate [93, 94, 95, 96] and counter-argument [97, 98, 99] text generation tasks are also related to our problem setting, where the generated text is aimed to refute the original post spreading hate and any generic argument, respectively. Some proposed models fine-tune large scale unsupervised language models on the hate-speech or argument text for text generation [93, 100]. Other models first generate a set of candidate counter-hate/counter-argument replies, and then select one based on the relevance to the original post in a generate-then-retrieve or identify-substitute manner [94, 98, 99]. Meanwhile, some related counter-hate/counter-argument datasets have also been released [101, 102, 98]. However, it should be noted that compared to counter-misinformation response generation, the task of counter-hate generation does not necessitate responses to be evidence-based. Similarly, the counter-argument generation is a generic task (e.g., arguing whether immigration is good) and is not specific to misinformation. Additionally, large annotated and curated datasets exist for counter-hate and counter-argument [101, 102], which is not the case for counter-misinformation generation. To fill these gaps, we both curate two novel datasets and propose a counter-misinformation generator which can refute misinformation while being polite and providing evidence.

1.2.4 Deep Sequence Embedding-based Classification Models

To determine whether a user is malicious, existing methods usually focus on building deep sequence embedding models to encode the sequential information and use the embedding for downstream applications [103, 104, 105, 106]. For example, Facebook creates a temporal embedding from users' sequence of posts, then predicts users' dynamic embedding when users write a new post, and finally uses these embeddings for fake account detection [25]. However, vulnerabilities of these deep user sequence embedding-based classi-

classification models have not been explored. On the other hand, many works formulate classification of the sequence of a user's posts as a sequential text classification [105, 107] or document classification problem [107]. Researchers first utilize Convolutional Neural Network (CNN) and RNN to capture the sequential reliance between text posts and encode the text features for detection [105], and later turn to transformer-based [108] architectures due to the superior performance [109, 110]. However, these sequential text classification models can be vulnerable to adversarial attacks, which are relatively unexplored.

1.2.5 Attack and Defense on Deep Sequence Embedding-based Classification Models

Generating adversarial text to attack text classifiers is an important task due to its contribution to model robustness [111]. These methods can mainly be grouped into two categories: (1) *Modification-based attacks*: these approaches mainly make minor modifications to existing text to generate new text. Modifications include changing or adding characters, words, or phrases [112, 113, 114, 115]. However, these models have various shortcomings: they are incapable of fully leveraging a user's rich history of posts, they can not generate original content, and their modifications can be easily detected by finding misspelled words and improperly manipulated sentences [116]. (2) *Generation-based attacks*: these methods (e.g., TextGAN [117]) generate a new piece of text to achieve the attack goal. An attack model called Malcom [118] generates new fake reply comments to news articles to fool detectors. This model achieves high success in fooling the detector. However, these attack models have some shortcomings: they are not designed to leverage a user's rich history of posts and the generated text is not personalized to the user. Some researchers use the sequence of posts to generate adversarial text for attack [119], but the posts are from different users and the attack target is the recommendation system rather than deep sequence embedding-based classifiers.

To defend machine learning models against adversarial attacks, the prevalent strategy involves employing min-max optimization. This approach aims to minimize the maximum

adversarial loss, which represents the worst-case scenario, by computing it with adversarial examples to bolster the robustness of deep learning models [120]. Such a method has been a cornerstone in enhancing model resilience across various domains, including computer vision and natural language processing (NLP). In the field of computer vision, adversarial examples are typically generated using techniques such as the Fast Gradient Sign Method [121], Projected Gradient Descent [122], or through the use of Generative Adversarial Networks [123]. These methods have been effective in altering classification outcomes by introducing subtle, often imperceptible changes to the input data. Conversely, in NLP, adversarial examples are crafted through various means, including the replacement of characters or words in the input text, or by adding noise to input token embeddings [124, 125]. These techniques aim to introduce or modify input data in a manner that can be learned by the models to improve the robustness. Despite these advancements, directly applying these adversarial defense methods to deep sequence embedding-based classifiers poses a challenge. This is because these classifiers operate on sequential data, where attackers meticulously design the subsequent entity in the sequence, rather than merely modifying existing elements. This distinct nature of sequential data necessitates the development of specialized defense mechanisms tailored to protect against attacks specifically designed for sequence-based models.

CHAPTER 2

CHARACTERIZING CROWD-GENERATED COUNTER-MISINFORMATION

2.1 Introduction¹

To combat misinformation, professional fact checkers play an important role in controlling the spread of misinformation on online platforms [126]. During the COVID-19 infodemic, the International Fact Checking Network (IFCN) verified over 6,800 false claims related to the pandemic until May 20, 2020. Social media platforms use these fact checks to flag and sometimes remove misinformation content. However, false information still prevails on social platforms because the ability of fact checking organizations to use social media to disseminate their work can be limited [127]. For example, on Facebook, content from the top 10 websites spreading health misinformation had almost four times as many estimated views as equivalent content from reputable organizations (e.g., CDC, WHO).²

In addition to professional fact checkers, ordinary citizens (a.k.a non-expert ordinary users, or crowds), who are concerned about misinformation, can play a crucial role in organically curbing its spread and impact. Compared to professional fact checkers, crowds, who are users of the platform where misinformation appears, have the ability to directly engage with people who propagate false claims either because of ignorance or for a malicious purpose. They can back up their arguments using professional fact checks and trusted sources, whenever available. The cohort of ordinary citizens is also commonly referred to as *crowd*. Thus, the role of crowds or citizens who are concerned about misinformation can be critically important. The goal of this work is to study the nature and extent of the role that concerned citizens play in responding to misinformation.

We use a broad definition of *misinformation* which includes falsehoods, inaccuracies,

¹This chapter is based on the paper published in IEEE BigData 2021 [13]

²https://secure.avaaz.org/campaign/en/facebook_threat_health/

rumors, decontextualized truths, or misleading leaps of logic, all regardless of the intention of the spreader [128, 11]. In this chapter, we focus on COVID-19 related misinformation on Twitter and utilize a data-driven approach to investigate how fact checks and other organic user responses attempt to refute and counter it. We explore two popular misinformation topics: *fake cures* and *5G conspiracy theories* [129]. Fake cures for COVID-19 included drinking water, eating garlic or ginger, and salt, and 5G conspiracy theories state that 5G technology is responsible for the spread of COVID-19 or that COVID-19 does not exist and people are getting sick due to 5G radiations.

A study of fact checks and concerned citizen responses to misinformation presents several challenges. These include collection of data relevant to misinformation topics and development of a classifier for automated detection of tweets that contain misinformation or responses that refute them. Analysis of tweets that counter misinformation is also necessary to gain insights into how misinformation is refuted. We address these questions and make the following contributions.

- We create a dataset³ of tweets related to misinformation and responses that counter it, by first collecting 6,840 fact checking stories from 95 fact checking websites. We focused on two most popular COVID-19 topics and collected 155,468 tweets from 105,772 users over a period of 4 months.
- We develop a text-based classifier for tweets that represent misinformation and their rebuttals. To seed the classifier, two researchers hand-labeled 4,800 tweets into misinformation, rebuttal and other categories. The classifier created from the labeled data achieves an F1 score that is between 0.7 and 0.8, identifying 33,237 misinformation claims, 33,413 refutations, and 87,377 irrelevant tweets.
- Our analysis of the dataset revealed two types of users responsible for refuting misinformation tweets: professional fact-checkers, i.e., people or organizations who sys-

³http://claws.cc.gatech.edu/covid_counter_misinformation.html

tematically fact check a claim, and concerned citizens, i.e., the crowd that organically counters false claims. We find that 96% of refuting tweets are generated by concerned citizens, which highlights their importance.

- Our study reveals that professional fact-checks are not only low in tweet volume, but also receive a significantly lower number of retweets. Also, we find that 67% of tweets that counter misinformation do not contain URL-based evidence such as fact check websites.
- Our analysis of the methods used for countering misinformation reveals two broad classes. One class of tweets is evidence-based, which includes links to fact checks or other trusted sources. The other class is opinion-based and does not include verifiable evidence. We find that opinion based tweets include more assertive words, use more negative words, are more abusive, and exhibit negative emotions and anger.
- Finally, we find that citizen responses come from older and more reputed accounts, while misinformation tends to be spread by more recently created accounts. This could potentially be due to the orchestrated nature of some misinformation campaigns.
- Reproducibility: Our code and datasets are available for download at http://claws.cc.gatech.edu/covid_counter_misinformation.html.

Overall, our findings show that countering of misinformation cannot be studied by considering only fact checks. Our work reveals interesting insights and can potentially lead to a better understanding of how best to leverage the organic countering of misinformation to support and amplify the results of work done by fact-checkers. It could also help in the development of tools and mechanisms that can empower concerned citizens to combat misinformation.

2.2 Data Curation

We collected data on two of the most popular COVID-19 misinformation topics: *fake cures* and *5G conspiracy theories* [129]. Below we describe our process of selecting these topics and finding misinformation and counter-misinformation tweets on these topics.

IFCN Dataset

To have a verified and complete list of COVID-19-related misconception statements, we built a misconception dataset by leveraging the work of fact checkers. Specifically, first, we extracted 6,840 false statements fact-checked by IFCN’s CoronaVirusFacts/DatosCoronaVirus alliance.⁴ From these statements, we selected English statements related to two widely fact-checked misinformation topics [129]: (1) *Fake cures* that include drinking water, eating garlic or ginger, and salt. e.g., “the coronavirus can be cured with a bowl water with ginger”; (2) *5G conspiracy theories* that state 5G technology is responsible for the spread of COVID-19 or that COVID-19 does not exist and people are getting sick due to 5G radiations. Our final IFCN dataset consisted of 57 claims associated with fake cures and 32 claims related to 5G.

COVID-19 Tweet Dataset

Similar to [130], we utilized a keyword-based method to collect COVID-19-related tweets. After generating the IFCN dataset, we identify a collection of keywords related to the entities mentioned in the data. These keywords belong to two sets:

- *COVID-19 keywords*: this set includes widely-used terms associated with COVID-19, such as: COVID-19, covid, corona virus, coronavirus [130].
- *Misconception entity keywords*: 5G, drinking water, ginger, garlic and salt. This set was extracted from widely mentioned terms in IFCN misconception statements related to our selected topics.

⁴<https://www.poynter.org/ifcn-covid-19-misinformation/>

Next, we created a COVID-19 tweet dataset by combining data from multiple sources. First, we retrieved tweets from a publicly available dataset [130], which uses Twitter’s Streaming API and selected tweets containing the aforementioned keywords. Since the streaming API only collects 1% of the tweets,⁵ we used Twitter Search API⁶ to augment our dataset by querying the aforementioned keywords.

Together, these data sources returned around 8 million tweets spanning 4 months from January 21, 2020 to May 20, 2020. Similarly to other research, to simplify our analysis, we filter out non-English tweets [131] and retweets. Finally, we get 155,468 tweets from 105,772 users for the selected two topics.

Professional Fact-Checking Tweet Dataset

In addition to the above tweets, we also collected tweets from fact-checking organizations’ Twitter accounts. We manually compiled a list of Twitter handles of the organizations mentioned in the IFCN dataset. Then we used Python’s Tweepy API⁷ to extract tweets posted between January 21, 2020 and May 20, 2020. Later, we manually filtered these tweets by keeping only those tweets that contained an explicit link to one of the 89 claims listed in the IFCN dataset. This data served as our professional fact-checking tweet dataset.

Hand-labeled Annotation

Traditional methods usually use links to low-credibility sources to label tweets as misinformation [132]. However, in our COVID-19 tweet dataset, only around 10% of the tweets have links. Since there are no ground-truth labels for the remaining 90%, we manually label a subset of tweets based on tweet content and create a textual classifier to label the entire dataset.

⁵<https://developer.twitter.com/en/docs/labs/sampled-stream/overview>

⁶<https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets>

⁷<https://www.tweepy.org/>

Label Definition

We provide the definitions of the labels to make the annotation more clear. Based on previous misinformation research [128, 11], we define three labels:

- **Misinformation Tweets:** Since in this research we are interested in identifying a broader set of COVID-19 misinformation tweets, we consider misinformation tweets to include falsehoods, inaccuracies, rumors, decontextualized truths, or misleading leaps of logic, all regardless of the intention of the spreader. See Tweets 1 & 2 in Table 2.1.

- **Counter-Misinformation Tweets:** Tweets in this category refute false claims and can be further categorized in one of the following two groups:.

- **Professional Fact-check Tweets:** These counter-misinformation tweets are generated by professional fact-checkers and were extracted from the professional fact-checking tweet dataset. See Tweet 3 in Table 2.1.

- **Concerned Citizen Tweets:** Tweets that counter, question, or refute misinformation and are not generated by a professional fact-checking organization. See Tweets 4, 5 & 6 in Table 2.1. In the analysis we further categorize these tweets into: (a) *Opinion-based citizen responses* and (b) *Evidence-based citizen responses*. The aim is to distinguish between tweets that do not use any checkable evidence and those that link to some evidence.

- **Irrelevant Tweets:** Any tweet that does not include misinformation or counter-misinformation. The content may still discuss COVID-19 . See tweet 7 in Table 2.1.

Annotation Process

The annotation task aims to generate ground-truth for tweets by categorizing them as misinformation, counter-misinformation, or irrelevant. The set of professional fact-checking tweets is deterministic and rule-based, so it does not require hand-labeling or building a classifier.

To hand-label tweets, two authors annotated 4,800 tweets (specifically, 2,400 tweets for each topic with 800 tweets for misinformation tweets, counter-misinformation tweets, and irrelevant tweets). To assess agreement levels, a random sample of 300 tweets was annotated by both annotators. The inter-rater agreement measured by Kappa score [133] between the two annotators was 0.782, which shows substantial agreement.

Annotation process: This was a time consuming and difficult process because social media data is noisy and this process can be subjective. Hence, we conducted this process over several iterations. After each iteration a meeting was conducted to discuss dubious tweets, following the setup used by other research [134, 135]. Specifically, to annotate the tweets we use the following process. We start by asking the question: *Does this tweet refer to a misconception statement that was verified by IFCN’s fact-checking alliance?* If the answer is yes, we determine whether the tweet is supporting the misconception claim or countering it. For instance, Tweet 1 in Table 2.1 supports the verified false claim which states that garlic could prevent COVID-19, so we annotate it as misinformation tweet. When a tweet counters a false claim, such as when people categorically state that garlic has no effect on coronavirus (see Tweet 4 in Table 2.1), we annotate it as a counter-misinformation tweet. If the tweet is neither misinformation nor counter-misinformation, such as Tweet 7 in Table 2.1, which is just talking about the surging increase in prices of garlic, we annotate the tweet as an irrelevant tweet.

Labeling was not always straightforward. For instance, there were tweets which did not refer to any of the false claims that were verified by IFCN. We did not discard these tweets because they could still spread or counter misinformation [136]. If those tweets included falsehoods, inaccuracies, rumors, decontextualized truths and misleading leaps of logic, we still labeled them as misinformation tweets. For example, Tweet 2 in Table 2.1 seems to be spreading a questionable story about the installation of 5G antennas, so we annotated it as misinformation. If the tweet is judged to be countering a questionable story (see Tweet 5 in Table 2.1), we annotated it as a counter-misinformation tweet.

Table 2.1: Examples of tweets and assigned labels in misinformation and counter-misinformation studies.

No	Tweet	Topic	Label	Reason
1	Garlic onion and ginger are very nutritious, which has great potential to prevent corona virus.	Fake cures	Misinformation	Misinformation probability = 0.79; counter-misinformation probability = 0.01; and irrelevant probability = 0.06.
2	Corona virus is a way for corporations to install 5G without us being around #theory #coronavirus	5G	Misinformation	Misinformation probability = 0.63; counter-misinformation probability = 0.12; and irrelevant probability = 0.03.
3	A conspiracy theory falsely linking 5G to the coronavirus is getting traction on social media. https://www.politifact.com/factchecks/2020/apr/03...	5G	Professional fact-check	The tweet contains the link mentioned in the IFCN dataset.
4	@X Major Newspaper had to put this out... Garlic: No effect on coronavirus. Sesame oil on the body: No effect on coronavirus. Herbal remedies: No effect on coronavirus. Smoking: No effect on coronavirus.	Fake cures	Opinion-based Citizen Response	Counter-misinformation probability = 0.69; misinformation probability = 0.27; irrelevant probability = 0.07; and no external link.
5	how do some people seriously believe that coronavirus is not a virus and it's the government trying to kill us with 5g network	5G	Opinion-based Citizen Response	Counter-misinformation probability = 0.77; misinformation probability = 0.13; irrelevant probability = 0.06; and no external link.
6	Garlic won't keep the coronavirus at bay. Neither will saltwater gargling or cow dung https://www.scmp.com/week-asia/health-environment/article/3049261/garlic-cant-keep...	Fake cures	Evidence-based Citizen Response	Counter-misinformation probability = 0.97; misinformation probability = 0.06; irrelevant probability = 0.01 and has an external link with evidence.
7	#Indonesia is grappling with surging garlic prices as the fast-spreading COVID spurs fears over supply disrupt in #China.	Fake cures	Irrelevant	Irrelevant probability = 0.96; misinformation probability = 0.02; and counter-misinformation probability = 0.06.

Misinformation and Counter-Misinformation Tweet Classifier

In this section, we describe the text-based classifier that we developed to classify the tweets in our COVID-19 tweet dataset as misinformation, counter-misinformation, or irrelevant tweets. Specifically, we first build the tweet representation for each tweet and then use the hand-labeled tweets to train the text-based classifier.

Tweet Representation

Embedding-based Representation We embed each tweet by the popular BERT text-embedding model [137]) to capture the semantic meaning of tweets. We selected BERT embedding because it obtained better performance than other models in our preliminary testing (e.g., GloVe [138]). We first remove all Twitter- and web-specific content such as URLs, usernames, hashtags and emojis, and then feed the remaining tweet text into the

BERT model. After this processing, we select the final hidden layer result in BERT model as the resulting 768-dimensional tweet representation.

Hashtag-based Representation Different from the embedding-based approach, in the tweet text, since people usually use hashtags (e.g., #5G) to highlight important information, we used hashtags as information indicators to represent tweets. The hashtag-based representation counts the number of occurrences of each hashtag in the tweet text and selects the widely-mentioned hashtags in the dataset. In total, 768 hashtags were selected by popularity (i.e., their number of occurrence), keeping the same 768 dimension as that of the aforementioned embedding-based representation to have a fair comparison. We also varied the number of hashtags used (from 100 to 1,000), which achieved a comparable performance.

Classifier Creation Given this standard three-class classification task, i.e., misinformation vs. counter-misinformation vs. irrelevant tweets, we trained three separate one-vs-all Logistic Regression classifiers. Note that we can also plug in any other classification model, such as SVM. Our empirical testing showed that the results of other models such as SVM were similar to the logistic regression results. Due to the relatively small dataset, we utilized the embedding-based BERT model rather than the fine-tuning-based BERT model or other advanced deep neural network models [137]) to avoid the possible over fitting problem. Each one-vs-all classifier is trained with hashtag-based, embedding-based and hashtag+embedding features, where the two feature vectors were concatenated. We conducted five-fold cross-validation on the manually-annotated tweets. The performance of the models was measured using precision, recall and F1 scores (see Table 2.2 and Table 2.3). We report these scores together with the corresponding standard deviations.

As observed from Table 2.2 and Table 2.3, by using the tweet embedding feature, we achieved the best performance with respect to precision, recall and F1 score. We noticed that the decreased performance in the combination of hashtag and embedding represen-

Table 2.2: Misinformation and counter-misinformation tweet classification performance for 5G topic.

Feature	Precision (σ)	Recall (σ)	F1 score (σ)
Misinformation Tweet Detection			
Hashtag	0.386±0.01	0.859±0.02	0.533±0.01
Embedding	0.763±0.03	0.788±0.02	0.775±0.02
Hashtag+Embedding	0.758±0.02	0.782±0.02	0.770±0.02
Counter-Misinformation Tweet Detection			
Hashtag	0.626±0.06	0.220±0.04	0.325±0.05
Embedding	0.807±0.05	0.823±0.01	0.814±0.02
Hashtag+Embedding	0.800±0.05	0.816±0.01	0.807±0.02
Irrelevant Tweet Detection			
Hashtag	0.608±0.03	0.266±0.02	0.369±0.02
Embedding	0.845±0.02	0.795±0.05	0.818±0.03
Hashtag+Embedding	0.834±0.02	0.786±0.05	0.809±0.03

Table 2.3: Misinformation and counter-misinformation tweet classification performance for fake cures topic.

Feature	Precision (σ)	Recall (σ)	F1 score (σ)
Misinformation Tweet Detection			
Hashtag	0.381±0.01	0.785±0.05	0.513±0.02
Embedding	0.698±0.03	0.693±0.04	0.694±0.03
Hashtag+Embedding	0.691±0.02	0.687±0.04	0.688±0.02
Counter-Misinformation Tweet Detection			
Hashtag	0.563±0.09	0.178±0.02	0.270±0.02
Embedding	0.760±0.05	0.759±0.05	0.758±0.03
Hashtag+Embedding	0.748±0.05	0.744±0.04	0.744±0.03
Irrelevant Tweet Detection			
Hashtag	0.475±0.02	0.295±0.03	0.363±0.03
Embedding	0.677±0.03	0.678±0.05	0.676±0.03
Hashtag+Embedding	0.675±0.03	0.678±0.04	0.676±0.03

tations was achieved when unrelated features were added to the tweets. Therefore, we used embedding feature together with logistic regression classifier to label all tweets. To achieve high-confidence labelling results, we set a very strict rule, which states that if a tweet is classified as misinformation tweet, the score needs to be higher than 0.5 and less than 0.5 for the other models (i.e., countering-misinformation and irrelevant). We used the same rule to detect counter-misinformation and irrelevant tweets. We use the 0.5 threshold because this value is widely-used in binary classification tasks [139].

Applying the classifier as described above returns a dataset containing misinformation, counter-misinformation, and irrelevant tweets, as shown in the table 6.2. Among counter-misinformation, we find 96% are sent by crowds.

Table 2.4: Misinformation and counter-misinformation twitter Data statistics

	Count
Total number of tweets	155,468
Number of Misinformation Tweets	33,237
Number of Counter-misinformation Tweets	34,854
Number of Irrelevant Tweets	87,377

2.3 Counter-misinformation Characterization

Types of Counter-Misinformation Tweets One of the aims of this work is to understand whether citizens use professional fact-checks as evidence for countering misinformation. We find that 67% of citizen responses do not contain any evidence to refute misinformation and we refer to these tweets as ‘*opinion-based citizen responses*’. On the other hand, only 33% of citizen response tweets have URLs, out of which only 4% point to fact-checking websites, showing the low use of the work done by professional fact checkers. To check where these URLs refer to, we randomly sampled 300 of these tweets and manually verified their URL. We found that 204 of them provided reliable evidence, of which 96 contained links to high-credibility sources, and the rest contained links to articles, videos, or websites that directly refute the misinformation. Similar to Chen et al. [140], we defined high-

credibility and low-credibility sources by compiling a list from several recent research papers [141, 142]. Since the proportion of links pointing to reliable evidence is very high, we label all counter-misinformation tweets with URLs as ‘*evidence-based citizen responses*’.

The frequency of high-credibility sources in counter-misinformation citizen responses is evidenced in Figure 2.1, which shows that 7 of the Top 20 most frequent sources are known high credibility. With respect to misinformation, we find that 5 of the Top 20 most-frequent sources are listed as low-credibility (see Figure 2.1). This finding confirms previous work which showed a strong correlation between misinformation and low-credibility sources [141].

In summary, in this section, we find that while concerned citizens use URLs to refute misinformation often, professional fact-checks are rarely used as evidence. Research needs to investigate techniques that would make fact-checks more accessible to concerned citizens.

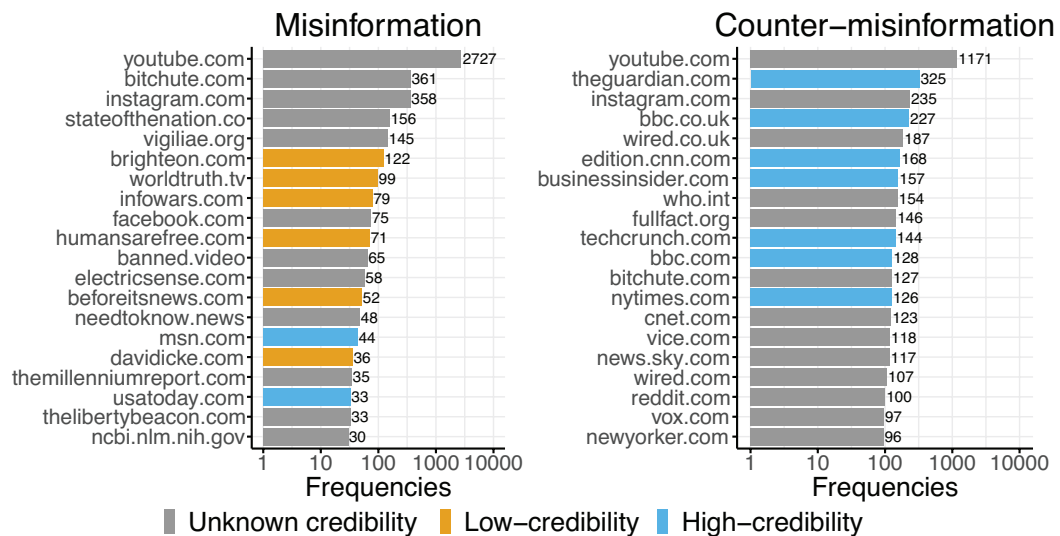


Figure 2.1: Most frequent external sources in misinformation and counter-misinformation tweets. High-credibility sources are more common in counter-misinformation tweets.

Characterizing the spread of Counter-Misinformation To get a better understanding of the extent to which misinformation is being countered, we investigate tweet volume, engagement, and spread.

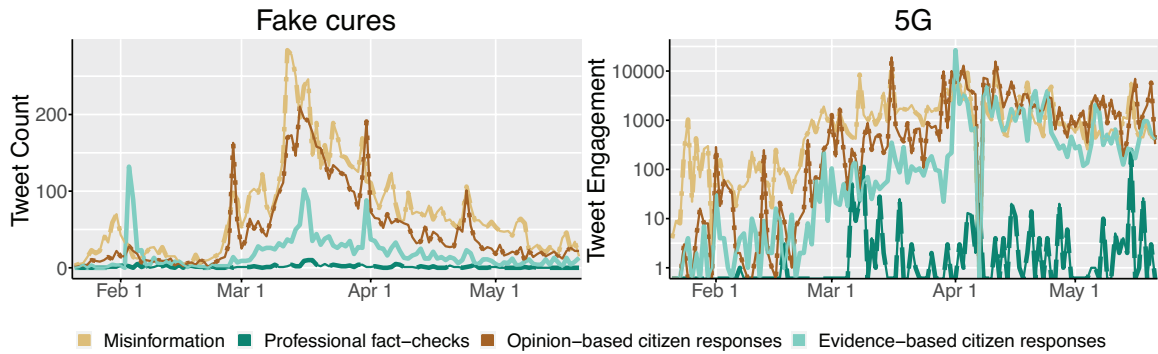


Figure 2.2: Tweet volume and engagement of misinformation and counter-misinformation over time. Both volume and engagement of countering misinformation are comparable to misinformation.

Misinformation has higher volume Figure 2.2 (left) shows the volume in terms of number of tweets of fake cures topic over the chosen four-month period. To provide an accurate representation of tweet volume and engagement, we selected unique tweets and discarded retweets. First, we note that for both topics, the volume of professional fact-checks is always significantly lower than that of the other groups. Second, we find that for 5G, the volume of misinformation tweets is comparable to the volume of counter-misinformation tweets (Wilcoxon pairwise comparison $p > 0.05$). Next, we note that among concerned citizens, the volume of tweets that counter misinformation without using professional fact-checking URLs as evidence is several times higher than those that use evidence. These findings show that there is a comparable amount of people that are countering misinformation and that tweets by professional fact-checkers are still not popular across the Twitter community.

Finally, an important observation is that there is a strong positive correlation between misinformation and counter-misinformation tweet counts, both for fake cures (Spearman correlation $\rho = 0.88$, $p < 0.001$) and 5G ($\rho = 0.85$, $p < 0.001$). Importantly, for both topics, we found a lower, but still positive, correlation between misinformation and fact-checks (fake cures: $\rho = 0.45$, $p < 0.001$, 5G: $\rho = 0.52$, $p < 0.001$). These findings show that misinformation is being countered at the same rate as it is being spread, similar to the findings made in Mendoza et al. [46].

Counter-misinformation and misinformation attract similar tweet engagement

We investigate engagement volume because engaging with a tweet increases its visibility. Tweet engagement is defined as liking or favoriting the tweet and sharing a tweet with or without an additional comment. Figure 2.2 (right) shows the engagement results for the 5G topic. We used Wilcoxon pairwise test to compare the differences across the types of information. First, we note that the engagement of misinformation and counter-misinformation tweets is similar ($p > 0.05$). This means that concerned citizens are obtaining a comparable level of engagement as users that spread misinformation. This is not the case with professional fact-checkers because in all cases they obtained significantly lower amount of engagement than misinformation and counter-misinformation. This indicates that work is needed to increase the visibility of professional fact-checking efforts. Finally, when restricting our analysis to only retweets, we find that most of the tweets received 0 retweets (71% - 80%), while only few tweets were retweeted more than 100 times (0.2% - 0.6%). This finding shows that few counter-misinformation and misinformation tweets get propagated through the Twitter community.

Overall, in this section, we find that concerned citizens are producing a similar volume of tweets as misinformation spreaders. Similarly, engagement levels are comparable and most tweets do not receive more than 100 retweets. Moreover, for 5G, counter-misinformation tweets receive more engagement than misinformation tweets. Professional fact-checks receive a significantly lower volume of engagement than misinformation and concerned citizen tweets.

Characterizing Textual Properties of Counter-Misinformation In this section, we analyze the linguistic and textual properties of misinformation, counter-misinformation, and fact-checks. We investigate their sentiment, politeness, word use and LIWC characteristics. This investigation will provide insights on how linguistic and textual properties could be used to detect and counter misinformation.

Sentiment and politeness of citizen responses

Here we analyze the sentiment and politeness [143] of counter-misinformation responses made by citizens. Prior research has also found it to be useful in detecting misinformation [144].

Figure 2.3 compares the distribution of sentiment and politeness across different types of content. For sentiment, we find that tweets by professional fact-checkers are neutral. On the other hand, tweets spreading misinformation about fake cures of COVID-19 are significantly positive, potentially hailing the ‘cure’ that works like a miracle. However, counter-misinformation tweets tend to be more neutral.

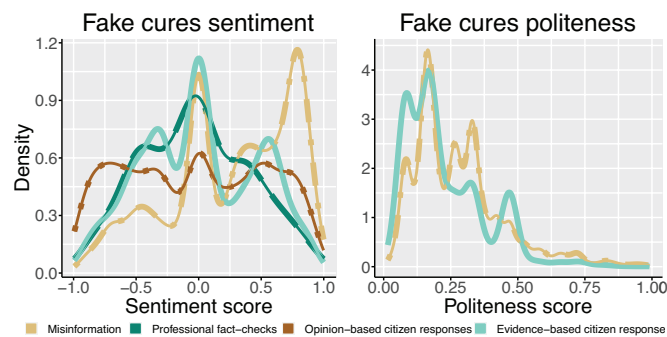


Figure 2.3: Sentiment and politeness of misinformation and counter-misinformation for fake cures. Misinformation tweets are more positive and professional fact-checks are neutral (left). Evidence-based citizen responses are less polite than misinformation (right).

As seen in Figure 2.3, with regards to politeness, we find that misinformation tweets are significantly more polite than countering-misinformation tweets ($p < 0.001$). Surprisingly, professional fact-checks are quantified as less polite as well. This is because these tweets do not contain much text, and the formal and restrained nature of these tweets could be leading to this outcome [145]. Previous research [146] defined polite text to be direct, clear, unambiguous and concise [146]. This means that misinformation tweets seem to be exhibiting more of these characteristics than the other tweets.

Digging deeper into the citizen responses to fake cures misinformation, we find that evidence-based citizen responses are very similar to professional fact-checking tweets in terms of sentiment, most of them being neutral. On the other hand, opinion-based citizen responses tend to express more sentiment, exhibiting a uniform distribution through the

spectrum of sentiment scores.

The main outcome of these findings is that COVID-19 misinformation is mostly positive, this contrasts with previous research which showed that misinformation is mostly negative [147]. This finding is further evidence that misinformation spreaders are continuously changing their tactics [11]. The lack of politeness, positive and negative sentiments in professional fact-checking tweets is also an interesting finding. This finding requires further research to investigate whether manipulating these properties could lead to higher volume and engagement.

Psycholinguistic Characteristics of Citizen Responses We analyze textual properties using LIWC dictionary [148], which accounts for the psycholinguistic properties of the text.

First, opinion-based citizen response tweets (see Figure 2.4) are significantly more self centered, use ‘you’ and third person pronouns (i.e., ‘he’ or ‘she’) more than misinformation and evidence-based responses ($p < 0.001$). We also find that opinion-based response tweets have a significantly higher amount of swear words than the other tweets ($p < 0.001$). Together, these two findings show that citizens are engaging in a more direct, confrontational, and argumentative conversation with misinformation spreaders. While swearing and confrontation used in countering misinformation contrasts with the findings from previous research [147], our analysis shows that it is a common technique used by citizen responders to call out misinformation spreaders [149].

Second, our analysis shows that evidence-based responses are quite similar to professional fact-checking tweets. For instance, we did not find significant differences when comparing all five pronouns ($p > 0.05$). We find similar values for impersonal pronouns, authentic words, dictionary words, words greater than 6 letters and tone. They have similar amount of words that contain discrepancy, tentative, achievement, power, reward, assent, fillers, netspeak and non-fluence. Also, we find that professional fact-checkers infrequently use personal and impersonal pronouns, especially the ‘i’ and ‘he’/‘she’.

Third, the textual properties of opinion-based citizen responses and misinformation tweets have highly similar levels of achievement, anxiety, positive emotions, netspeak and non-fluence (see Fig 2.4). Furthermore, they have similar amounts of adverbs, articles, auxiliary verbs, prepositions, dictionary words, function, long words, and tone. Opinion-based citizen responses have significantly more insight words, differentiation words, and tentative words ($p < 0.001$). On the other hand, misinformation tweets exhibit significantly more words related to cause ($p < 0.001$). For fake cures, counter-misinformation without evidence show significantly higher affiliation and tweets that counter-misinformation with evidence show significantly higher risk ($p < 0.001$).

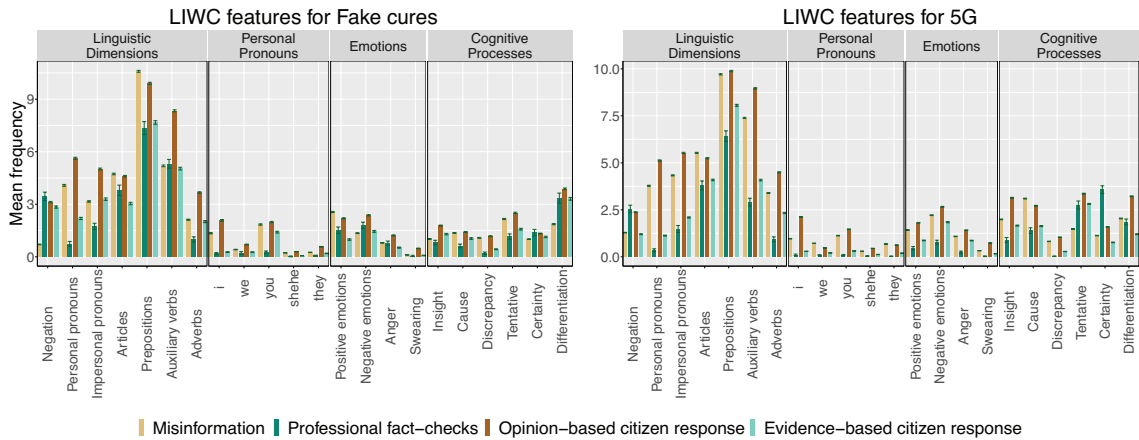


Figure 2.4: LIWC results of misinformation and counter-misinformation for fake cures (left) and 5G (right) tweets. Opinion-based citizen responses have similar linguistic properties as misinformation.

In summary, we find that opinion-based counter-misinformation portray insight, exhibit negative emotions and anger by using more negative words, including swear words. This implies that concerned citizens are engaging in a more direct, confrontational, and argumentative conversation and do not shy away from calling out misinformation spreaders [149].

User Characteristics of Concerned Citizens

To have a better understanding of concerned citizens who respond to misinformation we investigate their accounts' age, their popularity and whether they have verified accounts.

Misinformation spreaders have more recent accounts than concerned citizens First, we find that accounts that spread misinformation (fake cures: 5.7 years; 5G: 6 years) are significantly ($p < 0.001$) more recent than concerned citizens accounts (fake cures: 6.5 years; 5G: 6.8 years). Second, we find that professional fact-checking accounts (fake cures: 22%; 5G: 8%) and concerned citizens accounts (fake cures: 7% years; 5G: 4%) are significantly more likely ($p < 0.001$) to be verified accounts than misinformation (1%). These findings show that accounts that spread misinformation are typically more recent [127] and concerned citizens are more likely to have verified accounts than misinformation spreaders.

Concerned citizens are more popular Account popularity is a relevant user characteristic because when a user posts a tweet, it becomes visible to all of their followers. We find that users that spread misinformation have significantly ($p < 0.001$) fewer followers (fake cures: 6,052 and 5G: 6,651) than concerned citizens (fake cures: 40,295; 5G: 22,133). Professionals that fact-checked fake cures had significantly more followers (= 85,586) than concerned citizens and misinformation spreaders. Contrarily, professionals that fact-checked 5G conspiracy theories had similar number of followers as misinformation spreaders (8,169 vs 6,651; $p > 0.05$) and significantly fewer followers than concerned citizens (8,169 vs 22,133; $p < 0.001$). These findings show that counter-misinformation has the potential of reaching a wider audience than misinformation. Despite receiving low volume and engagement, some professional fact-checking accounts have a considerably large audience.

In summary, in this section, we find that concerned citizens are more popular and have older accounts, indicating that counter-misinformation has the potential to reach more users and are more trusted. These accounts are more likely to have verified accounts than misinformation spreaders.

2.4 Conclusion

In this chapter, we focus on COVID-19 related misinformation on Twitter and utilize a data-driven approach to investigate how fact checks and other organic user responses attempt to refute such misinformation.

Overall, our work shows that countering of misinformation cannot be studied by only considering fact checks because 96% of all refutations are being done by concerned citizens (i.e., the crowd). Moreover, professional fact-checks are not only low in tweet volume, but also receive a significantly lower number of retweets as opposed to those received by the crowd. Our analysis also reveals that the crowd uses two methods to counter-misinformation. The first approach is evidence-based, which includes links to fact-checks or other trusted sources. The second is opinion-based, which does not include checkable evidence. We find that opinion-based tweets include more assertive language, use more negative words, are more abusive, and exhibit negative emotions and anger.

Our results reveal interesting insights and can potentially lead to a better understanding of how to leverage the organic countering of misinformation to support and amplify fact-checking efforts best. More importantly this work could also lead to development of tools and mechanisms that can empower concerned citizens to combat misinformation.

Finally, the work presented has a number of limitations that could be addressed in the future. First, we only focus on the Twitter platform. In future work, one can extend our analysis to different platforms, like WhatsApp, Facebook and Instagram, since previous research found that misinformation flows across multiple platforms [150]. Second, we only studied two misinformation topics. These two topics help us understand the role of fact checks and organic citizens, but there could be variations across areas that are targets of misinformation (e.g., climate change issues). In fact, we did find evidence of differences across topics which needs to be fully explored in the future.

CHAPTER 3

CHARACTERIZING AND PREDICTING USER RESPONSE TO CROWD-GENERATED COUNTER-MISINFORMATION REPLIES

3.1 Introduction¹

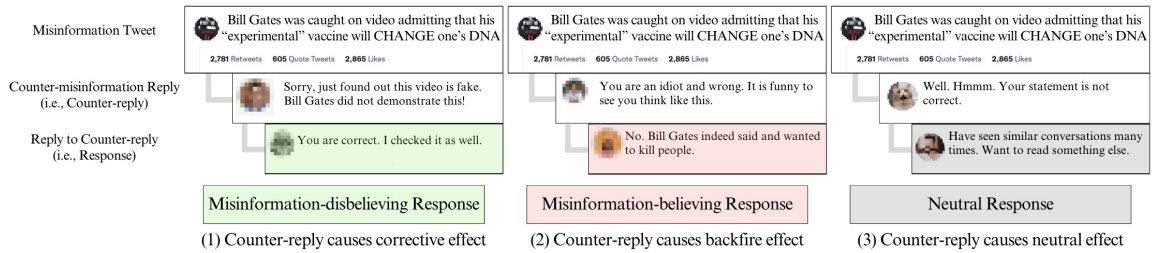


Figure 3.1: Examples of user responses to social correction. Here, the social correction is the counter-misinformation *reply* posted by ordinary users (the second row), and the user response is the *reply* to the counter-misinformation reply (the third row).

Even if we find that most of the counter-misinformation comes from crowds in the previous chapter, little is known about the real-world user response toward this correction. Understanding such responses is beneficial because 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) [67], 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) [152], 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 [153], 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.

¹This chapter is based on the paper published in ACM WebSci 2024 [151]

Despite its advantages, characterizing and predicting social correction is challenging because of multiple reasons. First, current research predominantly utilizes simulated experiments or user studies [47, 48, 49], approaches that may not accurately mirror real-world scenarios. Second, relevant research works and datasets [77, 78] do not contain conversational-style narratives with tripled misinformation posts, counter-replies, and responses, as shown in Figure 3.1. 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 can play 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 [77] or two-class sentiment [78] 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 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>.

3.2 Dataset

Definition

Misinformation Tweet We deploy a broad definition of misinformation which includes inaccuracies, falsehoods, rumors, or misleading leaps of logic [12]. Based on the existing work [154, 80], 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 [154, 155, 36]. A misinformation tweet is represented as m .

Counter-misinformation Reply (i.e., Counter-reply) Inspired by existing research works on social correction [80, 74], a direct reply to a misinformation tweet m is considered as a counter-misinformation reply (i.e., counter-reply as shown in Figure 3.1 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 [13, 156, 157, 80, 74], 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”).

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 3.1. 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 [79, 77, 156], we can group responses into three categories, as shown in Figure 3.1:

- 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.”).

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 $c = [c_1, c_2, \dots, c_n]$ posted in direct reply to m . Our goal is to build a classifier \mathcal{F} such that it can output a label $\mathcal{F}(c_i), i \in \{1, 2, \dots, n\}$, 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)?²

Dataset Curation

Misinformation Tweet Collection and Classification In our study, we followed an existing approach [80] and used the Anti-Vax dataset from [154], 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 3.2 and the current approach by [154]. Initially, 13,432 annotated tweets (4,836 misinformation, and 8,596 non-misinformation) were extracted from [154]. Using these tweets, we trained a BERT-based text classifier [137], 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.

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.

²Note 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 3.3; and (2) similar existing research works also do not emphasize this kind of dual labels [96].

To identify counter-replies, we follow the definition of counter-reply in Section 3.2 and build on existing works of counter-reply classification [80, 36]. Particularly, we first crawl a combined 2,479 annotated replies (1,425 counter-replies, and 1,054 non-counter-replies) from [80, 36]. Next, we train a RoBERTa-based lower-case counter-reply classifier [80] 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.

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 train a text-based classifier for annotation. Following the existing works of building tweet text classifiers [156], 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 3.2. 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)

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 3.2: Illustration of prompts used in GPT-4 annotation.

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.

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 [158, 159], especially the GPT-4 model [160]. Building on the existing research works regarding ChatGPT-based text annotation in computation social science domains [159], we adopt the well-performed few-shot in-context-learning diagram for GPT-4 annotation [158, 159]. 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 3.2. Then, we use this prompt to label our remaining annotated responses using a suggested moderate temperature of 0.5 in GPT-4 [158]. 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

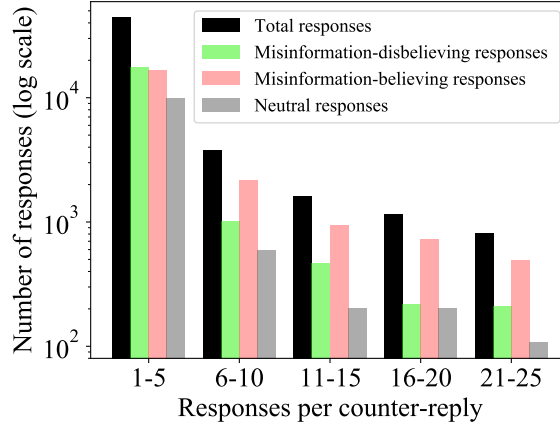


Figure 3.3: Distributions of the total number of responses (black), number of misinformation-disbelieving responses (green), number of misinformation-believing responses (red), and number of neutral responses (gray) per counter-reply, each presented on a log scale.

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 3.3.

3.3 User Response Characterization

Taxonomy of User Response

Different users respond differently to misinformation correction messages [77, 78]. 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 [161], we exhaustively analyze manually annotated responses and finally create the taxonomy of user responses in Table 3.1, presenting user actions employed to demonstrate the correspond-

Table 3.1: Taxonomy of user responses based on employed user actions within each response type.

Response type	User actions employed to demonstrate the corresponding response type	Examples
Ratio		
Misinformation-disbelieving response	Endorse those who counter-reply	“You are right”
	Confirm the counter-misinformation	“I checked the information as well and it is correct”
	Debunk or counter the misinformation again	“Again, the first tweet is misinformative and a bait!”
	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”
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”
	Reject the counter-reply	“What you said does not make sense to me. The reasoning in your reply is faulty.”
	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”
	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”
	Add new types of related misinformation	“Besides changing our DNA, the vaccine is actually developed by Bill Gates to depopulate the people. Take caution!”

ing response type. As we can see in Table 3.1, 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.

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 [80, 74] 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 [80]. 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:** Counterreplies 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 [96] 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 3.4.

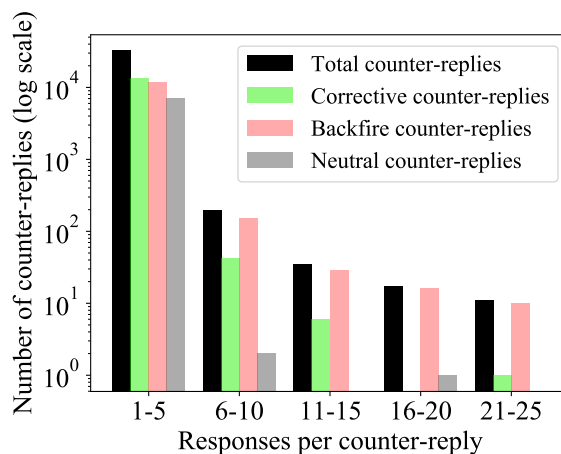


Figure 3.4: Distributions of the total number of counter-replies (black), number of corrective counter-replies (green), number of backfire counter-replies (red), and number of neutral replies (gray) based on response per counter-reply, each presented on a log scale.

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 [67, 162], we focus on these two kinds of counter-replies for the comparison analysis. Particularly, we build on related tweet analysis works [80] and analyze the linguistic, engagement, and poster attribute features [80] as well as the counter-misinformation property [36] 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 [36].

Table 3.2: List of linguistic, engagement, poster, and counter-misinformation property attributes for the counter-reply analysis.

Attribute type	List of attributes
Reply linguistic	Number of words in the reply. VADER [163] positive sentiment, negative sentiment, and compound sentiment of the reply. For each of the 65 dimensions of the LIWC [148] 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. Whether the replier is verified (1) or not (0). 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 [164]. 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 [36]. Evidence score of the reply, derived by checking the existence of high-credibility and fact-checking URLs in the reply [13].

Table 3.2 displays the full list of attributes we statistically study³ within each of these four categories.

Linguistic Attribute Analysis First, we find that corrective replies are 5.409% more positive ($p < 0.005$)⁴ 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

³This statistical test was performed using Welch’s unequal variances t -test between corrective and backfire counter-replies.

⁴All p -values are calculated using the Welch’s unequal variances t -tests.

words. No statistical significance is found between these two groups. After LIWC lexicon analysis [148], 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 [165]. Additionally, corrective replies mention more words related to family while backfire replies say more death-related emotions ($p < 0.05$).

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 [166], thus having corrective effects. In turn, we can also interpret that the misinformation-disbelieving responses make the corrective counter-reply more convincing, finally having more likes and retweets [167]. This mutually-reinforced effect demonstrates the importance of engagement attributes in the analysis.

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 ($p < 0.001$).

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 [36, 67]: politeness, evidence, and refutation.

Following the existing work [164, 80], 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 [36, 67]. Regarding evidence, we check the existence of high-credibility and fact-checking URLs in counter-replies, as suggested by [13]. 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 [67]. 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 [36] 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 $d = 0.202^5$). Even if higher refutation scores are expected in corrective replies [67], our result is still explainable considering when we add more refutation statements in replies, the emotions of some audience can be triggered [168]. 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.

3.4 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 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.

⁵Cohen's d here refers to the unweighted Cohen's d values.

Table 3.3: Classification performance of whether a counter-reply will have a corrective, backfire, or natural effect.

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

Dataset

To answer the above research question, we use the aforementioned dataset in Section 3.3. 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 3.3. The sizes of the three sets are 13, 482, 11, 893, and 7, 005.

Experiment Setup

After choosing the dataset, we follow similar approaches in tweet prediction tasks [169] by building a multi-class classifier. We utilize the set of attributes described in Section 3.3 as features. As shown in the existing tweet prediction work [13], the semantic information from textual embedding benefits the prediction task. Thus, we also generate the embedding vector for each reply using RoBERTa [170]. Finally, we concatenate the above feature vectors to form a reply feature vector to comprehensively represent the reply and use it for classification.

Classifier Creation and Evaluation

Following similar tweet or general text classification tasks [37], 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 3.3. 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 [13]. 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.

3.5 Conclusion

In this chapter, 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-performing 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.

CHAPTER 4

GENERATING COUNTER-MISINFORMATION TO EMPOWER CROWDS

4.1 Introduction¹

As we discussed in the previous chapters, non-expert social media users, i.e., ordinary users or crowd, act as eyes-on-the-ground who proactively question and counter misinformation, including emerging misinformation [13, 49, 171, 172, 57, 6, 74]. They complement fact checkers who can only verify a handful of stories after they have gone viral [21, 173]. We show that 96% of counter-misinformation responses are made by ordinary users, while professionals account for the rest [13].

Recent research on social correction [80], i.e., countering of misinformation claims by other social media users, has proven to be as effective as professional correction [51], curbs misinformation spread [38, 52, 53], and works across topics [42, 54, 55, 56, 57, 58], platforms and demographics [55, 59, 60, 61]. Corrections work [67, 87, 174, 88, 89] without causing an increase in misperception (i.e., the backfire effect has not been replicated) [85, 175, 176]. While corrections are not expected to convince everyone, they are most effective in reducing misinformation consumers' misperceptions [177, 52, 57, 53, 178]. Thus, empowering users to effectively correct misinformation promises a scalable solution towards information integrity. This solution is independent of, but complements, the efforts of social media platforms to detect misinformation via the crowd, e.g., Twitter Birdwatch [73].

Alarmingly, we find that linguistic analyses of in-the-wild crowd-generated counter-responses revealed that 2 out of 3 counter-misinformation posts are rude and do not use fact-checking evidence to support their counter-response [13]. Uncivil counter-responses

¹This chapter is based on the MisinfoCorrect [36] paper published in WWW 2023.

can lead to reduced trust in the correcting user [179, 180] and result in arguments [181, 182, 183]. This implies an urgent need to empower crowds so that they counter misinformation more effectively.

Thus, in this chapter, we seek to facilitate healthy misinformation correction by the crowd, which includes being objective, evidenced, and polite – properties that have been shown to be effective [64, 51]. To do so, we propose to create a counter-misinformation response generator, which generates desirable counter-response for a misinformation post (as illustrated in Figure 4.1). Our study is focused on countering misinformation on Twitter,

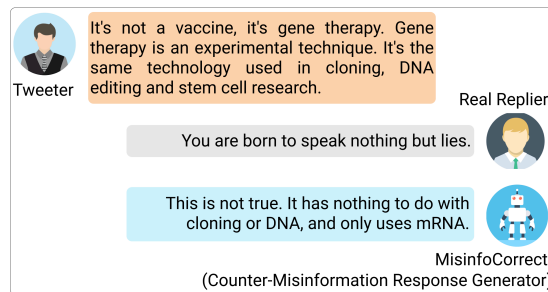


Figure 4.1: An overview of counter-misinformation response generation task in MisinfoCorrect.

given its prominence in the spread of online misinformation.

Generating effective counter-misinformation responses poses several challenges. First, there is no existing dataset containing pairs of annotated misinformation posts and counter responses. Second, there is no counter-misinformation response generator model. The closest research works in fact-checking generator [90] are non-conversational and related research in counter-hate speech/counter-argument generator [93, 94, 95, 97] do not apply directly since they are not evidence-based or not specific to misinformation. Third, counter-misinformation responses are effective if they have the following desirable properties: objective and evidenced [64, 51], makes rational arguments [65], refutes fallacy in reasoning [66], and polite [62, 63]. Off-the-shelf text generator models do not directly generate counter-responses with this desiderata. Four, bot-generated or template-based responses are not effective since they are non-personalized and non-contextualized with

respect to the false claims made in the misinformation post. Thus, the counter-response needs to be relevant to the misinformation post.

We propose to create two novel datasets containing misinformation and counter-responses (solution to challenge 1) – one collected from in-the-wild social media responses from Twitter and another created by crowd-sourcing from college students. We focus on four popular COVID-19 vaccine misinformation topics on Twitter (e.g., Bill Gates created vaccines to depopulate people [184, 185], and vaccines can cause infertility [155, 186], contain microchip [187], alter DNA [188, 189]). To create the in-the-wild dataset, for each misinformation topic, we collect all the replies to misinformation tweets identified in prior research [154]. We annotate associated replies to identify the responses that counter the tweet along with their textual attributes of refuting, politeness, and factuality. Finally, we have 754 misinformation tweet and countering response pairs. For the crowd-sourced response generation, we recruit and train 17 college students to write counter-misinformation replies when given misinformation posts. In total, we collect 591 crowdsourced replies.

Next, we propose a reinforcement learning-based framework, called MisinfoCorrect, that learns to generate counter-misinformation responses that are polite, evidenced, and refute misinformation (solutions to challenges 2 and 3). Specifically, this agent utilizes a policy network on a transformer-based language model adapted from GPT-2 [190]. During training, we reward the generation that increases the politeness and refutation attitude. Additionally, we ensure text fluency and relevancy to the misinformation post by adding fluency and relevance rewards in the reinforcement learning framework (solution to challenge 4).

MisinfoCorrect is evaluated against five representative baselines on the task of counter-misinformation response generation. Quantitative and qualitative experiments show that it can outperform the baselines by generating high-quality counter-responses.

To summarize, our contributions are as follows:

- We create two large novel and annotated datasets containing misinformation and counter-

response pairs from social media (in-the-wild) and generated via crowd-sourcing (in-lab). Together, both datasets contain 1,345 counter-misinformation responses.

- We propose a reinforcement learning based counter-response generation framework, where the counter-response is especially rewarded for being polite, evidenced, and refuting misinformation.
- Results on actual COVID-19 vaccine misinformation conversations show that the proposed model outperforms existing representative baselines.

The code and data is accessible on <https://github.com/claws-lab/MisinfoCorrect>.

4.2 Problem Definition

Given a misinformation post m , we aim to build a text generator g such that it can output counter-response $\hat{c} = g(m)$, which has certain desirable properties \mathcal{P} .

The **desirable properties** of \hat{c} are motivated by research works from social scientists, journalists and psychologists regarding misinformation correction, which shows that counter responses are effective if they have the following desirable properties: politeness [62, 63], objective and evidenced [64, 51], make rational arguments [65], convey the competence of the commenter [65], and refute fallacy in reasoning [66, 67]. More elaborately, the desirable properties include:

- *Refuting*: the response explicitly refutes the the misinformation to correct the misinformation spreader. The expressed refutation via explicitly and objectively refuting misinformation in counter response can reduce misinformation’s impact [64].
- *Evidence*: the response contains supporting sentences to back up the refutation. Evidenced-based responses can more effectively debunk the misleading claims, and likely reduce the belief of misinformation poster [67]. More importantly, people are more willing to agree with a countering response when it is evidence-based [67].
- *Politeness*: the response is polite to avoid possible backfire. When countering mis-

information, uncivil responses can aggravate the misinformation poster, while it is more likely that the misinformation spreader favorably considers the true information when responses are polite [62, 63].

Beyond these specific requirements in misinformation correction domain, other textual properties are also required in generated text:

- *Fluency*: the generated text should be fluent in expression such that it is natural for people to read and understand.
- *Relevance*: the response should be relevant to the misinformation post and ensure coherent expression.

4.3 Counter-response Datasets: In-the-Wild and Crowdsourced

We create two novel counter-response datasets, first containing in-the-wild social media counter-responses and second containing crowdsourced in-lab counter-responses.

Misinformation Topics

We focus on COVID-19 vaccine misinformation due to its impact across the world. We mainly choose four popular misinformation topics to which a large number of users have been exposed [13, 184, 185, 187, 155, 186, 188, 189]. These misinformation topics gained popularity from December 2020 when the COVID-19 vaccines were approved by the FDA [184], in essence,

- Bill Gates conspiracy theories [184, 185]: This includes conspiracies claiming that Bill Gates created the COVID-19 vaccine to depopulate people or he holds the patents for COVID-19 vaccine to profit from the vaccine sales.
- COVID-19 vaccines contain microchips to track people [187].
- COVID-19 vaccines cause infertility and prevent pregnancy in women [155, 186].
- COVID-19 vaccines alter DNA or the vaccine is gene therapy [188].

In-the-wild Social Media Counter-Response Dataset

Misinformation Tweet and Response Collection

Our dataset builds on 14, 123, 473 COVID-19 vaccine-related tweets crawled by [154] from Dec 1, 2020 to July 31, 2021. Since we are more focused on responses rather than tweets themselves, we only keep tweets having at least one response, resulting in 1, 609, 069 tweets.

To identify misinformation tweets, we first create a COVID-19 vaccine misinformation tweet classifier using BERT [137] based on tweet annotations provided by [154]. This classifier has a performance in precision, recall and F1 scores of 0.972, 0.979 and 0.975, respectively. Then, we use this classifier to classify all remaining non-annotated tweets. Finally, we have 141,766 classified misinformation tweets. We crawl all their direct replies, resulting in 793, 828 replies.

Next, we filter tweets to retain those within the scope of our misinformation topics (Section 4.3), with at least one of the following (non case sensitive) keywords in the tweet textual string: “bill gates”, “fertility”, “pregnancy”, “pregnant”, “gene”, “dna”, “gene therapy”, and “microchip”, resulting in 1, 655 tweets with 26, 190 responses.

To create a high-quality dataset, we manually annotate all the classified 1, 655 misinformation tweets by the textual content to remove false positives and only focus on original tweets (no retweets) and English-language content, as is common practice [13, 131].

Finally, this dataset contains 798 misinformation tweets and associated 11, 970 responses.

Annotating Counter-Misinformation Replies and Training the Classifier

Naturally, not all responses to misinformation tweets counter it. Therefore, to develop a counter-response dataset, we create the following procedure.

Training a counter-response classifier: Since annotating all 11, 970 responses manually

is labor-intensive, we leverage existing work by [156] to create a belief versus disbelief classifier in social media responses. Specifically, following their pipeline, we create the classifier using RoBERTa [170] and train it on their annotated responses. Since the topics of the original data and trained classifier in [156] are different from ours, we annotated additional responses. Specifically, two students annotated 500 randomly-selected responses from the unlabeled 11,970 responses, resulting in an inter-rater agreement score of 0.7033 measured by percent agreement. This gave 244 responses expressing belief and 118 expressing disbelief, while the remaining were neither expressing belief or disbelief. We used these annotated responses to fine-tune the disbelief classifier to our data and topic. Conducting five-fold cross validation, the classification performances of the classifier per precision, recall and F-1 scores were 0.695, 0.687 and 0.691, respectively. Finally, we use the fine-tuned classifier to identify all potential disbelief replies among all the 11,970 responses. This resulted in 2,852 responses classified as disbelief or counter-response. Then, we manually verify all the classified responses through the textual content to remove all false positives. Finally, 754 true counter-responses are identified, which we use in our work.

Annotating Linguistic Properties of Counter-Responses

Two students annotated 50 counter-responses as per the three desired properties [67, 62, 63]:

- Refuting: is the response explicitly rejecting the false claim or the misinformation spreader?
- Evidence: does the response contain evidence or supporting words or sentences to back up the counter-response?
- Politeness: Is the reply rude, neutral, or polite like having a soft and friendly tone in the expression?

The measured inter-rater agreement score by percent agreement is 78%. Disagreements were discussed and a final label was given. Next, each annotator annotated the remaining counter-responses to assign final labels to them.

Finally, this results in 754 annotated (misinformation tweet, counter-response) pairs from 238 misinformation tweets. The distribution of the linguistic properties of counter-responses is shown in Table 4.1.

Table 4.1: Statistics of 754 social media counter-responses in MisinfoCorrect.

	Politeness		Evidence?		Refutes?
Polite	51	Yes	181	Yes	588
Neutral	415	No	573	No	166
Rude	288				

As per the statistics, in-the-wild counter-responses are very low quality – 38.19% responses are rude, 75.99% do not have evidence, and 22.02% do not explicitly refute the misinformation. This indicates they may not be effective. This further reinforces the critical and timely need for our research to develop an effective counter-response generator.

Crowdsourced In-lab Counter-Misinformation Responses

The above statistics show that most in-the-wild responses are rude and lack evidence. As a result, it will be challenging to train an effective counter-response text generator model using this data alone. Thus, we create an alternate dataset via crowdsourcing. Motivated by similar text generation for healthy and social good online communication [101, 191, 93], we recruit users familiar with Twitter to generate counter-misinformation responses that have the desired properties mentioned earlier in Section 4.2.

Ethics: This protocol was approved by Georgia Tech’s IRB.

Procedure: We use the following three-step process:

First, we recruited 20 college undergraduate and graduate students majoring in engineering domains in March 2022. During the screening, subjects provided background information including: (1) Highest education level: high-school, bachelors, masters, or doctorate; (2) Fluency in English: basic, intermediate, advanced (fluent or native speaker);

(3) Familiarity with the concept of online misinformation on Twitter: not familiar, somewhat familiar, highly familiar; and (4) Witnessed countering misinformation online: yes or no.

Out of these, 17 participants met the criteria of having least high-school education, being fluent in English, highly familiar with online misinformation, and having seen online debunking.

Second, each subject is provided written guidance about writing an effective counter-misinformation response governed by existing literature [67, 67, 62, 63]. Representative counter-misinformation examples are shown that are manually selected by the authors from the in-the-wild dataset (Section 4.3). Each subject is given up to 50 randomly-selected COVID-19 vaccine misinformation tweets (from the in-the-wild social media dataset) identified in Section 4.3. These tweets span all four misinformation topics (Section 4.3) to ensure diverse responses from different subjects.

After filtering out 90 written responses that do not satisfy any desirable properties (Section 4.2), we finally created a high-quality counter-misinformation response dataset containing 591 crowd-generated responses. A representative example is shown below:

Misinformation Post: It’s not a vaccine, it’s gene therapy. Gene therapy is an experimental technique. It’s the same technology used in cloning, DNA editing, and stem cell research.

In-the-wild Counter-response: You are born to speak nothing but lies.

Crowdsourced Counter-response: Sorry to see you think in this way. It is not correct. The vaccine is not gene therapy. Instead, it uses mRNA to generate spike protein to protect people. Please do not say the misinformation again.

4.4 MisinfoCorrect: A Counter-Response Generation Model

Here we describe our proposed counter-response generation model that leverages the two datasets to generate counter-responses for a given misinformation post. The generated counter-responses should have the desirable properties described in Section 4.2.

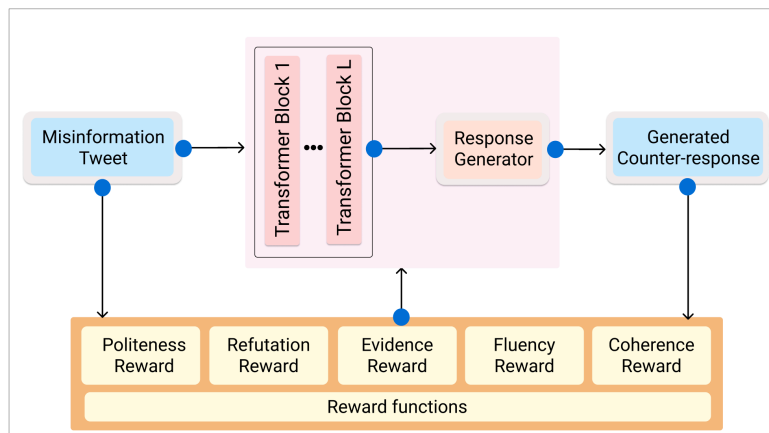


Figure 4.2: The overview of the MisinfoCorrect framework.

A Reinforcement Learning Framework

We choose a reinforcement learning-based approach due to its success in a variety of controllable text generation tasks [192, 191] and other tasks [193]. Moreover, we utilize reinforcement learning (RL) on top of a GPT-2 transformer-based text generation model since it is capable of generating quality example with limited number of examples derived from its strong generation power and is widely-used in text generation task [191]. By this design, we can bias the text generation process such that the generated counter-response is of high quality. Figure 4.2 presents the overview of MisinfoCorrect. Below we describe the components of the RL agent:

State:

The misinformation post provides the conversational text. The RL agent takes the misinformation post as the input to enhance the quality of counter-response text so that the response is relevant to the misinformation claims. Formally, the state $s \in \mathcal{S}$ is the same as the content of the misinformation post m , i.e., $s = m$. Our policy uses a string containing s for representation, which is also widely used in BERT-like models [137].

Action:

Given state s , the agent generates a candidate counter response \hat{c} . This generation action is represented as a lying in the whole action space \mathcal{A} , $a \in \mathcal{A}$, which is composed of all arbitrary-length sentences. We represent g as the text generator, and the action is $a = g(s)$.

Policy:

The policy is based on the transformer language model with the task of masked multi-head self-attention layers on GPT-2 [190, 108]. The input is an encoded representation of the state s and output is the action a . The generation task is framed as a language modeling problem where the goal is to generate \hat{c} that maximizes the conditional probability $p(\hat{c}|m)$. When using transformer component of GPT-2, we first encode our input string “m”. Then, after transforming the encoded representation as a vocabulary-sized vector using a softmax layer, we have a probability distribution over the entire vocabulary tokens. Next, top- p sampling method is used with the probability distribution to sequentially output a sequence of tokens to form a sentence. When the sampling process selects a special end-of-sequence token, the generation process stops. This generates the candidate counter-response \hat{c} .

Reward:

Research has shown that counter-misinformation responses are effective if they are polite, provide evidence, and explicitly refute the misinformation (Section 4.2). We design multiple novel reward functions to encourage the generated response to have these properties along with ensuring that the generated text is fluent, coherent, and relevant to the misinformation post. We describe the rewards below.

- **Politeness Reward:** Polite counter-responses are preferred (Section 4.2). We quantify the preference toward politeness as a politeness reward $r_{politeness}$ and create a politeness classifier $f_{politeness}$ using BERT [137] to measure politeness of text leveraging existing work [164]. The classifier fine-tuned and tested in our data in Section 4.3 has a classifi-

cation performance measured via precision, recall and F1 score of 0.8864, 0.9512, 0.8001. The politeness reward is formally computed as $r_{politeness} = f_{politeness}(\hat{c})$.

- **Refutation Reward:** Counter-responses that explicitly refute the misinformation are more effective (Section 4.2). Thus, we define the refutation reward $r_{refutation}$ to reward the actions that increase refutation of \hat{c} and penalize actions that decrease the refutation of \hat{c} . Following similar disbelief and polarity classification research works [194, 156], we build the refutation classifier $f_{refutation}$ using BERT [137] which measures whether the text expresses refutation. However, distinct from [156], who only use the response text for classification, we use both the tweet and generated response as input. The reason is that the refutation relationship would be better predicted by capturing the relative stance between the tweet and its response. We quantify the refutation reward as $r_{refutation} = f_{refutation}(m, \hat{c})$. In our experiments, the refutation classifier is first trained on the annotated data by [156]. Then, we fine-tuned and tested it on our data (Section 4.3), which finally achieves reasonable performance in precision, recall and F1 score with values of 0.7917, 0.8085, 0.7999, respectively.

- **Evidence Reward:** Responses containing evidence are more effective in countering misinformation [67]. Thus, we seek to generate response that provides textual evidence. We do not seek to provide a fact-checking URL as evidence, since readers are unlikely to click and read an external article from social media platforms [195, 196]. To effectively quantify the presence of evidence in responses, we consider the counter-response content where the response counters the misinformation post with supporting and relevant sentences.

We create an evidence classifier $f_{evidence}$ to predict whether the response provides evidence that counters the misinformation post. The classifier is trained by combining two sets of evidence-providing responses – first is the in-the-wild social media counter-responses that contain evidence (Section 4.3), and second is the subset of crowdsourced responses (Section 4.3) with evidenced responses. Finally, we create a balanced dataset of 573 evidenced-responses and 573 non-evidenced-responses to train the classifier.

We use BERT [137] as the classifier which takes both the post and response as inputs in a pair-wise setting [197] to measure the post-response pairwise relationship. After five-fold cross validation, the performance score of precision, recall and F1 score is 0.8864, 0.9512, 0.9176. The output of the classifier is the evidence reward, $r_{evidence}$, computed as $r_{evidence} = f_{evidence}(m, \hat{c})$.

- **Fluency Reward:** The agent needs to ensure that the response is fluent and grammatically correct. Thus, we want to reward actions that generate fluent outputs and penalize ones that result in non-fluent responses. To achieve this goal, following the previous work [198], we design the fluency reward $r_{fluency}$ which is the inverse of perplexity of the generated countering reply \hat{c} as $r_{fluency} = p_{GPT-2}(\hat{c})^{\frac{1}{M}}$, where p_{GPT-2} is the GPT-2 language model for English and M is the number of words in \hat{c} .

- **Coherence Reward:** Given a misinformation post, the generated response should be relevant to the post. We design a coherence reward $r_{coherence}$ which is computed via semantic similarity between m and \hat{c} as $r_{coherence} = sim(m, \hat{c})$, where sim measures the semantic similarity between two posts. In practice, we utilize the embedding from BERT model of the two text pieces [137] and compute their cosine similarity.

Total reward: Finally, the total reward is as

$$r = \alpha * r_{politeness} + \beta * r_{refutation} + \gamma * r_{evidence} + \theta * r_{fluency} + \lambda * r_{coherence} \quad (4.1)$$

where $\alpha, \beta, \theta, \gamma, \lambda$ are weights indicating the importance of rewards.

Optimization and Training

Warm-up start: We first use the pre-trained weights of DialoGPT [199] to initialize the weights in the transformer-based GPT-2 language model. Next, motivated by the warm-up approaches in reinforcement learning for dialogue generation by [192], we use the warm-start strategy on the paired data of misinformation posts and countering replies.

Reward Increment Training for Reinforcement Learning: To train the agent in the rein-

forcement learning framework, we take advantage of the existing reward increment training approach where the non-negative factor, offset reinforcement and characteristic eligibility are considered in the standard reinforcement learning setting [200]. In our setting for simplicity, we consider the reward r from the generated post and the probability of generating this post given the misinformation post, $p(\hat{c}|m)$. Finally, the loss function \mathcal{L} is computed as $\mathcal{L}(\theta) = -r * \log p(\hat{c}|m)$, where θ is the set of model parameters. We use \log to facilitate computation. Meanwhile, we utilize the negative of the reward to deploy the conventional gradient descent approach in experiments. Adam is used as the optimizer for model training [201].

4.5 Experimental Evaluation

We examine the performance of the proposed counter-misinformation response generation model. In particular, we focus on answering the following questions:

- **1:** Can the proposed model generate counter-misinformation responses of high quality with the desirable properties (Section 4.2)?
- **2:** What is the impact of using in-the-wild data versus crowdsourced data on the generated text output?
- **3:** What is the contribution of each component of the proposed method?
- **4:** Is the text generated good as evaluated by humans?

Baselines

We compare our model with representative dialog generation baselines and the work in fact-checking text generation:

- **Fact-checking Text Generation (FC-GEN)** [90]: The fact-checking text generation model takes in the tweets and replies for generation using gated recurrent unit.
- **DialoGPT** [199]: A dialogue generation model built on GPT-2 framework and pre-trained on Reddit conversations.

- Deep latent sequence model (**Seq2Seq**) [202]: A encoder-decoder model for general dialog text generation.
- **BART** [203]: An large pre-trained language model framework for sequence-to-sequence text generation.
- **Partner** [191]: A reinforcement-learning-based text rewriting method to output text.

Evaluation Metrics

To quantitatively evaluate the performance of the model, we use several metrics to measure both the effectiveness of the counter response and the text quality as follows:

- **Politeness**: We use the politeness classifier $f_{politeness}$ to test the level of politeness expressed in generated responses (Section 4.4).
- **Refutation**: We use the trained refutation classifier $f_{refutation}$ to measure refutation score, as defined in Section 4.4.
- **Evidence**: We use trained evidence classifier $f_{evidence}$ (Section 4.4) to measure how much evidence the reply provides.
- **Perplexity**: Following previous research [204, 198], we use pretrained GPT-2 language model to quantify perplexity to evaluate the expressed text fluency.
- **Relevance**: Following previous research [205], we compute the semantic similarity using BERT [137] to capture the coherence between posts and generated responses.

Evaluation of the Proposed Model

We train all the models with counter-responses from both *social media dataset* (Section 4.3) and *crowdsourced counter-responses* (Section 4.3). Specifically, we create a “clean” social media dataset by only selecting counter-responses with at least one dimension among politeness, refutation, and evidence labeled as positive. This is because training with low-quality counter-responses will lead to poor generation results. In addition, we use all crowdsourced counter-responses as they are all manually-verified to be polite, refuting,

Table 4.2: Effectiveness of MisinfoCorrect. Performance comparison of counter-response generators when trained on social media and crowdsourced responses.

Method	Polite. \uparrow	Refut. \uparrow	Evid. \uparrow	Perpl. \downarrow	Rele. \uparrow
DialoGPT	0.874	0.831	0.693	10.010	0.930
Seq2seq	0.794	0.794	0.621	13.403	0.948
BART	0.824	0.827	0.623	11.909	0.870
Partner	0.892	0.898	0.702	9.781	0.871
FC-GEN	0.815	0.714	0.594	14.971	0.810
MisinfoCorrect	0.915	0.931	0.723	8.010	0.960

Table 4.3: Examples of generated counter-responses by MisinfoCorrect and baseline methods.

Misinformation Post: It’s not a vaccine, it’s gene therapy. Gene therapy is an experimental technique. It’s the same technology used in cloning, DNA editing, and stem cell research.
MisinfoCorrect (proposed): This is not true. And, the vaccine is not gene therapy. It has nothing to do with cloning or DNA, and only uses mRNA for immunization goal. Please stop this misinformation.
DialoGPT: This is so unbelievably wrong. It is not gene therapy. The vaccine does not change DNA.
FC-GEN: It is misinformation. The vaccine is not gene therapy not gene therapy.

and evidenced.

The results comparing the generation models are shown in Table 4.2.

As can be seen, our proposed model generates the best counter-responses. When compared with baselines, our model has the highest politeness, refutation and evidence scores while still maintaining significantly lower perplexity and comparable relevance scores to ensure text of high quality. Table 4.3 illustrates responses generated by the proposed model and other baselines. As we can see, compared to other methods, MisinfoCorrect can generate text of desirable properties.

Table 4.4: Effectiveness of MisinfoCorrect. Performance comparison of counter-response generators when trained on social media responses only.

Method	Polite. \uparrow	Refut. \uparrow	Evid. \uparrow	Perpl. \downarrow	Rele. \uparrow
DialoGPT	0.762	0.726	0.623	12.039	0.940
Seq2Seq	0.734	0.641	0.473	14.312	0.820
BART	0.723	0.721	0.607	13.079	0.893
Partner	0.781	0.709	0.632	11.993	0.825
FC-GEN	0.714	0.663	0.515	15.102	0.782
MisinfoCorrect	0.854	0.797	0.643	10.110	0.938

Impact of Dataset Quality

Here we examine the impact of the dataset quality on the quality of generated response. We train the model using *only a “clean” social media responses* (i.e., responses that are evidenced, refuting, neutral, or polite) and no crowdsourced counter-responses. The performance results are shown in Table 4.4. First, we observe that compared to Table 4.2, the quality of responses generated by each model degrades. This highlights the importance of collecting crowdsourced data, which is of higher quality compared to social media data. Second, we note that our proposed model still generates the best counter-responses as per all metrics, except in relevance, in which it performs the second best.

Ablation Study

We examine the contribution of key components for effective counter-response generation (i.e., politeness, refutation and evidence rewards) in MisinfoCorrect on social media and crowdsourced responses data. We compare the model variations when using RL:

- *Base MisinfoCorrect model (Base)*: this model is the basic GPT-2 model fine-tuned on our dataset in a dialog manner as DialoGPT [199], but without using any rewards for training.
- *Base + politeness reward*: we only consider the politeness reward
- *Base + refutation reward*: we only consider the refutation reward
- *Base + evidence reward*: we only consider the evidence reward.
- *MisinfoCorrect model*: this is the complete model with all the reward functions.

Table 4.5: Ablation study of MisinfoCorrect.

Method	Polite. \uparrow	Refut. \uparrow	Evid. \uparrow	Perpl. \downarrow	Rele. \uparrow
Base	0.874	0.831	0.693	10.010	0.930
+ politeness	0.953	0.724	0.627	8.952	0.877
+ refutation	0.794	0.968	0.623	9.138	0.856
+ evidence	0.853	0.825	0.753	8.912	0.913
MisinfoCorrect	0.914	0.930	0.723	8.010	0.960

The results are shown in Table 4.5. When we only use the politeness, refutation or evidence reward function in the reinforcement learning framework, the corresponding politeness, refutation and evidence score is the highest and shows a significant increase compared to the Base model without any reward. When all the reward functions are combined in the MisinfoCorrect framework, there is a slight drop in each of the individual politeness, refutation, and evidence metrics, but it still has the second highest values along each dimension. This indicates that the MisinfoCorrect model finds a balance between the competing rewards during training.

Qualitative Evaluation

Experimental Setup: In addition to the quantitative evaluation of response generation, we follow previous research works [37] and also conducted human evaluation experiments to qualitatively examine the model performance. In particular, we recruited 10 subjects following the same procedure described in the counter-response annotation process (Section 4.3). Each subject is presented 30 data points, where each data point consists of one misinformation post and two counter-responses, and then asked “which response is better when countering the misinformation post: the first, the second, or are they equally effective?”. We test three settings: (1) the real counter-response versus the generated response by MisinfoCorrect; (2) the generated response by MisinfoCorrect versus the closest method, i.e., fact-checking generator (FC-GEN) [90]; (3) the generated response by MisinfoCorrect versus the most methodologically comparable baseline, i.e., DialoGPT [199]. We do not inform the subjects which response is generated by which method. Within each setting,

we randomly pick 50 data points for comparison, and each data point is annotated by two users. In the analysis of the results, we only summarize the data points on which the two users provide the same label, i.e., disagreement cases are discarded. In total, we received 300 data points in human evaluation for the three settings.

Ethics: This protocol was approved by Georgia Tech’s IRB.

Results: We get the following result:

(1) ***Real response versus MisinfoCorrect:*** In 46 out of 50 cases, both annotators provided the same answers. Among these, response generated by MisinfoCorrect were preferred in 76% cases, while in 6.5% cases, both responses were rated as equal. Real responses were preferred in the remaining cases.

(2) ***FC-GEN versus MisinfoCorrect:*** Annotators agreed in 44 out of 50 cases. Among these, MisinfoCorrect was preferred in 61.36% cases, 18.2% cases were equal, while 20.5% responses by FC-generator were better.

(3) ***DialoGPT versus MisinfoCorrect:*** Annotators agreed in 41 out of 50 cases. Among these, 36.6% cases prefer MisinfoCorrect, 36.6% cases are equal, and 26.8% cases prefer DialoGPT.

From all three comparison results, we can see that responses generated by MisinfoCorrect are preferred over the responses generated by the competing methods and the real responses. One representative example in Table 4.3 also illustrates the difference between these models and real responses. Altogether, the qualitative results show the potential for MisinfoCorrect in a real application to empower users to counter misinformation.

4.6 Conclusion

Overall, this work shows the potential to build on the recent advancements in generative text models for counter-misinformation response generation. Our proposed model showed promise by generating responses that were qualitatively and quantitatively better than real responses and other generated responses.

Given that we aim to mitigate the harm that can be caused by misinformation, our workflow should also minimize exposure to misinformation, e.g., when we engaged users in the tweet/reply annotation and reply writing process. We did not expose ordinary social media users to misinformation. Misinformation was shown to crowdsourcing workers to label tweets and write counter-replies. However, we followed several safeguards. First, we informed the crowd-workers up-front that the content may be misinformation. We also provided them with fact-checking resources. Ethical concerns were considered throughout the research and we ensured that our work was approved by the Georgia Tech IRB office before the research was conducted. Second, we acknowledge that certain misinformation topics can be controversial (e.g., climate change [206]). We need to ensure that we process misinformation in an unbiased way. To achieve this, we can crawl datasets from diverse groups rather than focusing on one specific group, open source our data and codebase for auditing, and filter out any potentially biased results when using the model in practice. We followed these guidelines to conduct our research on misinformation as fairly as possible.

The future work lies in three directions: (i) deploying and evaluating the model in practice, (ii) collecting data from professional fact-checkers as expert-generated counter-responses and compare the model performance against the current setup, and (iii) developing multi-lingual and multi-modal model to generate visual counter-responses.

CHAPTER 5

EVALUATING THE ROBUSTNESS OF DEEP SEQUENCE EMBEDDING-BASED DETECTORS

5.1 Introduction¹

Besides relying on crowds to combat online harmful information (e.g., misinformation) as shown in the previous chapters, we can turn to AI-powered detectors as well. This is critical as Web platforms, such as e-commerce, social media, and crowdsourcing platforms, have gained popularity, they are increasingly targeted by malicious actors for their gains [25, 207, 208]. The proliferation of undesirable users, such as fake accounts [25], spammers [209, 210], fake news spreaders [211, 118], abnormal users [212], vandal editors [213], fraudsters [207], and sockpuppets [208], poses a threat to the safety and integrity of online communities. To give an example, on Facebook, roughly 5% of monthly active users in 2019 were fake accounts [25]. Similarly, on Amazon, 63% reviews on beauty products were from fraudulent users [214]. Thus, the identification of malicious accounts by AI detectors is a critical task for all web and social media platforms.

Deep user sequence embedding-based classification models are increasingly gaining popularity for platform integrity tasks, including the TIES model at Facebook [25]. These models train a deep learning model to generate user embeddings by utilizing the temporal sequence of actions and post content of a user. The user embedding is then used to make predictions about the user. For example, Figure 6.1 shows a deep user sequence embedding-based classification model trained to identify malicious users from the user’s sequence of posts (top row).

However, deep learning models can be vulnerable to adversarial attacks [26]. While

¹This chapter is based on the PETGEN [37] paper published in ACM KDD 2021.

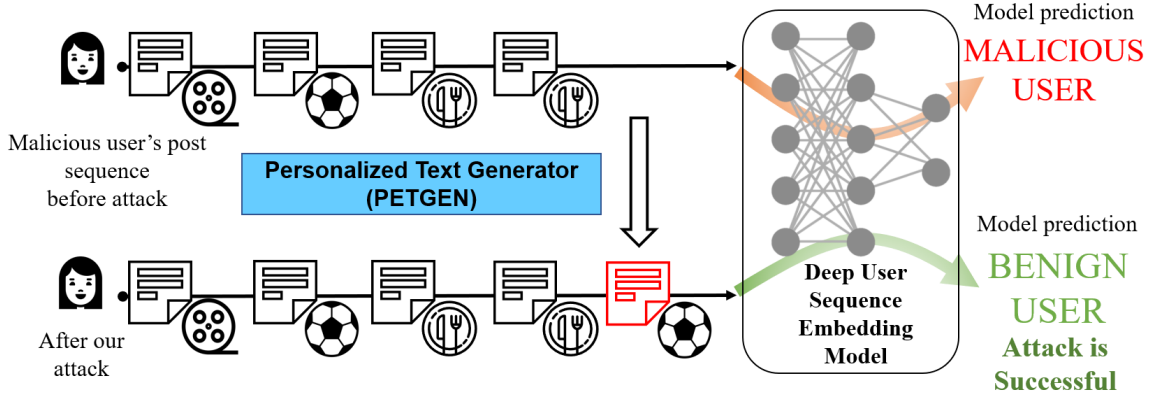


Figure 5.1: Application Setting of PETGEN: Deep user sequence embedding-based classification models are used to detect malicious users (top row). However, an evasion attack by an adversary by creating a new fake post can lead the same model to misclassify it as a benign user (bottom row). Our method, PETGEN, generates personalized text posts to adversarially attack the classifier.

adversarial attacks on deep learning models have received a lot of attention in graph representation learning, natural language processing, and computer vision domains [26], the vulnerability of deep user sequence embedding-based classification models remains unknown. For example, in Figure 6.1, the malicious user can create a new post, so that the entire user sequence is misclassified as benign by the classifier (bottom row). Thus, identifying the vulnerabilities of deep user sequence classification models is crucial to improve the models for real-world robustness.

We conduct an adversarial evasion attack on deep user sequence embedding-based classification models. Our **attack setting** is as follows: given a pre-trained deep user sequence classification model \mathcal{F} (trained to classify users as malicious or benign), a user's sequence of posts, and a target topic context, the goal of the attacker is to generate a new post on the target context such that the entire user sequence is now misclassified by \mathcal{F} .

Generating a fake attack post poses three major **challenges**. First, how can the text generation process effectively use the user's post sequence, such that the generated post aligns with the user's historical posts on similar contexts? Second, how to generate adversarial text that can fool a sequence embedding-based classifier? Finally, how to generate text that is personalized? Specifically, how can the text capture the user's writing style, be aware of

user’s recent vs past interests, and be knowledgeable about target context.

In this work, we create a Personalized Text Generation attack framework, called PETGEN, to generate adversarial text to attack deep user sequence embedding-based classification models. PETGEN is an end-to-end model. It leverages the sequential history of user posts (solution to challenge 1) by utilizing the relationship between the user’s historical posts and the target context, and builds a context-biased user sequence embedding. This is used to generate an initial version of the attack post. Next, the model adopts a multi-stage multi-task learning approach to manipulate the text to effectively attack the classification model (solution to challenge 2) and personalize the text to the user’s writing style, recent interests, and make the text relevant to the global discussions in the target topic context (solution to challenge 3). This step outputs the final attack text of PETGEN.

We evaluate the attack effectiveness and text quality of our model. We use two popular datasets: Yelp fake reviewer dataset [210] and Wikipedia vandal editor dataset [213], both with ground truth malicious users. We evaluate two popular deep user sequence embedding-based classification models: TIES, a model that is used in production at Facebook [25] and HRNN, a sequence classification model that uses sequential text embedding [105]. We compare PETGEN against five baseline and recent attack models that can generate attack text. Experiments reveal several key findings. First, both deep user sequence classification models are vulnerable to the fake text generation attack. Their model performance drops with even one generated post. Second, PETGEN generates attack text that results in a larger classification performance drop compared to existing attack methods. Third, the text generated by PETGEN has higher quality and is more personalized than existing attack methods. Fourth, PETGEN is highly effective in both the white-box setting (when the attacker has access to the details of the classification model) and the black-box attack setting (when the attacker does not know anything about the classification model). Finally, human evaluators rate text generated by PETGEN as being more realistic over text generated by existing generation-based attack methods.

Overall, our main contributions are:

- **New attack setting:** To the best of our knowledge, we are the first to investigate the problem of text generation attack on deep user sequence embedding-based classifiers, where adversaries generate a new piece of text added at the end of post sequence to fool the sequential classifier.
- **Attack model:** We create PETGEN, a multi-stage multi-task personalized text generation model that can generate attack that can effectively attack the sequence classifier and generate high-quality personalized text.
- **Effectiveness:** Extensive experiments on two datasets show that our methods can outperform five strong baselines in terms of the attack performance. Moreover, our method generates text with higher quality, both in terms of quantifiable metrics and as evaluated by human evaluators.

The code and data are at: <http://claws.cc.gatech.edu/petgen>.

5.2 Problem Definition

In this section, we formally define our problem as follows:

Preliminaries: We are given N users $U = \{u_1, \dots, u_N\}$ and a set of user ground truth labels $\mathcal{Y} = \{y_u\}$, where $y_u = 0$ means user u is a benign user and $y_u = 1$ means u is a malicious user. For each user u , we are given a sequence of chronologically ordered posts $P_u^{1:T} = \{p_u^1, \dots, p_u^t, \dots, p_u^T\}$, $P_u^{1:T} \in \mathcal{R}^{T \times d}$ where $p_u^t \in \mathcal{R}^d$ denotes user u 's post at time t and d is the number of tokens in the post. Each post has an associated context, describing the topic, background, or metadata of the post in detail. So, the sequence of contexts is $C_u^{1:T} = \{c_u^1, \dots, c_u^t, \dots, c_u^T\}$, $C_u^{1:T} \in \mathcal{R}^{T \times d'}$ where c_u^t is the topic context of post p_u^t and d' is the number of tokens in context. We are given a pre-trained deep user sequence embedding-based classification model \mathcal{F} , which generates user u 's predicated label $\mathcal{F}(P_u^{1:T})$. Model \mathcal{F} is trained to predict $\mathcal{F}(P_u^{1:T}) = y_u, \forall u \in U$.

Table 5.1: Table of major notations used in PETGEN.

Notation	Description
p_u^t	User u 's post at time t
$P_u^{1:T}$	User u 's sequence of past T posts
\hat{p}_u^{T+1}	User u 's generated post at time $T + 1$
c_u^t	User u 's context for post p_u^t
$C_u^{1:T}$	User u 's sequence of contexts $c_u^t, t \in \{1, \dots, T\}$
b_u	The target context for user u
y_u	The ground truth label of user u
\mathcal{G}	The text generator
\mathcal{F}	The pre-trained user sequence classifier

Attacker goal: Given user u 's sequence of posts $P_u^{1:T}$, contexts $C_u^{1:T}$, ground truth label y_u , and target context b_u , we aim to generate next post \hat{p}_u^{T+1} , such that $\mathcal{F}([P_u^{1:T}, \hat{p}_u^{T+1}]) = 1 - y_u$. Here $[P_u^{1:T}, \hat{p}_u^{T+1}]$ represents a sequence where the post \hat{p}_u^{T+1} is concatenated at the end of the sequence $P_u^{1:T}$. Thus, the goal of the attacker is to flip the prediction result of the classifier on the user's original post sequence. Our modeling goal is to train a text generator \mathcal{G} that generates the post \hat{p}_u^{T+1} using the user's historical posts. Thus, $\hat{p}_u^{T+1} = \mathcal{G}(P_u^{1:T}, C_u^{1:T}, b_u)$. We list the symbols in Table 6.1.

5.3 Methodology

System Overview

We propose an end-to-end personalized text generation system, called PETGEN, to attack deep user sequence classification models. Specifically, the input is the user's historical post sequence, corresponding contexts, the target context, and the pre-trained user sequence classifier \mathcal{F} . PETGEN has two major modules: in the first module, it leverages the user sequence and target context to generate sequence-aware contextual text. In the second module, this text is fine-tuned using a multi-stage multi-task learning setting such that it achieves the attack goal of fooling the classifier, adopts the user's writing style and ensures relevance to recent posts and to the target context. The resulting text is the output of PETGEN, which can successfully attack the target classifier. The overview of the system

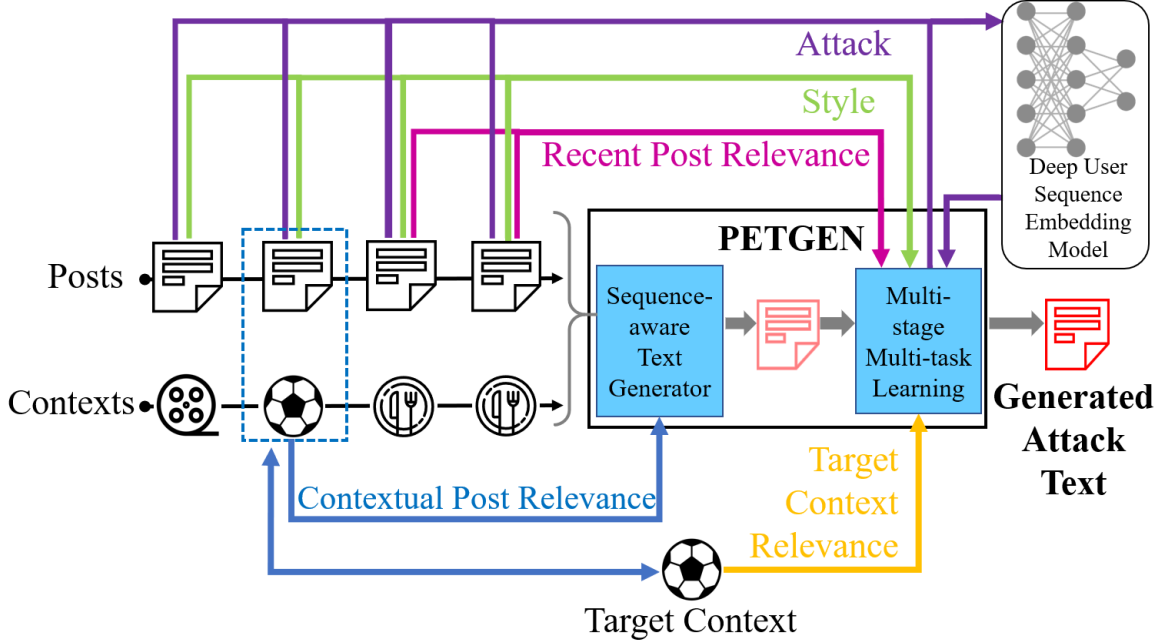


Figure 5.2: Overview of the PETGEN architecture: The sequence-aware text generator utilizes the sequence of post and context to generate text that maintains the contextual post relevance. Then, the multi-stage multi-task learning module fine-tunes the text by different tasks to generate attack text.

is shown in Figure 6.2.

Sequence-aware Conditional Text Generator In this module, PETGEN generates text on the input target context given a user’s sequence of historical posts and contexts. The goal of this module is to generate text such that the text incorporates the user’s historical views on the target context, as expressed in the past posts with contexts similar to the target context. Thus, among all the posts in the user’s sequence, the text generator should give more importance to posts that are on the same or similar context as the target context, motivated by multi-document summarization [215].

Here we treat the text generation process as a conditional language model which can leverage additional information [118, 216]. To this end, we propose a conditional text generation model incorporating the sequential post relevance through an attention mechanism, as shown in Figure 5.3. Specifically, $\mathcal{G}(P_u^{1:T}, C_u^{1:T}, b_u)$ is a conditional text generator that outputs next post \hat{p}_u^{T+1} , by sampling one token in one step. The output is based on (1)

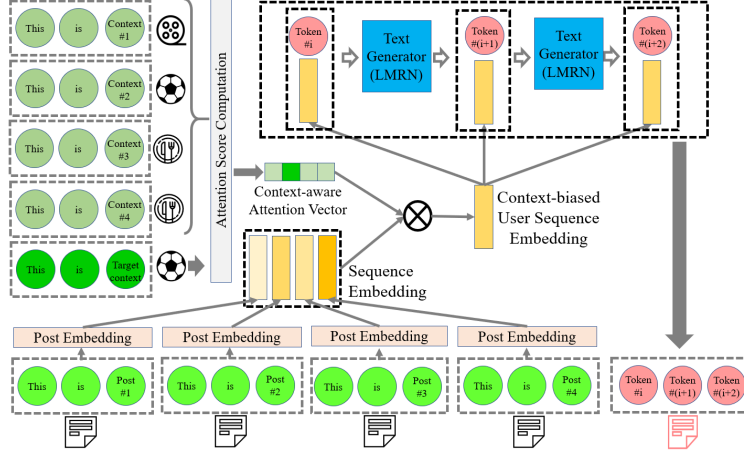


Figure 5.3: The overview of the sequence-aware conditional text generator in PETGEN. We first create the sequence embedding from the post embedding of each post in a sequence. We also compute the attention score between the target context and the user’s historical contexts to capture their pairwise relevance, resulting in a context-aware attention vector. After multiplying the generated sequence embedding and attention vector, we get the context-biased user sequence embedding. We concatenate it with the generated tokens for sequence-aware conditional text generation.

the sequence of posts $P_u^{1:T}$, (2) the sequence of context $C_u^{1:T}$, (3) the target context b_u , (4) previously generated tokens.

We select Relational Memory Recurrent Network (RMRN) as the basic text generation model g of \mathcal{G} , following previous work [118, 117], as RMRN models have shown remarkable performance in generating long text posts. Like traditional recurrent networks, g can convert each post in the sequence into a post embedding, obtained by the hidden state of g :

$$e_u^t = g(p_u^t) \quad (5.1)$$

where e_u^t is the embedding vector of the post $p_u^t, \forall t \in 1, \dots, T$.

To generate personalized text that is aware of the user sequence, we bias the text generator towards historical user posts that are contextually-relevant to the target context. This will ensure that the generated text has similar views as what the user has expressed in the past on the same context [215]. Specifically, we create an attention vector to quantify the contextual importance of each post in text generation. The attention vector is generated

by calculating the similarity between the target context b_u and each post’s context c_u^t . We create a context similarity function $A(\cdot)$ to capture the relationship as:

$$a_u^t = A(\text{Vect}(b_u), \text{Vect}(c_u^t)) \quad (5.2)$$

where $a_u^t, t \in \{1, \dots, T\}$ is the resulting attention score of the post p_u^t and it ranges from 0 to 1. $\text{Vect}(\cdot)$ is a function to transfer text into vector. Following the similar vectorization method in the previous works [118], we use the Latent Dirichlet Allocation model trained on the whole text to compute the vector representation. A high value of a_u^t means c_u^t is highly related to the target context b_u . Thus, the generated text should be more influenced by the corresponding post p_u^t . The attention vector is used to generate a Context-biased User Sequence Embedding vector s_u as follows:

$$s_u = \sum_{t \in \{1, \dots, T\}} \frac{\exp(a_u^t)}{\sum_{t \in \{1, \dots, T\}} \exp(a_u^t)} e_u^t \quad (5.3)$$

Thus, s_u is a representation of the user sequence which is biased towards user’s historical posts with similar contexts as the target context.

We use s_u in the text generation process to generate personalized and contextually-relevant text. Specifically, we combine s_u and the embedding vector of the generated token by addition to generate the next token. This ensures that each generated token is user sequence-aware. Formally, we have:

$$\hat{p}_u^{T+1}(i+1) \leftarrow \text{RM RN}(\text{LayerNorm}(\text{FeedFoward}(s_u) + \text{Embed}(\hat{p}_u^{T+1}(i)))) \quad (5.4)$$

where Embed is the embedding layer for tokens, FeedFoward is a feedforward layer to match dimensions during addition, LayerNorm is a normalization layer, and $\hat{p}_u^{T+1}(i)$ is a token at step i when generating \hat{p}_u^{T+1} . Note that a post has d tokens and thus the generation is done for d steps. The first token is initialized randomly. As we can see, each token is influenced by both the previous token and context-biased user embedding vector.

Finally, when outputting a token, each token is sequentially sampled using the conditional probability and the probability of the whole post can be presented as follows:

$$p(\hat{p}_u^{T+1} | P_u^{1:T}; C_u^{1:T}; b_u; \theta_{\mathcal{G}}) = \prod p(\hat{p}_u^{T+1}(i) | \hat{p}_u^{T+1}(i-1), \hat{p}_u^{T+1}(i-2), \dots, \hat{p}_u^{T+1}(1); P_u^{1:T}; C_u^{1:T}; b_u) \quad (5.5)$$

where $\theta_{\mathcal{G}}$ are the parameters of \mathcal{G} . Similar to the training of conditional language model [118, 216], we train \mathcal{G} by using Maximal Likelihood Estimation (MLE) with teacher-forcing and minimize the loss of negative log-likelihood for all posts based on the corresponding posts and contexts. To optimize the generator, we use the following objective function:

$$\min_{\theta_{\mathcal{G}}} L_{\mathcal{G}}^{GEN} = - \sum_{u \in U} \hat{p}_u^{T+1} \log p(\hat{p}_u^{T+1} | P_u^{1:T}; C_u^{1:T}; b_u, \theta_{\mathcal{G}}) \quad (5.6)$$

Finally, after training, the generator can output user u 's next post as:

$$\hat{p}_u^{T+1} = \mathcal{G}(P_u^{1:T}, C_u^{1:T}, b_u) \quad (5.7)$$

In our experiments, we use cosine similarity as the context similarity function $A(\cdot)$ to compute the attention score. Next, when training the generator \mathcal{G} , we use the last post as $(T + 1)$ -th post, the second last as T -th post and so on so forth. Additionally, since the sampling process is nondifferentiable, we use Gumbell-softmax relaxation trick to solve this problem [217, 117].

Multi-Stage Multi-Task Learning

In this module, the generated text post \hat{p}_u^{T+1} is modified to make the text realistic, personalized, and achieve the attack goal. We set it up as a multi-task learning module, which has four key tasks.

Style Task The generated post will only be personalized if it mimics the writing style

of the user. This is especially important when advanced classifiers, such as those deployed in practice [25], are equipped with a robust detector that detect posts that are way too different from the user’s previous writing style and the account is flagged as being malicious. Therefore, keeping the writing style similar is important for a successful attack. To achieve this goal, we create the style task to tune the generator \mathcal{G} .

We construct a text-GAN model for text style transfer, where a post style discriminator \mathcal{D} is deployed to co-train with \mathcal{G} by a Relativistic GAN loss [117, 218]. In particular, the discriminator \mathcal{D} determines whether the generated post \hat{p}_u^{T+1} by \mathcal{G} is less realistic than user’s historical post $p_u^t, \forall t \in [1, T]$ while the generator \mathcal{G} targets to generate realistic post to fool the discriminator \mathcal{D} . Formally, we have two objective functions to alternatively refine \mathcal{D} and \mathcal{G} :

$$\begin{aligned} \min_{\theta_{\mathcal{G}}} L_{\mathcal{G}}^{STY} &= -E_{(P_u^{1:T}, C_u^{1:T}, b_u) \sim p(P^{1:T}, C^{1:T}, B)} \log(\sigma(\mathcal{D}(p_u^t) - \mathcal{D}(\hat{p}_u^{T+1}))) \\ \min_{\theta_{\mathcal{D}}} L_{\mathcal{D}} &= -E_{(P_u^{1:T}, C_u^{1:T}, b_u) \sim p(P^{1:T}, C^{1:T}, B)} \log(\sigma(\mathcal{D}(\hat{p}_u^{T+1}) - \mathcal{D}(p_u^t))) \end{aligned} \quad (5.8)$$

where σ is a sigmoid function, $\theta_{\mathcal{D}}$ are the parameters of \mathcal{D} , $B = \{b_u\}$ is the set of all users’ target contexts, $P^{1:T} = \{P_u^{1:T}\}, C^{1:T} = \{C_u^{1:T}\}$ is the set of all users’ posts and contexts. In our experiment, we use multi discriminative representations [117] as the architecture of the discriminator \mathcal{D} .

Attack Task The primary goal of the generated text is to fool the target sequential classifier. Thus, we create the attack task to tune generator G to achieve this goal. The sequential classifier \mathcal{F} is originally trained using a binary cross entropy loss over the training data:

$$\min_{\theta_{\mathcal{F}}} L_{\mathcal{F}} = -\frac{1}{N} \sum_u y_u \log \mathcal{F}(P_u^{1:T}) + (1 - y_u) \log(1 - \mathcal{F}(P_u^{1:T})) \quad (5.9)$$

In a white-box attack, we directly use the trained classifier \mathcal{F} . In a black-box attack, we train a surrogate classifier \mathcal{F}' to mimic the predictions of \mathcal{F} . Note that once trained, both \mathcal{F} and \mathcal{F}' are not modified. Without loss of generality, we refer to the classifier we aim to attack as \mathcal{F} .

The classifier \mathcal{F} is utilized to tune generator \mathcal{G} such that the generated post \hat{p}_u^{T+1} fools the classifier into making incorrect predictions about the user $\mathcal{F}([P_u^{1:T}, \hat{p}_u^{T+1}]) = 1 - y_u$. Formally, we create the following objective function to optimize:

$$\begin{aligned} \min_{\theta_G} L_G^{ATT} = & -\frac{1}{N} \sum_u (1 - y_u) \log \mathcal{F}([P_u^{1:T}, \hat{p}_u^{T+1}]) \\ & + y_u \log(1 - \mathcal{F}([P_u^{1:T}, \hat{p}_u^{T+1}])) \end{aligned} \quad (5.10)$$

After the attack task is successful, the generated post will fool the classifier into predicting malicious users as benign users, and vice-versa.

Target Context Relevance Task Given a target context to generate a post, the attacker must ensure that the generated post is on-topic and is knowledgeable about the context. Otherwise, the generated post can be simply flagged as off-topic by a human or an automated topic detector. To ensure that the generated post is relevant to the target context, we minimize the mutual information gap between the target contexts $\{b_u\}$ of all users and the generated posts $\{\hat{p}_u^{T+1}\}$ of all users $u \in U$. A non-parametric Maximum Mean Discrepancy (MMD) based on the Reproducing Kernel Hilbert Space (RKHS) is utilized to effectively estimate this kind of distance [219]. Thus, we optimize the following objective function:

$$\begin{aligned} \min_{\theta_G} L_G^{CTX} = & MMD(\{b_u\}, \{\hat{p}_u^{T+1}\}) \\ = & \left\| \frac{1}{N} \sum_u \phi(b_u) - \frac{1}{N} \sum_u \phi(\hat{p}_u^{T+1}) \right\|_{\mathcal{H}} \end{aligned} \quad (5.11)$$

where \mathcal{H} is a universal RKHS, and ϕ is transfer function to change the space to the target RKHS space.

In experiments, the target context of the generated post is set to be the same as the context of the ground truth post at time T+1.

Recent Post Relevance Task This task ensures continuity and smoothness between the generated post and the most recent posts made by the user. This is important because real

users typically express such consistency in the real world [220]. Here, we quantify it as relevance towards recent posts, calculated as the mutual information distance between the generated post and the latest k posts of the user. Similar to the target context relevance task, we optimize such information gap by the following objective function:

$$\begin{aligned} \min_{\theta_G} L_G^{REC} &= MMD(\{P_u^{T-(k-1):T}\}, \{\hat{p}_u^{T+1}\}) \\ &= \left\| \frac{1}{N} \sum_u \sum_k \phi(p_u^{T-1-k}) - \frac{1}{N} \sum_u \phi(\hat{p}_u^{T+1}) \right\|_{\mathcal{H}} \end{aligned} \quad (5.12)$$

where k is the number of recent posts that have an impact on the next post generation. k is a hyper-parameter, which we typically set to 3 (more details are in the appendix).

Multi-stage Multi-task Learning Algorithm To achieve the personalized text generation objective, we optimize for the four tasks of style, attack, target context relevance and recent post relevance in a multi-stage process. Thus, we deploy the multi-stage multi-task learning framework to optimize:

$$\min_{\theta_{\mathcal{F}}} L_{\mathcal{F}}; \min_{\theta_{\mathcal{D}}} L_{\mathcal{D}}; \min_{\theta_G} (L_G^{STY} + L_G^{ATT} + L_G^{CTX} + L_G^{REC}) \quad (5.13)$$

where Eqn 13 is reflected in the while loop in the overall algorithm as presented in Algorithm 1. Finally, after tuning by the multi-task learning framework, the text generator finally generates personalized high-quality text for adversarial attack against the target sequential classifier.

5.4 Experiments

In this section, we examine the performance of the proposed PETGEN by conducting extensive experiments. Specifically, we aim to answer the following questions:

1. Is PETGEN able to successfully attack the deep user sequence classification model under both white-box and black-box attack settings?

Algorithm 1: PETGEN Algorithm

Input: a sequence of a user’s posts and associated contexts, the target context and the user’s label ;

Output: the user’s next post;

Train \mathcal{G} with contextual post relevance by MLE loss (Eqn 5.6);

while *Not Converge* **do**

 Train \mathcal{G} with \mathcal{D} on the Style Task (Eqn 5.8);

 Train \mathcal{G} on the Attack Task (Eqn 5.10);

 Train \mathcal{G} on the Target Context Relevance Task (Eqn 5.11);

 Train \mathcal{G} on the Recent Post Relevance Task (Eqn 5.12);

end

Table 5.2: Statistics of datasets used in PETGEN

Dataset	Yelp	Wikipedia
Number of users	3,940	794
Number of benign users	2,016	397
Number of malicious users	1,924	397
Total number of posts	35,123	11,547
Median posts per user	9	15

2. Beyond the attack performance, what is the quality of generated text, specifically its the relevance to the target context, contextual posts, and recent posts?
3. What is the contribution of the sequence-ware conditional text generator module and the multiple learning modules of PETGEN towards its performance?
4. When compared with other attack methods, is the text generated by PETGEN realistic

Table 5.3: *White-box attack performance* of PETGEN and existing methods on HRNN and TIES classifiers. PETGEN is the most effective attack (lowest F1 and highest Atk score).

Model	HRNN classifier				Min. imp. of PETGEN		TIES classifier				Min. imp. of PETGEN	
	Wikipedia		Yelp				Wikipedia		Yelp			
	F1↓	Atk↑	F1↓	Atk↑	F1↓	Atk↑	F1↓	Atk↑	F1↓	Atk↑		
Without attack	0.601	-	0.636	-	-	-	0.617	-	0.686	-	-	-
Copycat	0.550	21.3	0.610	8.0	9.836%	26.761%	0.513	16.3	0.625	11.5	6.823%	47.239%
Hotflip	0.581	21.2	0.591	9.5	6.937%	27.358%	0.514	15.0	0.641	10.3	7.004%	60.000%
UniTrigger	0.495	24.5	0.602	7.8	4.242%	10.204%	0.515	15.7	0.679	9.1	7.184%	52.866%
TextBugger	0.550	21.4	0.610	8.3	9.836%	26.168%	0.520	16.3	0.637	11.0	8.077%	47.239%
Malcom	0.479	25.5	0.570	18.0	1.044%	5.882%	0.560	18.0	0.538	21.8	6.877%	33.333%
PETGEN (proposed)	0.474	27.0	0.55	21.2	-	-	0.478	24.0	0.501	35.8	-	-

Table 5.4: *Black-box attack performance* of PETGEN and existing methods on HRNN and TIES classifiers. PETGEN is the most effective attack (lowest F1 and highest Atk score).

Model	HRNN classifier				Min. imp. of PETGEN		TIES classifier				Min. imp. of PETGEN	
	Wikipedia		Yelp				Wikipedia		Yelp			
	F1↓	Atk↑	F1↓	Atk↑	F1	Atk	F1↓	Atk↑	F1↓	Atk↑	F1	Atk
Without attack	0.601	-	0.636	-	-	-	0.617	-	0.686	-	-	-
Copycat	0.53	22.1	0.609	9.0	3.585%	8.597%	0.615	15.0	0.618	12.0	6.016%	64.167%
Hotflip	0.538	22.3	0.585	11.1	5.019%	7.623%	0.642	13.8	0.635	11.0	9.969%	79.091%
UniTrigger	0.529	22.0	0.624	7.5	3.403%	9.091%	0.601	17.9	0.601	15.0	3.827%	31.333%
TextBugger	0.545	21.0	0.607	9.5	6.239%	14.286%	0.627	14.0	0.617	12.2	7.815%	61.475%
Malcom	0.524	20.0	0.573	17.5	2.481%	20.000%	0.599	19.9	0.573	15.4	3.316%	27.922%
PETGEN (pro-posed)	0.511	24.0	0.53	22.3	-	-	0.578	33.0	0.554	19.7	-	-

enough from a human perspective?

Datasets

We evaluate the proposed method on real data from two popular platforms: Wikipedia and Yelp. Their statistics are shown in Table 6.2.

(a) **Wikipedia dataset:** This dataset consists of Wikipedia users (or editors) making edits on Wikipedia articles [213]. There are two types of editors: benign editors and vandal editors. Vandal editors were identified and removed from the Wikipedia platform by administrators. For each editor, the sequence of edits he or she made on Wikipedia articles is available. We consider each edit as one post. For each post, the leading paragraph of the edited page is set as the context of the post.

(b) **Yelp dataset:** This dataset consists of Yelp users giving reviews to restaurants [210]. Users are either benign reviewers or fraudulent reviewers. Fraudulent reviewers are identified by Yelp’s proprietary classification algorithm. For each reviewer, the sequence consists of its reviews on restaurants. Each review is one post. To create the context for each post, other reviews given on the same restaurant by other benign users are concatenated.

In both datasets, to ensure user sequences have enough information, we remove users with less than 5 posts and posts with less than 5 tokens. We use the latest 20 posts to create a user sequence.

Baselines

We compare PETGEN with five representative state-of-the-art adversarial text generation models.

(a) **Copycat**: Copycat randomly selects one post with similar context from the users' historical posts as the generated post. Three following baselines (Hotflip, UniTrigger, and TextBugger) use the Copycat post in their own attack.

(b) **Hotflip** [112]: Hotflip modifies the post generated by Copycat. It first detects the most important word in the post, based on the gradient of each input token with respect to the sequential classifier, and then swaps the most important word with a similar one.

(c) **Universal adversarial Trigger (UniTrigger)** [113]: UniTrigger generates an input-agnostic and fixed-length sequence of tokens to attack the classifier when concatenated to the end of an existing post. We turn to the topic modeling function in this specific application setting, similar to that adopted in prior work [118]. Particularly, we retrieve first topic-dependent words and contexts by the topic model and then prepend these universal prefix to a post.

(d) **TextBugger** [114]: TextBugger first uses various methods like deletion and swap to find carefully crafted tokens in a post and replaces some parts of the post with these tokens for attack.

(e) **Malcom** [118]: Malcom is the current state-of-the-art model in adversarial text generation to fool classifiers. It leverages the conditional language model to generate a new post where the attack and relevancy objective functions are deployed.

Evaluation Metrics To comprehensively evaluate text generation result, we use several metrics to measure attack effectiveness and text quality.

(a) **Attack Effectiveness: F1 score after attack (F1)**: This measures the classifier performance of the classifier. We compare the change in F1 score after the attack, compared to when there is no attack. If the resulting F1 score after the attack drops considerably, then the attack is successful. **Attack Rate (Atk)**: It measures the efficacy of the attack regarding changing predictions of the classifier. Specifically, a $M\%$ attack rate means the attack

Table 5.5: Comparison the quality of text generated by PETGEN and other attack strategies. PETGEN generates higher quality text in all but one case across all metrics.

Attack Model	Wikipedia Dataset								Yelp Dataset							
	HRNN				TIES				HRNN				TIES			
	BLEU↑	TCS↑	RS↑	CPS↑	BLEU↑	TCS↑	RS↑	CPS↑	BLEU↑	TCS↑	RS↑	CPS↑	BLEU↑	TCS↑	RS↑	CPS↑
Copycat	0.378	0.362	0.188	0.171	0.406	0.383	0.211	0.221	0.810	0.524	0.302	0.299	0.802	0.476	0.271	0.270
Hotflip	0.333	0.363	0.191	0.203	0.365	0.385	0.211	0.234	0.785	0.527	0.309	0.309	0.782	0.479	0.275	0.273
UniTrigger	0.213	0.397	0.214	0.192	0.239	0.410	0.230	0.223	0.737	0.527	0.325	0.326	0.725	0.463	0.273	0.272
TextBugger	0.341	0.372	0.192	0.172	0.374	0.393	0.214	0.226	0.771	0.520	0.311	0.312	0.768	0.478	0.280	0.279
Malcom	0.914	0.312	0.175	0.240	0.878	0.484	0.209	0.213	0.849	0.540	0.349	0.354	0.856	0.515	0.321	0.291
PETGEN	0.893	0.463	0.275	0.281	0.896	0.233	0.254	0.474	0.852	0.544	0.401	0.410	0.870	0.519	0.397	0.398

can fool the classifier $M\%$ of the time on the sequences that the classifier has previously correctly labeled.

(b) Text Quality: BLEU: Like previous works on text generation [117], we deploy BLEU to indicate the quality of generated post by comparing them with testing data. Higher scores indicate better text. **Target Context Similarity (TCS):** We compute the similarity between the generated posts and the target context as follows:

$\frac{1}{N} \sum_u \text{cosine}(\text{Vect}(b_u), \text{Vect}(\hat{p}_u^{T+1}))$, where $\text{cosine}(\cdot)$ is the cosine similarity function and N is the number of users. Higher scores indicate more relevant text. $\text{Vect}(\cdot)$ is the previously-defined LDA-based function to transfer text into vector. **Recent Post Similarity (RS):** Similar to target context similarity, recent post similarity score computes the distance between the generated post and the most recent k posts as:

$\frac{1}{N} \sum_u \sum_{t \in \{T-(k-1), \dots, T\}} \text{cosine}(\text{Vect}(p_u^t), \text{Vect}(\hat{p}_u^{T+1}))$. **Context Post Similarity (CPS):** Similarly, the context post similarity computes the similarity between the generated post and the posts in the user sequence that are of similar context as the target context. This is calculated as:

$\frac{1}{N} \sum_u \sum_{t \in \{1, 2, \dots, T\}} a_u^t * (\text{cosine}(\text{Vect}(p_u^t), \text{Vect}(\hat{p}_u^{T+1})))$, where a_u^t is the previously mentioned attention score which captures the relationship between the contexts c_u^t and b_u of posts p_u^t and \hat{p}_u^{T+1} respectively.

Target Classification Models

We target two deep user sequence classification models to test the generality of our attack.

(1) **Hierarchical Recurrent Neural Network (HRNN)** is a model where the sequential

Table 5.6: Ablation studies of PETGEN showing the contribution of each component in PETGEN.

Model	Wikipedia Dataset						Yelp Dataset					
	F1↓	Atk↑	BLEU↑	TCS↑	RS↑	CPS↑	F1↓	Atk↑	BLEU↑	TCS↑	RS↑	CPS↑
PETGEN Base Text Generator	0.479	26.5	0.899	0.375	0.268	0.247	0.625	11.7	0.857	0.382	0.349	0.187
w/ Style	0.576	21.1	0.895	0.390	0.218	0.249	0.59	17.5	0.871	0.481	0.324	0.301
w/ Attack against TIES	0.478	25.0	0.894	0.368	0.216	0.216	0.499	45.3	0.843	0.476	0.357	0.250
w/ Attack against HRNN	0.465	27.5	0.895	0.388	0.240	0.249	0.530	29.5	0.846	0.445	0.315	0.157
w/ Recent Post Relevance	0.486	23.8	0.887	0.463	0.275	0.267	0.592	17.7	0.851	0.495	0.43	0.215
w/ Target Context Relevance	0.483	23.9	0.887	0.459	0.258	0.258	0.571	18.0	0.830	0.559	0.361	0.203
w/ Contextual Post Relevance	0.566	21.2	0.705	0.397	0.225	0.276	0.554	19.2	0.845	0.514	0.331	0.451
PETGEN against HRNN	0.474	27.0	0.893	0.463	0.275	0.281	0.550	21.2	0.852	0.544	0.401	0.410
PETGEN against TIES	0.478	24.0	0.896	0.474	0.233	0.254	0.501	35.8	0.870	0.519	0.397	0.398

pattern of the input text is captured by the hierarchical structure for accurate classification [103]. In HRNN, each user post is first converted to a vector and the sequence of user post vectors is converted into a compact user embedding. This user embedding is used for user classification.

(2) **Temporal Interaction Embeddings (TIES)** is a model used by Facebook for malicious account detection. We use the temporal embedding component of the TIES model for classification (as there is no graph structure in our datasets). Note that TIES is the state-of-the-art deep user sequence embedding-based classification model for malicious user detection.

Experiment Setup We split the dataset by five-fold cross-validation and report the average numbers. By default, we set $k = 3$ as the number of recent- k posts (more details on impact of value k are in the appendix), the number of tokens in a post and a context to be $d = d' = 30$ and the learning rate as $1e-5$. We use Adam as the optimizer with mini-batch size of 64 [221].

Adversarial Attack on Sequential Post Classification

In this section, we evaluate the proposed attack model on both white-box classifiers and black-box classifiers.

Attack on White-Box Classifiers. In a white-box attack, the attacker has access to the model parameters of the target classifiers. Thus, they attack the trained model directly. The results comparing the performance of PETGEN with baseline models is shown in Table 5.3

on both Wikipedia and Yelp datasets, with both the HRNN and TIES models as classifiers. The table also shows the results of the classification models without any attack.

We have several important findings. First, without any attack, the TIES model has a higher model performance (F1 score) compared to the HRNN model on both the datasets. Second, under attack, the model performance of both TIES and HRNN reduces, showing the vulnerability of both these models to text generation attacks. Next, comparing all attacks, PETGEN attack results in the lowest F1 score and highest attack rate on both datasets, making it the most successful attack. On the TIES classifier, PETGEN has at least 6.82% improvement over all baselines in terms of F1 score and on the HRNN classifier, at least 1.04% improvement on F1. This is important as TIES is the state-of-the-art classifier that is being used at Facebook. Successfully attacking TIES shows the strength of our PETGEN attack. Finally, we find PETGEN attacks TIES more efficiently than HRNN with larger drop in F1 score and higher attack rate over baselines. A possible reason is that the more complex deep sequential model like TIES can provide more signal in computing cross entropy loss, finally enabling the attacker to learn more about how to downgrade the performance.

Attack on Black-Box Classifiers. In the black-box setting, the attacker does not have access to the parameters of the sequential post classifier. Thus, we train a surrogate HRNN classifier to mimic the classification of the original black-box classifier. The text generation attack methods create the fake post using this surrogate classifier, and then this generated text is used to attack the original black-box classifier. The results of the performance drop on black box classifiers is shown in Table 5.4.

First, as before, we see that PETGEN beats all the existing attack methods in terms of F1 score and attack rate. Next, similar to the result in the white-box setting, PETGEN can more effectively attack the HRNN and TIES models compared to existing attack approaches. Finally, comparing attacks on the same model under white-box and black-box setting, it is harder for the attackers to attack the black-box classifier. For all models, the drop in F1 score is lower during black-box attack compared to white-box attack.

Personalized Text Generation

Beyond the attack performance, we present text quality of the generated post in Table 5.5. As we can see, `PETGEN` always generates post with higher quality in all four evaluation metrics compared to the other five baseline methods. This is reflected in the BLUE score and in the relevance of the generated post to the previous posts of the user and the target context.

The reasons of higher quality text generation is the following. Compared to the four word-perturbation attack methods, namely, Copycat, Hotflip, UniTriggr and Textbugger, our method `PETGEN` is an end-to-end text generation framework that can effectively pick a less diverse set of words that are highly relevant to the target context, historical post, and recent post. This enables `PETGEN` to output text with higher quality. Compared to the Malcom model, `PETGEN` deploys the context-aware text generator and the learning task of recent post relevance to leverage the historical post and recent post information for generation. It makes text more real and personalized, thus having higher scores on all text quality metrics.

Evaluating Consistency in Attacker Goal: To further examine the effectiveness of our attack models, we compare the sentiment of generated adversarial post with that of the original post under the same context. We use Vader [163] to compute the sentiment score on the posts in the Yelp dataset. We find that 70.8% of generated posts have the same sentiment as the original post, indicating that the attacker’s generated post has the same positive or negative tone as desired to uprank or downrank a restaurant.

Ablation Study To examine the effectiveness of each component in `PETGEN`, we conduct the ablation study where we test the performance of different variants of `PETGEN`, and the results are in Table 5.6. The simplest model is simply the `PETGEN` base text generator, which is the traditional RMRN text generator and no other modules are used. As we can see, `PETGEN` with all the modules always performs the best or the second best among all other variants for all six metrics. Comparing the different variants, we find the attack

task can help decrease the F1 score and increase the attack rate, making the adversarial attack successful. Meanwhile, the task of post style, target context relevance and recent post relevance can enhance the target context and recent post similarity score. The sequence-aware text generation setting to capture the contextual historical post relevance increases the context post similarity score.

Human Evaluation on Generated Text To better evaluate the quality of the generated text, we conduct human evaluations. Specifically, we test whether posts generated by PETGEN are more realistic compared to those generated by Malcom (the SOTA end-to-end adversarial text generation method). We recruit two non-author evaluators and give them each 50 pair of posts, generated for 50 randomly selected user sequences. In each pair, one post is generated by PETGEN and the other by Malcom. The evaluators are not told which post is generated by which method. Their task is to mark which of the two posts is more realistic, or whether they are equally (un-)realistic.

We get the following result. The two reviewers achieve an inter-rater agreement score of 0.66 and 40% posts are labeled as equally realistic. Among the remaining posts, reviewers label 58.33% posts by PETGEN more realistic than Malcom. From this result, we can see our method is able to outperform Malcom in generating realistic posts, and has great potential in real-world applications.

5.5 Conclusion

We created a new attack framework to evaluate the robustness of deep user sequence classification models and showed its effectiveness. This work has some shortcomings. First, it is currently only applicable for posts in the English language, while social media posts can be in any language. Second, the model can only work with sequences, while does not incorporate complex structures, such as graphs. Third, the attack is restricted to generating new posts. Other attack capabilities can be explored in the future.

CHAPTER 6

IMPROVING THE ROBUSTNESS OF DEEP SEQUENCE EMBEDDING-BASED DETECTORS

6.1 Introduction¹

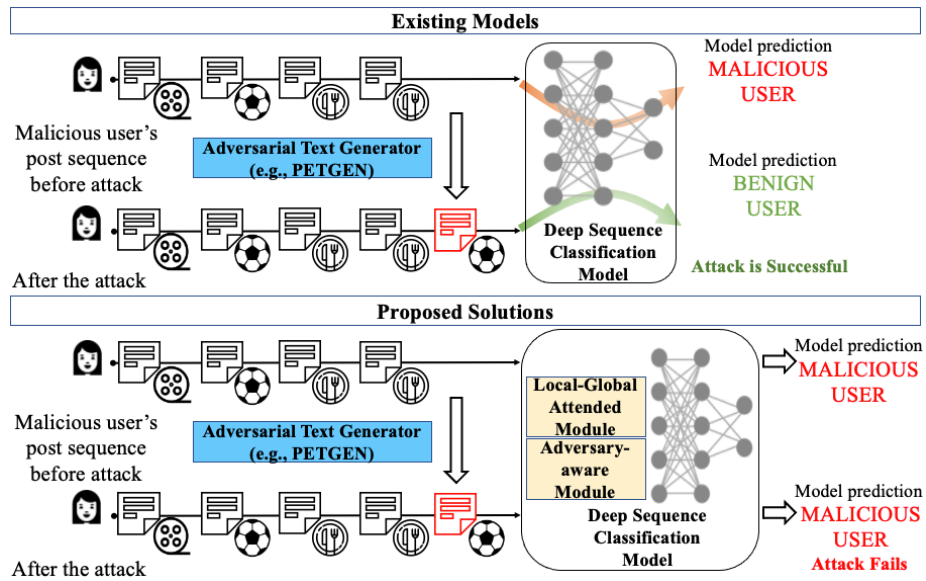


Figure 6.1: Deep sequence classification models are used to detect malicious users. However, the next post attack by an adversary by creating a new fake post can lead the same model to misclassify it as a benign user. Our proposed solution built on the local-global attended and adversary-aware modules can accurately and robustly identify malicious users.

As we found in the previous chapter, the deep sequence-embedding based classification models are vulnerable to adversarial attacks, as illustrated in Figure 6.1. The potential reason can be that existing deep user sequence embedding-based classification models are sensitive to the changes in the input sequence. Thus, improving the model robustness is also crucial.

Building an effective and robust classification model is non-trivial because (1) Existing models rely on the RNN and its variants or transformer-based BERT[137] to process posts

¹This chapter is based on the paper submitted to ACM TKDD 2024 [222].

one by one, thus neglecting to recognize and leverage two different levels of information – the post level and the sequence level (Challenge 1). (2) When designing bad actor detectors, existing research usually has not considered the setting where bad actors are able to write adversarial posts to bypass the classifier, therefore making the model vulnerable in real-world applications (Challenge 2).

To bridge these gaps, we propose a novel deep user sequence embedding-based classification model. Our setting is as follows: Given a user’s sequence of posts, the goal of the model is to accurately predict the label of the sequence, even if the sequence is modified by attackers through the next post attack [37]. In the design, we create a Transformer-based Adversary-aware Local-Global Attended classification model to effectively and robustly categorize user sequences. It first leverages the transformer encoder block to encode each post bidirectionally, thus building a comprehensive post embedding. Next, the model adopts the transformer decoder block to model the sequence of post embeddings by attention mechanism to generate the sequence embedding (solution to challenge 1). Finally, the sequence embeddings of original sequences and modified sequences by mimicked attackers are fed into a contrastive-learning-enhanced classification layer for sequence prediction. The modified sequences by mimicked attackers enhance the knowledge of the classifier such that it can be stable to adversarial attacks (solution to challenge 2).

We extensively evaluate the classification effectiveness and robustness of our model to show its superiority. Similar to Chapter 5, we employ two popular datasets: Yelp fake reviewer dataset [210] and Wikipedia vandal editor dataset [213], both with ground truth malicious users. We compare three popular deep user sequence embedding-based classification models: TIES, a model that is used in production at Facebook [25], HRNN, a sequence classification model that uses sequential text embedding [105], and one advanced transformer-based adapted BERT model for sequence classification [137]. On the other hand, we also add two defense-involved classification models: Fine-tuning-based defense [223] and Mixup-based Data-Augmentation-based defense [224]. The experiments

demonstrate that our model outperforms all compared methods in the F1 score in both datasets when under attack by the state-of-the-art next post attack by PETGEN [37] and Large Language Model (e.g., LLaMA [225]).

In summary, our main contributions are:

- We propose a transformer-based local-global attended pipeline to model both the post and sequence information to comprehensively capture the sequence representation.
- We propose a contrastive-learning-based adversary-aware training module for classification, thus enhancing the model knowledge to make it stable against adversarial attacks.
- Extensive experiments demonstrate that our method can outperform representative compared methods with the lowest F1 drop under state-of-the-art attack.

6.2 Problem Definition

In this section, we formally define our problem as follows:

Preliminaries: We are given N users $U = \{u_1, \dots, u_N\}$ and a set of user ground truth labels $\mathcal{Y} = \{y_u\}$, where $y_u = 0$ means user u is a benign user and $y_u = 1$ means u is a malicious user. For each user u , we are given a sequence of chronologically ordered posts $P_u^{1:T} = \{p_u^1, \dots, p_u^t, \dots, p_u^T\}$, $P_u^{1:T} \in \mathcal{R}^{T \times d}$ where $p_u^t \in \mathcal{R}^d$ denotes user u 's post at time t and d is the number of tokens in the post.

Our Goal: We aim to build a deep user sequence embedding-based classification model \mathcal{F} , which can accurately generate user u 's predicated label $\mathcal{F}(P_u^{1:T})$ such that $\mathcal{F}(P_u^{1:T}) = y_u, \forall u \in U$, even if in the **adversary-aware setting**. Specifically, given user u 's sequence of posts $P_u^{1:T}$, an attacker can use the off-the-shelf text generation model (e.g., LLAMA[225], ChatGPT [226], PETGEN [37]) to generate next post p_u^{T+1} , which may flip the prediction result of the classifier on the user's original post sequence $\mathcal{F}(P_u^{1:T})$. Particularly, after concatenating the new post to the existing sequence, the user has a new post

sequence $[P_u^{1:T}, p_u^{T+1}]$. However, the classifier can still accurately predict the label of the user in the face of the new post as $\mathcal{F}([P_u^{1:T}, p_u^{T+1}]) = y_u$. We list the symbols in Table 6.1.

Table 6.1: Table of notations used in the defense model

Notation	Description
p_u^t	User u 's post at time t
$P_u^{1:T}$	User u 's sequence of past T posts
p_u^{T+1}	User u 's generated post at time $T + 1$
y_u	The ground truth label of user u
\mathcal{F}	The deep user sequence embedding-based classifier
W_p	The position embedding matrix
W_e	The token embedding matrix
Emb_u	The sequence embedding of user u

6.3 Methodology

System Overview

In this work, we propose a Transformer-based Adversary-aware Local-Global Attended sequence classification system. Specifically, given the input of the user's historical post sequence, our system outputs the predicted label. It has two major modules: in the first module, it leverages both transformer encoder and decoder blocks to generate the local-global attended sequence embedding. In the second module, it utilizes contrastive learning to enhance the model capability by distinguishing the embedding of users from that of mimicked attackers. Through this design, our model can accurately predict the label of a given sequence. The overview of the system is shown in Figure 6.2.

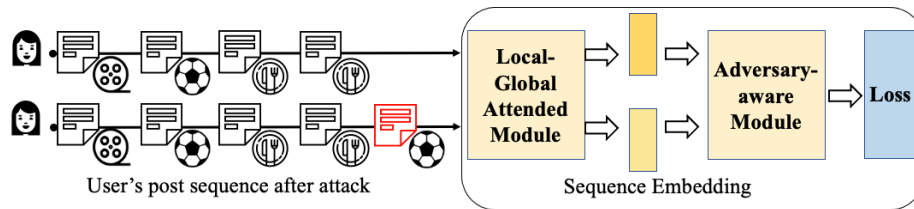


Figure 6.2: Overview of the proposed defense model.

Local-Global Attended Module

In this module, the goal is to generate the local-global attended embedding such that the embedding can comprehensively represent both the local and global information of the sequence. To achieve this goal, motivated by the self-attention mechanism in transformer [108], we build the encoder- and decoder-based dual transformer architecture, as shown in Figure 6.3. Finally, our model generates the sequence embedding given a user's sequence of historical posts.

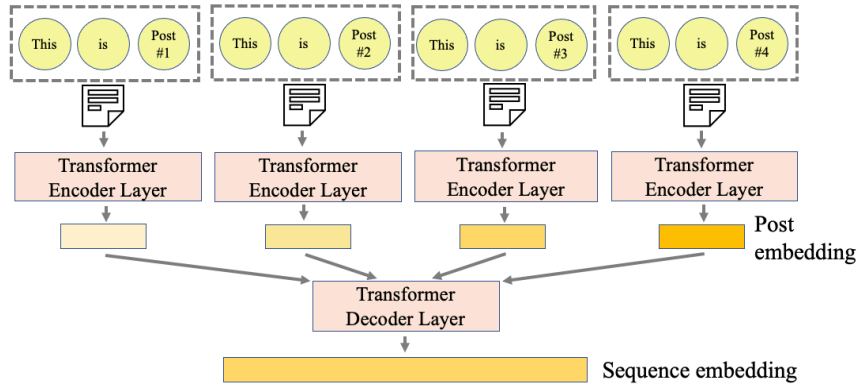


Figure 6.3: Local-global attended module in the defense model

Transformer Encoder-based Local Module at the Post Level To represent the information in each post, we need to fully understand the post by encoding the text bi-directionally rather than unidirectionally. Building on the bidirectional property of transformer encoder block [108], we first pass each post p_u^t to the embedding layer represented by W_e meaning the token embedding matrix. Then, we add the token embedding and the position embedding matrix W_p to form the initial token representation h_0^i . Finally, we pass h_0^i to transformer encoder blocks through n layers and obtain the post embedding h_l^i . This process is mathematically formulated as:

$$\begin{aligned}
 h_0^i &= p_u^i W_e + W_p \\
 h_l^i &= \text{transformer_encoder_block}(h_{l-1}^i) \quad \forall l \in [1, n]
 \end{aligned}
 \tag{6.1}$$

Transformer Decoder-based Global Module at the Sequence Level After passing a sequence of post $\{p_u^1, \dots, p_u^t, \dots, p_u^T\}$ to the transformer encoder-based module, we obtain a list of post embeddings, $h_l^1, h_l^2, \dots, h_l^i, \dots$. To capture the sequential pattern in the list of post embeddings, we need to chronologically process the posts and attend to the informative post in the sequence. Motivated by the transformer decoder design where masked self-attention is deployed to efficiently and effectively model the sequential relationship, we adopt this approach. Particularly, we first create a matrix $H = [h_l^1, h_l^2, \dots, h_l^i]$ to represent the list of post embeddings. After add the position embedding matrix W_p , we have the initial matrix H_0 . Finally, we pass H_0 to transformer decoder blocks through n layers and obtain the sequence embedding of the user Emb_u . This process is formulated as:

$$\begin{aligned}
 H_0 &= H + W_p \\
 H_l &= \text{transformer_decoder_block}(H_{l-1}) \quad \forall l \in [1, n] \\
 Emb_u &= H_l
 \end{aligned} \tag{6.2}$$

Adversary-aware Module

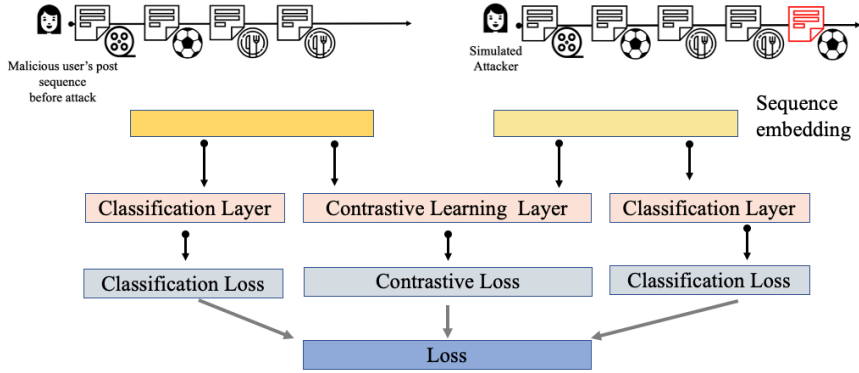


Figure 6.4: Adversary-aware module in the defense model

Classification Loss After passing a user post sequence $\{p_u^1, \dots, p_u^t, \dots, p_u^T\}$ to the local-global attended module, we have Emb_u . Similarly, for the modified user post sequence by adversarial attacks $[P_u^{1:T}, p_u^{T+1}]$, we have Emb_u^{attack} .

To classify the user sequence, we pass the embedding vector through a linear layer and add a softmax classifier on top of the embedding as follows:

$$p(y_u|Emb_u) = softmax(W Emb_u) \quad (6.3)$$

where W is the weight matrix for the linear layer. For the classification task, we minimize the cross entropy as the loss function:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N y_u \log p(y_u|Emb_u) + (1 - y_u) \log(1 - p(y_u|Emb_u)) \quad (6.4)$$

We have similar loss functions for the modified post sequence by attackers as:

$$\begin{aligned} \mathcal{L}_{CE}^{attack} = & -\frac{1}{N} \sum_{i=1}^N y_u \log p(y_u|Emb_u^{attack}) \\ & + (1 - y_u) \log(1 - p(y_u|Emb_u^{attack})) \end{aligned} \quad (6.5)$$

Finally, for the classification loss, we add the two loss functions together as:

$$\mathcal{L}_{CLF} = \mathcal{L}_{CE} + \mathcal{L}_{CE}^{attack} \quad (6.6)$$

Contrastive Loss To enhance the robustness of models against adversarial attacks, an intuitive strategy is to ensure that the sequence-level representations of a user’s original sequence and its adversarial counterpart are as similar as possible. This approach aims to make the model more resistant to noise by minimizing discrepancies between genuine and adversarially modified sequences. Conversely, sequences that do not form a paired relationship should be distinctly separated in the representation space, enhancing the model’s ability to discriminate between authentic and manipulated inputs. In pursuit of this goal, we incorporate contrastive learning as a supplementary regularization technique during training. Recent advancements in contrastive learning, exemplified by MoCo [227] and SimCLR [228], utilize a variety of data augmentation methods, such as random cropping and

color distortion, to generate positive pairs for training. MoCo introduces a queue mechanism for managing a pool of negative examples, whereas SimCLR leverages in-batch sampling to gather negative examples. These methods train models by optimizing the InfoNCE loss, a strategy that fosters learning by distinguishing between similar (positive) and dissimilar (negative) pairs of data points. In our approach, we adopt the SimCLR framework for generating positive and negative pairs, along with its loss function, to implement a contrastive learning objective tailored to our specific needs. This methodological choice allows us to effectively model the desired relationships between original and adversarial sequences, thereby improving the noise resilience of our deep learning models.

Particularly, we first pass the sequence embedding to a non-linear projection layer for the sequence representation transfer as follows:

$$\begin{aligned} z_u &= W_2 \text{ReLU}(W_1 \text{Emb}_u) \\ z_u^{\text{attack}} &= W_2 \text{ReLU}(W_1 \text{Emb}_u^{\text{attack}}) \end{aligned} \quad (6.7)$$

Where W_1 and W_2 are the weight matrices.

With a batch of N user sequence examples and their corresponding adversarial examples, for each positive pair (z_u and z_u^{attack}), there are $2(N - 1)$ negative pairs, i.e., all the rest of the examples in the batch are negative examples. Here, we use z_i to denote one in the $2(N - 1)$ examples and we have z_u and z_i as the negative pair. The contrastive objective is to identify the positive pair and we compute the contrastive loss using the InfoNCE loss as:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \left(\frac{\exp(\text{sim}(z_u, z_u^{\text{attack}}))}{\sum \exp(\text{sim}(z_u, z_i))} \right) \quad (6.8)$$

where $\text{sim}(u, v) = \frac{u^T v}{\|u\|_2 \|v\|_2}$ denotes the cosine similarity between two vectors.

Finally, we perform a multi-task learning diagram and take a weighted average of the classification loss and the contrastive loss as:

$$\mathcal{L} = w_{\text{contrastive}} \cdot \mathcal{L}_{\text{InfoNCE}} + (1 - w_{\text{contrastive}}) \cdot \mathcal{L}_{\text{CLF}} \quad (6.9)$$

where $w_{\text{contrastive}}$ indicates the weight of contrastive loss in the final loss computation.

When training the model, we minimize the loss for optimization through back-propagation using the Adam optimizer [229].

6.4 Evaluation

In this section, we conduct extensive experiments to examine the performance of the proposed defense model. Specifically, we aim to answer the following Research Questions:

- **RQ1:** Is the proposed model able to effectively identify malicious users on online platforms?
- **RQ2:** Under the attack setting, can the defense model successfully defend adversarial attacks and output the correct predictions?
- **RQ3:** What is the contribution of each module in the defense model towards its performance?

Datasets

Table 6.2: Dataset statistics used in the defense model

Dataset	Yelp	Wikipedia
Number of users	3,940	794
Number of benign users	2,016	397
Number of malicious users	1,924	397
Total number of posts	35,123	11,547
Median posts per user	9	15

We conducted our evaluation using real-world datasets from two widely-used platforms: Wikipedia and Yelp, with their detailed statistics presented in Table 6.2.

(a) **Wikipedia dataset:** Comprising user (or editor) contributions to Wikipedia articles, this dataset differentiates between benign and vandal editors, the latter of whom were identified and subsequently removed by Wikipedia administrators [213]. Each user’s contribution

history is recorded as a sequence of edits (i.e., an edit is a post in the context) made to articles, providing a basis for analysis.

(b) **Yelp dataset:** This dataset captures Yelp users’ reviews of restaurants, distinguishing between benign reviewers and fraudulent ones. Fraudulent reviewers are flagged by Yelp’s internal classification algorithms [210]. Here, a user sequence is composed of reviews (i.e., a review is a post in the context) made by the user, offering insights into their review patterns.

For both datasets, we implemented a filtering criterion to ensure the data’s richness and relevance: users with fewer than 5 posts and posts containing fewer than 5 tokens were excluded. To construct a user sequence, we utilized the most recent 20 posts from each user, aiming to capture the most current and relevant user behavior for our analysis.

Evaluation Metrics

To evaluate experiment results, we deploy several metrics to measure

(a) **Classification Performance:** Following the conventional measurement of classification models, we report the precision, recall, and F1-score.

(b) **Robustness Performance:** Similar to other adversarial attack and defense works [37], we report the F1 score after the attack, which will compare the F1 score before and after attack. If the resulting F1 score after the attack drops considerably, then the attack is successful.

Compared Classification Models

In this work, we compare three popular deep user sequence classification models and two defense models against adversarial attacks.

(1) **Hierarchical Recurrent Neural Network (HRNN)** [103] captures the sequential pattern of the input text by the hierarchical neural network structure for accurate classification. In HRNN, each user post is first transformed to a vector, and the sequence of user post

vectors is converted into a compact user embedding, which is finally used for user classification.

(2) **Temporal Interaction Embeddings (TIES)** [25] is developed and deployed by Facebook/Meta to detect malicious accounts. To adapt it in our sequence classification setting, we utilize the temporal embedding component of the TIES model for classification (as there is no graph structure in our datasets). Specifically, TIES converts a user post sequence into a user embedding vector by a sequence encoding layer and a pooling layer. (3) **Feature-based BERT (F-BERT)** [137] is a feature-based approach utilizing the advances of bidirectional transformers to encode and represent text. Here, to adapt to the sequence classification setting, we first obtain the BERT embedding for each post and pass the sequence of embedding to another layer transformer for classification.

(4) **Fine-Tuning-based defense (FT-Defense)** [223]: In this approach, we deploy the multi-stage fine-tuning approach based on adversarial attack data to improve the model robustness. Particularly, we first use the original data to train the model and have a frozen reference model. Next, we feed the adversarial attack data to fine-tune the model. In this step, for each data point, we have a prediction difference loss from the frozen reference model and active fine-tuned model to refine the model through back-propagation. Then, we repeat the fine-tuning process until convergence.

(5) **Mixup-based Data-Augmentation-based defense (MDA-Defense)** [224]: Instead of using the original and attack data in different steps as in FT-defense, this method will combine the two sets of data using the widely deployed mixup data augmentation strategy and use them to train the model together.

Experiment Setup

During the experiment, we split the dataset by five-fold cross-validation and report the average numbers. The number of tokens in a post is 30, and the learning rate is $1e - 3$.

We use Adam as the optimizer with the mini-batch size of 32 [221]. For the transformer encoder and decoder blocks, we set the embedding size and number of attention heads as 128 and 2.

RQ1: Effectiveness of Classifiers

In this section, we evaluate the defense model on Wikipedia and Yelp datasets. The classification result is presented in Table 6.3 and Table 6.4.

Table 6.3: Comparison of classification performance by different defense models on Wikipedia dataset

Method		Precision	Recall	F1
Sequence-based	HRNN	0.652	0.667	0.658
	TIES	0.689	0.645	0.666
	F-BERT	0.553	0.619	0.582
Defense-involved	FT-Defense	0.627	0.746	0.682
	MDA-Defense	0.683	0.667	0.671
Our Model		0.615	0.820	0.701

Table 6.4: Comparison of classification performance by different defense models on Yelp dataset

Method		Precision	Recall	F1
Sequence-based	HRNN	0.659	0.684	0.675
	TIES	0.680	0.687	0.683
	F-BERT	0.565	0.646	0.601
Defense-involved	FT-Defense	0.642	0.661	0.656
	MDA-Defense	0.677	0.700	0.688
Our Model		0.707	0.710	0.708

As we can see, our proposed model has the highest F1 score among all compared methods on two datasets, showing its superiority. Particularly, compared to sequence-based solutions, our model has the highest precision, recall, and F1 score in most cases except the precision is slightly lower on the Wikipedia dataset. The potential reason is the smaller data size of the Wikipedia dataset, which makes the training of our model less satisfactory. When trained and tested on the larger Yelp dataset, our model beats all sequence-based

approaches on all metrics of precision, recall, and F1 score. Second, when comparing our model with the defense-involved approaches, we have similar findings - our model is the best in most cases, especially on the larger Yelp dataset.

RQ2: Robustness of Classifiers

To evaluate the robustness of classifiers under adversarial attacks, we first mimic the behavior of malicious users by creating a new post using text generation models. Next, we concatenate the new post to the corresponding post sequence and feed it into the classification again. We report the mentioned F1 score after the attack in Section 6.4. In practice, we leverage the state-of-the-art text attack method against the deep sequence embedding-based classifiers (i.e., PETGEN [37]) to mimic the attack behaviors. On the other other, due to the advance of large language models (LLM), we also deploy LLM to generate the adversarial text for attack goals. In our experiment, we select the widely-used LLaMA [225] model. The F1 score after attack on Wikipedia and Yelp datasets are reported in Table 6.5 and Table 6.6, respectively.

Table 6.5: Comparison of robustness performance by different defense models on Wikipedia dataset. Here, we report the F1 score after the attack.

Method		Without Attack	PETGEN	LLaMA
Sequence-based	HRNN	0.658	0.605	0.628
	TIES	0.666	0.591	0.581
	F-BERT	0.582	0.573	0.570
Defense-involved	FT-Defense	0.682	0.620	0.631
	MDA-Defense	0.671	0.641	0.665
Our Model		0.701	0.695	0.697

The results show that after the attack, our model has the lowest decrease in F1 score and maintained the highest F1 score among all compared methods. This demonstrates the highest robustness of our our model solution. Particularly, compared to the sequence-based methods, our model only decreases 0.713% on average in F1 score while others worsen with even 12.76% decrease in F1 score. Second, when comparing with defense-involved

Table 6.6: Comparison of robustness performance by different defense models on Wikipedia dataset. Here, we report the F1 score after the attack.

Method		Without Attack	PETGEN	LLaMA
Sequence-based	HRNN	0.675	0.641	0.628
	TIES	0.683	0.661	0.659
	F-BERT	0.601	0.587	0.571
Defense-involved	FT-Defense	0.656	0.650	0.652
	MDA-Defense	0.688	0.673	0.664
Our Model		0.708	0.700	0.682

methods, we find that our model still works better even if MDA-Defense is the second best with only 2.68% decrease on average in F1 score. Interestingly, we also notice that defense-involved solutions are better than sequence-based ones when under attack. This is reasonable since defense-involved solutions explicitly take adversarial examples in the training stage to attend to the attack scenario.

RQ3: Ablation Studies

To examine the effectiveness of each module in our model, we conduct the ablation study where we test the performance of different variants of our model. Particularly, we examine:

- Our defense model: The full model with two modules.
- - w/o Local-Global Attended Module: We remove the local-global attended module in our model.
- - w/o Adversary-aware Module: We remove the adversary-aware module in our model

The results on Wikipedia and Yelp datasets are in Table 6.7 and Table 6.8, respectively.

As shown in the results, our model containing two modules always performs the best among all other variants when measured by the F1 score. When removing any module, the performance drops. This demonstrates the contribution of each module and the necessity of having two modules work together. Particularly, when comparing different variants, we find

Table 6.7: Ablation studies of our defense model on Wikipedia dataset, measured by F1 score.

Method	Without Attack	PETGEN	LLaMA
Our Model	0.701	0.695	0.697
- w/o Local-Global Attended Module	0.673	0.661	0.623
- w/o Adversary-aware Module	0.690	0.652	0.619

Table 6.8: Ablation studies of our defense model on Yelp dataset, measured by F1 score.

Method	Without Attack	PETGEN	LLaMA
Our model	0.708	0.700	0.682
- w/o Local-Global Attended Module	0.664	0.652	0.642
- w/o Adversary-aware Module	0.678	0.647	0.631

that removing the local-global attended module leads to 4.443% decrease on average while removing the adversary-aware module causes 3.878% decrease. The potential explanation is that the local-global attended module increases the model predictive capability while the adversary-aware module improves the model robustness. In our scenario of the user sequence classification, the local-global attended module weighs more since the predictive capability is the building block of a well-performed system.

6.5 Conclusion

In this chapter, we introduce a novel deep learning framework for user sequence classification, specifically designed to identify malicious users and bolster defenses against adversarial attacks. Our model applies a transformer encoder block to bidirectionally encode each post, creating detailed post embeddings that capture nuanced textual features. Subsequently, these embeddings are processed through a transformer decoder block. This block leverages an attention mechanism to discern and model the sequential patterns inherent in the post embeddings, culminating in the generation of sequence embeddings. Finally, the post embedding is passed through the contrastive learning-enhanced classification layer for the prediction task. Extensive results on two real-world datasets of Yelp and Wikipedia demonstrates the superiority of our proposed method.

CHAPTER 7

CONCLUDING REMARKS

In this concluding chapter, we discuss how our research can be used in a real-world setting. We also discuss its limitations and potential future research directions.

7.1 Opportunities for Real-World Impact

Misinformation has been a major challenge in our society for a long time. We believe multiple stakeholders may have a role in addressing it. Social media platforms like Facebook/Meta can be in the forefront to identify and limit the spread of misinformation. Government agencies, which have worked to educate citizens about cyber threats, can raise awareness about misinformation and encourage users to verify any claims and combat false information. Ordinary users who may not be experts can develop the mindset to recognize and counter misinformation, and our solutions can be used to empower them in this endeavor. Such education, awareness, and active countering efforts can enable us to fight against the threats posed by misinformation.

To ensure the real-world impact of our research in combating misinformation, first, it is necessary that strategies are developed to incentivize crowds to counter misinformation. Although we observed the positive role of crowds in combating misinformation, we also need to ensure that crowds indeed do it voluntarily. Some existing research works support such behavior [51, 57, 52]. Besides, it may be possible to design media literacy games where crowds are invited to play to learn about the harm of misinformation and the importance of countering misinformation. This can ensure that when they later see similar or the same misinformation, chances are high that they will proactively combat it. We can also turn to the friends of the misinformation spreaders, who may be more motivated to help and correct their friends. Second, we can direct our resources to social media (micro-

)influencers within the crowd, who have a larger audience than ordinary users. Particularly, if these social media influencers are encouraged to debunk popular misinformation with polite and evidence-based posts, the visibility of counter-misinformation is enhanced and the harm caused by misinformation can be mitigated. Third, we also admit that certain misinformation beliefs are deeply rooted in the minds of certain people. It is hard to change these beliefs whatever counter-misinformation we present to them. In this case, we can turn to the followed social media influencers or the friends of these people and invite them to counter the misinformation for their followers or friends. By doing so, we increase the credibility of the counter-misinformation and this can sometimes work. While the misinformation spreader's stance may not change in some cases, the observers of misinformation are less likely to believe the misinformation if it is debunked or countered.

Our proposed text generation model MisinfoCorrect can support several different use cases in real-world scenarios. It can be made available via a web portal or an API, where a user can input a misinformation post and our model will generate one or more counter-misinformation replies. As mentioned before, the social media influencers and the friends of misinformation spreaders can use MisinfoCorrect to debunk misinformation. Admittedly, one may wonder whether using the generated counter-responses can lead to online arguments. Our model is intended to encourage users who voluntarily and proactively already counter misinformation to do so in a polite and respectful manner – recall that in Chapter 2, we show 96% of all counter-misinformation responses are already generated by ordinary users, but 2 out of 3 times their responses are rude and abusive. Since our model generates polite responses, it has a lower chance of leading to online arguments. When using MisinfoCorrect in practice, to prevent generating unreasonable replies for unknown topics of misinformation or for non-misinformation tweets, we can add a filtering step so that the model will only output a counter-reply if the tweet is a misinformation post on the topic(s) it is trained on.

Another use case can be anticipated for the proposed attack model PETGEN. It can be

utilized by companies to do red-teaming exercises. Particularly, when companies are developing their malicious user detection models, they can use PETGEN to identify the vulnerabilities of their detectors and then improve their models. We should also take extreme caution since adversaries may use our PETGEN codebase for their gains. To alleviate this potential harm, we design the next-generation adversary-aware deep sequence embedding-based classifier in Chapter 6 where we use both state-of-the-art PETGEN attack model and large language models [225] for adversarial training. In this case, companies can directly use our proposed defense mechanism to improve their classifiers to defend against adversarial attacks.

Finally, we rely on AI to build user sequence classification and text generation models in our research. However, AI models can be biased to some degree. To address this issue, when we create the dataset to train AI models, diverse sources of such data must be used. Also, when evaluating the model, we should not only use automatic classifiers but also have human evaluations for fair and comprehensive comparison. These efforts can help reduce inaccurate and biased outputs.

7.2 Limitations

There are several limitations to our work. When leveraging AI to empower crowds to combat misinformation in Chapter 2, 3, and 4, we only focus on the Twitter platform when characterizing and assisting crowds who counter misinformation. Some previous research found that misinformation flows across multiple platforms [150]. Second, we only studied COVID-19-related misinformation topics ranging from 5G to fake cures to vaccines, but there could be variations across other areas (e.g., climate change). Third, we only focus on one language-English. The multi-lingual language models should be considered, especially for low-resource languages. Fourth, we only focus on textual information and do not investigate image, video, and audio content, e.g., visual counter-misinformation content. Generalization across platforms, topics, languages, and modalities is needed. When

researching misinformation, we do not collect data from professional fact-checkers (e.g., expert-annotated data points and expert-written counter-responses). This resource is valuable to potentially improve the quality of our work. When evaluating the quality of generated text by PETGEN and MisinfoCorrect, we rely on automatic classifiers. However, this machine-based evaluation has limits and could be faulty, e.g., false positives and negatives. This may lead to inaccurate comparison results between models. More human evaluations are needed for a comprehensive comparison. Limited real-world deployments can also help evaluate the model in practice. This is especially important for MisinfoCorrect because we still do not know whether the generated counter-misinformation responses can work in real-world Twitter conversations. One possible option is to conduct a field study on Twitter by replying to misinformation tweets with our generated responses and monitoring the behavior change of tweet posters or followers (e.g., whether the poster deletes misinformation posts or the posters and followers share less misinformation after viewing the counter-response.)

In our evaluation of the deep sequence embedding-based classification models through our proposed PETGEN attack model, PETGEN is currently only applicable for posts in the English language, while social media posts can be in any language. Second, PETGEN can only work with sequences, while it does not incorporate complex structures, such as graphs. Third, the attack is restricted to generating new posts.

For the proposed defense model, our work also has some limitations. A primary constraint is the model’s current design, which only accommodates English language posts, potentially overlooking the rich diversity of global social media discourse that encompasses a multitude of languages. Additionally, the model’s architecture is tailored exclusively for sequential data, thereby omitting the potential insights that could be gleaned from more complex data structures like graphs. This focus on sequences also means the model’s defensive capabilities are specifically tuned to counter the creation of new posts by attackers, leaving room for future enhancements to address a broader spectrum of adversarial strate-

gies, including but not limited to, more sophisticated manipulation techniques that might exploit other aspects of user behavior or platform interaction. Besides, we examine our method on relatively small datasets (Note that it is still the largest public dataset in the existing literature on bad actor detection sequential models). These datasets should be enriched in the future.

7.3 Conclusions

In this thesis, we presented diverse state-of-the-art AI techniques to combat misinformation with the twin goals of empowering crowds and evaluating detectors. Specifically, in Chapter 2, to understand the role of crowds who counter misinformation, we first characterize these crowds by analyzing the crowd-generated counter-misinformation contents and characteristics of these crowds. We then investigate the user responses to these crowd-generated counter-misinformation replies in Chapter 3. Among our findings, we note that 2/3 of crowd-generated counter-misinformation is rude or non-evidenced. To address this issue, we propose MisinfoCorrect, a reinforcement-learning-based text generation framework that can generate polite, evidenced, and refuting counter-response in Chapter 4. Besides relying on crowds, we also turn to automatic classifiers to combat misinformation. Given that the vulnerability of these classifiers is rarely studied, we examined the robustness of existing deep sequence embedding-based detectors and devised PETGEN, an end-to-end personalized text generation framework that can both successfully attack the existing detectors and generate high-quality text in Chapter 5. This demonstrates the vulnerability of the existing detection systems. To address this vulnerability issue, we propose a transformer-based adversary-aware local-global attended framework to improve the robustness of the detectors in Chapter 6.

Through these efforts, we demonstrate that it is feasible to use AI to combat misinformation by empowering crowds and evaluating detectors.

7.4 Future Work

There are still ongoing important research problems that have not been addressed in this thesis.

- **Multi-platform and Multimodal Countering:** Current social media-related work predominantly focuses on a few platforms like Twitter [96, 230, 231] and Sina Weibo [77]. However, crowds countering misinformation may behave differently across various platforms due to variations in user demographics and engagement dynamics. Exploring how crowds counter misinformation across multiple platforms and whether countering on one platform influences others is essential for a comprehensive understanding of crowd-driven misinformation mitigation. Additionally, the crowd-generated counter-misinformation is not limited to text alone; it can also involve images or videos to enhance the persuasiveness of their debunking efforts. Investigating these multimodal aspects benefits the design of effective countering content.
- **Multilingual and Topic-specific Countering:** Most research works concentrate on either a single language (e.g., English [231] or Chinese [77]) or a specific misinformation topic (e.g., COVID-19 [50]). However, misinformation spans many languages and topics, leading to the need for diverse countering actions. Analyzing how crowds in under-representative languages combat various topics reveals variations in countering strategies across languages and topics, thus contributing to a more comprehensive understanding of countering misinformation. On the other hand, existing research on the profiles of crowds focuses on demographic factors such as education, political leanings, and media literacy. However, it overlooks misinformation topic-specific factors. For instance, an individual having a background in health education may be effective at countering health misinformation but susceptible to believing in climate change misinformation. Therefore, exploring these topic-specific factors can

enhance our understanding of human factors involved in countering misinformation.

- **Comprehensive User Response to Counter-misinformation by Crowds:** We admit the limited number of cases where one user replies to misinformation tweets for countering, and another user responds to the counter-misinformation replies. However, if we crawl large-scale datasets, this will not be an issue. Additionally, the reply-to-reply signal is either positive or negative while likes and retweets indicate endorsement. In the future, we can consider combining user engagements (e.g., likes and retweets of counter-replies) with user responses 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 those responding to the counter-replies to investigate the potential phenomenon of networked “echo chamber” [153]. 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.
- **Enhanced Counter-misinformation Text Generation** To improve the text generation results for countering misinformation, we can (i) deploy and evaluate the model in practice, (ii) collect data from professional fact-checkers as expert-generated counter-responses and compare the model performance against the current setup in Misinfo-Correct, and (iii) develop a multi-lingual and multi-modal model to generate visual counter-responses.
- **Field study of crowd-generated and machine-generated counter-misinformation**
The field study to examine the effect of crowd-generated and machine-generated

counter-misinformation (e.g., on mental health [232]) is needed to confirm the impact of counter-misinformation beyond our large-scale data-driven study.

- **Extended Next Post Attack Beyond Sequences:** The PETGEN model can only work with sequences, while it does not incorporate complex structures, such as graphs. Furthermore, the attack is limited to generating new posts. Other attack capabilities can be explored in the future (e.g., considering the user connections).
- **Opportunities from Large Language Models:** We can use large language models to assist users in drafting high-quality and personalized counter-misinformation replies, or mimic bad actors to write human-like posts for the red-teaming test when building the robust malicious user detection model.

REFERENCES

- [1] M. Walker and K. E. Matsa, “News consumption across social media in 2021,” 2021.
- [2] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
- [3] D. M. Lazer *et al.*, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
- [4] F. Pierri *et al.*, “Online misinformation is linked to early covid-19 vaccination hesitancy and refusal,” *Scientific reports*, vol. 12, no. 1, pp. 1–7, 2022.
- [5] P. Ball and A. Maxmen, *The epic battle against coronavirus misinformation and conspiracy theories*, <https://www.nature.com/articles/d41586-020-01452-z>, May 2020.
- [6] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason, “Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing,” *IConf. 2014 Proc.*, 2014.
- [7] A. Arif, L. G. Stewart, and K. Starbird, “Acting the part: Examining information operations within #BlackLivesMatter discourse,” *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, Nov. 2018.
- [8] L. G. Stewart, A. Arif, and K. Starbird, “Examining trolls and polarization with a retweet network,” in *Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web*, vol. 70, 2018.
- [9] G. Verma, A. Bhardwaj, T. Aledavood, M. De Choudhury, and S. Kumar, “Examining the impact of sharing covid-19 misinformation online on mental health,” *Scientific Reports*, vol. 12, no. 1, pp. 1–9, 2022.
- [10] C. Fuchs and C. Fuchs, “Bill gates conspiracy theories as ideology in the context of the covid-19 crisis,” 2021.
- [11] S. Kumar and N. Shah, “False information on web and social media: A survey,” *arXiv preprint arXiv:1804.08559*, 2018.
- [12] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, “Misinformation in social media: Definition, manipulation, and detection,” *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 2, pp. 80–90, 2019.

- [13] N. Micallef, B. He, S. Kumar, M. Ahamad, and N. Memon, “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, 2020, 748–757 (The first two authors are the co first-authors).
- [14] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, “Misinformation and its correction: Continued influence and successful debiasing,” *Psychological science in the public interest*, vol. 13, no. 3, pp. 106–131, 2012.
- [15] M. Goindani and J. Neville, “Social reinforcement learning to combat fake news spread,” R. P. Adams and V. Gogate, Eds., vol. 115, PMLR, Jan. 2020, pp. 1006–1016.
- [16] C. Budak, D. Agrawal, and A. E. Abbadi, “Limiting the spread of misinformation in social networks,” ACM Press, 2011, p. 665, ISBN: 9781450306324.
- [17] J. Zhu, S. Ghosh, and W. Wu, “Robust rumor blocking problem with uncertain rumor sources in social networks,” *World Wide Web*, vol. 24, pp. 229–247, 1 Jan. 2021.
- [18] I. Litou, V. Kalogeraki, I. Katakis, and D. Gunopulos, “Efficient and timely misinformation blocking under varying cost constraints,” *Online Social Networks and Media*, vol. 2, pp. 19–31, Aug. 2017.
- [19] Z. Wang and Y. Guo, “Rumor events detection enhanced by encoding sentimental information into time series division and word representations,” *Neurocomputing*, vol. 397, pp. 224–243, Jul. 2020.
- [20] M. M. Haque, M. Yousuf, A. S. Alam, P. Saha, S. I. Ahmed, and N. Hassan, “Combating misinformation in bangladesh: Roles and responsibilities as perceived by journalists, fact-checkers, and users,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–32, 2020.
- [21] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand, “Scaling up fact-checking using the wisdom of crowds,” *Science Advances*, vol. 7, 36 Sep. 2021.
- [22] Y. Wang *et al.*, “Weak supervision for fake news detection via reinforcement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 516–523.
- [23] Y. Wang *et al.*, “Eann: Event adversarial neural networks for multi-modal fake news detection,” in *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 2018, pp. 849–857.

- [24] S. Antoniadis, I. Litou, and V. Kalogeraki, “A model for identifying misinformation in online social networks,” in *OTM Confederated International Conferences” On the Move to Meaningful Internet Systems”*, Springer, 2015, pp. 473–482.
- [25] N. Noorshams, S. Verma, and A. Hofleitner, “Ties: Temporal interaction embeddings for enhancing social media integrity at facebook,” in *SIGKDD*, 2020, pp. 3128–3135.
- [26] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” in *EuroS&P*, IEEE, 2016, pp. 372–387.
- [27] L. Sun *et al.*, “Adversarial attack and defense on graph data: A survey,” *IEEE Transactions on Knowledge and Data Engineering*, 2022.
- [28] Z. Zhou, H. Guan, M. M. Bhat, and J. Hsu, “Fake news detection via nlp is vulnerable to adversarial attacks,” *arXiv preprint arXiv:1901.09657*, 2019.
- [29] L. Chen, P. Chen, and Z. Lin, “Artificial intelligence in education: A review,” *Ieee Access*, vol. 8, pp. 75 264–75 278, 2020.
- [30] K.-H. Yu, A. L. Beam, and I. S. Kohane, “Artificial intelligence in healthcare,” *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.
- [31] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, “Combating fake news: A survey on identification and mitigation techniques,” *ACM Transactions on Intelligent Systems and Technology*, vol. 10, pp. 1–42, 3 May 2019.
- [32] B. Guo, Y. Ding, L. Yao, Y. Liang, and Z. Yu, “The future of misinformation detection: New perspectives and trends,” *arXiv preprint arXiv:1909.03654*, 2019.
- [33] W. Shahid, Y. Li, D. Staples, G. Amin, S. Hakak, and A. Ghorbani, “Are you a cyborg, bot or human?—a survey on detecting fake news spreaders,” *IEEE Access*, vol. 10, pp. 27 069–27 083, 2022.
- [34] S. Shaar, N. Georgiev, F. Alam, G. Da San Martino, A. Mohamed, and P. Nakov, “Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 2069–2080.
- [35] X. Zeng, A. S. Abumansour, and A. Zubiaga, “Automated fact-checking: A survey,” *Language and Linguistics Compass*, vol. 15, no. 10, e12438, 2021.

- [36] B. He, M. Ahamad, and S. Kumar, “Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation,” in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2698–2709.
- [37] B. He, M. Ahamad, and S. Kumar, “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*, 2021, pp. 575–584.
- [38] A. Friggeri, L. Adamic, D. Eckles, and J. Cheng, “Rumor cascades,” in *8th International AAAI Conference on Weblogs and Social Media*, 2014.
- [39] S. Kumar, R. West, and J. Leskovec, “Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes,” in *Proceedings of the 25th international conference on World Wide Web*, 2016, pp. 591–602.
- [40] E. K. Vraga and L. Bode, “Using expert sources to correct health misinformation in social media,” *Science Communication*, vol. 39, no. 5, pp. 621–645, 2017.
- [41] T. G. van der Meer and Y. Jin, “Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source,” *Health Communication*, vol. 35, pp. 560–575, 2020.
- [42] L. Bode, E. K. Vraga, and M. Tully, “Do the right thing: Tone may not affect correction of misinformation on social media,” *Harvard Kennedy School Misinformation Review*, 2020.
- [43] I. Litou, V. Kalogeraki, I. Katakis, and D. Gunopulos, “Efficient and timely misinformation blocking under varying cost constraints,” *Online Social Networks and Media*, vol. 2, pp. 19–31, 2017.
- [44] M. M. Bhuiyan, A. X. Zhang, C. M. Sehat, and T. Mitra, “Investigating “who” in the crowdsourcing of news credibility,” in *Computational Journalism Symposium*, 2020.
- [45] K. Roitero *et al.*, “The covid-19 infodemic: Can the crowd judge recent misinformation objectively?” In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1305–1314.
- [46] M. Mendoza, B. Poblete, and C. Castillo, “Twitter under crisis: Can we trust what we rt?” In *Proceedings of the first workshop on social media analytics*, 2010, pp. 71–79.

- [47] P. Borah, B. Irom, and Y. C. Hsu, “‘it infuriates me’: Examining young adults’ reactions to and recommendations to fight misinformation about covid-19,” *Journal of Youth Studies*, pp. 1–21, Aug. 2021.
- [48] J. Kirchner and C. Reuter, “Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, pp. 1–27, CSCW2 Oct. 2020.
- [49] M. Tully, L. Bode, and E. K. Vraga, “Mobilizing users: Does exposure to misinformation and its correction affect users’ responses to a health misinformation post?” *Social Media + Society*, vol. 6, p. 205 630 512 097 837, 4 Oct. 2020.
- [50] J. Veeriah, “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*, vol. 13, pp. 372–383, 7 2021.
- [51] H. Seo, A. Xiong, S. Lee, and D. Lee, “If you have a reliable source, say something: Effects of correction comments on covid-19 misinformation,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 896–907.
- [52] J. Colliander, “‘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*, vol. 97, pp. 202–215, 2019.
- [53] S. Wijenayake, D. Hettiachchi, S. Hosio, V. Kostakos, and J. Goncalves, “Effect of conformity on perceived trustworthiness of news in social media,” *IEEE Internet Computing*, vol. 25, pp. 12–19, 1 Jan. 2021.
- [54] E. K. Vraga and L. Bode, “I do not believe you: How providing a source corrects health misperceptions across social media platforms,” *Information, Communication & Society*, vol. 21, no. 10, pp. 1337–1353, 2018.
- [55] E. K. Vraga and L. Bode, “Addressing covid-19 misinformation on social media preemptively and responsively,” *Emerging infectious diseases*, vol. 27, no. 2, p. 396, 2021.
- [56] L. Bode and E. K. Vraga, “In Related News, That was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media,” *Journal of Communication*, vol. 65, no. 4, pp. 619–638, Jun. 2015. eprint: <https://academic.oup.com/joc/article-pdf/65/4/619/22320531/jjnlcom0619.pdf>.
- [57] L. Bode and E. K. Vraga, “See something, say something: Correction of global health misinformation on social media,” *Health Communication*, vol. 33, pp. 1131–1140, 9 Sep. 2018.

- [58] E. K. Vraga and L. Bode, *Correction as a solution for health misinformation on social media*, 2020.
- [59] E. Vraga, M. Tully, and L. Bode, “Assessing the relative merits of news literacy and corrections in responding to misinformation on twitter,” *New Media & Society*, p. 1 461 444 821 998 691, 2021.
- [60] E. K. Vraga, S. C. Kim, J. Cook, and L. Bode, “Testing the effectiveness of correction placement and type on instagram,” *The International Journal of Press/Politics*, vol. 25, no. 4, pp. 632–652, 2020.
- [61] E. K. Vraga, L. Bode, and M. Tully, “The effects of a news literacy video and real-time corrections to video misinformation related to sunscreen and skin cancer,” *Health communication*, pp. 1–9, 2021.
- [62] M. S. Steffens, A. G. Dunn, K. E. Wiley, and J. Leask, “How organisations promoting vaccination respond to misinformation on social media: A qualitative investigation,” *BMC public health*, vol. 19, no. 1, pp. 1–12, 2019.
- [63] P. Malhotra, K. Scharp, and L. Thomas, “The meaning of misinformation and those who correct it: An extension of relational dialectics theory,” *Journal of Social and Personal Relationships*, vol. 39, no. 5, pp. 1256–1276, 2022.
- [64] Y. Tanaka and R. Hirayama, “Exposure to countering messages online: Alleviating or strengthening false belief?” *Cyberpsychology, Behavior, and Social Networking*, vol. 22, pp. 742–746, 11 Nov. 2019.
- [65] G. Orosz, P. Krekó, B. Paskuj, I. Tóth-Király, B. Bóthe, and C. Roland-Lévy, “Changing conspiracy beliefs through rationality and ridiculing,” *Frontiers in Psychology*, vol. 7, Oct. 2016.
- [66] A. Stojanov, “Reducing conspiracy theory beliefs,” *Psihologija*, vol. 48, pp. 251–266, 3 2015.
- [67] M.-p. S. Chan, C. R. Jones, K. Hall Jamieson, and D. Albarracín, “Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation,” *Psychological science*, vol. 28, no. 11, pp. 1531–1546, 2017.
- [68] D. B. Margolin, A. Hannak, and I. Weber, “Political fact-checking on twitter: When do corrections have an effect?” *Political Communication*, vol. 35, no. 2, pp. 196–219, 2018.
- [69] S. Grandhi, L. Plotnick, and S. R. Hiltz, “By the crowd and for the crowd: Perceived utility and willingness to contribute to trustworthiness indicators on social

media,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 1–24, GROUP Jul. 2021.

- [70] M. Mosleh, C. Martel, D. Eckles, and D. Rand, “Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment,” in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.
- [71] A. Pal, Y. Alton, *et al.*, “Rumor analysis & visualization system,” in *Proceedings of the international multi conference of engineers and computer scientists*, 2019.
- [72] Y. Sun, J. Oktavianus, S. Wang, and F. Lu, “The role of influence of presumed influence and anticipated guilt in evoking social correction of covid-19 misinformation,” *Health Communication*, pp. 1–10, Feb. 2021.
- [73] N. Pröllochs, “Community-based fact-checking on twitter’s birdwatch platform,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 794–805.
- [74] K. Miyazaki, T. Uchiba, K. Tanaka, J. An, H. Kwak, and K. Sasahara, ““ this is fake news”: Characterizing the spontaneous debunking from twitter users to covid-19 false information,” *arXiv preprint arXiv:2203.14242*, 2022.
- [75] C. Drolsbach and N. Pröllochs, “Diffusion of community fact-checked misinformation on twitter,” *arXiv preprint arXiv:2205.13673*, 2022.
- [76] J. Allen, C. Martel, and D. G. Rand, “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*, 2022, pp. 1–19.
- [77] X. Wang, F. Chao, G. Yu, and K. Zhang, “Factors influencing fake news rebuttal acceptance during the covid-19 pandemic and the moderating effect of cognitive ability,” *Computers in human behavior*, vol. 130, p. 107 174, 2022.
- [78] Y. Zhang *et al.*, “Investigation of the determinants for misinformation correction effectiveness on social media during covid-19 pandemic,” *Information Processing & Management*, vol. 59, no. 3, p. 102 935, 2022.
- [79] X. Wang, F. Chao, and G. Yu, “Evaluating rumor debunking effectiveness during the covid-19 pandemic crisis: Utilizing user stance in comments on sina weibo,” *Frontiers in Public Health*, vol. 9, p. 770 111, 2021.

- [80] Y. Ma, B. He, N. Subrahmanian, and S. Kumar, “Characterizing and predicting social correction on twitter,” in *15th ACM Web Science Conference 2023*, 2023.
- [81] Y. Chuai, H. Tian, N. Pröllochs, and G. Lenzini, “The roll-out of community notes did not reduce engagement with misinformation on twitter,” *arXiv preprint arXiv:2307.07960*, 2023.
- [82] B. Nyhan and J. Reifler, “When corrections fail: The persistence of political misperceptions,” *Political Behavior*, vol. 32, no. 2, pp. 303–330, 2010.
- [83] P. Schmid and C. Betsch, “Benefits and pitfalls of debunking interventions to counter mRNA vaccination misinformation during the covid-19 pandemic,” *Science Communication*, vol. 44, no. 5, pp. 531–558, 2022.
- [84] C. Peter and T. Koch, “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*, vol. 38, no. 1, pp. 3–25, 2016.
- [85] B. Swire-Thompson, J. DeGutis, and D. Lazer, “Searching for the backfire effect: Measurement and design considerations,” *Journal of applied research in memory and cognition*, vol. 9, no. 3, pp. 286–299, 2020.
- [86] B. Wang and J. Zhuang, “Rumor response, debunking response, and decision makings of misinformed twitter users during disasters,” *Natural Hazards*, vol. 93, pp. 1145–1162, 2018.
- [87] N. Walter, J. J. Brooks, C. J. Saucier, and S. Suresh, “Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis,” *Health Communication*, vol. 36, no. 13, pp. 1776–1784, 2021.
- [88] N. Walter and S. T. Murphy, “How to unring the bell: A meta-analytic approach to correction of misinformation,” *Communication Monographs*, vol. 85, no. 3, pp. 423–441, 2018.
- [89] E. Porter and T. J. Wood, “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*, vol. 118, no. 37, e2104235118, 2021.
- [90] N. Vo and K. Lee, “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*, 2019, pp. 335–344.

- [91] N. Vo and K. Lee, “Standing on the shoulders of guardians: Novel methodologies to combat fake news,” in *Disinformation, Misinformation, and Fake News in Social Media*, Springer, 2020, pp. 183–210.
- [92] N. Micalef, M. Sandoval-Castañeda, A. Cohen, M. Ahamad, S. Kumar, and N. Memon, “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, 2022, pp. 651–662.
- [93] S. S. Tekiroglu, Y.-L. Chung, and M. Guerini, “Generating counter narratives against online hate speech: Data and strategies,” *arXiv preprint arXiv:2004.04216*, 2020.
- [94] W. Zhu and S. Bhat, “Generate, prune, select: A pipeline for counterspeech generation against online hate speech,” *arXiv preprint arXiv:2106.01625*, 2021.
- [95] Y.-L. Chung, S. S. Tekiroglu, and M. Guerini, “Towards knowledge-grounded counter narrative generation for hate speech,” *arXiv preprint arXiv:2106.11783*, 2021.
- [96] B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar, “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*, 2021, pp. 90–94.
- [97] M. Alshomary, S. Syed, A. Dhar, M. Potthast, and H. Wachsmuth, “Counter-argument generation by attacking weak premises,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1816–1827.
- [98] X. Hua, Z. Hu, and L. Wang, “Argument generation with retrieval, planning, and realization,” *arXiv preprint arXiv:1906.03717*, 2019.
- [99] C. Hidey and K. McKeown, “Fixed that for you: Generating contrastive claims with semantic edits,” 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)*, 2019, pp. 1756–1767.
- [100] B. Schiller, J. Daxenberger, and I. Gurevych, “Aspect-controlled neural argument generation,” *arXiv preprint arXiv:2005.00084*, 2020.
- [101] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, “A benchmark dataset for learning to intervene in online hate speech,” *arXiv preprint arXiv:1909.04251*, 2019.
- [102] A. Saakyan, T. Chakrabarty, and S. Muresan, “Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic,” *arXiv preprint arXiv:2106.03794*, 2021.

- [103] Y. Zhao, Y. Shen, and J. Yao, “Recurrent neural network for text classification with hierarchical multiscale dense connections,” in *IJCAI*, 2019.
- [104] S. Kumar, X. Zhang, and J. Leskovec, “Predicting dynamic embedding trajectory in temporal interaction networks,” *SIGKDD*, 2019.
- [105] J. Y. Lee and F. Dernoncourt, “Sequential short-text classification with recurrent and convolutional neural networks,” *NAACL*, 2016.
- [106] C. Esposito, V. Moscato, and G. Sperli, “Detecting malicious reviews and users affecting social reviewing systems: A survey,” *Computers & Security*, vol. 133, p. 103 407, 2023.
- [107] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *NAACL*, 2016.
- [108] A. Vaswani *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [109] X. Dai, I. Chalkidis, S. Darkner, and D. Elliott, “Revisiting transformer-based models for long document classification,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 7212–7230.
- [110] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint arXiv:2004.05150*, 2020.
- [111] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, “Adversarial attacks on deep-learning models in natural language processing: A survey,” *ACM TIST*, vol. 11, no. 3, pp. 1–41, 2020.
- [112] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, “Hotflip: White-box adversarial examples for text classification,” *ACL*, 2017.
- [113] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, “Universal adversarial triggers for attacking and analyzing NLP,” in *EMNLP*, 2019.
- [114] J. Li, S. Ji, T. Du, B. Li, and T. Wang, “Textbugger: Generating adversarial text against real-world applications,” 2018.
- [115] X. Jia, S. Li, H. Zhao, S. Kim, and V. Kumar, “Towards robust and discriminative sequential data learning: When and how to perform adversarial training?” In *SIGKDD*, 2019, pp. 1665–1673.
- [116] D. Pruthi, B. Dhingra, and Z. C. Lipton, “Combating adversarial misspellings with robust word recognition,” *ACL*, 2019.

- [117] W. Nie, N. Narodytska, and A. Patel, “Relgan: Relational generative adversarial networks for text generation,” in *ICLR*, 2018.
- [118] T. Le, S. Wang, and D. Lee, “Malcom: Generating malicious comments to attack neural fake news detection models,” *ICDM*, 2020.
- [119] H.-Y. Chiang, Y.-S. Chen, Y.-Z. Song, H.-H. Shuai, and J. S. Chang, “Shilling black-box review-based recommender systems through fake review generation,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 286–297.
- [120] J. Wang *et al.*, “Adversarial attack generation empowered by min-max optimization,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 020–16 033, 2021.
- [121] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [122] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *International conference on machine learning*, PMLR, 2018, pp. 274–283.
- [123] P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-gan: Protecting classifiers against adversarial attacks using generative models,” *arXiv preprint arXiv:1805.06605*, 2018.
- [124] X. Liu, P. He, W. Chen, and J. Gao, “Multi-task deep neural networks for natural language understanding,” *arXiv preprint arXiv:1901.11504*, 2019.
- [125] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, “Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp,” *arXiv preprint arXiv:2005.05909*, 2020.
- [126] A. L. Wintersieck, “Debating the truth: The impact of fact-checking during electoral debates,” *American Politics Research*, vol. 45, no. 2, pp. 304–331, 2017.
- [127] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, “An exploratory study of covid-19 misinformation on twitter,” *arXiv:2005.05710*, 2020.
- [128] D. M. J. Lazer *et al.*, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
- [129] J. S. Brennen, F. Simon, P. N. Howard, and R. K. Nielsen, “Types, sources, and claims of covid-19 misinformation,” *Reuters Institute*, vol. 7, 2020.

- [130] E. Chen, K. Lerman, and E. Ferrara, “Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set,” *JMIR Public Health Surveill*, vol. 6, no. 2, May 2020.
- [131] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Singh, and S. Young, “Detecting covid-19 misinformation on social media,” 2020.
- [132] K. Sharma, S. Seo, C. Meng, S. Rambhatla, and Y. Liu, “Covid-19 on social media: Analyzing misinformation in twitter conversations,” *arXiv preprint arXiv:2003.12309*, 2020.
- [133] C. Schuster, “A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales,” *Educational and Psychological Measurement*, vol. 64, no. 2, pp. 243–253, 2004.
- [134] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie, “Towards detecting rumours in social media,” in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [135] F. Alam *et al.*, “Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society,” *arXiv preprint arXiv:2005.00033*, 2020.
- [136] V. Qazvinian, E. Rosengren, D. Radev, and Q. Mei, “Rumor has it: Identifying misinformation in microblogs,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1589–1599.
- [137] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [138] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proc. of the 2014 conference on empirical methods in natural language processing*, 2014, pp. 1532–1543.
- [139] E. A. Freeman and G. G. Moisen, “A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa,” *Ecological modelling*, vol. 217, no. 1-2, pp. 48–58, 2008.
- [140] W. Chen, D. Pacheco, K.-C. Yang, and F. Menczer, “Neutral bots reveal political bias on social media,” *arXiv preprint arXiv:2005.08141*, 2020.
- [141] C. Shao *et al.*, “Anatomy of an online misinformation network,” *PLoS ONE*, vol. 13, no. 4, e0196087, 2018.

- [142] M. Avram, N. Micallef, S. Patil, and F. Menczer, “Exposure to social engagement metrics increases vulnerability to misinformation,” *Harvard Kennedy School Misinformation Review*, Jul. 2020.
- [143] J. P. Chang, C. Chiam, L. Fu, A. Z. Wang, J. Zhang, and C. Danescu-Niculescu-Mizil, “Convokit: A toolkit for the analysis of conversations,” *arXiv preprint arXiv:2005.04246*, 2020.
- [144] O. Ajao, D. Bhowmik, and S. Zargari, “Sentiment aware fake news detection on online social networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2507–2511.
- [145] M. Maros and L. Rosli, “Politeness strategies in twitter updates of female english language studies malaysian undergraduates,” *3L: Language, Linguistics, Literature®*, vol. 23, no. 1, 2017.
- [146] P. Brown, S. C. Levinson, and S. C. Levinson, *Politeness: Some universals in language usage*. Cambridge university press, 1987, vol. 4.
- [147] S. Jiang and C. Wilson, “Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media,” *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, Nov. 2018.
- [148] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: Liwc 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.
- [149] S. A. Memon and K. M. Carley, “Characterizing covid-19 misinformation communities using a novel twitter dataset,” *arXiv preprint arXiv:2008.00791*, 2020.
- [150] M. Cinelli *et al.*, “The covid-19 social media infodemic,” *arXiv preprint arXiv:2003.05004*, 2020.
- [151] B. He, Y. Ma, M. Ahamad, and S. Kumar, “Corrective or backfire: Characterizing and predicting user response to social correction,” in *Proceedings of the 16th ACM Web Science Conference*, 2024, pp. 149–158.
- [152] B. Swire-Thompson, N. Miklaucic, J. P. Wihbey, D. Lazer, and J. DeGutis, “The backfire effect after correcting misinformation is strongly associated with reliability,” *Journal of Experimental Psychology: General*, vol. 151, no. 7, p. 1655, 2022.
- [153] M. Del Vicario *et al.*, “Echo chambers in the age of misinformation,” *arXiv preprint arXiv:1509.00189*, 2015.

- [154] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew, “Anti-vax: A novel twitter dataset for covid-19 vaccine misinformation detection,” *Public health*, vol. 203, pp. 23–30, 2022.
- [155] J. Abbasi, “Widespread misinformation about infertility continues to create covid-19 vaccine hesitancy,” *JAMA*, vol. 327, no. 11, pp. 1013–1015, 2022.
- [156] S. Jiang, M. Metzger, A. Flanagan, and C. Wilson, “Modeling and measuring expressed (dis) belief in (mis) information,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 315–326.
- [157] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, and S. Singh, “Covidlies: Detecting covid-19 misinformation on social media,” 2020.
- [158] F. Gilardi, M. Alizadeh, and M. Kubli, “Chatgpt outperforms crowd-workers for text-annotation tasks,” *arXiv preprint arXiv:2303.15056*, 2023.
- [159] C. Ziemis, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, “Can large language models transform computational social science?” *arXiv preprint arXiv:2305.03514*, 2023.
- [160] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, “Gpteval: A survey on assessments of chatgpt and gpt-4,” *arXiv preprint arXiv:2308.12488*, 2023.
- [161] G. Buchanan, R. Kelly, S. Makri, and D. McKay, “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*, 2022, pp. 243–253.
- [162] C. Wittenberg and A. J. Berinsky, “Misinformation and its correction,” *Social media and democracy: The state of the field, prospects for reform*, vol. 163, 2020.
- [163] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *ICWSM*, vol. 8, 2014.
- [164] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, “A computational approach to politeness with application to social factors,” *arXiv preprint arXiv:1306.6078*, 2013.
- [165] U. K. Ecker *et al.*, “The psychological drivers of misinformation belief and its resistance to correction,” *Nature Reviews Psychology*, vol. 1, no. 1, pp. 13–29, 2022.
- [166] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, “Tweeting is believing? understanding microblog credibility perceptions,” in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 441–450.

- [167] D. Boyd, S. Golder, and G. Lotan, “Tweet, tweet, retweet: Conversational aspects of retweeting on twitter,” in *2010 43rd Hawaii international conference on system sciences*, IEEE, 2010, pp. 1–10.
- [168] M. Bayer, W. Sommer, and A. Schacht, “Reading emotional words within sentences: The impact of arousal and valence on event-related potentials,” *International Journal of Psychophysiology*, vol. 78, no. 3, pp. 299–307, 2010.
- [169] H. M. Zahera, I. A. Elgendy, R. Jalota, and M. A. Sherif, “Fine-tuned bert model for multi-label tweets classification.,” in *TREC*, 2019, pp. 1–7.
- [170] Y. Liu *et al.*, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [171] N. Vo and K. Lee, “The rise of guardians: Fact-checking url recommendation to combat fake news,” in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 275–284.
- [172] X. Zhou, K. Shu, V. V. Phoha, H. Liu, and R. Zafarani, ““this is fake! shared it by mistake”: Assessing the intent of fake news spreaders,” in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3685–3694.
- [173] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, “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*, 2018, pp. 324–332.
- [174] N. Walter, J. Cohen, R. L. Holbert, and Y. Morag, “Fact-checking: A meta-analysis of what works and for whom,” *Political Communication*, vol. 37, no. 3, pp. 350–375, 2020.
- [175] A. Guess and A. Coppock, “Does counter-attitudinal information cause backlash? results from three large survey experiments,” *British Journal of Political Science*, vol. 50, no. 4, pp. 1497–1515, 2020.
- [176] T. Wood and E. Porter, “The elusive backfire effect: Mass attitudes’ steadfast factual adherence,” *Political Behavior*, vol. 41, no. 1, pp. 135–163, 2019.
- [177] L. Bode and E. K. Vraga, “Correction experiences on social media during covid-19,” *Social Media + Society*, vol. 7, p. 205 630 512 110 088, 2 Apr. 2021.
- [178] H. Seo, A. Xiong, S. Lee, and D. Lee, “(in) effectiveness of accumulated correction on covid-19 misinformation,” 2021.

- [179] L. Flekova, D. Preoțiu-Pietro, and L. Ungar, “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)*, 2016, pp. 313–319.
- [180] K. Thorson, E. Vraga, and B. Ekdale, “Credibility in context: How uncivil online commentary affects news credibility,” *Mass Communication and Society*, vol. 13, no. 3, pp. 289–313, 2010.
- [181] G. M. Masullo and J. Kim, “Exploring “angry” and “like” reactions on uncivil facebook comments that correct misinformation in the news,” *Digital Journalism*, vol. 9, pp. 1103–1122, 8 Sep. 2021.
- [182] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, “Community interaction and conflict on the web,” in *Proceedings of the 2018 world wide web conference*, 2018, pp. 933–943.
- [183] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Anyone can become a troll: Causes of trolling behavior in online discussions,” in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 1217–1230.
- [184] K. Sharma, Y. Zhang, and Y. Liu, “Covid-19 vaccine misinformation campaigns and social media narratives,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 920–931.
- [185] S. Evanega, M. Lynas, J. Adams, K. Smolenyak, and C. G. Insights, “Coronavirus misinformation: Quantifying sources and themes in the covid-19 ‘infodemic’,” *JMIR Preprints*, vol. 19, no. 10, p. 2020, 2020.
- [186] A. L. Hsu, T. Johnson, L. Phillips, and T. B. Nelson, “Sources of vaccine hesitancy: Pregnancy, infertility, minority concerns, and general skepticism,” in *Open forum infectious diseases*, Oxford University Press US, vol. 9, 2022, ofab433.
- [187] I. Skafle, A. Nordahl-Hansen, D. S. Quintana, R. Wynn, E. Gabarron, *et al.*, “Misinformation about covid-19 vaccines on social media: Rapid review,” *Journal of medical Internet research*, vol. 24, no. 8, e37367, 2022.
- [188] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, “Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa,” *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.
- [189] T. Nuzhath *et al.*, “Covid-19 vaccination hesitancy, misinformation and conspiracy theories on social media: A content analysis of twitter data,” 2020.

- [190] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [191] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, “Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach,” in *Proceedings of the Web Conference 2021*, 2021, pp. 194–205.
- [192] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, “Deep reinforcement learning for dialogue generation,” *arXiv preprint arXiv:1606.01541*, 2016.
- [193] J. Chen, T. Lan, and C. Joe-Wong, “Rgmcomm: Return gap minimization via discrete communications in multi-agent reinforcement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 17 327–17 336.
- [194] V. Agarwal, S. Joglekar, A. P. Young, and N. Sastry, “Graphnli: A graph-based natural language inference model for polarity prediction in online debates,” in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2729–2737.
- [195] M. Glenski, C. Pennycuff, and T. Weninger, “Consumers and curators: Browsing and voting patterns on reddit,” *IEEE Transactions on Computational Social Systems*, vol. 4, no. 4, pp. 196–206, 2017.
- [196] M. Glenski, S. Volkova, and S. Kumar, “User engagement with digital deception,” in *Disinformation, Misinformation, and Fake News in Social Media*, Springer, 2020, pp. 39–61.
- [197] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [198] X. Ma, M. Sap, H. Rashkin, and Y. Choi, “Powertransformer: Unsupervised controllable revision for biased language correction,” *arXiv preprint arXiv:2010.13816*, 2020.
- [199] Y. Zhang *et al.*, “Dialogpt: Large-scale generative pre-training for conversational response generation,” *arXiv preprint arXiv:1911.00536*, 2019.
- [200] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” *Advances in neural information processing systems*, vol. 12, 1999.
- [201] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

- [202] Y. Zhu *et al.*, “Texygen: A benchmarking platform for text generation models,” in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1097–1100.
- [203] M. Lewis *et al.*, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019.
- [204] N. Dai, J. Liang, X. Qiu, and X. Huang, “Style transformer: Unpaired text style transfer without disentangled latent representation,” *arXiv preprint arXiv:1905.05621*, 2019.
- [205] X. Xu, O. Dušek, I. Konstas, and V. Rieser, “Better conversations by modeling, filtering, and optimizing for coherence and diversity,” *arXiv preprint arXiv:1809.06873*, 2018.
- [206] K. M. d. Treen, H. T. Williams, and S. J. O’Neill, “Online misinformation about climate change,” *Wiley Interdisciplinary Reviews: Climate Change*, vol. 11, no. 5, e665, 2020.
- [207] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, “Rev2: Fraudulent user prediction in rating platforms,” in *WSDM*, 2018, pp. 333–341.
- [208] S. Kumar, J. Cheng, J. Leskovec, and V. Subrahmanian, “An army of me: Sockpuppets in online discussion communities,” in *WWW*, 2017, pp. 857–866.
- [209] Y. Dou, G. Ma, P. S. Yu, and S. Xie, “Robust spammer detection by nash reinforcement learning,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 924–933.
- [210] S. Rayana and L. Akoglu, “Collective opinion spam detection: Bridging review networks and metadata,” in *SIGKDD*, 2015, pp. 985–994.
- [211] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
- [212] R. Chalapathy and S. Chawla, “Deep learning for anomaly detection: A survey,” *arXiv preprint arXiv:1901.03407*, 2019.
- [213] S. Kumar, F. Spezzano, and V. Subrahmanian, “Vews: A wikipedia vandal early warning system,” in *SIGKDD*, ACM, 2015.

- [214] T. Dobrilova, \NoCaseChange{<https://review42.com/what-percentage-of-amazon-reviews-are-fake/>}, [Online; accessed 02-02-2021], 2020.
- [215] Y. Liu and M. Lapata, “Hierarchical transformers for multi-document summarization,” in *ACL*, 2019.
- [216] A. M. Lamb, A. G. A. P. Goyal, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio, “Professor forcing: A new algorithm for training recurrent networks,” in *NeurIPS*, 2016, pp. 4601–4609.
- [217] E. Jang, S. Gu, and B. Poole, “Categorical reparameterization with gumbel-softmax,” *ICLR*, 2017.
- [218] I. J. Goodfellow *et al.*, “Generative adversarial networks,” *arXiv preprint arXiv:1406.2661*, 2014.
- [219] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, “Equivalence of distance-based and rkhs-based statistics in hypothesis testing,” *The Annals of Statistics*, pp. 2263–2291, 2013.
- [220] G. Redeker, *Coherence and structure in text and discourse*. 2000.
- [221] J. Heaton, “Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618,” *Genetic Programming and Evolvable Machines*, vol. 19, no. 1-2, 2017.
- [222] B. He, M. Ahamad, and S. Kumar, “Robad: Robust adversary-aware local-global attended bad actor detection sequential mode,” *In submission*, 2024.
- [223] S. Chhabra, P. Majumdar, M. Vatsa, and R. Singh, “Data fine-tuning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8223–8230.
- [224] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.
- [225] H. Touvron *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [226] T. Wu *et al.*, “A brief overview of chatgpt: The history, status quo and potential future development,” *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.

- [227] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
- [228] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.
- [229] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [230] A. Y. K. Chua, C.-Y. Tee, A. Pang, and E.-P. Lim, “The retransmission of rumor and rumor correction messages on twitter,” *American Behavioral Scientist*, vol. 61, pp. 707–723, 7 Jun. 2017.
- [231] A. Y. K. Chua and S. Banerjee, “A study of tweet veracity to separate rumours from counter-rumours,” ACM Press, 2017, pp. 1–8, ISBN: 9781450348478.
- [232] P. Ayranci, C. Bandera, N. Phan, R. Jin, D. Li, and D. Kenne, “Distinguishing the effect of time spent at home during covid-19 pandemic on the mental health of urban and suburban college students using cell phone geolocation,” *International journal of environmental research and public health*, vol. 19, no. 12, p. 7513, 2022.

VITA

Bing He is a Ph.D. Candidate in Computer Science at Georgia Institute of Technology, co-advised by Prof. Mustaque Ahamad and Prof. Srijan Kumar. He received B.S. in Electronic Engineering at the University of Electronic Science and Technology of China and M.S. in Computer Science at the University of Macau. He has published more than 10 articles in major data science and machine learning venues. His research interests include machine learning, natural language processing, and their application to combat misinformation on social media platforms.