BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

A MULTI-MODAL, MULTI-PLATFORM, AND MULTI-LINGUAL APPROACH TO UNDERSTANDING ONLINE MISINFORMATION

by

YUPING WANG

B.E., Huazhong University of Science and Technology, 2013
M.S., Zhejiang University, 2016

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2023

Approved by

First Reader

Gianluca Stringhini, PhD
Assistant Professor of Electrical and Computer Engineering

Second Reader

Manuel Egele, PhD
Associate Professor of Electrical and Computer Engineering

Third Reader

Lei Guo, PhD
Professor of Journalism
Fudan University

Fourth Reader

Ioannis Ch. Paschalidis, PhD
Distinguished Professor of Engineering
Professor of Systems Engineering
Professor of Electrical and Computer Engineering
Professor of Biomedical Engineering
Professor of Computing & Data Sciences

Acknowledgments

First, I would like to thank my advisor Prof. Gianluca Stringhini. He is patient, supportive, and warmhearted. Whenever I make some progress, he always praises me and I do feel encouraged. If I had any setbacks, he would always cheer me up. Also, he introduces me to his friends and collaborators, and so I can build connections with these outstanding scholars. He also spends tremendous time in instructing me on refining my presentations, as well as revising my writings. In addition to that, he also cares about my career after graduation and gives me suggestions on how to choose a job wisely. I feel fortunate to have him as my advisor, and I cannot thank him enough.

I would like to thank my dissertation committee members Prof. Manuel Egele, Prof. Lei Guo, and Prof. Ioannis Ch. Paschalidis. Thank you for your time in reading my dissertation and your prudent comments that help improve the quality of my dissertation greatly.

I also thank Prof. Jeremy Blackburn, Prof. Emiliano De Cristofaro, and Prof. Barry Bradley for their feedback on my papers.

I would also like to thank my labmates and friends Assel Aliyeva, Dr. Sadullah Canakci, Dr. Onur Zungur, Shiza Ali, Ioannis Angelakopoulos, William Blair, Muzammil Hussain, Beliz Kaleli, Chen Ling, Rasoul Jahanshahi, Alexander Bulekov, Pujan Paudel, Hammas Saeed, Dr. Yiyan Zhang, Dr. Savvas Zannettou, and Dr. Qianqian Ma.

Finally, I would like to thank my parents Danjie Wang and Liming Wang for their enduring love and firm support for me, which help me overcome all the difficulties during my PhD journey.

Yuping Wang

A MULTI-MODAL, MULTI-PLATFORM, AND MULTI-LINGUAL APPROACH TO UNDERSTANDING ONLINE MISINFORMATION

YUPING WANG

Boston University, College of Engineering, 2023

Major Professor: Gianluca Stringhini, PhD
Assistant Professor of Electrical and Computer Engineering

ABSTRACT

Due to online social media, access to information is becoming easier and easier. Meanwhile, the truthfulness of online information is often not guaranteed. Incorrect information, often called misinformation, can have several modalities, and it can spread to multiple social media platforms in different languages, which can be destructive to society. However, academia and industry do not have automated ways to assess the impact of misinformation on social media, preventing the adoption of productive strategies to curb the prevalence of misinformation. In this dissertation, I present my research to build computational pipelines that help measuring and detecting misinformation on social media. My work can be divided into three parts.

The first part focuses on processing misinformation in text form. I first show how to group political news articles from both trustworthy and untrustworthy news outlets into stories. Then I present a measurement analysis on the spread of stories to characterize how mainstream and fringe Web communities influence each other.

The second part is related to analyzing image-based misinformation. It can

be further divided into two parts: fauxtography and generic image misinformation. Fauxtography is a special type of image misinformation, where images are manipulated or used out-of-context. In this research, I present how to identify fauxtography on social media by using a fact-checking website (Snopes.com), and I also develop a computational pipeline to facilitate the measurement of these images at scale. I next focus on generic misinformation images related to COVID-19. During the pandemic, text misinformation has been studied in many aspects. However, very little research has covered image misinformation during the COVID-19 pandemic. In this research, I develop a technique to cluster visually similar images together, facilitating manual annotation, to make subsequent analysis possible.

The last part is about the detection of misinformation in text form following a multi-language perspective. This research aims to detect textual COVID-19 related misinformation and what stances Twitter users have towards such misinformation in both English and Chinese. To achieve this goal, I experiment on several natural language processing (NLP) models to investigate their performance on misinformation detection and stance detection in both monolingual and multi-lingual manners. The results show that two models: COVID-Tweet-BERT v2 and BERTweet are generally effective in detecting misinformation and stance in the two above manners. These two models are promising to be applied to misinformation moderation on social media platforms, which heavily depends on identifying misinformation and stance of the author towards this piece of misinformation.

Overall, the results of this dissertation shed light on understanding of online misinformation, and my proposed computational tools are applicable to moderation of social media, potentially benefitting for a more wholesome online ecosystem.

Contents

List of Tables

List of Figures

List of Abbreviations

AAAI	.....	Association for the Advancement of Artificial Intelligence
ACL	.....	Association for Computational Linguistics
ACM	.....	Association for Computing Machinery
AI	.....	Artificial Intelligence
ARR	.....	ACL Rolling Review
COVID-19	.....	Coronavirus disease 2019
CNN	.....	convolutional neutral networks
CSCW	.....	Computer-Supported Cooperative Work And Social Computing
GDELT	.....	Global Database of Events, Language, and Tone
ICWSM	.....	International AAAI Conference on Web and Social Media
IEEE	.....	Institute of Electrical and Electronics Engineers
IRB	.....	Institutional Review Board
NLP	.....	Natural Language Processing
PACM HCI	.....	Proceedings of the ACM on Human Computer Interaction
TABARI	.....	Textual Analysis by Augmented Replacement Instructions
URL	.....	Uniform Resource Locator

Chapter 1

Introduction

As a useful part of our daily life, social media delivers important information to every digital citizen. At the same time, this convenience inevitably allows information that is not correct, which is usually called misinformation (Wu et al., 2019), to spread. In some cases, misinformation can have severe consequences.

One example is pertinent to trolls on social media. During the 2016 U.S. presidential elections, Russian trolls exploited Twitter to influence the choices of voters (Luceri et al., 2020). Other instances include the false notion that bleach can treat the coronavirus, leading to life-threatening results for some people in the U.S. (Reimann, 2020), and the rumor that 5G technology can spread the coronavirus, causing chaos in several countries in Europe (Nic Fildes, Mark Di Stefano, and Hannah Murphy, 2020).

Despite misinformation being a significant threat to our society, we still lack approaches to detect it and comprehensively evaluate its impact on multiple social media platforms, and this limits our ability to develop policies to curb its spread. The reasons are at least twofold. On the one hand, the overwhelming amount of data makes traditional manual analysis impractical. For instance, (Guo et al., 2016) argued that when a research analyzing a dataset containing more than 100,000 social media posts, e.g. texts, images, it could be treated as a big data analyzing project, since the volume is already intractable for manual efforts. Nevertheless, every day social media data wildly exceed this limit. For example, in Chapter 4, I am dealing with 2.2B posts, obviously impossible to process manually.

On the other hand, misinformation can come in different modalities (text, images, and so on) and different languages (English, Chines, and so on). Therefore, measuring and detecting misinformation require various techniques spanning multiple modalities and languages. Although artificial intelligence (AI) technology, e.g. deep learning (LeCun et al., 2015), image processing (Krizhevsky et al., 2012), and text processing (Devlin et al., 2018) techniques, have emerged as a potential aid to misinformation research, AI techniques are unable to be directly applied to misinformation research because misinformation research requires domain knowledge.

Thus, to solve the above problems, I make the following **thesis statement: Using a multi-modal, multi-platform, and multi-language approach can improve detecting misinformation and measuring the spread of misinformation at scale.** To fulfill this thesis statement, my research develops techniques to detect misinformation and builds computational pipelines that adapt AI techniques to specific misinformation research tasks with multiple modalities and languages, with the goals of highlighting misinformation and measuring the impact of misinformation on social media. In particular, my research focuses on two important media forms: text and images.

- For text, I have built a computational pipeline to identify distinct stories among numerous news articles posted on social media by using the GDELT dataset. Employing the tool, I characterize the spread of stories on social media and how different social media influence each other (Chapter 3). The results are published in the *Proceedings of 2021 IEEE International Conference on Big Data* (*Big Data*). For detection of multi-lingual misinformation, I experiment on several natural language processing (NLP) models to investigate their performance on misinformation detection and stance detection on our curated dataset in both monolingual and multi-lingual manners, and the results show that certain

models are generally effective in detecting misinformation and stance in the two above manners (Chapter 6). The results are presented in a paper under revision according to the reviews from the *ACL Rolling Review* (*ARR*) October 2022 round.

- For images, I have established computational pipelines to process a specific type of image misinformation called fauxtography (Chapter 4) and generic image misinformation, respectively. The latter one is used to characterize the image misinformation on Twitter in the COVID-19 pandemic (Chapter 5). The results are published in the *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media* (*ICWSM* 2021) and *The Proceedings of the ACM on Human Computer Interaction* (*PACM HCI*), CSCW1, April 2023 Issue, respectively.

This dissertation is split into four main chapters, which are described below.

News Stories. In Chapter 3, I present my research on the measurement study of sharing of political news stories on Web communities, including both mainstream communities like Twitter and Reddit and fringe communities dominated by extremists like 4chan, Gab, and the /r/The_Donald subreddit (Zannettou et al., 2017; Hine et al., 2017; Zannettou et al., 2018a). A news story is commonly defined in communication theory, where every news "story" is composed of several story "segments" (Marcelino et al., 2019). For example, a news story would be the news coverage of a foreign trip by the US President, with the various segments being the single events related to this trip, e.g., what people he met or what activities he took part in.

Previous work studying the news ecosystem on the Web has mostly focused on a single platform (usually Twitter) and has looked at references to single news articles at a time (Ratkiewicz et al., 2011; Rye et al., 2020; Shao et al., 2018; Zhao et al., 2011). Moreover, efforts to study the intertwined relationship between news coverage

and social media discussions have been limited to direct quotes from news articles posted on Twitter (Leskovec et al., 2009), or on how Web communities influence each other in spreading *single* news URLs (Zannettou et al., 2017). In fact, news articles from different news outlets can report the same story, and when users read a news article from one news domain, they may choose to view the same story from another news domain (Garimella et al., 2021), and disseminate the news story by using a news article from the latter domain. Thus, while direct quotes cannot model such type of users' sharing behavior, looking at the sharing of news stories instead provides a more comprehensive view of the news sharing phenomenon. Due to the lack of publicly available measurement infrastructures able to track how news stories are discussed online, research that aimed at answering more complex research questions on how certain news stories are discussed online had to limit the number of social media posts being analyzed and to rely on qualitative analysis (Conover et al., 2011b; Starbird, 2017; Wilson et al., 2018). The lack of measurement techniques able to satisfy these requirements limited the scope of past research studying online news discussion.

In this chapter, I fill the aforementioned gap by developing a measurement pipeline that groups together related political articles into stories and traces their discussion on multiple online communities, which is used to measure the influence of a Web community on other ones. One of the results shows that small/fringe communities like /r/The_Donald have a much larger *external* influence, affecting the posting of stories at a rate much larger than their relative size would suggest, which highlights the significant negative impact of fringe communities on mainstream communities.

Fauxtography. In Chapter 4, I focus on *fauxtography* (Cooper, 2007), i.e., news images that have been modified or miscaptioned to change their intent, often with the goal of spreading a false sense of the events they purport to depict. Although previous research efforts have proposed detection tools for fauxtography (Zhang et al., 2018;

Zlatkova et al., 2019), to the best of my knowledge, the *impact* of on news discussion has not been studied. In particular, I investigate two research questions:

- RQ1: Does sharing fauxtography increase engagement on social media?
- RQ2: Do fauxtography images have a life beyond their questionable verisimilitude (their appearance of being being real)? I.e., do new variants and memes using them appear on social media?

To answer these questions, I develop a computational analysis pipeline geared to identify posts containing fauxtography at scale, measure the engagement of users sharing and favoring such posts, and understand how these images are used on three different social media platforms (Twitter, Reddit, and 4chan). In addition to the quantitative analysis, I also conduct a case study by focusing on three selected case studies of fauxtography which spawned new variants. I conclude this chapter by discussing implications for dealing with fauxtography in the wild, considering the current environment of social media moderation.

COVID-19 misinformation images. In Chapter 5, I aim to shed light on how images are used to spread COVID-19 misinformation on Twitter and focus on the following research questions:

- RQ1: Do COVID-19 misinformation images generate more user engagement?
- RQ2: What are the temporal properties of COVID-19 misinformation images? Do COVID-19 misinformation images have a longer lifespan and longer burst times than non-misinformation images?
- RQ3: What are the characteristics of users who post COVID-19 misinformation images?

To this end, I collect 2.3M COVID-19 related tweets posted between March 1, 2020 to June 16, 2020, and then download 340K images included in those tweets. To facilitate manual analysis of these messages, I build a computational pipeline based on perceptual hashing techniques and clustering algorithms to group visually similar images together. I then develop a codebook to characterize COVID-19 misinformation images, identify five different types of COVID-19 misinformation images, and build a dataset of over 2.8K COVID-19 misinformation images posted on Twitter. I then perform a quantitative analysis on the tweets that contain COVID-19 misinformation images to answer the research questions above.

The results shed light on how images are used to spread COVID-19 misinformation on Twitter. Most interestingly RQ1 contradicts what was found by previous research on misinformation, which found that tweets containing false information receive more engagement (Vosoughi et al., 2018; Wang et al., 2021). A potential reason is that past research followed a top-down approach, only looking for false stories that had been fact-checked, while this approach is bottom-up, identifying groups of misinformation images as they are posted online. I argue that more discussion is needed within the misinformation research community to better understand the advantages and disadvantages of different data approaches, and the biases that these choices might introduce in research results.

COVID-19 multi-lingual misinformation on Twitter. In Chapter 6, I focus on COVID-19 multi-lingual misinformation on Twitter. Misconceptions and misinformation permeate public discourse online and offline, regardless of the language that people might speak. In particular, if we accumulate enough knowledge for one misconception in one language, we may be able to transfer it to moderating the same misconception in another language, which is helpful for multi-lingual social media platforms like Twitter.

As a result, we first compile an English and Chinese bilingual annotated Twitter dataset relevant to COVID-19 misconceptions[1] with a total number of 6,000 tweets. The annotation is composed of identifying whether one specific misconception is contained in one tweet and what stance the author of this tweet has toward this misconception. Based on this dataset, then I experiment on several natural language processing (NLP) models to investigate their performance on misinformation detection and stance detection in both monolingual and multi-lingual manners. The results show that two models: COVID-Tweet-BERT v2 and BERTweet are generally effective in detecting misinformation and stance in the two above manners. These two models are promising to be applied to misinformation moderation on social media platforms, which heavily depends on identifying misinformation and stance of the author towards this piece of misinformation.

Takeaways. The contributions of my research are fourfold.

- The proposed computational pipeline facilitates the application of AI techniques to more specific practical tasks.
- The created tools help solve the big data difficulties faced by social science scholars in areas like communication research.
- My research can be used to better pinpoint misinformation on social media.
- The results of my measurement study can be useful to social media platform moderators and policy makers as evidence for policy-making, and the proposed pipelines can also be used to evaluate the effectiveness of existing counter measures, potentially benefitting for a more wholesome online ecosystem.

Ethics. In all the research presented in this dissertation, I only use publicly available data, and I do not interact with users. Therefore, the research presented in Chapter 3,

[1]The misconceptions are related to ginger/garlic, hydroxychloroquine, and bioweapon

Chapter 4, and Chapter 5 is not considered human subjects research by the Institutional Review Board (IRB) of Boston University. The research presented in Chapter 6 involves data annotation approach dealing with crowdsourced annotators to label my dataset. The research protocol was reviewed by the IRB of Boston University, and I obtained an IRB exemption for this research at Boston University. Also, I follow standard ethical guidelines (Rivers and Lewis, 2014), encrypt data at rest, and make no attempt to de-anonymize users.

Since there are human faces in my dataset in Chapter 5, which create privacy concerns, in this research I blur the face of people portrayed in images unless a) they are public figures, b) the image is a drawing, or c) the image comes from stock photography.

Chapter 2

Related Work

In this chapter, I present the related work to help readers better understand the context of my work. This chapter is split into three parts: definition of misinformation, text analysis, and image analysis.

Readers may want to refer to (Kumar and Shah, 2018; Zhou and Zafarani, 2020; Alam et al., 2021; Fung et al., 2022) for surveys in the area of misinformation research.

2.1 Definition of misinformation

Research on false information typically distinguishes between information that is spread with malicious intent (i.e., *disinformation*) and incorrect claims that are genuinely believed by whoever is posting them (i.e., *misinformation*) (Wu et al., 2019; Lazer et al., 2018; Wilson et al., 2018; Wilson and Starbird, 2020). While this distinction is important to understand the goal of people posting false information online and to design appropriate defenses, it is very challenging to infer the intent with which a piece of false information is posted online. For this reason, in this dissertation, I adopt the definition of misinformation proposed by Wu et al., which defines misinformation as "informative content that contains incorrect or inaccurate information" (Wu et al., 2019), regardless of the purpose with which it was posted. There are some other similar terms. For instance, a rumor is to a piece of information that is unverified at the time of propagation and it is usually widely circulated (Guo and Zhang, 2020). In particular, a rumor can be confirmed to be true eventually. Another instance is fake

news referring to news reports that contain intentionally false information (Shu et al., 2019). The research of these two example terms are out of scope of this dissertation, and interested readers can refer to surveys in this area like (Kumar and Shah, 2018).

2.2 Text analysis

Though misinformation study has multiple branches in various information forms like images and videos (Reis et al., 2020; Ling et al., 2022), text misinformation has always been a critical research area (Wu et al., 2019; Kumar and Shah, 2018). In the following subsections, I review four research directions germane to my research: textual misinformation spread on social networks, misinformation detection, stance detection, and news trustworthiness at the level of news outlets.

2.2.1 Textual misinformation spread on social networks

A large body of research has studied propagation of textual misinformation on social media, which is related to my work presented in Chapter 3. (Vosoughi et al., 2018) showed that fake news spread faster than true news on Twitter. By investigating the discussions on mass shooting events on Twitter, (Starbird, 2017) revealed that alternative news outlets actively propagate alternative narratives, while (Wilson et al., 2018) studied information operations through the lens of the "Aleppo Boy" narrative, and showed that some news media collaborate to spread alternative narratives. Also, (Zannettou et al., 2019b) analyzed disinformation campaigns carried out by state-sponsored actors, characterizing their influence on social networks, while (Jiang et al., 2020) analyzed user comments to characterize the public's (dis)belief towards news items. (Flintham et al., 2018) surveyed users consuming news on social network and found that both sources and content play key roles in how they evaluate news veracity. Other studies in this area showed that the spread of misinformation is amplified by the involvement of bots, trolls, and political astroturfing (Ratkiewicz

et al., 2010; Ratkiewicz et al., 2011; Stringhini et al., 2015; Shao et al., 2018; Wang and Paschalidis, 2016; Saeed et al., 2022).

2.2.2 Misinformation detection

Various misinformation research projects focus on text misinformation posted on social media, which is pertinent to the content in Chapter 6. One research direction is to develop automated approaches to detect false information, which typically are based on machine learning and natural language processing techniques (Wu and Liu, 2018; Shu et al., 2017; Castillo et al., 2011; Wang, 2017; Paudel et al., 2022). Such techniques typically combine natural language processing (NLP) techniques, machine learning techniques, and even structure features together to detect misinformation on social media (Shu et al., 2019; Lin et al., 2022).

To gauge the effectiveness of these misinformation detection approaches, multiple misinformation datasets are built, ranging from fake news articles (Nørregaard et al., 2019; Kazemi et al., 2021) to rumors on social media platforms like Twitter (Ma et al., 2017) and its Chinese counterpart Sina Weibo (Chen et al., 2020b; Leng et al., 2021; Hu et al., 2020). In particular, numerous COVID-19 related misinformation datasets have emerged after the outbreak of COVID-19 pandemic (Memon and Carley, 2020; Hossain et al., 2020; Lin et al., 2022; Phillips et al., 2022; Alam et al., 2020; Nakov et al., 2021; Kar et al., 2020).

The common characteristic of these datasets is that they mix various misconceptions in one dataset, with a few instances for each misconception. In general, these datasets are designed to test if misinformation detection approaches can successfully identify hidden misinformation, which can be used to alarm as early as possible after misinformation begins to propagate (Lin et al., 2022). Although the dataset compiled in Chapter 6 overlaps with these datasets in terms of misinformation identification, the goal of Chapter 6 differs from the above mentioned datasets. In fact, the misinfor-

mation detection dataset presented in Chapter 6 is composed of three independent parts, each of which is related to a specific claim. Misinformation detection systems built on one such claim can be used to judge whether one unseen tweet contains that specific false claim, which can be applied to flag misinformation posts for popular and important topics on social media (Zannettou, 2021).

2.2.3 Stance detection

To better understand users on social media, a research direction aims to investigate the stance expressed in social media posts. This research direction is related to my research in Chapter 6.

Stance detection does not necessarily centered on misinformation (Ullah et al., 2021; Hardalov et al., 2021; Küçük and Can, 2020; Hardalov et al., 2022; Mohammad et al., 2016; Lai et al., 2020). However, it has a pivotal role in misinformation research as tweets promoting misconceptions may deserve some moderation (Zannettou, 2021).

In this research area, classical text classification models, including convolutional neutral networks (CNN) (Kim, 2014), FastText (Joulin et al., 2017), BiLSTM (Schuster and Paliwal, 1997), and BERT (Devlin et al., 2018) have been used in inferring stances.

Similar to misinformation detection research, several stance detection datasets have been created to test the effectiveness of stance detection approaches. Most of these datasets are monolingual. To name a few, one of the most widely used stance detection dataset is SemEval Task 6 (Mohammad et al., 2016) which annotates 4,870 English tweets associated with 6 common targets in U.S. . After the outbreak of COVID-19 pandemic, several stance detection datasets have been proposed to understand how people react to COVID-19 administration policies (Glandt et al., 2021), COVID-19 misconceptions (Hossain et al., 2020; Weinzierl et al., 2021; Hou et al., 2022), and vaccines (Poddar et al., 2022) based on English tweets.

Only a few stance detection datasets are multi-lingual. On the basis of two stance

detection datasets (Mohammad et al., 2016; Taulé et al., 2017), Lai et al., created a dataset with stances towards election issues and referendums in English, Spanish, Catalan, French, and Italian (Lai et al., 2020). By collecting comments towards more than 100 political issues from election candidates in Switzerland, researchers built a stance detection dataset composed of English, German, French and Italian text (Vamvas and Sennrich, 2020). All of the above multi-lingual datasets fall in political areas, while recently researchers created a German, French, and English dataset with respect to public opinions towards COVID-19 vaccines (Chen et al., 2022b).

These multi-lingual stance detection datasets are all composed of text from Indo-European languages. It is still unknown how cross-lingual stance detection works in two languages that belong to two different language families. To the best of my knowledge, our compiled dataset in Chapter 6 is the first to fill this gap by providing a stance detection dataset with languages from two language families with multiple common targets.

Three papers (Micallef et al., 2020; Kim et al., 2022; Mutlu et al., 2020) cover stances towards using ginger/garlic or hydroxychloroquine to treat COVID-19 and these datasets only have text in English. The annotation instructions are consistent between Chinese and English, and those external datasets cannot be used as part of my research in chapter 6 since their annotation instructions are different from this research, which may impair the cross-lingual stance detection results (Ng and Carley, 2022; Bozarth and Budak, 2020).

2.2.4 News trustworthiness at the level of news outlets

Since fact-checking all news articles that appear on social media is a daunting task, and it is unfeasible for fact-checking organizations to cover them all, another line of research considers the trustworthiness of entire news outlets, instead of focusing on

single articles (Lazer et al., 2018; Budak, 2019). This research direction is relevant to my research presented in Chapter 3. For example, researchers investigated the spread of articles from untrustworthy news outlets during the 2016 US presidential election (Grinberg et al., 2019; Budak, 2019), and both concluded that although untrustworthy news outlets had a large influence on social media users, news articles written by trustworthy news outlets were still more widely shared than those from untrustworthy news outlets (Grinberg et al., 2019; Budak, 2019). (Pennycook and Rand, 2019) assessed the trustworthiness of a news outlet based on laymen's evaluations, showing that crowdsourced judgements are successful in assessing trustworthy news sources, although not as much as professional fact-checkers. By inspecting narratives around two distinct political themes, (Starbird, 2017; Wilson et al., 2018) showed that untrustworthy news outlets often coordinated with each other when shaping discourses around specific topics.

(Gentzkow and Shapiro, 2006; Gentzkow et al., 2015) showed that news outlets can report news in a biased way and mislead the readers; Soraka et al. (Soroka et al., 2019) found that people tend to be more "attracted" by negative news stories. By analyzing Web browsing histories of online users, (Garimella et al., 2021) showed that online news consumers prefer news outlets aligned with their own political leanings. To reduce bias, (Babaei et al., 2018) proposed a method to identify "purple news," i.e., news that can be unanimously accepted by readers who have opposite political leanings.

My work differs from this line of work not only in terms of methodology but also because, besides Twitter, I also study fringe, impactful communities like Gab and 4chan. Moreover, I analyze the influence of fringe communities on mainstream ones, which may help to better understand the influence dynamics of false news sharing.

2.3 Image analysis

More recently, the research community has begun to look at the interplay between images and misinformation. This research direction is related to my research presented in Chapter 4 and Chapter 5.

Prior work has also studied fauxtography, aiming to detect false images. (Zhang et al., 2018) built a fauxtography detector called "FauxBuster" based on machine learning techniques, while (Bayar and Stamm, 2016) used deep learning to detect manipulated images. Similarly, (Zlatkova et al., 2019) extracted various features from images and text, and use machine learning to assess the authenticity of specific claims. Furthermore, they describe which features are the most effective in verifying the authenticity of the claims. However, my research presented in Chapter 4 is the first work, to my best knowledge, that investigates what impact fauxtography has on various social media platforms at scale.

Other research used computational approaches to study the spread of image-based misinformation on social media. Previous work showed that images are commonly used on social media to spread misinformation (Du et al., 2020; Qu et al., 2022), hateful content (Zannettou et al., 2018b; Kiela et al., 2021), and negative emotions (Ali et al., 2022). (Dewan et al., 2017) presented a pipeline to extract themes and sentiments conveyed in images, and highlight several instances where images were used to share disinformation. Additionally, misinformation images are commonly used in political settings. Previous research found that these images were prevalent in public WhatsApp groups during election campaigns in India and Brazil (Garimella and Eckles, 2020; Resende et al., 2019; Reis et al., 2020), and that state-sponsored influence campaigns made wide use of images too (Zannettou et al., 2020a; Ng et al., 2022).

(Javed et al., 2022; Javed et al., 2020) studied COVID-19 textual and image misinformation shared in public WhatsApp groups and on Twitter in Pakistan, finding

that the spread of this content appears to be organic and not carried out by bots on Twitter. (Lee et al., 2021) looked at manipulated and misleading data visualizations used to push false narratives surrounding COVID-19. Compared to these previous work, the study in Chapter 5 is more general as it examines any type of image used to carry out COVID-19 misinformation, and it does not focus on a single country but look at the entirety of Twitter.

(Zannettou et al., 2018b) used visual similarity (i.e., perceptual hashing) and images annotated by the website KnowYourMeme to identify and study image memes posted on several mainstream and fringe social media platforms. As follow up work, (Ling et al., 2021) performed a mixed-method analysis of popular memes, looking for which indicators contribute to their virality.

In the spirit of this past research, in Chapter 5, I study how images containing misinformation on COVID-19 are shared on Twitter. Unlike previous work, which relied on a top-down approach of looking for images that have been fact-checked or labeled by external organizations, I follow a bottom-up approach, grouping together images that look similar, developing a codebook to characterize image-based misinformation, annotating, and analyzing them.

Chapter 3

News stories

3.1 Introduction

The Web has facilitated the growth of fast-paced, online-first news sources. It has also allowed users to actively contribute to and shape the discussion around the news. This creates an environment where journalists are not necessarily the arbiters of how a news story develops and spreads. In today's "hybrid" media system (Chadwick, 2011), the popularity of a news story is also influenced by how users discuss it. Although such discussions usually happen organically, various actors from polarized online communities or state-sponsored troll farms might also attempt to manipulate them, e.g., by pushing (Ferrara, 2017) or weaponizing (Zannettou et al., 2017) certain narratives.

To get a comprehensive understanding of how news stories are discussed online and of the communities that are influential in spreading news stories the research community needs measurement techniques able to not only identify news articles that are related to the same story, but also keep track of how these articles are discussed on multiple online communities. The lack of measurement techniques able to satisfy these requirements limited the scope of past research studying online news discussion. In fact, previous work studying the news ecosystem on the Web has mostly focused on a single platorm (usually Twitter) and has looked at references to single news articles at a time (Ratkiewicz et al., 2011; Rye et al., 2020; Shao et al., 2018; Zhao et al., 2011). Moreover, efforts to study the intertwined relationship between news coverage

and social media discussions have been limited to direct quotes from news articles posted on Twitter (Leskovec et al., 2009), or on how Web communities influence each other in spreading *single* news URLs (Zannettou et al., 2017). In fact, news articles from different news outlets can report the same story, and when users read a news article from one news domain, they may choose to view the same story from another news domain (Garimella et al., 2021), and disseminate the news story by using a news articles from the latter domain. Thus, while direct quotes cannot model such type of users' sharing behavior, looking at the sharing of news stories instead provides a more comprehensive view of the news sharing phenomenon. Due to the lack of publicly available measurement infrastructures able to track how news stories are discussed online, research that aimed at answering more complex research questions on how certain news stories are discussed online had to limit the number of social media posts being analyzed and to rely on qualitative analysis (Conover et al., 2011b; Starbird, 2017; Wilson et al., 2018).

In this chapter, I fill the aforementioned gap by developing a measurement pipeline, which is shown in Figure 3·1, that groups together related articles into stories and traces their discussion on multiple online communities. This research follows a common definition of news stories used in communication theory, where every news "story" is composed of several story "segments" (Marcelino et al., 2019). For example, a news story would be the news coverage of a foreign trip by the US President, with the various segments being the single events related to this trip, e.g., what people he met or what activities he took part in. This measurement pipeline consists of different components to 1) collect data, 2) extract named entities, 3) group articles belonging to the same story, and 4) estimate the influence of a Web community on other ones.

I then instantiate the pipeline by focusing on a mix of mainstream and fringe communities – namely, Twitter, Reddit, 4chan's Politically Incorrect board (/pol/),

and Gab – and extract 38M posts including 15.6M unique news URLs, spanning a period of almost three years. I use named entity extraction to analyze what types of news these communities discuss. I also study the interplay between news discussion and the trustworthiness of the news sources cited using NewsGuard (NewsGuard, 2019d), a trustworthiness assessment site compiled by professional fact-checkers.

Next, I perform community detection to group together related articles into news stories and study how they are discussed on different Web communities. To this end, I use GDELT, a dataset that labels global news events (Leetaru and Schrodt, 2013). Because of GDELT's focus on politics, this measurement also concentrates on political news stories. Finally, to study the influence that different Web communities have on each other in spreading news stories, I use Hawkes Processes (Hawkes, 1971). These allow me to estimate which news stories are organically discussed on the Web and for which ones certain communities exercise a significant influence in spreading them.

In summary, this chapter makes the following contributions:

- I develop a measurement pipeline that overcomes the limitations of previous work by grouping news articles into stories and tracing their discussion online, which can help researchers perform more comprehensive analyses of online news discussion.

- I find that when discussing news, different Web communities post URLs to news outlets with varying levels of trustworthiness. In particular, Gab and /r/The_Donald (subreddit) prefer untrustworthy ones compared to Reddit, Twitter, and 4chan.

- I find that some communities are particularly influential in the dissemination of news stories. While large ones like Twitter and Reddit have consistent influence, small/fringe communities like /r/The_Donald have a much larger

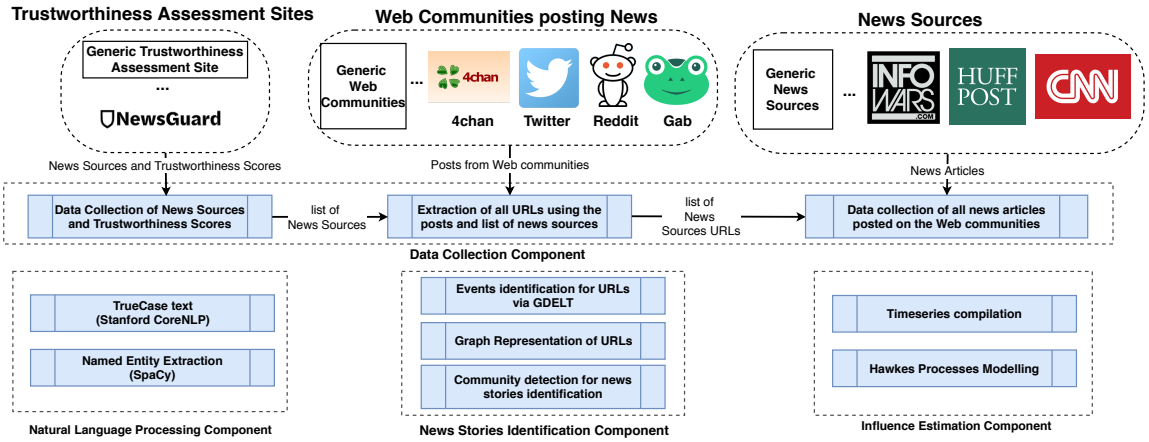


Figure 3·1: High-level overview of the proposed processing pipeline.

external influence, affecting the posting of stories at a rate much larger than their relative size would suggest.

3.2 Datasets

In this section, I describe how I build my dataset, which is the first component of my pipeline, which is used to select a set of suitable news sources, determine their trustworthiness, and retrieve posts on different social networks, including URLs to these sources. While this methodology is general and can be used with any data source, in the following, I describe it along with the specific services selected for this work.

News Sources. Previous work has mostly relied on pre-determined lists of websites (Zannettou et al., 2017; Grinberg et al., 2019; Budak, 2019; Allcott et al., 2019). However, this incurs important limitations, as the popularity of news sites varies over time (Scheitle et al., 2018), and low-reputation sites are often ephemeral. For instance, out of the 54 "alternative" news sites studied in (Zannettou et al., 2017), only 24 are still active as of November 2021. Thus, I opt to take a more systematic approach. First, I gather popular domains using the top 30K websites from the

Majestic list (Majestic, 2019; Scheitle et al., 2018) as of February 2019. Then, I use the VirusTotal API (VirusTotal, 2020), a service that provides domain categorization, to select only the domains categorized as *news and media* or *news*. Note that VirusTotal's categorization is not exempt from mistakes; e.g., domains like `ananova.com`, `adbusters.org`, and `cagle.com` are misclassified as news sites. To further refine this list, I use the NewsGuard API (NewsGuard, 2019d), a service that ranks news sources based on their trustworthiness, and restrict this analysis to news sites that are rated by NewsGuard as of February 2019, i.e., before it became a paid service. As a result, I obtain a total of 1,073 news websites.

Source Trustworthiness. I also use NewsGuard (NewsGuard, 2019d) to characterize the trustworthiness of a news Website, as it provides credibility/transparency scores. The scores are based on nine journalistic criteria, focused on different aspects (e.g., whether a news site consistently publishes false content, uses deceptive headlines, etc.) that do not take into account political leanings and range from 0 to 100. If the score is greater than or equal to 60, the news source is labeled as trustworthy and untrustworthy otherwise (NewsGuard, 2019b). For example, I found trustworthy news outlets, including `wsj.com`, `npr.com`, and `reuters.com`, and untrustworthy news outlets including `zerohedge.com`, `sputniknews.com`, and `infowars.com`. NewsGuard's evaluation is conducted by a team of experts (NewsGuard, 2019b), and this manual vetting provides us with reasonable confidence in its accuracy.

Note the threshold of 60 is pre-defined by NewsGuard, and its assessment evaluation is designed to fix the threshold first and then assign the points for each criterion according to this threshold (NewsGuard, 2019b). Any change of this threshold would also need to be accompanied by a reevaluation of the points for each criterion, which is out of the scope of this research. Therefore, I stick to this threshold.

Also, NewsGuard has been working with researchers (Resnick et al., 2018; Zhou

et al., 2020; Nørregaard et al., 2019), libraries, Web browsers, news outlets (NewsGuard, 2019a), and service providers to increase the transparency of news credibility assessments (NewsGuard, 2019a). News outlets are evaluated transparently with details shown in the corresponding "nutrition label" page where readers can find the reasons for judgment (NewsGuard, 2019c).

Web Communities. I retrieve social media posts that include URLs from the 1,073 news sources in this dataset. This selection is based on highlighting the interplay and the influence between different online communities instead of political leanings. While this pipeline can include any platforms, in this research, I focus on a few Web communities: Twitter, Reddit, 4chan, and Gab. That is, I study both mainstream communities like Twitter as well as "fringe" ones like 4chan. In particular, I turn to 4chan's Politically Incorrect board (/pol/) as prior work shows it is an influential actor with respect to the dissemination of news (Zannettou et al., 2017) and memes (Zannettou et al., 2018b). I also include Gab because it is an emerging community marketed as a "free speech haven," which is heavily used by alt- and far-right users (Zannettou et al., 2018a). As for Reddit, I also choose to study /r/The_Donald as a separate community since previous research has highlighted its influence in spreading information on the Web (Flores-Saviaga et al., 2018; Zannettou et al., 2017).[1]

Table 4.1 provides a summary of the number of posts and unique news URLs for each community. Next, I describe the data that I collect for each Web community in detail.

Twitter. I collect tweets made available through the 1% Streaming API between Jan 1, 2016 and Oct 31, 2018. That is once the tweets are generated, they get sampled by the API immediately and returned to us for storage. Note that, due to failure on the data collection infrastructure, I have some gaps in this dataset; specifically, 1) Dec

[1]Note that /r/The_Donald was banned by Reddit in 2020 (The New York Times, 2020). However, this research was done before this ban.

Table 3.1: Overview of this datasets. For each community, I report the number of posts that include URLs to trustworthy and untrustworthy news sources.

Community	#Posts		#Unique URLs	
	Trust.	Untrust.	Trust.	Untrust.
Twitter	7,123,715	686,497	3,893,357	291,354
Reddit	23,605,406	1,342,429	11,170,005	612,213
/r/The_Donald	528,142	190,742	385,384	122,204
4chan	458,431	75,705	275,422	37,472
Gab	2,369,149	2,265,336	749,547	385,317
Total	33,556,092	4,369,923	14,636,451	984,812

4-11, 2016; 2) Dec 25, 2016 to Jan 08, 2017; 3) Dec 17, 2017 to Jan 28, 2018; and 4) Sep 20 to Oct 31, 2018. I extract all tweets containing URLs to one of the news sources I study, collecting 7M tweets containing URLs to trustworthy news sources and 686K tweets containing URLs to untrustworthy news sources. The total number of unique (news) URLs is 3.9M and 291K for, respectively, trustworthy and untrustworthy news.

Reddit and /r/The_Donald. For Reddit, I use the monthly dumps available from Pushshift (Baumgartner et al., 2020). I collect all submissions and comments from Jan 1, 2016 to Oct 31, 2018, and extract all submissions and comments that include a URL to the news sources in my data. For the whole Reddit dataset, I find 24M and 1.3M posts with trustworthy and untrustworthy news URLs, respectively, while for /r/The_Donald, 528K posts with trustworthy and 191K with untrustworthy news URLs. The number of unique URLs is 11.2M and 612K, respectively, for trustworthy and untrustworthy news for the entirety of Reddit and 385K and 122K for /r/The_Donald. Note that the Reddit dataset also includes /r/The_Donald, as I aim to study the dynamics of Reddit as a whole. Nevertheless, even though /r/The_Donald is a minimal subset, and as such, it has a negligible effect on the analysis, I remove it from Reddit for this influence estimation presented in Section 3.5.2.

4chan's /pol/. I obtain all posts on 4chan's Politically Incorrect board (/pol/) between Jun 30, 2016 and Oct 31, 2018 from (Papasavva et al., 2020). I extract all posts containing URLs to one of the news sources, collecting 458K and 76K posts with URLs to trustworthy and untrustworthy news sources, respectively, for a total of 275K and 37K, respectively, unique URLs.

Gab. I use the data collection methodology presented in (Zannettou et al., 2018a) to collect Gab posts from Aug 10, 2016, and Oct 31, 2018. Once again, I extract posts that include URLs to trustworthy and untrustworthy news sources, collecting 2.4M posts containing trustworthy news URLs and 2.3M posts containing untrustworthy news URLs.

In total, I extract 15.6M unique URLs, 14.6M pointing to trustworthy and 985K to untrustworthy news sources posted on the five Web communities during Feb 2019. Note that all the posts collected from these platforms are obtained in real-time. I note that the Twitter and Reddit datasets start a few months earlier (January 2016) than Gab and 4chan. This is because the authors of (Papasavva et al., 2020) began collecting 4chan data in June 2016, and 4chan data is ephemeral; therefore, it is not possible to retrieve older posts. Gab, on the other hand, was launched in August 2016.

News Content. Next, I collect the *content* of the 15.6M news articles using the Newspaper library for Python3 (Lucas Ou-Yang, 2020), which, given a URL, retrieves the text from an article. I collected this data between February and April 2019.

Note that I am unable to retrieve the content of about 1.4M unique articles due to server-side problems of news outlet websites (e.g., the article was deleted from the server or the server was down at that time) and about 1M unique articles because of paywalls. For the latter, manual inspection shows that paywalls typically trigger a set of standard sentences being displayed instead of the actual news content (e.g., "sign up for a new account and purchase a subscription"). Thus, I parse results to exclude

articles containing these sentences. In the end, I gather the text of 13M unique articles, 12M from trustworthy sources, and almost 1M from untrustworthy ones.

To assess the extent of the problem of partially downloaded articles, I randomly select one URL from each of the top 300 most popular trustworthy domains out of 1036 trustworthy domains, whose URLs account for 92.5% of all trustworthy URLs, and one URL from each of the 37 untrustworthy domains. Then I download the text of these articles by using the Newspaper Library, and then manually compare the text with the content of the Web page loaded in a regular Web browser. Among 337 articles from the trustworthy and untrustworthy news outlets, the library downloads the full content of articles for 312 of them (92.6% of the total), while, for 25, the text is partially downloaded. The articles that could only be partially downloaded all belong to trustworthy news outlets.

To investigate the amount of content that is downloaded in these partial articles, I inspect the content of the articles from the top three domains for which the problem occurs, namely "nytimes.com"(4.47% of all URLs in this dataset), "cnn.com" (2.24%), and "nyti.ms" (1.54%), and find that the first 5, 8, and 4 paragraphs of the articles are downloaded, respectively. Assuming that all articles from the 25 outlets for which I could only download partial content present this problem, this would account for 12% of the URLs in this dataset.

Following the "inverted pyramid" style (Wikipedia, 2021) in journalism, the first paragraph of a news article usually provides the most important information about the story, with details appearing in later paragraphs. Because of this relatively standard style, I expect the issues related to partial content download to be minimal. Note that in this analysis pipeline I only use the text of articles to extract named entities, and I manually find that several named entities are included in the partially downloaded content of those articles. See Section 3.4.2 for more details.

3.3 Methodology

Having built my dataset, I show the methods that I use to analyze the dataset below.

3.3.1 Natural Language Processing

I first describe the NLP component, which I use to extract meaningful named entities that are referenced both on news articles and on discussions on several Web communities. This NLP component involves two models: 1) a true case model that predicts and converts text into its correct case (e.g., "donald trump is the president" is converted to "Donald Trump is the President"); and 2) a named entity detection model that extracts known named entities from the text along with an associated category (i.e., whether the extracted entity is a person, an organization, etc.). The former is necessary since the latter is case-sensitive.

True Case Model. I use *TrueCaseAnnotator* from the Stanford CoreNLP toolkit (Manning et al., 2014). This converts the case of the text to match as it should appear in a well-edited format (e.g., "united states" becomes "United States"), using a discriminative model built on Conditional Random Fields (CRFs) (Lafferty et al., 2001).

Named Entity Extraction Model. To obtain named entities, I rely on the SpaCy library (spaCy, 2019) and the *en_core_web_lg* model. I choose this model since it is trained on the largest available dataset. The named entity detection model leverages millions of Web entries consisting of blogs, news articles, and comments to detect and extract a wide variety of entities from text, ranging from people to countries and events (see (spaCy, 2019) for a list of all the supported types of entities). The model relies on Convolutional Neural Networks (CNNs), trained on the OntoNotes dataset (Weischedel et al., 2019), as well as Glove vectors (Pennington et al., 2014) trained on the Common Crawl dataset (Common Crawl Repository, 2019).

3.3.2 News Stories Identification

The news articles in this dataset cover various aspects ranging from politics to entertainment. Among all categories, I focus on politics because previous work showed that these stories are often discussed differently in different online communities (Zannettou et al., 2017) and are often used to spread disinformation narratives (Starbird, 2017; Wilson et al., 2018; Zannettou et al., 2019b). Therefore, I design a news stories identification component to group political news articles covering the same "story."

I use the definition by Marcelino et al. (Marcelino et al., 2019), whereby every news "story" is composed of several story "segments." In a nutshell, I perform three tasks: 1) I identify segments using the GDELT dataset (Leetaru and Schrodt, 2013); 2) I build a graph where news articles are nodes and edges are common segments discussed in them; and 3) I perform community detection on the graph to identify articles that discuss the same story.

Event Identification with GDELT. In this study, I use events identified by GDELT (Leetaru and Schrodt, 2013) in an article as a news story "segment." GDELT is a dataset containing event information for articles (published between Oct 30, 2015 and Nov 3, 2018) covering political news stories. GDELT's focus on politics makes it the ideal candidate for this analysis pipeline. I find 31M unique news URLs belonging to the 1,073 domains that I study in the GDELT dataset, composed of 30M unique trustworthy news URLs and 712K unique untrustworthy news URLs. For each URL, GDELT lists *events* (e.g., "Egyptian Minister of Foreign Affairs Mohamed Orabi attended the summit yesterday" (Leetaru and Schrodt, 2013)), which are each assigned a globally unique identifier ("Event ID"). The event extraction is performed at the sentence level by an automated coding engine called TABARI (Leetaru and Schrodt, 2013), which identifies the actors involved in the event, the action performed, and where the event happened. The result of this is that two different sentences

referring to the same event are given an identical event ID, and each sentence is an *event mention* of this event ID (GDELT, 2015). When identifying an event from a sentence, GDELT also gives a confidence score to the event mention, ranging from 10 to 100% (in 10% increments), representing the confidence that this sentence indeed corresponds to that event ID (GDELT, 2015).

To extract the segments associated with the news articles in this dataset, I first look up which of the URLs in this dataset are present in GDELT after a number of pre-processing steps, such as expanding shortened URLs, removing the query string as well as the `www` prefix or slash suffixes from the URLs. Then, I extract the list of corresponding events for each matched URL. I find 3M unique URLs, corresponding to 15.8M Web community posts, in the dataset, and these 3M unique URLs comprise 24.6M event mentions (i.e., story segments). The reason why GDELT only covers a fraction on the URLs in this dataset is that GDELT only focuses on political news, leaving out other topics like entertainment.

Graph Representation of Stories. After labeling news URLs with events that are relevant to them, I build a graph linking single articles with common events they cover. In other words, if two articles share one event, these two articles are "related" and the more events two articles share, the closer their content is. The graph is built as follows: 1) I treat each URL (stripped of its parameters) as a node. 2) GDELT associates a confidence score with each event, which indicates how confident the platform is that the article is actually discussing such event. Through manual inspection, I find that events with low confidence include considerable noise and would bias these results. I therefore discard events with a confidence lower than a threshold c. 3) If two URLs share at least one event ID, I build an edge between the two nodes. 4) The edge weight is computed as the number of unique event IDs that two URLs share. 5) Two URLs being connected by an edge does not necessarily mean the two URLs *really* share

common events due to incorrect event extractions by GDELT with low confidence scores. To reduce the number of wrong edges, I therefore remove edges with a weight lower than a threshold e before performing community detection. I discuss how I selected an optimal value for c and e later in this section.

Community Detection. Two URLs that share one or more events are not guaranteed to cover the same story. To further refine the association between common events and news stories, I apply community detection on the graph. I then consider URLs to belong to the same news story if they are part of the same community. I apply the Louvain Method (Blondel et al., 2008), which allows us to efficiently find communities in sparse graphs like the one I am dealing with: this graph is composed of 3M nodes and 1.6M edges. In the following, I discuss how I select the optimal values for the thresholds c and e for these experiments.

Selecting the Optimal Thresholds. As previously discussed, this community detection approach relies on two thresholds, one to discard event IDs that have a low confidence in GDELT (c) and one to remove edges between news articles that share a small number of events (e). In the following, I describe the experiments that I performed to select the optimal values for these two thresholds.

I first discard URLs with more than 60 unique event IDs, since I find that these results are due to errors in GDELT's crawling process; when manually inspecting these URLs, I find that the content is mostly homepages of news outlets, including numerous headlines, and therefore flagged with multiple spurious event IDs. Further, I remove communities whose URLs are from a single domain only, since by manually looking at the clusters, I find that these URLs are published on the same day and share several events even though their texts are totally different.

To determine the optimal parameter combination of confidence threshold c and story edge weight threshold e, I follow a grid search approach. I select values for the

confidence threshold c from 10% to 100% with 10% increments and weight thresholds for $e \in \{1, 2, 3, 4\}$. In total, I obtain $4 \times 10 = 40$ parameter combinations. For each parameter combination, I apply the Louvain Method, obtain the corresponding communities of URLs, and randomly select 20 communities with sizes larger than 10 for further inspection. These samples are independently reviewed by two annotators to determine if the articles in them belong to the same story. To ensure the accuracy of subsequent analysis, I set a condition that a parameter combination is an optimal parameter candidate only if all the 20 communities in the sample with the corresponding parameter combination have a precision (i.e., the number of correctly grouped articles vs. all articles in the community) above 90%. After comparing their results, the two annotators agreed that among the 40 parameter combinations, the parameter combinations $\{(c = 60\%,\ e = 3),\ (c = 70\%,\ e = 3),\ (c = 80\%,\ e = 3),\ (c = 90\%,\ e = 3),\ (c = 100\%,\ e = 3),\ (c = 60\%,\ e = 4),\ (c = 70\%,\ e = 4)$, and $(c = 80\%,\ e = 4)$ satisfy the condition of being candidates for the optimal parameter. Note parameter combinations $(c = 90\%,\ e = 4)$ and $(c = 100\%,\ e = 4)\}$ only have 12 communities with sizes larger than 10, respectively, and all of these communities have a precision above 90%. Since the parameter combination $(c = 60\%,\ e = 3)$ produces the most communities (in total 43,312) among all the optimal parameter candidates, I settle on this parameter combination, and set the parameter confidence threshold $c = 60\%$ and the story edge weight threshold $e = 3$.

To further verify the performance of this approach using the chosen optimal parameter combination, I randomly select 400 stories, and check them. 397 out of 400 stories (99%) have a precision above 90%. I therefore conclude that this parameter selection is good for these purposes.

Alternatives to GDELT. Note that I also tested alternatives to GDELT as external "ground truth." More specifically, I group articles based on the TF-IDF (Manning et al., 2008) of their text and DBSCAN clustering (Ester et al., 1996). However,

manual analysis reveals that the performance of these methods is substantially worse. An alternative approach would have been to use topic modeling, e.g., LDA (Blei et al., 2003). However, these methods are most effective when modeling topics that are broader than fine-grained news stories and are therefore less appropriate than this approach in this case. The reason is that features from LDA, as well as TF-IDF, are obtained at the *word* level. So keywords shared by two different stories can interfere with the clustering result while features from my method are obtained from the *sentence* level (i.e., events in the stories), which avoids this issue. One example is a pair of stories found in my result: "Donald Trump's call to punish flag burners caters to voters in his base" and "air conditioning company Carrier says it has a deal with Trump to keep jobs in Indiana." Both stories are from Nov 29, 2016 (and thus cannot be distinguished by date), and their text share some keyword candidates, e.g., "donald," "trump," "president-elect," and "tweet," which make it difficult for LDA and TF-IDF to distinguish between them.

3.3.3 Influence Estimation

After grouping articles into news stories, I am interested in studying the temporal characteristics of how these stories are discussed in various communities of interest. More specifically, I aim to understand and measure the interplay between multiple Web communities with respect to the news stories they share. To do so, I create a time series that captures the cascades of each news story per Web community. After obtaining the time series, I model the interplay between Web communities using a statistical framework known as Hawkes Processes (Hawkes, 1971), which lets us quantify the influence that each Web community has on the others with respect to the dissemination of news stories.

Timeseries compilation. As a first step, I organize the data into timeseries. For each Web community of interest, I focus on all the news stories that are in the overlap

time, i.e., between Aug 10, 2016 and Sep 19, 2018. Also, I exclude news stories in which at least one URL falls in the gap described in "Twitter" part of in Section 3.2. This step ensures all the URLs in the stories can appear in different Web communities, avoiding underestimation of the influence of a specific Web community. In total, the number of stories used for estimation is 31,065 out of the 43,312 stories obtained. Next, for each news story i, on each Web community k, I build a time-series $u_{ik}(t)$, whose value is the number of occurrences of news URL related to a specific news story in hour t.

Hawkes Processes are self-exciting temporal point processes (Hawkes, 1971) that describe how events (in this case, posts including news URLs pertaining to a news story) occur on a set of processes (in this case, Web communities). A Hawkes model consists of K point processes, each with a *background rate* of events $\lambda_{0,ik}$. Note that the events considered for Hawkes processes are a set of posts made on Web communities and do not have to be confused with the event IDs from the GDELT dataset that I used to identify the news stories. For us, the point processes will be the timeseries $\{u_{ik}|k = 1, \ldots, K\}$ for a given story i. The background rate is the expected rate at which events referring to a story will occur on a process *without* influence from the processes modeled or previous events; this captures stories mentioned for the first time, or those seen on a process I do not model and then occur on a process I do model. An event on one process can cause an *impulse response* on other processes, which increases the probability of an event occurring above the processes' background rates. The shape of the impulse response determines how the probability of these events occurring is distributed over time. Hawkes Processes are used for various tasks like modeling the influence of specific accounts (Alvari and Shakarian, 2019; Zannettou et al., 2019b; Zannettou et al., 2019a), quantifying the influence between Web communities (Zhou et al., 2013; Zannettou et al., 2017; Zannettou et al., 2018b),

and modeling information diffusion (Soni et al., 2019; Kong, 2019; Guo et al., 2015; Lukasik et al., 2016). Here, I use them to quantify the influence between multiple Web communities with respect to the dissemination of news stories.

Model fitting. I assume a Hawkes model that is fully connected: each process can influence all the others, as well as itself, which describes behavior where a user on a Web community sees a news story and re-posts it on the same platform. Fitting a Hawkes model to a series of events on a set of processes provides us with the values for the background rates for each process, along with the probability of an event on one process causing events on other processes. Background rates also let us account for the probability of an event caused by external sources of information—i.e., a Web community that I do not model. Thus, while I only model the influence for a limited number of Web communities, the resulting probabilities are affirmatively attributable to each of them as the influence of the greater Web is captured by the background rates. This also helps to address the limitations of these datasets. In particular, the tweets that are not included in the Twitter 1% Streaming API are absorbed into the background rate of each community, avoiding erroneous attribution of an event to a different community. According to this hypothesis, the influence of Twitter is underestimated. To fit a Hawkes model for each news story, I use the approach described in (Linderman and Adams, 2014; Linderman and Adams, 2015), which uses Gibbs sampling to infer the parameters of the model from the data, including the background rates and the shape of the impulse responses between the processes.

Influence. Overall, this enables us to capture the interplay between the posting of news stories across multiple Web communities and quantify the influence that each Web community has on each other. More precisely, I use two different metrics: 1) ***raw influence***, which can be interpreted as the percentage of news story appearances that are created on a *destination* Web community in response to previously occurring

appearances on a *source* Web community; and 2) ***normalized influence*** (or efficiency), which normalizes the raw influence with respect to the number of news story appearances on the source Web community, hence denoting how efficient a community is in spreading news stories to other Web communities.

3.4 General Characterization

3.4.1 News Sources

Using the NewsGuard scores, I find that out of the 1,073 news sources in this dataset, 1,036 (96.6%) are labeled as trustworthy (e.g., the New York Times, the Washington Post) and only 37 (3.4%) untrustworthy (e.g., Infowars, Breitbart), i.e., they have a score of less than 60/100. However, out of the 15.6M unique URLs in this dataset, 985K are to untrustworthy and 14.6M trustworthy news sources. That is, over 6.3% of posted URLs are from untrustworthy news, even though these only account for 3.4% of the sources.

3.4.2 Named Entities

Next, I describe the named entities extracted as per the methodology described in Section 3.3.1. Note that although GDELT does offer extracted entities in their metadata, I find that their labeling is not suitable for these purposes. More specifically, GDELT relies on two databases of public figures, which were last updated in 2010 (The Computational Event Data System, 2014). So, for example, "Trump" does not appear in any of the entity metadata. Instead, I use TrueCaseAnnotator, SpaCy, and *en_core_web_lg*. Next, I describe the named entities extracted from the news articles in this dataset and then move to the ones extracted from the posts on Web communities containing news URLs.

Articles. Using the article text (Section 3.2) and this pipeline's NLP component

(Section 3.3.1), I extract the named entities for each article in this dataset during Apr 2019. The most popular entities referenced in both trustworthy and untrustworthy news are related to US politics. For example, Donald Trump is referenced in 18% and 27% of the trustworthy and untrustworthy news articles, respectively.

Web Communities. I also study the named entities that appear in posts, including news URLs, and the entities are extracted during Apr 2019. Note that these entities are *not* related to the text of the news article pointed by the URL, but rather the comment it was posted with. Similar to the named entities detected on the news articles, most of the entities appearing in the posts are related to world events and politics and, in particular, US politics. For instance, one of the most popular entities is "Trump," with 7.6%, 9%, 9.4%, 13.2%, and 2.9%, for Twitter, Reddit, /r/The_Donald, 4chan, and Gab, respectively.

There are also some interesting differences between top entities across the communities: e.g., on 4chan, several entities are Jewish and Israel related (Israel with 8%, Jews 5%, Jewish 4.8%, and Israeli with 3.7%), while, on /r/The_Donald and Gab, I find Islam-related entities ("Muslim" with 4.6% on /r/The_Donald and 1.1% on Gab).

As the top named entities are mostly related to politics, to see if any type of named entities are missing, I select 10 common named entities in each of the following categories: politics, sports, entertainment, business, locations, health, science, technology, and religion. I find all of them in the results of the named entity extraction from both articles and posts, which shows that this analysis is representative.

3.4.3 News Domains in Web Communities

Next, I study the popularity of the news sources on each Web community. Overall, I find that 8.7%, 5.0%, 25.7%, 12.6%, and 48.7% of the total occurrences point to untrustworthy URLs for Twitter, Reddit, /r/The_Donald, 4chan, and Gab, respectively, while the rest point to trustworthy URLs.

On Twitter and Reddit, the most popular domains are mainstream, trustworthy news sources like The New York Times, The Washington Post, and CNN. On /r/The_Donald, the most popular news source is Breitbart, which was considered an untrustworthy news source from NewsGuard at the time of this experiments (57 score). I also find other untrustworthy news sources, e.g., the Gateway Pundit (20 score), Zerohedge (0 score), and Infowars (25 score), among the top 12 most popular news sources. I observe this phenomenon to an even greater extent on Gab: the three most popular news sources are untrustworthy, with Breitbart being included in almost 15% of Gab posts with news URLs. For 4chan, I find mostly trustworthy news sources in the top 20, with the exception of Breitbart, the Russian state-sponsored RT (32.5 score), and Zerohedge. The most popular news source is actually the Daily Mail, which has a 64.5 NewsGuard score.

Overall, I find that /r/The_Donald and Gab are particularly polarized communities that extensively share news from untrustworthy sources, while Reddit and Twitter, which are more mainstream, do so to a much lesser extent. 4chan seems to be somewhat in the middle of the two: I find a substantial number of URLs from both trustworthy and untrustworthy news sources, which is perhaps surprising considering that 4chan is one of the most "extreme" communities on the Web (Hine et al., 2017).

I also study trustworthiness at the granularity of specific URL appearances. For each URL appearance, I extract the news source and assign its NewsGuard score. In Figure 3·2, I plot the resulting CDF. Note that Gab shares substantially more URLs from less trustworthy sources (with a median score of 64.5), followed by /r/The_Donald (median 82.5). 4chan shares substantially more URLs from trustworthy sources (with a median of 95) compared to Gab and /r/The_Donald, and its median matches the one from Twitter (95). Finally, Reddit users predominantly share URLs from trustworthy sources (with a median of 100). To confirm these observations, I perform χ^2 tests

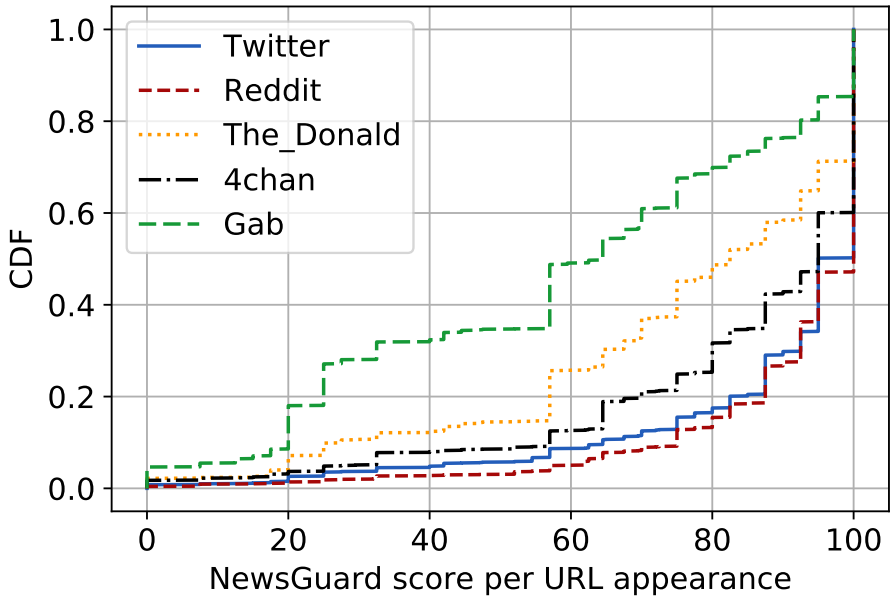


Figure 3·2: CDF of the NewsGuard score for each news URL.

of independence on the proportion of trustworthy and untrustworthy news shared in each community. The results reject the null hypothesis that there is no difference in the rate of trustworthy and untrustworthy shared by Web communities ($p < 0.01$, with statistics values higher than 10,000 for all experiments).

3.4.4 Main Takeaways

I find that different Web communities discuss different types of news; e.g., 4chan focuses more than others on Jewish and Israel-related news, and /r/The_Donald and Gab on news about Muslims. Also, users on /r/The_Donald and Gab prefer to cite untrustworthy news outlets to support their discussion.

3.5 Analyzing News Stories

In this section, I set to understand how news stories, rather than single URLs, are discussed on different Web communities. First, I describe the news stories identified, focusing on the differences across platforms. Then, I study the influence that different

communities have on each other using Hawkes Processes, and some case studies.

3.5.1 News Stories

Out of the 15.6M unique news URLs, corresponding to 38M posts, I extract from Twitter, Reddit, /r/The_Donald, 4chan, and Gab, 3.2M unique news URLs of them, corresponding to 15.8M posts, appear in the GDELT dataset. After the story identification, I extract 43,312 unique stories: 21,878 on Twitter, 42,783 on Reddit, 4,943 on /r/The_Donald, 5,007 on 4chan, and 9,929 on Gab. These correspond to 109,153 unique URLs, 105,143 trustworthy and 4,010 untrustworthy. Trustworthy URLs occur 130,235 times on Twitter, 469,058 on Reddit, 9,249 on /r/The_Donald, 14,184 on 4chan, and 46,559 on Gab. Untrustworthy URLs occur 2,759 times on Twitter, 9,221 on Reddit, 990 on /r/The_Donald, 656 on 4chan, and 11,469 on Gab. Recall that GDELT is focused on political stories with national and international relevance. As such, I expect news about local matters as well as sports or entertainment not to be included in this dataset.

Comparing this to the results in Section 3.4.3, I note that untrustworthy URLs appear in 2.1% of all news story posts on Twitter (compared to 8.7% for all news URL occurrences), while for Reddit, this is 1.9% (compared to 5.4% overall), 4.4% for /r/The_Donald (27% overall), 4.4% for 4chan (14% overall), and 21% for Gab (49% overall). This might indicate that although some communities prefer to use untrustworthy news URLs to support their discussion when discussing political news stories, they still prefer to quote trustworthy ones. The analysis in Section 3.5.3 suggests that this is sometimes done to give trustworthiness to a claim.

I then look at the *lifespan* of news stories on different Web communities, i.e., the time between the first and the last time a URL from a given story is posted on a platform. Fig. 3·3 plots the CDF of the news story lifespans on the five Web communities. The vast majority of stories on all platforms are short-lived. Overall,

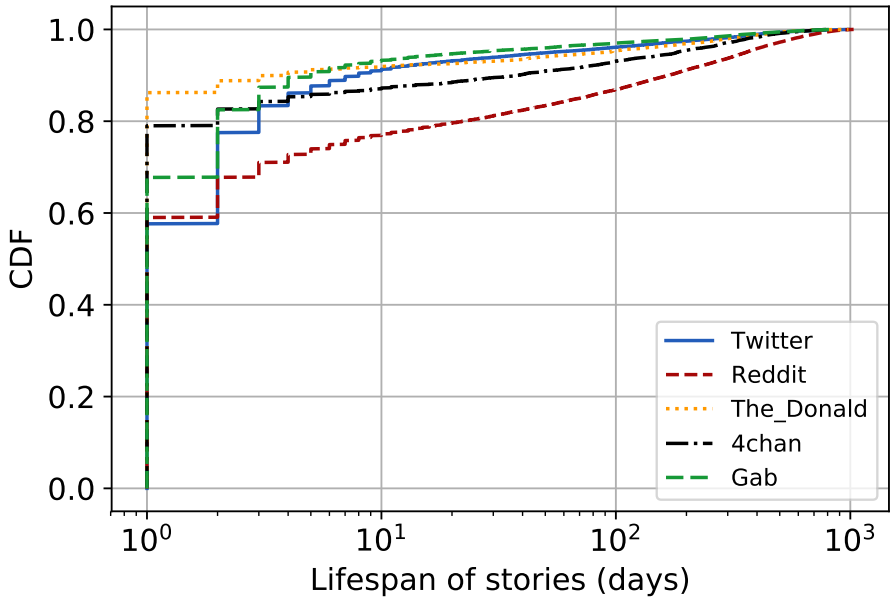


Figure 3·3: CDF of lifespan of stories on Web communities.

Reddit users discuss news stories the longest, as almost one out of ten (8.8%) last for 200 days or more. Interestingly, 4chan comes next, with 4.6% of the stories being discussed on that platform for 100 days or more. This is interesting, considering that posts on 4chan are ephemeral (i.e., they are deleted after a few days), with news content disappearing from the platform on a regular basis; hence, the fact that the 4chan community keeps discussing the same story for long periods of time indicates that new threads about it are constantly created.

3.5.2 Influence Estimation

I now study the influence of Web communities with respect to news stories. As discussed in Section 3.3.3, I create a Hawkes model for all the stories that are in the overlap time among all the Web communities, and more precisely, 31,056 stories. Note that each model consists of five processes, one per community. Then, for each story, I fit a Hawkes model using Gibbs sampling.

Table 3.2 reports the overall number of events (i.e., appearances of news stories)

modeled with Hawkes Processes for each Web community. Looking at the raw number of events, I observe that, unsurprisingly, Reddit and Twitter are the communities with the most events in the selected 31,056 news stories. Fitting a Hawkes model provides us with the parameters for the background rates and impulse responses of each process; thus, I am able to quantify the influence that each Web community has on each other. Figure 3·4 shows the influence estimation results, capturing how influential and efficient Web communities are in spreading news stories.

Overall, I make several observations. First, in terms of raw influence (see Figure 3·4(a)), Twitter and Reddit are the most influential Web communities, mainly because of the large number of news story appearances that they produce. Moreover, out of the three smaller communities (4chan, Gab, and /r/The_Donald), /r/The_Donald is the second most influential Web community for news stories that appear on Twitter and Reddit, which is particularly interesting since the overall number of news story appearances on /r/The_Donald is substantially smaller compared to the ones on 4chan (see Table 3.2). Finally, in terms of efficiency (see Figure 3·4(b)), /r/The_Donald is by far the most efficient Web community in making news stories appear on other Web communities.

The last column in Figure 3·4(b) is the sum of normalized influence from the specific source community to the rest of the platforms (i.e. excluding self-influence). Generally, the larger the percentage, the larger the overall external influence from the source community to all the others; for instance, the cell corresponding to /r/The_Donald's sum of normalized influence to other Web community is the largest, suggesting /r/The_Donald is the most efficient Web community in spreading news stories to other Web communities.

Note that statistical tests, e.g., to elicit confidence intervals for the influence probabilities, are hard to compute for Hawkes processes as, to the best of my knowledge,

Table 3.2: Number of events modeled via Hawkes Processes.

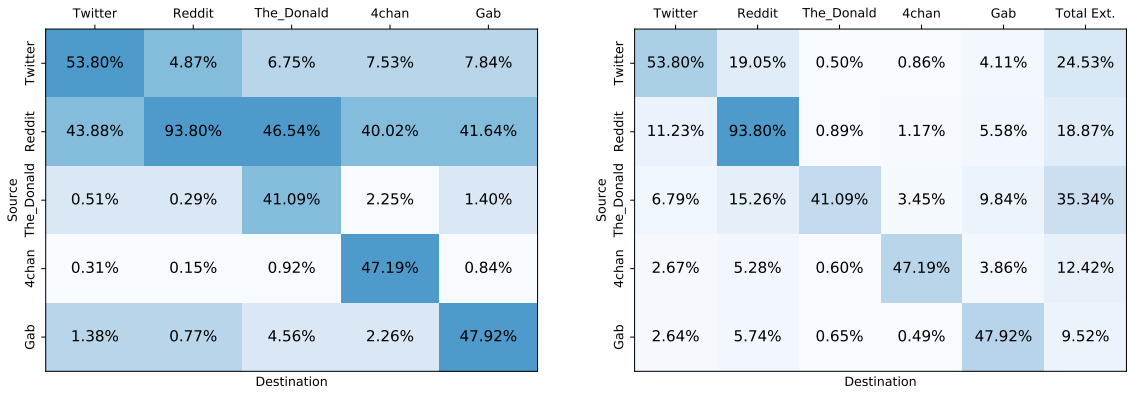
Twitter	Reddit	/r/The_Donald	4chan	Gab	Total
72,608	291,285	5,496	8,202	37,686	415,277

there is no statistical tool that is both meaningful and tractable. More specifically, goodness of fit for Hawkes process exists, but it has not been implemented or tested at scale, and therefore I leave it as part of future work.

3.5.3 Selected Case Studies

Since this influence estimation is done on a per-story basis, I can identify stories for which each community is the most influential. I take a closer look at some case studies I believe to be particularly interesting. To do so, I calculate the overall external influence of each community and extract the top 20 most influential stories (i.e., I do not include a community's self influence). From this, I obtain a set of 100 unique stories, which are the ones where either Twitter, Reddit, /r/The_Donald, 4chan, or Gab is the most influential community on the other communities. Below, I present some case studies from these 100 stories.

2016 Presidential election. 37 of the 100 stories are what I consider 2016 US Presidential election related. Although the five Web communities all extensively discuss the election, I find that different communities push different narratives and topics onto other platforms. For instance, Twitter influences the other platforms to discuss a conspiracy theory espousing that leaking the conversation of Michael Flynn with the Russian ambassador to the U.S. about sanctions on Russia, which led to Michael Flynn's resigning as National Security Advisor, is the plot of Democrats to attack the Trump administration (e.g., (McClatchy DC Bureau, 2017)). Reddit pushes a story with respect to Jared Kushner, the son-in-law of Trump, having undisclosed contacts with Russia during the 2016 U.S. Presidential campaign (e.g., (Chicago



(a) (b)

Figure 3·4: Influence estimation results: a) Raw influence between source and destination Web communities, which can be interpreted as the expected percentage of events created on the destination community because of previously occurring events on the source community; and b) Normalized influence (efficiency) of each Web community, which can be interpreted as the influence per news story appearance.

Tribune, 2016)). /r/The_Donald is influential in spreading a story suggesting that then Secretary of State Hillary Clinton silenced a security firm hired by State Department to investigate alleged security failures at the Benghazi embassy (e.g., (The Daily Caller, 2016)). 4chan spread a story seemingly confirming Trump's claim of election fraud via dead people voting in Pennsylvania (e.g., (The Washington Post, 2016)). Gab was influential for a story reporting Trump's wins in Wisconsin and Pennsylvania were confirmed via recount (e.g., (New York Post, 2017)).

Immigration. Over the past few years, immigrant and refugee crises have been often covered in the news, as also confirmed by previous research (Boudemagh and Moise, 2017). In this dataset, I find 11 news stories for which one Web community has influenced its spread on other platforms. On Twitter, I find a story with articles about migrants stuck in US airports as a consequence of President Trump's immigration ban in 2017 (e.g., (The New York Times, 2017)) and on Reddit, there is a story about an Iraqi immigrant lying that his mother would have received life saving medical

treatment in the US had their been no travel ban (e.g., (Amy Lange, 2016)). For 4chan, I find misrepresentations of remarks made by the Mexican government at the NAFTA summit, interpreted as an acceptance to pay for the border wall (e.g., (Infowars, 2018)). Among the stories for which Gab is influential, there is one defending the Muslim Ban by listing seven facts (e.g., (JOHN HAYWARD, 2017)) and another related to the U.S. Supreme Court upholding Trump's Muslim Ban on June 26, 2018 (e.g., (Lawrence Hurley, 2018)).

3.5.4 Main Takeaways

I find that Twitter and Reddit are the most influential Web communities w.r.t. discussing news stories. However, /r/The_Donald is the most efficient considered its small(er) size. My analysis also shows that different communities influence discussion on different stories. In particular, while Twitter and Reddit do so for major events reported by mainstream news along "neutral" narratives, more polarized communities are influential in discussing stories with specific narratives, from celebrating or criticizing political figures (New York Post, 2017; The Daily Caller, 2016), to promoting anti-immigration rhetoric (JOHN HAYWARD, 2017; Lawrence Hurley, 2018) or distributing false news and conspiracy theories (Infowars, 2018; The Washington Post, 2016).

3.6 Discussion and Conclusion

I analyzed the sharing and the spreading of online news, showing that different communities present fundamental differences; for instance, Gab and /r/The_Donald "prefer" untrustworthy news sources (e.g., on Gab, 48.9% of all news URLs are from untrustworthy sources, compared to the 8.7% for Twitter). I also found that smaller Web communities can appreciably influence the news discussion on larger ones, with /r/The_Donald being very effective in pushing news stories on Twitter and the rest of

Reddit.

Comparison between similar concepts used in my work and communication research. Note stories and mutual influence are also researched in communication research (Guo and Vargo, 2020; Guo and Zhang, 2020; Harder et al., 2017; Welbers, 2016). Below I will discuss the two concepts used in my research and communication research, respectively.

For the concept "story," the same aspects is that my definition of "story" is in consistent with prior work in communication research (Guo and Vargo, 2020; Harder et al., 2017; Welbers, 2016). I use levels of granularity to differentiate "events," "stories," and "themes." "News story" is defined in the intermediate level of granularity between events and themes, and stories are larger components combined by events.

The difference is that how I define "event" compared with prior communication research, which is actually from different systems of taxonomy. In the context of communication research, one news article can be coded as "event" (Welbers, 2016). Nevertheless, in my work, a "segment," which is a "news event" in GDELT[2], is a sentence in one news article. For instance in the news relevant to the retired Supreme Court justice Stephen Breyer [3], the sentence "Biden praised retiring Justice Stephen Breyer $\cdots$" is an event.

As a result, the news articles that are grouped into one story by our method are likely to share parts of similar text. These stories can contain events happened in only one day, but it can also contain events that last for longer.

However, for news articles that actually belong to one story but do not have similar sentences, my method may not be able to group them together. For instance, in (Guo

[2]In addition to GDELT global knowledge graph (GKG) dataset, it also has a dataset called GDELT Event Database: `https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/`, which is what I use for this research

[3]`https://www.usnews.com/news/best-states/south-carolina/articles/2022-01-27/at-least-3-judges-eyed-as-biden-mulls-supreme-court-pick`

and Vargo, 2020), if one news article is about "House approved spending bills to start Trump's border wall" and another news article is about "Trump pressured Mexico on border wall payment," both of which belong to the story "Trump's promise to build a border wall between the United States and Mexico." If the news article about "House approved spending bills to start Trump's border wall" mentions "Trump pressured Mexico on border wall payment" as a background introduction in the text, then using my method may be able to group the two news articles together. If not, then my method may not cluster them into one story.

The mutual influence between news outlets, known as intermedia agenda-setting (IAS) theory in communication research (Guo and Vargo, 2020; Guo and Zhang, 2020), differs the mutual influence in my work in the aspects that IAS theory focuses on the content *producers* on the Internet, e.g., New York Times and Washington Post. In particular, IAS theory is used to find out between two content producers, does one producer influence the other one in producing content of specific topics. While in my work, I do not take into consideration of the mutual influence among content producers. Instead, I focus on the content *disseminators* and *audience* on the Internet, which are mostly ordinary users of different Web communities. In particular, I am interested in discovering if there are any Web communities that are extremely influential in pushing news stories to other Web communities. For instance, the subreddit /r/The_Donald is influential in pushing conspiracy news stories to other Web communities, which helps the spread of misinformation.

Limitations. While I did my best to gather a view of online news discussion that was as comprehensive as possible, this dataset of news websites only includes English news websites as identified by the Majestic list and NewsGuard. Moreover, I focused my analysis on four social networks, i.e., leaving out others like Facebook, due to the difficulty of collecting data. In addition, the Twitter data is 1% of all the Twitter at

the time of my execution of this research in 2019. I still argue my contribution to the pipeline of mutual influence of social media from news stories' perspective. Finally, I relied on the GDELT dataset, which, as discussed, presented noise and crawling errors and on a named entity recognition model that is mostly trained on well-edited text like news articles. However, I took several steps to mitigate these issues by performing a sensitivity analysis that allowed us to build accurate communities of news articles to form the news stories that I analyzed. My approach could serve as the foundation for a wealth of research not only in computer science but also in journalism and political science.

Chapter 4

Fauxtography

4.1 Introduction

Recent years have seen an increase in false information published online and spread through social media (Kumar and Shah, 2018). An important aspect of news consumption is that users not only pay attention to text, but also to the accompanying images in the article. In fact, research in psychology shows that images play a crucial role in both how readers perceive certain issues (Zillmann et al., 1999) and in which articles individuals choose to read (Zillmann et al., 2001). Therefore, it is not surprising that images may be manipulated or misrepresented to mislead users.

In this chapter, I focus on *fauxtography* (Cooper, 2007), i.e., news images that have been modified or miscaptioned to change their intent, often with the goal of spreading a false sense of the events they purport to depict. Although previous research efforts have proposed detection tools for fauxtography (Zhang et al., 2018; Zlatkova et al., 2019), to the best of my knowledge, the *impact* of fauxtography on news discussion has not been studied. In particular, I set out to investigate two research questions:

- RQ1: Does sharing fauxtography increase engagement on social media?
- RQ2: Do fauxtography images have a life beyond their questionable verisimilitude (their appearance of being being real)? I.e., do new variants and memes using them appear on social media?

To answer these questions, I develop a computational analysis pipeline, shown

in Figure 4·1, geared to identify posts containing fauxtography at scale, measure the engagement of users sharing and favoring such posts, and understand how these images are used on different social media platforms. First, I gather 2.6 billion posts from three social media platforms (Twitter, Reddit, and 4chan) as well as 32M news articles published by over 1,000 news websites. Then, I extract all images appearing in these posts and articles, and use perceptual hashing (Monga and Evans, 2006) to match them to images labeled as fauxtography by the fact-checking site Snopes. In total, I identify 61K posts containing fauxtography shared by users over the two year period from 2016 to 2018.

To address RQ1, I analyze the reactions to posts containing fauxtography or images fact-checked as true by Snopes on social media, compared with the reaction to posts by the same users with no image or with general images characterized. I find that including fauxtography or images fact-checked as true by Snopes in posts does increase user engagement on social media. On the other hand, posting links to news articles that contain fauxtography (rather than posting images directly) does not increase engagement on Twitter, while it does yield more interactions on Reddit. Surprisingly, the extent to which an image is misleading – e.g., whether it is completely false or just partially true – does not significantly affect engagement, suggesting that the increased engagement is driven by the inflammatory and controversial nature of fauxtography more than its verisimilitude.

For RQ2, I search for variants of fauxtography images that each appear on all of the social media platforms. My intuition is that instances of fauxtography are likely have some sort of inherent exploitability making them suitable as a base for new memes. Visual memes have become important to the spread of racist and political ideology (Du et al., 2020; Zannettou et al., 2020b; Zannettou et al., 2018b) and have been used by state-sponsored actors to wage information warfare (Abidin, 2020; Zannettou et al.,

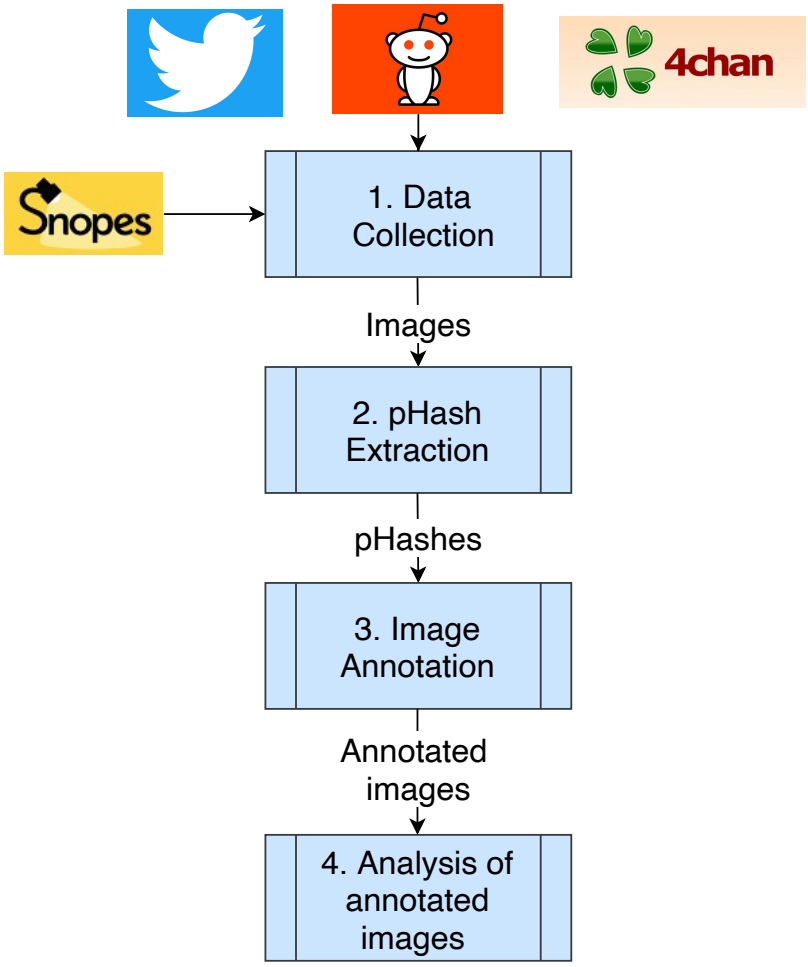


Figure 4·1: Overview of my computational analysis pipeline.

2020a). I find evidence of fauxtography images being turned into memes and being manipulated in ways not related to their original verisimilitude.

Finally, by focusing on three selected case studies of fauxtography which spawned new variants, I will discuss implications for dealing with fauxtography in the wild, considering the current environment of social media moderation.

4.2 Fauxtography

The term fauxtography was first coined by (Cooper, 2007) in the context of the 2006 Lebanon war, as combination of the word *faux* (French for false) and *photography*.

Cooper defines fauxtography as "visual images, especially news photographs, which convey a questionable (or outright false) sense of the events they seem to depict." Fauxtography usually involves manipulated images aiming to influence the emotions of viewers. Therefore, it involves deception, often realized by directly manipulating the images, captions, or overall the narrative associated with the image.

To better explain what fauxtography is, I provide two examples. Figure 4·2 shows a picture of a protester in the UK holding a sign reading "Black Lives Matter," which was manipulated to instead read "Lincoln Was Racist." Online sources also erroneously claimed that the person holding the sign was a Missouri State University student at a US protest. Figure 4·3 shows an image that was not manipulated, but that has often been used out of context and miscaptioned to imply that migrants on a caravan to the US in 2018 had burned the American flag. In reality, this photo was taken at an anti-Trump protest in the US and the flag is actually a Trump banner, not a US flag. These examples demonstrate two important characteristics of fauxtography, distinguishing them from "simple" fake images: 1) they are related to news or public affairs, and 2) users who see them can be fooled relatively easily if the images are not fact-checked.

4.3 Dataset

In this section, I present how I collect data to build my dataset, which is the first component of my pipeline, as shown in Figure 4·1. My study relies on two types of data sources: 1) *images* from Web communities and news articles posted on them; and 2) *annotation sources* to identify which images are fauxtography. For the former, I use Twitter, Reddit, and 4chan, and in particular images shared between July 1, 2016 and October 31, 2018; basic statistics are reported in Table 4.1. As an annotation source, I use `Snopes.com`, and specifically images posted on its fauxtography section.

(a) Manipulated (b) Original

Figure 4·2: This picture originally depicted a UK protester holding the "Black Lives Matters" sign. It was manipulated so that the sign says "Lincoln was Racist" and the person has been mischaracterized as being a Missouri State University student. See `https://www.snopes.com/fact-check/abe-lincoln-racist-protest-sign/`

Platform	#Posts	#Image URLs	#Images Obtained
Twitter	2,213,019,239	701,806,921	435,244,799
Reddit	295,460,914	78,682,398	61,703,316
4chan	99,614,382	27,044,132	23,379,630
News Articles	32,200,604	28,654,146	27,360,218
Snopes	2,286	16,206	7,835

Table 4.1: Overview of my datasets.

This allows me to identify all images that Snopes classified as fauxtography between the early 2000s and October 2019. Note that this analysis pipeline supports any Web community and any annotation source; however, in the following, I provide details of the sources used in this research.

Images shared on Web communities. First, I collect images posted publicly on Twitter, 4chan, and Reddit. For Twitter, I collect data using the 1% Streaming API, with tweets stored as they were posted, in real time. In total, I parse 2.2B tweets, 702M of which contain at least one image. Note that the Twitter API does not return images directly, but rather a URL pointing to the image. I download the images

Figure 4·3: Miscaptioned image used to falsely claim that people in the migrant caravan burnt an American flag. See `https://www.snopes.com/fact-check/caravan-burning-flag/`

in March 2020 and are able to retrieve 435M of them. The remaining images are unavailable, either because the image URL had changed, the tweet was deleted, or because the account that posted it was suspended.

For Reddit, I use the Pushshift dataset (Baumgartner et al., 2020). I obtain 295M posts, 79M of which contain images. Of these, I successfully retrieve 62M images, with the rest having been deleted. For 4chan, I use the dataset from (Papasavva et al., 2020) and obtain 100M posts from 4chan's Politically Incorrect board (/pol/). The dataset does not include the images posted on /pol/ (only an md5 checksum), hence I use `4plebs.org`, an archival service, to collect the images. Overall, I collect 27M image URLs from 4plebs, from which I am able to download 23M images.

Images from news articles posted on Web communities. On most social networks, when a user shares a news article, the platform often automatically generates a preview for it. Typically, this includes the main image of the article (i.e., the one appearing on the top). The preview is important with respect to users' image sharing behavior, thus, I complement the image data collection with images included on news articles shared on Twitter, Reddit, and 4chan.

Original Labels	True	Mostly True	Mostly False	False	Miscaptioned	Legend	Outdated	Satire	Unproven	Mixture
Labels used in this research	Merged True		Merged False			*Not considered*				

Table 4.2: Overview of the fauxtography labels assigned by Snopes and of the grouping that I use for the analysis in this research.

To do so, I use a systematic approach to create a list of news outlets; I start from the top 30K Majestic (Majestic, 2019) websites released as of February 2019, and use the VirusTotal API (VirusTotal, 2020), a domain categorization service, to get domains categorized as "news" and "news and media." Note that the news outlet labeling given by VirusTotal is not always accurate, e.g., domains like `adbusters.org` are incorrectly classified as news outlets. To solve this problem, I use the NewsGuard API (NewsGuard, 2019d) to refine the which domains are actually news outlets, and only select those listed in NewsGuard as of February 2019. In total, I identify 1,037 news outlets.

I then collect posts containing URLs to the 1,037 news outlets posted on Twitter, Reddit, and 4chan, gathering a total of 32M news articles, with approximately 29M including image URLs. Note that I only consider the top image from each article, which is the image that appears on top of the article. To collect the images, I use the Newspaper3k Library (Lucas Ou-Yang, 2020) to parse the HTML of the 32M news articles, and then extract the URL of the top image identified by Newspaper3k. I am able to download 27M images from the 29M image URLs in the news articles.

Snopes. As mentioned, I annotate images using Snopes, a website dedicated to fact-checking news, which has a special section dedicated to fauxtography. [1] Each entry in this section consists of a topic and a claim associated to an image, which is rated by Snopes using ten possible labels, listed in Table 4.2. For this analysis, I merge these labels into two groups: Merged True (True and Mostly True) and Merged False (Mostly False, False, and Miscaptioned), as illustrated by Table 4.2. The former

[1] `https://www.snopes.com/fact-check/category/photos/`

category includes cases where although part of the claim might be inaccurate, the usage of the image is still correct, whereas, the latter indicates that the usage of an image for a given claim is problematic.

I collect data from the Snopes fauxtography category from the very beginning of the site (early 2000s) to October 2019, obtaining 2,286 articles. These include 16K URLs to images, out of which I successfully download 7.8K (the rest of the URLs are no longer available).

4.4 Methodology

After building the dataset described in Section 4.3, I present the remaining part of my computational pipeline in this section, which consists of three components, as depicted in Figure 4·1: 1) pHash extraction, 2) image annotation, 3) image analysis.

4.4.1 pHash Extraction

Having collected all images as described in Section 4.3, the next step in this pipeline is to convert the images to a format that I can easily work with. To do so, I apply the Perceptual Hashing (pHash) algorithm (Monga and Evans, 2006) using the ImageHash library (Buchner, 2020), which generates a hash for each image in such a way that visually similar images have minor differences in their hashes. The algorithm is robust to image transformations (e.g., slight rotation, skew).

4.4.2 Image Annotation

Next, I annotate and identify the images that relate to fauxtography. To do this, I perform pairwise comparisons between the pHashes of images obtained from the various Web communities (including news articles) and images obtained from image annotation sites, such as Snopes. I calculate the Hamming distance between a pair of pHashes (i.e., an image from Snopes and an image shared on Twitter) and I assume

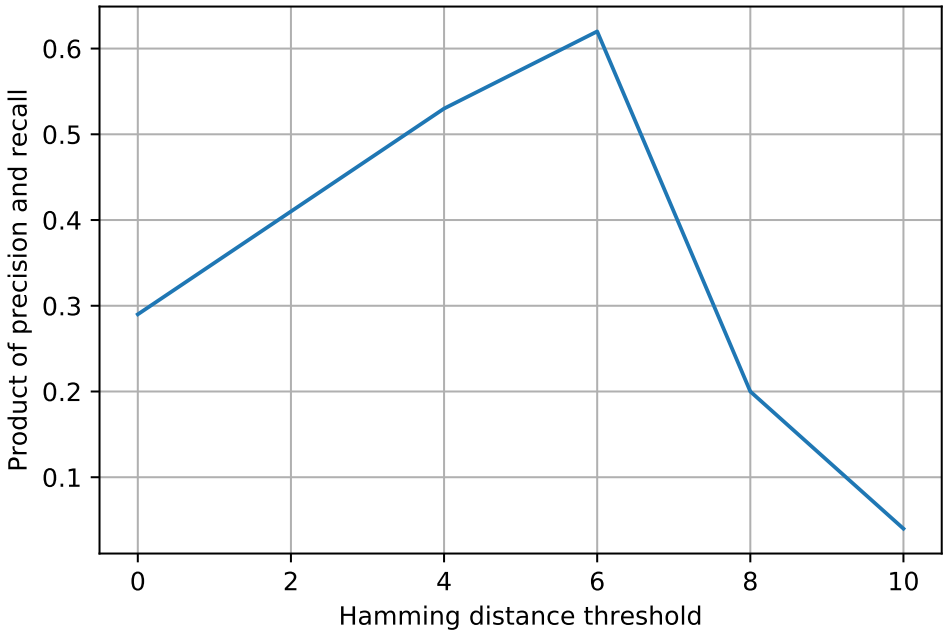


Figure 4·4: Product of precision and recall at different pHash Hamming distance thresholds in the image annotation process.

that an image is related to fauxtography if the Hamming distance is less than or equal to a pre-defined threshold, which I set below. Previous work (Zannettou et al., 2018b) shows that pHash is ineffective when dealing with images that are dominated by a single background color (e.g., screenshots on a white background), thus, I remove from this dataset images from annotation sites dominated by a single background color (i.e., screenshots, images of sky, etc.). Overall, this leaves is with 5,789 Snopes images for subsequent analysis.

Setting the pHash threshold. I consider two images to be visually similar if the Hamming distance is below a certain threshold. I vary the threshold from 0 to 10 and perform manual inspection of the matched images between the top images of news articles shared on all three Web communities and the corresponding Snopes images.[2] I consider a match to be correct if a human annotator considers them visually similar.

[2]Empirically, I find that any pair of images with Hamming distance above 10 consists of extremely dissimilar images.

For each value of the Hamming Distance, I calculate the product of precision and recall for all pairs. In total, I manually check 76,067 pairs of matched images.

The result of the pHash threshold selection process is shown in Figure 4·4: the maximum product of precision and recall is obtained at Hamming distance 6 (0.89 precision and 0.69 recall), hence, I use 6 as the threshold to determine if two images are similar. At this threshold, I find that 2,129 fauxtography images from Snopes have at least one match in posts on one of the social networks or in this dataset. In total, I find 45,567 tweets, 10,916 submissions and comments from Reddit, 2,987 posts from 4chan, and 1,633 news articles that include fauxtography.

4.5 RQ1: Impact on Engagement

To understand if including fauxtography in social media posts increases engagement, I first look at whether posts on Twitter containing fauxtography produce more retweets and likes than other posts. I next look at submissions on Reddit, where I use the scores that a submission receives and the length of threads as engagement metrics. Finally, I look at posts on Twitter and Reddit that do not include fauxtography directly, but that rather include links to news articles containing fauxtography. Note that I do not analyze the 4chan data here because the small number of data points makes statistical analysis unsuitable (301 threads fauxtography images in total for images that are shared directly, and 38 images for news articles containing fauxtography).

4.5.1 Twitter

As I use the Twitter streaming API for data collection, my data contains real time activity, i.e., tweets are gathered as soon as they are posted. This makes the dataset less than ideal to assess the engagement received by tweets, because the number of retweets and likes reported by the API represents short-term behavior. To gain a

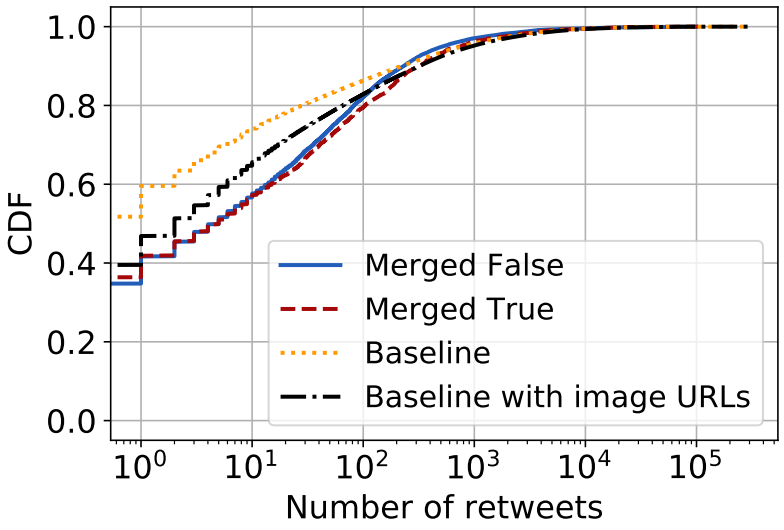


Figure 4·5: CDF of number of retweets on tweets sharing directly images.

better view of the long-term engagement, I leverage a process called *hydration*[3]: given a tweet ID, I retrieve the latest version of the current number of retweets and likes for that tweet. I hydrate the tweets in this dataset between June and July 2020.

Tweets can be classified as original tweets, retweets, and quote tweets. After hydration, I find that I cannot retrieve the actual retweet and likes count of regular retweets.[4] Therefore, to assess engagement for retweets, I retrieve the latest version of the original tweet that generated the retweet.

As discussed earlier, Snopes provides detailed labels to characterize fauxtography. For my experiments, I combine similar ratings together and form a binary system with two classes, Merged True and Merged False (see Table 4.2).

To understand whether posts containing fauxtography produce more engagement

[3]`https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/get-statuses-lookup`

[4]The "retweet_count" field in the metadata of the retweet, representing how many times a tweet is retweeted, is always equal to the "retweet_count" field in the metadata of the corresponding original tweet. In addition, the field "favorite_count," i.e., how many times a tweet is liked, is always 0 in the metadata of a retweet even if users press "like" on the retweet instead than on the original tweet, and the "favorite_count" of the original tweet is increased instead.

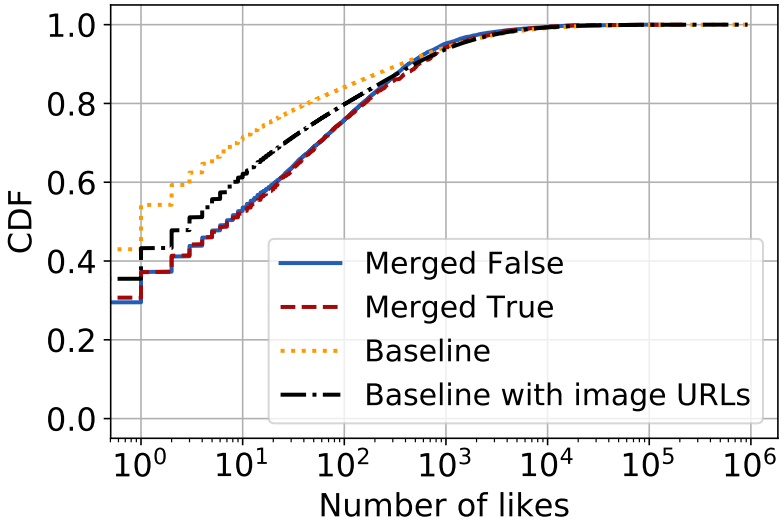


Figure 4·6: CDF of number of likes on tweets sharing directly images.

on Twitter, I extract two baselines: a set of random tweets and a set of tweets containing images that are not labeled as fauxtography. I then compare the engagement distribution of these tweets to posts containing fauxtography. I identify 9,858 Twitter users that shared tweets containing Merged True and Merged False fauxtography. I collect 9,771 tweets containing fauxtography rated as Merged False, 2,183 tweets containing fauxtography rated as Merged True. Then, I construct two baselines deriving from all the tweets shared from these 9,858 Twitter users: 1) 1,720,197 tweets that do not include images; and 2) 782,391 tweets that include non-fauxtography images.

Figures 4·5 and 4·6 show the cumulative distribution functions (CDFs) of the retweets and likes received by the four types of tweets, respectively. I observe that tweets containing an image from my fauxtography dataset (whether true or false) are more likely to produce more retweets and likes than my baseline tweets: 42% tweets containing Merged False fauxtography and 43% tweets containing Merged True fauxtography have been retweeted more than 10 time, while only 26% tweets from the generic baseline of random tweets have been retweeted more than 10 times. Similarly,

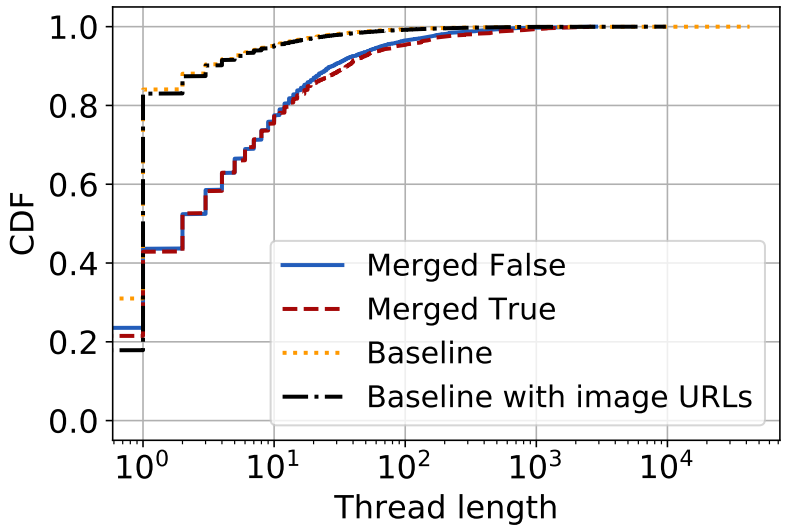


Figure 4·7: CDF of Reddit submission thread length on submissions sharing directly images.

46% tweets containing Merged False fauxtography and 47% tweets containing Merged True fauxtography have been liked more than 10 times, while only 29% tweets from the generic baseline have been liked more than 10 times.

To assess differences between these distributions, I run two sample Kolmogorov-Smirnov tests (K-S test) (Lindgren, 1993). I first compare to the baseline set of random tweets. I find that the differences between the following distributions are statistically significant at the $p < 0.05$ level: Merged False tweets compared to the baseline (D=0.182), and Merged True tweets compared to the baseline (D=0.185) when examining retweets. As for likes, I have statistically significant differences between the distribution of Merged False tweets compared to the baseline (D=0.187), and for Merged True tweets compared to the baseline (D=0.192). In all cases, $p \ll 0.001$ I thus reject the null hypothesis that tweets with fauxtography images receive the same level of engagement as random tweets.

There is reason to believe that tweets containing images get more engagement overall (Li and Xie, 2020). To lend further evidence to the observation that my images

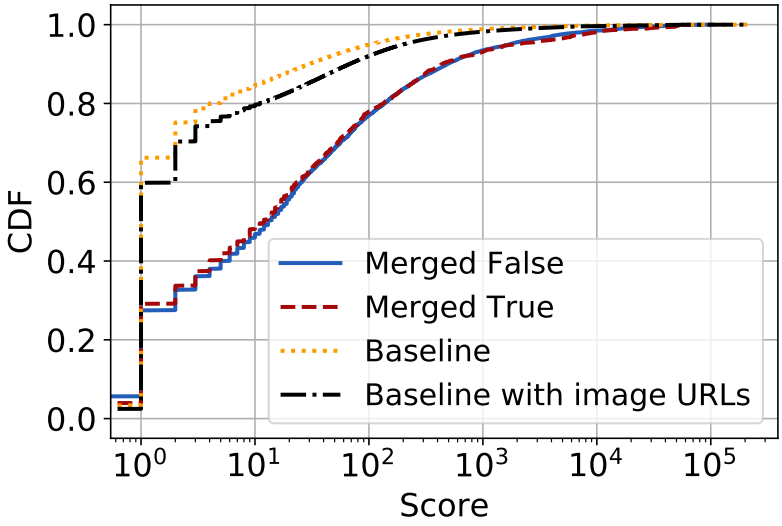


Figure 4·8: CDF of Reddit submission score on submissions sharing directly images.

in my fauxtography dataset are likely to receive more engagement than random images, I next compare the fauxtography distributions to the baseline of tweets with images in Figures 4·5 and 4·6. Again, I observe that tweets with images from my fauxtography dataset are more likely to be retweeted and liked than those with other images: only 34% of tweets containing non-fauxtography image have been retweeted more than 10 times, and only 38% have been liked more than 10 times. Using 2-sample K-S tests, I reject the null hypothesis that tweets containing non-fauxtography images and those with fauxtography have the same probability of receiving engagement ($p \ll 0.001$ in all cases). For retweets, I have D=0.0811 for Merged False tweets compared to non-fauxtography image baseline, and D=0.0896 for Merged True tweets compared to non-fauxtography image baseline. For likes I have D=0.0854 for Merged False tweets compared to the non-fauxtography image baseline, and D=0.0968 for Merged True tweets compared to the non-fauxtography image baseline.

Finally, a question remains as to whether or not the verisimilitude of a fauxtography image affects its engagement. I compare the distribution of engagement between tweets

with Merged True and Merged False images. In this case, I reject the null hypothesis that there is a difference with respect to retweets (D=0.0380, $p = 0.011$), however I am *unable* to reject the null hypothesis of differences with respect to likes (D=0.0219, $p = 0.36$). One explanation for this result is that images in these fauxtography dataset are usually quite controversial, with a sensationalist tone. I speculate that this tends to drive engagement, regardless of the underlying verisimilitudeof the image itself.

4.5.2 Reddit

For Reddit, I run analogous experiments using the length of a thread and the score of a submission as engagement metrics. Reddit calculates the score of a post as the difference between the number of upvotes and downvotes that it receives. On Reddit, the initial post in a thread is the "submission," and other posts in that thread are "comments." The length of a thread is obtained from the "num_comments" metadata field, and the score (i.e., the number of upvotes minus the number of downvotes) is obtained from the "score" field in submission metadata. Note that the "score" field is a precise value[5] while upvote and downvote values are fuzzed.

First, I identify 4,883 users that shared submissions containing fauxtography. These users shared 5,444 submissions containing Merged False fauxtography and 1,522 submissions containing Merged True fauxtography, respectively. Then, I construct two baselines based on the same set of Reddit users: 1) 7,248,595 submissions that do not include images; and 2) 3,367,222 submissions that include non-fauxtography images.

Figures 4·7 and 4·8 plot the CDF of thread length and score (respectively) for each of the four sets of submissions just described. From the plots, I note that 23% of submissions containing Merged False or Merged True fauxtography resulted in threads with more than 10 comments, while this is true for only 4.6% of non-image

[5] `https://www.reddit.com/wiki/faq\#wiki_how_is_a_submission.27s_score_determined.3F`

submissions and 4.8% submissions containing non-fauxtography images. Similarly, 53% submissions containing Merged False fauxtography and 53% submissions containing Merged True fauxtography have scores higher than 10, but only 15% of generic non-image submissions and 20% submissions containing non-fauxtography images have a score above 10. This suggests that fauxtography images produce more engagement than the baseline, regardless of whether the random post contains an image or not.

The differences in these distributions are again statistically significant as confirmed via 2-sample K-S tests. For the length of threads, I have D=0.404 for Merged False submissions compared to the no-image baseline, and D=0.411 for Merged True submissions compared to the no-image baseline. For likes, I have D=0.426 for Merged False submissions compared to the no-image baseline, and D=0.414 for Merged True submissions compared to the no-image baseline. When comparing to the non-fauxtography image baseline, I have D=0.394 for Merged False submissions and D=0.401 for Merged True submissions when looking at the length of threads. For likes, I have D=0.381 for Merged False submissions compared to the non-fauxtography image baseline, and D=0.368 for Merged True submissions compared to the non-fauxtography image baseline. In all cases, I find $p \ll 0.001$

Similar to Twitter, I am *unable* to reject the null hypothesis that there is no difference in engagement between true and false fauxtography images on Reddit. In the case of Reddit it is important to note that I am unable to reject the null hypothesis for both types of engagement; $D = 0.0220$, $p = 0.61$ for thread length and $D = 0.0225$, $p = 0.51$ for submission score. Again, this suggests that the engagement generated by fauxtography images is independent of the verisimilitude of the image.

4.5.3 News URLs

I now look at the engagement generated by posts that have links to news articles that include fauxtography rather than directly including fauxtography. On Twitter, I

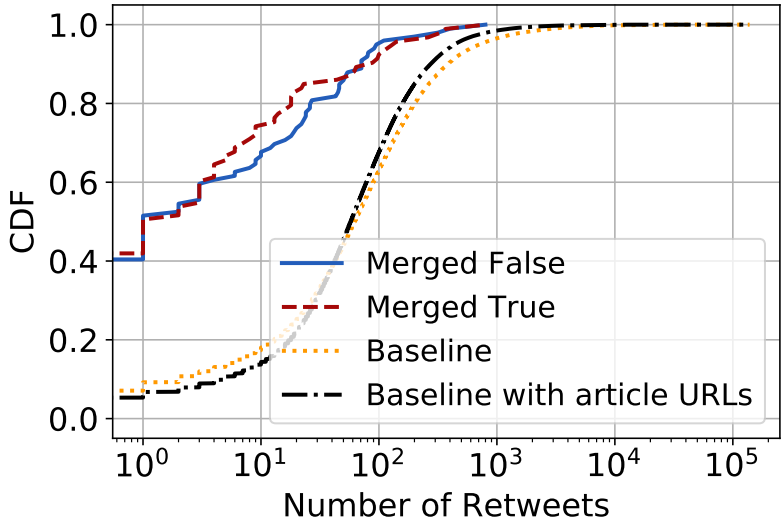


Figure 4·9: CDf of number of retweets on tweets sharing news articles.

identify 100 tweets with links to news articles that contain Merged False fauxtography images and 94 tweets with links to articles that contain Merged True fauxtography images. On Reddit, I identify 431 submissions with links to articles that contain Merged False fauxtography and 272 submissions with links to articles that contain Merged True fauxtography.

Once again, I compare the engagement of posts containing links to news articles containing fauxtography to a generic baseline of 492,604 tweets on Twitter and 19,704,911 Reddit submissions, respectively, and to a baseline of 239,079 tweets and 9,554,016 posts containing generic news URLs. The baselines are constructed by collecting all non-fauxtography posts posted by the users who made at least one fauxtography related submission on Twitter or Reddit. Figures 4·9 and 4·10 show the retweets and likes of tweets containing fauxtography news URLs. Contrary to what observed previously, these tweets do not receive more engagement than baselines. More precisely, on Twitter, 32% tweets containing Merged False fauxtography and 26% tweets containing Merged True fauxtography have been retweeted more than 10 times, while 82% generic tweets and 85% tweets containing non-fauxtography news

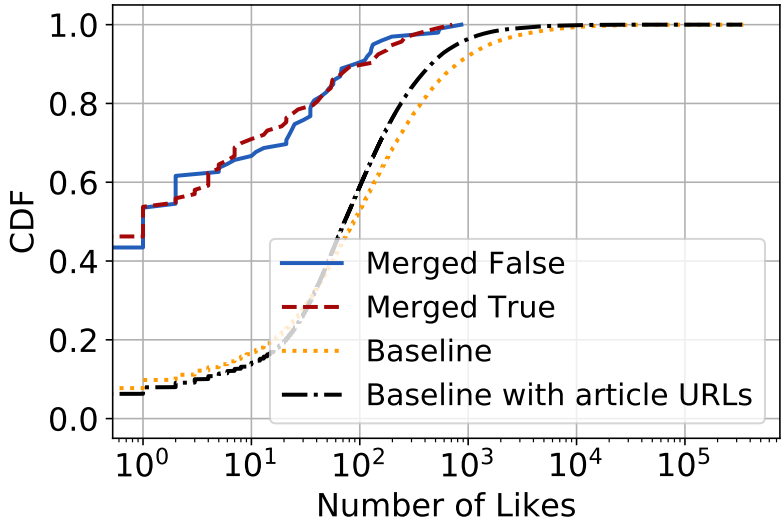


Figure 4·10: CDf of number of likes on tweets sharing news articles.

URLs been retweeted more than 10 times. Furthermore, only 33% tweets containing fauxtography rated as Merged False and 29% tweets containining fauxtography rated as Merged True have been liked more than 10 times, while 83% of generic tweets and 86% tweets contain generic non-fauxtography news URLs have been liked more than 10 times.

On Reddit, on the other hand, I find that posts containing links to fauxtography news articles still receive more engagement. As Figures 4·11 and 4·12 show, 16% of submissions containing Merged False fauxtography and 14% of submissions containing Merged True fauxtography have thread lengths longer than 10, while the same is true only for 1.3% of generic submissions and 1.6% non-fauxtography news URL submissions.

On Twitter, I confirm differences in these distributions via the 2-sample K-S test for fauxtography submissions compared to baseline submissions, where for the number of retweets I have D=0.506 for Merged False tweets compared to the non-fauxtography baseline, and D=0.584,for Merged True tweets compared to the generic baseline. For likes, I have D=0.509 for Merged False tweets compared to the generic baseline, and

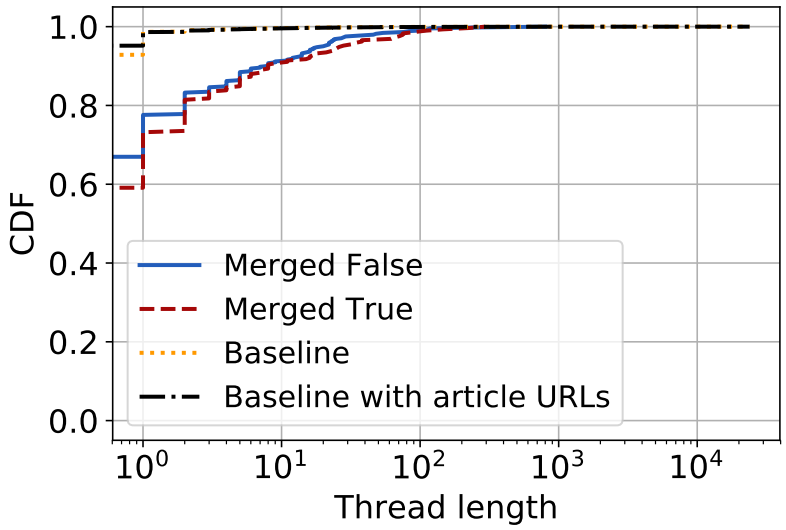


Figure 4·11: CDF of Reddit submission thread length on submissions sharing news articles.

D=0.545 for Merged True tweets compared to the generic baseline. Looking at the non-fauxtography news article baseline, I have D=0.536 for Merged False tweets and D=0.614, for Merged True tweets when looking at retweets. For likes, I have D=0.534 for Merged False tweets, and D=0.571 for Merged True tweets. In all cases, $p \ll 0.001$ leading me to reject the null hypothesis that there are no differences between these distributions.

On Reddit, when looking at the length of threads I have D=0.275 for Merged False submissions compared to the generic baseline, and D=0.358 for Merged True submissions compared to the generic baseline. For Scores, I have D=0.260 for Merged False submissions compared to the generic baseline, and D=0.339 for Merged True submissions compared to the generic baseline. When looking at the non-fauxtography news article baseline, I have D=0.271 for Merged False submissions and D=0.354 for Merged True submissions for the length of threads. For Scores, I have D=0.282 for Merged False submissions compared to the non-fauxtography image baseline, and D=0.362 for Merged True submissions compared to the non-fauxtography image

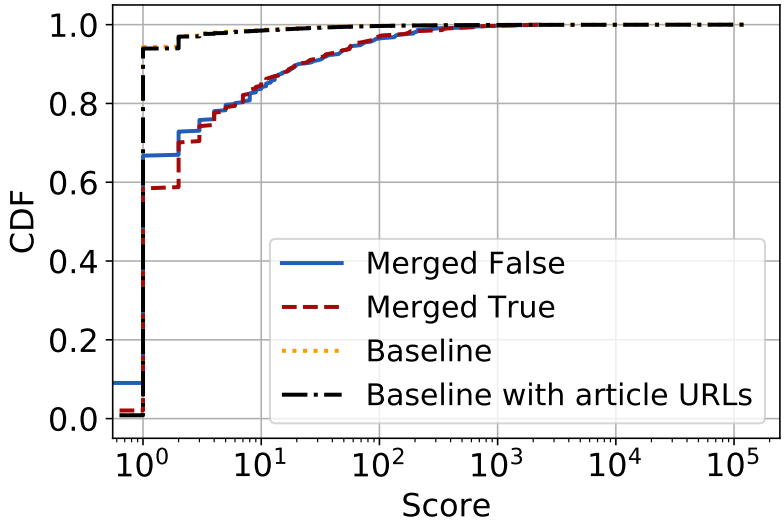


Figure 4·12: CDF of Reddit submission score on submissions sharing news articles.

baseline. In all cases, $p \ll 0.001$ leading me to reject the null hypothesis that there are no differences between these distributions.

Note, however, I am unable to reject the null hypothesis that there are differences in engagement between Merged True and Merged False tweets and submissions. On Twitter, a KS test gives me D=0.0998 ($p = 0.7$) for retweets and D=0.0562 ($p \approx 1.0$) for likes. On Reddit, I obtain D=0.0826 ($p = 0.17$) for the length of threads and D=0.0791 ($p = 0.21$) for scores.

While most of the results for this experiment are consistent with what I previously found with regards to directly sharing fauxtography or images fact-checked as true by Snopes, interestingly, tweets containing links to news articles with fauxtography attract less engagement than other news links. One possible reason is that when sharing news URLs, many confounding factors can come into play with regards to enticing users into interacting with the tweet, for example clickbait titles and the content of the article. For Reddit, the results show that using news articles to share fauxtography can increase engagement, which is consistent with the results found

sharing images directly.

4.5.4 Takeaways

My analysis provides evidence that posts directly containing fauxtography images or images fact-checked as true by Snopes do indeed generate higher engagement on both Twitter and Reddit. However, when it comes to sharing links to news articles that make use of fauxtography or images fact-checked as true by Snopes, I find that they generate significantly *less* engagement on Twitter but significantly *more* engagement on Reddit. Further, except in the case of retweets on Twitter, I am unable to reject the null hypothesis that posts containing images or links to news stories using fauxtography or images fact-checked as true by Snopes receive the same levels of engagement. Twitter users seem more resistant to engaging with links to news stories that use fauxtography or images fact-checked as true by Snopes, but more likely to engage with tweets containing images themselves. Reddit users were more likely to engage with any fauxtography or images fact-checked as true by Snopes related content.

These differences pose interesting problems for social media platforms; for example, fact-checking efforts that focus on links to news articles (some of which have been implemented by Twitter are likely to have little effect on the spread of fauxtography in general as the images themselves still achieve relatively high engagement.

4.6 RQ2: Fauxtography's Evolutionary Nature

Previous work has indicated that memes exhibit some evolutionary properties, with new variants frequently emerging. Since, by definition, my fauxtography dataset includes images that have spread wide enough to warrant fact checking, I posit that some might have found life beyond fauxtography. Thus I ask: do fauxtography images become memes with different variants?

Platform	#CommonFalse-NoRating Images	#FalseImages with variation	#FalseImages w/o any variation	#FalseImages w/o variation-RandomImages	#FalseImages with variation-SameImage
Twitter	162	86	76	1291	70
Reddit	145	70	75	625	63
4chan	58	25	33	269	117

Table 4.3: Statistics for false images variations in Twitter, Reddit, and 4chan.

To answer this question, I relax the distance threshold used to detect instances of fauxtography images from 6 to 8 and examine the resulting images matches. I further focus on fauxtography images that were labeled only False, to focus on the role of fauxtography in spreading false information. I find 238 source images labeled "False" from my Snopes dataset that appear at least once on all Twitter, Reddit, and 4chan. I note that although measuring engagement on 4chan is problematic enough that I do not include details in Section 4.5 , 4chan is a key player in the meme ecosystem; thus I include it in this analysis.

For each source image, I manually determine whether each image within distance 8 is a variant. Table 4.3 provides details on the number of instances of variants across each platform. I observe that, of the 238 source images appearing on all three platforms, there were an additional 162 images on Twitter within distance 8 that I confirmed were indeed a match for a source image. Of these 162, 86 were sufficiently different from the source image to be deemed a variant, while 76 were essentially the same as the source image (i.e., they can be considered false negatives due to my threshold selection). An additional 1,291 images with distance 8 were completely unrelated (i.e., true negatives). For each of the three platforms, I see relatively similar numbers.

4.6.1 Case Studies

I find that 13 source images have variants that appear at least once on all three platforms I study (although not necessarily the same variant). A manual inspection

(a) Al Franken source image

(b) Al Franken Twitter variant

(c) Al Franken Reddit variant

(d) Al Franken 4chan variant

Figure 4·13: Variations of common Fauxtography images relevant to Al Franken on all three platforms.

shows that variants of these 13 source images correspond to memes. I examine three representative and particularly well-known cases in Figure 4·13, 4·14 and 4·15. My intuition is that particularly powerful fauxtography images are likely to take on a life of their own and become memes.

The first source image (Figure 4·13(a)) is a picture of Al Franken inappropriately touching Leeann Tweeden's breasts while she slept. The image is real, and was taken in 2006 on a C-17 cargo plane on their return from a USO tour in Afghanistan. This source image played a crucial role in then US Senator Al Franken's retirement from politics. The image was particularly controversial due it coming to light at the height of the #MeToo movement (Garber, 2017) as well as claims that it was related to a sketch that had been performed on the USO tour. The image is labeled as a false

(a) Trump Marine 1 original

(b) Marine 1 Twitter variant

(c) Marine 1 Reddit variant

(d) Marine 1 4chan variant

Figure 4·14: Variations of common Fauxtography images relevant to Trump on all three platforms.

instance of fauxtography due to a widely circulated claim that the photographer that took the picture said it was staged. However, Franken fully admitted to the picture to be be real and not staged, accepted responsibility for what was ultimately irresponsible behavior, and resigned.

The variant of this image on Twitter (Figure 4·13(b)) replaces Franken's face with that of Roy Moore, an Alabama political figure that lost a hotly contested race against

(a) GWB reading book original

(b) GWB Twitter variant

(c) GWB Reddit variant

(d) GWB 4chan variant

Figure 4·15: Variations of common Fauxtography images relevant to George W. Bush on all three platforms.

Democrat Doug Jones for Jeff Sessions's US Senate seat after he was appointed US Attorney General. The text added to the image is related to allegations of sexual assault and pedophilia by Roy Moore, and Gloria Aldred's involvement in the incident. The variation on Reddit (Figure 4·13(c)) is much less political, merely replacing Franken's head with that of 80s sitcom character Alf.

On 4chan (Figure 4·13(d), the variant replaces Franken's face with with that of Donald Trump, replaces Leeann Tweeden's head with Stephen Colbert's head, and places two scoops of ice cream over Tweeden's breasts. This is likely in reference to Stephen Colbert's comments on sexual harassment (Van Hoozer and Peuchaud, 2020)

and Trump's alleged routine of receiving two scoops of ice cream for desert when everyone else at the table receives only one (Mercia, 2017) (e.g., Colbert's nickname for Trump, "Donnie Two Scoops").

The second source image Figure 4·14(a) shows a rear view of former President Donald Trump entering Marine One. Based on Snopes this image is a "slightly manipulated" image (Trump's posterior has been enhanced) originally taken by a Reuters' photographer while Trump boarded Marine One at Joint Base Andrews in Maryland. This photo was generally considered unflattering for Trump, as can be seen in the Twitter variant (Figure 4·14(b)), which uses Trump's buttocks to replace the two "Os" in Google's logo. This is indicative of some of the derision expressed online towards Trump's physical appearance. The Reddit variant (Figure 4·14(c)) introduces Vladimir Putin embracing Trump by "grabbing his butt." The 4chan variant (Figure 4·14(d)) is somewhat different, and replaces the saluting Marine guard with a saluting North Korean soldier with a rifle slung over his shoulder. The end of the rifle barrel is depicted as being inserted into Trump's buttocks.

The final source image I examine (Figure 4·15(a)) shows George W. Bush at a book reading at school in Houston in 2002. Snopes labels it as false because a manipulated version showing Bush holding the book upside down with a false caption was being spread on the Web. On Twitter (Figure 4·15(b)), I see a variant that has a non-manipulated version of the image, but has added text that implies Bush is telling the student about how right when the world needed it, "Q" (from the Qanon conspiracy) appeared to save us all. The variant that appears on Reddit (Figure 4·15(c)) is the manipulated variant where it appears Bush is holding the book upside down. Finally, I see a variant on 4chan (Figure 4·15(d)) that uses the manipulated version with the upside down book, and adds large arrows pointing to the stars behind the students along with the text "WTF!" It is not entirely sure what this variant is trying to

express, but based on my understanding of 4chan, I suspect it is conspiracy theory related.

4.6.2 Takeaways

Fauxtography is a complicated issue, in large part due to its visual nature and the Web's propensity for not just spreading visual information, but modifying it. The Franken image, which is not altered in any way and has a known provenance, is easily exploited for uses completely unrelated to its use in fauxtography. Similarly, the Bush image shows that even relatively innocuous pictures manipulated in subtle ways can become further manipulated to politicize them. The Trump image shows how even slight manipulations of real photos can elicit numerous meme variants.

This raises serious concerns about how to mitigate the relatively low-tech problem of fauxtography. For example, none of the variants I found were particularly convincing in terms of being real photos; the majority were very clearly manipulated, as is common for memes. What is there to fact check about a fictional TV alien groping a sleeping woman, after all? However, these variants tend to carry the same fundamental idea as the source image that *was* fact checked, and thus can still cause damage. Although issues like this warrant future exploration, at minimum, they calls into question the efficacy of fact checking *visual* mis/disinformation.

4.7 Discussion & Conclusion

In this chapter, I presented a data-driven study of fauxtography on social media. I found that including fauxtography or images fact-checked as true by Snopes in social media posts increases user engagement. This highlights the need to take images into account when developing disinformation mitigations. At the same time, I showed that fauxtography images are often taken out of context and turned into memes, which highlights the challenges faced in automatically identifying image-based disinformation.

Next, I discuss the implications of my findings and highlight some limitations of my study.

Implications of these findings. The fact that sharing fauxtography on social media increases user engagement highlights how image-based disinformation cannot be overlooked, and that any effort to curb the problem should take not only text into account, but also images. At the same time, I showed that fauxtography images are often used as memes on social media, blurring the line between the intention to mislead and satire. This opens up a number of problems when moderating fauxtography, since it is challenging to automatically determine the intention with which an image is posted, which is often context specific. Crucially, my study also highlighted the fact that the verisimilitude of fauxtography images does not have an impact on the engagement that they receive. This suggests that the "clickbait" power of these images is what drives engagement, and raises questions on the effectiveness of mitigations based on fact-checking labels and user warnings.

Limitations. Naturally, this study is not without limitations. First, my image analysis pipeline allows me to identify images that are very similar to fauxtography images, but is unable to verify if the image is used in the misleading setting flagged by Snopes. For example, I am unable to tell if miscaptioned images are being used in a miscaptioned context. Similarly, for manipulated photos, it is possible that my analysis pipeline identifies the unmodified picture as a fauxtography one. This motivates future work combining my analysis pipeline with semantic analysis techniques to study the context in which fauxtography is used. Third, the identification of news outlets using the top 30K Majestic websites excludes many small local news outlets. Since I expect that local news outlets have less fastidious fact-checking as compared to larger venues, this suggests my analysis will tend to underestimate the spread of fauxtography on the Web. Additionally, Snopes select images that are generally most cared by their

readers[6], which may also cause the potential bias. It may partially explain the reason why fauxtography images or images fact-checked as true by Snopes are more popular than the baseline images, but there may also still exist other factors. For instance, some fauxtography images are created to mislead readers about the impression of specific political figures, and these images get promoted by organized operations. It requires further research to determine the reasons that fauxtography images are more popular than the baseline images.

Finally, collecting images at scale from the Web present challenges. In particular, I found that many images were no longer available when I attempted to download them. Still, I believe that the scale of my dataset is large enough to allow us to gain a comprehensive view of the use of fauxtography on social networks.

[6] https://www.snopes.com/faq/decide-fact-check/

Chapter 5

COVID-19 Image misinformation

5.1 Introduction

People who spend time online are constantly bombarded by a deluge of information, consisting not only of text, but also of visuals like images, GIFs, and memes. With limited time, expertise, and investigative means, people usually have to take this information at face value and cannot reliably determine if it is true or not. The COVID-19 pandemic has exacerbated this problem, with a lack of knowledge about the virus allowing misinformation to spread in the early stage of the pandemic (Miller, 2020; Pennycook and Rand, 2021).

A wealth of research has been conducted in the past three years to better understand the dynamics of COVID-19 related misinformation and its effect on our society and on public health measures (Chen et al., 2020a; Ferrara, 2020; Micallef et al., 2020; Ziems et al., 2020; Tahmasbi et al., 2021; Tasnim et al., 2020). Most of this research has focused on textual content shared on social media; misinformation, however, is not solely composed of text but also of visuals. Images are more immediate than text, and can convey more complex messages than what can be contained in short social media posts (e.g., tweets) (Ling et al., 2021). As a result, COVID-19 related image misinformation is particularly dangerous, because it can become viral and severely impact our society, for example by encouraging people not to protect themselves properly or promoting false cures. Despite the dangers posed by image-based COVID-19 misinformation, the research community has spent limited efforts to understand

the problem, by either only analyzing images that appeared in news articles and were fact-checked (Brennen et al., 2021) or by focusing on false information spread within a single country (Javed et al., 2020). In (Lee et al., 2021), researchers discuss how COVID-19 case illustrations are used by COVID-19 skeptics to support their own beliefs.

In this chapter, I aim to shed light on how images are used to spread COVID-19 misinformation on Twitter. To this end, I collect 2.3M COVID-19 related tweets posted between March 1, 2020 to June 16, 2020. I then download 340K images included in those tweets. To facilitate manual analysis of these messages, I build a computational pipeline based on perceptual hashing techniques and clustering algorithms to group visually similar images together. I then develop a codebook to characterize COVID-19 misinformation images, identify five different types of COVID-19 misinformation images, and build a dataset of over 2.8K COVID-19 misinformation images posted on Twitter. I then perform a quantitative analysis on the tweets that contain COVID-19 misinformation images to answer the following research questions:

- RQ1: Do COVID-19 misinformation images generate more user engagement?
- RQ2: What are the temporal properties of COVID-19 misinformation images? Do COVID-19 misinformation images have a longer lifespan and longer burst times than non-misinformation images?
- RQ3: What are the characteristics of users who post COVID-19 misinformation images?

For RQ1, I compare the reactions (retweets and likes) to tweets containing COVID-19 misinformation images with baseline tweets, as well as baseline tweets containing random images posted by the same set of users. I find that tweets containing COVID-19 misinformation images do not receive significantly more engagement on Twitter.

For RQ2, I compare the lifespan of COVID-19 misinformation images on Twitter with that of non-misinformation images, finding that tweets containing COVID-19 misinformation images are shared for longer periods of time, and they also tend to have longer burst times.

For RQ3, I apply a mixed approach to characterize the users who post COVID-19 misinformation images. I find that these users are quite diverse and from all over the world. Additionally, I find that a large portion of the US users in this dataset supports either the Republican Party or Democratic Party, and I find that users who support the Democratic and the Republican parties post a similar amount of tweets with misleading or false COVID-19 images. At a first glance, this is in contrast with previous work. For example, Lazer et al. (Lazer et al., 2021) shows that registered Republicans are far more likely to share COVID-19 misinformation by citing URLs from fake news outlets than registered Democrats during the pandemic. This analysis does however find that the type of COVID-19 misinformation images shared by supporters of the two parties is different. While pro-Republican users often promote COVID-19 conspiracy theories about the origin of the virus and advocate for the use of hydroxychloroquine to treat COVID-19, pro-Democrat users share false or misleading claims surrounding the response to the pandemic adopted by the Trump administration, as well as manipulated or forged images intended as satire.

These results shed light on how images are used to spread COVID-19 misinformation on Twitter. Most interestingly RQ1 contradicts what was found by previous research on misinformation, which found that tweets containing false information receive more engagement (Vosoughi et al., 2018; Wang et al., 2021). A potential reason is that past research followed a top-down approach, only looking for false stories that had been fact-checked, while my approach is bottom-up, identifying groups of misinformation images as they are posted online. I argue that more discussion is needed within

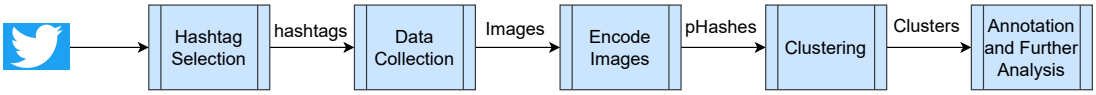


Figure 5·1: Overview of my computational analysis pipeline.

the misinformation research community to better understand the advantages and disadvantages of different data approaches, and the biases that these choices might introduce in research results.

5.2 Dataset

A summary of my research pipeline is shown in Figure 5·1. In this section, I describe how I build my dataset, which is the first step of this pipeline and I will describe the other steps in Section 5.3

I need to collect Twitter data related to the COVID-19 pandemic and download the images contained in those tweets. To this end, I follow the same snowball sampling approach conducted by past research on hateful tweets (Chatzakou et al., 2017). For my analysis, I leverage data from the public 1% Twitter Streaming API (Kwak et al., 2010; Chen et al., 2022a; Pfeffer et al., 2018).

Hashtag selection. First, I collect all public tweets returned by the API during the month of March 2020. I then aim to identify popular hashtags that are included in COVID-19 related tweets. To this end, I start by extracting all tweets that contain three hashtags: "COVID," "coronavirus," and "COVID19." I then proceed by identifying other hashtags that co-occur with these three, similarly to what was done by (Chatzakou et al., 2017). I finally select the 100 most popular co-occurring hashtags; adding these to my initial set, I have 103 total hashtags. The full list of hashtags has been shared anonymously at the following link.[1]

Data collection. By using the Twitter streaming API (which provides a 1% sample

[1] `https://bit.ly/2YTJvh9`

of all public tweets), I obtain a total of 505,346,347 tweets between March 1, 2020 and June 16, 2020. Then by using the 103 identified popular COVID-19 related hashtags in the hashtags selection step, I extract all tweets containing any of these hashtags from these 505M tweets. This gives me a total of 2,335,844 COVID-19 related tweets. Of these, 370,465 tweets contain image URLs, of which I am able to successfully download 339,952 images, which are shared by 339,891 tweets in June 2020. The tweets not included in the COVID-19 related dataset will also be used in RQ1 to establish baselines.

Note that I collect my own data instead of using existing datasets like the one compiled by Chen et al. (Chen et al., 2020a) because for this analysis I need to compare tweets with images containing COVID-19 misinformation with baseline tweets posted by the same users sharing COVID-19 misinformation image tweets, to avoid bias generated by considering tweets from users with varying numbers of followers (Wang et al., 2021; Zannettou, 2021). The baseline obtained from the Twitter Streaming API is more general since it contains tweets related to COVID-19, as well as tweets unrelated to COVID-19, which are missing in existing COVID-19 Twitter datasets (Chen et al., 2020a). Therefore, I rely on data collected from the Twitter Streaming API instead. Please see Section 5.4.1 for more details about how I build the baselines.

5.3 Methodology

After I built the dataset described in Section 5.2, I follow a mixed-method approach to annotate data for this research. As shown in Figure 5·1, I first apply perceptual hashing to all the images in the dataset and then clustering to group together all images that look similar. Next, to identify images that contain misinformation I develop a codebook to facilitate the labeling of COVID-19 misinformation images and their categorization. Finally, I analyze the identified COVID-19 misinformation

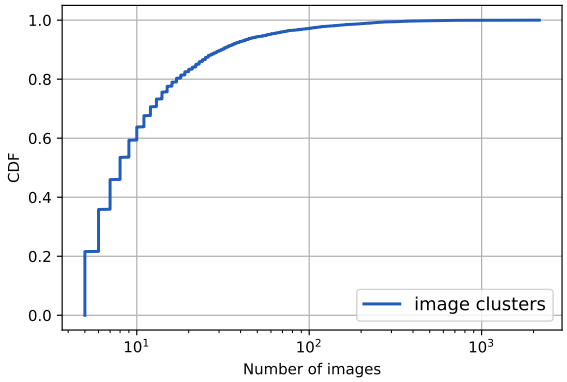


Figure 5·2: The cumulative distribution function (CDF) of image cluster sizes.

images to answer my research questions. In the following, I discuss each phase of these steps in detail.

5.3.1 Grouping visually similar images

Images on social media do not appear in isolation, but rather are often re-shared and modified, becoming memes (Du et al., 2020). Therefore, to analyze how images are used in online discourse, I need techniques to group together not only occurrences of the same identical image, but also of visually similar images which might be variations of the same meme. This also reduces the workload required for human annotators when labeling images as misinformation, as I will discuss in Section 5.3.2. To identify and group together visually similar images I use the method used by (Garimella and Eckles, 2020) and (Zannettou et al., 2018b). I first calculate perceptual hashes (pHashes) for the images in my dataset, and then use clustering to group similar images together. In the rest of this section, I describe these two steps in detail.

Encode images. To find images that are visually similar I apply perceptual hashing to encode them. This technique returns binary vectors (pHashes) of 64 bits that are numerically close if the images are visually similar (Monga and Evans, 2006). This allows me to find not only images that are identical, but also minor variations or

meme derivatives of an image. I use the version of perceptual hashing provided by the ImageHash library (Buchner, 2020), which previous work showed is robust to image transformations (e.g., slight rotation, skew) (Zannettou et al., 2018b).

Image clustering. After obtaining the pHash values for all images in this dataset, I apply clustering to group together similar images. To this end, I use the DBSCAN clustering technique (Ester et al., 1996), which was used for similar purposes in previous work (Zannettou et al., 2018b). I learn from the experience of parameter selection from (Zannettou et al., 2018b), setting the Hamming distance difference threshold for pHash values of two images to be considered visually similar to 6 and the minimum number of elements in a cluster to 5. The detailed process followed to select these thresholds is described in Appendix A. My implementation of DBSCAN is based on (Pedregosa et al., 2011).

As I described above, I obtain 339,952 images in total. Of these, 78,348 are obtained from original tweets, 261,604 are from retweets. Note that at this stage of my analysis I do not distinguish between images posted as original tweets and retweets. All images are fed into the DBSCAN clustering algorithm.

After clustering images, I group 148,987 images into 7,773 clusters. The cumulative distribution function (CDF) of the size of these clusters is shown in Figure 5·2. The median size of these clusters is 8, but there is a long tail of clusters that are much larger, with 10% of the clusters containing 31 or more images. In the subsequent annotation, I will focus on the images contained in clusters only. The reason is that I am interested in understanding how misinformation images are re-shared on Twitter, and if images did not cluster together it is safe to assume that they did not become popular on Twitter. In fact, considering that the data is a uniform 1% sample of all public tweets and that the minimum cluster size is 5, we can assume that images that do not form clusters in my dataset were likely shared publicly on Twitter less

Figure 5·3: Little or indistinguishable visual dissimilarity of two images with distinct pHash values in my dataset.

than 500 times at the time when the data was collected. In Appendix B I perform an additional analysis showing that tweets containing images that did not get clustered by my approach attracted significantly fewer retweets and likes than those that formed clusters, confirming this intuition. Other work in this area also focused on subsets of all images when performing annotation, by either only annotating the most shared images (Resende et al., 2019) or by taking a sample of images for annotation (Reis et al., 2020; Garimella and Eckles, 2020). Therefore, I believe that the selection criteria that I use to filter my data is appropriate for my purposes.

Evaluating image similarity within the same clusters. Before continuing with my annotation and analysis, it is of paramount importance to understand whether the clusters produced by my approach are of good quality. In other words, I want to assess whether images that are clustered together are either instance of the same original image or minor variations of it. To this end, I first look at how many clusters only contain identical images. I find that 6,128 out of 7,773 clusters only contain images with identical pHash values, which indicates all the images within the same cluster are visually identical (Resende et al., 2019; Zannettou et al., 2020a).

I then manually inspect the remaining 1,645 clusters that contain images with

Figure 5·4: Minor variations or meme derivatives of an image of two images with distinct pHash values in my dataset.

different pHash values. I find that these clusters fall within three categories:

- Clusters containing images that although having distinct pHash values, they appear identical to the human eye. This is due to very small differences in the images. One such example is shown in Figure 5·3.

- Clusters containing images that are derivatives (i.e., minor variations) of other images in the cluster. An example of two images falling in this category is shown in Figure 5·4. As it can be seen, the meaning of the images is similar (adopt simple precautions to protect yourself from COVID-19 and do not listen to politicians), but while one image is targeted at a US audience, the second one is targeted at a Mexican one.

- Clusters containing images that do not appear visually similar to the human eye, despite having close pHash values. This is due to limitations in the pHash algorithm. For example, images where the background of a certain color dominates might be mistakenly grouped together. I consider these as false positives of my clustering approach.

After inspecting all clusters, I find that 105 image clusters contain false positives. This translates into my approach having an accuracy of 98.6%, giving us confidence

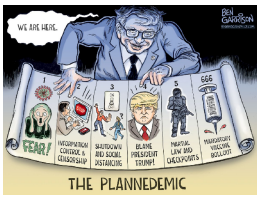
that my approach is reliable in grouping together similar images. Note that since the false positive clusters appear visually different to a human annotator, they are ignored in the later stages of my analyses to avoid biases in the results. In total, I have 7,668 clusters left, containing 146,192 images.

5.3.2 Identifying COVID-19 misinformation images

Because whether an image contains misinformation or not often depends on context, it is challenging to automatically label images as misinformation. To overcome this limitation, another PhD student and I manually annotate every image cluster in my dataset with the goal of building a credible ground-truth dataset (Cortis and Davis, 2021; Akyürek et al., 2020; Park et al., 2020).

In this section, we develop a codebook to guide the thematic annotation process for COVID-19 images on Twitter (Braun and Clarke, 2006). As mentioned previously, we use the definition of misinformation proposed by (Wu et al., 2019), which defines misinformation as "informative content that contains incorrect or inaccurate information." We divide the development of this codebook into two phases based on this definition. First, We use binary labeling to evaluate whether or not images are related to COVID-19. If so, then We call these images "*informative*." As a further step, We characterize the images that contain misinformation.

We use the following three steps to create a codebook and perform annotation: 1) My collaborator and I separately evaluate this dataset and provide preliminary codes based on thematic coding (Braun and Clarke, 2006). 2) We then discuss these preliminary codes and go through multiple rounds, utilizing a subset of the data to create a complete codebook. The procedure is repeated until the codebook reaches a point where future iterations would not improve it anymore. 3) We classify the remainder of our dataset and discuss differences until a satisfactory consensus is obtained.



(a) Conspiracy theories on Bill Gates.

(b) Fauxtography of a lion wandering the streets of Russia.

(c) Misinformation on using a malaria drug to treat COVID-19.

(d) Wrong understanding of the threat severity of COVID-19.

(e) Other false claims.

Figure 5·5: Types of COVID-19 misinformation images identified by my codebook.

I next describe my process and my codebook in more detail.

Phase I: Labeling informative images. As previously stated, the initial stage of our annotation process is concerned with selecting informative images, which are images related to COVID-19. We begin by selecting 1,000 clusters at random from our dataset of 7,668 clusters. My collaborator and I review and discuss every image in these clusters to develop a shared understanding of what an informative image looks like. We agree on the following criteria for an informative image based on this initial dataset:

- The image has no words or contain words in English or Chinese. We focus on these two languages because these are the two languages spoken by the researchers that annotated the dataset.

- The image must contain sufficient visual cues or words connected to the COVID-19 pandemic, such as RNA virus, public figures during the pandemic, and medical elements such as physicians, hospitals, face masks, and so on.

As long as one image in a cluster is informative, then we label this image cluster as informative. This is reasonable, because as we showed in Section 5.3.1, the accuracy of my clustering approach is high, and those clusters that did not produce good results were manually removed before proceeding to this annotation. Note the only goal of determining "informative" images is to filter out a smaller set of image candidates for us to manually identify COVID-19 misinformation images.

After my collaborator and I independently label the 1,000 image clusters as either informative or non-informative by checking every image in these image clusters, we calculate the Cohen's Kappa between the annotators and find perfect agreement ($\kappa = 0.991$) (McHugh, 2012). This shows that two annotators strongly agree with each other, verifying the codebook's validity. After establishing that the codebook is mutually agreed by the annotators, the rest of the images in this collection are labeled by me by checking every image in these image clusters. Finally, out of 7,668 clusters, we identify 2,316 informative clusters containing 39,367 images.

Note that if the text in a tweet is related to COVID-19 this does not necessarily imply that the images included in this tweet are also related to COVID-19. For example, a tweet discussing the pandemic in the text might include a generic image like a reaction GIF.

Phase II: Characterizing COVID-19 misinformation images. While identifying images as informative is important for comprehending the problem, not all informative images contain misinformation. In fact, a wealth of good information is posted on social media to urge individuals to take responsibility and take measures for halting the pandemic. In this phase, we want to identify the features of COVID-19

Type	#Cluster	#Images
Conspiracies on COVID-19.	91	1,403
Wrong understanding of the threat severity of COVID-19.	10	294
Fauxtography.	27	455
Wrong medical advice.	41	605
Other false claims.	23	478

Table 5.1: Overview of cluster numbers and number of images for each of the misinformation types.

misinformation images, by analyzing image themes. We start by having both annotators look over the labeled informative images from Phase I. We follow a loose approach with the objective of getting a basic idea of the themes present in the dataset. We then convene to discuss our findings.

Eventually, the annotators identify five types of misinformation, the specifics of which are presented below.

1. *Conspiracies on COVID-19.* Conspiracy theories, particularly those involving science, medicine, and health-related issues, are common since long before the COVID-19 pandemic (Oliver and Wood, 2014). A large body of research has demonstrated that conspiracy theories can cause individuals to reject information from competent authorities, raising worries about the potential for popular conspiracy theories to diminish people's willingness to follow public health advice (Uscinski et al., 2020; Freeman et al., 2020; Banai et al., 2020). In this work we define conspiracy images as visuals that provide a theory that explains an occurrence or set of circumstances as the product of a hidden scheme by generally strong conspirators (van Prooijen and Douglas, 2018). Figure 5·5(a) shows an example of this type of misinformation.

2. *Fauxtography.* Previous research indicates that a minority of misinformation is created from scratch (Brennen et al., 2020). A prominent example of information that is repurposed with the goal of misleading is fauxtography, where a photo

is altered or miscaptioned to convey false information (Wang et al., 2021). An example of this type of misinformation is shown in Figure 5·5(b), where a news screenshot presents a lion wandering on the street. The caption says that "Russia unleashed more than 500 lions on its streets to ensure that people are staying indoors during the pandemic outbreak." In reality, the picture is depicting a lion that was roaming the streets of a South African city after being released by a local film company in 2016, and it has nothing to do with COVID-19. [2]

3. *Wrong medical advice.* Our dataset contains a number of instances of incorrect medical advice, in accordance with previous work studying health misinformation during the COVID-19 pandemic (Memon and Carley, 2020). Examples include not wearing masks or fake cures against COVID-19. According to past research, verified Twitter handles (including organizations/celebrities) are also active in either generating (new tweets) or distributing misinformation (retweets) (Shahi et al., 2021). In Figure 5·5(c), the screenshot of a news broadcast claims "Malaria drug can treat coronavirus." [3]

4. *Wrong understanding of the threat severity of COVID-19.* There are so-called COVID-19 skeptics (Lee et al., 2021), who do not believe COVID-19 poses a serious threat (Motta et al., 2020; Wray, 2020), putting anyone who believes this message in peril. Examples include claiming COVID-19 is a fabrication or that COVID-19 is a common flu. Figure 5·5(d) shows an example of this type

[2] `https://www.snopes.com/fact-check/russia-release-lions-coronavirus/`

[3] Note that as the medical consensus evolves, so does the common knowledge of what is wrong medical advice. For example, the US Food and Drug Administration (FDA) issued an emergency use authorization (EUA) authorizing the use of hydroxychloroquine to treat certain COVID-19 patients between March 28, 2020 and June 15 (`https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-revokes-emergency-use-authorization-chloroquine-and`). Since my data collection period partially overlaps with this EUA, the consensus around hydroxychloroquine changed during my study. Nonetheless, following my definition of misinformation, I still consider this type of content as misinformation, as the use of this drug to cure COVID-19 was later debunked.

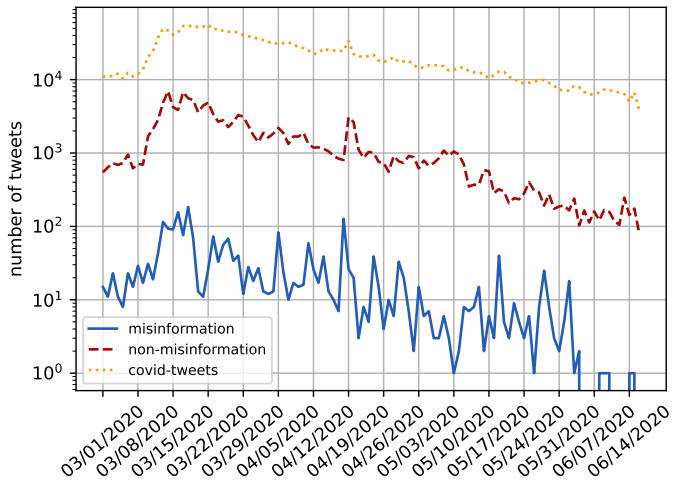


Figure 5·6: Number of tweets containing COVID-19 misinformation images, non-misinformation images, and overall COVID-19 tweets in my dataset appearing every day during my observation period.

of misinformation.

5. *Other false claims.* The text on images may also contain other false claims. One example is the image shown in Figure 5·5(e), claiming that the Trump family refused to donate money for COVID relief. In reality, former US President Donald Trump donated his salary from the last quarter of 2019 to combat COVID-19.[4]

We have limited resources to determine the goal of each cluster of pictures in disseminating misinformation, as well as the accuracy of the misinformation. Therefore, to label images as misinformation we use information gathered from trustworthy fact-checking websites like AP News and WHO Mythbusters.[5]

We label an image as misinformation if it falls into at least one of the five categories. As a result, a single image might belong to two or more categories (Javed et al., 2020).

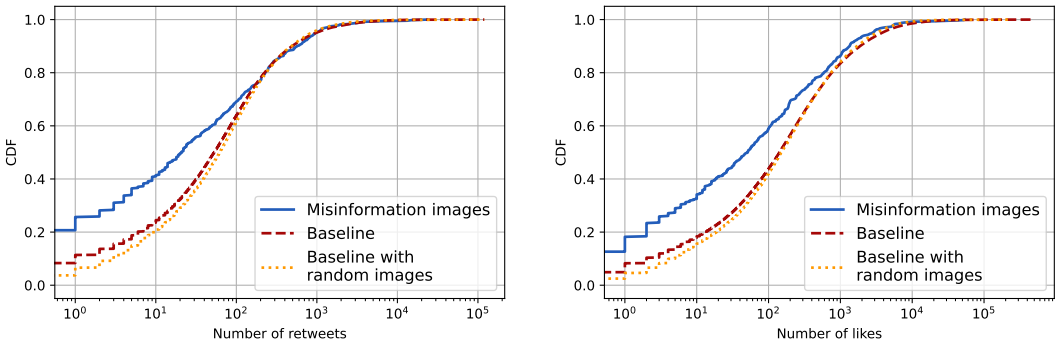
[4] `https://www.cnbc.com/2020/03/03/trump-donates-his-2019-q4-salary-to-help-combat-coronavirus.html`

[5] `https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters`

All the 2,316 informative clusters generated during the first phase are annotated by the two researchers. Similar to Phase I, the two annotators inspect all the images in these image clusters. As long as one image is labeled as misinformation, we label the image cluster as misinformation, and we check every image in the corresponding cluster to see if this image is an image with misinformation. After some deliberation, the two annotators agree on 165 image clusters to be COVID-19 misinformation image clusters, containing 2,418 images posted by 2,404 users. Further inspection reveals only one image among these images clusters is not a COVID-19 misinformation image, which is also ignored. At last, we have 165 image clusters identified as COVID-19 misinformation image clusters, containing 2,417 images posted by 2,403 users. Table 5.1 provides an overview of the number of clusters identified for each category and of the number of images in each of the categories.

5.4 Results

The 165 COVID-19 misinformation image clusters contain 2,417 images included in 2,417 tweets in total while the remaining 7,503 non-misinformation image clusters include 143,774 images shared in 143,755 tweets. Figure 5·6 shows the time occurrences of tweets containing both types of images in my data. In Figure 5·6, I also plot the time occurrence of the 2.3M COVID-19 tweets that are selected from the general tweets, as explained in Section 5.2. As it can be seen, the occurrence of three types of tweets increased at the beginning of the observation period and then gradually declined. The highest number of tweets containing non-misinformation images appeared on March 14, 2020 with a total of 6,956 occurrences, while the day with the highest number of tweets containing COVID-19 misinformation images was March 18, 2020 when the total number of occurrences was 184. As for COVID-19 tweets, the highest number of instances appeared on March 22, 2020, with 55,951 occurrences in total. After these



(a) Retweets (b) Likes

Figure 5·7: CDFs of retweets and likes for tweets with COVID-19 misinformation images vs. baseline tweets.

dates, the number of tweets gradually decreased.

In the following, I present the analysis to answer my three research questions.

5.4.1 RQ1: Do COVID-19 misinformation images generate more user engagement?

I use the number of retweets and likes that Twitter posts receive to characterize user engagement. The raw Twitter data collected using the Twitter streaming API contains real-time activity, i.e., the streaming API gathers tweets as soon as they are posted. However, tweets get re-shared and liked over time, and only looking at this early snapshot is not enough to evaluate the engagement collected by tweets. To comprehensively assess long-term engagement, I re-download the tweets in my dataset based on their tweet IDs, the process of which is called *hydration*.[6] This enables us to know the actual number of retweets and likes of a tweet at the time of hydration.

Twitter posts can be classified as original tweets, retweets, and quote tweets. The difference between quote tweets and retweets is that quote tweets contain some extra text from the users who quote. After hydration, I find that due to limitations in the

[6]`https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/get-statuses-lookup`

Twitter API, I cannot retrieve the actual number of retweets and likes of normal retweets (Wang et al., 2021).[7] For this reason, the assessment of engagement retweets and likes of a normal retweet post is conducted by hydrating the original tweet that produced the retweet post (Wang et al., 2021).

To investigate whether tweets containing COVID-19 misinformation images receive more engagement, I extract a set of random tweets posted by the same set of users who share the tweets with COVID-19 misinformation images. This is to eliminate potential bias introduced by comparing users with a different number of followers (Wang et al., 2021; Zannettou, 2021). My goal is to compare the engagement distribution of baseline tweets to tweets containing COVID-19 misinformation tweets.

To assemble the data for this analysis, I first take the tweet IDs for all tweets in my dataset that contain COVID-19 misinformation images. To avoid duplication potentially introduced by retweets, for those tweets that are retweeted I take the tweet IDs of the original tweets instead. I then deduplicate this set to ensure that each tweet ID is only considered once in this experiment. After this process, I obtain 635 unique tweets containing COVID-19 misinformation images for hydration shared by 565 users. Finally, I hydrate these tweet IDs in April 2021. This is done to ensure that my analysis considers the latest number of retweets and likes. After this process, I obtain 483 unique tweets posted by 429 users, and 152 tweets are not available.

To build the set of baseline tweets for comparison, I obtain all tweets that are contained in the 505M tweets obtained by using Twitter streaming API (See Section 5.2) and are posted by the 429 users who shared COVID-19 misinformation images and follow the same process described above, while in addition removing any tweet that is already considered as part of the COVID-19 image misinformation tweets. The amount of these deduplicated tweets is 63,283. In total, I obtain 59,644 unique baseline

[7] In the Twitter JSON files, the field "retweet_count" of a retweet is equal to the field "retweet_count" of the corresponding original tweet, and the field "favorite_count" of a retweet, which shows the number of likes the tweet receives is always 0, even if the retweet receives a like.

tweets after hydration, and 3,639 unique tweets are not available. The hydration is conducted at the same time described above, in April 2021.

The CDFs of the retweets and likes produced by tweets containing COVID-19 misinformation images (labeled as "Misinformation images") and by baseline tweets (labeled as "Baseline") are shown in Figure 5·7(a) and 5·7(b), respectively. My observation is that baseline tweets are more likely to produce more engagement than tweets containing COVID-19 misinformation images: Baseline tweets receive a median of 53 retweets while COVID-19 misinformation images receive a median of 21 retweets. Similarly, baseline tweets receive a median of 142 likes while COVID-19 misinformation images receive a median of 51 likes.

To evaluate the difference between these distributions, I use two-sample Kolmogorov-Smirnov tests (K-S test) (Lindgren, 1993). I compare the tweets containing COVID-19 misinformation with baseline tweets, and the results show that the difference between these two categories is statistically significant at the $p < 0.01$ level with $D = 0.181$ and $D = 0.178$ for retweets and likes, respectively. Thus, I reject the null hypothesis that tweets containing COVID-19 misinformation images receive the same level of engagement as baseline tweets.

Previous research showed that tweets containing images are more likely to receive engagements on social media (Li and Xie, 2020; Wang et al., 2021). To reduce this bias, similar to previous work (Wang et al., 2021; Resende et al., 2019), I further add one baseline which is composed of 28,390 tweets that contain random images posted by the same 429 users. This set of images is drawn from all baseline tweets that contain images, and these images do not include COVID-19 misinformation images that I identified by my approach (Wang et al., 2021; Resende et al., 2019; Zannettou, 2021). Again, I plot the CDFs of the retweets and likes produced by the added baseline, which is labeled as "Baseline with random images." I observe that tweets containing

COVID-19 misinformation images also tend to produce less engagement than those with other images: Baseline tweets with random images receive a median of 62 retweets while COVID-19 misinformation images receive a median of 21 retweets. Similarly, baseline tweets with random images receive a median of 151 likes while COVID-19 misinformation images receive a median of 51 likes.

I evaluate the difference between the two distributions by using a two-sample K-S test. The comparison is conducted between tweets with COVID-19 misinformation images and tweets with random images, and the results show that the difference between these two categories is statistically significant at the $p < 0.01$ level with $D = 0.216$ and $D = 0.205$ for retweets and likes, respectively.

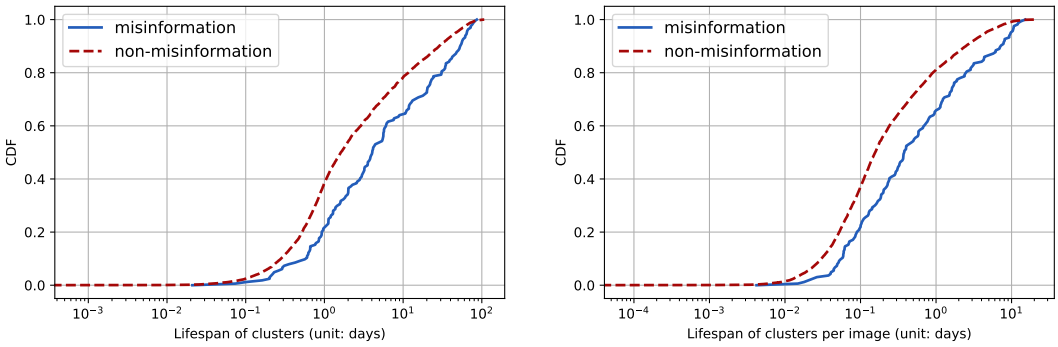
Note that I do not directly compare tweets that contain COVID-19 misinformation images and tweets that contain non-misinformation images that are included in the 7,503 clusters, because the overlap of users sharing them is low: of the 429 users who share COVID-19 misinformation images, only 130 share a total of 420 non-misinformation images. Instead, I use random images, which include all the images other than the misinformation images shared by the same set of 429 users.

This result shows that the tweets containing COVID-19 misinformation images are not more popular than baseline tweets, as well as baseline tweets with random images.

Takeaways of RQ1. From RQ1, I find that COVID-19 misinformation images do not produce as many engagements as the two baselines. In Section 5.6 I discuss the implications that this finding has for the field of misinformation studies.

5.4.2 RQ2: What are the temporal properties of COVID-19 misinformation images? Do COVID-19 misinformation images have a longer lifespan and longer burst times than non-misinformation images?

Another interesting research question beyond the number of likes or retweets that a COVID-19 misinformation image receives is understanding the temporal properties of



(a) Raw lifespan of COVID-19 misinformation images vs. non-misinformation images

(b) Normalized lifespan of COVID-19 misinformation images vs. non-misinformation images

Figure 5·8: CDF of the lifespan of COVID-19 misinformation images and non-misinformation images in my dataset.

COVID-19 misinformation images on Twitter. In RQ2 I aim to answer this question. In particular, I investigate the lifespan and burst time of COVID-19 misinformation images compared to those of non-misinformation images, respectively.

I define the time between the first tweet containing an image in a cluster and the last tweet containing an image from the same cluster posted as the *lifespan* of an image. The lifespan comparison is conducted only between tweets with images to eliminate the effect caused by tweets without images.

Figure 5·8(a) shows the CDFs of the raw lifespan of COVID-19 misinformation images and non-misinformation images, which corresponds to 165 COVID-19 misinformation image clusters and 7,503 non-misinformation image clusters, respectively. We can see that COVID-19 misinformation images tend to linger on Twitter longer than non-misinformation images: Non-misinformation images have a median raw lifespan of 1.62 days while COVID-19 misinformation images have a median raw lifespan of 4.05 days.

I further use a two-sample K-S test to verify the difference between the two distributions. The result shows that the difference is statistically significant between the two distributions at the $p < 0.01$ level with D=0.202. Therefore, I reject the null

hypothesis that COVID-19 misinformation images and non-misinformation images have the same level of lifespan.

Noticing that the size of clusters may influence the lifespan of clusters, I normalize the raw lifespan of clusters by the number of images for each cluster and present the CDFs of the normalized lifespan of COVID-19 misinformation images and non-misinformation images in Figure 5·8(b). Still, I find that the normalized lifespan of COVID-19 misinformation images is more likely to last longer than non-misinformation images: Non-misinformation images have a median normalized lifespan of 0.16 days while COVID-19 misinformation images have a median normalized lifespan of 0.38 days. Similarly, I use a two-sample K-S test to check the differences between the two distributions. The result shows that the difference is statistically significant between the two distributions at the $p < 0.01$ level with D=0.220. I conclude both the raw and normalized lifespan of COVID-19 misinformation images are longer than that of non-misinformation ones.

I further analyze the burst time of images (Resende et al., 2019). I define burst time as the time between two consecutive shares of two images from one cluster, similar to (Resende et al., 2019). Figure 5·9 shows the CDFs of burst times of misinformation images and non-misinformation images. The burst time of misinformation images tends to be longer than that of non-misinformation images: The median burst time of the misinformation image is 0.705 hours while the median burst time of the non-misinformation image is 0.276 hours. Again, I inspect the difference between the two CDFs by using a two-sample K-S test, the result of which indicates that the difference is statistically significant between the two distributions at the $p < 0.01$ level with D=0.165. This allows us to reject the null hypothesis that COVID-19 misinformation images and non-misinformation images have the same level of burstiness.

Note that in RQ1, I use COVID-19 misinformation images compared with random

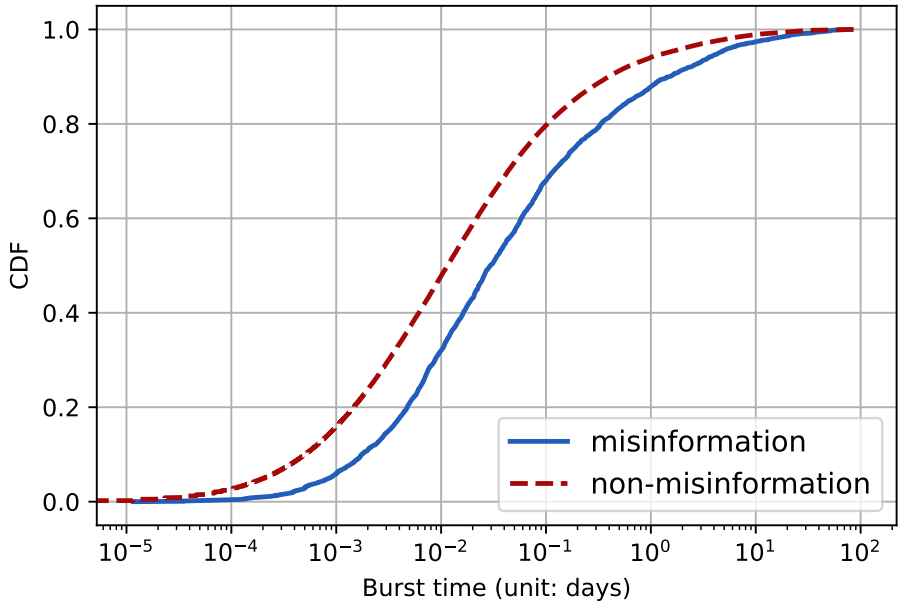


Figure 5·9: Burst time of COVID-19 misinformation and non-misinformation images appearing on Twitter.

images while in RQ2 I use COVID-19 misinformation images compared with non-misinformation images. However in RQ2, I do not compare COVID-19 misinformation images with random images for temporal properties because as I described in Section 5.2, I only download images that are included in COVID-19 tweets, and I do not download random images in June 2020. When I analyze the temporal properties of COVID-19 misinformation images in 2021, various random images are not available for us to download. The quality of random image clusters obtained in this way is affected by the unavailable images, and it introduces bias. In addition, compared with the size of general tweets obtained by Twitter Streaming API and the size of COVID-19 tweets (505M vs. 2.3M), I may have 200 times more clusters for manual inspection to ensure the quality of these clusters, which makes it infeasible for me to remove clusters that are incorrectly clustered. Therefore, I use the existing non-misinformation image clusters to do the comparison, which takes into account the effect that images may

	Users	Tweets
United States	1,282 (45.8%)	1,354 (46.9 %)
India	261 (9.3%)	263 (9.1%)
United Kingdom	142 (5.1%)	142 (4.9%)
Malaysia	99 (3.5%)	99 (3.4%)
South Africa	90 (3.2%)	91 (3.2%)

Table 5.2: Top 5 countries, dependencies, and areas of special sovereignty of users sharing COVID-19 misinformation images.

bring, and it is still a good indicator to characterize that COVID-19 misinformation images may have longer lifespans and burst times.

Takeaways of RQ2. As for temporal properties of COVID-19 misinformation images, I find that they tend to have longer lifespans, regardless of raw or normalized, and burst times compared with non-misinformation images. This suggests that COVID-19 misinformation images may linger longer on Twitter and have a longer-term negative effect on Twitter users.

5.4.3 RQ3: What are the characteristics of users who post COVID-19 misinformation images?

In the first two research questions, I analyzed the characteristics of COVID-19 misinformation images and of the tweets discussing them. In this section, I switch the attention to the users posting COVID-19 misinformation images, looking for their characteristics and for patterns. I first look at geographical information to better understand the geographic makeup of the users in my dataset. Next, I look at the profile description of users, looking for popular hashtags. Finally, I look at the political leanings of the users in the US posting COVID-19 misinformation images on Twitter. From the 165 image clusters that contain COVID-19 misinformation images, I have 2,417 tweets that contain COVID-19 misinformation images, which include 165 original tweets, and 2,252 retweets. The metadata of these retweets contains 475 unique original tweets. After removing the overlapping original tweets, in total, I have 2,887 tweets containing

COVID-19 misinformation images posted by 2,801 users.

Analysis of user locations. I look at the users who post COVID-19 misinformation images based on their location (see previous studies like (Zheng et al., 2018; Ajao et al., 2015)). To do so, I manually checks all the 2,801 users to determine their home locations by using location features and indicators from users' profiles in the collected Twitter user metadata, which include location fields, self-descriptions in bios, and flag icons (Zheng et al., 2018; Ajao et al., 2015). This information is at the granularity level of countries, dependencies, and areas of special sovereignty, e.g., Puerto Rico and Hong Kong. [8] If the profiles do not provide enough information, then I infer the home locations of these users by using the content (Han et al., 2014) of their tweets posted in 2020, as well as interaction information (e.g., mentions, retweets, or replies) generated in 2020 between these users and users whose home locations are explicit (Ajao et al., 2015; Backstrom et al., 2010; Jurgens, 2013; McGee et al., 2013). If such information is still not enough to infer their location, I classify them as "unknown."

After the manual inspection, I find that the location for 2,656 users is known, where for 1,845 users the location is obtained from their metadata, and the other 811 users are inferred from their tweets. For the remaining 145 users, the location is unknown. The home locations of users who post COVID-19 misinformation images are more than 10 countries, dependencies, and areas of special sovereignty, including English-speaking countries, e.g., US, UK, and Canada, as well as non-English speaking countries, e.g., China, Indonesia, Mexico. I present the Top 5 countries, dependencies, and areas of special sovereignty that the home locations of users who post COVID-19 misinformation images are in Table 5.2.

From Table 5.2, I can see that users from the top 5 countries, dependencies, and areas of special sovereignty account for 66.9% of all users who share tweets with COVID-19 misinformation images in my dataset. Users from the US are the most,

[8] `https://www.state.gov/dependencies-and-areas-of-special-sovereignty/`

and their portion is almost half of all the users. US users also share nearly half of all the tweets that contain COVID-19 misinformation images. I also observe that in the top 5 countries from Table 5.2, English is a *de facto* official language,[9] which is partly because the hashtags I select are mostly from English and I exclude images that contain text other than Chinese and English from being COVID-19 misinformation image candidates.

Note that location estimation methods described above may not work for the estimation of the city or finer level of location granularity, however, I still argue that this location estimation approach with the full available information can be used as the best guess for the country level of location granularity, i.e., the most likely one, which is useful to help us understand the constitution of user locations.

Analysis of user bios. I start by measuring the most common hashtags used in their user biographies among all the 2,801 users who post COVID-19 misinformation images, as previous work shows that hashtags in bios can reveal user characteristics, for example, their political leaning (Conover et al., 2011a). Only 324 users in my dataset (11.6%) include hashtags in their bios. The top 20 hashtags in user bios are shown in Table 5.3. As we can see several hashtags suggest the user's support of Republican presidential candidate Trump, e.g., #MAGA ("Make America Great Again"), #KAG ("Keep America Great"), #Trump2020, #BuildTheWall, etc. as well as hashtags that are commonly associated to conservatism, e.g., #Conservative and #NRA ("National Rifle Association"). We can also find hashtags that are associated with the Democratic party or Anti-Trump movements, e.g., #Biden2020, #Resist (Times, 2017), and #FBR (often stands for following back resistance). Interestingly, there are also hashtags related to QAnon, which is a US far-right movement promoting various conspiracy theories (Papasavva et al., 2021; Aliapoulios et al., 2021). Such hashtags include #Q,

[9] `https://en.wikipedia.org/wiki/List_of_countries_and_territories_where_English_is_an_official_language`

Hashtag	Count
#MAGA	82
#KAG	43
#WWG1WGA	38
#Resist	32
#2A	26
#Trump2020	18
#TheResistance	18
#FBR	17
#1A	13
#Patriot	12
#Resistance	12
#Q	10
#TRUMP2020	9
#NRA	9
#QAnon	8
#followbackhongkong	8
#Biden2020	7
#Conservative	7
#resist	7
#BuildTheWall	7

Table 5.3: Top 20 hashtags in the bios of user profiles

#QAnon, #WWG1WGA ("When we go one, we go all," which is a popular slogan among QAnon adherents (Papasavva et al., 2021)).

This result indicates that many accounts posting COVID-19 misinformation images are focused on US politics. This is in line with the fact that the majority of the users in my dataset are from the US (1,282, as shown in Table 5.2).

Analysis of US users' political leanings. To better understand their political leanings, one PhD student and I further annotate the US-based users with their political leanings (Conover et al., 2011a; Kulshrestha et al., 2019).

We determine to manually check the US-based users who post COVID-19 misinformation images based on their profiles in my collected Twitter user metadata and their tweets posted in 2020 to annotate their political leanings (Zannettou, 2021; Conover et al., 2011a; Kulshrestha et al., 2019). To do so, we have developed a codebook which is shown below:

- *"Pro-Republican."* We use "Pro-Republican" to represent the political leanings

of users who identify themselves as Trump supporters or Republican supporters who are not against Trump. The pro-Republican users often use keywords like "MAGA," "KAG," and "Trump2020" to show their support for Trump or keywords like "Republican" and "Conservative" to show their support for the Republican Party in their profiles and tweets.

- *"QAnon."* Among the pro-Republicans users[10], we code those who support the QAnon movement as "QAnon." QAnon adherents often describe themselves with keywords like "QAnon" and "WWG1WGA" in their profiles and tweets.

- *"Pro-Democrat."* We code users who identify themselves as Democratic supporters or who are against Trump as users whose political leaning are "Pro-Democrat." The pro-Democrat users often use keywords like "Democrat" and "Biden2020" to show their support for the Democratic Party and keywords like "Resist" and "FBR" to show they are against Trump in their profiles and tweets.

- *"Neutral."* If a user supports neither Republicans nor Democrats in their profile and tweets, we classify their political leaning as "neutral."

- *"N/A."* If there is no clue in the profile to show the political leaning of a user, and their profile is suspended or their account is protected, then we classify their political leaning as "N/A."

Following the approach of similar work (Zannettou, 2021; Conover et al., 2011a; Kulshrestha et al., 2019), we randomly select 300 users from the 1,282 US-based users and independently code these users based on their profiles in my collected Twitter user metadata and their tweets posted in 2020 to determine their political leanings. To evaluate the agreement between the two annotators, we need to calculate Cohen's

[10] One of the core beliefs among QAnon adherents is to support Trump. See `https://en.wikipedia.org/wiki/QAnon` for further explanations.

	Users	Tweets
pro-Republicans	508 (39.6%)	538 (39.7 %)
QAnon	111 (8.7%)	117 (8.6%)
pro-Democrats	546 (42.6%)	586 (43.2%)
Neutral	118 (9.2%)	118 (8.7%)
N/A	110 (8.6%)	112 (8.3%)

Table 5.4: Political leanings of users post misinformation images tweets among users from the United States

Kappa between the annotation results of the two annotators. Since QAnon users are a subset of Pro-Republican users and Cohen's Kappa does not apply to the case in which one item is given two labels, we split the annotation process into two parts:

- In the first part, we only focus on the four categories "Pro-Republican," "Pro-Democrat," "Neutral," and "N/A." Then we calculate the Cohen's Kappa between the results of the two annotators and find a very high agreement ($\kappa = 0.890$), which suggests the two annotators strongly agree with each other, verifying the validity of this codebook about the four categories "Pro-Republican," "Pro-Democrat," "Neutral," and "N/A."

- After that, the two annotators discuss and determine 152 users out of these 300 users as Pro-Republican. Then in the second part, the two annotators code these 152 users as "QAnon" or not independently. We again calculate the Cohen's Kappa between the QAnon coding results of the two annotators and we find the two annotators highly agree with each other ($\kappa = 0.941$), confirming the validity of this codebook about the category "QAnon."

After establishing that the codebook is reliable, the remaining users are annotated by me. The annotation result is shown in Table 5.4, where we can see that more than 80% of users who post COVID-19 misinformation images in the US present indicators that they support a political party. Also, I find that the number of tweets with

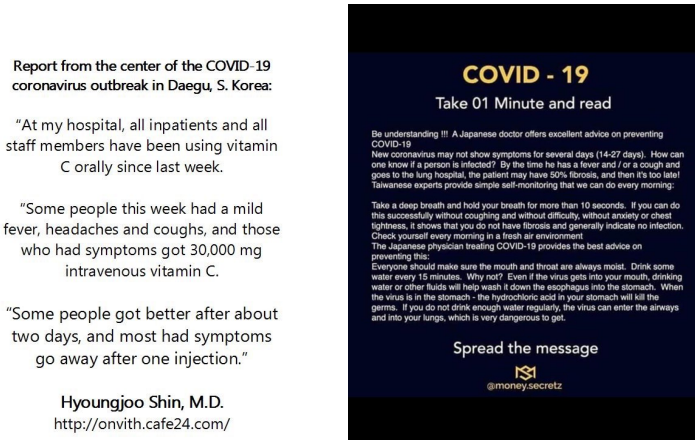


Figure 5·10: Images spreading COVID-19 treatment rumors in my dataset.

COVID-19 misinformation images shared by pro-Republican and pro-Democrat users is close. In Section 5.5 I further investigate what type of COVID-19 misinformation images is shared by the supporters of the two political parties, finding that supporters of the two political parties share different types of COVID-19 misinformation images. I also find that although QAnon supporters are not mainstream pro-Republican, they are still a non-negligible part of the users who support the Republican party: QAnon theory adherents are around 21.9% of all users who are pro-Republicans.

Takeaways of RQ3. Overall, I find that users who post COVID-19 misinformation images are from numerous countries and regions, and nearly half of them are from the US. Among the users from the US, I find most of them show explicit political leanings, and the numbers of tweets posted by users who support Republicans and Democrats are close to each other. In Section 5.5 I further analyze the type of misinformation images shared by users with different political leanings.

5.5 Case Studies

In this section, I present three case studies to better illustrate the types of COVID-19 misinformation narratives that I observe in my dataset. First, I examine some narratives on false or unconfirmed treatments for COVID-19. I next proceed to look at viral conspiracy theories about Bill Gates, who is blamed for creating the virus. Finally, I show examples of COVID-19 misinformation images shared by pro-Republican and pro-Democrat users, showing that users with different political leanings share different types of COVID-19 misinformation images.

False and unconfirmed treatments for COVID-19. Throughout the COVID-19 pandemic, several rumors about cheap and accessible treatments against COVID-19 have emerged on the Web. However, these treatments are not effective against the virus, or have never been confirmed by rigorous clinical trials. I see 33 such image clusters in my dataset. The image on the left of Figure 5·10 encourages people to take vitamin C to treat COVID-19. This treatment was debunked by institutes like the NIH.[11] Another alternative treatment example is shown on the right of Figure 5·10, claiming that drinking water can prevent people from getting COVID-19. Likewise, this treatment is also ineffective as explained by BBC.[12]

Conspiracies on Bill Gates. Bill Gates has been the target of conspiracy theories since the beginning of the COVID-19 pandemic. According to The New York Times and Zignal Labs, he was named 1.2 million times in the two months leading up to the initial global pandemic in 2020 (Daisuke Wakabayashi, Davey Alba, and Marc Tracy, 2020). These alternative narratives are directed toward Bill Gates himself and the Bill and Melinda Gates Foundation.

[11] https://ods.od.nih.gov/factsheets/DietarySupplementsInTheTimeOfCOVID19-Consumer/\#:~:text=Research\%20hasn't\%20clearly\%20shown,and\%20minerals\%20to\%20work\%20properly.

[12] https://www.bbc.com/future/article/20200319-covid-19-will-drinking-water-keep-you-safe-from-coronavirus

Figure 5·11: Images promoting conspiracies on Bill Gates in my dataset.

(a) (b) (c)

Figure 5·12: COVID-19 misinformation images shared by users supporting the Democratic party.

Bill Gates conspiracies are aimed at him and his philanthropic work in advancing global health issues. In my dataset, I find 14 clusters related to Bill Gates conspiracy theories. Two examples of such images are shown in Figure 5·11. In the left image, Bill Gates is accused of testing the COVID-19 vaccine on Indian children without their consent. In the right one, a fauxtography image is miscaptioned to claim to portray a rally demanding the arrest of Bill Gates for crimes against humanity.

Misinformation images shared by pro-Republican and pro-Democrat users. The COVID-19 pandemic has become a polarizing issue in the US and has been at the center of competing narratives during the 2020 General Election. As shown in Section 5.4.3, in my analysis I found that pro-Democrat and pro-Republican users share

a similar amount of tweets with COVID-19 misinformation images. In my dataset, I see pro-Democrat and pro-Republican users involve in sharing 49 and 92 COVID-19 misinformation image clusters, respectively, for a total of 586 tweets shared by pro-Democrat users and 538 tweets shared by pro-Republican users. This is somewhat surprising since previous work (Lazer et al., 2021) shows that the amount of COVID-19 misinformation articles shared by pro-Republican users exceeds tremendously than those shared by pro-Democrat users, and one would expect that this ratio might represent the general trend for COVID-19 misinformation image sharing between Republicans and Democratic supporters on Twitter. However, in my case, I find that with respect to sharing COVID-19 misinformation images, the amounts of tweets shared by pro-Republican and pro-Democrat users are much closer. In this section I aim to shed light on this finding, analyzing what type of COVID-19 misinformation images is shared by users with different political affiliations.

I find that it is rare for Republican and Democrat users to share the same COVID-19 misinformation images on Twitter: only 17 image clusters are shared by both pro-Democrat and pro-Republican users, and there are 32 out of 49 image clusters and 75 out of 92 image clusters that are only shared by pro-Democrat and pro-Republican users respectively. I present six examples that are shared mutually exclusive between the pro-Democrat and pro-Republican users in Figure 5·12 and Figure 5·13.

Among the 32 image clusters shared by pro-Democrat users, the most shared ones are false or misleading claims surrounding the Trump administration's response to the pandemic, where 9 image clusters belong to this category. Figure 5·12(a) shows an image shared by a pro-Democrat user. The caption criticizes President Trump for refusing to accept WHO-supplied test kits, a claim that has been debunked as not true.[13] The image comes with a tweet stating "This all day" and three hashtags

[13]Please see `https://www.factcheck.org/2020/03/biden-trump-wrong-about-who-coronavirus-tests/` for fact-checking

"#COVID19US," "#TrumpRefusedTestKits," and "#TrumpIsTheWORSTPresidentEVER." This indicates a tendency of pro-Democrat users to share false claims for political gain, for example by making the Trump Administration's response to the pandemic appear worse than it actually was. Another common COVID-19 misinformation image type is related to various false claims targeting issues other than Trump, which has 5 clusters. One such image example in Figure 5·12(b) falsely claims pandemics happen every 100 years, which is also proved to be incorrect.[14]

Additionally, I find several examples of fauxtography images shared by pro-Democrat users. These images are also shared with the goal of criticizing the Trump administration's response to COVID-19 but often have a satirical angle. 7 out of 32 image clusters in this group belong to this category. One example is shown in Figure 5·12(c), which is a meme showing former President Trump screaming at dead bodies around him to draw attention to his approval rate. This image is a composition of a photo documenting the tragic scene of dead bodies inside a truck posted by a New York nurse[15] and a popular meme showing Trump yelling at a boy mowing the White House lawn.[16] The purpose of this fauxtography is likely to satirize the former's president alleged obsession with approval ratings during a time of great tragedy.

On the other hand, I find 27 out of 75 clusters indicate that pro-Republican users post COVID-19 misinformation images advocating for conspiracy theories and false claims about China or Democrats. For example, Harvard University Professor Charles Lieber was charged for hiding his links with a Chinese University just after the outbreak of the COVID-19 pandemic in 2020,[17] which caused a conspiracy theory that the coronavirus might be a bioweapon of China, as shown in Figure 5·13(a). The

[14]Please see https://www.statesman.com/story/news/politics/elections/2020/04/10/fact-check-has-pandemic-occurred-every-100-years/984128007/ for fact-checking

[15] https://www.dailymail.co.uk/news/article-8167283/Horrifying-moment-dead-bodies-loaded-refrigerated-truck-forklift.html

[16]https://knowyourmeme.com/memes/trump-yelling-at-lawn-mowing-boy

[17]https://en.wikipedia.org/wiki/Charles_M._Lieber

(a) (b) (c)

Figure 5·13: COVID-19 misinformation images shared by users supporting the Republican party.

associated text in the tweet reads "#CoronaVirus arrests$\cdots$," which indicts the user connected the charge with COVID-19. Another example, shown in Figure 5·13(b), is a conspiracy theory about the origin of COVID-19, claiming that House Speaker Nancy Pelosi took part in releasing the coronavirus to help President Biden get elected. The text associated with this image has the hashtag "#Coronavirus, " and asserts that the exaggeration of COVID-19 is the biggest political fraud in history. Another common narrative shared by pro-Republican users promoted the use of hydroxychloroquine to treat COVID-19,[18] which was later proven to be ineffective, a narrative that was also embraced by former President Trump, making it a controversial political issue, unlike other COVID-19 treatment rumors. An example is an image in Figure 5·13(c), which falsely claims that hydroxychloroquine is effective against COVID-19 and conjecture that people do not want to use hydroxychloroquine because they are reluctant to admit Trump is correct. The user who posted the image tagged such response as "#TrumpDerangementSyndrome," in the associated Tweet text.

When looking at users who support QAnon, I find the tendency to support

[18] `https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-hydroxychloroquine`

(a) (b)

Figure 5·14: COVID-19 misinformation images shared by users who are QAnon adherents.

conspiracy theories surrounding COVID-19. I find 53 COVID-19 misinformation image clusters posted by QAnon adherent users, and 33 of them promote conspiracies, in line with the actions of the community that were observed by previous work (Aliapoulios et al., 2021; Wong, 2018). The image in Figure 5·14(a) attacks the role of the CDC, claiming that it is a vaccine company. The user posted this image with a passage of text with similar content as the text on the image. Another image example is shown in Figure 5·14(b) and the associated text they posted reads " the #Covid_19 is Bilderberg after dinner entertainment #Agenda21." Here Bilderberg meeting is a secret annual meeting participated by powerful people around the world[19], and Agenda 21 is a goal set by the United Nations for global development.[20] This suggests this QAnon adherent user believes that the outbreak of the COVID-19 pandemic is planned by the United Nations and other powerful entities, aiming to construct the so-called "New World Order."

[19] `https://en.wikipedia.org/wiki/Bilderberg_meeting`
[20] `https://en.wikipedia.org/wiki/Agenda_21`

5.6 Discussion

In this section, I first discuss the implications of my findings for the field of misinformation research. I then reason on what my results mean for platforms aiming to moderate and mitigate misinformation. Finally, I discuss the limitations of my study.

Implications for misinformation research. I present a codebook that aims to characterize the various types of COVID-19 misinformation images. As the pandemic progresses and new narratives emerge, this codebook can be used by other misinformation researchers to better characterize new narratives and how they propagate on social media.

Although this dataset is rather small, I make several findings that are of interest to the misinformation research community.

Unlike previous works on textual (Vosoughi et al., 2018) and visual (Wang et al., 2021) misinformation, I find that image-based COVID-19 misinformation does not receive more engagement than baseline tweets, as well as baseline tweets with random images on Twitter. A possible reason is that those previous works focused on news stories and images that had been analyzed and debunked by fact-checking organizations. While analyzing false news and images that are popular enough to be fact-checked is useful to understand the misinformation phenomenon, this can introduce a selection bias where smaller false narratives that do not meet the threshold for fact-checking are simply ignored.

From the perspective of political leanings, I find that users supporting the Republican and Democratic parties post a similar rate of COVID-19 misinformation images. This is interesting, because it may be different than what was found by previous work (Lazer et al., 2021). Upon further inspection, I find that the supporters of the two parties share different types of false or misleading images. For instance, I find that Democrats often share misleading and false claims about the Trump administration's

response to the pandemic, as well as manipulated images (i.e., fauxtography) used as a satirical commentary to such response. Republican users, on the other hand, prominently share conspiracy theories about the virus, claiming that it was manufactured by powerful entities (e.g., China, the Democratic Party, United Nations) to achieve various nefarious goals. Republican supporters also often share images promoting the use of hydroxychloroquine to treat COVID-19. These findings show that there are different types of COVID-19 misinformation images that become popular on social media, some more dangerous than others, and that social network companies should take this into account when choosing how to mitigate them (Zannettou, 2021). These results also highlight the need to develop approaches able to identify the *intent* with which a misleading image is posted. For instance, posts that are advocating for the adoption of some unproved cure are potentially dangerous to the community and should be considered by platforms for moderation, while posts that satirize such cures are not a threat to public health. At the moment, the research community lacks automated methods to distinguish between the two.

Instead of following a top-down approach dictated by fact-checking websites, in this research, I follow a bottom-up one, where my data analysis and clustering identify COVID-19 misinformation images that are worth being studied. What I find is that COVID-19 misinformation images are in general not more popular than tweets with random images, but they are shared for longer periods of time compared with non-misinformation images. Going forward, I argue that misinformation researchers should combine the different approaches, relying on solid ground truth provided by fact-checkers, but also relying on real-world data to identify stories that might not meet the threshold to be fact-checked, but nonetheless are discussed on social media for long periods of time.

Implications for platform moderation. This research highlights the types of

image-based misinformation that are shared on Twitter during the COVID-19 pandemic. I find a set of challenges and opportunities that can be adopted by platforms like Twitter aiming to curb the misinformation problem.

First, I find that conspiracy narratives that target politicians and wrong medical advice are particularly common on Twitter. These narratives make it challenging for the public to be properly informed, and can hamper public measures like mask mandates and vaccination requirements. In this setting, a promising avenue is using soft moderation measures (Zannettou, 2021), where misinformation tweets are not taken down but labels providing additional information and resources are provided to the user.

Second, I find that COVID-19 misinformation images on Twitter have a longer lifespan. In particular, the same misinformation image can be used for long periods of time and even resurface to promote new false narratives. For example, images of Bill Gates promoting vaccinations have been popular ground for conspiracy theorists well before the COVID-19 pandemic, and have risen back to popularity as efforts to develop the COVID-19 vaccine mounted. This gives an opportunity for platforms to develop image-based mitigation able to identify emerging misinformation narratives that use new versions of old images.

This approach can allow platforms to identify and moderate misinformation images that are pervasive but are never shared by popular accounts, and therefore risk remaining under the radar with current approaches. I argue that by looking at images that are shared hundreds of times on Twitter (i.e., form large clusters following my approach) moderators could identify and curb emerging misinformation narratives, potentially before a popular account re-shares that information and makes it visible to even more users.

Finally, I find that memes are often used to promote false narratives on Twitter.

This opens up a number of challenges for platforms because the line between satire and harmful information is often blurry and very context-dependent. I argue that Twitter and other social media platforms should invest in technology that can identify the purpose and the context in which an image is used to improve automated misinformation detection systems.

Limitations of my study. In this research, I provide a comprehensive characterization of COVID-19-related misinformation images. Nonetheless, this work still comes with limitations. First, the set of hashtags that I use is English only. Since the pandemic is global, this unavoidably missed tweets posted by people speaking other languages. Future work might replicate my analysis of data from other languages. Second, COVID-19 tweets may not use the hashtags that I selected, causing my analysis to miss relevant images. Third, my selection criteria for human annotation is that images need to cluster together and the cluster size needs to be five or higher. I have to do this to keep the analysis manageable for the two annotators and to ensure the quality of clusters. This means that I may miss less popular images; given that my data source is the Twitter 1% sample, I expect these images in smaller clusters to be shared less than 500 times on Twitter at the time when I collect data. Also, since some Twitter users have already been suspended from Twitter and their profiles do not provide enough information, it prevents us to determine the characterization of some users. This study also inherits the limitations of using the public Twitter Streaming API, which only provides us with a view of 1% of all public tweets. While this dataset is partial, I believe that it still helps us to infer general trends in the share of COVID-19 misinformation images on Twitter during the observation period.

In addition to that, my work annotates images that are grouped into clusters. These images are generally more popular, which makes my results biased toward images that go viral. Future research may find efficient ways to annotate images that

are less popular and are likely to be pervasive on social networks. Another issue is that the number of COVID-19 misinformation images identified by my study is small, and therefore further research is needed to investigate if these results generalize to larger datasets and settings. Though the research has the above limitations, I believe this research still helps people better understand what impact the misinformation images have on Twitter, and provides insight on further Twitter moderation.

5.7 Conclusion

In this chapter, I build a large dataset of misinformation images related to COVID-19 posted on Twitter between March and June 2020. I develop a codebook to characterize the various types of COVID-19 misinformation images related to the virus, from false medical advice to conspiracy theories. I then use this dataset to understand how COVID-19 misinformation images are used on social media. I find that these images do not receive more retweets and likes than tweets with random images. On the other hand, COVID-19 misinformation images are shared for longer periods of time than non-misinformation images on Twitter. I also find that COVID-19 misinformation images are shared by users who support the Democratic and the Republican party in similar numbers, but there is a difference in the type of images that the two groups share. While Democratic users often share misleading facts on the Trump administration's response to the pandemic, together with manipulated satirical images to critique this response, Republican users often share conspiracy theories about the origin of the virus, as well as images advocating for false treatment of hydroxychloroquine against COVID-19. My findings help researchers gain a better understanding of image-based misinformation on social media, and identify a number of challenges and opportunities for further research in this space.

Chapter 6

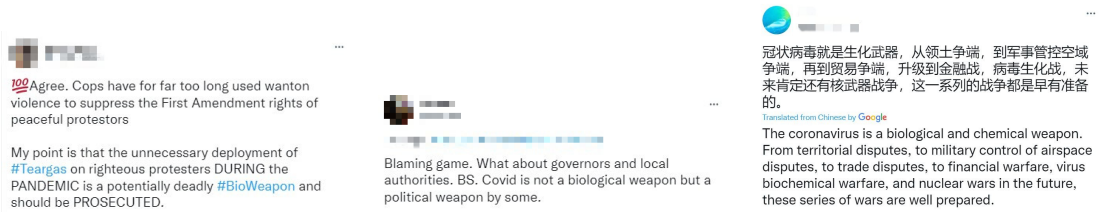
COVID-19 multi-lingual misinformation on Twitter

6.1 Introduction

In the past three years, the COVID-19 pandemic has been an important health emergency worldwide. Misconceptions and misinformation permeate public discourse online and offline, regardless of the language that people might speak, which is disastrous to our society. To minimize their impact, it is desired to apply moderation on posts that convey these misconceptions (Zannettou, 2021), which requires effective techniques to detect misinformation and the stance that a user has towards that piece of misinformation,

For instance, a widespread misconception during the COVID-19 pandemic is that COVID-19 is a bioweapon (Poynter, 2020). A naive approach to detect this misconception in a tweet is to search whether this tweet contains keywords like "COVID-19," "bioweapon." However, this method may not work. A counterexample is shown in Figure 6·1(a), where the author of this tweet believes that teargas used towards protesters in the pandemic is a bioweapon. Though it contains keywords like "pandemic" and "bioweapons," it does not connect coronavirus with bioweapons.

For more effective moderation, it is also important to identify the stance of a tweet and moderate the tweet that support the misinformation on multi-lingual social media platforms (Zannettou, 2021). Such examples can be seen in Figures 6·1(b) and 6·1(c)



(a) Irrelevant (b) Refute (c) Support

Figure 6·1: Examples of Twitter misinformation moderation candidates.

which shows a Chinese tweet with its English translation. We can see that the author in Figure 6·1(b) refutes the misconception that COVID-19 is a bioweapon, which is, however, promoted by the author in Figure 6·1(c). As a result, the tweet shown in Figure 6·1(c) should be moderated to minimize its malicious effect on other Twitter users. However, platforms like Twitter may only be able to moderate posts in a limited number of languages manually, and posts in other languages need to be monitored by some other automatic ways at scale.

One possible way to alleviate the multi-lingual challenge is that if we accumulate enough knowledge for one misconception in one language, we may be able to transfer it to moderating the same misconception in another language, which is beneficial for multi-lingual social media platforms.

Following this method, we compile an English and Chinese bilingual annotated Twitter dataset relevant to COVID-19 misconceptions[1] with a total number of 6,000 tweets by using crowdsourcing. The annotation is composed of identifying whether one specific misconception is contained in one tweet and what stance the author of this tweet has toward this misconception. Based on this dataset, I experiment on several natural language processing (NLP) models to investigate their performance on misinformation detection and stance detection in both monolingual and multi-lingual manners. I find that two models: COVID-Tweet-BERT v2 and BERTweet

[1]The misconceptions are related to ginger/garlic, hydroxychloroquine, and bioweapon

	#Ginger/Garlic	#HCQ	#Bioweapon
En	70,164	543,097	128,348
Zh	4,712	5,514	11,432

Table 6.1: Overview of the whole dataset. "En" stands for "English" and "Zh" stands for "Chinese."

are generally effective in detecting misinformation and stance detection in the two above manners, and the multi-lingual detection performances of these two models are comparable with the monolingual detection ones. In certain cases, the multi-lingual detection of these two models even outperforms the monolingual detection. The results reveal that both the models COVID-Tweet-BERT v2 and BERTweet are promising to be applied to misinformation moderation on social media platforms, which heavily depends on identifying misinformation and stance of the author towards this piece of misinformation. The detailed steps are shown in the following sections.

6.2 Dataset Construction

In this section I describe how we build the annotated dataset. In a nutshell, I first construct an unlabeled dataset by using BrandWatch and Twitter API. Then I take a sample of this dataset and use crowdsourcing to annotate this sample. I show the details in the following subsections.

6.2.1 Builidng an unlabeled dataset

To begin with, we are interested in three COVID-19 relevant misinformation conceptions debunked by institutes like WHO (WHO, 2021; WHO, 2020) and fact-check agencies like IFCN (Poynter, 2020):

1. Ginger/garlic can be used to cure COVID-19.

2. Hydroxychloroquine (hcq) is a cure for COVID-19.

3. COVID-19 is a bioweapon.

I query English Twitter data through BrandWatch, which can provide a full access to the Twitter data for specific keywords, over the period from Jan 1, 2020 to Dec 31 2020 based on a list of keywords related to COVID-19 compiled in (Chen et al., 2020a) combined with a list of misconception related keywords, composed of "hydroxychloroquine," "bioweapon," "ginger," and "garlic." The query requires each tweet contains at least one keyword from the list of COVID-19 related keyword list and at least one keyword from the list of misconception related keywords, which ensures every tweet is likely to be related to one of the three misconceptions, e.g., the example in Figure 6·1(b). The query result does not contain text of the tweets directly, instead it contains some other key information of a tweet like its tweet ID and the ID of its root tweet, which is the first tweet of a thread of tweets. Similarly, to obtain Chinese tweets, I manually translate every keyword in the COVID-19 related keyword list and the misconception related keyword list into Chinese. Then I query Chinese tweets based on these two Chinese keyword lists following the same way as English tweets. In total, I obtain 6,797,155 unique IDs of tweets, including their root tweets. To obtain the text of tweets, I hydrate these tweet IDs. Except 1,900,366 unavailable tweets, I obtain 4,896,789 tweets.

I further clean this set of tweets by keeping tweets that contain at least one keyword from the list of misconception related keywords regardless of English or Chinese. The reason for this step is that the above mentioned set of tweets contains root tweets, which may be related to the three misconceptions if they contain a misconception related keyword. Thus they can be cleaned by reserving the tweets that contain one of the keywords from the list of misconception related keywords. After this step, I have 3,191,199 tweets in total. Then I further remove all the retweets, and there are 762,167 original tweets and quote tweets left. The statistics of this dataset can be

	#M/#NM	#S/#R/#N
Ginger/Garlic (En)	434/566	277/96/627
Ginger/Garlic (Zh)	159/841	106/32/862
HCQ (En)	705/295	324/225/451
HCQ (Zh)	634/366	372/148/480
Bioweapon (En)	645/355	497/86/417
Bioweapon (Zh)	745/255	531/87/382

Table 6.2: Overview of the labeled dataset. #M and #NM stand for the number of tweets that are related to and not related to the specific misconception, respectively. #S,#R, and #N stand for the number of tweets that whose stance is support, refute, and none, respectively

seen in Table 6.1. Given the sheer scale of the dataset, we can only annotate a sample of these tweets to compile a labeled dataset, which is explained in the next subsection.

6.2.2 Annotating sampled data

I sample 1,000 tweets for each of the 3 misconceptions in both English and Chinese tweets. In total, I obtain a sampled dataset with 6,000 tweets.

For each misconception, two PhD students and I develop a English codebook, which is translated to Chinese by one Chinese PhD student and me manually, to guide the annotation.

For each tweet, we are interested in the following three questions:

1. Is this tweet related to a misconception that we focus on? The answer can be either yes or no. If the answer is no, the answer to the other two will be not applicable.

2. What is the user's attitude towards this tweet? The attitude can be spread, refute, or none if their answer to the first question is yes. Otherwise, the answer is not applicable (N/A).

3. Is this tweet related to a conspiracy theory? The answer can be yes or no if their answer to the first question is yes and N/A otherwise.

A conspiracy theory is an explanation for an event that happens because of a plot planned by powerful people or organizations (Wikipedia, 2022). Note in this research, I do not use the answers related to conspiracy theories, which will be left for future exploration. I attach a English version of the codebook in the Appendix for the misconception related to hydroxychloroquine.

To annotate English tweets, I use Amazon Mechanical Turk to find people working on annotation. To ensure a high quality annotation, the two PhD students mentioned above and I develop a golden set of tweets and check the annotations of crowdsourced workers against them. To this end, I sample 200 tweets among 1,000 tweets for each misconception, and the two PhD students and I annotate them. The agreed annotation is then used as the answers to the golden set. To ensure the quality of the crowdsourcing work, I require every worker has a rating more than 90%, having finished at least 100 tasks, and answer correctly at least 6 questions out of 9 questions towards 3 tweets in the golden set.

I ask the crowdsourced workers to work on the 800 unlabeled tweets for each English misconception. Each tweet is annotated by 5 workers (Guo et al., 2020; Ma and Olshevsky, 2020), and I select the label for each question by using a majority vote. If one final label of a question of a tweet earns at most two votes, the three researchers who annotate the golden set examine this tweet together again and annotate this tweet for all the three questions as the final labels. For example, this situation happens when two vote for support, two for refute, and one for none for the second question of a tweet.

For Chinese tweets, I do not seek to annotate them on Amazon Mechanical Turk since this is an English-centric community and it is likely that I cannot find enough people eligible as suggested in (Park et al., 2021). Instead, I recruit annotators from users in the groups of Chinese social media platform Tencent QQ.

We build a golden set of 60 tweets for each Chinese misconception annotated by two PhD students and me as well. We ask every Chinese crowdsourced annotator to finish a test containing 3 tweets in the golden set. Same with English tweets, only a person who answers correctly at least 6 questions out of 9 questions towards 3 tweets in the test is eligible for annotation. Then we ask Chinese crowdsourced annotators to work on the remaining 940 unlabeled Chinese tweets for each misconception. Again, if one final label earns at most two votes, the three researchers examine the corresponding tweet together and annotate all the three questions of this tweet to obtain the final labels.

By doing so, we construct an annotated dataset of 6,000 tweets in both English and Chinese. Note if the answer to the second question of a tweet is N/A, then we consider the stance toward this tweet as none since it implies this tweet is irrelevant to the specific misconception that we focus on and therefore the stance is none, which is a common practice in this area (Hossain et al., 2020). The statistics of this dataset is shown in Table 6.2. I randomly split the 1,000 tweets in each of the misconception dataset in both English and Chinese into training data, validation data, and test data with a split of 760, 120, and 120 tweets, respectively.

Remarks. The crowdsourced annotators were compensated for their time. In particular, each Amazon Mechanical Turk worker was paid 3 U.S. dollers (USD) for annotating every 20 tweets, corresponding to around 12-15 USD per hour. When Amazon Mechanical Turk workers annotating English tweets, I did not explicitly tell them this project was for the academic use only while I explained explicitly to Chinese workers that this work was for research only. For Chinese workers, I negotiated with them for their wages individually. The wage ranged from 300-500 Chinese Yuan (CNY) for annotating every 1,000 Chinese tweets. The wages for both Amazon Mechanical Turk workers and Chinese annotators were appropriate. The Amazon Mechanical

Turk workers were not necessarily from one specific country while it was likely that all Chinese workers lived in China.

6.3 Methodology

In this section, I describe my experimental method. First, I describe the models I use, and then I show how I process the Twitter data. Finally, I explain how to choose the hyperparameters and how to evaluate the performance of my experiments.

6.3.1 NLP models

In this section I describe the NLP models that I conduct experiments on. I have several BERT based models as baselines, including XLM-R, XLM-T, COVID-Tweet, and BERTweet (Conneau et al., 2020; Barbieri et al., 2021; Müller et al., 2020; Nguyen et al., 2020). Though BERT (Devlin et al., 2018) and its variants achieve generally state-of-the-art performance on various NLP tasks (Wang et al., 2018), I include several non-BERT models for a comprehensive comparison, which include CNN, FastText, and BiLSTM (Kim, 2014; Joulin et al., 2016a; Schuster and Paliwal, 1997).

- **XLM-R**: XLM-Roberta is a transformer-based cross-lingual language model (Conneau et al., 2020). It is pretrained by corpus from 100 languages, and it is a state-of-the-art multi-lingual language model in various tasks, and widely used as baselines in multi-lingual tasks.

- **XLM-T**: On the basis of XLM-R base model, XLM-T is further pretrained on tweets in more than 30 languages (Barbieri et al., 2021), which is designed to multi-lingual tasks for corpus in Twitter. It also has a sentiment analysis version pretrained on 8 language tweets, called XLM-T-sentiment, which I use as well (only applicable to stance detection).

- **COVID-Tweet-BERT**: COVID-Tweet-BERT is a pretrained Model which is further pretrained on COVID related tweets (Müller et al., 2020). This model has two versions, and I use both of them as baselines, referred to as CTBv1 and CTBv2, respectively.

- **BERTweet**: BERTweet is a model trained for Tweet related tasks. I adopt the version that is further pretrained on COVID related tweets (Nguyen et al., 2020). I refer to this model as BTweet.

The non-BERT models include CNN, FastText, and BiLSTM.

- **CNN**: Convolutional neural network is first used for visual tasks (Krizhevsky et al., 2017). Then it is demonstrated to work in NLP tasks as well (Kim, 2014).

- **FastText**: FastText is developed by Facebook (Meta) as a text classification model (Joulin et al., 2016a; Joulin et al., 2016b), which I refer to as FT.

- **BiLSTM**: BiLSTM is a variant of LSTM model (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997). It is also a frequently used model for various NLP tasks.

All of the experiments are conducted in Google Colab pro plus. All the BERT models are implemented by using HuggingFace and non-BERT models are implemented by using PyTorch (Wolf et al., 2020; Paszke et al., 2019).

6.3.2 Approaches to process Chinese tweets

Since in my experiments I aim to perform cross-lingual zero-shot learning and some of the models are monolingual, I use two types of approaches to process Chinese tweets.

- **Original Chinese text with multi-lingual models**: The original Chinese tweets are experimented on XLM-R, XLM-T, and XLM-T-sentiment directly.

- **Automatic translation**: I use Google Cloud Translation API (Google, 2022) to convert the corresponding Chinese tweets into English and processed by all the models. Note although XLM-R, XLM-T, and XLM-T-sentiment can process original Chinese text directly, in my experiments, I experiment the translated Chinese tweets, as well as the original Chinese tweets, on them for a complete comparison. To avoid confusion, I refer to XLM-R, XLM-T, and XLM-T-sentiment as XLM-R-original, XLM-T-original, and XLM-T-sentiment-original, respectively when they process **original Chinese text** directly.

6.3.3 Data preprocessing

For each tweet in the dataset, I transfer the contractions into their full forms, remove emojis and URLs, and convert uppercase letters to lowercase letters. For non-BERT models, I use spaCy and Glove to tokenize tweets and convert them into vectors (spacy, 2022; Pennington et al., 2014). For BERT models, I tokenize and convert tweets into vectors following the instructions from their developers and truncate the length of tweet into 128 if necessary.

6.3.4 Hyperparameters

In this section, I discuss the hyperparameters that I select when training the various models. In general, these hyperparameters are selected following prior work in this area or default values (Glandt et al., 2021; Wolf et al., 2020; Joulin et al., 2016a; Trevett, 2021).

- **CNN**: The number of filters is 100 with a filter sizes of 2,3,4,5. The embedding dimensional number is 100 and the dropout is equal to 0.5.

- **FT**: The embedding dimensional number is 100.

- **BiLSTM**: The number of hidden dimensions, embedding dimensions and layers are 256, 100 and 2, respectively. The network has a dropout 0.5.
- **BERT based models**: All of the BERT based models are fine-tuned by using an AdamW optimizer with an initial learning rate $5 \times e^{-5}$.

The non-BERT models are trained for 120 epochs while the BERT models are trained for 15 epochs.

6.3.5 Evaluation

The training and validation data in the same language are used in the training epochs. The trained model then is evaluated on the test data in the same language, as well as on the test data in the other language. I use four metrics to assess the results: accuracy (acc), Macro average F1 (F1), Macro average precision (Pr), and Macro average Recall (Re), respectively. Each evaluation is based on the parameters that achieve the best F1 score in the training epoch. I run every experiment three times and report the average of the results of the experiments.

6.4 Results

In this section, I report the experimental results. The misinformation detection results are shown in Table 6.3 while the stance detection results are shown in Table 6.4. In Tables 6.3 and 6.4, I only present the models that achieve the best performance at least in one column among all the models in the same pair of training data and test data to highlight the models that achieve the best performance with regard to each specific metric. The full list of results is shown in Appendix D. I provide discussions for the two tasks in the following two subsections, respectively.

	Misinformation: Ginger/Garlic Train on English								Misinformation: Ginger/Garlic Train on Chinese							
	Train on English & Test on English				Train on English & Test on Chinese				Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Pr	Re	Acc	F1	Pre	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc
CTBv1	0.880	0.881	0.879	0.881	0.860	0.836	**0.904**	0.919	0.856	0.835	0.883	0.925	**0.878**	**0.884**	**0.876**	**0.879**
CTBv2	**0.886**	**0.889**	**0.888**	**0.886**	0.774	0.765	0.864	0.836	0.841	0.811	0.891	0.911	0.859	0.865	0.858	0.861
BTweet	**0.886**	0.886	0.887	**0.886**	**0.861**	**0.886**	0.840	**0.936**	**0.868**	0.835	**0.917**	**0.928**	0.837	0.861	0.834	0.842
FT	0.824	0.840	0.821	0.828	0.791	0.805	0.780	0.903	0.618	**0.940**	0.588	0.883	0.439	0.778	0.545	0.575
	Misinformation: Hydroxychloroquine Train on English								Misinformation: Hydroxychloroquine Train on Chinese							
	Train on English & Test on English				Train on English & Test on Chinese				Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Pr	Re	Acc	F1	Pre	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc
CTBv1	0.797	0.803	0.792	0.861	**0.811**	**0.833**	**0.800**	**0.839**	**0.824**	**0.834**	**0.817**	**0.847**	0.753	**0.835**	0.718	0.856
CTBv2	**0.825**	0.825	**0.825**	0.878	0.793	0.790	0.796	0.814	0.501	0.605	0.550	0.694	0.462	0.723	0.512	0.781
BTweet	0.817	**0.841**	0.800	**0.881**	0.757	0.803	0.740	0.803	0.800	0.812	0.792	0.828	**0.810**	0.823	**0.799**	**0.872**
	Misinformation: Bioweapon Train on English								Misinformation: Bioweapon Train on Chinese							
	Train on English & Test on English				Train on English & Test on Chinese				Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Pr	Re	Acc	F1	Pre	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc
CTBv1	**0.850**	0.871	**0.839**	0.827	**0.888**	**0.890**	**0.887**	**0.894**	0.561	0.516	0.637	0.719	0.516	0.469	0.583	0.683
CTBv2	0.840	**0.882**	0.822	**0.861**	0.861	0.861	0.864	0.867	0.726	0.720	0.756	0.811	0.664	0.683	0.688	0.767
BTweet	0.846	0.861	0.836	**0.861**	0.871	0.870	0.876	0.878	0.872	**0.892**	0.862	0.883	**0.791**	**0.858**	**0.773**	**0.825**
XLM-T	0.776	0.786	0.770	0.797	0.877	0.876	0.879	0.883	**0.880**	**0.892**	**0.872**	**0.889**	0.755	0.764	0.750	0.778

Table 6.3: Averages of results for misinformation detection corresponding to best performance models. Note XLM-T corresponds to the mode that processes the translated Chinese text. Please see Section 6.3.2 for more details.

6.4.1 Misinformation detection

From Table 6.3, we can see using automatic translation methods outperforms using original Chinese tweets processed by multi-lingual model methods in cross-lingual cases in general. Additionally, for most misconceptions, CTBv1, CTBv2, and BTweet can achieve the best performance and non-BERT models rarely perform best in terms of the four metrics. Therefore, in practice, it is highly recommended to use CTBv1, CTBv2, and BTweet. In addition, whether transferring from English to Chinese or from Chinese to English, the zero-shot cross-lingual performance are close to the performance of same language performance. With respect to misinformation detection, zero-shot learning can be used in practice in a bidirectional manner between English and Chinese tweets, highlighting potential uses when moderating multi-lingual content.

6.4.2 Stance detection

As can be seen in Table 6.4, the best performance achieved for the stance detection drops compared with the misinformation detection. Still, for most misconceptions, BTweet and CTBv2 achieve the best performance, both of which are recommended to use in practice. Another observation is that zero-shot learning is more effective when

	Stance: Ginger/Garlic Train on English								Stance: Ginger/Garlic Train on Chinese							
	Train on English & Test on English				Train on English & Test on Chinese				Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Pr	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc
CTBv1	0.693	0.712	0.682	0.775	**0.751**	**0.702**	**0.830**	0.889	0.472	0.443	0.517	0.886	0.521	0.554	0.530	0.711
CTBv2	**0.765**	**0.776**	**0.763**	**0.800**	0.712	0.699	0.741	0.875	0.650	**0.764**	0.649	**0.914**	**0.591**	0.687	0.573	**0.744**
BTweet	0.671	0.665	0.681	0.750	0.733	0.692	0.797	0.881	**0.657**	0.623	**0.708**	0.886	0.588	0.608	**0.581**	0.708
CNN	0.556	0.631	0.537	0.700	0.508	0.635	0.469	**0.897**	0.446	0.433	0.471	0.867	0.432	0.611	0.454	0.589
FT	0.627	0.765	0.581	0.711	0.392	0.419	0.383	0.861	0.360	0.521	0.358	0.897	0.338	**0.707**	0.380	0.628
	Stance: Hydroxychloroquine Train on English								Stance: Hydroxychloroquine Train on Chinese							
	Train on English & Test on English				Train on English & Test on Chinese				Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Pr	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc
CTBv1	0.772	0.772	0.777	0.775	0.702	0.710	**0.705**	**0.731**	0.562	0.544	0.605	0.667	0.455	0.441	0.510	0.528
CTBv2	**0.795**	**0.811**	**0.792**	**0.800**	**0.705**	**0.712**	**0.705**	0.728	0.363	0.364	0.414	0.536	0.302	0.273	0.373	0.408
BTweet	0.707	0.719	0.711	0.711	0.674	0.710	0.668	0.711	**0.669**	**0.711**	**0.659**	**0.742**	**0.620**	**0.640**	**0.629**	**0.625**
	Stance: Bioweapon Train on English								Stance: Bioweapon Train on Chinese							
	Train on English & Test on English				Train on English & Test on Chinese				Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Pr	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc	F1	Pr	Re	Acc
BTweet	**0.751**	**0.780**	0.739	**0.778**	**0.757**	**0.764**	**0.758**	**0.808**	**0.770**	**0.752**	**0.799**	**0.797**	**0.664**	**0.661**	**0.730**	**0.700**
XLM-T	0.744	0.734	**0.765**	0.767	0.639	0.663	0.635	0.708	0.618	0.622	0.641	0.703	0.599	0.585	0.677	0.619
XLM-T-Original	**0.751**	0.746	0.757	0.758	0.439	0.494	0.464	0.617	0.624	0.630	0.644	0.706	0.479	0.504	0.490	0.547

Table 6.4: Averages of results for stance detection corresponding to best performance models. Note XLM-T and XLM-T-Original correspond to the modes that process the translated and original Chinese text, respectively. Please see Section 6.3.2 for more details.

transfer from English to Chinese than vice versa. Although this is a drawback, it is still possible for zero-shot learning to be used in moderation platforms like Twitter since it is likely that moderators are more familiar with English than Chinese.

6.4.3 Error Analysis

I follow the practice in (Glandt et al., 2021) and conduct a qualitative error analysis to help readers better understand the results.

I choose the case in stance detection related to "ginger/garlic," and use one of the best models CTBv2 to demonstrate how the model performs. For each test tweet, it can be predicted by models trained on the same language as well as on the cross-lingual manner and all the Chinese tweets mentioned here are translated automatically.

Examples are shown in Table 6.5. Typically, CTBv2 performs well when the stance towards the misconception is none as can be seen in tweet No.1 and No.3. However, CTBv2 stumbles when the meaning of a tweet is vague. One such example is tweet No.2. The human annotators label it as a tweet supports the efficacy of garlic in treating COVID-19 probably because the tweet mentions *the effect is very good*, which the human annotators believe it refers to the effect of garlic. However, one can also

No.	Tweet	Label	$Pred_{zh\_zh}$	$Pred_{en\_zh}$
1	@(username) cry, cry, cry! why is there no iced one, i like it the most! minced garlic and egg yolks are also good! it's all because of the pandemic!	None	None	None
2	@(username) in fact, there are gauze materials that can be used to make masks by yourself.china has a population of 1.4 billion, which cannot be produced and consumes resources. it should teach people all over the country to make masks at home on tv. masks can be sandwiched, dry tea leaves or wormwood leaves/dried garlic chips wait, the effect is very good, and it can block virus droplets.	Support	None	None
No.	Tweet	Label	$Pred_{en\_en}$	$Pred_{zh\_en}$
3	ginger loves covid and rapists	None	None	None
4	my mom think ginger tea gone keep me from getting covid lmaaoo	Refute	Support	Support

Table 6.5: Error analysis for ginger/garlic stance examples. Tweets No.1 and 2 are translated from Chinese. I hide usernames to protect their privacy. $Pred_{zh\_zh}$, $Pred_{en\_zh}$, $Pred_{en\_en}$, and $Pred_{zh\_en}$ stand for the predicted label obtained in the "train on Chinese & test on Chinese", "train on English & test on Chinese", "train on English & test on English", and "train on Chinese & test on English" manners, respectively.

argue that this could refer to the effect of masks. Such controversial tweet prevents the model from predicting correctly. Another potential reason leads to an erroneous prediction is online slang and its variant. As seen in tweet No.4, lmaaoo is a variant of lmao (Dictionary, 2018), showing the author of the tweet that they do not believe ginger is a cure for COVID-19. The variant of this slang may be distant even to a pretrained model, making the model predict incorrectly.

6.5 Discussion

In this section, I discuss the implication, the limitations, potential risks, and privacy issues of this research.

6.5.1 Implications

The results of experiments show CTBv2 and BTweet, i.e., COVID-Tweet-BERT v2 and BERTweet, are generally capable of detecting misconceptions expressed in a tweet and detecting the stance of the author toward this misconception when used in both monolingual and multi-lingual manners. By applying these models, content moderators may pinpoint tweets that are likely to spread certain specific misconceptions, making

it appropriate for moderation at scale on Twitter and surveillance of suspicious tweets pertinent to misconceptions in low-resource language.

6.5.2 Limitations

I do not have a large-scale hyperparameter search and I use good ones suggested by previous work (Glandt et al., 2021; Wolf et al., 2020; Joulin et al., 2016a; Trevett, 2021). The implements are based on Google Colab Pro Plus, which is influenced by Google's policy of use, and GPUs are not always guaranteed to be the same. In particular, Google changes GPUs available to Google Colab Pro Plus on Sep 29, 2022. However, all the experiments using automatic translated Chinese tweets are conducted using nVidia V100 with a memory 16GB before Sep 29, 2022 and all the experiments using original Chinese tweets are conducted after Sep 29, 2022 and all the experiments are based on nVidia T4. This may affect the fairness of comparsion.

6.6 Conclusion

In this chapter, I compile an annotated English and Chinese Twitter dataset for COVID-19 misinformation and stance detection. On the basis of this dataset, I experiment on various models to investigate their performance on misinformation detection and stance detection in both monolingual and multi-lingual manners. In general, I find that by using automatic translation, the cross-lingual zero-shot learning performs better compared with original Chinese text processed by multi-lingual models. Among models that are studied in cross-lingual zero-shot learning manners, both COVID-Tweet-BERT v2 and BERTweet are promising to be effective in practical use. This sheds light on multi-lingual misinformation research for other languages in the future. However, multi-lingual models are still indispensable when one language is not applicable for automatic translation. Also, from the error analysis in Section 6.4.3, we can see the performance of the models may be easily influenced by the vagueness of

sentences and online slangs, which can inspire pretraining research for large language models, explainability research, and practical use of these models.

Chapter 7

Conclusion

In this chapter, I conclude my dissertation. First, I summarize the contributions of my research in each chapter. I then propose some potential directions for future research.

7.1 Contributions

In this dissertation, I focus on investigating misinformation from a multi-platform, multi-modal, and multi-lingual perspective. In particular, my research focuses on the measurement of the impact of misinformation and detection of misinformation. For the measurement research, I study how news stories, fauxtography, and COVID-19 misinformation images impact the social media. For the misinformation detection research, my research investigates how to detect multi-lingual misinformation and user stances in tweets. The detailed contributions are listed below.

News stories. In this chapter, I analyzed the sharing and the spreading of online news, showing that different communities present fundamental differences; for instance, Gab and /r/The_Donald "prefer" untrustworthy news sources (e.g., on Gab, 48.9% of all news URLs are from untrustworthy sources, compared to the 8.7% for Twitter). I also found that smaller Web communities can appreciably influence the news discussion on larger ones, with /r/The_Donald being very effective in pushing news stories on Twitter and the rest of Reddit.

Fauxtography. In this chapter, I presented a data-driven study of fauxtography on social media. I found that including both fauxtography images and images that are

fact-checked as true in social media posts increases user engagement. This highlights the need to take images into account when developing disinformation mitigations. At the same time, I showed that fauxtography images are often taken out of context and turned into memes, which highlights the challenges faced in automatically identifying image-based disinformation.

COVID-19 Image misinformation. In this chapter, I built a large dataset of misinformation images related to COVID-19 posted on Twitter between March and June 2020. I developed a codebook to characterize the various types of COVID-19 misinformation images related to the virus, from false medical advice to conspiracy theories. I then used this dataset to understand how COVID-19 misinformation images are used on social media. I found that these images do not receive more retweets and likes than tweets with random images. On the other hand, COVID-19 misinformation images are shared for longer periods of time than non-misinformation images on Twitter. I also found that COVID-19 misinformation images are shared by users who support the Democratic and the Republican party in similar numbers, but there is a difference in the type of images that the two groups share. While Democratic users often share misleading facts on the Trump administration's response to the pandemic, together with manipulated satirical images to critique this response, Republican users often share conspiracy theories about the origin of the virus, as well as images advocating for false treatment of hydroxychloroquine against COVID-19. My findings help researchers gain a better understanding of image-based misinformation on social media, and identify a number of challenges and opportunities for further research in this space.

COVID-19 multi-lingual misinformation on Twitter. In this chapter, I compiled an annotated English and Chinese Twitter dataset for COVID-19 misinformation and stance detection. On the basis of this dataset, I experimented on a variety of NLP

models to study their performance on misinformation detection and stance detection in both monolingual and multi-lingual manners. In general, I found that by using automatic translation, the cross-lingual zero-shot learning performs better compared with original Chinese text processed by multi-lingual models. Among models that were studied in cross-lingual zero-shot learning manners, both COVID-Tweet-BERT v2 and BERTweet are promising to be effective in practical use. This sheds light on multi-lingual misinformation research for other languages in the future. However, multi-lingual models are still indispensable when automatic translation is not available for one language. Also, from the error analysis in Section 6.4.3, we can see the performance of the models may be easily influenced by the vagueness of sentences and online slangs, which can inspire pretraining research for large language models, explainability research, and practical use of these models.

Overall, the results of this dissertation shed light on understanding of online misinformation, and my proposed computational tools are applicable to moderation of social media, potentially benefitting for a more wholesome online ecosystem.

7.2 Future directions.

There are several research directions left for future exploration.

News stories. In chapter 3, I proposed one method to cluster political news into stories. However, due to limitations in GDELT, that method is only effective on political news stories. This calls for clustering approaches that apply to generic news, including sports news, local news, entertainment news, and so on, with a satisfactory performance.

To achieve that goal, more advanced NLP techniques are potential candidates. For instance, by using proper prompts, GPT-3 may be able to determine whether two news articles belong to one news story. Alternatively, self supervising learning

methods may also be a good fit for this task.

I characterize the mutual influence among Web communities by using the raw influence and normalized influence of news articles regardless of their trustworthiness. This limits our understanding of the role of news sources in pushing news stories to distinct Web communities. Further analysis may be conducted to reveal to what extent trustworthy or untrustworthy news articles have on various Web communities. We may also pinpoint the most influential news outlet and the most efficient one in influencing Web communities.

Another interesting research topic is to determine the authenticity of statements in news articles. Currently, research in this field judges the trustworthiness of a news article by its source. This may not reflect the quality of a news article appropriately. One improvement would be checking the authenticity of the extracted statements in news articles, which may better characterize the genuineness of a news article.

Fauxtography. In Chapters 4 and 5, I investigated the engagement impact of fauxtography. Though fauxtography is misleading, it is still unknown to what extent social media users believe fauxtography is real. The trustworthiness of a fauxtography image may be influenced by factors including the source of this image, the background knowledge of a social media user, and so on. In particular, the topic category of fauxtography may play a role in the engagement of sharing fauxtography on social media platforms. For instance, fauxtography relevant to US presidential candidates may draw numerous attention during the US presidential elections.

This problem is difficult to solve by using mere computational approaches since it requires opinions from social media users. Thus, to better understand fauxtography, it may be interesting to conduct interviews with users from a diverse background, and by examining their reactions towards these fauxtography images, we can have a more thorough knowledge of this issue and understand the underlying reason why people

prefer to share fauxtography.

Another interesting research direction is to develop methods to automatically detect fauxtography, as well as general multi-modal misinformation. As shown in Chapter 2, previous research has devoted numerous efforts in investigating methods to detecting fauxtography. However, there is still space for further improvement of the detection of images with misinformation and we need to come up with more effective approaches to solve this issue.

Multi-lingual COVID-19 misinformation detection on Twitter. In Chapter 6, I established a Twitter dataset with two languages. By applying the same annotation method, it is possible to build a COVID-19 misinformation Twitter dataset with more languages. This may further benchmark the performance of various multi-lingual NLP models on cross-lingual misinformation and stance detection tasks. It may also help us better understand how misinformation spreads among people speaking different languages, and how these users respond to these misinformation topics.

In future, it would be interesting to investigate how to enhance the performance of NLP models on identifying COVID-19 misinformation on Twitter. One possible way is to use ChatGPT to determine whether one tweet contains a misconception or not with proper prompts. Other potential methods include designing novel large language models pretrained by relevant tweets and developing methods to better pinpoint misinformation in tweets.

Appendix A

Parameter Selection for Clustering

There are two parameters in DBSCAN algorithm to be determined: maximum distance ε and the minimum size of a cluster minPts (Schubert et al., 2017). The maximum distance ε determines the maximum distance (with an arbitrary distance measure) between two points that are considered neighbors for each other in one cluster. The minimum size of a cluster minPts specifies the minimum number of elements in a cluster. The members in Group with fewer elements than minPts are treated as noise points.

In practice, previous work applies a heuristic method to determine the parameter minPts and the minPts are set to default value 4 for two-dimensional data (See Section 4.2 of paper (Ester et al., 1996) and Section 4.1 of paper (Schubert et al., 2017)). In Section 4.1 of paper (Schubert et al., 2017), it is suggested that for datasets with high dimensions, it could improve results by increasing minPts. Since the vectors in my dataset are 64-bit binary vectors (i.e., vectors with sixty-four dimensions), and I determine the distance between two vectors by using Hamming distance (i.e., the number of bits that are different between two vectors), and I refer to the parameters

Distance	#Cluster	#Images clustered	%Noise
2	7,590	144,398	57.5%
4	7,674	146,612	56.5%
6	7,773	148,987	56.2%
8	7,854	152,368	55.2%
10	7,827	161,522	52.5%

Table A.1: Overview of cluster parameter performance I.

Distance	%Correctly grouped clusters
2	99.5%
4	99.5%
6	99.5%
8	97%
10	86%

Table A.2: Overview of cluster parameter performance II.

used in (Zannettou et al., 2018b), where minPts are set heuristically as 5, I decide to set the minPts as 5. Then I turn to determine the parameter ε.

Following a similar parameter selection method described in appendix A of paper (Zannettou et al., 2018b), I find that when clustering by varying the parameter ε, the percentage of noise does not change a lot. The result is shown in Table A.1.

Next, I randomly select 200 clusters for each threshold and manually check the clusters. I find that among candidates for distance 2, 4, and 6, 199 clusters are totally correctly clustered among 200 clusters for each threshold, while for distance 8 and 10, 194 and 172 clusters are totally correctly clustered, respectively. The proportion of correctly grouped clusters for each threshold is shown in Table A.2.

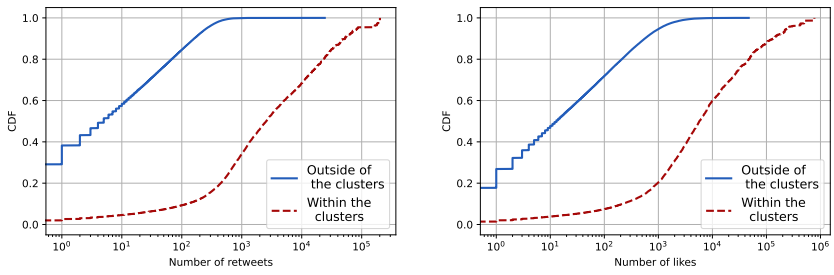
First, between the thresholds 8 and 10, I select 8 because it has a higher percentage of correctly grouped clusters. Then among thresholds 2, 4, and 6, I select 6 because threshold 6 has a lower percentage of noise while maintaining the same percentage of correctly grouped clusters. Finally, Between the final two threshold candidates 6 and 8, I select 6 because it has a higher percentage of correctly grouped clusters while the percentage of noise differ little between the two thresholds.

Appendix B

Evaluate popularity of images in the clusters vs images not in the clusters

In total, I have 339,891 tweets that contain images within 2.3M COVID-19-related tweets. As described in Section 5.4.1, I hydrate the original tweets, quoted tweets, and the original tweet part of the retweets in October 2020. In total, I have 1,424,481 unique tweets for hydration and after hydration, I obtain 1,272,929 unique tweets, among which I have 122,679 unique tweets containing images that are grouped into clusters, while I obtain 165,859 unique tweets containing images that are outside of the clusters. Note that compared with the hydration conducted in April 2021, which is used for investigating RQ1, the hydration conducted in October 2020 does not take general tweets into consideration. Therefore, I cannot use the hydrated tweets in October 2020 to answer RQ1.

I plot the CDFs of the two types of images in Figure B·1. The median for the



(a) Retweets (b) Likes

Figure B·1: Engagement of tweets that contain images within the clusters and tweets that contain images outside of the clusters, respectively.

retweets of tweets that contain images within the clusters and tweets that contain images outside of the clusters are 2,576 and 5, respectively while the median for the likes of tweets that contain images within the clusters and tweets that contain images outside of the clusters are 5,733 and 13, respectively. The further two-sample K-S tests show that differences between these two categories are statistically significant at the $p < 0.01$ level with $D = 0.75$ and $D = 0.81$ for retweets and likes, respectively. These results reject the null hypothesis that tweets containing images within the clusters receive the same level of engagement as tweets containing images outside of the clusters. In the cases of both likes and retweets, tweets that contain images within the clusters tend to have more engagement than tweets that contain images outside of the clusters. I conclude that tweets containing images within the clusters are more likely to generate more engagement than tweets containing images outside of the clusters.

Appendix C

Codebook for the misconception related to hydroxychloroquine in English

C.1 Overview

Welcome! In this task, you will see a short tweet during the COVID-19 pandemic. It will be related to hydroxychloroquine/chloroquine (HCQ for short). Please read the tweet carefully and label it based on the three questions asked.

C.2 Warning

We want to reward workers who take time to provide high quality answers, so we are trying to meet a $12-15/hour wage. We will cross-validate your answers with other workers. If you speed through or provide nonsense answers constantly, we will have to decline the HIT or block you on Mechanical Turk. We are very reluctant to do that so please pay attention.

C.3 Notes

1. Please make your decision on the meaning. Do not infer, imagine, or over-interpret.

2. Do not use specialized background knowledge (e.g., If the tweet mentioned someone without introducing his/her identity, you do not have to search for that person). Otherwise, the answer is not available (N/A).

3. Do not use information learned from the other tweets.

4. Please take the meaning of the hashtags, mentioned accounts into consideration when reading the content.

5. If the tweet contains non-English, please just decide based on the English part

6. After you submit your answer, you are more than welcome to continue work on the remaining HITs in this project! Thank you very much for your help.

C.4 Detailed Instructions & Examples

1. Does this tweet explicitly or implicitly talk about hydroxychloroquine/chloroquine (HCQ for short) as a treatment or potential treatment of COVID-19?

A. Yes

B. No (if no, please select "Not applicable" for the following two questions)

Instruction & Examples are shown in Table C.1 Add tables.

Instructions	Examples
You should answer "Yes" as long as the tweet mentions the information; that is, even if the tweet refutes the statement that hydroxychloroquine/chloroquine can treat COVID-19, you should still answer yes.	"Repeated studies show #Hydroxychloroquine doesnt work for #COVID19 patients"
You should answer No if the tweet does NOT mention hydroxychloroquine/chloroquine can treat COVID-19 or mentions hydroxychloroquine/chloroquine can treat other diseases.	Hydroxychloroquine is effective against non-resistant strains of Malaria. It has long been known to cause cardiac arrests, but generally thats better than malaria!

Table C.1: Instructions & examples for the first question.

Instructions	Examples
You should answer "Support" when the tweet support the use of hydroxychloroquine/chloroquine as an effective (or potentially effective) treatment of COVID-19 for the general public	We can go support to work; if you get the virus doctors should treat you with hydroxychloroquine. #COVID19
You should answer "Refute" when the tweet does NOT support the use of hydroxychloroquine/chloroquine as an effective (or potentially effective) treatment of COVID-19 for the general public	"Dr. Fauci, an immunologist & Trump's chief at NIAID, says hydroxychloroquine IS NOT effective in preventing coronavirus
You should answer "None" when the tweet has no clear attitude, just jokes around, or cites an objective description without commenting	"Dr. Brian Tysons First-Person Account of Treating COVID-19 with Hydroxychloroquine The Economic Standard"

Table C.2: Instructions & examples for the second question.

2. Considering the overall attitude of the author, does this tweet support or refute the use of hydroxychloroquine/chloroquine as an effective (or potentially effective) treatment of COVID-19 for the general public?

A. Support

B. Refute

C. None

D. Not applicable

Instruction & Examples are shown in Table C.2

3. Does this tweet associate the use/non-use of hydroxychloroquine/chloroquine and COVID-19 and some secret plots by powerful actors, such as governments, politicians, companies (e.g., pharmacies), public figures (e.g., Anthony Fauci or Bill Gates), or other organizations (e.g., CDC, FDA), etc.?

A. Yes

B. No

C. Not applicable

Instruction & Examples are shown in Table C.3

Instructions	Examples
You should answer Yes if the tweet associates the use/non-use of hydroxychloroquine/chloroquine and COVID-19 and some secret plots by powerful actors, such as governments, politicians, companies (e.g., pharmacies), public figures (e.g., Anthony Fauci or Bill Gates), or other organizations (e.g., CDC, FDA), etc.?	In case you are wondering why #Hydroxychloroquine isn't universally being used? Big Pharma makes no money.
You should answer No if there is no such association	"Patients with rheumatic disease who were taking #hydroxychloroquine had a lower risk of #COVID19 infection than patients taking other disease-modifying anti-rheumatic drugs"

Table C.3: Instructions & examples for the third question.

Appendix D

Full result tables

The full results of experiments in Chapter 6 are shown in this appendix.

Ginger/Garlic	Train on English & Test on English				Train on English & Test on Chinese			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.880	0.881	0.879	0.881	0.860	0.836	**0.904**	0.919
CTBv2	**0.886**	**0.889**	**0.888**	**0.886**	0.774	0.765	0.864	0.836
BTweet	**0.886**	0.886	0.887	0.886	**0.861**	**0.886**	0.840	**0.936**
XLM-T	0.829	0.832	0.828	0.831	0.815	0.808	0.824	0.908
XLM-R	0.838	0.839	0.837	0.839	0.791	0.756	0.870	0.875
CNN	0.804	0.819	0.802	0.808	0.757	0.745	0.772	0.875
FT	0.824	0.840	0.821	0.828	0.791	0.805	0.780	0.903
BiLSTM	0.778	0.791	0.777	0.783	0.722	0.743	0.706	0.875
XLM-T-Original	0.841	0.841	0.840	0.842	0.644	0.643	0.771	0.733
XLM-R-Original	0.816	0.816	0.816	0.817	0.714	0.687	0.795	0.817

Table D.1: Misinformation: Ginger/Garlic Train on English

Ginger/Garlic	Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.856	0.835	0.883	0.925	**0.878**	**0.884**	**0.876**	**0.879**
CTBv2	0.841	0.811	0.891	0.911	0.859	0.865	0.858	0.861
BTweet	**0.868**	0.835	**0.917**	**0.928**	0.837	0.861	0.834	0.842
XLM-T	0.823	0.800	0.855	0.906	0.784	0.810	0.784	0.792
XLM-R	0.802	0.801	0.812	0.903	0.756	0.804	0.757	0.769
CNN	0.774	0.850	0.734	0.908	0.590	0.765	0.629	0.653
FT	0.618	**0.940**	0.588	0.883	0.439	0.778	0.545	0.575
BiLSTM	0.771	0.760	0.785	0.883	0.736	0.746	0.735	0.741
XLM-T-Original	0.739	0.755	0.726	0.881	0.773	0.823	0.774	0.786
XLM-R-Original	0.829	0.803	0.865	0.908	0.676	0.786	0.690	0.708

Table D.2: Misinformation: Ginger/Garlic Train on Chinese

Hydroxychloroquine	Train on English & Test on English				Train on English & Test on Chinese			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.797	0.803	0.792	0.861	**0.811**	**0.833**	**0.800**	**0.839**
CTBv2	**0.825**	0.825	**0.825**	0.878	0.793	0.790	0.796	0.814
BTweet	0.817	**0.841**	0.800	**0.881**	0.757	0.803	0.740	0.803
XLM-T	0.786	0.822	0.763	0.864	0.684	0.720	0.677	0.747
XLM-R	0.758	0.794	0.739	0.847	0.625	0.682	0.627	0.711
CNN	0.725	0.759	0.707	0.825	0.725	0.740	0.719	0.767
FT	0.711	0.730	0.699	0.811	0.606	0.653	0.604	0.697
BiLSTM	0.755	0.762	0.748	0.833	0.730	0.754	0.719	0.775
XLM-Tweet-Original	0.782	0.773	0.793	0.842	0.729	0.733	0.760	0.739
XLM-R-Original	0.605	0.594	0.621	0.786	0.561	0.537	0.600	0.667

Table D.3: Misinformation: Hydroxychloroquine Train on English

Hydroxychloroquine	Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	**0.824**	**0.834**	**0.817**	**0.847**	0.753	**0.835**	0.718	0.856
CTBv2	0.501	0.605	0.550	0.694	0.462	0.723	0.512	0.781
BTweet	0.800	0.812	0.792	0.828	**0.810**	0.823	**0.799**	**0.872**
XLM-T	0.792	0.817	0.779	0.825	0.711	0.804	0.680	0.836
XLM-R	0.788	0.813	0.775	0.822	0.717	0.806	0.686	0.839
CNN	0.794	0.804	0.788	0.822	0.651	0.647	0.657	0.747
FT	0.654	0.719	0.646	0.736	0.526	0.674	0.545	0.783
BiLSTM	0.790	0.819	0.775	0.825	0.604	0.602	0.607	0.717
XLM-T-Original	0.786	0.818	0.771	0.822	0.591	0.723	0.585	0.792
XLM-R-Original	0.758	0.793	0.746	0.803	0.599	0.623	0.602	0.736

Table D.4: Misinformation: Hydroxychloroquine Train on Chinese

Bioweapon	Train on English & Test on English				Train on English & Test on Chinese			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	**0.850**	0.871	**0.839**	**0.827**	**0.888**	**0.890**	**0.887**	**0.894**
CTBv2	0.840	**0.882**	0.822	0.861	0.861	0.861	0.864	0.867
BTweet	0.846	0.861	0.836	0.861	0.871	0.870	0.876	0.878
XLM-T	0.776	0.786	0.770	0.797	0.877	0.876	0.879	0.883
XLM-R	0.648	0.659	0.671	0.744	0.706	0.761	0.738	0.783
CNN	0.758	0.822	0.743	0.796	0.745	0.763	0.740	0.771
FT	0.735	0.756	0.726	0.767	0.662	0.677	0.658	0.694
BiLSTM	0.780	0.811	0.767	0.808	0.634	0.636	0.644	0.642
XLM-T-Original	0.796	0.840	0.781	0.825	0.573	0.670	0.688	0.575
XLM-R-Original	0.514	0.486	0.582	0.686	0.557	0.556	0.584	0.697

Table D.5: Misinformation: Bioweapon Train on English

Bioweapon	Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.561	0.516	0.637	0.719	0.516	0.469	0.583	0.683
CTBv2	0.726	0.720	0.756	0.811	0.664	0.683	0.688	0.767
BTweet	0.872	**0.892**	0.862	0.883	**0.791**	**0.858**	**0.773**	**0.825**
XLM-T	**0.880**	**0.892**	**0.872**	**0.889**	0.755	0.764	0.750	0.778
XLM-R	0.829	0.868	0.818	0.850	0.721	0.743	0.715	0.756
CNN	0.846	0.850	0.843	0.856	0.698	0.704	0.695	0.725
FT	0.818	0.818	0.818	0.828	0.626	0.646	0.655	0.628
BiLSTM	0.804	0.807	0.802	0.817	0.608	0.607	0.613	0.625
XLM-T-Original	0.714	0.766	0.699	0.800	0.636	0.715	0.634	0.708
XLM-R-Original	0.539	0.519	0.583	0.764	0.478	0.478	0.552	0.669

Table D.6: Misinformation: Bioweapon Train on Chinese

Ginger/Garlic	Train on English & Test on English				Train on English & Test on Chinese			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.693	0.712	0.682	0.775	**0.751**	**0.702**	**0.830**	0.889
CTBv2	**0.765**	**0.776**	**0.763**	**0.800**	0.712	0.699	0.741	0.875
BTweet	0.671	0.665	0.681	0.750	0.733	0.692	0.797	0.881
XLM-T	0.642	0.645	0.642	0.747	0.671	0.636	0.731	0.888
XLM-T-sentiment	0.596	0.615	0.584	0.725	0.694	0.641	0.787	0.883
XLM-R	0.252	0.203	0.333	0.608	0.314	0.297	0.333	0.892
CNN	0.556	0.631	0.537	0.700	0.508	0.635	0.469	**0.897**
FT	0.627	0.765	0.581	0.711	0.392	0.419	0.383	0.861
BiLSTM	0.661	0.669	0.655	0.742	0.573	0.546	0.620	0.867
XLM-T-Original	0.680	0.686	0.691	0.769	0.679	0.635	0.787	0.822
XLM-T-sentiment-Original	0.550	0.614	0.555	0.683	0.311	0.364	0.449	0.567
XLM-R-Original	0.576	0.573	0.590	0.675	0.386	0.373	0.492	0.756

Table D.7: Stance: Ginger/Garlic Train on English

Ginge/Garlic	Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.472	0.443	0.517	0.886	0.521	0.554	0.530	0.711
CTBv2	0.650	**0.764**	0.649	**0.914**	**0.591**	0.687	0.573	**0.744**
BTweet	**0.657**	0.623	**0.708**	0.886	0.588	0.608	**0.581**	0.708
XLM-T	0.544	0.511	0.649	0.861	0.534	0.559	0.537	0.714
XLM-T-sentiment	0.601	0.779	0.583	0.892	0.535	0.609	0.518	0.725
XLM-R	0.314	0.297	0.333	0.892	0.252	0.203	0.333	0.608
CNN	0.446	0.433	0.471	0.867	0.432	0.611	0.454	0.589
FastText	0.360	0.521	0.358	0.897	0.338	**0.707**	0.380	0.628
BiLSTM	0.600	0.624	0.586	0.867	0.535	0.539	0.533	0.658
XLM-T-Original	0.473	0.459	0.494	0.844	0.314	0.276	0.374	0.631
XLM-T-sentiment-Original	0.534	0.563	0.512	0.883	0.588	0.684	0.556	0.725
XLM-R-Original	0.434	0.415	0.478	0.856	0.397	0.401	0.432	0.661

Table D.8: Stance: Ginger/Garlic Train on Chinese

Hydroxychloroquine	Train on English & Test on English				Train on English & Test on Chinese			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.772	0.772	0.777	0.775	0.702	0.710	**0.705**	**0.731**
CTBv2	**0.795**	**0.811**	**0.792**	**0.800**	**0.705**	**0.712**	**0.705**	0.728
BTweet	0.707	0.719	0.711	0.711	0.674	0.710	0.668	0.711
XLM-T	0.613	0.643	0.610	0.631	0.546	0.587	0.549	0.628
XLM-T-sentiment	0.563	0.568	0.565	0.575	0.494	0.509	0.501	0.542
XLM-R	0.190	0.133	0.333	0.400	0.207	0.150	0.333	0.450
CNN	0.516	0.539	0.517	0.531	0.437	0.467	0.448	0.500
FT	0.473	0.510	0.481	0.511	0.523	0.631	0.508	0.589
BiLSTM	0.451	0.460	0.457	0.483	0.461	0.468	0.465	0.508
XLM-T-Original	0.591	0.601	0.589	0.594	0.476	0.678	0.498	0.606
XLM-T-sentiment-Original	0.563	0.590	0.560	0.575	0.383	0.361	0.417	0.525
XLM-R-Original	0.495	0.494	0.536	0.561	0.401	0.447	0.443	0.536

Table D.9: Stance: Hydroxychloroquine Train on English

Hydroxychloroquine	Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.562	0.544	0.605	0.667	0.455	0.441	0.510	0.528
CTBv2	0.363	0.364	0.414	0.536	0.302	0.273	0.373	0.408
BTweet	**0.669**	**0.711**	**0.659**	**0.742**	**0.620**	**0.640**	**0.629**	**0.625**
XLM-T	0.614	0.632	0.618	0.672	0.474	0.549	0.489	0.483
XLM-T-sentiment	0.595	0.624	0.593	0.675	0.376	0.504	0.427	0.433
XLM-R	0.629	0.665	0.624	0.667	0.475	0.560	0.490	0.489
CNN	0.610	0.614	0.611	0.644	0.406	0.418	0.408	0.417
FT	0.620	0.706	0.627	0.644	0.368	0.542	0.444	0.422
BiLSTM	0.579	0.589	0.576	0.625	0.376	0.379	0.377	0.392
XLM-T-Original	0.659	0.670	0.654	0.700	0.335	0.434	0.417	0.403
XLM-T-sentiment-Original	0.590	0.594	0.594	0.625	0.284	0.302	0.372	0.383
XLM-R-Original	0.637	0.639	0.639	0.686	0.367	0.463	0.377	0.383

Table D.10: Stance: Hydroxychloroquine Train on Chinese

Bioweapon	Train on English & Test on English				Train on English & Test on Chinese			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.551	0.310	0.528	0.747	0.094	0.257	0.328	0.093
CTBv2	0.267	0.223	0.333	0.670	0.035	0.018	0.333	0.055
BTweet	**0.751**	**0.780**	0.739	**0.778**	**0.757**	**0.764**	**0.758**	**0.808**
XLM-T	0.744	0.734	**0.765**	0.767	0.639	0.663	0.635	0.708
XLM-T-sentiment	0.712	0.745	0.703	0.742	0.465	0.468	0.480	0.667
XLM-R	0.233	0.201	0.333	0.497	0.239	0.198	0.321	0.458
CNN	0.632	0.723	0.611	0.694	0.455	0.451	0.473	0.656
FT	0.536	0.634	0.535	0.697	0.394	0.380	0.415	0.567
BiLSTM	0.650	0.651	0.674	0.700	0.394	0.393	0.399	0.475
XLM-T-Original	**0.751**	0.746	0.757	0.758	0.439	0.494	0.464	0.617
XLM-T-sentiment-Original	0.728	0.768	0.703	0.767	0.379	0.429	0.430	0.575
XLM-R-Original	0.535	0.576	0.557	0.642	0.248	0.372	0.356	0.467

Table D.11: Stance: Bioweapon Train on English

Bioweapon	Train on Chinese & Test on Chinese				Train on Chinese & Test on English			
	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy
CTBv1	0.302	0.263	0.377	0.568	0.087	0.056	0.337	0.122
CTBv2	0.403	0.423	0.428	0.600	0.375	0.388	0.411	0.547
BTweet	**0.770**	**0.752**	**0.799**	**0.797**	**0.664**	**0.661**	**0.730**	**0.700**
XLM-T	0.618	0.622	0.641	0.703	0.599	0.585	0.677	0.619
XLM-T-sentiment	0.632	0.623	0.657	0.717	0.554	0.576	0.558	0.583
XLM-R	0.661	0.678	0.654	0.775	0.607	0.619	0.603	0.697
CNN	0.653	0.686	0.654	0.728	0.481	0.564	0.506	0.586
FT	0.487	0.491	0.517	0.703	0.426	0.788	0.458	0.561
BiLSTM	0.605	0.679	0.585	0.700	0.543	0.647	0.513	0.608
XLM-T-Original	0.624	0.630	0.644	0.706	0.479	0.504	0.490	0.547
XLM-T-sentiment-Original	0.659	0.653	0.666	0.692	0.501	0.498	0.536	0.525
XLM-R-Original	0.217	0.161	0.333	0.483	0.222	0.167	0.333	0.500

Table D.12: Stance: Bioweapon Train on Chinese

References

Lawrence Hurley (2018). U.S. top court upholds Trump travel ban targeting Muslim-majority nations. `https://www.reuters.com/article/us-usa-court-immigration/u-s-top-court-backs-trump-on-travel-ban-targeting-muslim-majority-nations-idUSKBN1JM1U9`.

Abidin, C. (2020). Meme factory cultures and content pivoting in Singapore and Malaysia during COVID-19. *Harvard Kennedy School Misinformation Review*.

Ajao, O., Hong, J., and Liu, W. (2015). A survey of location inference techniques on twitter. *Journal of Information Science*, 41(6):855–864.

Akyürek, A. F., Guo, L., Elanwar, R., Ishwar, P., Betke, M., and Wijaya, D. T. (2020). Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G. D. S., Shaar, S., Firooz, H., and Nakov, P. (2021). A survey on multimodal disinformation detection. *arXiv preprint arXiv:2103.12541*.

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Martino, G. D. S., Abdelali, A., Durrani, N., Darwish, K., et al. (2020). 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*.

Ali, S., Razi, A., Kim, S., Alsoubai, A., Gracie, J., De Choudhury, M., Wisniewski, P. J., and Stringhini, G. (2022). Understanding the digital lives of youth: Analyzing media shared within safe versus unsafe private conversations on instagram. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Aliapoulios, M., Papasavva, A., Ballard, C., De Cristofaro, E., Stringhini, G., Zannettou, S., and Blackburn, J. (2021). The gospel according to q: Understanding the qanon conspiracy from the perspective of canonical information. *arXiv preprint arXiv:2101.08750*.

Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2).

Alvari, H. and Shakarian, P. (2019). Hawkes Process for Understanding the Influence of Pathogenic Social Media Accounts. In *arXiv:1902.01970*.

Amy Lange (2016). Detroit family caught in iraq travel ban, says mom died waiting to come home. `https://www.fox5dc.com/news/detroit-family-caught-in-iraq-travel-ban-says-mom-died-waiting-to-come-home`.

Babaei, M., Kulshrestha, J., Chakraborty, A., Benevenuto, F., Gummadi, K. P., and Weller, A. (2018). Purple feed: Identifying high consensus news posts on social media. In *AIES*.

Backstrom, L., Sun, E., and Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70.

Banai, I. P., Banai, B., and Mikloušić, I. (2020). Beliefs in covid-19 conspiracy theories predict lower level of compliance with the preventive measures both directly and indirectly by lowering trust in government medical officials.

Barbieri, F., Anke, L. E., and Camacho-Collados, J. (2021). Xlm-t: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The Pushshift Reddit Dataset. In *ICWSM*.

Bayar, B. and Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. In *ACM IH*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan).

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

Boudemagh, E. and Moise, I. (2017). News media coverage of refugees in 2016: A gdelt case study. In *ICWSM*.

Bozarth, L. and Budak, C. (2020). Toward a better performance evaluation framework for fake news classification. In *ICWSM*.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Brennen, J. S., Simon, F., Howard, P. N., and Nielsen, R. K. (2020). Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 7(3):1.

Brennen, J. S., Simon, F. M., and Nielsen, R. K. (2021). Beyond (mis) representation: visuals in covid-19 misinformation. *The International Journal of Press/Politics*, 26(1):277–299.

Buchner, J. (2020). A python perceptual image hashing module: Imagehash. `https://github.com/JohannesBuchner/imagehash`. Accessed: 2021-04-08.

Budak, C. (2019). What happened? the spread of fake news publisher content during the 2016 us presidential election. In *The WebConf*.

Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *WWW*.

Chadwick, A. (2011). The hybrid media system. In *ECPR*.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Chen, E., Lerman, K., and Ferrara, E. (2020a). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Chen, K., Chen, A., Zhang, J., Meng, J., and Shen, C. (2020b). Conspiracy and debunking narratives about covid-19 origination on chinese social media: How it started and who is to blame. *arXiv preprint arXiv:2011.08409*.

Chen, K., Duan, Z., and Yang, S. (2022a). Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences*, 41(1):114–130.

Chen, N., Chen, X., Zhong, Z., and Pang, J. (2022b). " double vaccinated, 5g boosted!": Learning attitudes towards covid-19 vaccination from social media. *arXiv preprint arXiv:2206.13456*.

Chicago Tribune (2016). Trump revealed highly classified information to Russian diplomats, U.S. officials say. `http://www.chicagotribune.com/news/nationworld/ct-trump-revealed-classified-information-russians-20170515-story.html`.

Common Crawl Repository (2019).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011a). Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 192–199. IEEE.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011b). Political polarization on twitter. In *ICWSM*.

Cooper, S. (2007). A Concise History of the Fauxtography Blogstorm in the 2006 Lebanon War. *American Communication Journal*, 9.

Cortis, K. and Davis, B. (2021). A dataset of multidimensional and multilingual social opinions for maltas annual government budget. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 971–981.

Daisuke Wakabayashi, Davey Alba, and Marc Tracy (2020). Bill Gates, at Odds With Trump on Virus, Becomes a Right-Wing Target. `https://www.nytimes.com/2020/04/17/technology/bill-gates-virus-conspiracy-theories.html`.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dewan, P., Suri, A., Bharadhwaj, V., Mithal, A., and Kumaraguru, P. (2017). Towards Understanding Crisis Events On Online Social Networks Through Pictures. In *ASONAM*.

Dictionary, U. (2018). `https://www.urbandictionary.com/define.php?term=lmao`.

Du, Y., Masood, M. A., and Joseph, K. (2020). Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *ICWSM*.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*.

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, 22(8).

Ferrara, E. (2020). What types of covid-19 conspiracies are populated by twitter bots? *arXiv preprint arXiv:2004.09531*.

Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., and Moran, S. (2018). Falling for fake news: investigating the consumption of news via social media. In *ACM CHI*.

Flores-Saviaga, C. I., Keegan, B. C., and Savage, S. (2018). Mobilizing the trump train: Understanding collective action in a political trolling community. In *ICWSM*.

Freeman, D., Waite, F., Rosebrock, L., Petit, A., Causier, C., East, A., Jenner, L., Teale, A.-L., Carr, L., Mulhall, S., et al. (2020). Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in england. *Psychological medicine*, pages 1–13.

Fung, Y. R., Huang, K.-H., Nakov, P., and Ji, H. (2022). The battlefront of combating misinformation and coping with media bias. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4790–4791.

Garber, M. (2017). Al Franken, That Photo, and Trusting the Women. `https://www.theatlantic.com/entertainment/archive/2017/11/al-franken-thatc-and-trusting-the-women/545954/`. Accessed: 2021-04-08.

Garimella, K. and Eckles, D. (2020). Images and misinformation in political groups: Evidence from whatsapp in india. *arXiv:2005.09784*.

Garimella, K., Smith, T., Weiss, R., and West, R. (2021). Political polarization in online news consumption. *arXiv:2104.06481*.

GDELT (2015). The GDELT Event Database Data Format Codebook V2.0.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2).

Gentzkow, M., Shapiro, J. M., and Stone, D. F. (2015). Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1.

Glandt, K., Khanal, S., Li, Y., Caragea, D., and Caragea, C. (2021). Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1.

Google (2022). `https://cloud.google.com/translate`.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425).

Guo, F., Blundell, C., Wallach, H., and Heller, K. (2015). The bayesian echo chamber: Modeling social influence via linguistic accommodation. In *Artificial Intelligence and Statistics*.

Guo, L., Mays, K., Lai, S., Jalal, M., Ishwar, P., and Betke, M. (2020). Accurate, fast, but not always cheap: Evaluating crowdcoding as an alternative approach to analyze social media data. *Journalism & Mass Communication Quarterly*, 97(3):811–834.

Guo, L. and Vargo, C. (2020). fake news and emerging online media ecosystem: An integrated intermedia agenda-setting analysis of the 2016 us presidential election. *Communication Research*, 47(2):178–200.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., and Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, 93(2):332–359.

Guo, L. and Zhang, Y. (2020). Information flow within and across online media platforms: An agenda-setting analysis of rumor diffusion on news websites, weibo, and wechat in china. *Journalism Studies*, 21(15):2176–2195.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Hardalov, M., Arora, A., Nakov, P., and Augenstein, I. (2021). A survey on stance detection for mis-and disinformation identification. *arXiv preprint arXiv:2103.00242*.

Hardalov, M., Arora, A., Nakov, P., and Augenstein, I. (2022). Few-shot cross-lingual stance detection with sentiment-based pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36.

Harder, R. A., Sevenans, J., and Van Aelst, P. (2017). Intermedia agenda setting in the social media age: How traditional players dominate the news agenda in election times. *The International Journal of Press/Politics*, 22(3):275–293.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*.

Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., and Blackburn, J. (2017). Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S., and Singh, S. (2020). Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Hou, Y., van der Putten, P., and Verberne, S. (2022). The covmis-stance dataset: Stance detection on twitter for covid-19 misinformation. *arXiv preprint arXiv:2204.02000*.

Hu, Y., Huang, H., Chen, A., and Mao, X.-L. (2020). Weibo-cov: A large-scale covid-19 social media dataset from weibo. *arXiv preprint arXiv:2005.09174*.

Infowars (2018). Mexico Agrees to Pay for Wall. `https://www.infowars.com/mexico-agrees-to-pay-for-wall/`.

Javed, R. T., Shuja, M. E., Usama, M., Qadir, J., Iqbal, W., Tyson, G., Castro, I., and Garimella, K. (2020). A first look at covid-19 messages on whatsapp in pakistan. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 118–125. IEEE.

Javed, R. T., Usama, M., Iqbal, W., Qadir, J., Tyson, G., Castro, I., and Garimella, K. (2022). A deep dive into covid-19-related messages on whatsapp in pakistan. *Social Network Analysis and Mining*, 12(1):1–16.

Jiang, S., Metzger, M., Flanagin, A., and Wilson, C. (2020). Modeling and measuring expressed (dis) belief in (mis) information. In *ICWSM*.

JOHN HAYWARD (2017). Seven Inconvenient Facts About Trumps Refugee Actions. `https://www.breitbart.com/politics/2017/01/29/trumps-immigration-pause-sober-defenses-vs-hysterical-criticism/`.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016a). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016b). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Kar, D., Bhardwaj, M., Samanta, S., and Azad, A. P. (2020). No rumours please! a multi-indic-lingual approach for covid fake-tweet detection. In *2021 Grace Hopper Celebration India (GHCI)*, pages 1–5. IEEE.

Kazemi, A., Garimella, K., Gaffney, D., and Hale, S. A. (2021). Claim matching beyond english to scale global fact-checking. *arXiv preprint arXiv:2106.00853*.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C. A., Bull, P., Lipstein, G., Nelli, T., Zhu, R., et al. (2021). The hateful memes challenge: competition report. In *NeurIPS 2020 Competition and Demonstration Track*, pages 344–360. PMLR.

Kim, M. G., Kim, M., Kim, J. H., and Kim, K. (2022). Fine-tuning bert models to classify misinformation on garlic and covid-19 on twitter. *International Journal of Environmental Research and Public Health*, 19(9):5126.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kong, Q. (2019). Linking Epidemic Models and Hawkes Point Processes for Modeling Information Diffusion. In *WSDM*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Küçük, D. and Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., and Karahalios, K. (2019). Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 22(1):188–227.

Kumar, S. and Shah, N. (2018). False information on web and social media: A survey. In *arXiv:1804.08559*.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Lai, M., Cignarella, A. T., Farías, D. I. H., Bosco, C., Patti, V., and Rosso, P. (2020). Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Lazer, D., Ruck, D. J., Quintana, A., Shugars, S., Joseph, K., Grinberg, N., Gallagher, R. J., Horgan, L., Gitomer, A., Bajak, A., et al. (2021). The covid states project# 18: Fake news on twitter.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, C., Yang, T., Inchoco, G. D., Jones, G. M., and Satyanarayan, A. (2021). Viral visualizations: How coronavirus skeptics use orthodox data practices to promote unorthodox science online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Leetaru, K. and Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979-2012. In *ISA Annual Convention*.

Leng, Y., Zhai, Y., Sun, S., Wu, Y., Selzer, J., Strover, S., Zhang, H., Chen, A., and Ding, Y. (2021). Misinformation during the covid-19 outbreak in china: cultural, social and political entanglements. *IEEE Transactions on Big Data*, 7(1):69–80.

Leskovec, J., Backstrom, L., and Kleinberg, J. M. (2009). Meme-tracking and the Dynamics of the News Cycle. In *KDD*.

Li, Y. and Xie, Y. (2020). Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19.

Lin, H., Ma, J., Chen, L., Yang, Z., Cheng, M., and Chen, G. (2022). Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. *NAACL 2022*.

Linderman, S. W. and Adams, R. P. (2014). Discovering Latent Network Structure in Point Process Data. In *ICML*.

Linderman, S. W. and Adams, R. P. (2015). Scalable Bayesian Inference for Excitatory Point Process Networks. In *arXiv:1507.03228*.

Lindgren, B. (1993). *Statistical Theory*, volume 22.

Ling, C., AbuHilal, I., Blackburn, J., De Cristofaro, E., Zannettou, S., and Stringhini, G. (2021). Dissecting the meme magic: Understanding indicators of virality in image memes. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–24.

Ling, C., Blackburn, J., De Cristofaro, E., and Stringhini, G. (2022). Slapping cats, bopping heads, and oreo shakes: Understanding indicators of virality in tiktok short videos. In *14th ACM Web Science Conference 2022*, pages 164–173.

Lucas Ou-Yang (2020). Newspaper3k: Article scraping & curation. `https://newspaper.readthedocs.io/en/latest/`. Accessed: 2021-04-08.

Luceri, L., Giordano, S., and Ferrara, E. (2020). Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 417–427.

Lukasik, M., Srijith, P., Vu, D., Bontcheva, K., Zubiaga, A., and Cohn, T. (2016). Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *ACL*.

Ma, J., Gao, W., and Wong, K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Ma, Q. and Olshevsky, A. (2020). Adversarial crowdsourcing through robust rank-one matrix completion. *Advances in Neural Information Processing Systems*, 33:21841–21852.

Majestic (2019). The Majestic Million List. `https://majestic.com/reports/majestic-million`. Accessed: 2019-01-28.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *ACL*.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*.

Marcelino, G., Semedo, D., Mourão, A., Blasi, S., Mrak, M., and Magalhaes, J. (2019). A benchmark of visual storytelling in social media. In *ICMR*.

McClatchy DC Bureau (2017). House majority leader told his colleagues in 2016: 'I think Putin pays' Trump. `https://www.mcclatchydc.com/news/politics-government/article151133157.html`.

McGee, J., Caverlee, J., and Cheng, Z. (2013). Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 459–468.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

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

Mercia, D. (2017). Trump gets 2 scoops of ice cream, everyone else gets 1 – and other top lines from his Time interview. `https://www.cnn.com/2017/05/11/politics/trump-time-magazine-ice-cream/index.html`. Accessed: 2021-04-08.

Micallef, N., He, B., Kumar, S., Ahamad, M., and Memon, N. (2020). The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. *arXiv preprint arXiv:2011.05773*.

Miller, J. M. (2020). Do covid-19 conspiracy theory beliefs form a monological belief system? *Canadian Journal of Political Science/Revue canadienne de science politique*, 53(2):319–326.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Monga, V. and Evans, B. L. (2006). Perceptual image hashing via feature points: performance evaluation and tradeoffs. *IEEE transactions on Image Processing*, 15(11).

Motta, M., Stecula, D., and Farhart, C. (2020). How right-leaning media coverage of covid-19 facilitated the spread of misinformation in the early stages of the pandemic in the us. *Canadian Journal of Political Science/Revue canadienne de science politique*, 53(2):335–342.

Müller, M., Salathé, M., and Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Mutlu, E. C., Oghaz, T., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., Ozmen, O., and Garibay, I. (2020). A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19. *Data in brief*, 33:106401.

Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Mansour, W., et al. (2021). Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 264–291. Springer.

New York Post (2017). Trump's win in Wisconsin confirmed after vote recount. `https://nypost.com/2016/12/12/trumps-win-in-wisconsin-confirmed-after-vote-recount/`.

NewsGuard (2019a). Inside NewsGuard's First Year Fighting Misinformation.

NewsGuard (2019b). Rating Process and Criteria.

NewsGuard (2019c). Sample nutrition labels.

NewsGuard (2019d). The Internet Trust Tool. `https://www.newsguardtech.com/`. Accessed: 2021-04-08.

Ng, L. H. X. and Carley, K. M. (2022). Is my stance the same as your stance? a cross validation study of stance detection datasets. *Information Processing & Management*, 59(6):103070.

Ng, L. H. X., Moffitt, J., and Carley, K. M. (2022). Coordinated through aweb of images: Analysis of image-based influence operations from china, iran, russia, and venezuela. *arXiv preprint arXiv:2206.03576*.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Nic Fildes, Mark Di Stefano, and Hannah Murphy (2020). How a 5g coronavirus conspiracy spread across europe. `https://www.ft.com/content/1eeedb71-d9dc-4b13-9b45-fcb7898ae9e1`.

Nørregaard, J., Horne, B. D., and Adalı, S. (2019). Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 630–638.

Oliver, J. E. and Wood, T. (2014). Medical conspiracy theories and health behaviors in the united states. *JAMA internal medicine*, 174(5):817–818.

Papasavva, A., Blackburn, J., Stringhini, G., Zannettou, S., and Cristofaro, E. D. (2021). is it a qoincidence?: An exploratory study of qanon on voat. In *Proceedings of the Web Conference 2021*, pages 460–471.

Papasavva, A., Zannettou, S., De Cristofaro, E., Stringhini, G., and Blackburn, J. (2020). Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *ICWSM*.

Park, C. Y., Yan, X., Field, A., and Tsvetkov, Y. (2020). Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. *arXiv preprint arXiv:2010.10820*.

Park, C. Y., Yan, X., Field, A., and Tsvetkov, Y. (2021). Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. In *ICWSM*, pages 479–490.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Paudel, P., Blackburn, J., De Cristofaro, E., Zannettou, S., and Stringhini, G. (2022). Lambretta: Learning to rank for twitter soft moderation. *arXiv preprint arXiv:2212.05926*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *EMNLP*.

Pennycook, G. and Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *PNAS*, 116(7).

Pennycook, G. and Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*.

Pfeffer, J., Mayer, K., and Morstatter, F. (2018). Tampering with twitters sample api. *EPJ Data Science*, 7(1):50.

Phillips, S. C., Ng, L. H. X., and Carley, K. M. (2022). Hoaxes and hidden agendas: A twitter conspiracy theory dataset: Data paper. In *Companion Proceedings of the Web Conference 2022*, pages 876–880.

Poddar, S., Mondal, M., Misra, J., Ganguly, N., and Ghosh, S. (2022). Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 782–793.

Poynter (2020). `https://www.poynter.org/?ifcn_misinformation=studies-show-the-coronavirus-was-engineered-to-be-a-bioweapon`.

Qu, J., Li, L. H., Zhao, J., Dev, S., and Chang, K.-W. (2022). Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. In *WWW Companion*.

Ratkiewicz, J., Conover, M., Meiss, M. R., Goncalves, B., Patil, S., Flammini, A., and Menczer, F. (2010). Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. In *arXiv:1011.3768*.

Reimann, N. (2020). Some americans are tragically still drinking bleach as a coronavirus cure. `https://www.forbes.com/sites/nicholasreimann/2020/08/24/some-americans-are-tragically-still-drinking-bleach-as-a-coronavirus-cure/?sh=421223aa6748`.

Reis, J. C., Melo, P., Garimella, K., Almeida, J. M., Eckles, D., and Benevenuto, F. (2020). A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 903–908.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019). (mis) information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference*, pages 818–828.

Resnick, P., Ovadya, A., and Gilchrist, G. (2018). Iffy quotient: A platform health metric for misinformation.

Rivers, C. M. and Lewis, B. L. (2014). Ethical research standards in a world of big data. *F1000Research*.

Rye, E., Blackburn, J., and Beverly, R. (2020). Reading In-Between the Lines: An Analysis of Dissenter. In *ACM IMC*.

Saeed, M. H., Ali, S., Blackburn, J., De Cristofaro, E., Zannettou, S., and Stringhini, G. (2022). Trollmagnifier: Detecting state-sponsored troll accounts on reddit. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2161–2175. IEEE.

Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S. D., and Vallina-Rodriguez, N. (2018). A long way to the top: significance, structure, and stability of internet top lists. In *ACM IMC*.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

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

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1).

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019). defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1).

Soni, S., Ramirez, S. L., and Eisenstein, J. J. (2019). Detecting Social Influence in Event Cascades by Comparing Discriminative Rankers. In *SIGKDD Workshop on Causal Discovery*.

Soroka, S., Fournier, P., and Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *PNAS*, 116(38).

spaCy (2019). Industrial-Strength Natural Language Processing. `https://spacy.io/`.

spaCy (2019). Named Entity Recognition.

spacy (2022). `https://spacy.io/`.

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*.

Stringhini, G., Mourlanne, P., Jacob, G., Egele, M., Kruegel, C., and Vigna, G. (2015). Evilcohort: Detecting communities of malicious accounts on online services. In *24th USENIX Security Symposium (USENIX Security 15)*.

Tahmasbi, F., Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y., and Zannettou, S. (2021). go eat a bat, chang!: On the emergence of sinophobic behavior on web communities in the face of covid-19. In *Proceedings of the Web Conference 2021*, pages 1122–1133.

Tasnim, S., Hossain, M. M., and Mazumder, H. (2020). Impact of rumors and misinformation on covid-19 in social media. *Journal of preventive medicine and public health*, 53(3):171–174.

Taulé, M., Martí, M. A., Rangel, F. M., Rosso, P., Bosco, C., Patti, V., et al. (2017). Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*, volume 1881, pages 157–177. CEUR-WS.

The Computational Event Data System (2014). Dictionaries. `http://eventdata.parusanalytics.com/software.dir/dictionaries.html`.

The Daily Caller (2016). SOURCES: China Hacked Hillary Clinton's Private E-mail Server. `https://dailycaller.com/2018/08/27/china-hacked-clinton-server`.

The New York Times (2017). Judge Blocks Trump Order on Refugees Amid Chaos and Outcry Worldwide. `https://www.nytimes.com/2017/01/28/us/refugees-detained-at-us-airports-prompting-legal-challenges-to-trumps-immigration-order.html`.

The New York Times (2020). Reddit, acting against hate speech, bans 'the_donald' subreddit. `https://www.nytimes.com/2020/06/29/technology/reddit-hate-speech.html`.

The Washington Post (2016). Trump supporter charged with voting twice in Iowa. `https://www.washingtonpost.com/news/post-nation/wp/2016/10/29/trump-supporter-charged-with-voting-twice-in-iowa/`.

Times, N. Y. (2017). "resist" is a battle cry, but what does it mean? `https://www.nytimes.com/2017/02/14/us/politics/resist-anti-trump-protest.html`.

Trevett, B. (2021). Pytorch sentiment analysis. `https://github.com/bentrevett/pytorch-sentiment-analysis`.

Ullah, A., Das, A., Das, A., Kabir, M. A., and Shu, K. (2021). A survey of covid-19 misinformation: Datasets, detection techniques and open issues. *arXiv preprint arXiv:2110.00737*.

Uscinski, J. E., Enders, A. M., Klofstad, C., Seelig, M., Funchion, J., Everett, C., Wuchty, S., Premaratne, K., and Murthi, M. (2020). Why do people believe covid-19 conspiracy theories? *Harvard Kennedy School Misinformation Review*, 1(3).

Vamvas, J. and Sennrich, R. (2020). X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*.

Van Hoozer, S. and Peuchaud, S. (2020). "Speaking of Sexual Harassers Who Should Resign Tomorrow... Donald Trump": A Feminist Rhetorical Analysis of Stephen Colbert's Late Show Monologues. *The Journal of Popular Culture*, 53(1):34–57.

van Prooijen, J.-W. and Douglas, K. M. (2018). Belief in conspiracy theories: Basic principles of an emerging research domain. *European journal of social psychology*, 48(7):897–908.

VirusTotal (2020). VirusTotal. `https://www.virustotal.com`. Accessed: 2021-04-08.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, J. and Paschalidis, I. C. (2016). Botnet detection based on anomaly and community detection. *IEEE Transactions on Control of Network Systems*, 4(2):392–404.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *arXiv:1705.00648*.

Wang, Y., Tamahsbi, F., Blackburn, J., Bradlyn, B., De Cristofaro, E., Magerman, D., Zannettou, S., and Stringhini, G. (2021). Understanding the use of fauxtography on social media. In *International Conference on Web and Social Media*.

Weinzierl, M., Hopfer, S., and Harabagiu, S. M. (2021). Misinformation adoption or rejection in the era of covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), AAAI Press*.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Sameer Pradhan, L. R., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., and Houston, A. (2019). OntoNotes Release 5.0. `https://catalog.ldc.upenn.edu/LDC2013T19`.

Welbers, K. (2016). *Gatekeeping in the Digital Age*. PhD thesis.

WHO (2020). `https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-food-safety-and-nutrition`.

WHO (2021). `https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-hydroxychloroquine`.

Wikipedia (2021). `https://en.wikipedia.org/wiki/Inverted_pyramid_(journalism)`.

Wikipedia (2022). `https://en.wikipedia.org/wiki/Conspiracy_theory`.

Wilson, T. and Starbird, K. (2020). Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 1(1).

Wilson, T., Zhou, K., and Starbird, K. (2018). Assembling strategic narratives: Information operations as collaborative work within an online community. In *CSCW*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wong, J. C. (2018). What is qanon? explaining the bizarre rightwing conspiracy theory. *The Guardian*.

Wray, M. (2020). corona challenge: Tiktok star films herself licking airplane toilet seat. `https://globalnews.ca/news/6718358/tiktok-toilet-seat-lick-coronavirus/`.

Wu, L. and Liu, H. (2018). Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *WSDM*.

Wu, L., Morstatter, F., Carley, K. M., and Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.

Zannettou, S. (2021). " i won the election!": An empirical analysis of soft moderation interventions on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 865–876.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., and Blackburn, J. (2018a). What is gab: A bastion of free speech or an alt-right echo chamber. In *The WebConf Companion*.

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Suarez-Tangil, G. (2018b). On the origins of memes by means of fringe web communities. In *ACM IMC*.

Zannettou, S., Caulfield, T., Bradlyn, B., De Cristofaro, E., Stringhini, G., and Blackburn, J. (2020a). Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter. In *ICSWM*.

Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G., and Blackburn, J. (2017). The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM IMC*.

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Blackburn, J. (2019a). Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *The WebConf Companion*.

Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., and Blackburn, J. (2019b). Who Let The Trolls Out?: Towards Understanding State-Sponsored Trolls. In *WebSci*.

Zannettou, S., Finkelstein, J., Bradlyn, B., and Blackburn, J. (2020b). A quantitative approach to understanding online antisemitism. In *ICWSM*.

Zhang, D. Y., Shang, L., Geng, B., Lai, S., Li, K., Zhu, H., Amin, M. T., and Wang, D. (2018). Fauxbuster: A content-free fauxtography detector using social media comments. In *IEEE Big Data*.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*.

Zheng, X., Han, J., and Sun, A. (2018). A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671.

Zhou, K., Zha, H., and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*.

Zhou, X., Mulay, A., Ferrara, E., and Zafarani, R. (2020). Recovery: A multimodal repository for covid-19 news credibility research. *arXiv:2006.05557*.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Ziems, C., He, B., Soni, S., and Kumar, S. (2020). Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*.

Zillmann, D., Gibson, R., and Sargent, S. L. (1999). Effects of photographs in news-magazine reports on issue perception. *Media Psychology*, 1(3).

Zillmann, D., Knobloch, S., and Yu, H.-s. (2001). Effects of photographs on the selective reading of news reports. *Media Psychology*, 3(4).

Zlatkova, D., Nakov, P., and Koychev, I. (2019). Fact-checking meets fauxtography: Verifying claims about images. In *EMNLP-IJCNLP*.

CURRICULUM VITAE

Yuping Wang

7.1 Contact Information

Division of Systems Engineering
College of Engineering
Boston University
8 Saint Mary's St.
Boston, MA 02246 USA

Tel: (978) 387-4157
E-mail: yupingw@bu.edu
Website: https://yupingw.github.io/
Office: Photonics Center, Room 310

7.2 Research Interests

Disinformation, Computational Social Science, Social Media Analysis, Social Computing, and Data Science

7.3 Education

Boston University Sep 2017 - Jan 2023
Boston, USA

Ph.D., Systems Engineering

- Advisor: Prof. Gianluca Stringhini

Zhejiang University Sep 2013 - Jun 2016
Hangzhou, China

M.S., Systems Analysis and Integration

- Advisor: Prof. Ji Xiang

Huazhong University of Science and Technology Sep 2009 - Jun 2013
Wuhan, China

B.E., Automation

7.4 Work Experience

Zhejiang University
Hangzhou, China

Research Assistant

- Advisor: Prof. Ji Xiang

7.5 Publications

- **Yuping Wang**, Chen Ling, and Gianluca Stringhini, Understanding the Use of Images to Spread COVID-19 Misinformation on Twitter, *PACM HCI*-2023. This work will be presented in *CSCW*-2023
- **Yuping Wang**, Savvas Zannettou, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, and Gianluca Stringhini, A Multi-Platform Analysis of Political News Discussion and Sharing on Web Communities, *IEEE BigData*-2021
- **Yuping Wang**, Fatemeh Tahmasbi, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, David Magerman, Savvas Zannettou, and Gianluca Stringhini, Understanding the Use of Fauxtography on Social Media, *ICWSM*-2021 (Accepted directly, top 3/90. ICWSM is a top-tier conference for computational social science, particularly social media.)
- **Yuping Wang**, Ji Xiang, Yanjun Li, Michael Z. Q. Chen, Controllability of Dynamic-Edge Multi-Agent Systems, *IEEE Transactions on Control of Network Systems* 5 (3), 857-867, DOI: 10.1109/TCNS.2017.2648513
- **Yuping Wang**, Yanjun Li, Ji Xiang, Controllability of Multi-agent Systems Coupled by Dynamic Edges, The 34th *Chinese Control Conference* 2015 (CCC2015), Hangzhou, China, July 2015, P6765-6770
- Chris Chao Su, Nina Cesare, Zhuo Chen, and **Yuping Wang**, Social Media Users and Their Tweeting Behaviors: Quantifying Sample Biases in a Conspiracy-Theory Dataset, *ICA Conference*-2022, Extended Abstract

7.6 Teaching Experience

- EK 381 Probability, Statistics, and Data Science for Engineers, Graduate Teaching Fellow, Spring 2021
- EK 500 Probability and Statistical methods, Graduate Teaching Fellow, Fall 2021

7.7 Honors and Awards

- Distinguished Systems Engineering Fellowship, BU, 2017
- Excellent Graduation Student, HUST, Jun 2013

7.8 Reviewer

- IEEE Transactions on Automatic Control
- ICWSM
- IEEE Internet Computing

7.9 Computer Skills

- Languages: Python, SQL, Matlab, C, HTML, CSS, JavaScript and LaTeX
- Operating Systems: Linux, Windows