FAKE NEWS ON SOCIAL MEDIA

FAKE NEWS ON SOCIAL MEDIA: FROM FAKE NEWS LIFECYCLE TO FAKE NEWS COMBAT CYCLE

By MONA NASERY, M.Sc., B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree

Doctor of Philosophy in Information Systems

McMaster University DOCTOR OF PHILOSOPHY in Information Systems (2024) Hamilton, Ontario, Canada

TITLE: Fake News on Social Media: From Fake News Lifecycle to Fake News Combat Cycle

AUTHOR: Mona Nasery, M.Sc., B.Sc. (McMaster University)

SUPERVISOR: Dr. Yufei Yuan, Dr. Ofir Turel

NUMBER OF PAGES: xiii, 177

Abstract

Social media platforms facilitate the spread of a large volume of information in split seconds. However, some false information is widely spread, generally called “fake news”. This can have significant negative impacts on individuals and societies. Thus, there is an urgent need to find effective mechanisms to combat fake news on social media. The first step to address this problem is to understand fake news clearly. To this end, this research first provides an overview of the fake news lifecycle and different types of false information. This dissertation includes two primary studies. The first study aims to understand various kinds of false information on social media, focusing on X. We analyzed the spread dynamics of different types of false tweets and user behaviour towards each type using advanced data analytics and NLP methods. Finally, we examined whether and how users’ responses affect the spread of false tweets. This study is important from several aspects. First, considering the rapid spread of fake news on social media, only a tiny fraction can be flagged by fact-checkers. Understanding the spread dynamics of diverse types of false information helps decide what kinds of false content to fact-check first. Second, analyzing users’ conversations provides insights into users’ behaviour. It shows what users think and how they react to a piece of information, which helps develop more efficient fake news detection and classification tools. The second study aims to provide a comprehensive approach to combat fake news on social media. We adopt the Straub Model of Security Action Cycle to the context of fighting fake news on social media. We use the framework to classify the vast literature on fake news into action cycle phases (deterrence, prevention, detection, and mitigation). Based on a systematic and inter-disciplinary literature review, we analyze the status and challenges in each stage of combating fake news and introduce future research directions. These efforts allow the development of a holistic view of the research frontier on fighting fake news online.

Acknowledgements

This PhD thesis would not have been possible without the guidance, support, and encouragement of many individuals I am deeply grateful to.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Yufei Yuan who has been an exceptional mentor throughout my PhD. This research would have never been completed without his constructive criticism, constant support, and kind-hearted nature. His positive mindset and continuous encouragement kept me motivated and hopeful, even during challenging times. I am very grateful to have had the opportunity to work under his supervision.

I am also profoundly thankful to my co-supervisor, Dr. Ofir Turel, Professor at Melbourne University. He dedicated a significant amount of time and effort to guide me. His invaluable insights and expertise greatly enriched my research. I am also grateful to Dr. Milena Head, whose positive attitude and feedback as a committee member added value to my research. I would also like to thank my previous supervisor, Dr. Brian Detlor. Although I eventually shifted my research direction, Dr. Detlor was always supportive and provided valuable guidance during the early stages of my PhD.

In addition, I am thankful to Dr. Jesse Shore, a research scientist at Meta and former assistant professor at Boston University. Although Dr. Shore was a committee member for a short period, his guidance was incredibly impactful. I would also like to acknowledge my thesis external reviewer, Dr. Joey George, a Distinguished Professor at Iowa State University, whose constructive feedback significantly improved the quality of my thesis.

Furthermore, I would like to thank the staff at McMaster University for creating a warm and welcoming environment. I also like to thank my colleagues and friends at DeGroote, especially Samira Farivar, Sepideh Ebrahimi, Ali Vaezi, and Somayeh Ghazalbash for their friendship and support. I am also grateful to my friends outside the school, whose companionship made my years in Toronto more enjoyable, especially when I was far from home and family. Special thanks go to Amir Ashouri, for his constant support and companionship during this journey.

Lastly, I wish to express my deepest gratitude to my family, especially my parents, for their constant support and encouragement throughout my academic journey. Their love and belief in me have been my greatest source of strength.

Table of Contents

List of Tables

List of Figures

Chapter 1

1 Introduction

1.1 Background and Motivation

Social media platforms such as X (formerly Twitter) provide a more accessible, cheaper, and faster way for individuals to consume and share news. Today, half of U.S. adults get news from social media at least sometimes[1]. However, these benefits come at a cost, namely a large volume of fake news on social media platforms. Fake news is news items that are false, regardless of the intentions of the news originator (Zhou & Zafarani, 2020). As such, they include misinformation (false or misleading information with no intention to deceive) and disinformation (false information to deceive people) (Lazer et al., 2018).

The spread of fake news on social media can severely impact individuals and societies. For example, in the context of COVID-19, fake news about ingesting fish tank cleaning products, alcohol, or injecting bleach to treat the virus can pose a severe threat to people's lives. The harmful impacts of fake news have been shown in various other contexts, such as politics (Allcott & Gentzkow, 2017), the economy (Kogan et al., 2019), and responses to natural disasters (Gupta, Lamba, Kumaraguru et al., 2013). Thus, there is an acute need for effective mechanisms to stop or limit the harmful consequences of fake news. Indeed, giant tech companies such as Google, Microsoft, Facebook, and X issued a joint

[1] https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/

statement to combat fake news about COVID-19[2]. In response, scholars have proposed numerous approaches to combat fake news. However, as I discuss later, such approaches primarily focused on one action, namely detection, and overlooked tackling the problem through different stages of fake news dissemination. The remainder of this section describes some of these challenges and gaps in the literature and explains how this research addresses them.

First, there needs to be an overall agreement on what to consider fake news. The term "fake news" can refer to different forms of false or inaccurate information (with or without intention), such as rumours, satire, conspiracy theories, and more. To better address the problem of fake news, the first step is to clearly understand its definitions, different types of fake news, and their characteristics. To this end, *the first part of this research* aims to provide a better understanding of fake news by providing various definitions of the term in the literature, different types of fake news, and an overview of the fake news ecosystem by identifying the stages of fake news lifecycle, namely, creation, spread, and impact. For each stage of the fake news lifecycle, I review relevant studies in the literature. The first part of this dissertation addresses some of the research gaps identified in the literature review. For example, a large body of research compared fake news to true news in terms of its propagation characteristics, content, or users' responses. However, they often didn't differentiate between different types of fake news and either looked at one kind or lumped them together.

[2] https://X.com/microsoft/status/1239703041109942272?lang=en

Similarly, in terms of content, most prior studies focused on comparing the content of true vs. fake news, and they often analyzed the content of fake news posts (and not users' comments on fake news). Very few studies analyze the content of users' replies to fake news (Shu, Mahudeswaran, Wang, Lee, et al., 2020). To the best of my knowledge, there has been no study analyzing the content of replies to different types of fake news.

To fill the above gap, the first part of this research discusses different types of false information and their characteristics. The four types of fake news studied in this research include Conspiracy theories, Clickbaits, Political/biased, and Misleading (other types such as rumours, satires, etc.). More specifically, **study 1** addresses the following research questions:

- ***RQ1:*** *Do the propagation characteristics of diverse types of false information differ?*
- ***RQ2:*** *How do users respond (emotionally and attitudinally) to different types of false information? Do responses differ among diverse types of false information?*
- ***RQ3:*** *Do users' responses (as related to RQ2), explain the differences in the spread patterns of diverse types of false information (as associated with RQ1)?*

Second, the large spread of fake news on social media severely impacts individuals and societies. Unfortunately, people often cannot correctly identify fake news from the truth. Also, manual fact-checking and debunking fake news cannot keep up with the large volume and fast spread of fake news on social media. As a result, a large body of research focused on automated fake news detection. However, these mechanisms only address

fake news when it already spread. Fake news can spread exponentially fast at the early stages and pose destructive impacts in a very short period. For example, a false tweet about Barack Obama being injured in a White House explosion (although debunked quickly) was "enough to wipe out $130 billion in stock value in a matter of seconds"[3]. Thus, it is essential to devise strategies to stop fake news after its spread and even before its creation. To the best of my knowledge, there has been a dearth of studies that provide a comprehensive picture of how to combat fake news on social media systematically. **Study 2** aims to address this gap by:

- *Providing a comprehensive framework to combat fake news on social media in four stages: deterrence, prevention, detection, and mitigation/remedy. For each stage, based on a systematic and inter-disciplinary review, I describe the challenges, existing approaches in the literature, their limitations, and introduce opportunities for future research.*

1.2 Research Structure

I researched this topic by comprehensively reviewing the literature about fake news on social media. Based on this review, there are two main streams of research in the fake news literature: 1) Studies focusing on fake news characteristics, including propagation-based, content-based, context-based, user-based, and feedback-based features (e.g., users' responses). There have been fewer attempts to study these characteristics from the perspective of ***fake news lifecycle***. Thus, the first part of my review looks at studies about

[3] https://business.time.com/2013/04/24/how-does-one-fake-tweet-cause-a-stock-market-crash/

fake news in different stages of its lifecycle, from when it is created until it propagates and impacts individuals and societies, 2) Studies focusing on ***combating fake news*** on social media. Most studies in this research stream focused on fake news detection approaches. We propose a framework to fight fake news during the whole lifecycle of fake news, which comprises four stages: deterrence, prevention, detection, and mitigation. Accordingly, I divided my proposed research into two main parts. The first part aims to address some research gaps found in the first stream of research (e.g., the fake news lifecycle), and the second part addresses a research gap in the second stream (e.g., combating fake news). Figure 1 summarizes the research structure and position of the two parts of my proposed research under two main streams of literature.

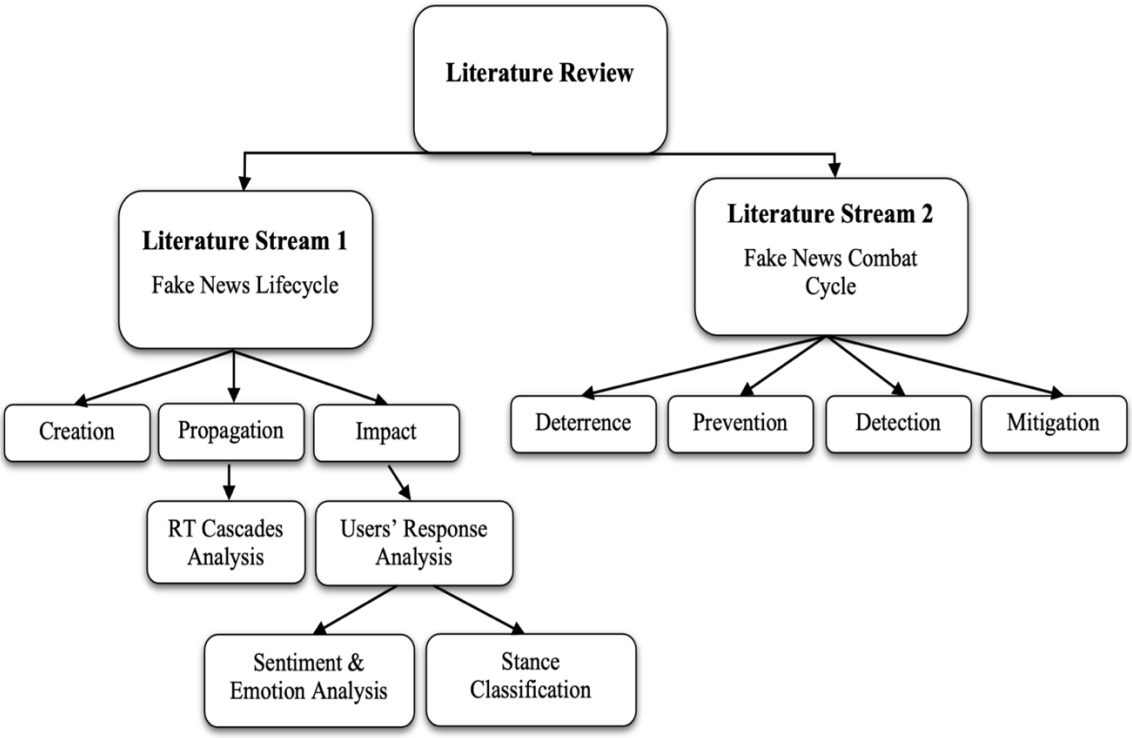

Figure 1: Research Structure

Chapter 2

2 Fake News Typology and Lifecycle

In this section, I start by reviewing how the term “fake news” is defined in the literature and present the definition adopted in this paper. Next, I review different types of false information, often used interchangeably as “fake news” in the literature. Note that there are several types of false information, and the so-called term “fake news” can be considered as one type of “false information”. Finally, I describe different stages in the fake news life cycle on social media.

2.1 Fake News Definition

The term “fake news” has gained widespread attention mainly after the 2016 U.S. Presidential Campaign. There has been no overall agreement on the definition of fake news. This is because the term "fake news" covers a wide range of (with or without intention) false or inaccurate information, such as deceptive stories, rumours, satires, and conspiracy theories. Therefore, this section aims to provide an overview of how the term “fake news” has been used and defined in the literature. Also, in the next section, we provide clear definitions and examples of different types of fake news or terms closely related to fake news. Allcott & Gentzkow (2017) define fake news as “a news article that is intentionally false and is verifiable”. Several other studies (e.g., Bondielli & Marcelloni, 2019; Kim et al., 2019; Shu et al., 2017) adopted this definition. This is, however, a narrow definition of fake news, which emphasizes the information's

authenticity and intention. There are also broader definitions of fake news, which do not restrict the intention of the information/news. For example, Zhou & Zafarani (2020) broadly defined fake news as false news. Table 1 shows different definitions of fake news. In this research, we purposefully adopt the broad definition of fake news provided by Sharma et al. (2019): *"a news article or message published and propagated through media, carrying false information regardless the means and motives behind it"*. The broad definition of fake news allows us to cover different types of fake news and related terms, such as rumours, misleading news, and conspiracies.

Table 1: Definitions of Fake News in the Literature

Fake News Definition	**Reference(s)**
Fake news is false news (broad definition)	
A news article or message published and propagated through media, carrying false information regardless of the means and motives behind it (broad definition).	(Sharma et al., 2019)
News article that is intentionally and verifiably false (narrow definition)	(Allcott & Gentzkow, 2017), (Bondielli & Marcelloni, 2019), (A. Kim et al., 2019a)
Fabricated information that mimics news media content in form but not in organizational process or intent	(Lazer, et al., 2018)
False stories disguised as a credible news source for political or financial gain	(Shin et al., 2018), (Silverman, 2017)
Information presented as a news story that is factually incorrect and designed to deceive the consumer into believing it is true	(Golbeck et al., 2018)

2.2 Different Types of False Information

Several terms and concepts linked to fake news have been frequently used in the literature. For example, Tandoc Jr et al., (2018) identified six ways that the term "fake news" has been used in the literature: satire, parody, fabrication, manipulation, propaganda, and advertising. A good distinction between fake news and different terms

related to fake news is provided based on three characteristics: *intention* to deceive or mislead others, *authenticity* (whether it includes non-factual information), and whether the information is *news* (Zhou & Zafarani, 2020). For example, based on intention, false information can be divided into two broad categories: misinformation and disinformation. **Misinformation** refers to "inadvertent sharing of false information" (there is no intention). **Disinformation**, on the other hand, refers to the "deliberate creation and sharing of false information" (S. Kumar & Shah, 2018; Wardle, 2017). Rubin et al., (2015) identified three types of fake news: serious fabrications (tabloids and yellow journalism), large-scale hoaxes (deliberate falsification causing harm), and humorous fakes (satire and parody). Table 2 presents different types of fake news and the associated definitions. It differentiates between various kinds of fake news based on two main dimensions: (1) the authenticity or facticity of the news stories (does it rely on facts? Is it based on factual or non-factual statement?), and (2) the intention to deceive or mislead readers/users.

Truthfulness	Intention	Relevant Terms & Definitions
False	*Malicious*	**Disinformation**: False information with the intention to deceive (S. Kumar & Shah, 2018; Wardle, 2017)
		Hoax: Reports of false information disguised as proper news (Bondielli & Marcelloni, 2019; Rubin et al., 2015). A false story used to masquerade the truth, originating from the verb hocus, meaning "to cheat" (Nares 1822). News stories that contain facts that are either false or inaccurate and are presented as legitimate facts (Zannettou et al., 2019) A deliberately fabricated falsehood made to masquerade as truth (Wikipedia).
		Serious Fabrication: Prototypical form of fake news, i.e. articles with a malicious intent that often become viral through social media (Bondielli & Marcelloni, 2019; Rubin et al., 2015)
		Propaganda: News stories which are created by a political entity to influence public perceptions (Tandoc et al., 2018)
	No Malicious intention	**Misinformation**: False or misleading information without the intention to deceive (S. Kumar & Shah, 2018; Wardle, 2017)
		Parody: Use of non-factual and fabricated content to inject humor (Tandoc et al., 2018)
		Satire[4]: News stories that are factually incorrect, but the intent is not to deceive but rather to call out, ridicule, or expose behavior that is shameful, corrupt, or otherwise "bad (Golbeck et al., 2018). Mock news programs, which typically use humor or exaggeration to present audiences with news updates (Tandoc et al., 2018) News stories "in a format typical of mainstream journalism but rely heavily on irony and humor to emulate a genuine news source, mimicking credible news sources and stories" (Rubin et al., 2015)
True	*Malicious*	**Misleading Content**: Misleading use of information to frame an issue (Sharma et al., 2019)

[4] The truthfulness of satire depends on which definition we adopt. For example, Tandoc et al., (2018) considered satire as facts and stated that "their being fake only refers to their format", while Golbeck et al., (2018) considered satires as "factually incorrect" stories. In this paper, we adapted the latter.

		Irrelevant Context: Using true information in an unrelated context to mislead people
	No Malicious intention	**Real News or True Information**: genuine news or true information based on facts.
The True or False is unknown	*Malicious*	**Conspiracy Theory (CT)**: A proposed explanation of some events in terms of the significant causal agency of a relatively small group of persons acting in secret (Keeley, 1999). Causal narratives of an event as a covert plan orchestrated by a secret cabal of people (or organizations) instead of a random or natural happening (Banas & Miller, 2013; Douglas & Sutton, 2008)
		Clickbait: Use of misleading headlines to entice readers to click on links under false pretenses (Ireton & Posetti, 2018). Article titles or social media posts whose aim is to attract readers to follow a link to the actual article page (Bondielli & Marcelloni, 2019; Y. Chen et al., 2015)
	No Malicious intention	**Rumor*:*** Stories whose truthfulness is ambiguous or never confirmed (Zannettou et al., 2019). Circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety (Zubiaga et al., 2018)

Table 2: Classification and definitions of different types of false information

2.2.1 Types of False Information in this Research

The primary data in this research is collected by Sharma et al., (2020), which includes four types of false information: *unreliable, conspiracy, clickbait*, and *political/biased* (please note that false content may belong to multiple types). The criteria for assigning false news to the four mentioned types are described in Sharma et al., (2020). Section 3.3 provides further information about the data, including the data collection procedure and data characteristics. The four types of false information in this study are further explained below. In addition, Table 3 provides a comparison of these four types across various dimensions.

2.2.1.1 Conspiracy Theories

Sunstein & Vermeule, (2008) defined conspiracy theory as "an effort to explain some event or practice by reference to the machinations of powerful people, who have also managed to conceal their role". Conspiracies are explanations to reduce the complexity of reality by explaining significant social or political events as secret plots conceived by powerful individuals or organizations (Bessi, Petroni, et al., 2015). More definitions of conspiracy theories are provided in Table 2. Conspiracies about 9/11, climate change (e.g., denial of global warming), the death of Princess Diana, the assassination of John F. Kennedy, and the origins of AIDS and COVID-19 viruses (e.g., they are produced in government labs) are just some examples of conspiracy theories in different domains.

Conspiracies are often used when people are not able to find the reason behind some events, such as a virus pandemic. Therefore, they provide explanations by offering some speculations to blame the government, authorities, or secret powerful groups, or maybe just to confirm their pre-existing beliefs (Sunstein & Vermeule, 2008). The intentions behind the conspiracies are always nefarious (Keeley, 1999). Regarding the characteristics of people who are more prone to believe conspiracies, research found that they have higher levels of narcissism but low self-esteem (Cichocka et al., 2016). Also, believing in conspiracy theories is associated with higher levels of paranoia (Imhoff & Lamberty, 2018). Finally, people who are more prone to fall for conspiracies often have an intuitive thinking style rather than rational thinking (Swami & Furnham, 2014).

Previous research have shown the harmful impacts of conspiracies in many aspects, such as promoting political violence driven by anger (Jolley & Paterson, 2020), committing

crime (Jolley et al., 2019), or undermine public health measures (Hornsey et al., 2018, 2021; Imhoff & Lamberty, 2020), such as Covid-19 vaccine hesitancy (Hornsey et al., 2020).

2.2.1.2 Clickbaits

This category includes exaggerated or misleading headlines and/or body purposed to attract attention for reliable and/or unreliable information. In this research, clickbait contains unreliable information. Chen et al., (2015) suggested referring to clickbait as misleading content or false news. They defined clickbait as contents that seek to attract the attention of readers and lure them into clicking on a link to a website through tactics such as sensationalist stories, eye-catching headlines and images that work as bait. The main objective of clickbait is to increase the click-through rate and, therefore, increase the advertising revenue (Shu et al., 2017). More definitions of clickbait are provided in Table 2. Clickbait headlines create a "*curiosity gap*", which encourages the readers to click on the link to address their curiosity. The characteristics of curiosity: its intensity, transience, association with impulsivity, and tendency to disappoint (Loewenstein, 1994) lead to cognitively induced deprivation - a knowledge gap - which motivates exploring activity from the reader (Chen et al., 2015).

Previous research investigated the psychological appeal of clickbaits. Blom & Hansen, (2015) examined how clickbaits use two forms of *forward referencing* – discourse deixis and cataphora – to lure the readers to click on the article links (Chakraborty et al., 2016). Chen et al., (2015) studied clickbaits characteristics and examined potential methods for the automatic detection of clickbaits. They argued that clickbaits can be identified

through certain linguistic patterns combined with readers behavior as predictive variables. For example, clickbaits often use *exaggeration* and *sensationalism* to produce an information gap, attract public attention, and encourage them to click. The sensationalism of a news item can be evaluated from several aspects such as sentiment, punctuation, and similarity between the news headline and its main body (Zhou et al., 2020). Other characteristics of clickbaits are *low quality* and *high informality* (e.g., using swear words such as "damn").

Cognitive studies have argued that clickbait is an enabler of *attention distraction*. As the readers keep switching to new articles after being baited by the headlines, the attention residue from these constant switches result in *cognitive overload*, deterring the readers from reading more informative and in-depth news stories (Chakraborty et al., 2016). Unfortunately, clickbaits help fake news attract more clicks (i.e., visibility) and further gain public trust, as indicated by the *attentional bias,* which states that the public trust to a certain news article will increase with more exposure, as facilitated by clickbaits (Zhou et al., 2020).

2.2.1.3 Political/Biased

This category includes political and biased news, written in support of a particular point of view or political orientation, for reliable and/or unreliable information such as propaganda (Sharma et al., 2020). Political false tweets are a subset of false information that specifically revolves around political topics, including elections, political candidates, policies, and current events. These tweets often contain misleading or fabricated information with the goal of influencing public opinion, shaping political narratives, or

promoting a particular agenda (Lazer et al., 2018). Examples of political false tweets include false claims about political figures, or fake news stories related to elections.

The spread of political false tweets may be influenced by partisan strength and political ideologies. Also, "consistent with theories of selective exposure, people differentially consume false information that reinforces their political views" (Guess, Nyhan, et al., 2020). Prior research found that the rapid spread of political false information is fueled by partisan polarizations, echo chambers, algorithmic amplification, and the engagement of politically motivated users (Del Vicario et al., 2016; Kitchens et al., 2020). Political false tweets can have significant implications for public discourse, democratic processes, and societal trust. They can undermine the credibility of political institutions, polarize communities, and fuel distrust in media sources (Lazer et al., 2018).

2.2.1.4 Misleading (Others)

This category is defined to include *false, questionable, rumors* and *misleading news*. In addition, *satire* is also included in this category because satire has the potential to perpetuate misinformation (Zimdars, 2016) or be used as a cover for the spread of misinformation (Sharma et al., 2019). Also, satire is more similar in complexity and style to fake news than to true news (Horne & Adali, 2017).

Table 3: Comparison of different types of false tweets in our study across various characteristics

Fake News Type	Characteristics (Style, ideology, etc.)	Why people believe it? (Motivations)	Main Purpose
Conspiracy Theory	• rooted in people's beliefs and interests. • target morals more than clickbait • intuitive thinking (based on feelings and instincts, rather than logic)	• pre-existing beliefs • cognitive biases • social influence • high level of paranoia & narcissism • lack of trust (Goertzel, 1994) • lack of control	• to blame the government, authorities, or some secret powerful groups • to confirm their pre-existing beliefs
Clickbait	• eye catching • extreme sentiment & exaggeration • sensational content • information gap between the title and the body	• curiosity • attentional bias • forward referencing (discourse deixis & cataphors) (Chakraborty et al., 2016)	• mainly monetary (e.g., to increase the click-through rate and advertising revenue)
Political/Biased	• biased towards a specific political party	• political/ideological orientations	• to support a political party or ideology

2.3 Fake News Lifecycle

The life cycle of any online news has three basic stages: from the time it is created and published on online platforms until it disseminates and impacts individuals and societies. The remainder of this section further explains each stage. Additionally, Table 4 provides sample references in the literature categorized by the stage of fake news lifecycle they focused on and by the social media platform they used.

2.3.1 Stage 1: Fake News Creation

The first stage in the lifecycle of any news (true or false) is the creation stage. Identifying the creators of fake news on social media is important because it can help in stopping a large number of fake news from the origin (creation stage), before it spreads on social media.

2.3.1.1 Who Creates Fake News on Social Media?

Fake news on social media platforms can be created automatically (e.g., by social bots) or manually by real humans. Many researchers believe that focusing on the source/creator is one of the best ways to detect fake news. There has been extensive research on the analysis of malicious accounts on social media. Overall, we can categorize creators of fake news into two broad categories: Social bots and humans. Both humans and bots deceive people by creating an illusion of consensus towards the fake content, for example, by repeating it multiple times (S. Kumar & Shah, 2018).

Social Bots (Non-humans): Social bots are fake accounts created by a single individual or a program to promote (true or false) information. They control an account on a particular online social network, and are able to perform several activities (e.g., posting a message) with the goal of influencing users and spreading false information. Several studies found that social bots are responsible for a considerable amount of X political chatter (Bessi & Ferrara, 2016; Shao et al., 2017). Shu et al., (2020) randomly sampled 10,000 users who posted fake and real news on X. Their findings confirmed the presence of bots in X to create and spread fake news and that bots are more likely to post tweets related to fake news than real users. However, most users (~ 78%) who post fake news

are still more likely to be humans than bots, which is consistent with the findings in (Vosoughi et al., 2018).

Malicious online accounts (e.g., sock puppets) on social media intentionally create fake content to promote false information, influence individuals, and manipulate public opinion. Sock puppet refers to a person or group with false online identity with the purpose of deceiving others. Even though there are sock puppets without deceptive intention, sock puppets with deceptive intentions are twice as common (S. Kumar et al., 2017). Sock puppet accounts engage with ordinary users in online discussions and agree with each other to amplify their opinion and oppose those who disagree with the information.

Humans: Whether fake news is created manually or automatically (using bots), humans who aim to deceive online users are the ultimate creators of false information because even social bots are programmed by humans. Despite the findings about the role of bots in creation and spread of fake news, some recent studies found that humans, and not bots, were the main actors for creation and spread of false information on X. For example, in an analysis of over 126,000 false information on X (Vosoughi et al., 2018) showed that humans were responsible for spread of false information on X, not bots.

2.3.1.2 Characteristics of Fake News Creators

Analysis of the news source or creator is an important aspect of detecting fake news on social media. Identifying the characteristics of creators/sources of fake news helps in debunking such accounts before fake news spreads. Previous research has shown that

malicious accounts have different characteristics and behaviour compared to legitimate accounts. For example, Chu et al., (2012) showed that the numbers of followers and friends (followings) are good indicators for identifying malicious or fake accounts. They proposed an equation to measure *account reputation* with the number of followers and friends/followings. According to their findings, the number of followers of the legitimate accounts is close to the number of friends/followings, except for a few nodes representing celebrities, famous influencers, and organizations with much more followers than friends/followings. In contrast, bots usually have much more friends/following than followers.

Prior research have shown that the *credibility of the creator* of a post/news can be a good indicator for the credibility of information, which ultimately can be used to identify false information (Castillo et al., 2011; Ma et al., 2015; Wu et al., 2015; F. Yang et al., 2012; Zubiaga, Liakata, & Procter, 2016). There are several characteristics of the source/creator accounts that can be useful to assess the user credibility such as: the *age* of the account, analysis of the *account/creator activity* (e.g., number of posts, shares, etc.), and *temporal* analysis (e.g., the frequency of posting, sharing, replying, etc.). For example, the *age* of fake news accounts tend to be shorter than the real ones (Allcott & Gentzkow, 2017). One example is many fake news accounts during the 2016 US presidential election which no longer exist. Account activity is another factor to differentiate bots, fake, and malicious accounts from real accounts with benign intentions. For example, Kumar et al., (2017) used activity features such as number of posts, time difference between two consecutive posts, and tenure time (number of days from user's first post) to identify sockpuppets.

Regarding the temporal behavior, bots usually use regular and periodic timing (Chu et al., 2012) while human behavior is more complex (Gianvecchio et al., 2008).

2.3.2 Stage 2: Fake News Propagation

Once fake news is created, it spreads quickly on social media in the form of shares and re-shares (e.g., tweets and retweets on X). A large body of research has focused on fake news propagation in social media (Amoruso et al., 2020; Del Vicario et al., 2016; Shao et al., 2017, 2018; Vosoughi et al., 2018; Zubiaga, Liakata, Procter, et al., 2016). The propagation dynamics of information on social media can provide valuable insights to detect true from false information (Castillo et al., 2011). Fake news often mimics real news in content, and therefore, fake news detection only based on the news content is not very effective. However, the propagation features of false news are shown to be different from real news (Shao et al., 2017; Vosoughi et al., 2018). A large-scale study of false information on X found that tweets about false stories spread significantly farther (more number of users retweeted), deeper (more number of retweet hops), faster (more number of retweets in a shorter time), and broader (more number of users at the same depth) than those about true stories (Vosoughi et al., 2018). Several other studies had similar findings and showed that false stories spread faster (Doerr et al., 2012; L. Zeng et al., 2016; Zubiaga, Liakata, Procter, et al., 2016) and deeper (Friggeri et al., 2014) than true stories.

The differences between propagation characteristics of true and false news can provide useful clues for fake news detection. A large number of studies incorporated propagation features into their machine learning or deep learning models for the automatic detection of fake news (Bian et al., 2020; Castillo et al., 2011; Ma et al., 2015; Shu,

Mahudeswaran, Wang, & Liu, 2020; Wu et al., 2015). Most early works in this direction used propagation features in supervised classifiers to identify fake news (Castillo et al., 2011; F. Yang et al., 2012), and later in deep neural models (Ma et al., 2016; Ruchansky et al., 2017). Some studies developed kernel-based methods, which model information structure as propagation trees to detect true from false rumors by comparing their tree-based similarities (Ma et al., 2017; Wu et al., 2015). More recently, Shu et al., (2020) built hierarchical propagation networks of fake and real news and analyzed them in terms of structural (e.g., tree depth, height, etc.) and temporal aspects (e.g., tweets lifetime, speed). Their study shows the effectiveness of propagation network features for fake news detection.

While most prior research in this stream studied the propagation characteristics of "true" vs "false" information, less attention is given to different types of false information. Only a handful of studies examined the differences in the spread of true and false stories across topics or different types of false information. For example, the findings of Vosoughi et al., (2018) that false news spread faster, deeper, and broader than true news, were found to be more pronounced for political fake news than false information in any other context such as natural disasters, science, terrorism, or financial information. A few studies focused on specific types of false information and compared information consumption patterns of conspiracy theories versus scientific stories (Bessi, Coletto, et al., 2015; Del Vicario et al., 2016, 2016). Their findings depict that conspiracy stories are spread more slowly and showed a positive relation between lifetime and cascade size (number of users in the cascade). In contrast, science news reaches a higher level of diffusion more

quickly, and there is no relationship between lifetime and cascade size (i.e., a longer lifetime doesn't correspond to a higher level of interest). However, none of these studies compared the spread of different types of false information (e.g., conspiracies vs. clickbait). The first part of this thesis aims to address this gap in the literature.

False information comes in various forms, with different characteristics and purposes (Rubin et al., 2015; Tandoc Jr et al., 2018). Not all types of false stories spread or behave the same way. Ignoring the distinctions between different types of false information may lead to inefficient or unnecessary intervention design (Babcock, Beskow, et al., 2019). In addition, understanding the differences between types of false stories may be useful for developing future classification tools and for improved decision-making regarding whether to and how to respond to each type of false information. Recently, a few studies found that different types of false information act differently (Babcock et al., 2018; Babcock, Beskow, et al., 2019; Volkova & Jang, 2018). For example, Volkova & Jang, (2018) compared retweet patterns of different types of false information and find that clickbait, hoaxes, and propaganda are retweeted at higher volumes in a shorter period of times compared to base rumors. A case study of X conversation about the Black Panther movie showed the difference in the diffusion speed of false stories across different story types, response types, and communities (Babcock et al., 2018). Also, the results of a subsequent study found that different types of false stories differ in the role users play, the amount and timing of activity, the use of hashtags, and the possible presence of bot like accounts (Babcock, Beskow, et al., 2019). The types of false information in the last two studies were: fake attack, satire attack, fake scene, and Alt-right.

The first part of this thesis adds to this line of research by comparing some propagation characteristics (e.g., cascade size, speed, lifetime) of different types of false information, namely conspiracies, clickbait, political, misleading (other types).

2.3.3 Stage 3: Fake News Impact

The spread of false information can have severe and far-reaching impacts on many aspects of our lives, including but not limited to politics (Allcott & Gentzkow, 2017; Aro, 2016; Fisher et al., 2016), economy (Carvalho et al., 2011; Kogan et al., 2019; Rapoza, 2017), and responses to natural disasters (Gupta, Lamba, Kumaraguru, et al., 2013; Takayasu et al., 2015). In politics, one severe instance of the impact of fake news was the "Pizzagate" incident, where the spread of political fake news during the 2016 USA presidential election led to violence and public shootings (Fisher et al., 2016). The propagation of false information on social media during natural disasters, such as the spread of fake images of Hurricane Sandy, created panic and chaos among the people (Gupta, Lamba, Kumaraguru, et al., 2013). In terms of economy, fake news propagation has also impacted financial markets. Carvalho et al. (2011) noted a false report of bankruptcy of a United Airlines parent company in 2008 caused the stock price to drop by as much as 76% in a matter of minutes; although the stock rebounded after the news was identified as false, it closed 11.2% lower than the previous day and the negative effect persisted for 6 more days.

The impact of false information can be measured **using engagement statistics** such as view count, share count, and more. For example, Kumar et al., (2016) measured the impacts of hoaxes in Wikipedia in terms of their *survival time* (How long they survive

before they get debunked), view count (how often they are viewed), and *spread* (#links clicked by readers to reach the hoax article). Their findings suggest that most hoaxes (about 90%) on Wikipedia are identified and flagged quickly within an hour of patrol. Also, hoaxes are viewed less frequently compared to non-hoaxes (85% get fewer than 10 views per day), but a non-negligible number (1% of hoaxes) get a lot of views (at least 100 views per day). Silverman, (2016) analyzed engagement of true and fake news on Facebook during the 2016 US Presidential election. They measured engagement in terms of the number of shares, *reactions, and comments* on a post and found that fake news got significantly higher engagement compared to real news.

User engagements (likes, shares, replies, or comments) contain rich information captured in the *propagation structure* (e.g., retweet cascades/trees), *temporal information* in timestamps of engagements, *textual information in user replies (comments),* and *user profile information* by the user involved in the engagement. Previous research used user engagements to better understand user behavior and the underlying characteristics of different types of information (e.g., true vs. false news). For example, it is observed that fake news tends to receive more negative and questioning responses than true news (Qian et al., 2018). Also, sentiment of replies to true news are more neutral, compared to replies to for fake news, which tended more toward negative sentiments (Shu, Mahudeswaran, Wang, Lee, et al., 2020). Several studies have also used stance classification to understand users' position (e.g., agree, disagree, or neutral) towards a piece of information (Kaliyar et al., 2021; Müller et al., 2023; Umer et al., 2020). The first part of

this thesis adds to this line of research by analyzing users' reactions to (different types of) false information in terms of sentiments, emotions, and users' stance toward fake news.

Table 4: Relevant Studies in each Stage of Fake News Lifecycle. The type of false information addressed in each article is also specified: F: Fake news, R: Rumor, CT: Conspiracy Theory, H: Hoax, SP: Satir & Parody, MT: Multiple Types

Lifecycle / Platform	Creation	Propagation/Spread	Impact
X	*(Bessi & Ferrara, 2016)[R], (Shao et al., 2017)[F], (Chu et al., 2012)[SP], (Davis et al., 2016), (Shu, Mahudeswaran, Wang, Lee, et al., 2020)[F],*	*(Andrews et al., 2016)[R], (Arif et al., 2016)[R], (Doerr et al., 2012)[R], (Gupta, Lamba, & Kumaraguru, 2013)[F], (Jin et al., 2014, 2013)[R], (Mendoza et al., 2010)[R], (Oh et al., 2010)[R], (Vosoughi et al., 2018)[R]*[*], (Shao et al., 2017, 2016)[F], (Zubiaga, Liakata, Procter, et al., 2016)[R], (Kwon et al., 2013)[R], (L. Zeng et al., 2016)[R], (Bessi & Ferrara, 2016)[R], (Babcock, Beskow, et al., 2019)[MT], (Babcock, Cox, et al., 2019)[MT]*	*(Vosoughi et al., 2018)[R], (Rapoza, 2017)[F], (Gupta, Lamba, Kumaraguru, et al., 2013)[F], (Takayasu et al., 2015)[R],*
Facebook	*(Allcott & Gentzkow, 2017)[F], (Anagnostopoulos et al., 2014)[CT],*	*(Bessi, Petroni, et al., 2015)[CT], (Anagnostopoulos et al., 2014)[CT], (Del Vicario et al., 2016)[CT], (Friggeri et al., 2014)[R], (Silverman, 2016)[F]*[*], (Guess, Nyhan, et al., 2020)[F]*[*],	*(Friggeri et al., 2014)[R], (Allcott & Gentzkow, 2017)[F], (Silverman, 2016)[F],*
Other	*(F. Yang et al., 2012)[R](Sina Weibo), (Wu et al., 2015)[R](Sina Weibo)*		*(S. Kumar et al., 2016)[H](Wikipedia), (Carvalho et al., 2011)[F](Internet), (Kogan et al., 2019)[F](financial news platforms),*
Multi-platform	*(Ma et al., 2015)[R] (X & Sina Weibo),*	*(Zannettou et al., 2017)[F],*	

[*] (Vosoughi et al., 2018) studied the spread of **verified** true and false rumors.
[*] even though (Silverman, 2016) used the term "fake news", the sources of the fake news in their study was either from websites that publish **hoaxes** or **hyperpartisan** websites.
[*] (Guess, Nyhan, et al., 2020) used fake news (factually dubious content)

Chapter 3

3 Comparative Analysis of False Tweets

3.1 Motivation and Research Questions

A large body of fake news literature focused on the spread of fake news on social media (Pierri & Ceri, 2019; Vosoughi et al., 2018). However, most studies didn't differentiate between different types of fake news. They either looked at one type of fake news or lumped them together. For example, Vosoughi et al., (2018) studied the spread of more than 12,000 rumours on X and showed that fake news (rumours) spread significantly farther, faster, deeper, and more broadly. On the other hand, some other studies found that fake news (conspiracies) spread more slowly and the lifetime increases with cascade size (Del Vicario et al., 2016; Quattrociocchi et al., 2016). One potential explanation for the conflicting findings could be that the former study compared the spread of "**rumours**" and true information on X whereas the latter studies examined the spread of "**conspiracies**" vs. scientific information on Facebook.

Similarly, in terms of content, most prior studies focused on comparing the content of true vs fake news. Moreover, most prior research analyzed the content of fake news posts. There are very few studies analyzing the content of the replies to fake news (Shu, Mahudeswaran, Wang, Lee, et al., 2020), and to the best of our knowledge, there has been no study analyzing the content of replies to different types of fake news. Analyzing the content of the conversations in X is important because it provides further context to

the fake news post. It shows what users think, how they react to fake news, and whether they can realize the fake content. Not only is this helpful in understanding the differences (if any) in propagation characteristics of diverse types of fake news, but it can also help in predicting user behaviour or fake content. Further, previous studies have theorized the relationship between user behaviours and their perceived beliefs on the information on social media (A. Kim & Dennis, 2019). For example, the behaviours of likes and retweets are more emotional, while replies are more rational. Analyzing the replies to different types of fake news can also provide insights into user interaction related to different types of fake news.

Importance: *First*, different types of fake news have different design, purposes, and impacts, and their propagation features, such as speed, depth, or virality, may not be the same. Therefore, aggregating different types of false information into one "false" category, as most prior studies did, could ignore the differences important to understanding the conflicting results in the literature. *Second*, fact-checking is a costly process and considering the rapid spread of fake news on social media platforms, only a small fraction of false information can be flagged by fact-checking organizations. Therefore, deciding which types of false information to fact-check first is important as it helps remove the false content with potentially larger negative impact before false information with lower or negligible impact. *Finally*, understanding the nuances in the spread and content of diverse types of false information can inform the development of more accurate detection models and effective mitigation strategies and enhance our ability to counteract the impact of false information.

To this end, study 1 aims to address the following research questions:

- ***RQ1:*** *Do the propagation characteristics of diverse types of false information differ?*
- ***RQ2:*** *How do users respond (emotionally and attitudinally) to different types of false information? Do reactions differ among diverse types of fake news?*
- ***RQ3:*** *Do users' responses (as related to RQ2), explain the differences in the spread patterns of diverse types of fake news (as related to RQ1)?*

3.2 X (Twitter) Terminologies

A post on X can be one of three types: an original (source) tweet, a retweet, or a quote. An original (or source) tweet is content posted by a user on X. A retweet refers to the reshare of a post (therefore, the content of the tweet and retweet is the same post). These are generally posted by other users than the one who posted the original. A quote is a retweet with a comment added by the quoting user. The X API includes methods that can be used to find out if a post is an original or if it is a quote. In the case that it is a quote, information is also presented regarding which tweet was quoted. This information however is not included when it comes to retweets, then we can only know what the original tweet was. Therefore, if a post is a retweet of a retweet, we cannot know who the intermediary retweeter was. Figure 2 shows an example tweet and a retweet on X.

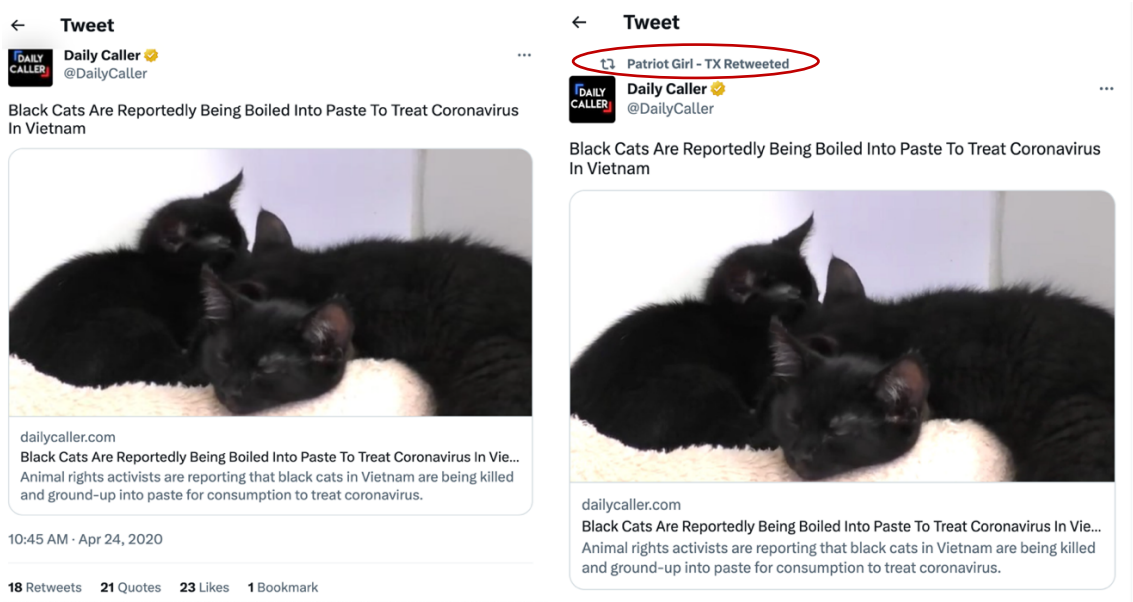

Figure 2: Example Tweet and Retweet on X. The figure on the left shows a source tweet posted by the user "DailyCaller", and the figure on the right shows one of its retweets (share) by the user "Patriot Girl".

3.2.1 Information Propagation & Retweet Cascades

A **retweet cascade** refers to a group of nodes consisting of the root node (an initial user post) and some reshare (retweet) of the post by other users. Figure 3 shows the list of retweeters of the sample tweet in Figure 2. The X API provides information about who retweeted a tweet (list of retweeters), but it doesn't show intermediary retweets (i.e., retweet of a retweet). Instead, it links all retweets to the source tweet. Therefore, it is not possible to directly construct the retweet cascade from the data collected from the official X API. One way to model information diffusion (e.g., to infer retweet cascades) is by using self-exciting point process models such as Hawkes processes (M. Kim et al., 2020; Zhao et al., 2015). In self-exciting processes, the occurrence of past events increases the likelihood of future events. Rizoiu et al., (2017) formulated information diffusion in X as a self-exciting point process, in which they modelled: *magnitude of influence*, tweets by

users with many followers tend to get retweeted more; *memory over time*, that most retweeting happens when the content is fresh, and *content quality*. Point-process models are useful to address a range of problems, such as explaining the nature of the underlying process, simulating future events, and predicting the likelihood and volume of future events.

To build the retweet cascades, we used the evently library (Kong et al., 2021) in R, which models reshare cascades using Hawkes processes. The evently library has a function to extract cascades from JSON formatted raw tweets. In the context of X, each retweet is considered an event in the point process. A simulated process is represented by a dataframe (table), where each row consists of an event time, which indicates the event happening time, and magnitude, which is the event mark information. In the context of retweet diffusion cascades, the first row is the original tweet, and all following events are its retweets. ***Time*** records the relative time (in seconds) of each retweet to the original tweet, and ***magnitude*** refers to the local influence of the user (here computed by the followers' count of the user). Kong et al., (2021) denoted a cascade observed up to time T as H (T) = $\{t0, t1, \ldots\}$, where $ti \in$ H (T) are the event times relative to the first event ($t0 = 0$). They denoted cascades with additional information about events (event marks). The mark (m) or magnitude of each event (retweet) models the user influence for each tweet. They used the notation $H_m(T) = \{(t0, m0), (t1, m1), \ldots\}$, where each event is a tuple of an event time and an event mark. The event intensity function in a Hawkes process is defined as:

$$\lambda(t) = \lambda_0(t) + \sum_{T_i<t} \phi_{m_i}(t - T_i)$$

$\lambda_0(t)$ is the arrival rate of immigrants' events into the system. The original tweet is the only immigrant event in a cascade, therefore $\lambda_0(t) = 0, \forall_t > 0$.

$$\lambda(t \mid \mathcal{H}(T)) = \sum_{t_i<t} \phi(t - t_i)$$

where $\phi : R^+ \rightarrow R^+$ is a kernel function capturing the decaying influence from a historical event. Exponential function $\phi_{EXP}(t) = \kappa\theta e^{-\theta t}$ and the power law function $\phi_{PL}(t) = \kappa(t + c)^{-(1+\theta)}$ are two most widely adopted kernel functions. Further information about retweet cascades and Hawkes processes is provided in (Rizoiu et al., 2017).

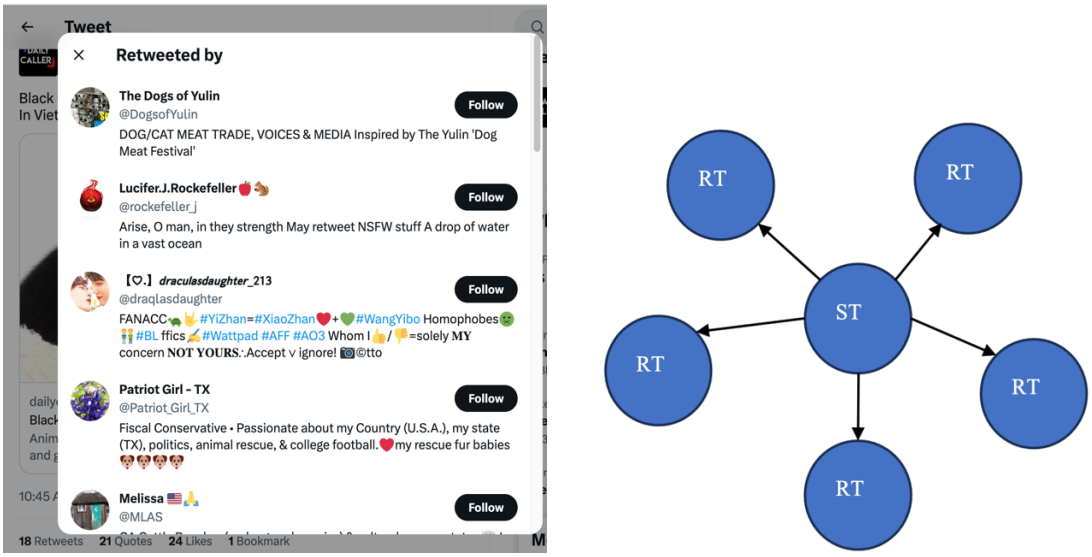

Figure 3: Users who retweeted a tweet (left), and a retweet cascade (star graph) as provided by the X API (right) ST: source tweet, and RT: retweet.

Having retweet cascades, we compare the propagation of different types of false information by looking at the following *propagation* and *temporal* features:

Cascade size: number of nodes (users) in the cascade. It shows the number of users sharing a tweet. The cascade size corresponds to the number of unique users involved in the cascade because users can only retweet (share) a tweet once on X.

Cascade lifetime: the duration of the propagation of each false information event. It is the time distance between the original (source) tweet and the last retweet.

$$Cascade\ Lifetime = Last\ Retweet\ Time - Original\ Tweet\ Time$$

This metric provides insight into the rate of dissemination of information through retweets over time. It's a straightforward way to gauge how quickly a tweet is spreading through the X network. By comparing the number of retweets per hour across different types of false information tweets, we can analyze the relative propagation speeds of each type and assess differences in their dissemination dynamics.

Propagation Speed: Analyze how quickly false information spreads through the network. We measure propagation speed by the number of retweets per hour, using the following formula:

$$Tweet\ Propagation\ Speed\ = \frac{Total\ Number\ of\ Retweets}{Time\ Span\ (in\ Hour)}$$

3.2.2 Conversation Threads (Reply Trees)

In X, a user can post a tweet (we call this a source tweet or root tweet), and other users who see the tweet can reply to it. A conversation thread consists of a source tweet and all its replies (users' comments). Replies can be direct if a user replies to the source tweet, or indirect if a user is replying to another user (other than the root user) in the same

conversation. Table 5 presents a sample conversation including a source tweet and users' replies to it. Analyzing users' conversations provides valuable insights into what users think or how they react to a topic. Users express their emotions and opinions about a tweet through replies. For example, a user may provide a negative, neutral, or supporting comment on the source tweet. In addition, replying to a tweet requires the user to invest additional effort in composing an original response, whereas retweeting requires relatively little effort. Comparing the rate of replies and the rate of retweets may give an indication of the level of engagement of users with that tweet. That is, a tweet with a higher ratio of replies might be considered more influential than a tweet with retweets alone.

Previously, X API (version 1) did not provide the functionality to collect reply trees in their entirety. The main technical challenge in collecting tweets related to the same conversation was that the X API only provided a link from the reply to the original tweet, but not vice versa. Thus, given a root tweet, one could not simply query for all subsequent replies. Therefore, prior studies used several other methods to build conversation threads, e.g., developing scrappers to collect reply trees (Garland et al., 2022) or collecting all posts and mentions, scanning the entire dataset and using the reply-to field to link posts to replies recursively (Saveski et al., 2021). Recently, the X API version 2 provided the "conversation_id", a unique identifier of the conversations[5]. When Tweets are posted in response to a Tweet (known as a reply), or in response to a reply (i.e., reply to a reply), there is now a defined conversation_id on each reply, which

[5] https://developer.X.com/en/docs/X-api/conversation-id

matches the Tweet ID of the original Tweet that started the conversation. Using the new version of the X API (version 2), it is possible to retrieve and reconstruct an entire conversation thread (reply trees) to better understand what users said and how conversations and ideas evolve around a topic.

Table 5: Example Conversation including a source tweet (first row) and its replies (all subsequent rows)

False Tweet	***James Delingpole: Chloroquine known as effective against coronavirus since 2005***
Replies	Zinc's available in supplement form. Chloroquine delivers zinc more effectively through the human cell wall to shut off the human cell's copying machine that the virus takes advantage of.
	Is corona virus known to this medication on 2005?
	The problem is that zinc is an ion, and it has a difficult time crossing the lipid bilayer of the cell. Chloroquine has the ability to take the ion and transport it through the cell. Chloroquine enhances zinc inside of the cell.
	Chloroquine is a zinc ionophore. A virus invades the cells and hijacks the machinery that reproduces. It then starts to make copies of itself. Zinc has the potential to kill viruses because it shuts down the "copy machine" so that it can't reproduce anymore.
	Hello! There was no Corona Virus in 2005!
	Not verified medically yet, but being looked at.
	Vitamin D3 plays a vital part in natural protection against URTI's and most in Northern latitudes suffer a deficiency in the winter. If the sun is shining get out and get some. Supplement of 10k units or 10ug/day also helps. It will provide 20%greater protection against infection
	Why isn't it being used then? It's bewildering

3.3 Data Collection Methodology

Primary Data: The primary data in this research is fake news about COVID-19, a list of tweet IDs (false tweets) labelled based on the type of false information (labels: unreliable, conspiracy, clickbait, political/biased). This primary data is collected by Sharma et al.,

(2020) and contains ~ **65k source tweets** (i.e., tweets that are not retweets or replies). They used the X streaming API to fetch tweets related to COVID-19 and categorized/labelled fake news into four types: unreliable, conspiracy theories, clickbait, and political/biased. They categorized false information based on three fact-checking sources: Media Bias/Fact Check[6], NewsGuard[7], and Zimdars (2016). For Media Bias/Fact Check, they included the list of questionable news sources with reported low factual content into the unreliable categorization. They included news sources listed by NewsGuard for publishing false content related to COVID-19 into the unreliable categorization. In the case of Zimdars (2016), they included tags fake, rumour, unreliable, and satire in the unreliable categorization. They included tags conspiracy and junksci (pseudoscience, naturalistic fallacies) in the conspiracy categorization, clickbait tag in the clickbait categorization, and tags bias and political to the political/biased. The collection period was from March 1, 2020, to June 5, 2020 (our data is from March 1, 2020, to April 24, 2020).

3.3.1 Data Collection Methodology for Retweets

Due to X policy, the USC researchers could only share tweet IDs with us. I used the X API and Twarc Python library to get metadata about tweets such as the tweets' text, public metrics of tweets and users (e.g., number of likes, number of retweets, number of followers, etc.). I also collected **all retweets** (to build retweet cascades) and **replies** (to create conversation threads) for my tweet IDs in our primary data. Using X API, it is only

[6] https://mediabiasfactcheck.com/
[7] https://www.newsguardtech.com/covid-19-resources/

possible to get the 100 most recent retweets for a list of tweet IDs, not all retweets). However, getting all retweets for a list of users is possible. My workaround (see Figure 4) was to extract all unique users and collect all retweets for the users. Next, I filtered all non-relevant retweets and kept only the retweets that their source tweet existed in the primary data. Figure 5 shows the methodology for collecting all retweets of the tweets in our primary dataset, and Table 6 shows a summary of the dataset statistics. Below is the summary of the steps I followed to collect all retweets for the source tweets in my data:

- Out of ~**65k source tweet** IDs in the primary data, I could collect metadata for 40,552 tweets (this is because some tweets or accounts may have been removed).
- Next, I only **kept the tweets with more than one retweet.** This filter is rational because we want to understand the spread of false tweets. If a tweet has no retweet, it means it didn't spread. Among the ~40k source tweets, only ~5k have more than one retweet.
- Among those ~5k tweets with more than one retweet, there are **2422 unique users**. Therefore, I collected all retweets for only these 2422 unique users.
- Finally, I removed all extra retweets of users and kept only the retweets for which their source tweet matched the tweet IDs in our original data, i.e., the tweet IDs for which we wanted to get retweets. (Recall, that the X API didn't allow us to collect all retweets for a list of tweets directly, but we could collect retweets for a list of users (authors of those tweets).

Since some users have thousands of retweets, the total number of retweets for 2422 users is enormous. To save time and space, I split the file containing usernames into multiple

files, each containing 100 users, and processed 100 users at a time. For example, as shown in Figure 5, there were over 2 million retweets for the first file (100 users). As explained earlier, we are only interested in users' retweets whose source tweets exist in our primary dataset. This way, the final number of retweets (matched retweets) for the first file (100 users) is reduced from ~2 million (~8GB) to ~23k (~12 MB). In other words, among the 2 million retweets of the first 100 users, only ~23k retweets are relevant for our study (the source tweets of only ~23k retweets exist in our primary COVID-19 dataset).

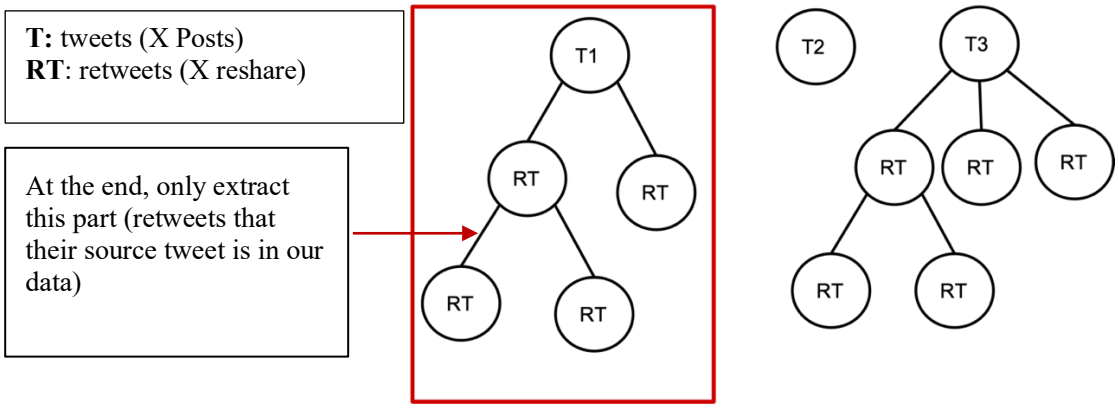

Figure 4: Illustration of extracting all retweets of the source tweets in our dataset. For example, suppose a user has three tweets in her profile: T1, T2, and T3. Using X API, we could get all retweets of the users (authors of the source tweets). Then, we kept only those retweets that their source tweet existed in our dataset (in this example, we assume that only T1 exists in our primary COVID-19 dataset).

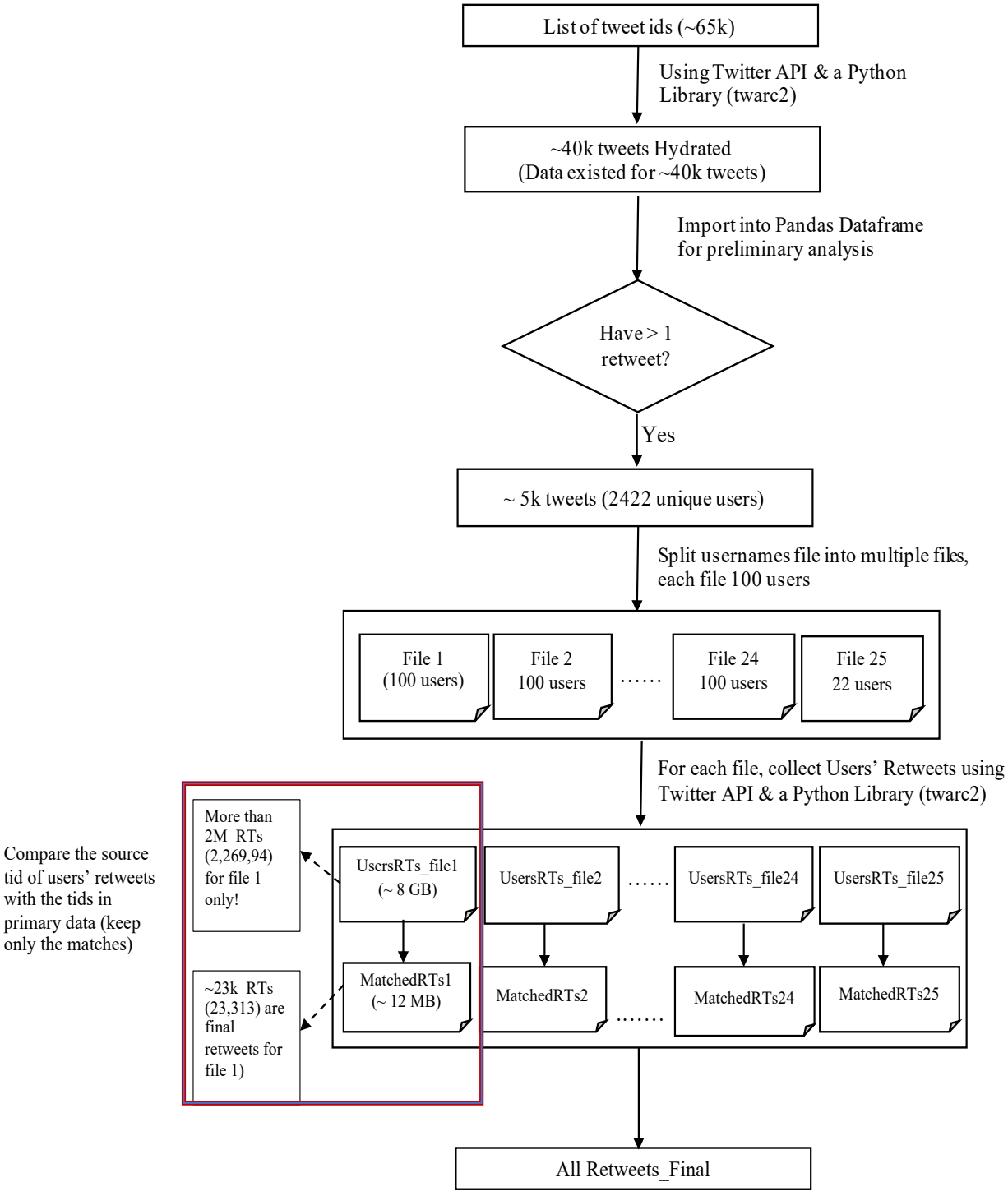

Figure 5: Methodology for collecting all retweets of the tweets in our primary COVID-19 dataset.

3.3.2 Data Collection Methodology for Replies (Conversations)

I used the X API (v2) and a Python library (Twarc) to collect all replies to the source tweets in my primary data. As mentioned in section 3.2.2, the new version of X API (v2) provides a new field, "conversation id", which is a unique identifier for each conversation (i.e., a source tweet and all its replies have the same conversation ID). Below is the summary of the steps I followed to collect replies and build the conversation threads:

1. Import tweet Object into Pandas dataframe.
2. **Condition**: tweets that have at least one reply & are not replies to other tweets
3. **Result**: Out of ~40k tweets, 3652 tweets > 1 reply and are not replies themselves
4. **Collecting Replies**: Using twarc2 python library and conversation_id (the tweet id of the source of conversation) → Result stored in a CSV file (replies.csv)
5. There are **26,673 replies**, out of which **19,443 are direct replies.**
6. Out of ~26k reply tweets, there are **2899 conversation threads** (similar to retweet cascades, conversation threads are reply tweets with a common single origin/source tweet)

Figure 6 presents the methodology for collecting all replies to the source false tweets in our primary dataset.

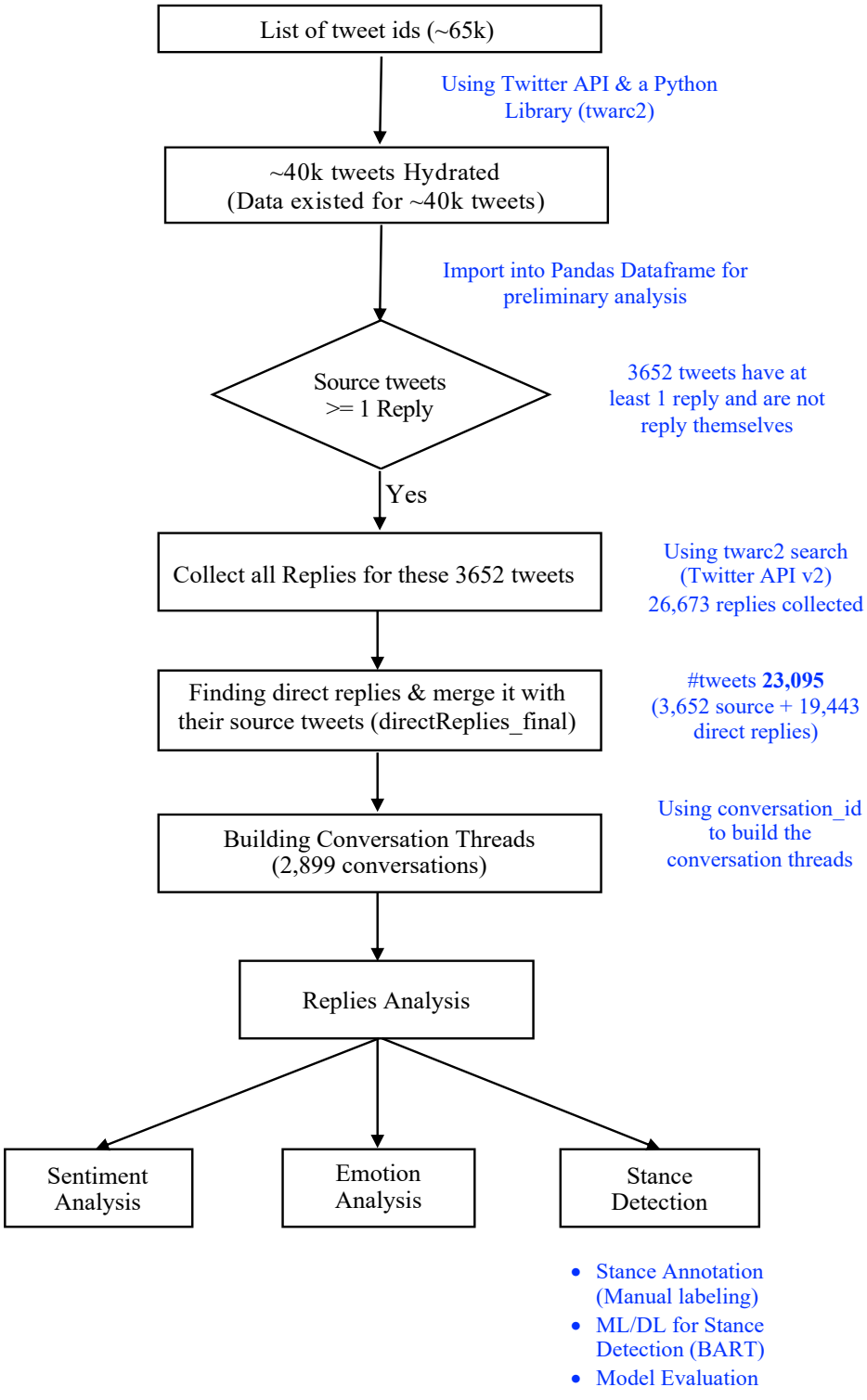

Figure 6: Methodology for collecting replies of the tweets (and building conversation threads) in our primary Covid-19 dataset, and conversation analysis.

3.3.3 Datasets Statistics

3.3.3.1 Retweets Dataset Statistics

For this dissertation, I collected retweets for a subset of the data (the first five files, which correspond to retweets of ~500 users). Table 6 and Table 7 provide summary statistics of the primary data and retweet dataset (data for which we collected retweets), respectively.

Table 6: Primary Dataset Statistics

number of tweets (primary data)	~ 65k
number of hydrated tweets	~ 40k
number of tweets with more than 1 retweet	5254
number of unique users (authors of tweets with >1 retweet)	2422

Table 7: Summary Statistics of the Collected Retweet Dataset

	File 1	File 2	File 3	File 4	File 5	Total
all retweets (RTs)	2,269,94	2,084,080	1,803,709	2,033,482	953,803	~ 9M
source tweets	~ 85k	135,725	64,975	89,077	68,692	~ 450k
retweeters	~ 400k	310,613	298,986	344,858	249,703	~ 1.6M
matched source tweets	763	1387	1733	2092	195	6170
RTs of matched tweets	23,313	9508	9315	5838	2245	50,219
unique retweeted users	88	90	82	89	80	429

There are about 9 million retweets for the first five files, whereas only a few thousand (50,219) are relevant to our study. These are users' retweets that their source tweet existed in our primary data (i.e., the COVID-19 tweets for which we wanted to collect retweets). We collected all their retweets for 6170 (source) tweets.

Regarding the types of false information, each tweet can have multiple types. For example, a tweet can be labelled clickbait and political/biased. Table 8 shows the number of source tweets for each category (type of false tweets). Since we are interested in analyzing the spread of false tweets, we only considered tweets that have more than one retweet (because if a tweet has no retweet it didn't spread. Also, to build retweet cascades, it makes sense to consider tweets with more than one retweet). Table 9 presents summary statistics of source tweets with more than one retweet, by type of false information.

Table 8: Count of each type of false tweets in the hydrated source tweets

Types of False Tweet	Count
clickbait, political/biased	12,144
unreliable	**10,500**
unreliable, political/biased	6,162
unreliable, conspiracy, political/biased	4,237
political/biased	**2,459**
conspiracy	**2,132**
unreliable, clickbait	1,797
conspiracy, political/biased	427
unreliable, clickbait, political/biased	399
unreliable, conspiracy	146
clickbait	**71**
conspiracy, clickbait	45
conspiracy, clickbait, political/biased	30
unreliable, conspiracy, clickbait	3
Total hydrated tweets	**40,552**

Table 9: Number of each type of false tweet for tweets with more than one retweet

Types of False Tweet	Count
clickbait, political/biased	1,949
Unreliable	**1,403**
unreliable, political/biased	691
unreliable, conspiracy, political/biased	430
unreliable, clickbait	291
political/biased	**258**
Conspiracy	**136**
unreliable, clickbait, political/biased	42
conspiracy, political/biased	37
unreliable, conspiracy	14
Clickbait	**4**
conspiracy, clickbait, political/biased	3
conspiracy, clickbait	2
unreliable, conspiracy, clickbait	0
Total (tweets with > 1 retweet)	**5,260**

Table 10: Number of Retweets and Cascades for each type of false tweet

Type of false tweet	Retweets Count	Cascades Count
total RTs (retweets of matched tweets)	50219	2287
political/biased	34307	1554
clickbait	32489	1288
unreliable	19625	1030
conspiracy	1149	112
only-unreliable	10039	559
only-political/biased	1,702	117
only-conspiracy	129	18
only-clickbait	2	0

Table 10 shows the number of retweets and the number of cascades for various kinds of false information in our data. Please note that Table 8 and Table 9 provide summary statistics of the source tweets, whereas Table 10 provides similar information for the retweets. As mentioned earlier, a total of 50219 retweets were collected. Unfortunately, we couldn't collect metadata (such as retweets) for all tweet IDs in our dataset because of the recent changes in Twitter API policy (currently, there is no free API to collect all the metadata we need for this research). As shown in Table 10, "political/biased" and "clickbait" tweets have the largest number of retweets. Please note that the "only" prefix before labels refers to exclusive types. For example, the label "only-conspiracy" refers to all tweets that are only labelled as "conspiracy", whereas the "conspiracy" label is inclusive, i.e., it can include tweets that are a mix of conspiracy and other types such as "conspiracy, unreliable", "conspiracy, clickbait", etc. In our data analysis, we mainly focus on the exclusive types (i.e., only-political, only-conspiracy, etc.) for comparative analysis. In terms of the number of cascades, there are 1639 retweet cascades in total.

There are more retweet cascades for tweets of type “political/biased”, “clickbait”, and “unreliable” compared to “conspiracy”. However, this is mainly because there are a much smaller number of conspiracy tweets in our data than other types.

3.3.3.2 Replies Dataset Statistics

As mentioned earlier, tweet labels (types of false tweets) are not mutually exclusive, and each tweet can have more than one label. For instance, a tweet which exhibits characteristics of “politicalBiased” and “clickbait” categories is labelled as “politicalBiased, clickbait”. Since this study aims to understand the nuances among diverse types of false information, I only include tweets that belong to one category. The only exception is the “clickbait” tweets for which I considered “clickbait, politicalBiased” and “clickbait, unreliable” tweets because there are very few “only-clickbait” tweets (only two tweets with two replies). Table 11 presents the summary statistics of tweets we use for user conversation/replies analysis (tweets with at least one reply).

Types of False Tweet	Count	#Replies	#Conversations
clickbait politicalBiased	**1254**	**7910**	**1002**
unreliable	**997**	**3477**	**793**
unreliable politicalBiased	507	2881	373
unreliable conspiracy politicalBiased	354	827	232
unreliable clickbait	**190**	**3091**	**157**
politicalBiased	**165**	**915**	**115**
conspiracy	**112**	**185**	**79**
conspiracy politicalBiased	30	86	21
unreliable clickbait politicalBiased	25	56	17
unreliable conspiracy[8]	11	8	8
conspiracy clickbait politicalBiased	4	4	3
clickbait	2	2	2
unreliable conspiracy clickbait	1	1	1
Total number of tweets with replies	**3,652**	**19,443**	**2,803**

Table 11: Summary statistics of tweets with at least one reply, by type of false information

3.4 Data Analysis

This section provides data analysis results to answer research questions. As mentioned before, this research compares different types of false information in terms of some propagation characteristics and users' responses to each type. Thus, the analysis has two main parts: 1) propagation analysis and 2) users' response analysis (in terms of users' emotions and attitudes). Regarding users' emotions, I use sentiment and emotion analysis, and for users' attitudes, I use users' stance analysis.

[8] The reason why the number of replies for this type is less than the number of tweets (which means some tweets do not have replies) is that at the time of data collection those tweets had replies while later (when collecting replies), some replies may have been deleted.

3.4.1 Propagation Analysis (RQ1)

In this section, we present the results of some exploratory data analysis. First, let's look at the overall tweet timeline during the tweet data collection period (i.e., from March 9, 2020, to April 24, 2020). Figure 7 shows the number of tweets (with 60 minutes frequency) over the seven weeks (47 days).

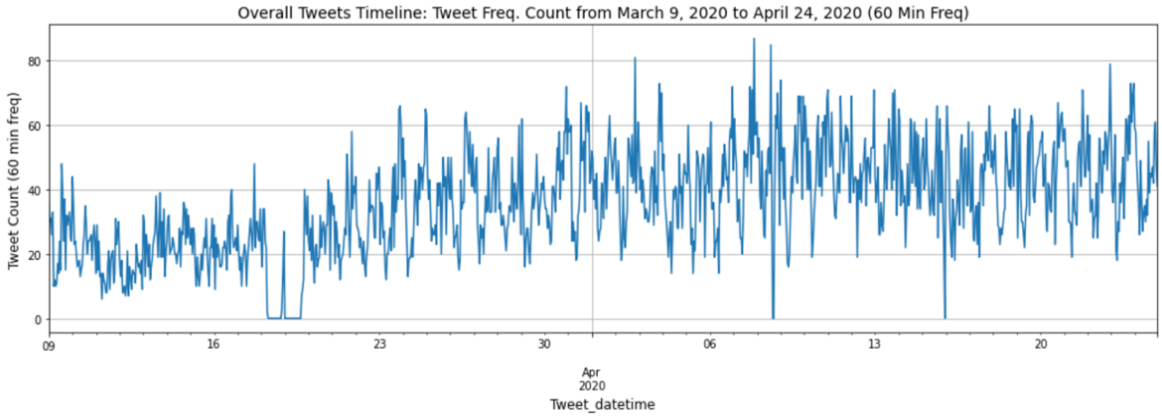

Figure 7: Overall (source) tweets timeline in our dataset

Figure 8 shows the distribution of tweets lifetime (in hours), which shows the retweeting period and is defined as the time distance between the source tweet posting time and the time of the last retweet. As can be seen, most retweets occur within the first couple of hours after the source tweet is posted, and the lifetime of most tweets is less than a day (there are almost no retweets after 24 hours. In other words, most tweets die after ~ 24 hours).

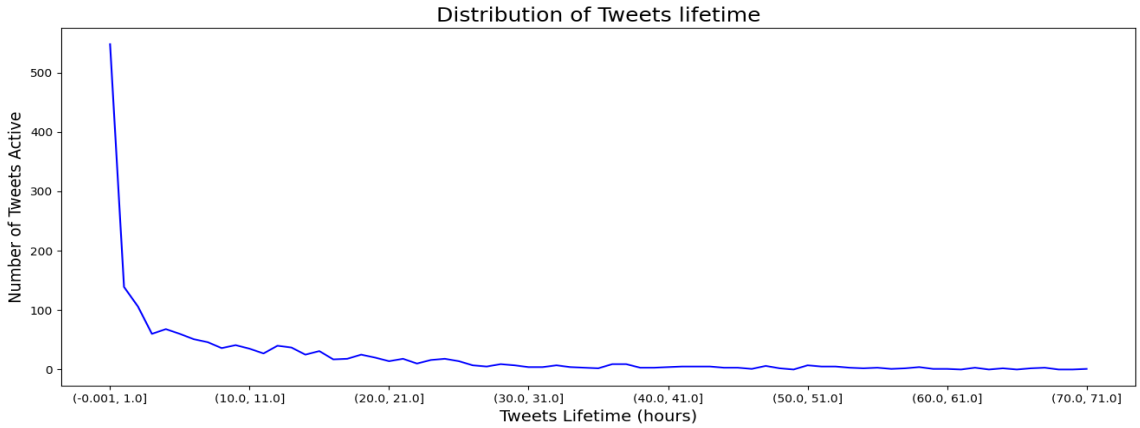

Figure 8: Distribution of (source) tweets lifetime

3.4.1.1 Retweet Cascades Analysis

In this section, we compare retweet cascades for different types of false tweets in terms of cascade size and lifetime. We computed cascade size as the number of users in the cascade, i.e., the number of unique users retweeting a tweet (in X, each user can retweet a tweet only once), and cascade lifetime as the time distance between the source tweet and the last retweet in the cascade.

3.4.1.1.1 Central Tendency Measures

Table 12 provides the central tendency measures for three types of false information: "only-political," "only-unreliable," and "only-conspiracy." We also included "all false types" (all false tweets regardless of their type) to better compare each specific type with false tweets in general (i.e., when all false tweets are lumped into one "false" category without differentiating between various kinds).

Table 12: Central Tendency Measures for Cascade Size and Cascade Lifetime

	CASCADE SIZE			CASCADE LIFETIME(HOURS)		
	Mean	Median	Std	Mean	Median	Std
All False Tweets	21.95	5.00	88.25	14.91	4.01	44.17
Only-Political	30.16	7.00	129.45	18.77	7.48	33.11
Only-Unreliable	17.96	4.00	118.93	13.75	4.22	32.93
Only-Conspiracy	7.17	3.00	8.54	49.14	24.99	58.14

Cascade Size:

The mean cascade size for only-political false tweets (30.16) is considerably higher compared to only-unreliable (17.96), only-conspiracy (7.17), and all types combined (21.95). This means that, on average, *political false tweets tend to have larger cascade sizes* than other types. In other words, there are more users engaged in political cascades. Also, the "only-conspiracy" false tweets have the lowest mean cascade size, indicating less widespread dissemination. The median cascade size is relatively consistent across different types of false tweets, suggesting that regardless of the type, a significant portion of false tweets have relatively small cascade sizes. Also, the median cascade size for all four types of false tweets is generally lower than the mean, indicating that there are *a few large cascades skewing the mean towards higher values*. Finally, in terms of variability, "only-political" false tweets have the highest standard deviation, indicating a higher variability than other types, and "only-conspiracy" false tweets have the lowest standard deviation, showing less variability and a more consistent spread of cascade sizes.

Cascade Lifetime:

As shown in Table 12, only-conspiracy false tweets have the longest mean and median cascade lifetime (49.14 and 24.99 hours), followed by only-political (18.77, 7.48 hours). This indicates that ***conspiracy false tweets generally live longer than other types of false tweets***. Also, the only-conspiracy false tweets have the highest variability in cascade lifetime (Std = 58.14 hours), indicating a wide range of lifetimes. Similar to cascade size, there is a notable difference in the range of cascade lifetime across different types of false tweets, with only-unreliable false tweets having the broadest range (0.0 to 234.83 hours), indicating a wide range of lifetimes.

3.4.1.1.2 Boxplots

We also provide boxplots of cascade size and lifetime for different types of false tweets. Figure 10 presents the boxplots of cascade size and lifetime for different types of false tweets.

Cascade Size:

We found that only-political false tweets have a higher average cascade size, indicating a more widespread dissemination than other types. The boxplots further support this, as the median line for only-political is positioned higher within the box than other types, suggesting a larger data spread for this type. Also, the outlier points above the upper whisker line show the occurrence of larger cascades in the political category. In contrast, only-conspiracy false tweets have a much narrower spread, which is evident by the smaller size of the box and the shorter length of the whisker lines. The median line for the only-conspiracy type is closer to the bottom of the box, which shows a concentration of smaller cascade sizes. Also, the upper whisker is longer than the bottom whisker,

suggesting a positively skewed distribution with longer tails towards larger cascade sizes. Moreover, the absence of outliers and the higher minimum point suggest a more limited range of cascade sizes for only-conspiracy false tweets. Finally, only-unreliable false tweets have a more moderate spread of cascade sizes.

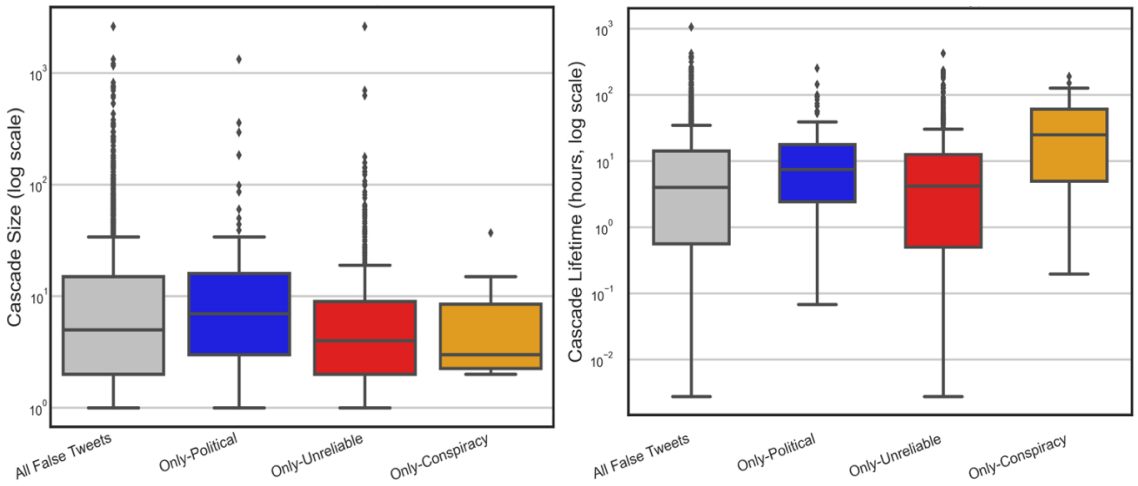

Figure 9: Boxplots of cascade size (left) and lifetime (right) for different types of false tweets

Cascade Lifetime:

The boxplots of cascade lifetime (Figure 9, right) further support the results of central tendency measures. The box for conspiracy false tweets is located higher, indicating longer lifespans of this type. Also, the median line for conspiracy false tweets is higher than the other types, further confirming the longer cascade lifetime of this type compared to other types of false tweets. In contrast, on average, political and unreliable false tweets generally have shorter cascade lifetimes. This is shown by their median line within the box, which is located lower than the conspiracy. Finally, the distribution of cascade lifetime for conspiracy false tweets is a positively skewed median. The median line for

the conspiracy type is closer to the top of the box, which shows a concentration of data points toward longer lifespans.

Overall, these observations highlight the differences in the spread of different types of false information distributions, with ***political false tweets having larger cascade sizes*** and ***conspiracy false tweets having longer lifetimes*** than other types of false tweets.

3.4.1.1.3 Probability Density Function (PDF)

While boxplots show the data's central tendency and spread, PDF plots provide a more detailed view of the distribution by showing the probability density of different values. PDF plots allow a more nuanced understanding of the distribution's shape and characteristics. Figure 10 and Figure 11 present the PDF of cascade size and cascade lifetime for different types of false tweets, respectively.

Cascade Size:

The PDF plots of all four types of false tweets show a positively skewed distribution, with most tweets having a smaller cascade size and a longer tail towards larger cascade sizes. Overall, political false tweets exhibit larger cascade sizes, ***while conspiracy false tweets tend to generate cascades with smaller sizes***. This is in line with our findings from central tendency measures and our observations from the boxplots.

In the PDF plots of only-political false tweets, the main tall peak suggests a cluster of small cascade sizes, which is common for viral content (most political false tweets have a cascade size of less than 200). There are also two small bumps, indicating small clusters of larger cascade sizes. The long tail suggests that a small number of highly viral political

false tweets result in large-scale propagation. In contrast, the PDF plot of only-conspiracy false tweets shows two distinct peaks (at x=~5 and x=~35), indicating a bimodal distribution. Also, a significant portion of conspiracy false tweets have smaller cascade sizes.

Overall, the PDF plots reveal that false information, especially political and unreliable content, is more likely to reach a larger audience, ranging from hundreds to thousands of people, similar to the top percentiles of false news cascades (Vosoughi et al., 2018).

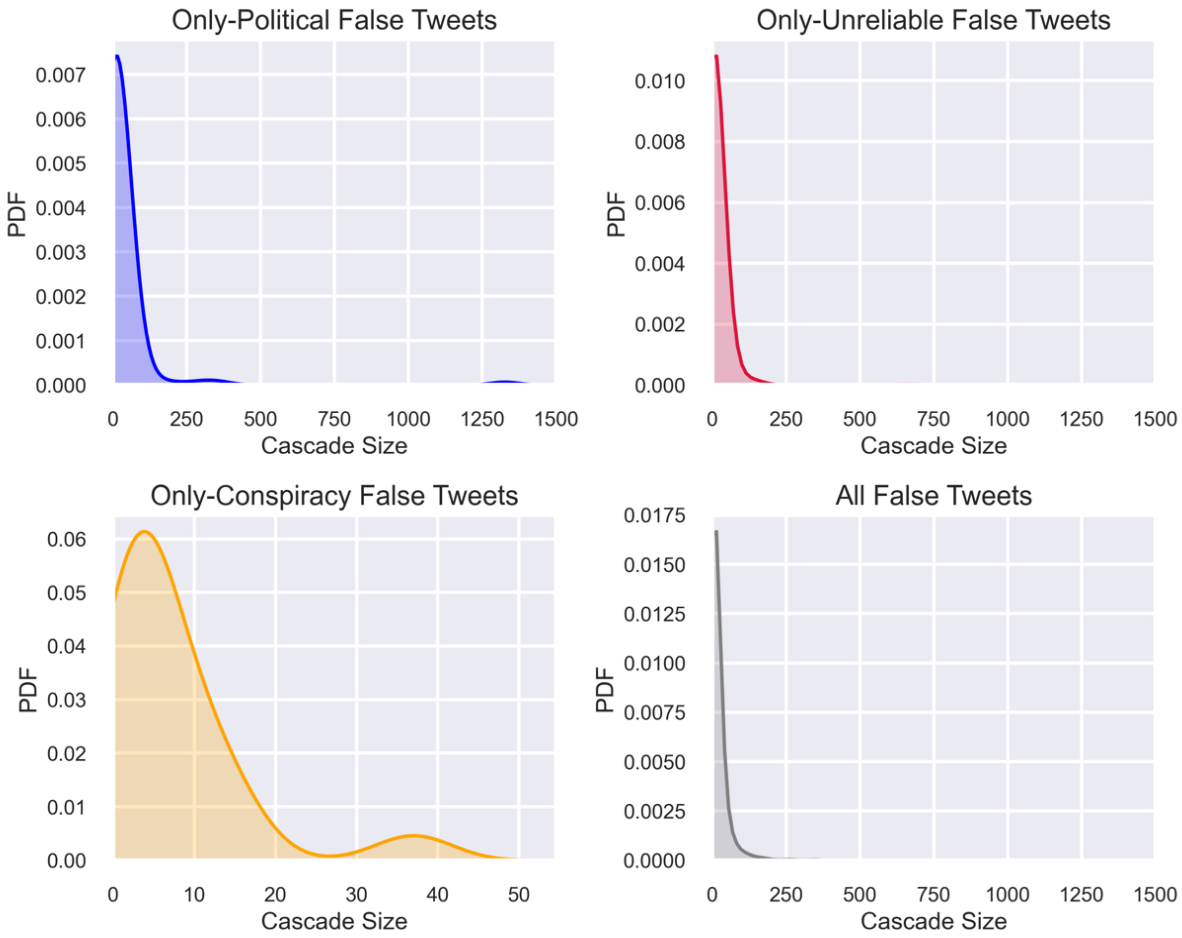

Figure 10: PDF of cascade size for different types of false tweets. Cascade size is the number of unique users in the cascade (users retweeting a tweet)

Cascade Lifetime:

Overall, Figure 11 reveal differences in the distribution and variability of cascade lifetimes among different types of false tweets. The PDF plot of political false tweets shows a peak at the beginning followed by a sharp decline, indicating a concentration of cascade lifetimes with shorter lifetimes. The unreliable and all false tweets have distributions similar to political but with a narrower shape, which suggests less variability in their lifetimes. In contrast, the PDF plot for conspiracy false tweets has a wider shape with a more gradual decline, which indicates a broader range of cascade lifetimes

compared to other types of false tweets. On average***, the lifetime of conspiracy false tweets is notably higher than other types of false tweets***. For example, the PDF values for conspiracy tweets in the range of X (lifetime) > 50 hours are higher than other types of false tweets. This indicates a higher probability of observing longer cascade lifetimes for conspiracies, suggesting that conspiracy cascades are more likely to maintain users' engagement over a longer period.

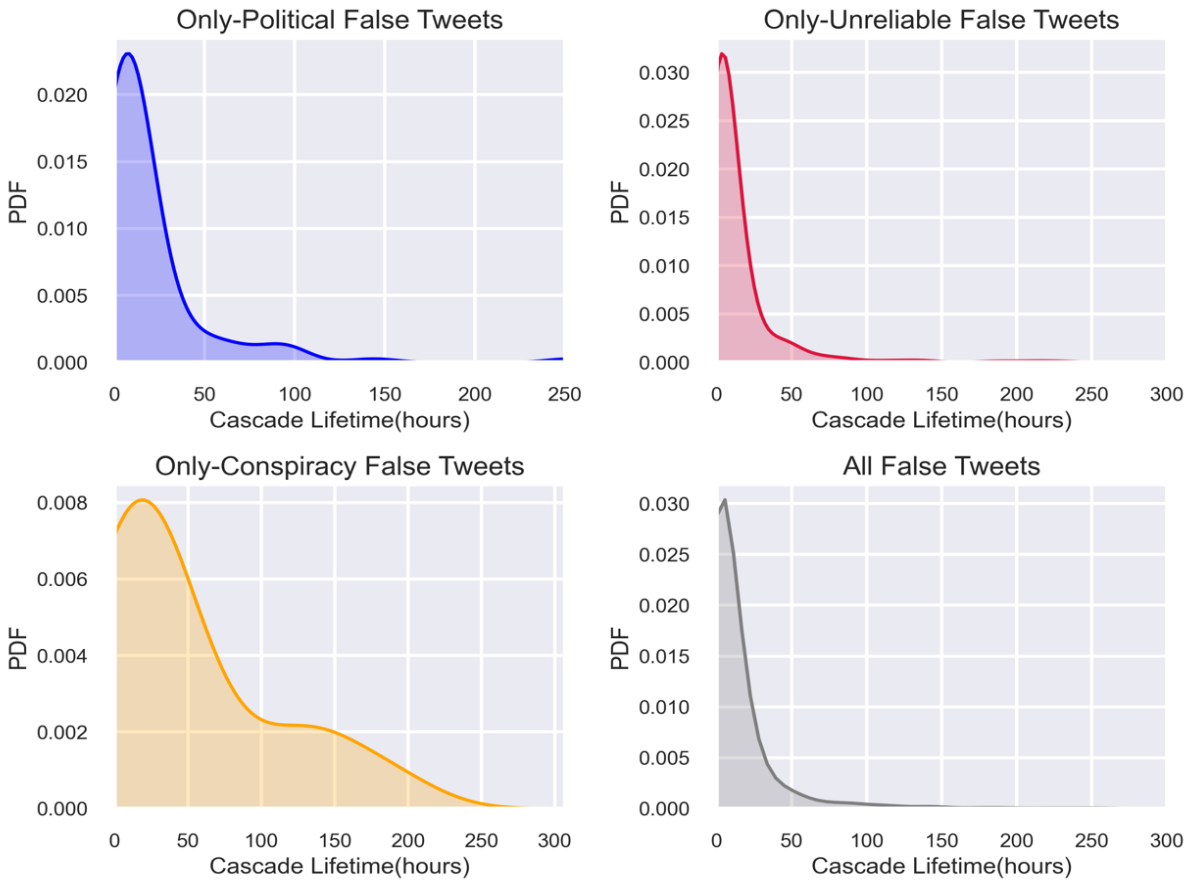

Figure 11: PDF of cascade lifetime for different types of false tweets. Cascade lifetime is computed by the time distance (in hours) between the source tweet and its last retweet.

3.4.1.1.4 Complementary Cumulative Distribution Function (CCDF)

Figure 12 provides a comparative analysis of cascade size and lifetime among different types of false tweets. Regarding the cascade size (left figure), most retweet cascades for all three types of false tweets have a small size (less than ten users), which means for most retweet cascades, less than ten users are retweeting a tweet. In general, only-political false tweets have a larger cascade size, indicating that they reach more people. However, a small percentage of only-unreliable false tweets have larger cascade sizes (see the tail of the plots). Finally, conspiracy false tweets exhibit the lowest cascade size compared to other types of false tweets. These findings align with the boxplot results (Figure 9) and PDF plots (Figure 10 and Figure 11).

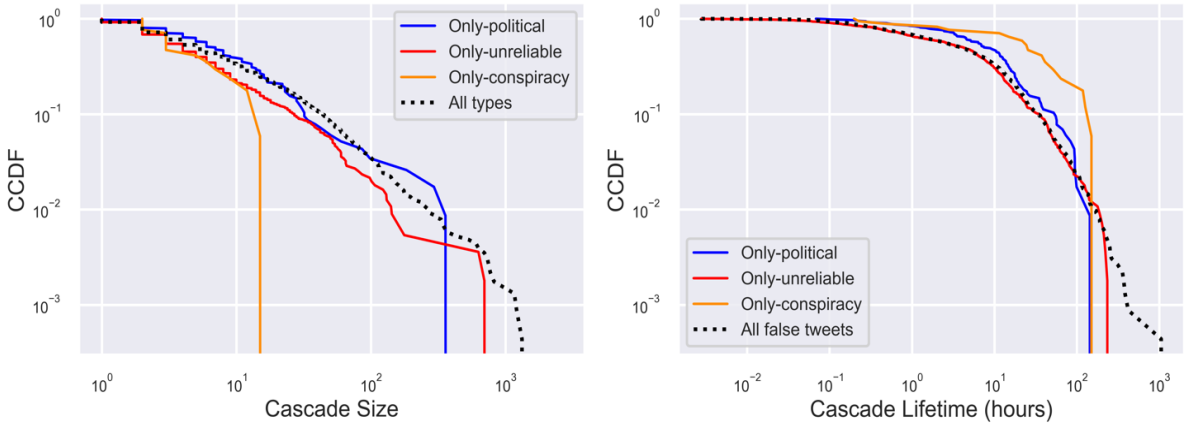

Figure 12: Comparison of different types of false tweets in terms of Empirical Complementary Cumulative Distribution Function (CCDF) of cascade size (left) and cascade lifetime (right)

In terms of cascade lifetime (Figure 12-right), only-conspiracy false tweets have notably the highest lifetime on average. This is shown by the higher CCDF curve of conspiracy cascades, indicating that more conspiracy false tweets have a higher probability of longer lifetimes (especially for lifetimes between ~10 and ~100 hours). However, looking at the

tails of the distributions, there is a small percentage of only-unreliable false tweets with considerably longer lifetimes. The tails in the CCDF plots represent the extreme values of the distributions.

3.4.1.2 Spread Speed Analysis

This section provides a comparative analysis of spread speed across different types of false tweets. We defined the spread speed as the rate at which tweets are retweeted (shared) over time, computed by dividing the number of retweets by the tweet lifespan (in hours). We can use the Complementary Cumulative Distribution Function (CCDF) plot to compare the spread speed of different false tweet types. The CCDF plot shows the probability that a random variable is greater than a given value. In the context of spread speed analysis, the CCDF of spread speed shows the probability that the spread speed of false tweets exceeds a specific value.

Figure 13 shows the CCDF of spread speed using absolute speed (left) and normalized speed values (right) for different types of false tweets: only-political, only-unreliable, and only-conspiracy. We included the CCDF of spread speed for all false tweets (regardless of the type) to better demonstrate the differences between the spread dynamics of each specific kind of false tweet and all false tweets in general.

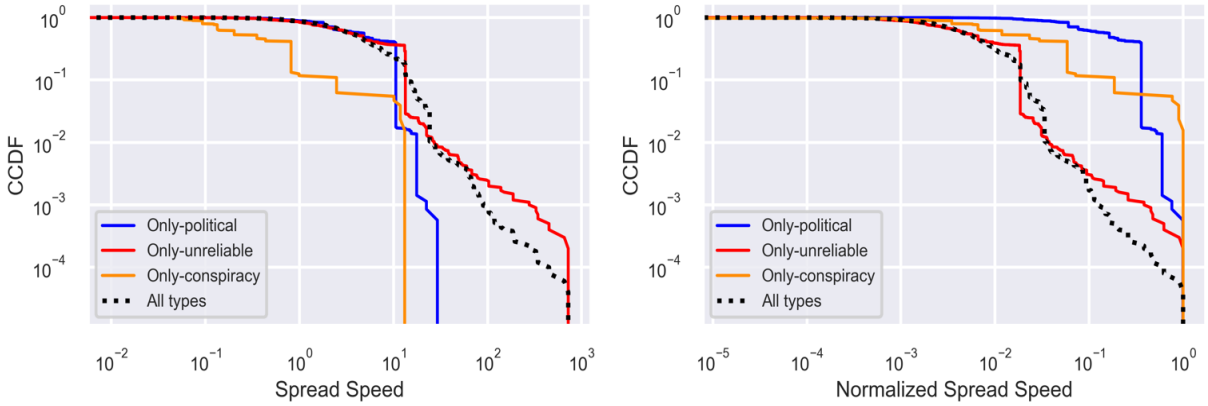

Figure 13: CCDF plots of spread speed using the absolute values (left) and normalized spread speed (right) for different types of false tweets

The CCDF plot of actual spread speed values shows absolute differences in dissemination rates, which allows for a direct comparison of spread speed across different types of false tweets. However, it may not provide a fair comparison, especially when sample sizes are significantly different, as it may be biased towards the larger group. In contrast, the CCDF plot of normalized spread speed values provides a fair comparison of the relative distribution of spread speeds across different types, regardless of the sample sizes. We provide both plots for a more comprehensive understanding of relative differences in spread speed across various types of false tweets.

CCDF of Spread Speed (Actual Values): In the CCDF plot of spread speed using actual values (Figure 13, left), the conspiracy false tweets exhibit the slowest spread speed, shown by its CCDF curve consistently lower compared to other types. This suggests that conspiracy false tweets tend to spread more slowly in terms of the absolute number of retweets per tweet lifetime. The slower spread speed of conspiracy types could be attributed to the complex nature of conspiracy content, which may require more time for

users to process or verify before sharing. The political and unreliable false tweets have similar CCDF at lower spread speeds (up to a speed of ~10). However, after that point, the CCDF curve for unreliable false tweets rises more rapidly, indicating its faster spread compared to other types. The unreliable false tweets exhibit a faster spread speed overall.

CCDF of Normalized Spread Speed: The CCDF plot of normalized spread speed for different types of false tweets (Figure 13, right) shows distinct spread dynamics among false tweet types. The unreliable false tweets consistently have the slowest normalized spread speed across the whole range of speeds. The political false tweets initially show a relatively higher spread speed, indicated by the higher CCDF curve. However, conspiracy false tweets eventually outpace political false tweets. This suggests that while most political false tweets spread faster, there are small percentages of conspiracy false tweets with the potential to reach a more intensive spread over time.

These findings underscore the nuanced differences between the spread dynamics across different types of false information. While prior research found that false information as a whole demonstrates significant reach compared to true information (Vosoughi et al., 2018), our analysis reveals differences in the spread patterns across different types of false information.

3.4.1.3 Statistical Tests: Kruskal-Wallis

So far, we have seen differences in the spread dynamics of different types of false information. We use the Kruskal-Wallis test to check whether the observed differences are statistically significant. The Kruskal-Wallis test compares the distributions of a

continuous variable (in the context of this study, cascade size, cascade lifetime, and spread speed) across multiple groups (in our case, different types of false tweets). The Kruskal-Wallis test is appropriate because it does not assume normality (This is advantageous when dealing with skewed or non-normal distributions) and can handle unequal sample sizes. It's often used as an alternative to the one-way analysis of variance (ANOVA) when the assumptions of ANOVA are not met. Table 13 presents the results of the Kruskal-Wallis test.

Table 13: Kruskal-Wallis test results

Variable	Kruskal-Wallis Statistic	p-value
Cascade Size	21.1313	9.88688e-05
Cascade Lifetime	26.4263	7.76512e-06
Spread Speed	255.975	3.33734e-55

***Cascade Size*:** The Kruskal-Wallis statistic is 21.1313 with a p-value of 9.88688e-05. This indicates a statistically significant difference in cascade size between at least two groups (types of false tweets).

***Cascade Lifetime*:** The Kruskal-Wallis statistic is 26.4263 with a p-value of 7.76512e-06. Similar to cascade size, this result shows a statistically significant difference in cascade lifetime between the groups.

***Spread Speed*:** The Kruskal-Wallis statistic is 255.975 with a p-value of 3.33734e-55. This extremely low p-value indicates a significant difference in spread speed between the groups. The high value of the Kruskal-Wallis statistic for spread speed compared to cascade size and cascade lifetime shows that there may be more variations or differences

in spread speed between different types of false tweets compared to cascade size and cascade lifetime.

Overall, the results of the Kruskal-Wallis test suggest statistically significant differences in cascade size, cascade lifetime, and spread speed across different types of false tweets, with the spread speed showing the most significant differences among the groups.

3.4.1.4 Statistical Test: Mann-Whitney U Test

The results of the Kruskal-Wallis test revealed significant difference differences in spread dynamics (cascade size, cascade lifetime and spread speed) across different types of false tweets. To determine which specific groups (types of false tweets) differ from each other, we also conduct a post-hoc test, pairwise Mann-Whitney U test. These tests help identify where the difference between different types of false tweets lies. The test results are provided in Table 14.

Table 14: Pairwise comparison of spread dynamics (cascade size, cascade lifetime, and spread speed) across different types of false tweets using the Mann-Whitney U test

	Cascade Size		Cascade Lifetime(h)		Spread Speed	
	U	p-value	U	p-value	U	p-value
All False Types vs only-Political	121176	0.0833	106754	0.0002	8.71486e+07	0.1006
All False Types vs only-Unreliable	707504	0.0001	657164	0.3028	2.46635e+08	0.0006
All False Types vs only-Conspiracy	23101.5	0.3679	11266	0.0009	5.7913e+06	0
only-Political vs only-Unreliable	39760.5	0.0002	40225	0.0001	1.64402e+07	0
only-Political vs only-Conspiracy	1311	0.0941	709	0.0262	405784	0
only-Unreliable vs only-Conspiracy	5069.5	0.9561	2665	0.0007	1.1591e+06	0

***Cascade Size*:** Overall, there is no statistically significant difference in cascade size between most types of false tweets. However, there are significant differences in cascade

size between "All False Types" and "only-Unreliable" (p-value = 0.001) and between "only-Political" and "only-Unreliable" (p-value = 0.002).

***Cascade Lifetime*:** In terms of cascade lifetime, there are significant differences in cascade lifetime between most types of false tweets (p-values < 0.001).

***Spread Speed*:** There are significant differences in spread speed between almost all types of false tweets (p-values < 0.001).

Overall, these results suggest that while there may not be notable differences in cascade size between most types of false tweets, there are clear distinctions in cascade lifetime and, especially, spread speed. Also, while the results of the Mann-Whitney U test show no statistically significant difference in cascade size between political and conspiracy false tweets (p-value = 0.0941), it is important to consider the U statistic as well, which is 1311. The U statistic indicates a moderate difference in cascade size between political and conspiracy false tweets. In addition, the CCDF plots of cascade size show a difference between conspiracy false tweets compared to both political and unreliable types (see Figure 12. We should note that while the results of the Mann-Whitney U test may not reveal statistically significant differences in cascade size between certain pairs of false tweet types, the CCDF plots provide a qualitative understanding of the differences in distributions of cascade size across different types of false. These differences may be due to several factors, such as the characteristics of the content, user, or other spread dynamics of various kinds of false tweets.

3.4.1.5 Statistical Test: K-S Tests

Table 15 presents the results of the Kolmogorov-Smirnov (K-S) test, which provides a pairwise comparison of the empirical CCDF of different types of false tweets: only-political, only-conspiracy, and only-unreliable.

Table 15: Pairwise comparison of eCCDF of different types of false tweets using the K-S test

Feature	Type 1	Type 2	Median Type 1	Mean Type 1	Std Type 1	Median Type 2	Mean Type 2	Std Type 2	D	p-value
#followers	onlyPol	onlyConspiracy	375944	666214	513913	11753	19308.5	23037.6	0.886937	3.46933e-115
#followers	onlyPol	onlyUnreliable	375944	666214	513913	141930	178968	177525	0.822825	0
#followers	onlyConspiracy	onlyUnreliable	11753	19308.5	23037.6	141930	178968	177525	0.674778	1.04141e-57
#followings	onlyPol	onlyConspiracy	859	30354.9	49960.1	4134	7499.94	6753.55	0.578575	6.59042e-40
#followings	onlyPol	onlyUnreliable	859	30354.9	49960.1	937	9533.08	27344.3	0.306208	8.25544e-218
#followings	onlyConspiracy	onlyUnreliable	4134	7499.94	6753.55	937	9533.08	27344.3	0.557685	1.02108e-37
#tweets	onlyPol	onlyConspiracy	68554	68766.2	66544.7	61579	97422.2	95207.6	0.320145	7.71203e-12
#tweets	onlyPol	onlyUnreliable	68554	68766.2	66544.7	40056	119920	209807	0.365339	6.81652e-313
#tweets	onlyConspiracy	onlyUnreliable	61579	97422.2	95207.6	40056	119920	209807	0.413185	3.89573e-20

The D statistic is defined as the absolute maximum distance between the CDFs of the two samples. The closer this number is to 0, the more likely the two samples were drawn from the same distribution. We reject the null hypothesis that the two samples were drawn from the same distribution if the p-value is less than our significance level. We choose a confidence level of 95%; that is, we reject the null hypothesis in favour of the alternative if the p-value is less than 0.05. The values of the D Statistic for the variables "#followers (followers count)", "#followings (followings count)", and "#tweets" show the distance between each pair of false tweet types tweets. Also, the p-values indicate significant differences in the distribution of #followers, #following, and #tweets between different pairs of false tweet types. Specifically:

Followers Count: The p-values suggest highly significant differences between political and conspiracy false tweets ($p < 0.001$), political and unreliable false tweets ($p < 0.001$), as well as conspiracy and unreliable false tweets ($p < 0.001$). These results show that the

distributions of followers count (#followers) significantly differ across all pairs of false tweet types.

Following Count: Similarly, we observe significant differences in the distribution of following count (#followings) between political and conspiracy false tweets ($p < 0.001$), political and unreliable false tweets ($p < 0.001$), and conspiracy and unreliable false tweets ($p < 0.001$). These findings suggest distinct patterns in the number of accounts followed by users engaging with different types of false tweets.

Tweet Count: There are significant differences in the distribution of tweet count (#tweets) between political and conspiracy false tweets ($p < 0.001$), political and unreliable false tweets ($p < 0.001$), and conspiracy and unreliable false tweets ($p < 0.001$). This implies that users interacting with different types of false tweets exhibit varying activity levels when posting their tweets.

Overall, the results suggest ***significant differences in user engagement patterns across different types of false tweets, particularly in terms of the number of followers, following, and tweets***. These differences may reflect underlying differences in the nature of false information, the characteristics of users sharing false content, and the spread dynamics of false information within social networks.

3.4.1.6 Why is the spread of different types of fake news likely to differ?

Our propagation analysis revealed that various types of false tweets exhibit different spread dynamics. Table 16 compares the spread characteristics of diverse types of false tweets.

Table 16: Comparison of the propagation characteristics of different types of false tweets.

Type of False	Cascade Size	Cascade Lifetime	Spread Speed
Political	**Largest** cascade size	Similar to Unreliable and All False	**Fastest** spread speed (Overall, most political false tweets spread faster than other types)
Conspiracy	**Smallest** cascade size	**Longest** lifetime (Most Conspiracies live between 10 to 100 hours. A small percentage of Conspiracies live more than 100 hours)	Small percentages of Conspiracies spread faster than all other types
Unreliable	Small percentages of Unreliable false tweets (less than 1%) have larger cascade sizes	Small percentages of Unreliable false tweets (less than 1%) have longer cascade lifetime	**Slowest** spread speed
All False	Similar to Unreliable	Similar to Unreliable	Similar to Unreliable

Cascade Size

Only-political false tweets tend to have larger cascade sizes on average (i.e., more users involved in political cascades). Conversely, conspiracy false tweets have smaller cascade sizes, suggesting a narrower reach but potentially deeper engagement among a more niche audience. The observed differences in cascade size among different types of false tweets can be due to various factors related to content, users' engagement, and platform. Political false tweets often contain polarized topics that attract widespread attention and discussion, leading to larger cascade sizes. Political content may also be more likely to cause strong emotions or reactions from users, increasing the likelihood of sharing.

Additionally, political false tweets may be supported by individuals or groups aiming to shape public opinion, further contributing to their larger cascade sizes. In contrast, conspiracy false tweets may have smaller cascade sizes due to their niche or fringe nature. Conspiracy theories often target specific groups or communities with shared beliefs or interests, resulting in a narrower audience base compared to political content. Additionally, conspiracy false tweets may face more skepticism or scrutiny from users, resulting in limited reach. We also observed lower variability in conspiracy cascades, which may reflect the relatively consistent engagement patterns within these niche communities, where information spreads gradually and among smaller networks.

Cascade-Lifetime

Conspiracy false tweets have the longest cascade lifetime on average, followed by political, unreliable, and a combination of all types. This indicates that conspiracy content may remain relevant for longer periods compared to other types of false information. The longer cascade lifetime for conspiracy false tweets may be due to the nature of conspiracy theories, which often involve *complex narratives that can sustain interest and engagement over extended periods*. In contrast, political and unreliable false tweets may have shorter lifetimes due to the rapid pace of political debates and the temporary nature of unreliable information. Also, the range of cascade lifetime varies widely between types, with only-unreliable false tweets having the broadest range. The wider range of cascade lifetime for unreliable false tweets may reflect the diverse nature of false information in this category, spanning from quickly debunked rumours to more persistent false narratives.

Spread Speed

Conspiracy false tweets often contain complex narratives or controversial claims which may need more time for people to process before they decide to share. Additionally, the content of conspiracy false tweets may be perceived as more contentious or questionable by users, leading to lower engagement and slower spread speeds. On the other hand, unreliable false tweets may include sensational or more emotional content to cause strong reactions from users, resulting in higher engagement and faster spread speeds. Finally, false political tweets often address current events, political issues, or ideological narratives. The spread speed of false political tweets may depend on factors such as content novelty and alignment with users' beliefs or political ideology.

Overall, the observed differences in the spread dynamics of different types of false tweets could be influenced by various factors such as the characteristics of the content (e.g., emotional appeal, sensationalism, etc.), user (e.g., engagement of influential users), or platform. To better understand these factors, we provide additional complementary analysis (e.g., user response analysis such as sentiment and emotion analysis and stance detection) in the remainder of this thesis.

In conclusion, our propagation analysis highlights the nuanced differences in the spread of different types of false information, suggesting the need for tailored approaches to combat false content based on its specific characteristics and propagation dynamics. For example, to combat political misinformation, we may need to focus on strategies to reduce the virality of false political content, while to counter conspiracy theories, we may

need to devise strategies to address individual's beliefs and motivations causing their propagation.

3.4.2 Users Response Analysis (RQ2)

In the previous section, we found that different types of false tweets have distinct spread dynamics (in terms of cascade size, cascade lifetime, and spread speed). In this section, we explore the content and users' responses to different types of false tweets to provide a more comprehensive understanding of the characteristics of various types of false information. The analysis in this section aims to answer our second research question: *How do users respond to different types of false information? Do responses differ among diverse types of false information?*

To answer RQ2, we conducted sentiment analysis, emotion analysis, and stance detection. These analyses provide a better understanding of the propagation analysis we provided in the previous section. First, sentiment analysis reveals the emotional tone of users towards false tweets, indicating whether they perceive the false content positively, negatively, or neutrally. Second, emotion analysis helps identify specific emotions in users' responses to false tweets, providing further context beyond sentiment. Finally, stance detection helps uncover users' attitudes and beliefs towards different types of false content (whether they agree, disagree, or neutral to a piece of content). Integrating these analyses with propagation metrics offers a more comprehensive understanding of how false information (in our case, different types of false tweets) spreads and how users engage emotionally and attitudinally.

First, I started by data cleaning and pre-processing the text. This includes removing nulls and duplicates, punctuation, special characters, URLs, etc. To answer research questions about how users react (emotionally and attitudinally) to fake news in general and to different types of fake news more specifically, I perform three types of analysis: *sentiment analysis* and *emotion analysis* to understand users' emotions about false tweets. Next, I use *stance classification* to identify users' attitudes (i.e., the position users hold towards a topic).

3.4.2.1 Sentiment Analysis

For sentiment analysis of replies, we used the VADER library. VADER (Valence Aware Dictionary and sEntiment Reasoner): a rule/lexicon-based, open-source sentiment analyzer pre-built library within NLTK. The tool is specifically designed for sentiments expressed in **social media**, and it uses a combination of a sentiment lexicon and a list of lexical features generally labelled according to their semantic orientation as positive or negative. VADER calculates the text sentiment and returns the probability of a given input sentence to be **positive**, **negative**, or **neural**. The tool can analyze data from various social media platforms, such as X and Facebook. Here are some of the main reasons to use VADER for sentiment analysis: 1) it does not require training data, 2) it understands the sentiment of text containing emoticons, slang, conjunctions, etc., 3) excellent for social media text, 4) it is an Open-source library.

Figure 14 shows the results of sentiment analysis for the content of different types of false tweets: only-political, only-conspiracy, only-unreliable, clickbait-political, and clickbait-unreliable. It can be observed that the most dominant sentiment for the contents

of almost all types of false tweets (except conspiracy) is negative, followed by neutral. The positive sentiment has the lowest percentage in the content of all types of false tweets.

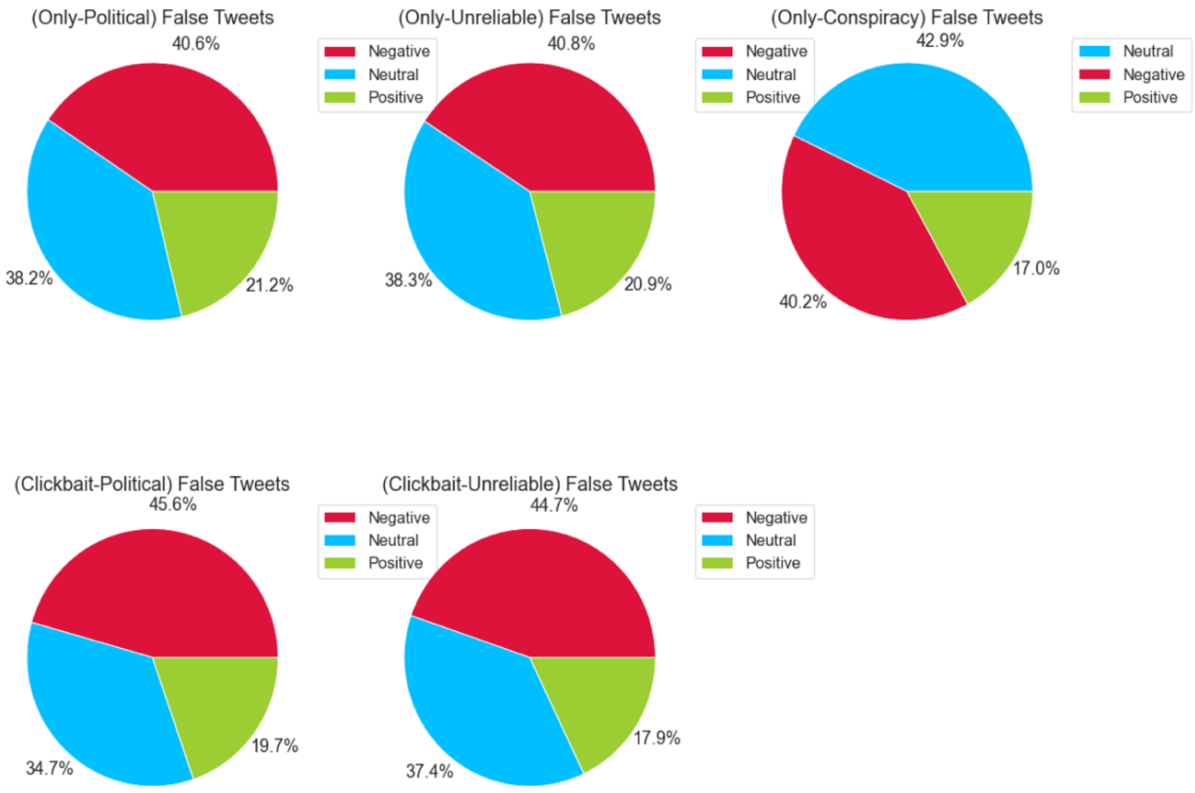

Figure 14: Sentiment of the Content of different types of false tweets

We also provided the sentiment of users' responses to different types of false tweets (please see Figure 15). While the dominant sentiment in the content of most types of false tweets is negative, the prominent sentiment in users' responses to those false tweets is neutral for most types of false tweets (except for the only-political type, where negative sentiment has the highest percentage).

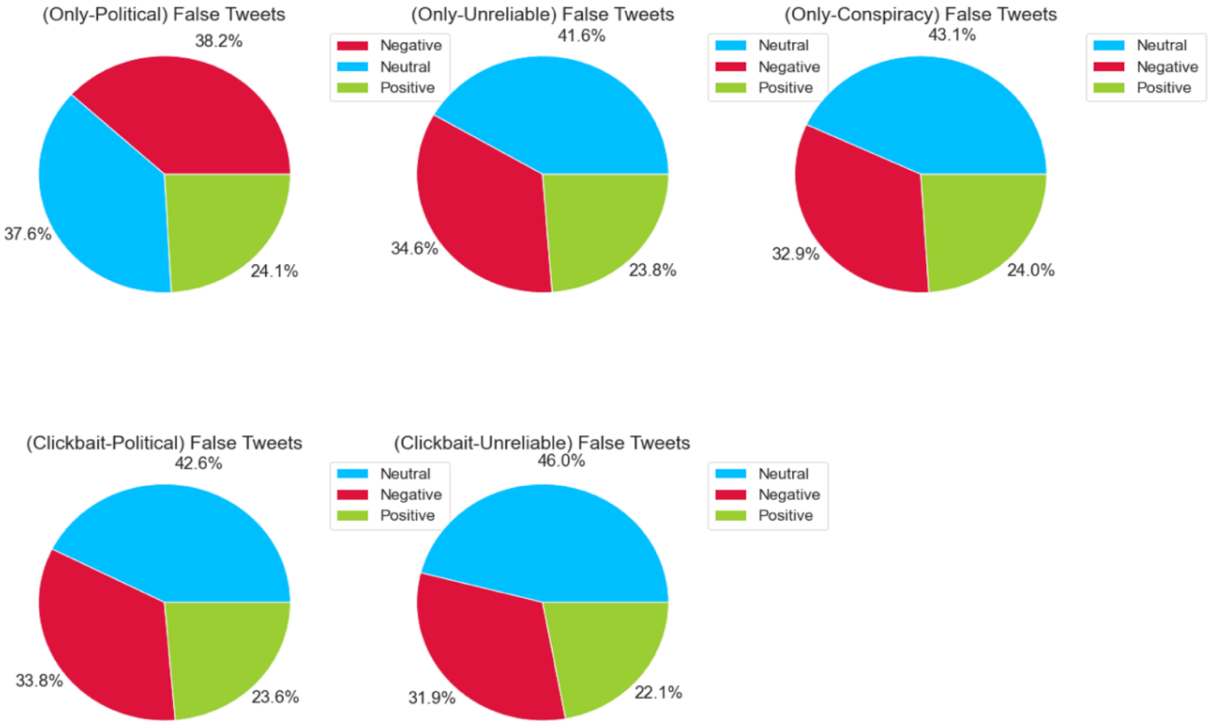

Figure 15: Sentiment of Users' Responses to different types of false tweets

3.4.2.2 Emotion Analysis

Emotion analysis is crucial to understanding users' responses to false tweets and discerning patterns in emotional expressions associated with different types of false information. This study employed the NRC (National Research Council) method for emotion analysis. The NRC method (Mohammad & Turney, 2013) is a lexicon-based approach that assigns a set of pre-defined emotions to words in a given text. The NRC lexicon is a comprehensive resource containing words annotated with emotion labels. Each word in the lexicon is associated with one or more emotions, capturing a broad spectrum of affective states. The NRC lexicon includes eight primary emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. The NRC method was applied for each false tweet in the dataset to extract emotion scores based on the frequency of

emotional words in the text. The scores indicate the intensity of each emotion expressed in the tweet. By aggregating these scores, a profile of emotional expressions for each type of false tweet (conspiracy, unreliable, political, clickbait-political, clickbait-unreliable) was created. Figure 16 presents the mean score of emotions for replies to different types of false tweets.

We generally observe that users' responses to various types of false tweets show different emotions. However, there are also some similarities. For instance, the two most dominant emotions in all three types of false tweets (“only-political,” “only-unreliable,” and “only-conspiracy”) are trust and anticipation. Also, the third most dominant emotion for "only-unreliable" and “only-conspiracy” false tweets is fear, and for "only-political" false tweets, it is *disgust*.

We can see that including a clickbait component increases certain emotions in users’ responses. For example, “clickbait-political” false tweets exhibit higher levels of almost all eight emotions, especially *fear, anger, sadness*, and *disgust*. Also, “clickbait-unreliable” false tweets evoke higher levels of emotions such as *anger, surprise, sadness,* and *disgust* compared to the “only-unreliable” category. This result aligns with the sensationalism nature of clickbaits, which evokes stronger emotional reactions.

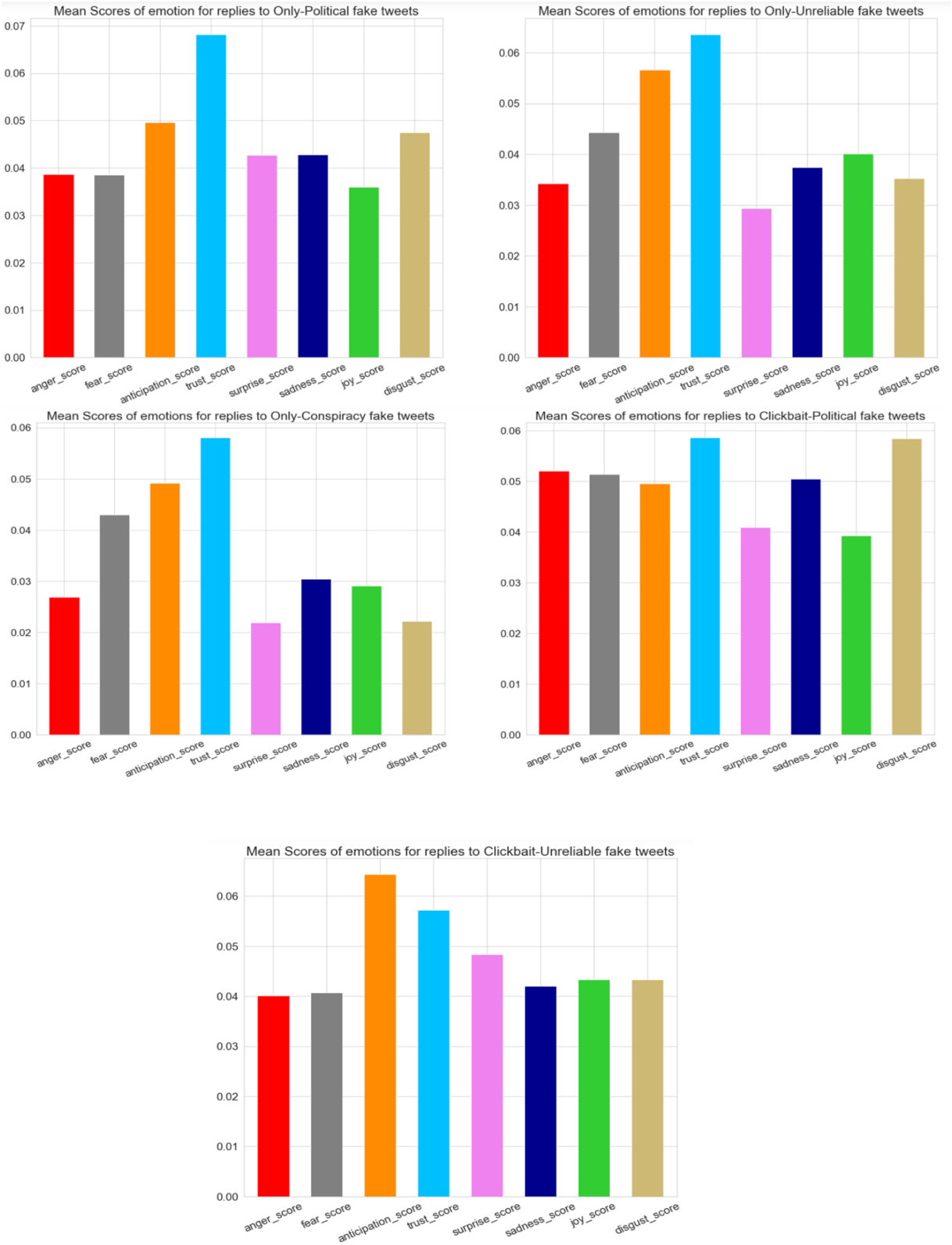

Figure 16: Emotion Analysis – Mean score of emotions for replies to different types of false tweets

Trust Dynamics in Clickbait Combinations: Interestingly, false tweets with "clickbait" in their labels induce different emotional responses than false tweets not labelled as clickbait. For example, when comparing "only-political" to "clickbait-political" false tweets, we observe that including a clickbait component influences trust dynamics, reducing *trust*. Similarly, users' responses to "clickbait-unreliable" false tweets exhibit lower trust scores compared to "only-unreliable" false tweets. One explanation could be that users clicking on articles with clickbait headlines may have certain expectations set by the headlines. However, in the case of clickbait, since the actual content does not align with these expectations or if the content is perceived as misleading, it can lead to a decrease in trust.

3.4.2.3 Stance Classification

The stance refers to the attitude or position a person holds towards the truthfulness of the target (e.g., a false tweet). Stance can be used to measure public opinions and help determine the veracity of information. In this research, we use the following three labels for stance classification (stance towards COVID-19 false information in our dataset):

- ***Support (agree)****:* when the author of the reply tweet supports the veracity of the target (the false tweet to which they are responding).
- ***Oppose (disagree)****:* when the author of the reply tweet rejects the veracity of the false tweet to which they respond.
- ***Neutral (No Stance):*** when the author of the reply tweet neither agrees nor disagrees with the target's false tweet to which they are responding (when the tweet is either neutral or irrelevant to the target).

Table 17 shows an example of stance classification for a sample conversation (a false tweet and its replies). Stance detection is difficult because of the complicated semantic meaning of the whole commenting sentence, which must be learned to understand its stance.

Table 17: Example Stance Classification for a sample Conversation. The first row shows a target false information (source tweet), and all subsequent rows are users' replies to the target false tweet.

Tweet	**Stance (label)**
Black Cats Are Reportedly Being Boiled Into Paste To Treat Coronavirus In Vietnam	*N/A (Source Tweet)*
Fake news	oppose
I'm Vietnamese and I promise we DONT EAT BLACK CAT. this is a fake news. We even oppose to eat dog or cat meat. And I really hate SO who eat cat or dog meat. So every country have many kind of person. Please understand that not all Vietnamese eat black cat. I'm so disappointed.	oppose
Fuck you. Show me your proof that these cats are from Vietnam. Using photos without credit is illegal and don't speculate if you're not scientists or local who witnessed the truth.	oppose
Where did you get this FAKE news from? I've never heard of sth like what you post, VNeses don't even know but how could you insistently affirm a false numerer on	oppose
You believe that fake new? Now I know why other dude burned 5G towers 🤣	oppose
Very very sick people!! Animal cruelty.	support
Oh, God. please...	support
Black cats are awesome!	neutral
THIS IS SICK BEYOND SICK.	support
Black cat lives matter!	neutral
Evil never ends.	support
Humans have not evolved. These people have no morality.	support
Stop eating pets, people.	support
What?!👀	neutral
Every cloud has a silver lining.	support
People are stupid. We deserve this pandemic.	support

The steps for stance classification are as follows:

1) ***Stance Annotation*:** refers to labelling the tweets based on their stance towards the target. Stance annotation can be **manual** (including crowdsourcing) or **automated**, but most prior works are manually annotated because it is more accurate and reliable. For this research, I manually annotated (labelled) **1100 reply tweets**. These tweets belong to a total of **20 conversation** threads. I plan to make the labelled data publicly available on my GitHub page.

2) ***Data Preprocessing***: Prepare the input for the classifiers. This includes data cleaning, such as removing punctuation and non-ASCII characters, lowercase, stop words, etc. (I kept hashtags and emojis because they contain valuable information about users' feelings.)

3) ***Annotation Assessment***: Various approaches are used to evaluate annotation quality, such as Percent agreement, Fleiss Kappa score, and Pairwise Cosine Similarity on the vector representation of the tweets. Alternatively, the annotations can be assessed by repeating rounds of labelling and revising with different annotators until an agreement is reached.

4) ***Classification Models (Classifiers)***: Various machine learning and deep learning models have been used in the literature for stance classification or detection. Example models include but not limited to SVM, Bag-of-Words, LSTM, and, more recently, transformer-based models such as BERT and SBERT are used. In this research, I used transformer-based deep learning models (Facebook Bart and BERT models) for Stance classification and fine-tuned it on my labelled data.

5) ***Model Evaluation***: There are several metrics to evaluate the performance of the models, such as F1 score, Macro-average F1 (for multi-class), Precision, and Recall.

Table 18 reviews the sample papers in the literature that used stance detection, including their method for stance annotation and assessment, the models used for stance classification, evaluation metrics, and the context of their dataset.

Dataset/Paper	Stance annotation	Annotation assessment	Models	Evaluation Metrics	Labels	Covid Context
RumorEval2019 (Gorrell et al., 2018)	Manual (paid)	Not provided	branchLSTM, NileTMRG	Macro-averaged **F1 Score**	Support, Deny, Comment, Query	No
COVID-CQ (Mutlu et al., 2020)	Manual	Pairwise Cosine Similarity	SVM, LR, MNB, SGD, GB, CNN		Favor, Against, Neutral	Yes (Chloroquine to treat covid-19)
COVIDLIES (Hossain et al., 2020)	Manual (experts)	percent agreement, Fleiss Kappa score	Linear, Bag-of-Words, Avg. GloVe, BiLSTM, SBERT	Precision, Recall, F1 Score	Agree, Disagree, No Stance	Yes (6761 tweets, 86 common misconceptions about covid-19)
COVMis (Hou et al., 2022)	Manual	Cohen's Kappa score	Linear, Bag-of-Words, Linear, Avg. GloVe, SBERT, SBERT (DA), BERTScore (DA) + SBERT (DA)	Precision, Recall, F1 Score, & Macro-average	Favor, Against, Neither	Yes (2631 tweets for 111 misinformation items about COVID-19)
COVID-19 Rumor Dataset (Cheng et al., 2021)	Manual	Repeat & revise until agreement	BERT, VAE		Support, Deny, Query, Comment	Yes (6834 rumors about covid-19)
STANCY (Popat et al., 2019)	Automated (Neural Network)	Not provided	LSTM, ESIM, MLP, $BERT_{BASE,}$,		Support, Oppose	No
SemEval2016 (Dias & Becker, 2016)	Automated (Rule-based)	Not provided	Weakly Supervised model		Against, Favor, None	No
KE-MLM (Kawintiranon & Singh, 2021)	Manual (MTurk)	Not provided	KE-MLM	Macro-Average F1 score + F1 score of each class	Support, Oppose, Neutral	No
This Research	Manual	Repeat & revise until agreement	DistilBERT, BERT, Bart	Macro-averaged F1 Score	Support, Oppose, Neutral	Yes

Table 18: Summary of relevant papers (publicly available datasets) on stance classification

3.4.2.3.1 Transformer-based Models for Stance Detection

Transformer-based methods are among these popular deep-learning methods (Yay et al., 2020). As is the case for many tasks related to NLP, a high percentage of recent work on stance detection employs transformer-based deep learning approaches, including but not limited to Bidirectional Encoder Representations of Transformers (BERT) (Devlin et al., 2018).

While BERT (Bidirectional Encoder Representations from Transformers) models have gained prominence in natural language processing tasks, I also used the Facebook BART model (BART: Denoising Sequence-to-Sequence Pre-training) for stance classification. With its sequence-to-sequence architecture, BART excels in capturing contextual relationships and generating coherent outputs. This adaptability makes it well-suited for stance detection, allowing for a more comprehensive analysis of the interplay between statements and their responses.

Below are the results of the BART model for stance classification trained and fine-tuned on our Covid-19 dataset.

	precision	**recall**	**f1-score**
support	0.89	0.81	0.85
oppose	0.13	0.50	0.21
neutral	0.75	0.21	0.33
accuracy			**0.74**
macro avg	0.59	0.51	0.46
weighted avg	0.84	0.74	**0.77**

- **Accuracy:** Accuracy measures the overall correctness of predictions across all classes. The overall accuracy is 0.77, indicating that the model correctly predicted the stance for 77% of the instances.
- **Precision:** Precision measures the accuracy of the positive predictions made by the model.
- **Recall:** Recall measures the ability of the model to capture all relevant instances of a class.
- **F1-Score:** The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance.
- **Macro Avg and Weighted Avg:** The macro avg is the unweighted average of precision, recall, and F1-score across all classes. The weighted avg considers the number of instances in each class, giving more weight to larger classes.

In summary, the model performs well in predicting the "support" class but poorly in predicting the "neutral" and "oppose" classes. This is because of the imbalance in the model's performance across different stances (there are many more stances of support compared to oppose or neutral). Therefore, we must further analyze and improve the model, especially for the underrepresented classes.

Handling Imbalance Dataset using Class Weights: Imbalanced datasets, where one class significantly outweighs the others, pose a challenge for model training. Our dataset is highly imbalanced towards the “support” label because most users’ responses agreed with their corresponding false tweet (i.e., most users supported the false tweet). In this

study, we confront the class imbalance by using class weights. In this method, we assign higher weights to the minority class (class with more samples) and lower weights to the majority class (class with fewer samples). We used the following formula to address the imbalanced classes in our data:

$$Wj = \frac{n\_samples}{(n\_classes \ * n\_samplesj)}$$

Where Wj is the weight for class j, n_samples is the total number of samples in our data, n_classes is the number of classes in our data (in our case, there are three classes: support, oppose, neutral), and n_samplesj is the number of samples in the respective class.

Assigning weights proportional to the inverse of the class frequencies helps to improve the bias introduced by the majority class. This adjustment ensures that the model considers each class with due importance, preventing it from favouring the overrepresented category.

3.4.3 Regression Analysis (RQ3)

This section provides the regression analyses conducted to explore the relationships between various factors and the spread characteristics of false tweets. Regression analysis is a powerful statistical method used to examine the influence of independent variables (e.g., tweets or users' characteristics) on a dependent variable (e.g., spread characteristics such as retweet count). The choice of variables for inclusion in the models was guided by the research questions to uncover insights into the dynamics of false tweet propagation.

We employed ordinary least squares (OLS) regression for each regression analysis, a widely used method that minimizes the sum of squared differences between observed and predicted values. OLS provides estimates for the coefficients of independent variables and assesses their significance.

The rationale for using regression was to identify the factors that influence the spread of false tweets and understand the role of user emotions and stances in spreading false information. In the remainder of this section, we provide the results of the regression analysis, accompanied by detailed interpretations of the results.

3.4.3.1 Multiple Regression with Emotion Scores

This subsection provides the results of multiple regression analysis incorporating emotion scores, such as anger and fear, as independent variables. The goal was to understand how the emotional tone of user responses influences the spread (retweet counts) of false tweets. Table 19 summarizes regression analysis to examine the impacts of two emotions (fear and anger) in users' responses on the spread of false tweets. As we can see, the R-squared value is very close to zero, indicating that the model does not explain much of the variability in retweet count. Also, the F-statistic is low, and the p-value is high (0.719), suggesting that the model is not statistically significant.

Overall, based on these results, the model, which includes 'fear' and 'anger' emotions, does not provide a significant explanation of the variance in spread (retweet counts). These emotional scores alone may not be strong predictors of the spread characteristics of

false tweets. Therefore, we will explore other factors and variables to improve the model's explanatory power in the following.

OLS Regression Results

Dep. Variable:	retweet_count	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.002
Method:	Least Squares	F-statistic:	0.3297
Date:	Sun, 14 Jan 2024	Prob (F-statistic):	0.719
Time:	22:43:42	Log-Likelihood:	-1305.4
No. Observations:	821	AIC:	2617.
Df Residuals:	818	BIC:	2631.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2080	0.045	4.615	0.000	0.120	0.296
anger_score	-0.0094	0.404	-0.023	0.982	-0.802	0.783
fear_score	-0.3302	0.463	-0.714	0.475	-1.238	0.578

Omnibus:	1683.661	Durbin-Watson:	1.927
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3600564.674
Skew:	15.673	Prob(JB):	0.00
Kurtosis:	325.911	Cond. No.	12.7

Table 19: Summary of Regression Analysis with emotions (fear and anger) as independent variables and retweet count as dependent variable

3.4.3.2 Linear Regression with Stance

The following table shows the result of linear regression with stance (users' stance towards false tweets) as the independent variable and retweet count as the dependent variable. The R-squared value is 0.006, suggesting that only a small portion of the variability in retweet counts is explained by the users' stance towards false tweets. Also, the p-value associated with the "stance" variable is 0.012. The low p-value indicates that the users' stance towards false tweets is statistically significant in predicting the retweet count. Additionally, the effect size (magnitude of the coefficient) is relatively small, indicating a modest impact.

The results indicate a statistically significant association between users' stance and retweet count. However, the small R-squared value suggests that the model explains only a small proportion of the variability in retweet counts. This could be due to the influence of other unmeasured factors, which we will explore in the following subsection.

OLS Regression Results

Dep. Variable:	retweet_count	R-squared:	0.006
Model:	OLS	Adj. R-squared:	0.005
Method:	Least Squares	F-statistic:	6.288
Date:	Mon, 15 Jan 2024	Prob (F-statistic):	0.0123
Time:	00:50:17	Log-Likelihood:	-1640.9
No. Observations:	1059	AIC:	3286.
Df Residuals:	1057	BIC:	3296.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0411	0.079	0.520	0.603	-0.114	0.196
stance	0.1206	0.048	2.508	0.012	0.026	0.215

Omnibus:	2061.809	Durbin-Watson:	1.847
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3880515.064
Skew:	14.309	Prob(JB):	0.00
Kurtosis:	298.169	Cond. No.	4.88

Table 20: Results of regression analysis with stance as independent variable and retweet count as dependent.

3.4.3.3 Multiple Regression with Additional Variables

A more comprehensive multiple regression analysis was conducted to better understand the factors that influence the spread of false tweets, including several variables like "like_count," "reply_count," etc. The results are provided in Table 21. This allowed us to evaluate the combined impact of various factors on the spread of false tweets. As we can see, the R-squared is 0.745, indicating that the model explains approximately 74.6% of the variance in the retweet count. Also, the F-statistic is 770.8, and the extremely low p-

value (4.18e-311) suggests that at least one of the predictors is significantly related to the retweet count.

OLS Regression Results

Dep. Variable:	retweet_count	R-squared:	0.745
Model:	OLS	Adj. R-squared:	0.744
Method:	Least Squares	F-statistic:	770.8
Date:	Mon, 15 Jan 2024	Prob (F-statistic):	4.18e-311
Time:	01:34:39	Log-Likelihood:	-920.00
No. Observations:	1059	AIC:	1850.
Df Residuals:	1054	BIC:	1875.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0291	0.041	-0.711	0.477	-0.109	0.051
stance	0.0213	0.025	0.868	0.386	-0.027	0.069
like_count	0.1566	0.005	30.958	0.000	0.147	0.167
author.followers_count	1.695e-05	9.15e-06	1.853	0.064	-9.97e-07	3.49e-05
author.following_count	9.414e-07	9.94e-06	0.095	0.925	-1.86e-05	2.04e-05

Omnibus:	720.721	Durbin-Watson:	1.826
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51697.413
Skew:	2.381	Prob(JB):	0.00
Kurtosis:	36.896	Cond. No.	1.46e+04

Table 21: Results of Multiple Regression Analysis

The model explains a substantial portion of the variability in retweet counts. The number of likes has a significant positive impact on retweet counts. For each additional like, retweet counts increase by 0.1566, and this effect is highly significant (p-value < 0.001).

3.5 Discussion (Study 1)

This study investigated the spread dynamics of diverse types of false information on social media platforms, focusing on X. Our analysis revealed differences in the propagation characteristics of various types of false tweets. For instance, we found that conspiracy false tweets exhibited longer cascade lifetimes compared to political and unreliable false tweets, indicative of their potential to linger and resonate within online

communities (Grinberg et al., 2019). Furthermore, while political false tweets showed rapid spread speeds, unreliable false tweets exhibited longer lifetimes and faster dissemination, underscoring the multifaceted nature of misinformation propagation.

Moving to users' responses (emotionally and attitudinally) towards different types of false tweets, our sentiment and emotion analysis revealed interesting insights into the varied reactions derived from different types of false tweets. Although the dominant sentiment in the content of false tweets is negative, users' responses often shifted towards a neutral tone, particularly across political and unreliable false tweets. The results of emotion analysis further supported distinct patterns, with trust and anticipation as dominant emotions across all types of false tweets. Interestingly, including clickbait further increases emotional responses, resulting in heightened levels of fear, anger, sadness, and disgust. This finding is aligned with the sensationalist nature of clickbait content (W. Chen et al., 2016).

Additionally, our regression analyses aimed to elucidate the factors influencing the spread of false tweets, revealing intriguing insights into the impact of user responses and engagement metrics. While emotional scores of emotions such as fear and anger alone did not emerge as strong predictors of spread characteristics, users' stance towards false tweets showed a statistically significant association with the spread of false tweets. Moreover, our analysis revealed the influence of users engagement metrics, such as the number of likes, on the spread of false tweets. This finding shows the importance of user interactions and their implications for information propagation. These findings

underscore the need for a multifaceted approach to address the spread of false information, including content and user characteristics.

By highlighting the unique propagation characteristics of diverse types of false tweets, study 1 offers valuable insights to improve detection models, mitigation strategies, and efforts to counteract the impact of fake news. **First**, fake news detection models can incorporate propagation features such as cascade size, lifetime, or spread speed to better distinguish between different types of false information. For example, models can use higher spread speed and larger cascade size to detect political false information and longer cascade lifetime for conspiracies. **Second**, understating the nuances in the spread of diverse types of false information can enhance mitigation strategies. For instance, considering the fast spread of political false tweets, rapid responses such as early detection can be used to limit their reach. On the other hand, given that conspiracies have longer lifetimes, they require more sustained efforts, such as continuous monitoring and spreading corrective information. **Finally,** the insights from this research can help fact-checkers to allocate their resources more effectively by prioritizing the rapidly spreading false information, especially in critical times (e.g., during elections or pandemics).

However, we acknowledge that this study has several limitations. First, different platforms may have different characteristics, such as user behaviours, content moderation policies, and network structures, which could influence the dynamics of false information propagation. We focused on a single social media platform (X), potentially limiting the generalizability of our findings. Second, the different sample sizes for diverse types of false tweets could introduce bias and may have influenced the robustness of our findings.

Third, our study primarily focused on quantitative analyses of spread patterns and user responses. Qualitative methods such as content analysis, interviews, or surveys could provide deeper insights into the contextual factors that could influence the spread of false information. Also, different types of false information differ along several dimensions; hence, it is difficult to theorize the direction of the total difference. Finally, our sentiment and emotion analysis offered valuable insights into user responses to false information. However, they may need to fully capture the complexity of user behaviour and motivations. Several other factors, such as users' decision-making processes, cognitive biases, and social influences, play critical roles in information sharing but still need to be directly addressed in our study.

One interesting direction for future research is to investigate the role of individual user characteristics and network structures in shaping the spread and consumption of false information. Another exciting area is to explore the psychological factors underlying user behaviour to provide a more comprehensive understanding of false information dissemination dynamics. Also, future research should investigate the spread dynamics of different types of false information across diverse platforms to provide a more comprehensive understanding of false information dynamics across diverse online ecosystems. Also, future studies should employ more extensive and balanced datasets to enhance the validity and generalizability of our findings.

Chapter 4

4 Combating Fake News on Social Media

4.1 Background and Motivation

There are several barriers to combat online fake news. First, fake news on social media spreads faster, farther, and deeper than true news (Vosoughi et al., 2018). That is, fake news can spread exponentially fast at early stages and pose harmful impacts in a very short time. Second, in many cases, it is difficult to identify whether the news is fake or not. Manual fact-checking and debunking fake news cannot keep up with the large volume and fast spread of fake news on social media. To address this, a large body of research focused on automated fake news detection. However, regardless of the type of algorithm for fake news detection (text-based, propagation-based, etc.), they are still not very effective.

Thus, it is important to devise strategies to stop fake news not only after its spread but also before its spread and even before its creation. Here, we aim to examine this broad landscape by focusing on all lifecycle stages of fake news dissemination. We specifically seek to provide a comprehensive picture of combating fake news on social media. This holistic view considers synergies among approaches and makes more careful and hopefully effective plans to tackle the problem. To this end, we adapt the Straub Model of Security Action Cycle to the context of combating fake news on social media. This model comprises four steps (countermeasures) to address security threats: deterrence,

prevention, detection, and mitigation/remedy. Notably, the Straub Model of Security is rooted in criminology and is hence not limited to the security context. It can be applied to any undesirable behaviour. Since creating or spreading fake news on social media is an undesirable behaviour with destructive impacts on individuals and societies, we propose similar steps to combat fake news on social media. Based on a thorough investigation of the relevant literature, we use this model to classify the vast literature on fake news. We believe that this framework helps readers grasp the whole picture of the research frontier.

We note that in recent years, there have been several attempts to review the literature on fake news from different perspectives. Table 28 in the appendix summarizes the various review papers on fake news, their combat stage, classification criteria, and the type of false information addressed in their review. Based on this review, we conclude that most existing reviews focus on fake news detection (Bondielli & Marcelloni, 2019; Sharma et al., 2019; K. Shu et al., 2017; K. Shu, Bernard, et al., 2019; Zhang & Ghorbani, 2020; Zhou & Zafarani, 2020; Zubiaga et al., 2018). Thus, there is a need to consider also other approaches, such as deterrence, and to conduct a systematic and multidisciplinary review on the full lifecycle of combating fake news on social media.

4.2 A Framework to Combat Fake News on Social Media

In this section, we describe a framework to combat fake news on social media. The adopted framework (shown in Figure 17) is inspired by the Straub Model of the Security Action Cycle (Straub & Welke, 1998). According to the model, the first step to address the system risks is to use “deterrents” such as administrative policies or employee

training. ***Deterrents*** are passive countermeasures to discourage individuals from engaging in illicit behaviour or committing a crime. Deterrence is applicable in the stage where the adversaries have intentions but have not yet taken any action to launch security attacks. If deterrents fail, the next step is to use "***preventives***". These are active countermeasures to impede or stop individuals from engaging in criminal activities or illegal behaviour. This means that prevention may happen when an abuser has taken an action, but the system will stop them. If an abuser overcomes the first two stages and engages in undesirable behaviour, then detection approaches should be used. ***Detection*** refers to the process of monitoring and identifying undesirable behaviour. Finally, an effective IS system should be able to ***mitigate*** or ***remedy*** the destructive impacts of undesirable behaviour. Remedy refers to the post-attack process or activities that reduce the negative impacts of undesirable behaviour.

In this paper, we apply the Straub model of the Security Action Cycle to the context of fake news on social media and propose similar steps to combat fake news on social media platforms. The rationale behind this is twofold. ***First***, similar to Information security threats that harm individuals, organizations, and society, creating or spreading fake news is also an undesirable phenomenon which can negatively affect many different entities such as individuals, organizations, political parties, and financial markets. Research shows the destructive and far-reaching impacts of fake news on many aspects of our lives, including but not limited to politics (Allcott & Gentzkow, 2017), businesses (Bakir & McStay, 2018; Petratos, 2021), healthcare (Carrieri et al., 2019), or people's responses to natural disasters such as Hurricane Sandy (Gupta, Lamba, Kumaraguru, et al., 2013).

Thus, both security and fake news represent undesirable behaviours that can be deterred, prevented, detected, and remedied. ***Second***, fake news can sometimes (and certainly not always) represent a security threat, which makes the application of models from the security domain to fake news (Botha & Pieterse, 2020). In some cases, the alluring nature of clickbait can be used to spread malicious software (E. Zeng et al., 2020).

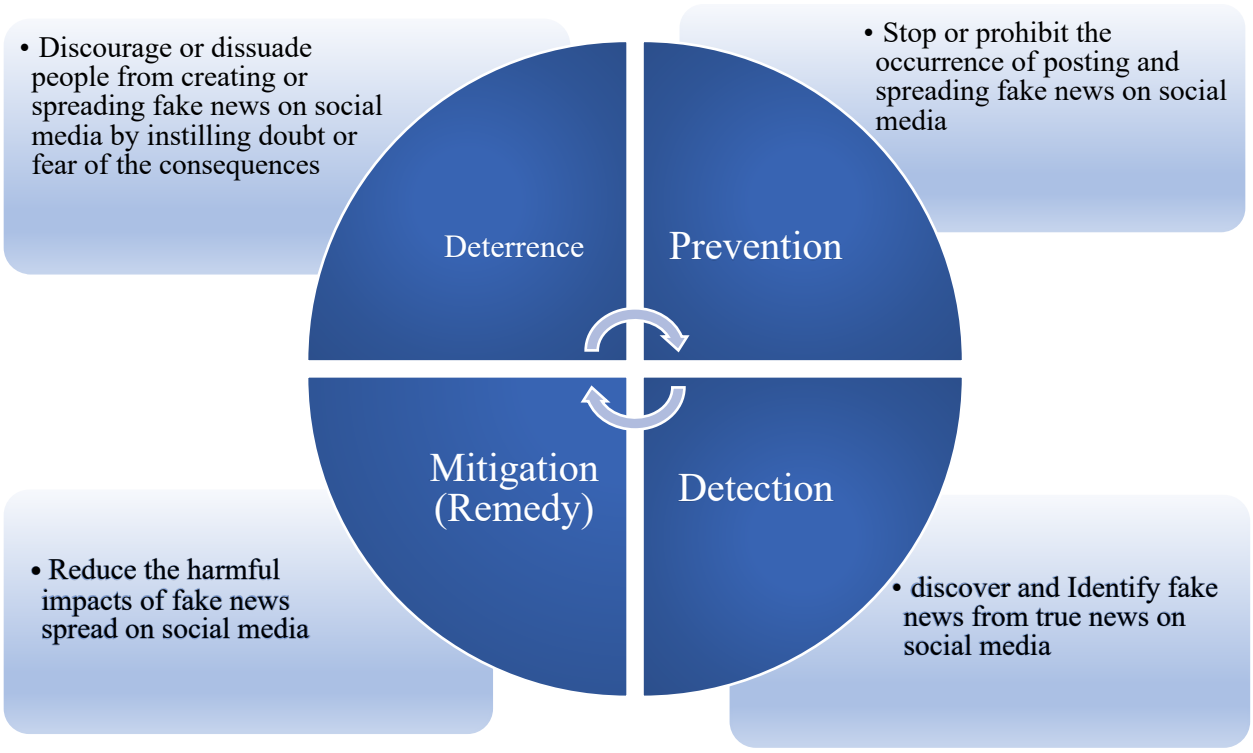

Figure 17: Framework to Combat Fake News on Social Media (Stages and Definitions)

Importantly, fake news and security issues do not always have the same attributes. We outline the similarities and differences between fake news and information security threats in Table 22 and Table 23.

Table 22: Example Application of the Framework in Security and Fake News Contexts

Combat Stage	Description	Examples in Security	Examples in Fake News
Deterrence	• The first step to cope with system risks (in this research, to combat fake news) is to use deterrents. Deterrents are passive countermeasures to discourage people from engaging in illicit behavior or committing a crime. Deterrents are passive in that in that they have no inherent provision for enforcement and depend on the willingness of users (Straub & Welke, 1998).	• Policies and guidelines for proper system use • Educate users (e.g., Security awareness programs) about the risks and threats in organizational environment and to emphasize the certainty and severity of sanctions for violation	• Establish laws, policies, and regulations by government, authorities, and social media platforms. • Educating users and increase their awareness about fake news and its destructive impacts. • Information literacy, media literacy, and other training programs
Prevention	• Preventives are "active countermeasures with inherent capabilities to enforce policy and ward off illegitimate use" (Gopal & Sanders, 1997; Straub & Welke, 1998).	• Locks on computers • Password access control	• Block or suspend malicious accounts. • Block or remove known fake content
Detection	• If deterrents and preventives don't work and the abuser penetrate the system (in our case, when fake news is already published and disseminated), the next step is to identify and detect misuse (in our case detecting fake news)	• System Audits to monitor computer use activities. • Transaction log reports • Virus scanning	• Fact-checking (Manual, Crowd-sourced, Automated) • Algorithmic Solutions (Machine learning, and other approaches)
Mitigation (Remedy)	• The last stage is to mitigate or reduce the harmful effects of abuse (in our case, reducing the negative impacts of fake news)	• Software recovery • Prosecution of perpetrators • Legal actions such as criminal and civil suits	• Minimize the spread of fake news by blocking certain nodes in the network (e.g., influential nodes) • Spreading true information • Platform interventions (account-level, and content-level) to stop or limit the spread of fake news

Table 23: Comparison of Fake News and Security Contexts

	Fake News	Security Attacks
Creators (Who)	• Bots • Malicious/fake accounts • Politicians, or governments, etc. • Normal people	• Hackers • Corporate spies • Terrorist groups • People with security knowledge (in contrast to fake news that can be propagated by any individual, security attacks can be done only by people who have relevant knowledge)
Motives (Why)	• Monetary motives (e.g., increase revenue or web traffic in case of clickbait), • Ideological motives, • Political motives (e.g., during elections)	• Financial/Monetary motives • Access data • Political motives (Hacktivism)
Intention	Anyone with or without malicious intent may spread fake news (e.g., many individuals may share fake news and misinformation without knowing it is false)	Often with malicious intent (however, sometimes security threats can occur because of carelessness, or compromised credentials
Where	Social media, messaging apps, peer-to-peer	Organizations, firms
Targets (Who) & Impacts	• **Individuals** (increase panic, distrust, conflict, radicalization/extremism), • **Societies** (echo-chambers, polarization, voting patterns), • **Organizations** (impact on the relationship between companies and consumers, destroy brand reputation).	• Often on **organizations** (e.g., economic loss, loss of customer and stakeholder trust, destroy brand reputation) • **Societies** (e.g., shortage of products or services, panic buying, etc.) **Individuals** (e.g., because of weak passwords, or storing their personal information on devices while using unsecure public networks).

Why people fall for it	• Ideological beliefs, Confirmation Bias, Naïve realism (people tend to believe they have the "true" perception of reality and those who disagree with them must be uninformed, irrational, or biased), • Social Normative Theory (the influence of other people that leads us to conform in order to be liked and accepted by them), • Intuitive or emotional response and lack of analytical thinking (Dual Process Theory), • Familiarity with the topic, • Social validation • Echo-chambers & personalized contents (people are often exposed to contents that agree with their beliefs)	• Lack of enough security measures (e.g., weakness in security policies) • System weaknesses (e.g., weakness in computer technologies such as network protocols (TCP/IP) or operating systems' weaknesses) • Individuals' sloppiness or negligence • Lack of knowledge
Example Impacts	To **manipulate public opinion**, **reducing trust** in governments, institutions, or experts. For example, in the context of Covid-19, fake news reduced trust in medical experts and doctors. Another example is Macedonian teenagers who were targeting Trump supporters in the 2016 US presidential election, although their motivation was financial (for advertising revenue). In some cases, such as the "Pizzagate" incident (Fisher et al., 2016), fake news resulted in **physical violence**.	Security attacks often impact organizations. For example, Microsoft Exchange Servers data breach in 2021 was one of the biggest cyberattacks of US history, which affected more than 30,000 US companies. Security attacks can also impact individuals and societies. For example, in case of the Colonial Pipeline ransomware attack in May 2021, millions of people experienced fuel shortages, and many airlines had to cancel or change flights due to jet fuel shortage.

These two tables demonstrate nuanced differences between security issues and fake news, but also point to key similarities, namely in the undesirability of the behavior, the problems it causes, and the potency of deterrence, prevention, detection, and remedy to reduce the behavior or its adverse outcomes. Given such similarities, and the possibility to apply the stages in Table 22 to fake news, we view the application of the Staub model to fighting fake news as reasonable.

4.3 Review Process Methodology

To find the relevant literature, we used two major online scientific databases, namely Google Scholar and Scopus. Google Scholar was linked to major online libraries and databases such as Web of Science, EBSCOhost, ProQuest, ACM Digital Library, and IEEE Xplore. We set six criteria to include or exclude articles in the literature review: 1) we selected articles written in English, 2) we included journal publications, and conference papers, as well as the grey literature to expand scholarly efforts and gain more practical insights about the fake news phenomenon (Adams et al., 2017), 3) since fake news research is a multidisciplinary topic, we included studies from various disciplines such as Information Systems (IS), Computer Science (CS), Information Security, Psychology, Social Science, etc. 4) we selected articles that focus on combating fake news on social media, conceptual papers about fake news, relevant literature review papers, and a few studies from the security literature (our theoretical foundation is based on a model from the security literature), 5) We also excluded studies about fake news

propagation, echo-chambers, filter bubbles, and polarization, 6) Finally, we did not limit our search to any specific time range.

To obtain more effective search results, we used the following keywords in our search query: ("fake news" OR "misinformation" OR "disinformation" OR “Rumor" OR “false information" AND ("combat" OR "fight") OR "deter fake news" OR “prevent fake news” OR "detect fake news" OR “mitigate fake news”. We searched document titles, abstracts, and keywords. This search strategy and selection criteria identified 1640 articles in Google Scholar and 925 articles in Scopus. After eliminating the overlapping materials and reading and skimming the abstracts, 245 papers were selected for further screening and reading the full text. Screening the full text also led to the elimination of 81 more papers. The final number of papers included in this review was 164 articles. We note that our literature search was by no means exhaustive, rather we tried to provide a representative summary of the relevant research to combat fake news on social media. Figure 18 shows the flow diagram for our literature review process.

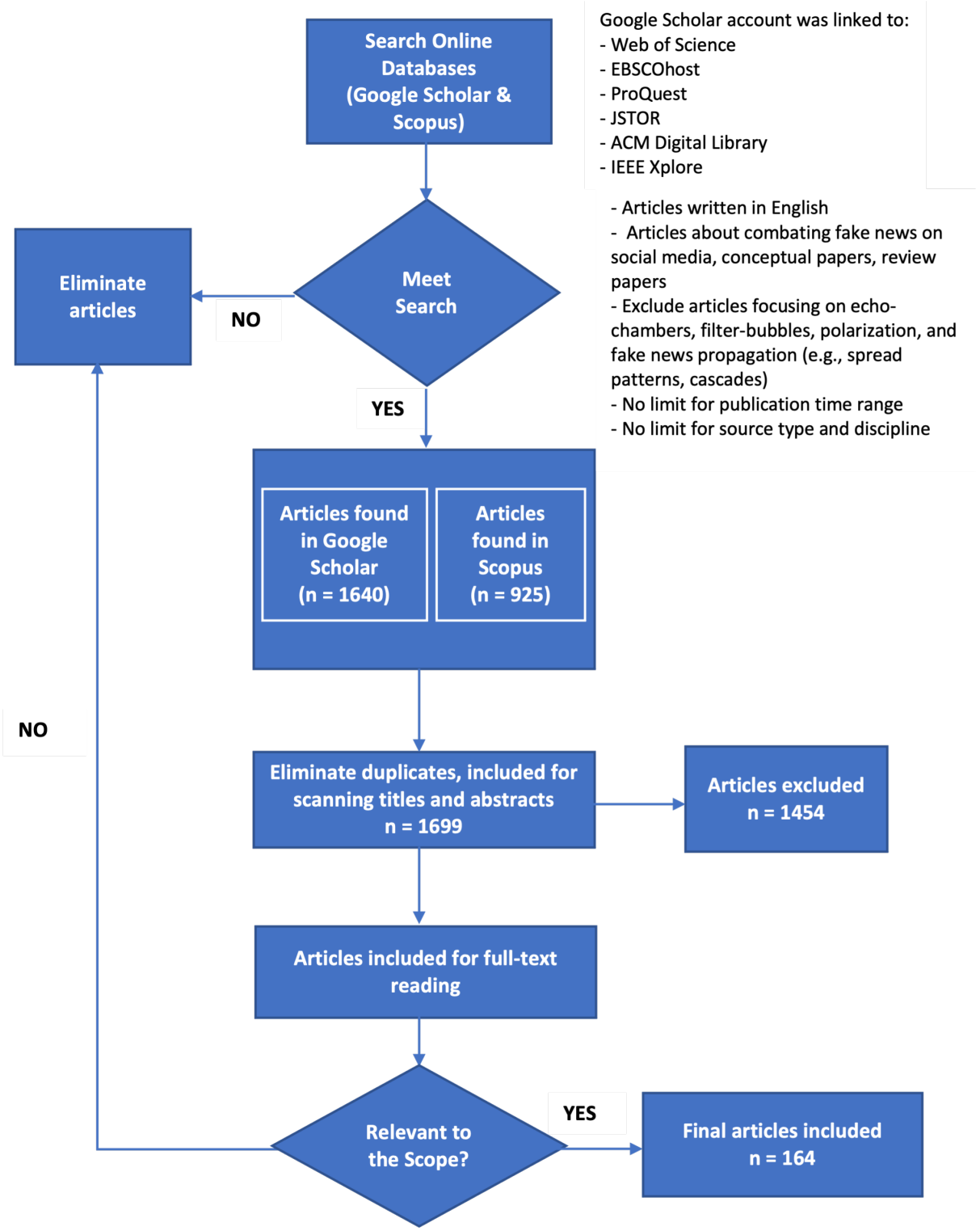

Figure 18: Flow Diagram for the Literature Review Process

4.3.1 Descriptive Statistics of The Articles

Figure 19 shows descriptive statistics about the articles reviewed in this study. First, Figure 19 (a) shows the year-wise distribution of articles reviewed in this research. This figure shows an increasing trend in the number of publications about fake news, which shows a growing interest in this topic, especially after the year 2016. This is largely due to the proliferation of fake news during the 2016 US. presidential election (Allcott & Gentzkow, 2017). Second, the figure on the top right (Figure 19 (c)) shows the number of review articles by the publisher, where the "ACM Digital Library", "Taylor & Francis", "Elsevier", "IEEE", and "Springer" are among the top 5 publishers. Third, Figure 19 (b) on the bottom left shows the distribution of reviewed papers by discipline. This figure shows that the reviewed articles about fake news come from a range of disciplines. The majority of the contribution comes from the Computer Science (36%) field, followed by the Information Systems (16%) field. Finally, we can see that most of the work on combating fake news on social media is focused on "detection", while "deterrent" strategies have gained less attention from academic scholars Figure 19 (d). We note that for this figure, we only included articles focusing on combating fake news and excluded other papers such as review papers, and theoretical papers from Information Security literature. Also, if a paper focuses on more than one stage, for example, all four stages, it is presented in all the stages of the pie chart. The reason for doing this is that if we considered a separate part in the pie chart for all combinations (e.g., deter & prevent, deter & detect, …), each part would have been very small (there are 14 possible combinations). Also, our goal is to show several studies (portion of the research) for each

combat stage. For example, if a paper addressed both detection and mitigation, it should appear in both "detection" and "mitigation" slices of the pie chart. However, based on our review, only a few papers focused on more than one stage.

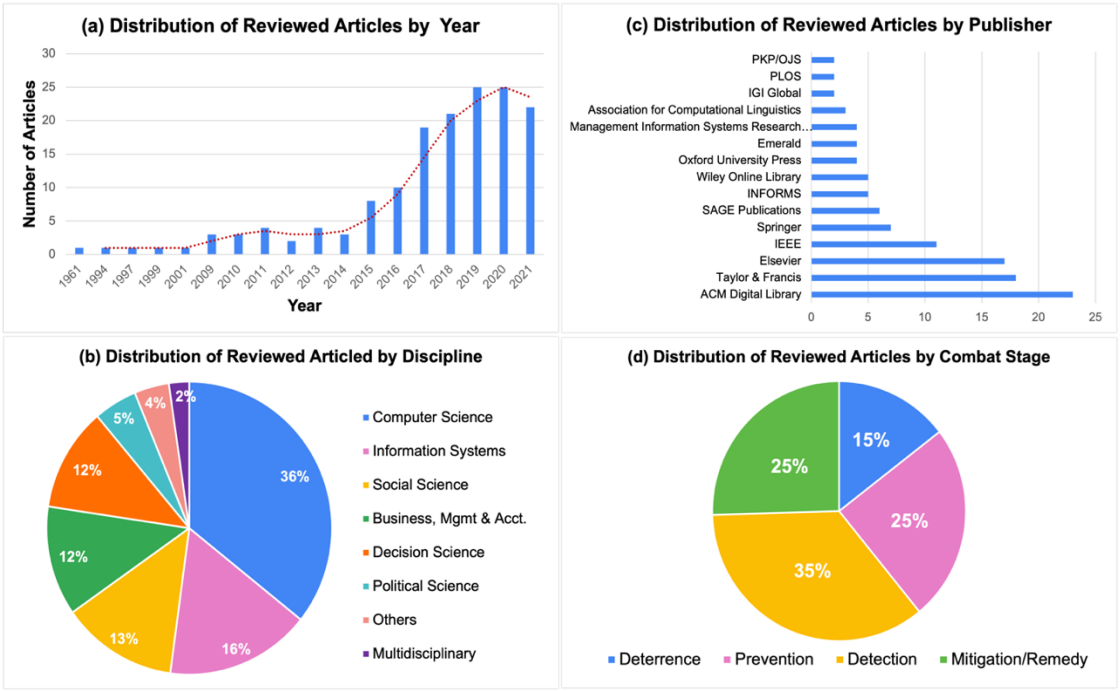

Figure 19: Descriptive Statistics of Reviewed Papers

As mentioned earlier, fake news is a multidisciplinary field and the articles reviewed in this research come from a variety of disciplines. However, the contribution of different fields varies across different stages of combating fake news. As depicted in Figure 20 (a), most of the reviewed articles related to fake news "deterrence" come from the field of "Social Science" (27%). In terms of fake news "prevention" (Figure 20 (b)), almost half of the articles belong to the "Social Science" and "Psychology" fields (29% and 19% respectively). While the Social science discipline has the highest contribution in fake news "deterrence" and "prevention" research, there is very little research (only 3%) in

fake news detection. Figure 20 (c), shows that the "Computer Science" discipline plays the dominant role in fake news detection studies (47%). Interestingly, more than 70% of reviewed papers on fake news detection are from "Computer Science" and "Information Systems". This is probably due to the technical nature of fake news detection on social media platforms. Finally, as shown in Figure 20 (d), the research on fake news mitigation is mainly covered in "Computer Science" (42%). In the following section, we will further explain each stage of combating fake news on social media.

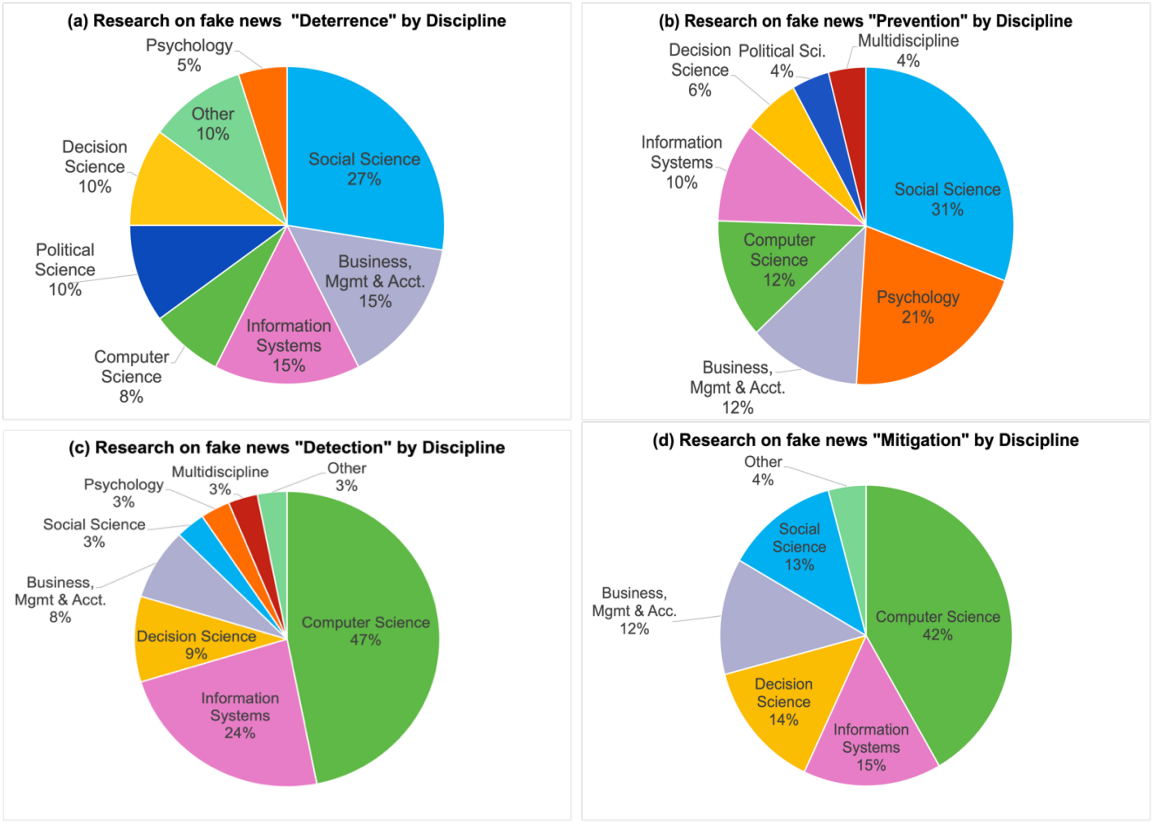

Figure 20: Distribution of the Reviewed Articles on Fake News Combat Stages across Disciplines

4.4 Stages to Combat Fake News on Social Media

In this section, we further discuss each stage of the fake news combat cycle in our framework namely, deterrence, prevention, detection, and remedy (mitigation). For each stage of the fake news combat cycle, we provide our definition for that stage, the challenges that exist to implement that stage, the existing approaches, limitations of current approaches, and directions for future research.

Table 24 provides a summary of the challenges, approaches, limitations, and future directions for each stage of combating fake news on social media. Also, Table 28 in the appendix provides a complete list of reviewed articles classified by fake news combat stage (please note that it only contains papers relevant to combating fake news, and excludes review papers, conceptual papers, etc.).

Table 24: Fake News Combat Stages, Challenges, Approaches, Limitations, and Future Opportunities

	Challenges	Approaches	Limitations & Future Research
Deterrence	•Several motivations for fake news creation and propagation •Difficult to discourage people from creating or posting fake news especially when it is politically or ideologically motivated. •Social media companies lack incentives to police their platforms.	•Establish laws, policies, and regulations on fake news by governments, authorities, and social media platforms. •Educate users to increase the awareness of regulations	•Fake news has not been legally treated as a crime and no agreement on which criteria to consider a fake news as a crime. •Regulation may be viewed as restriction of freedom of speech. •Why laws and regulations are less effective to deter fake news.

Prevention	•Difficult to apply preventives due to the nature of free information exchange. •Fake news prevention could be interpreted as censorship to against freedom of speech. •Governments and authorities may misuse fake news prevention against opposition for political purpose.	•Block and suspend malicious accounts on social media platforms. •Block or filter the known fake news on social media platforms. •Prebunking (inoculation against fake news by e.g., preemptive warnings)	•How to effectively prevent wide and fast spreading of fake news in social media? •How to distinguish and balance the fake news prevention and freedom of speech? •How to prevent true information to be mistakenly blocked •How to combat people's ideology biases in relation to fake news?
Detection	•Fake news is masqueraded as true news and humans are often unable to identify fake news. •People like to receive and share the news they like without considering if they are true or fake. •Social media facilitate the spread of massive news, and it is difficult to check every news piece.	•Manual detection (either by experts or through crowdsourcing). •Automated detection (computational fact-checking, algorithmic solutions using ML, propagation pattern, etc.) •Guidelines for fake news detection	•Manual detection is difficult and time consuming. •There are needs to further improve the effectiveness and applicability of algorithmic solutions (semi-supervised and unsupervised models, fake audio and video detection, the use of social contexts features) •Educate people to detect fake news
Remedy (Mitigation)	•Fake news causes significant damage to the individuals trust believe and the justice of democratic society. •It is difficult to make people disbelieve fake news and change behavior accordingly. •Continued Influence Effect (CEI), i.e., when discredited information (e.g., flagged fake news) continues to affect behavior and beliefs.	•Minimize the influence of fake news propagation. •Spreading truth through both social media and public media to discredit fake news. •Platform interventions to clean up fake news. •Execute legal sanctions against those who caused significant damage by creating and spreading fake news.	•Anti-fake news actions can backfire and increase the spread of fake news. •Platform interventions have also some limitations, e.g., people may perceive unflagged content as true •What is appropriate rule of multiple stakeholders such as governments, political parties, social media providers, organizations, and individuals to maintain healthy social media environment.

4.4.1 Deterrence

The first stage to combat fake news is deterrence defined as discouraging or dissuading people from creating or spreading fake news on social media by instilling doubt or fear of the consequences. Importantly, deterrents dissuade people from action through the threat of force and not the actual use of force. Since deterrence is about demotivating people, we first need to understand the motives behind creating fake news.

4.4.1.1 Deterrence Challenges

One of the main challenges of this stage is that there are different motivations to create or spread fake news on social media: 1) Political motives to influence public opinion, to advance a preferred candidate and political party, or to damage opponents, especially during election periods (Allcott & Gentzkow, 2017). 2) Economic/Financial motives to generate revenue and monetary profit. A common example is using clickbait headlines which entice/attract users to click through and subsequently generate revenues through increasing page traffic. 3) Ideological motives to promote ideological views. For example, the ISIS terrorist group uses social media platforms to promote their opinions through spreading propaganda (Zannettou et al., 2019), 4) Other Individual motives: These include malicious intents (to hurt others in various ways), influence (to get power or to manipulate public opinion), sow discord (confusion), and fun (Zannettou et al., 2019).

However, there are insufficient discouragement mechanisms to demotivate people from creating and spreading fake news on social media. This is in part because there is no clear governing body. Social media platforms as the main actors in this space have little

incentive to deter the production of fake news. At the same time, governments struggle to restrict freedom of speech and create perceptions of effective deterrence. The challenge lies in deterrent dependency on users' free will, and it is difficult to restrict or control it without effective "carrots and sticks". In the following subsections, we discuss the approaches to demotivate or deter users from creating and spreading fake news on social media.

4.4.1.2 Deterrence Approaches (Deterrents)

A common deterrent approach to fight against fake news is to establish laws, regulations and policies that clearly define sanctions and consequences for those who create and/or spread fake news on social media. According to the General Deterrence Theory, perceived certainty and severity of sanctions deter individuals from engaging in illegal behaviour or committing a crime (in criminology) or IS misuse intention (in IS security). The idea behind this is that people will avoid abusive behaviour (e.g., creating or spreading fake news) if they believe that the cost of their actions is higher than the benefits. Therefore, establishing laws, regulations, and policies is an important deterrent to dissuade people from creating or spreading fake news on social media. Although such attempts conflict with free speech ideas and ideals, some level of restriction on free speech is inevitable to discourage the creation and spread of fake news, rather than just preventing its spread (Helm & Nasu, 2021).

In recent years, there have been some attempts by governments, policymakers, legislators, and social media platforms to address the fake news problem. For example, Malaysia's government was one of the first to establish a law to combat fake news by

penalizing offenders with a 10-year jail sentence, a fine up to (£90,000) or both[9]. In 2018, the German parliament established a law, known as NetzDG, which oblige large social media companies to remove fake news and hate speech content within a 24-hour deadline or pay a penalty of up to 50 million euros[10]. In Italy, the anti-trust chief Giovanni Pitruzzella has called for the EU to establish rules to consider the penalty for companies that spread false content (Morgan, 2018). Following claims of Russia's meddling in the 2017 French presidential election, President Emmanuel Macron promised anti-fake news laws in 2018 to stop fake news (Nugent, 2018). A comprehensive list of anti-misinformation actions around the world is provided in (Funke & Flamini, 2022).

Another (non-legislative) deterrence approach is to use educational and training programs. Such programs dissuade users from illicit behaviours (create or spread false content in the context of fake news) by increasing awareness about regulations and policies, and the penalties associated with violating the laws. In security literature, it has been shown that the best way to ensure the viability of a security policy is to educate users about it to make sure they understand it and accept the necessary precautions (Whitman et al., 2001). IS research found that user's awareness of security policies and SETA (Security Education, Training, and Awareness) program deter IS misuse (D'Arcy et al., 2009). A similar study found that employees can better manage cybersecurity tasks when they are aware of their company's information security policy (Li et al., 2019). In the context of online fake news, governments in several countries took some steps to increase users' awareness about fake news through training and media literacy initiatives.

[9] https://www.theguardian.com/world/2018/mar/26/malaysia-accused-of-muzzling-critics-with-jail-term-for-fake-news
[10] https://www.reuters.com/article/us-singapore-politics-fakenews-factbox-idUSKCN1RE0XN

For example, in 2019, the federal government of Canada announced it was giving $7 million to projects aimed at increasing public awareness of online fake news[11]. In the same year, the Netherlands government launched a public awareness campaign to inform their citizens about the spread of fake news online.

4.4.1.3 Deterrence Limitations and Future Opportunities

There are several limitations in effectively implementing deterrence strategies, especially in the context of fake news. First, establishing laws and regulations to deter users from creating or spreading fake news in the context of fake news is more difficult and complex compared to security or criminology contexts. One limitation is that it is not easy to recognize fake news as a crime because there is not even an overall agreement on how to define fake news, or when to consider it as a crime. For example, in the context of politics, a content that the left party consider as true news may be considered as fake by the right party.

In general, there are not enough deterrent mechanisms against fake news on social media. There should be more effective laws, regulations, and policies by governments, authorities, and social media platforms to discourage users from creating/spreading fake news. Sanctions and penalties against fake news should be certain and severe to be effective as deterrents. However, the laws and regulations established by governments can be viewed against freedom of speech, especially by people who don't trust their governments and those who think these laws increase corruption and prevent their right to free speech. Research shows that regulations are not the preferred choice of the public to

[11] https://www.canada.ca/en/canadian-heritage/services/online-disinformation.html

combat fake news on social media as people may view regulations as a restriction to freedom of speech. Most people, even when they perceive fake news harmful to society, if they have a choice, they prefer non-regulatory solutions such as education over regulations (Jang & Kim, 2018). The authors explain that most people prefer education over regulations because they "do not want to sacrifice their freedom of speech to protect other's vulnerability".

Ultimately, more research is needed to understand *why* anti-fake news laws and regulations are less effective, *how* differences between laws affect the motivation and ability to generate fake news, and how, why, and when people respond differently to deterrence measures against fake news generation and spread. This line of work should also examine interactions of legislation and other means. As pointed out by Haciyakupoglu et al., (2018), legislation should be complemented by other means such as pre-emptive inoculation, immediate measures (e.g., fact-checking), and long-term measures (e.g., education and media literacy). We discuss all these measures and more in the remainder of this paper.

4.4.2 Prevention

If people choose to ignore the deterrents, the next stage is to use preventive actions, defined as *"active countermeasures with inherent capabilities to enforce policy and ward off illegitimate use"* (Gopal & Sanders, 1997; Straub & Welke, 1998). Applied to fake news, preventive actions are active countermeasures to prevent individuals from creating or spreading fake news on social media. In the context of fake news, blocking fake accounts or blocking fake content are examples of preventive countermeasures (users may create a fake account, but it will be blocked or removed). We further explain the prevention stage in the remainder of this section.

4.4.2.1 Prevention Challenges

Implementing preventive measures in the context of fake news is more challenging compared to the security context. One challenge is the debate over censorship and freedom of speech, which can be a potential explanation for the weakness of social media platforms in implementing effective preventive countermeasures. For example, preventive measures such as blocking or suspending social media accounts can be misinterpreted as censorship or as conflicting with freedom of speech ideals. The laws against fake news established by governments can especially be questioned by people who do not trust their governments and those who think these laws increase corruption and prevent their right to free speech. In fact, in some cases, governments and authorities may use preventive measures to censor the opposing views and further spread the information aligned with their views and benefits. In addition, prevention mechanisms vary based on the countries in which they are implemented. For example, some countries have taken stronger preventive measures and have more control over the information their people consume online. However, as mentioned earlier, there is a concern that the governments use it to further spread fake news. In the remainder of this section, we further explain and review the current preventive approaches to combat fake news on social media. Based on our review, we also discuss the research gaps and future opportunities for this stage of the fake news combat cycle.

4.4.2.2 Prevention Approaches (Preventives)

In recent years, there have been growing concerns about the role of social media in facilitating the spread of fake news and several studies called for actions by social media

platforms to fight against fake news (Flew et al., 2019; Hartley & Vu, 2020; Hemphill, 2019; Smyth, 2019). In response, social media platforms have taken some steps to prevent the spread of fake news by e.g., blocking fake and malicious accounts and updating their algorithms to remove incentives for users who promote false information. In terms of preventive measures, Facebook updated its recidivism policy to stop people who repeatedly violate its Community Standards from being able to create new pages or groups[12]. Following the 2020 presidential election campaign in the United States, Facebook banned deepfake media (manipulated videos or photos)[13] from its platform. X has accelerated its combat against fake accounts by suspending millions of fake and suspicious accounts in 2018 (over 70 million only in May and June). X's growing campaign against bots and trolls was driven by political pressure from the U.S. Congress following reports of manipulation by Russian disinformation during the 2016 presidential election (Timberg & Dwoskin, 2018). In addition, X announced a COVID-19 misinformation policy in response to a large volume of false and misleading information related to COVID-19. Depending on the severity of the violation, the consequences of violating this policy may include tweet deletion, labelling the tweet, and even account locks and permanent suspension of the accounts for severe or repeated violations of this policy[14].

In academia, several studies focused on platform interventions to fight fake news on social media. A type of platform intervention that restricts the accounts from publishing

[12] https://about.fb.com/news/2020/09/keeping-facebook-groups-safe/
[13] https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/
[14] https://help.X.com/en/rules-and-policies/medical-misinformation-policy

fake news is *"account-level intervention"*. Several attempts have been proposed in this direction such as algorithms to identify bots and malicious accounts (Sharma et al., 2019), and network monitoring which leverages a set of nodes to filter the information they receive and block what they identify as fake news (Amoruso et al., 2020; Kimura et al., 2009; Zhang et al., 2016). More recently, Ng et al., (2021) examined "fake news flags" as content-level and "forwarding restriction" as an account-level intervention to combat fake news. They found that the two types of interventions have different effects on fake news: flagging fake news leads to the more centralized and less dispersed spread of fake news while forwarding restriction leads to less direct and more indirect forwarding of fake news, compared with true news.

Another preventive approach is Prebunking Fake News by Inoculation. According to the Inoculation Theory (Papageorgis & McGuire, 1961), people can be inoculated against persuasion by being exposed to a refuted version of a counterargument beforehand. Just like vaccines, a sufficiently weakened dose of counterargument triggers the production of "mental antibodies", immunizing people to unwanted persuasion (Compton, 2013). Inoculation involves two elements: (a) forewarning – a warning of a forthcoming threat, designed to motivate resistance and defend one's attitudes, and (b) a pre-emptive refutation (or prebunking) of the persuasive arguments. Several studies have shown inoculation as an effective strategy to confer resistance against fake news on social media. For example, inoculation, based on logical communication and facts, reduces the influence of conspiracy persuasion by increasing the degree of skepticism towards conspiratorial claims (Banas & Miller, 2013). In the context of climate change,

inoculation has been shown to neutralize the influence of misinformation on a perceived consensus about climate change (Cook et al., 2017). Similarly, preemptive warnings help protect (inoculate) public attitudes about the scientific consensus against misinformation (Van der Linden et al., 2017). In the context of COVID-19, the theory of inoculation is shown to be an effective strategy to confer resistance against fake news (van Der Linden et al., 2020). Research shows that inoculation or prebunking fake news is more effective than debunking it, and preexposure warnings have a stronger effect than corrections (King et al., 2021). In other words, prevention is better than cure. For example, Jolley & Douglas, (2017) found that anti-conspiracy arguments that were present prior to conspiracy theories improved vaccination intention, but they were not effective if they came afterwards (once established, conspiracy theories become resistant to correction).

Similar to inoculation, a line of education research called "misconception-based learning" (McCuin et al., 2014) suggests that teaching approaches directly addressing and refuting misconceptions as well as explaining the facts, stimulate higher engagement with the content, which results in more effective and longer-lasting learning (Kowalski & Taylor, 2009). Misconception-based learning has shown to be one of the most effective means of reducing misconceptions (Ecker et al., 2017; Kowalski & Taylor, 2009), which has been successfully applied in various settings. For example, The Massive Open Online Course (MOOC) on climate science denial, which has reached more than thousands of students, used misconception-based learning to refute 50 of the most common myths about climate change (Schuenemann & Cook, 2015).

4.4.2.3 Prevention Limitations and Future Opportunities

There have been insufficient mechanisms and strategies to prevent the creation or spread of fake news on social media. Unlike the security context where using passwords or locks on computers can be used as a preventive measure, implementing preventives in the context of fake news is not that easy. As mentioned earlier, one issue is that preventives such as blocking social media accounts can be interpreted as censorship and against the freedom of speech. However, the harmful impacts of fake news may outweigh the benefits of free speech (Helm & Nasu, 2021). Therefore, one important direction for future research is to investigate the balance between freedom of speech and preventive measures against the creation or spread of fake news on social media. Another concern with preventive measures such as blocking accounts on social media is that it might unintentionally prevent the spread of truth if it mistakenly blocks legitimate accounts. Moreover, only a limited number of malicious accounts can be blocked compared to the large volume of fake news on social media. In addition, there have been a few studies on inoculation and education to prepare users to fight against fake news, mostly in the context of climate change (Lewandowsky & Van Der Linden, 2021; Van der Linden et al., 2017). Finally, prior studies mainly focused on passive inoculation where people are inoculated against the same information to which they will be exposed later. However, recent research shows that "active inoculation" where people are exposed to similar, but not the same information is more effective in creating resistance against fake news (Roozenbeek & Van Der Linden, 2019).

Ultimately, more research is needed on preventive measures: *when* they work, *why* they work, *for what types* of fake news or in what contexts they work, and for what *types of people* they work best. Findings from such studies can help social media providers apply effective restrictions. One more question that is relevant in this context is: how to prevent true information from being mistakenly blocked?

4.4.3 Detection

If fake news cannot be stopped at the first two stages, which means fake news is already spread on social media, the next stage is to detect fake news. We define detection as discovering and identifying fake news from massive news posted and shared on social media.

4.4.3.1 Detection Challenges

Detecting fake news on social media is a challenging task. First, fake news is always decorated as true news which makes its detection difficult. As pointed in (J. George et al., 2021):

> *FN is created with truth-subversive language, designed to play on emotion and connect with recipients by signaling authenticity and homophilic characteristics on the part of the originator. The objective of such strategies is to seed FN content effectively, and to increase the propagation of FN messages through social networks (p. 6)*

At the same time, people's ability to identify fake news is only slightly better than chance (Kumar et al., 2016; Ott et al., 2011; Rubin, 2010). More importantly, the term fake news has been highly polarized and misused, especially by politicians who label any piece of content that is not aligned with their view as "fake news" (Vosoughi et al., 2018).

Fake news detection is especially challenging in the context of social media where everyone can post any content, real or fake, with no cost or friction, resulting in a massive amount of news posted every day. It is difficult to monitor and detect all the fake news posted on social media. In general, people like to receive and share the news they like and believe what they like without considering if the content is true or fake (Moravec et al., 2019). In addition, social media platforms facilitate the spread of fake news through personalized recommendations which leads to the formation of "echo-chambers". Echo-chamber (Sunstein, 1999), refers to an effect when users in social media form groups with like-minded individuals where they are largely exposed to the information that confirms their own opinions (Shore et al., 2018). Echo chambers facilitate the spread of fake news, which can be explained through two psychological factors: social credibility (people tend to perceive a source as credible if others perceive it is credible) and frequency heuristic (when processing information, people favour information they have seen more frequently, even if it is fake) (Shu et al., 2017). In the remainder of this section, we review the existing fake news detection approaches and discuss the limitations and future research opportunities.

4.4.3.2 Detection Approaches

Fact-checking: One of the main approaches to detecting fake news on social media is through fact-checking. Fact-checking is the process of evaluating the authenticity of news by comparing the knowledge extracted from to-be-checked content with facts. There are three types of fact-checking. First, *"Expert-based fact-checking"* uses credible fact-checkers to manually assess the accuracy of the news. In recent years, several fact-

checking organizations such as PolitiFact[15], and Snopes[16] have emerged to verify the veracity of information. For example, PolitiFact's Truth-O-Meter provides six ratings, including true, mostly true, half true, mostly false, false, and pants on fire (i.e., the statement is not accurate and makes a ridiculous claim) to reflect the accuracy of a claim. The website also provides a "scorecard" to show the accuracy of statements based on the mentioned ratings. The Snopes website also has a similar rating scale with a few more labels such as unproven, miscaptioned, scam, etc. Second, *"Crowdsource-based Fact-checking"* uses a group of regular individuals to evaluate the accuracy of information. For example, Fiskkit[17] is a crowd-based fact-checking website where users can apply tags to judge the article's accuracy and view how others evaluated the article. The International Fact-Checking Network (IFCN)[18] has launched a huge crowdsourcing project (The "CoronaVirusFacts" Alliance database) that unites more than 100 fact-checking organizations worldwide to fight the COVID-19 infodemic. Recently, X introduced "Birdwatch", a crowdsourced fact-checking pilot that allows people to flag Tweets they perceive as misleading and write notes to provide additional context for why it may be misleading. Finally, *"Computational (Automated) Fact-checking"* uses computational solutions such as ML and NLP to automatically fact-check fake news. Two well-known examples are Truthy (Ratkiewicz et al., 2011) which track political memes in X and help detect misinformation, and Hoaxy (Shao et al., 2016), a platform for automatic tracking of fake news diffusion and its competition with fact-checking

[15] https://www.politifact.com/
[16] https://www.snopes.com
[17] http://fiskkit.com
[18] https://www.poynter.org/ifcn/

efforts on X. Some other examples include Factmata, an AI project by Google (Dale, 2017), ClaimBuster (Hassan et al., 2017), and ClaimRank (Gencheva et al., 2017) that use machine learning approaches for fact-checking. Kim et al., (2018) used both the crowd and expert knowledge to detect and prevent the spread of fake news. They developed CURB, a scalable online algorithm to decide which stories to send for fact-checking and when to do so. Table 25 shows a comparison of fact-checking approaches.

Table 25: Comparison of Fact-checking Approaches

Fact-checking	Advantage(s)	Drawback(s)
Expert-based (Manual)	• High accuracy (because it use experts)) • Expert-based fact-checking websites can be used as a public data repository for fake news research, e.g., LIAR (Wang, 2017) and FakeNewsNet (Shu et al., 2020)	• Slow • Costly • Low scalability (they cannot keep up with the large volume and rapid spread of fake news on social media)
Crowdsource-based (Manual)	• Faster than expert-based fact-checking • More scalable than expert-based	• Low accuracy (because it relies on regular people for verification) • Vulnerable to manipulation and misuse by adversaries • Less scalable than computational (automated) fact-checking
Computational (Automated)	• Faster than both expert-based and crowdsourced-based fact-checking • High scalability	• Less accurate than expert-based fact-checking

Automated Algorithmic Solutions: In recent years, there has been several survey papers reviewed the literature on fake news detection on social media and classified the approaches to detect fake news from different perspectives such as fake news component (content, user, context), methodology, etc. *From a data mining perspective,* fake news

detection methods are classified into *knowledge-based* and *style-based* methods (based on content features) and *stance-based* and *propagation-based* approaches (based on social context features) (Shu et al., 2017; Zhou & Zafarani, 2020). *From a methodology perspective*, there have been several categorizations. For example, fake news detection approaches can be broadly divided into *classification* (ML and DL), and *other approaches* (propagation pattern, retweet behaviour, etc.) (Bondielli & Marcelloni, 2019). Other categorizations based on methodology are *machine learning, systems* (systems that inform users about detected fake news), and *other models/algorithms* such as epidemiological models, Hawkes processes, etc. (Zannettou et al., 2019). Fake news detection approaches have also been categorized *based on fake news components* (content, user, context). For example, fake news detection methods can be divided into three types: *Content-based* (identify fake news based on the content of the information), *Feedback-based* (based on user responses on social media), and *Intervention-based* (actively identify and contain the spread of fake news and mitigate their impacts) (Sharma et al., 2019). Finally, a comprehensive review of fake news detection approaches is provided in Zhang & Ghorbani, (2020), where authors provided three different perspectives to classify fake news detection approaches: *Component-based* (creator/user, content, social context), *Data mining-based* (supervised, unsupervised), *and Implementation-based* (online, offline).

We provide a summary of review papers on fake news detection, their classification criteria, and the type(s) of fake news they addressed in their study (see Table 26).

Table 26: Review Papers on Fake News Detection, Classification Criteria, and Type of False

Fake News Review Papers	Classification Criteria for Fake News Detection Approaches	Type of False Information
(Shu et al., 2017)	• Content Models: knowledge-based, style-based • Context Models: stance-based, propagation-based	Fake News
(S. Kumar & Shah, 2018)	Based on algorithms: • Feature-based • Graph-based • Model-based (Temporal, Propagation models)	Fake News, Fake Reviews, Hoaxes
(Zubiaga et al., 2018)	No specific classification for rumour detection	Rumours
*(Shu, Bernard, et al., 2019)**	Based on Network: • Interaction network embedding • Temporal diffusion • Friendship network embedding • Knowledge network matching	Fake News
*(Zannettou et al., 2019)**	• Machine learning • Systems • Other Models/Algorithms	Rumours, Hoaxes, Conspiracy Theories, Satire, Clickbait, Fabricated
*(Sharma et al., 2019)**	• Content-based • Feedback-based (based on user responses) • Intervention-based (detection and mitigation)	Fake News, Rumour
(Bondielli & Marcelloni, 2019)	• Classification approaches (ML, DL) • Other approaches (Crowdsourcing, Diffusion patterns, etc.)	Fake News, Rumour
(X. Zhang & Ghorbani, 2020)	• Component-based (Creator analysis, Content analysis, Context analysis) • Data mining-based (Supervised learning, Unsupervised learning) • Implementation-based (Online/Real-time, Offline detection)	Fake News, Fake Review, Rumour or Satire
(Zhou & Zafarani, 2020)	• Knowledge-based (Manual fact-checking, Automated fact-checking) • Style-based (based on content) • Propagation-based (using News Cascades, Propagation Graphs) • Credibility-based (source credibility)	Fake News
(Collins et al., 2021)	Classified fake news detection into 8 categories: Experts/Fact-check approach, Crowdsourced, Hybrid (Expert-crowdsource, Human-Machine), ML, DL, NLP, Graph-based methods, Recommender Systems	Fake News (Clickbait, Propaganda, Satire & Parody, Hoax, other)
(Khan et al., 2021)	• Knowledge-based • Feature-based • Network Propagation • Hybrid Approach	Fake News (including Rumor & Clickbait detection)

Please note that, in this research, we have not provided a new classification of fake news detection approaches because this was previously done. However, we provide example references for different fake news detection approaches, classified based on the methodology in Table 27.

Guidelines for News Detection: It is also important to help users improve their ability to detect fake news. In recent years, numerous workshops, training programs, and courses have been developed to help people recognize fake news from true news. A common approach is to provide guidelines for people to detect fake news. These guides often suggest a checklist for evaluating a news source. The CAARP (currency, authority, accuracy, relevance, and purpose) test, SMART (source, motive, authority, review, two-source test), or SMELL (source, motive, evidence, logic and left-out) are just a few examples (Lim, 2020). Other examples include but are not limited to a research guide on "Fake News, Misinformation, and Propaganda" by Harvard University library, two research guides offered by the University of Toronto library, and the "LibGuide", a popular library guide offered by librarians at Indiana University to help students in evaluating the credibility of information (Banks, 2017).

4.4.3.3 Detection Limitations and Future Opportunities

A large body of research has focused on fake news detection approaches, especially through algorithmic solutions. However, there are still many limitations. First, there is a lack of large-scale publicly available datasets on fake news that can be used as a benchmark to compare different algorithms. Such datasets help build and evaluate models

in a situation similar to the real world. In recent years, some public datasets have been developed (Shu et al., 2020; Wang, 2017). Second, most existing detection algorithms use supervised learning based on labelled datasets for training and validation. In real-world scenarios, most data are either unlabeled or only a few labels are available, in which cases unsupervised or semi-supervised models should be applied. Also, unsupervised models can better handle large amounts of data in real-time, which is especially useful in the context of social media where a large volume of information is created and disseminated every day. Third, prior research in fake news detection has mainly focused on the content. However, the context can help to identify if the content is true or false. For instance, the person described in the news could not be in the place at the time mentioned. Although there have been some recent works using contextual features (Atanasova et al., 2019; Nguyen et al., 2020; Shu et al., 2019), social context features need to be further investigated for fake news detection. Finally, information on social media platforms comes in various formats such as text, audio, video, etc. Usually, pictures or video recordings can be used as evidence of truth. However, with the advances in information technology, especially artificial intelligence in recent years, it is easy to use photo editing or deepfake technology to make fake images or videos that appear authentic but are practically indistinguishable by humans (Westerlund, 2019). It is important to develop methods that can detect not only fake text but also fake audio or video (Yu et al., 2021).

Overall, most detection efforts, especially the algorithmic solutions are in computer science. Although fake news research has gained more attention among IS scholars in

recent years, it mainly focuses on user behaviour and the psychological and cognitive factors in sharing fake news (Kim & Dennis, 2019; Moravec et al., 2019, 2022; Turel & Osatuyi, 2021). Fake news is a multidisciplinary research field in nature, and we believe that there is an opportunity for IS scholars to further contribute to solving this problem. For instance, questions around *why* and *when* people believe algorithmic screening should be examined. There is also an opportunity to examine human-bot interactions in the process of screening fake news, and whether such approaches are superior to using just bots or just humans.

4.4.4 Remedy (Mitigation)

The remedy (mitigation) stage aims at reducing the destructive impacts of fake news diffusion on social media. In this research, we use the words remedy and mitigation interchangeably.

4.4.4.1 Remedy/Mitigation Challenges

Fake news causes significant damage to the trust and beliefs of individuals (Ognyanova et al., 2020). It has also had significant negative impacts on global issues faced by human society such as fighting COVID-19 pandemics (Shirish et al., 2021), or the recent war between Russia and Ukraine (e.g., deepfake videos of Putin or Zelenskyy circulating on social media amid the conflict[19]).

To reduce the negative impact of fake news, it is important to know why people believe fake news even when they are told it is fake. People's ideology and pre-existing beliefs play an important role in believability and the spread of fake news. In fact, people believe

[19] https://www.dw.com/en/fact-check-the-deepfakes-in-the-disinformation-war-between-russia-and-ukraine/a-61166433

what they want to believe, even when it makes no sense at all (Moravec et al., 2019). From a theoretical perspective, several theories explain this. First, the theory of *Confirmation Bias* (Nickerson, 1998) posits that people tend to believe what confirms their pre-existing beliefs. Second, according to the theory of *Naïve Realism* (Ross & Ward, 1996) people tend to believe they have the "true" perception of reality and those who disagree with them must be uninformed, irrational, or biased. Finally, people are also influenced by their peers, and they tend to share information that is more aligned with their peers' beliefs to gain social acceptance and affirmation, regardless of the veracity of that information (*Social Normative Theory*) (Deutsch & Gerard, 1955). In political contexts, partisanship and the political ideology of individuals are common explanations for why people believe fake news, i.e., people perceive fake news as accurate if it is consistent with their political ideology (Turel & Osatuyi, 2021).

4.4.4.2 Remedy/Mitigation Approaches

A common mitigation strategy is to minimize the influence of fake news by limiting the scope of its spread, e.g., by blocking certain nodes or links in the network. The goal is to minimize the impact of fake news spread on social media. The impact of fake news on social media can be assessed by the number of people who are affected by fake news. Blocking the flow of information from influential users in the network can significantly reduce the impact of fake news spread as these users have many followers. Indeed, finding a minimum subset of individuals who are neighbours with the rumour community can help in limiting the spread of the rumour to the rest of the network (Fan et al., 2013). Fake news can be contained by training a set of individuals in a network to help them

distinguish fake from true news and stop the spread of fake news (Kotnis & Kuri, 2014). Tong et al., (2017) addressed the rumour-blocking problem in online social networks by using a random-based approach. They evaluated their randomized algorithm on both real and synthetic social networks (Power2500, Wiki, Epinion, and YouTube) and showed that their algorithm outperforms the state-of-the-art rumour blocking algorithms such as greedy algorithm with the Monte Carlo simulation in terms of running time. Another example is the DRIMUX model (dynamic rumour influence minimization with user experience), which minimizes the influence of rumours by blocking a subset of nodes while considering users' experience (a time threshold that a particular node is willing to wait while being blocked) (Wang et al., 2017).

Another approach to mitigate the impacts of fake news is through increasing the spread of true information (Shu, Bernard, et al., 2019). To this end, most prior research used competing cascades which contain true information, to compete with the fake news cascade as the falsehood begins to spread through the network rather than after its diffusion. The goal is to make sure that true news reaches users who are exposed to fake news, to reduce the chance of believing fake news, and to make social media a more reliable source of information. Several models have been proposed in this direction. For example, Budak et al., (2011) models the spread of two cascades evolving simultaneously: "bad campaign" spreading bad information (fake news) and "good campaign" to counteract the effects of fake news. They identified a subset of individuals (k influential users) to spread true information to minimize the number of users who at the end of the propagation process adopt the bad campaign. One limitation of the

approach used in Budak et al., 2011 is that their model assumes if a user is exposed to a piece of news, then they will also share the news. In a similar notion, Nguyen et al., (2012) proposed a model which finds a small set of influential nodes (users) to spread "good information" to contain misinformation. Their findings depict that when the number of required nodes to spread true information is small, it is most effective to select influential nodes in large communities. However, when more nodes are required, selecting influential nodes from smaller communities is more effective in limiting the fake news spread. Wang et al., (2014) developed two strategies to select the smallest set of influential nodes decontaminated with true information to effectively contain the spread of fake news. Their experimental results using three datasets from X, Friendster, and a random synthetic network proved the performance benefits of their proposed strategies.

In IS, there has been a growing interest in platform interventions to fight fake news on social media, either through content-level interventions (interventions that only target a piece of content) or account-level interventions (interventions that target the accounts that post fake news) (Ng et al., 2021). We discussed the account-level interventions in the prevention section. Content-level interventions reduce the impact of fake news by triggering users' cognition, e.g., through flagging fake news or highlighting the source of the article. A common example is using "fake news flags". In IS, scholars mainly studied the effectiveness of flagging on changing users' beliefs and limiting the spread of fake news. In this vein, two different approaches to implement a fake news flag were examined; one designed to trigger system 1 ("automatic cognition" or "fast-thinking")

and the other to trigger system 2 ("deliberate cognition", or "slow-thinking") (Moravec et al., 2020). Both approaches are shown effective in reducing the believability of fake news and combining both approaches was about twice as effective. To understand whether some types of flagging is more effective than others, three flagging strategies were examined: fact-checker flags, peer-generated flags, and publishers' self-identified humour flags (Garrett & Poulsen, 2019). They found that publishers' self-identified flags were the most effective strategy in reducing people's beliefs and sharing intentions of fake news. In addition to fake news flags, "highlighting the source of the article" and "source rating" are other forms of content-level interventions proposed in the literature (Kim & Dennis, 2019). The authors showed that both changing the interface to highlight the source of the article, and source rating (showing low ratings for the source) can nudge users to be more skeptical of fake news and less likely to believe and spread any article. Finally, different rating mechanisms (*experts' ratings, users' article ratings*, and *users' source ratings*) influence user beliefs in news articles (Kim et al., 2019). It was found found that users perceive expert ratings as more cognitive and user ratings as more emotional.

4.4.4.3 Remedy/Mitigation Limitations and Future Opportunities

Unfortunately, corrective information does not necessarily change people's beliefs and can have the opposite effect (Flynn et al., 2017). In politics, not only correction may fail to reduce misperceptions, but it can also backfire and strengthen misperceptions among ideological subgroups holding those misperceptions (Nyhan & Reifler, 2010). In line with this, King et al., (2021) used X data to examine the dynamic interaction between

true and fake news and found that information correction does not reduce the spread of fake news. Instead, it backfires and increases the propagation of fake news on social media. These findings are in line with prior research that shows that any attempt to debunk fake news by confronting falsehood and truths facilitates the acceptance of fake news (Pennycook et al., 2018). This is because frequent exposure (in this case repeating fake news) increases familiarity, which in turn increases the chance of accepting fake news. More research is needed to clarify these contradictory findings. Timing of information correction is also important and different methods may be useful in different phases of fake news propagation. For example, He et al., (2015) proposed an optimization approach that combines two methods (*blocking rumours* at influential users and *spreading the truth* to clarify rumours). They showed that the method of *"spreading truth"* should play a dominant role in the start of rumour containment, whereas the method of "rumour *blocking"* should be used extensively when approaching the end of the rumour restraining phase. This is because the exposure to fake news increases as time passes. The more fake news is circulated and repeated, it increases users' familiarity and acceptance. As a result, the "spreading truth" method may be less effective after longer exposure to fake news.

In terms of content-level interventions, most prior research studied their effectiveness in terms of psychological and cognitive aspects such as believability. Believability is an important factor in studying fake news on social media and prior research found the strong effect of believability on users' actions such as read, like, share, and comment (Kim et al., 2019a). However, there are contradictory findings about the effectiveness of

content-level interventions (e.g., flagging fake news) in reducing users' belief in fake news, and there are several factors (e.g., prior beliefs or source reputation) that can weaken the effectiveness of such interventions. For example, although using fake news flags triggers more cognitive activity, it is shown that it cannot overcome the role of confirmation bias and users continue to believe what they want to believe, regardless of the truth of a news article (Moravec et al., 2019). Also, a trusted source with a high reputation can lower the impact of flags on reducing the believability of fake news (Figl et al., 2019). Finally, using fake news flags may cause an implied truth effect, meaning that it may lead people to believe that unflagged content is trustworthy (Pennycook, Bear, et al., 2020).

Even though prior research on the effectiveness of content-level platform interventions is inconclusive, such interventions are still helpful in combating fake news because they trigger users' cognition and nudge them to think more deeply before sharing content on social media (Moravec et al., 2022). However, there are many more opportunities for IS scholar to understand when, how and why people resist the temptation to spread fake news and have a stronger motivation to check news items before they share them.

4.5 Study 2 Discussion

Several strategies are proposed to combat fake news on social media, mostly focused on detection approaches. However, fake news detection—although necessary—is not enough to stop fake news on social media. First, manual detection of fake news is time consuming and labor intensive. Automated fake news detection addresses this issue but is

less accurate. Also, automated fake news detection often suffers from limited explainability. Second, fake news detection happens only after it is disseminated and consumed by people. Since fake news can have severe harmful impacts in a matter of seconds, it is necessary to devise strategies to stop fake news from happening in the first place. In this paper, we proposed a framework to address the problem of fake news not only after its propagation, but even before it is created. The framework, which is inspired by the Straub Model of Security Action Cycle includes four stages: deterrence, prevention, detection, and remedy/mitigation. A summary of the approaches to combat fake news on social media and example references for each stage is provided in **Error! Reference source not found.**.

We next pointed to similarities between fake news and security threats (section 4.2), but also acknowledged key differences between information security threats and fake news on social media. This makes the implementation of some of the countermeasures more challenging in the context of fake news. As mentioned earlier, one difference between fake news and information security is that in case of fake news, people want to believe false news that fit their ideology, while from a behavioral standpoint, information security threats are primarily due to people's sloppiness in detecting threats. Thus, one interesting direction for future research is to investigate the ways we can **combat people's ideology biases** in relation to fake news. Some of the countermeasure approaches in our framework can be helpful in reducing belief in false information. For example, **accuracy-promoting interventions** such as warnings or nudging users to think about information veracity before sharing it can impact judgements about fake news

credibility (Bryanov & Vziatysheva, 2021). One approach is **inoculation interventions** which aims at pre-emptively warning users to the threat of fake news and equipping them with the tools to combat it. For example, **media and information literacy** approaches to educate users about deception strategies (Cook et al., 2017) or guidelines to help people detect fake news can be helpful. For example, recent research finds that exposing users to simple guidelines to detect misinformation (e.g., "Be skeptical of headlines," "Watch for unusual formatting") improves fake news discernment rate among both nationally representative samples in the U.S. (by 26.5%) and in India (by 17.5%), regardless of whether the headlines are politically concordant or not (Guess, Lerner, et al., 2020). Another approach is **using labels or flags** to trigger critical thinking. To understand whether some type of flagging is more effective than others, three flagging strategies were examined: fact-checker flags, peer-generated flags, and publishers' self-identified humor flags (Garrett & Poulsen, 2019). They found that publishers' self-identified flags were the most effective strategy in reducing people's beliefs and sharing intentions of fake news.

Another challenge in combating fake news on social media is that there are several motivations for fake news creation and spread, while there are not enough demotivation strategies. Deterrents such as "establishing laws and regulations" can be used to demotivate people from creating or spreading fake news on social media. However, there are **several limitations in effectively implementing deterrents and preventive measures to combat fake news on social media**. ***First***, fake news has not been legally treated as a crime and there is no agreement on which criteria to consider when

recognizing fake news as a crime. Also, the term "fake news" has been highly politicized and for example, what is considered as fake news by Republicans may be considered true by Democrats and vice versa. Therefore, as long as fake news is not recognized as a serious threat and there is no overall agreement on what content to consider as fake news, it will be difficult to devise effective regulations and penalties against it. ***Second***, there are different types of fake news with different characteristics. Thus, a single strategy cannot be enough to address the variety of behaviors in the fake news context. Fake news can be created and propagated with intention to deceive (disinformation) or without malicious intention (misinformation). There should be a distinction between users who purposefully create and share fake news and those who erroneously share false content with good intentions. For example, deterrent strategies can be helpful to deter malicious users who share disinformation but may be less effective against those who may not know that the content they are sharing is false. However, laws and regulations against fake news can still be effective to some extent (even for users with no bad intentions) because they make users think more carefully before sharing any content on social media. Also, legislations and penalties can further focus on the fake news with more harmful impacts. For example, Burkina Faso's parliament adopted a law to punish the publication of "fake news information compromising security operations, false information about rights abuses or destruction of property, or images and audio from a "terrorist" attack."[20] ***Third***, the legal punishment of users who unintentionally share fake news is a violation of free speech. Therefore, in case of sharing false content without intention

[20] https://www.poynter.org/ifcn/anti-misinformation-actions/

(misinformation), other strategies such as users' inoculation or education may be more effective. Educating users to increase their awareness about fake news characteristics, its' destructive impacts, and ways to spot and counter it on social media has proved to be very helpful in combating fake news. ***Finally***, there is a major concern about the compatibility of fake news prevention and the right to free speech. Fake news deterrents and preventive measures can be interpreted as censorship or violations of the right for free speech. On the other hand, authorities may misuse the preventives to filter the opposing views or even filter the truth. There are many interesting research questions to explore here, such as: how to balance the prevention of fake news and freedom of speech? How to combat fake news while protecting free speech? How to prevent true information from being mistakenly blocked?

An additional challenge in combating fake news on social media is how to effectively reduce the harmful consequences of fake news. To address this, we described several remedies such as providing true information. However, it is difficult to persuade people to disbelieve fake news. Due to the Continued Influence Effect (CIE) of fake news (Johnson & Seifert, 1994), information correction often fails to disbelieve fake news and fake news continues to influence people's thinking even after correction. One explanation is that information correction often requires repeating fake news. The repetition of fake news increases familiarity, which in turn increases believability in fake news (Lewandowsky et al., 2012). An important question in this stage is how to help people disbelieve fake news once they consume it.

Ultimately, our review has led us to believe that fake news is a multidisciplinary problem that requires various expertise and should be addressed through collective efforts from different fields. The IS discipline can contribute significantly to the research on fake news. *IS scholars can draw on theories and empirical findings on the design, use, and impacts of IT artifacts at different levels of analysis (Gimpel et al., 2021; Kim et al., 2019; Kim & Dennis, 2019; Moravec et al., 2019, 2022)*. The focus on human-technology interactions, design elements and managerial practices that can influence it is a key feature of IS research and is also a cornerstone feature of research on fighting fake news. Thus, the IS scholars can contribute to all stages of the process.

In addition, *IS researchers can learn from the findings in related areas such as fake reviews* (Cheng Nie et al., 2022; Xiao & Benbasat, 2011), *social behaviours in online social networks* (Kuem et al., 2017) and *security* (Bulgurcu et al., 2010; D'Arcy et al., 2009). For example, since fake news shares some similar characteristics with the context of security, IS research can benefit from applying various approaches used in security to the context of fake news. Moreover, the research in psychology and social science can shed light on psychological and behavioural factors contributing to the creation and spread of fake news. They can help in better understanding why people believe fake news, understanding different types of fake news, and how to break echo-chambers and filter bubbles among like-minded users on social media. Other examples for future research include, but are not limited to: 1) How to combat people's ideology biases concerning fake news? 2) Why do some people continue to believe in fake news, even after it is flagged as false? (continued influence effect), 3) Why do sometimes some anti-

fake news actions such as information correction backfire and increase the spread of fake news? 4) How to balance the policies and regulations against fake news with the need for freedom of expression? 5) How do the various Information Communication Technologies (ICT) impact fake news detection? Such questions and beyond can be addressed by the IS research community.

4.6 Study 2 Limitations

While we review extant works on fake news, classify them by process stages, propose a framework to understand the vast literature on the topic, propose new research directions, and pave the way for future research, our study has limitations that should be acknowledged.

First, we analyze each stage of combating fake news separately. Considering that this part is already lengthy and complex, we include this point as a promising direction for future research. In essence, we present an important starting point for examining combinations of approaches. Nevertheless, there were some articles in our review which focused on more than one countermeasure. For example, Ng et al., (2021) used two types of platform interventions: 1) an account-level intervention (forwarding restriction) which is a "preventive" approach, and 2) a content-level intervention (flagging fake news) as a "mitigation/remedy". Also, several papers (mostly in Computer Science) studied both "detection" and "mitigation" to counter fake news (Kim et al., 2018; Papanastasiou, 2020; Sharma et al., 2019; Shu et al., 2019). In addition, Helm & Nasu, (2021) discussed three different countermeasures to combat fake news: 1) information correction

(mitigation), 2) blocking or removing contents/accounts (prevention), and 3) criminal sanctions (deterrence). However, this was a conceptual study, and they didn't examine different approaches. Based on our review, there is a dearth of studies on combining more than two stages (e.g., three or all four stages). Future research can look at these combinations to provide a more comprehensive picture. Instead of a micro-level look and seeing only each stage at once, future research can take a more holistic view to combat fake news.

Second, some of the countermeasures to combat fake news may be classified under more than one category. For example, *"Fake news influence minimization: limiting the scope of fake news spread by blocking certain nodes (users)"* is referred to as a mitigation strategy in all the highly cited papers (Sharma et al., 2019; Shu, Bernard, et al., 2019). Examples of articles using this approach are (Amoruso et al., 2020; Lin et al., 2019; Shrivastava et al., 2020). In this paper, we classified the articles using this approach under the **"mitigation"** strategy, to be consistent with the literature. However, one may also consider it as a preventive approach because it prevents the further spreading of fake news by blocking some nodes/users. Also, we classified "Flagging fake news" as a mitigation/remedy strategy because ***first***, it happens after fake news is detected (detection can be done manually by e.g., fact-checkers or automatically by e.g., ML & DL approaches). ***Second***, mitigation/remedy is defined as "reducing the harmful impacts of abuse (in the context of this study, reducing negative impacts of fake news". Research shows that flagging fake news reduces the impacts of fake news by triggering users' critical thinking.

Third, our focus was on combating fake news on social media platforms. However, fake news spreads through different channels: social networking sites such as X or Facebook, fake news websites, and peer-to-peer sharing via e.g., messaging apps such as WhatsApp, Telegram, etc. For fake news on social media platforms such as Facebook, given the right incentive, the platform can more easily implement certain control methods. For peer-to-peer sharing via messaging apps, neither the platform nor the government can easily insert itself in the process. In such cases, some of the countermeasures in our framework such as increasing users' awareness, inoculation, educational campaigns, and media literacy initiatives (mentioned in the deterrence and prevention stages) may help counter fake news. Future research can further investigate the approaches to address the peer-to-peer sharing of fake news.

Last, we acknowledge that we appropriated the countermeasures in the Straub Model as a set of containers for the individual articles. We are not commenting on the truth value of the framework or whether it is better or worse than any other framework, just that it provides sufficient value for decomposing the articles into logical categories for further analysis. We are not testing the framework as though it were a prediction or theory, we are just using it to provide a basis for analysis of the literature.

1.1 Study 2 Conclusion

Combating fake news on social media in an extremely complex and challenging problem which requires a multidisciplinary effort. Scholars across various disciplines from computer science and information systems to social science should work collaboratively

to address this serious issue. Our findings suggest that most of the fake news research have focused on *detection methods* and was mostly published in computer science outlets.

However, there is an opportunity and need to know more about deterring and blocking the creation and dissemination of fake news before the detection phase, and about reducing their harms and further limiting their spread after they are detected. We believe that the IS community take a more active role in addressing the fake news challenge and propose that efforts can be guided by the provided framework. Taking this holistic view can help IS scholars examine important research areas, and ultimately develop more comprehensive, synergetic multi-stage plans for combating fake news on social media. Fake news is an ongoing phenomenon. It is like a virus that will never disappear, but we need to keep fighting it.

Table 27: Approaches to Combat Fake News and Example References for Each Stage

	Combat Approaches	Sample Articles (References)
Deterrence	Establish Laws, Regulations, and Policies/Increase Public Awareness about Policies	*(Batchelor, 2017; Delellis & Rubin, 2018; Helm & Nasu, 2021; Jones-Jang et al., 2021), (Haciyakupoglu et al., 2018), (Hartley & Vu, 2020), (Morgan, 2018), (Nugent, 2018), (Jang & Kim, 2018), (Kreiss & McGregor, 2019) (Flew et al., 2019), (Hemphill, 2019), (Hensel & Kacprzak, 2021), (Smyth, 2019), (D'Arcy et al., 2009), (Li et al., 2019), (Whitman et al., 2001)*
Prevention	Inoculation (Prebunking)	*(Banas & Miller, 2013), (Cook et al., 2017), (Cook, 2016), (Jolley & Douglas, 2017),(Bolsen & Druckman, 2015), (Roozenbeek & Van Der Linden, 2019), (Roozenbeek & van der Linden, 2019), (Van der Linden et al., 2017), (Basol et al., 2020), (Lewandowsky & Van Der Linden, 2021)*
	Inoculating through Education/Misconception-based Learning	*(De Paor & Heravi, 2020), (McCuin et al., 2014), (Cook et al., 2014),(Cook, 2022), (Mihailidis & Viotty, 2017), (Kowalski & Taylor, 2009), (Tippett, 2010),(Banks, 2017), (Walton & Hepworth, 2011), (Batchelor, 2017), (Delellis & Rubin, 2018),(Jones-Jang et al., 2021), (Lewandowsky et al., 2017), (Ecker et al., 2017), (Lefkowitz, 2017), (Schuenemann & Cook, 2015)*
	Block Malicious Accounts on Social Media	*(Batchelor, 2017; Delellis & Rubin, 2018; Jones-Jang et al., 2021; Timberg & Dwoskin, 2018), (Coleman, 2021), (Amoruso et al., 2020), (Zhang et al., 2016), (Ng et al., 2021), (Chakraborty et al., 2016)*

Detection	Fact-checking (Manual, Crowdsourcing, Computational)		*(Wang, 2017), (Shu et al., 2020), (Hassan et al., 2017), (Babakar, 2018), (Gencheva et al., 2017), (Kim et al., 2018), (Ratkiewicz et al., 2011),(Shao et al., 2016), (Konstantinovskiy et al., 2021), (Shi & Weninger, 2016), (Ciampaglia et al., 2015), (Atanasova et al., 2019)*
	Automated Algorithmic Solutions	Machine Learning (ML) & Deep Learning (DL)	*ML: (Castillo et al., 2011), (Ma et al., 2015), (Kwon et al., 2017),(Hamidian & Diab, 2019), (Wu et al., 2015), (Shu, Wang, et al., 2019), (Yang et al., 2012), (Vosoughi et al., 2017), (Kumar et al., 2016),(Z. Jin et al., 2016), (Ahmad et al., 2020) DL: (Ma et al., 2016, 2018), (Qian et al., 2018), (Bian et al., 2020),(Wang et al., 2018), (Kaliyar et al., 2020), (Sahoo & Gupta, 2021),(Nasir et al., 2021), (Nguyen et al., 2020), (Yuan et al., 2021)*
		Other Methods (spread pattern, statistics, etc.)	*(Kim et al., 2018), (Papanastasiou, 2020), (Wang & Terano, 2015),(Wang et al., 2017), (Kumar & Geethakumari, 2014), (Chen et al., 2016)*
Mitigation (Remedy)	Minimizing the Influence of Fake News		*(Fan et al., 2013), (Kotnis & Kuri, 2014), (Wang et al., 2014), (Wang et al., 2017), (Kimura et al., 2009), (Tambuscio et al., 2015), (He et al., 2015)*
	Spreading Truth to discredit fake news		*(Budak et al., 2011), (Nguyen et al., 2012), (Tripathy et al., 2010),(Tong et al., 2017), (Yang et al., 2020), (He et al., 2015), (King et al., 2021)*
	Platform Interventions (Content-level)		*(Moravec et al., 2020; Moravec et al., 2019, 2022), (Kim et al., 2019a), (Kim & Dennis, 2019),(Figl et al., 2019), (Garrett & Poulsen, 2019),(Pennycook, Bear, et al., 2020; Pennycook, McPhetres, et al., 2020), (Ng et al., 2021), (Gimpel et al., 2021)*

Chapter 5

5 Research Contributions

5.1 Potential Contributions

This research aims to address the problem of fake news on social media. This research contributes to theory and research in several ways. ***First***, we provide an overview of definitions of fake news and several related terms in the literature. This will help scholars to have a better understanding of the term "fake news" and other relevant terms that are often used interchangeably in the literature. As mentioned earlier, various types of false information have different characteristics, intentions, and impacts. Understanding the distinctions between different types of false stories can help develop classification tools and more efficient interventions, e.g., prioritizing debunking fake news with more severe impacts over those with negligible impacts.

Second, while previous research focused on fake news characteristics from various perspectives (e.g., propagation, content, users, etc.), this research is among the first to study fake news from the perspective of its lifecycle. I study the fake news ecosystem by identifying the stages of the fake news lifecycle (creation, propagation, impact) and reviewing relevant research for each stage. This will also help researchers (especially if they are new to this topic) identify relevant studies in each stage.

Third, the first study contributes to research on the spread and impact stages of the fake news lifecycle. Our findings reveal that political false tweets have larger cascade sizes,

indicating that more users are involved in political cascades. Regarding cascade lifetime, our results show that conspiracy false tweets have the longest lifetime, indicating that conspiracy content may remain relevant for longer compared to other types of false information. Finally, we found that most political false tweets spread faster on average, although a small percentage of conspiracy false tweets have the potential to spread more intensively over time.

In terms of the impact of fake news, this research investigates users' reactions, both emotionally and attitudinally, to different types of false information. Although there is a large body of research on fake news propagation, most studies compare the spread of true vs. false news. The results of sentiment analysis did not find a notable difference in the sentiment of various types of false tweets or in the sentiment of users' responses to them (Figure 14 and Figure 15). However, users' responses to various types of false tweets show different emotions (Figure 16). For example, including a clickbait component increases certain emotions in users' responses but reduces *trust*. The results of this study also provide insights into how users' responses to different types of fake news (e.g., in terms of sentiment, various emotions or stance) impact their propagation (e.g., in terms of retweet count). A more detailed discussion of the findings of Study 1 is provided in section 3.5.

Fourth, the second study contributes to the research on combating fake news on social media. Although there are many studies about combating fake news, most of them focus on fake news detection approaches. However, regardless of the algorithm, detection approaches address the fake news problem after its diffusion. As mentioned earlier, it is

essential to devise strategies to stop fake news during the whole lifecycle of fake news, even before it is created. To this end, the second study in this thesis extends prior research by providing a framework (adopted from security) to combat fake news in four stages: deterrence, prevention, detection, and mitigation/remedy. The proposed framework provides a comprehensive picture of combating fake news on social media by addressing this problem in all stages of the fake news lifecycle. This research calls for interdisciplinary collaborations between researchers from different fields to address this problem.

Lastly, the second study contributes to research on fake news by providing a comprehensive and systematic review of 164 articles related to the four countermeasures of the framework. I also provided some descriptive statistics of the reviewed papers to better depict the current status of fake news combating research. In addition, I used the adapted framework to discuss the approaches to combat fake news on social media, the challenges involved, the limitations of the current approaches, and directions for future research. These should allow the IS community to take a more systematic and active role in combating fake news, not just in fake news detection. It also helps readers to grasp the whole picture of the research frontier.

This research has several **practical implications**. ***First***, the findings from the first study provide practical insights into how different types of false information spread and responded to, which helps develop future classification tools (to detect different types of false information) and also designing more efficient interventions (e.g., which type of fake news to fact-check or debunk first) to prevent the proliferation of fake news on

social media platforms. It improves the detection of fake news by understanding which type of false information to address first (e.g., those with faster spread or higher impacts). In practice, this is of great importance because it often costs millions of dollars for social media platforms to contain the spread of fake news. Fake news detection models can incorporate features such as a larger cascade size and high spread speed to identify political false tweets and use prolonged cascade lifetimes to detect conspiracy false tweets. In addition, there can be different mitigation strategies for different types of false content. For example, since conspiracy false tweets have a longer lifetime, mitigation efforts must be sustained over a longer period. This may include continuous monitoring and repeated spread of corrective information.

Second, social media platform managers can use the proposed framework in the second study as an effective approach to systematically combat fake news on their platforms. As mentioned before, most prior studies focused on fake news detection methods. A systematic approach to combating fake news can create synergies and allow for more careful and, hopefully, effective plans to tackle the problem.

Finally, fake news is a complex problem which is multidisciplinary. As mentioned earlier, fake news detection methods are essential, but more work is needed to curb the spread of fake news effectively. The second study provides insight into various approaches and strategies to combat fake news during all stages of its lifecycle (not only after it spreads). It also highlights the need for multidisciplinary efforts and collaboration between governments, policymakers, and managers of social media platforms to effectively address the ongoing problem of fake news on social media.

References

Adams, R. J., Smart, P., & Huff, A. S. (2017). Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews*, *19*(4), 432–454.

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, *2020*.

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Amoruso, M., Anello, D., Auletta, V., Cerulli, R., Ferraioli, D., & Raiconi, A. (2020). Contrasting the spread of misinformation in online social networks. *Journal of Artificial Intelligence Research*, *69*, 847–879.

Anagnostopoulos, A., Bessi, A., Caldarelli, G., Del Vicario, M., Petroni, F., Scala, A., Zollo, F., & Quattrociocchi, W. (2014). Viral Misinformation: The Role of Homophily and Polarization. *arXiv E-Prints*, arXiv-1411.

Andrews, C., Fichet, E., Ding, Y., Spiro, E. S., & Starbird, K. (2016). Keeping up with the tweet-dashians: The impact of'official'accounts on online rumoring. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 452–465.

Arif, A., Shanahan, K., Chou, F.-J., Dosouto, Y., Starbird, K., & Spiro, E. S. (2016). How information snowballs: Exploring the role of exposure in online rumor propagation. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 466–477.

Aro, J. (2016). The cyberspace war: Propaganda and trolling as warfare tools. *European View*, *15*(1), 121–132.

Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., & Glass, J. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, *11*(3), 1–27.

Babakar, M. (2018, May 29). *Crowdsourced Factchecking*. Full Fact. https://fullfact.org/blog/2018/may/crowdsourced-factchecking/

Babcock, M., Beskow, D. M., & Carley, K. M. (2019). Different faces of false: The spread and curtailment of false information in the black panther twitter discussion. *Journal of Data and Information Quality (JDIQ)*, *11*(4), 1–15.

Babcock, M., Cox, R. A. V., & Kumar, S. (2018). *Diffusion of pro-and anti-false information tweets: The Black Panther movie case*.

Babcock, M., Cox, R. A. V., & Kumar, S. (2019). Diffusion of pro-and anti-false information tweets: The Black Panther movie case. *Computational and Mathematical Organization Theory*, *25*(1), 72–84.

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154–175.

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, *39*(2), 184–207.

Banks, M. (2017). Fighting fake news. *American Libraries*, *48*(3–4), 18–21.

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1).

Batchelor, O. (2017). Getting out the truth: The role of libraries in the fight against fake news. *Reference Services Review*.

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs Conspiracy: Collective Narratives in the Age of Misinformation. *Plos One*, *10*(2), e0118093. https://doi.org/10.1371/journal.pone.0118093

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. *First Monday*, *21*(11–7).

Bessi, A., Petroni, F., Del Vicario, M., Zollo, F., Anagnostopoulos, A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Viral misinformation: The role of homophily and polarization. *Proceedings of the 24th International Conference on World Wide Web*, 355–356.

Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 549–556.

Biyani, P., Tsioutsiouliklis, K., & Blackmer, J. (2016). " 8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. *Thirtieth AAAI Conference on Artificial Intelligence*.

Blom, J. N., & Hansen, K. R. (2015). Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, *76*, 87–100.

Bolsen, T., & Druckman, J. N. (2015). Counteracting the politicization of science. *Journal of Communication*, *65*(5), 745–769.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55.

Botha, J., & Pieterse, H. (2020). Fake news and deepfakes: A dangerous threat for 21st century information security. *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and Publishing Limited*, 57.

Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news. *PLoS One*, *16*(6), e0253717.

Budak, C., Agrawal, D., & El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. *Proceedings of the 20th International Conference on World Wide Web*, 665–674.

Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. *MIS Quarterly*, 523–548.

Carrieri, V., Madio, L., & Principe, F. (2019). Vaccine hesitancy and (fake) news: Quasi-experimental evidence from Italy. *Health Economics*, *28*(11), 1377–1382.

Carvalho, C., Klagge, N., & Moench, E. (2011). The persistent effects of a false news shock. *Journal of Empirical Finance*, *18*(4), 597–615.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web*, 675–684.

Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 9–16.

Chen, W., Yeo, C. K., Lau, C. T., & Lee, B. S. (2016). Behavior deviation: An anomaly detection view of rumor preemption. *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 1–7.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as" false news". *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, 15–19.

Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., Xiao, X., Nazarian, S., & Bogdan, P. (2021). A COVID-19 rumor dataset. *Frontiers in Psychology*, *12*, 644801.

Cheng Nie, Zhiqiang (Eric) Zheng, & Sarkar, S. (2022). Competing with the Sharing Economy: Incumbents' Reaction on Review Manipulation. *MIS Quarterly*, *46*(3), 1573–1602. https://doi.org/10.25300/MISQ/2022/15666

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS One*, *10*(6), e0128193.

Cichocka, A., Marchlewska, M., & De Zavala, A. G. (2016). Does self-love or self-hate predict conspiracy beliefs? Narcissism, self-esteem, and the endorsement of conspiracy theories. *Social Psychological and Personality Science*, *7*(2), 157–166.

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

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2021). Trends in combating fake news on social media–a survey. *Journal of Information and Telecommunication*, *5*(2), 247–266.

Compton, J. (2013). Inoculation theory. *The SAGE Handbook of Persuasion: Developments in Theory and Practice*, *2*, 220–237.

Cook, J. (2016). Countering climate science denial and communicating scientific consensus. In *Oxford Research Encyclopedia of Climate Science*.

Cook, J. (2019). *Understanding and countering misinformation about climate change.*

Cook, J. (2022). Understanding and countering misinformation about climate change. *Research Anthology on Environmental and Societal Impacts of Climate Change*, 1633–1658.

Cook, J., Bedford, D., & Mandia, S. (2014). Raising climate literacy through addressing misinformation: Case studies in agnotology-based learning. *Journal of Geoscience Education*, *62*(3), 296–306.

Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS One*, *12*(5), e0175799.

Dale, R. (2017). NLP in a post-truth world. *Natural Language Engineering*, *23*(2), 319–324.

D'Arcy, J., Hovav, A., & Galletta, D. (2009). User awareness of security countermeasures and its impact on information systems misuse: A deterrence approach. *Information Systems Research*, *20*(1), 79–98.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). Botornot: A system to evaluate social bots. *Proceedings of the 25th International Conference Companion on World Wide Web*, 273–274.

De Paor, S., & Heravi, B. (2020). Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news. *The Journal of Academic Librarianship*, *46*(5), 102218.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Delellis, N. S., & Rubin, V. L. (2018). Educators' perceptions of information literacy and skills required to spot 'fake news.' *Proceedings of the Association for Information Science and Technology*, *55*(1), 785–787.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629.

Dias, M., & Becker, K. (2016). Inf-ufrgs-opinion-mining at semeval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 378–383.

Doerr, B., Fouz, M., & Friedrich, T. (2012). Why rumors spread so quickly in social networks. *Communications of the ACM*, *55*(6), 70–75.

Douglas, K. M., & Sutton, R. M. (2008). The hidden impact of conspiracy theories: Perceived and actual influence of theories surrounding the death of Princess Diana. *The Journal of Social Psychology*, *148*(2), 210–222.

Ecker, U. K., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, *6*(2), 185–192.

Fan, L., Lu, Z., Wu, W., Thuraisingham, B., Ma, H., & Bi, Y. (2013). Least cost rumor blocking in social networks. *2013 IEEE 33rd International Conference on Distributed Computing Systems*, 540–549.

Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., & Zha, H. (2017). Fake news mitigation via point process based intervention. *International Conference on Machine Learning*, 1097–1106.

Figl, K., Kießling, S., Rank, C., & Vakulenko, S. (2019). *Fake News Flags, Cognitive Dissonance, and the Believability of Social Media Posts*.

Fisher, M., Cox, J. W., & Hermann, P. (2016). Pizzagate: From rumor, to hashtag, to gunfire in DC. *Washington Post*, *6*.

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, *10*(1), 33–50.

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, *38*, 127–150.

Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014). Rumor cascades. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1).

Funke, D., & Flamini, D. (2022, January 21). A guide to anti-misinformation actions around the world. *Poynter*. https://www.poynter.org/ifcn/anti-misinformation-actions/

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, *11*(1), 3.

Garrett, R. K., & Poulsen, S. (2019). Flagging Facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. *Journal of Computer-Mediated Communication*, *24*(5), 240–258.

Gencheva, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., & Koychev, I. (2017). A context-aware approach for detecting worth-checking claims in political debates. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 267–276.

George, J. F., Gupta, M., Giordano, G., Mills, A. M., Tennant, V. M., & Lewis, C. C. (2018). The effects of communication media and culture on deception detection accuracy. *MIS Quarterly*, *42*(2), 551–575.

George, J., Gerhart, N., & Torres, R. (2021). Uncovering the Truth about Fake News: A Research Model Grounded in Multi-Disciplinary Literature. *Journal of Management Information Systems*, *38*(4), 1067–1094. https://doi.org/10.1080/07421222.2021.1990608

Gianvecchio, S., Xie, M., Wu, Z., & Wang, H. (2008). Measurement and Classification of Humans and Bots in Internet Chat. *USENIX Security Symposium*, 155–170.

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems*, *38*(1), 196–221.

Goertzel, T. (1994). Belief in Conspiracy Theories. *Political Psychology*, *15*(4).

Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., & Everett, J. B. (2018). Fake news vs satire: A dataset and analysis. *Proceedings of the 10th ACM Conference on Web Science*, 17–21.

Gopal, R. D., & Sanders, G. L. (1997). Preventive and deterrent controls for software piracy. *Journal of Management Information Systems*, *13*(4), 29–47.

Gorrell, G., Bontcheva, K., Derczynski, L., Kochkina, E., Liakata, M., & Zubiaga, A. (2018). Rumoureval 2019: Determining rumour veracity and support for rumours. *arXiv Preprint arXiv:1809.06683*.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545.

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480.

Gupta, A., Lamba, H., & Kumaraguru, P. (2013). $1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter. *2013 APWG eCrime Researchers Summit*, 1–12.

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. *Proceedings of the 22nd International Conference on World Wide Web*, 729–736.

Haciyakupoglu, G., Hui, J. Y., Suguna, V. S., Leong, D., & Rahman, M. F. B. A. (2018). *Countering fake news: A survey of recent global initiatives*.

Hamidian, S., & Diab, M. T. (2019). Rumor detection and classification for twitter data. *arXiv Preprint arXiv:1912.08926*.

Hartley, K., & Vu, M. K. (2020). Fighting fake news in the COVID-19 era: Policy insights from an equilibrium model. *Policy Sciences*, *53*(4), 735–758.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., & Nayak, A. K. (2017). Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, *10*(12), 1945–1948.

He, Z., Cai, Z., & Wang, X. (2015). Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks. *2015 IEEE 35Th International Conference on Distributed Computing Systems*, 205–214.

Helm, R. K., & Nasu, H. (2021). Regulatory responses to 'fake news' and freedom of expression: Normative and empirical evaluation. *Human Rights Law Review*, *21*(2), 302–328.

Hemphill, T. A. (2019). 'Techlash', responsible innovation, and the self-regulatory organization. *Journal of Responsible Innovation*, *6*(2), 240–247.

Hensel, P. G., & Kacprzak, A. (2021). Curbing cyberloafing: Studying general and specific deterrence effects with field evidence. *European Journal of Information Systems*, *30*(2), 219–235.

Horne, B., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1).

Hornsey, M. J., Chapman, C. M., Alvarez, B., Bentley, S., Salvador Casara, B. G., Crimston, C. R., Ionescu, O., Krug, H., Preya Selvanathan, H., & Steffens, N. K.

(2021). To what extent are conspiracy theorists concerned for self versus others? A COVID-19 test case. *European Journal of Social Psychology*, *51*(2), 285–293.

Hornsey, M. J., Finlayson, M., Chatwood, G., & Begeny, C. T. (2020). Donald Trump and vaccination: The effect of political identity, conspiracist ideation and presidential tweets on vaccine hesitancy. *Journal of Experimental Social Psychology*, *88*, 103947.

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. *Health Psychology*, *37*(4), 307.

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

Hou, Y., van der Putten, P., & Verberne, S. (2022). The COVMis-stance dataset: Stance detection on twitter for COVID-19 misinformation. *arXiv Preprint arXiv:2204.02000*.

Imhoff, R., & Lamberty, P. (2018). How paranoid are conspiracy believers? Toward a more fine-grained understanding of the connect and disconnect between paranoia and belief in conspiracy theories. *European Journal of Social Psychology*, *48*(7), 909–926.

Imhoff, R., & Lamberty, P. (2020). A bioweapon or a hoax? The link between distinct conspiracy beliefs about the coronavirus disease (COVID-19) outbreak and pandemic behavior. *Social Psychological and Personality Science*, *11*(8), 1110–1118.

Ireton, C., & Posetti, J. (2018). *Journalism, fake news & disinformation: Handbook for journalism education and training*. Unesco Publishing.

Jang, S. M., & Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in Human Behavior*, *80*, 295–302.

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on twitter. *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, 1–9.

Jin, F., Wang, W., Zhao, L., Dougherty, E., Cao, Y., Lu, C.-T., & Ramakrishnan, N. (2014). Misinformation propagation in the age of twitter. *Computer*, *47*(12), 90–94.

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). News verification by exploiting conflicting social viewpoints in microblogs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1).

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420.

Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, *47*(8), 459–469.

Jolley, D., Douglas, K. M., Leite, A. C., & Schrader, T. (2019). Belief in conspiracy theories and intentions to engage in everyday crime. *British Journal of Social Psychology*, *58*(3), 534–549.

Jolley, D., & Paterson, J. L. (2020). Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology*, *59*(3), 628–640.

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist*, *65*(2), 371–388.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765–11788.

Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, *61*, 32–44.

Kawintiranon, K., & Singh, L. (2021). Knowledge enhanced masked language model for stance detection. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4725–4735.

Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy*, *96*(3), 109–126.

Khan, T., Michalas, A., & Akhunzada, A. (2021). Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications*, *190*, 103112.

Kim, A., & Dennis, A. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, *43*(3).

Kim, A., Moravec, P., & Dennis, A. (2019a). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, *36*(3), 931–968.

Kim, A., Moravec, P., & Dennis, A. (2019b). When Do Details Matter? Source Rating Summaries and Details in the Fight against Fake News on Social Media. *Source Rating Summaries and Details in the Fight against Fake News on Social Media (September 6, 2019). Kelley School of Business Research Paper*, *19–52*.

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 324–332.

Kim, M., Paini, D., & Jurdak, R. (2020). Real-world diffusion dynamics based on point process approaches: A review. *Artificial Intelligence Review*, *53*, 321–350.

Kimura, M., Saito, K., & Motoda, H. (2009). Blocking links to minimize contamination spread in a social network. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *3*(2), 1–23.

King, K. K., Wang, B., Escobari, D., & Oraby, T. (2021). Dynamic Effects of Falsehoods and Corrections on Social Media: A Theoretical Modeling and Empirical Evidence. *Journal of Management Information Systems*, *38*(4), 989–1010.

Kitchens, B., Johnson, S. L., & Gray, P. (2020). Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption. *Mis Quarterly*, *44*(4), 1619–1649. https://doi.org/10.25300/MISQ/2020/16371

Kogan, S., Moskowitz, T. J., & Niessner, M. (2019). Fake news: Evidence from financial markets. *Available at SSRN 3237763*.

Kong, Q., Ram, R., & Rizoiu, M.-A. (2021). Evently: Modeling and Analyzing Reshare Cascades with Hawkes Processes. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 1097–1100.

Konstantinovskiy, L., Price, O., Babakar, M., & Zubiaga, A. (2021). Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats: Research and Practice*, *2*(2), 1–16.

Kotnis, B., & Kuri, J. (2014). Cost effective rumor containment in social networks. *arXiv Preprint arXiv:1403.6315*.

Kowalski, P., & Taylor, A. K. (2009). The effect of refuting misconceptions in the introductory psychology class. *Teaching of Psychology*, *36*(3), 153–159.

Kreiss, D., & McGregor, S. C. (2019). The "arbiters of what our voters see": Facebook and Google's struggle with policy, process, and enforcement around political advertising. *Political Communication*, *36*(4), 499–522.

Kuem, J., Ray, S., Siponen, M., & Kim, S. S. (2017). What Leads to Prosocial Behaviors on Social Networking Services: A Tripartite Model. *Journal of Management Information Systems*, *34*(1), 40–70. https://doi.org/10.1080/07421222.2017.1296744

Kumar, K. K., & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-Centric Computing and Information Sciences*, *4*(1), 1–22.

Kumar, S., Cheng, J., Leskovec, J., & Subrahmanian, V. S. (2017). An army of me: Sockpuppets in online discussion communities. *Proceedings of the 26th International Conference on World Wide Web*, 857–866.

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

Kumar, S., West, R., & Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. *Proceedings of the 25th International Conference on World Wide Web*, 591–602.

Kwon, S., Cha, M., & Jung, K. (2017). Rumor detection over varying time windows. *PloS One*, *12*(1), e0168344.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. *2013 IEEE 13th International Conference on Data Mining*, 1103–1108.

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. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lefkowitz, M. (2017, March). *Library tackles fake news with workshops, resources, advice*. Cornell Chronicle. https://news.cornell.edu/stories/2017/03/library-tackles-fake-news-workshops-resources-advice

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369.

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

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384.

Li, L., He, W., Xu, L., Ash, I., Anwar, M., & Yuan, X. (2019). Investigating the impact of cybersecurity policy awareness on employees' cybersecurity behavior. *International Journal of Information Management*, *45*, 13–24.

Lim, S. (2020). Academic library guides for tackling fake news: A content analysis. *The Journal of Academic Librarianship*, *46*(5), 102195.

Lin, Y., Cai, Z., Wang, X., & Hao, F. (2019). Incentive mechanisms for crowdblocking rumors in mobile social networks. *IEEE Transactions on Vehicular Technology*, *68*(9), 9220–9232.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). *Detecting rumors from microblogs with recurrent neural networks*.

Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1751–1754.

Ma, J., Gao, W., & Wong, K.-F. (2017). *Detect rumors in microblog posts using propagation structure via kernel learning*.

Ma, J., Gao, W., & Wong, K.-F. (2018). *Rumor detection on twitter with tree-structured recursive neural networks*.

Marabelli, M., Vaast, E., & Li, J. L. (2021). Preventing the digital scars of COVID-19. *European Journal of Information Systems*, *30*(2), 176–192.

McCuin, J. L., Hayhoe, K., & Hayhoe, D. (2014). Comparing the effects of traditional vs. Misconceptions-based instruction on student understanding of the greenhouse effect. *Journal of Geoscience Education*, *62*(3), 445–459.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we RT? *Proceedings of the First Workshop on Social Media Analytics*, 71–79.

Mihailidis, P., & Viotty, S. (2017). Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American Behavioral Scientist*, *61*(4), 441–454.

Mohammad, S. M., & Turney, P. D. (2013). CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moody, G. D., Siponen, M., & Pahnila, S. (2018). Toward a unified model of information security policy compliance. *MIS Quarterly*, *42*(1).

Moravec, P., Kim, A., & Dennis, A. (2020). Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media. *Information Systems Research*, *31*(3), 987–1006.

Moravec, P., Kim, A., Dennis, A., & Minas, R. (2018). Do you really know if it's true? How asking users to rate stories affects belief in fake news on social media. *How Asking Users to Rate Stories Affects Belief in Fake News on Social Media (October 22, 2018). Kelley School of Business Research Paper*, *18–89*.

Moravec, P. L., Kim, A., Dennis, A. R., & Minas, R. K. (2022). Do you really know if it's true? How asking users to rate stories affects belief in fake news on social media. *Information Systems Research*.

Moravec, P. L., Minas, R. K., & Dennis, A. (2019). Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense At All. *MIS Quarterly*, *43*(4), 1343–1360.

Moravec, P., Minas, R., & Dennis, A. (2018). Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business Research Paper*, *18–87*.

Morgan, S. (2018). Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy*, *3*(1), 39–43.

Müller, M., Salathé, M., & Kummervold, P. E. (2023). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, *6*, 1023281.

Mutlu, E. C., Oghaz, T., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., Ozmen, O., & 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.

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007.

Ng, K. C., Tang, J., & Lee, D. (2021). The Effect of Platform Intervention Policies on Fake News Dissemination and Survival: An Empirical Examination. *Journal of Management Information Systems*, *38*(4), 898–930.

Nguyen, N. P., Yan, G., Thai, M. T., & Eidenbenz, S. (2012). Containment of misinformation spread in online social networks. *Proceedings of the 4th Annual ACM Web Science Conference*, 213–222.

Nguyen, V.-H., Sugiyama, K., Nakov, P., & Kan, M.-Y. (2020). Fang: Leveraging social context for fake news detection using graph representation. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1165–1174.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nugent, C. (2018). *How France's Potential Law Banning Fake News Could Work*. Time. https://time.com/5304611/france-fake-news-law-macron/

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*.

Oh, O., Kwon, K. H., & Rao, H. R. (2010). An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010. *Icis*, *231*, 7332–7336.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv Preprint arXiv:1107.4557*.

Papageorgis, D., & McGuire, W. J. (1961). The generality of immunity to persuasion produced by pre-exposure to weakened counterarguments. *The Journal of Abnormal and Social Psychology*, *62*(3), 475.

Papanastasiou, Y. (2020). Fake news propagation and detection: A sequential model. *Management Science*, *66*(5), 1826–1846.

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

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

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

Petratos, P. N. (2021). Misinformation, disinformation, and fake news: Cyber risks to business. *Business Horizons*, *64*(6), 763–774.

Pierri, F., & Ceri, S. (2019). False news on social media: A data-driven survey. *ACM Sigmod Record*, *48*(2), 18–27.

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2019). STANCY: Stance classification based on consistency cues. *arXiv Preprint arXiv:1910.06048*.

Qian, F., Gong, C., Sharma, K., & Liu, Y. (2018). Neural User Response Generator: Fake News Detection with Collective User Intelligence. *IJCAI*, *18*, 3834–3840.

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo Chambers on Facebook. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2795110

Rapoza, K. (2017). Can 'fake news' impact the stock market? *Forbes News*.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams. *Proceedings of the 20th International Conference Companion on World Wide Web*, 249–252.

Rizoiu, M.-A., Graham, T., Zhang, R., Zhang, Y., Ackland, R., & Xie, L. (2018). # debatenight: The role and influence of socialbots on twitter during the 1st 2016 us presidential debate. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1).

Rizoiu, M.-A., Lee, Y., Mishra, S., & Xie, L. (2017). Hawkes processes for events in social media. In *Frontiers of multimedia research* (pp. 191–218).

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10.

Roozenbeek, J., & Van Der Linden, S. (2019). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580.

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and Knowledge*, *103*, 135.

Rubin, V. L. (2010). On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–10.

Rubin, V. L. (2019). Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions. *Journal of Documentation*.

Rubin, V. L., Chen, Y., & Conroy, N. K. (2015). Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4.

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806.

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983.

Saveski, M., Roy, B., & Roy, D. (2021). The structure of toxic conversations on Twitter. *Proceedings of the Web Conference 2021*, 1086–1097.

Schuenemann, K. C., & Cook, J. (2015). Using" Making Sense of Climate Science Denial" MOOC videos in a college course. *AGU Fall Meeting Abstracts*, *2015*, ED12A-01.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. *Proceedings of the 25th International Conference Companion on World Wide Web*, 745–750.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. *arXiv Preprint arXiv:1707.07592*, *96*, 104.

Shao, C., Hui, P.-M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PloS One*, *13*(4), e0196087.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1–42.

Sharma, K., Seo, S., Meng, C., Rambhatla, S., Dua, A., & Liu, Y. (2020). Coronavirus on social media: Analyzing misinformation in Twitter conversations. *arXiv Preprint arXiv:2003.12309*.

Shi, B., & Weninger, T. (2016). Fact checking in heterogeneous information networks. *Proceedings of the 25th International Conference Companion on World Wide Web*, 101–102.

Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, *83*, 278–287.

Shirish, A., Srivastava, S. C., & Chandra, S. (2021). Impact of mobile connectivity and freedom on fake news propensity during the COVID-19 pandemic: A cross-country empirical examination. *European Journal of Information Systems*, *30*(3), 322–341.

Shore, J., Baek, J., & Dellarocas, C. (2018). Network structure and patterns of information diversity on twitter. *MIS Quarterly*, *42*(3), 849–872. https://doi.org/10.25300/MISQ/2018/14558

Shrivastava, G., Kumar, P., Ojha, R. P., Srivastava, P. K., Mohan, S., & Srivastava, G. (2020). Defensive modeling of fake news through online social networks. *IEEE Transactions on Computational Social Systems*, *7*(5), 1159–1167.

Shu, K., Bernard, H. R., & Liu, H. (2019). Studying fake news via network analysis: Detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining* (pp. 43–65). Springer.

Shu, K., Dumais, S., Awadallah, A. H., & Liu, H. (2020). Detecting fake news with weak social supervision. *IEEE Intelligent Systems*, *36*(4), 96–103.

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

Shu, K., Mahudeswaran, D., Wang, S., & Liu, H. (2020). Hierarchical propagation networks for fake news detection: Investigation and exploitation. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 626–637.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 312–320.

Silverman, C. (2016). This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed News*, *16*.

Silverman, C. (2017). What exactly is fake news. *The Fake Newsletter*, *26*.

Smyth, S. M. (2019). The Facebook Conundrum: Is it Time to Usher in a New Era of Regulation for Big Tech? *International Journal of Cyber Criminology*, *13*(2), 578–595.

Straub, D. W., & Welke, R. J. (1998). Coping with systems risk: Security planning models for management decision making. *MIS Quarterly*, 441–469.

Sunstein, C. R. (1999). The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, *91*.

Sunstein, C. R., & Vermeule, A. (2008). *Conspiracy theories*.

Swami, V., & Furnham, A. (2014). 12 Political paranoia and conspiracy theories. *Power, Politics, and Paranoia: Why People Are Suspicious of Their Leaders*, 218.

Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W., & Takayasu, H. (2015). Rumor diffusion and convergence during the 3.11 earthquake: A Twitter case study. *PLoS One*, *10*(4), e0121443.

Tambuscio, M., Ruffo, G., Flammini, A., & Menczer, F. (2015). Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. *Proceedings of the 24th International Conference on World Wide Web*, 977–982.

Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news" A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153.

Timberg, C., & Dwoskin, E. (2018). Twitter is sweeping out fake accounts like never before, putting user growth at risk. *Washington Post*.

https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/

Tippett, C. D. (2010). Refutation text in science education: A review of two decades of research. *International Journal of Science and Mathematics Education*, *8*(6), 951–970.

Tong, G., Wu, W., Guo, L., Li, D., Liu, C., Liu, B., & Du, D.-Z. (2017). An efficient randomized algorithm for rumor blocking in online social networks. *IEEE Transactions on Network Science and Engineering*, *7*(2), 845–854.

Tripathy, R. M., Bagchi, A., & Mehta, S. (2010). A study of rumor control strategies on social networks. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1817–1820.

Turel, O., & Osatuyi, B. (2021). Biased Credibility and Sharing of Fake News on Social Media: Considering Peer Context and Self-Objectivity State. *Journal of Management Information Systems*, *38*(4), 931–958.

Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B.-W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). *IEEE Access*, *8*, 156695–156706.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008.

van Der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology*, *11*, 2928.

Volkova, S., & Jang, J. Y. (2018). Misleading or falsification: Inferring deceptive strategies and types in online news and social media. *Companion Proceedings of the The Web Conference 2018*, 575–583.

Vosoughi, S., Mohsenvand, M. 'Neo,' & Roy, D. (2017). Rumor gauge: Predicting the veracity of rumors on Twitter. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *11*(4), 1–36.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Walton, G., & Hepworth, M. (2011). A longitudinal study of changes in learners' cognitive states during and following an information literacy teaching intervention. *Journal of Documentation*.

Wang, B., Chen, G., Fu, L., Song, L., & Wang, X. (2017). Drimux: Dynamic rumor influence minimization with user experience in social networks. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2168–2181.

Wang, N., Yu, L., Ding, N., & Yang, D. (2014). *Containment of misinformation propagation in online social networks with given deadline*.

Wang, S., Moise, I., Helbing, D., & Terano, T. (2017). Early signals of trending rumor event in streaming social media. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, *2*, 654–659.

Wang, S., & Terano, T. (2015). Detecting rumor patterns in streaming social media. *2015 IEEE International Conference on Big Data (Big Data)*, 2709–2715.

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv Preprint arXiv:1705.00648*.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*, 849–857.

Wardle, C. (2017). Fake news. It's complicated. *First Draft*, *16*, 1–11.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11).

Whitman, M. E., Townsend, A. M., & Aalberts, R. J. (2001). Information systems security and the need for policy. In *Information security management: Global challenges in the new millennium* (pp. 9–18). IGI Global.

Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. *2015 IEEE 31st International Conference on Data Engineering*, 651–662.

Xiao, B., & Benbasat, I. (2011). Product-related deception in e-commerce: A theoretical perspective. *Mis Quarterly*, 169–195.

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on sina weibo. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 1–7.

Yang, L., Li, Z., & Giua, A. (2020). Containment of rumor spread in complex social networks. *Information Sciences*, *506*, 113–130.

Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A survey on deepfake video detection. *IET Biometrics*, *10*(6), 607–624.

Yuan, H., Zheng, J., Ye, Q., Qian, Y., & Zhang, Y. (2021). Improving fake news detection with domain-adversarial and graph-attention neural network. *Decision Support Systems*, *151*, 113633.

Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)*, *11*(3), 1–37.

Zeng, E., Kohno, T., & Roesner, F. (2020). Bad news: Clickbait and deceptive ads on news and misinformation websites. *Workshop on Technology and Consumer Protection*, 1–11.

Zeng, L., Starbird, K., & Spiro, E. S. (2016). Rumors at the speed of light? Modeling the rate of rumor transmission during crisis. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 1969–1978.

Zhang, H., Alim, M. A., Li, X., Thai, M. T., & Nguyen, H. T. (2016). Misinformation in online social networks: Detect them all with a limited budget. *ACM Transactions on Information Systems (TOIS)*, *34*(3), 1–24.

Zhang, H., Zhang, H., Li, X., & Thai, M. T. (2015). Limiting the spread of misinformation while effectively raising awareness in social networks. *International Conference on Computational Social Networks*, 35–47.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., & Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522.

Zhou, X., Jain, A., Phoha, V. V., & Zafarani, R. (2020). Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, *1*(2), 1–25.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1–40.

Zimdars, M. (2016). False, misleading, clickbait-y, and satirical “news” sources. *Google Docs*.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, *51*(2), 1–36.

Zubiaga, A., Liakata, M., & Procter, R. (2016). Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv Preprint arXiv:1610.07363*.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS One*, *11*(3), e0150989.

Appendix

Table 28: Reviewed Articles Classified by Fake News Combat Stage (this table only contains papers relevant to combat fake news, excluding review papers, conceptual papers, etc.)

	Article	Database(s)	Source	Source Type	Publisher	Citation	Year
Combat (Deterrence)	*(Helm & Nasu, 2021)*	Google Scholar	Human Rights Law Review	Journal	Oxford University Press	11	2021
	(Li et al., 2019)	Google Scholar/ Scopus/ ScienceDirect	International Journal of Information Management	Journal (IS/ CS/ SS)	Elsevier	172	2019
	(Rubin, 2019)	Google Scholar	Journal of documentation	Journal (CS/IS/SS)	Emerald	83	2019
	(Lewandowsky et al., 2017)	Google Scholar/ Scopus	Journal of applied research in memory and cognition	Journal (Psychology)	Elsevier	1099	2017
	(D'Arcy et al., 2009)	Google Scholar/ EBSCOhost/ ProQuest/ JSTOR	Information systems research	IS Journal	INFORMS	1557	2009
	(Moody et al., 2018)	Google Scholar/ EBSCOhost	MIS quarterly	IS Journal	ACM Digital Library	367	2018
	(Hensel & Kacprzak, 2021)	Google Scholar	EJIS (European Journal of Information Systems)	IS Journal	Taylor & Francis	11	2021
	(Hartley & Vu, 2020)	Google Scholar/ EBSCOhost/ ProQuest	Policy Sciences	Journal (SS)	Springer	89	2020
	(Flew et al., 2019)	Google Scholar	Journal of Digital Media and Policy	Journal	Intellect	102	2019
	(Hemphill, 2019)	Google Scholar	Journal of Responsible Innovation	Journal (Business/IS)	Taylor & Francis	8	2019
	(Morgan, 2018)	Google Scholar	Journal of Cyber Policy	Journal	Taylor & Francis	133	2018

	(Smyth, 2019)	Google Scholar/ ProQuest	International Journal of Cyber Criminology	Journal (SS)	Not provided	10	2019
	(Haciyakupoglu et al., 2018)	Google Scholar	Rajaratnam School of International Studies (RSiS)	Report	RSiS	61	2018
	(Jang & Kim, 2018)	Google Scholar/ Scopus/ ScienceDirect	Computers in Human Behavior	Journal (SS/CS)	Elsevier	321	2018
	(Whitman et al., 2001)	Google Scholar/ Scopus/	Information Security Management: Global challenges in the new millennium	Book	IGI Global	112	2001
	(Kreiss & McGregor, 2019)	Google Scholar/ EBSCOhost	Political Communication	Journal (SS)	Taylor & Francis	71	2019
Combat (Prevention)	*(Ng et al., 2021)*	Google Scholar/ Scopus/ EBSCOhost	Journal of Management Information Systems (JMIS)	Journal (IS)	Taylor & Francis	3	2021
	(Zhang et al., 2016)	Google Scholar/ Scopus/ACM	ACM Transactions on Information Systems	Journal	ACM	62	2016
	(Gopal & Sanders, 1997)	Google Scholar/ EBSCOhost/ ProQuest/JSTOR	JMIS (Journal of Management Information Systems)	Journal (IS)	Taylor & Francis	405	1997
	(Marabelli et al., 2021)	Google Scholar	EJIS (European Journal of Information Systems)	Journal (IS)	Taylor & Francis	28	2021
	(Banas & Miller, 2013)	Google Scholar/ EBSCOhost/Scopus/Web of Science	Human Communication Research	Journal (Psychology/Social Science)	Oxford Univ. Press	150	2013

	(Batchelor, 2017)	Google Scholar/ Scopus/Emerald	Reference Services Review	Journal (SS)	Emerald	118	2017
	(Jolley & Douglas, 2017)	Google Scholar/ Scopus/	Journal of Applied Social Psychology	Journal (Psychology)	Wiley	265	2017
	(Chakraborty et al., 2016)	Google Scholar/ Scopus	IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)	Conference	IEEE	372	2016
	(Cook et al., 2017)	Google Scholar/ Scopus	PloS one	Journal (Multidisciplinary)	Public Library of Science	541	2017
	(Cook, 2016)	Google Scholar	Oxford Research Encyclopedia of Climate Science	Book		86	2016
	(Bolsen & Druckman, 2015)	Scholar/EBSCOhost/Scopus	Journal of Communication	Journal (SS)	Oxford Univ. Press	188	2015
	(Roozenbeek & Van Der Linden, 2019)	Scholar/EBSCOhost/Scopus	Journal of Risk Research	Journal (Business/ Eng/SS)	Taylor & Francis	277	2019
	(Roozenbeek & van der Linden, 2019)	Google Scholar/ Scopus	Palgrave Communications	Journal (Economy, SS, Psychology)	Palgrave	296	2019
	(Basol et al., 2020)	Google Scholar/ Scopus	Journal of Cognition	Journal	Ubiquity Press	146	2020
	(Van der Linden et al., 2017)	Google Scholar	Global Challenges	Journal	Wiley Online Library	614	2017
	(Papageorgis & McGuire, 1961)	Google Scholar/ Scopus	The Journal of Abnormal and Social Psychology	Journal (Psychology)	American Psychological Association	233	1961

	(Lewandowsky & Van Der Linden, 2021)	Google Scholar/ Scopus	European Review of Social Psychology	Journal (Psychology/SS)	Taylor & Francis	113	2021
	(Ecker et al., 2017)	Google Scholar/ ScienceDirect	Journal of Applied Research in Memory and Cognition}	Journal (Psychology)	Elsevier	222	2017
	(Cook et al., 2014)	Google Scholar/ Scopus	Journal of Geoscience Education	Journal (SS/Geo)	Taylor & Francis	60	2014
	(Cook, 2019)	Google Scholar	Handbook of research on deception, fake news, and misinformation online	Book	IGI global	59	2019
	(De Paor & Heravi, 2020)	Google Scholar/ Scopus	The Journal of Academic Librarianship	Journal (SS)	Elsevier	58	2020
	(Delellis & Rubin, 2018)	Google Scholar/ Scopus	Proceedings of the Association for Information Science and Technology	Journal (SS)	Wiley	10	2018
	(Kowalski & Taylor, 2009)	Google Scholar/ Scopus	Teaching of Psychology	Journal (Psych/SS)	SAGE Publications	234	2009
	(Walton & Hepworth, 2011)	Google Scholar/ Scopus/ProQuest	Journal of Documentation	Journal (SS/ IS)	Emerald	124	2011
	(Jones-Jang et al., 2021)	Google Scholar/ Scopus	American Behavioral Scientist	Journal (Psych/SS)	SAGE Publications	272	2021
	(McCuin et al., 2014)	Google Scholar/ Scopus/ ScienceDirect	Journal of Geoscience Education	Journal (SS/ Education)	Taylor & Francis	48	2014
	(Mihailidis & Viotty, 2017)	Google Scholar/ Scopus	American Behavioral Scientist	Journal (Psych/SS)	SAGE Publications	439	2017

Combat (Detection)	*(Nasir et al., 2021)*	Google Scholar/ Scopus/ ScienceDirect	International Journal of Information Management Data Insights	Journal (IS, CS, SS)	Elsevier	126	2021
	(Sahoo & Gupta, 2021)	Google Scholar/ Scopus/ ScienceDirect	Applied Soft Computing	Journal (CS)	Elsevier	112	2021
	(Kwon et al., 2017)	Google Scholar/ Scopus	PLoS one	Journal	PLOS	322	2017
	(Kaliyar et al., 2020)	Google Scholar/ Scopus/ ScienceDirect	Cognitive Systems Research	Journal (CS, Psych., NeuroSci.)	Elsevier	140	2020
	(Ahmad et al., 2020)	Google Scholar/ Scopus	Complexity	Journal	Hindawi	124	2020
	(Atanasova et al., 2019)	Google Scholar/ Scopus/ACM	Journal of Data and Information Quality	Journal	ACM	38	2019
	(W. Chen et al., 2016)	Google Scholar/ Scopus/IEEE	IEEE Annual Information Technology, Electronics and Mobile Communication Conference	Conference	IEEE	27	2016
	(Shu, Wang, et al., 2019)	Google Scholar/ Scopus/ACM	ACM international conference on web search and data mining	Conference	ACM	373	2019
	(Wang & Terano, 2015)	Google Scholar/ Scopus/IEEE	IEEE Big Data	Conference	IEEE	67	2015
	(Castillo et al., 2011)	Google Scholar/ Scopus/ACM	WWW	Conference	ACM	2493	2011
	(Ma et al., 2015)	Google Scholar/ Scopus/ACM	ACM	Conference	ACM	496	2015

	(Ma et al., 2016)	Google Scholar/ Scopus	IJCAI International Joint Conference on Artificial Intelligence	Conference	AAAI Press	819	2016
	(Qian et al., 2018)	Google Scholar/ Scopus	IJCAI International JointConference on Artificial Intelligence	Conference	IJCAI	141	2018
	(Wu et al., 2015)	Google Scholar/ IEEE	IEEE 31st international conference on data engineering	Conference	IEEE	498	2015
	(Bian et al., 2020)	Google Scholar/ Scopus	AAAI Conference on Artificial Intelligence	Conference	PKP/OJS	175	2020
	(Ma et al., 2018)	Google Scholar/ Scopus	Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)	Conference	ACL (Association for Computational Linguistics)	337	2018
	(Y. Wang et al., 2018)	Google Scholar/ Scopus/ ACM	ACM SIGKDD International Conference on Knowledge Discovery & Data Mining}	Conference	ACL (Association for Computational Linguistics)	480	2018
	(Shao et al., 2016)	Google Scholar/ Scopus/ ACM	Proceedings of the 25th International Conference on World Wide Web	Conference	ACM	384	2016
	(F. Yang et al., 2012)	Google Scholar/ Scopus/ ACM	ACM SIGKDD	Workshop	ACM	557	2012
	(Vosoughi et al., 2017)	Google Scholar/ Scopus/ ACM	ACM transactions on knowledge discovery from data (TKDD)	Journal (CS)	ACM	185	2017

	(S. Kumar et al., 2016)	Google Scholar/ Scopus/ ACM	WWW	Conference	ACM	304	2016
	(Z. Jin et al., 2016)	Google Scholar/ Scopus	AAAI Conference on Artificial Intelligence	Conference	PKP/OJS	356	2016
	(J. Kim et al., 2018)	Google Scholar/ Scopus/ ACM	ACM conference on web search and data mining	Conference	ACM	197	2018
	(Papanastasiou, 2020)	Google Scholar/ Scopus	Management Science	IS Journal	INFORMS	81	2020
	(Shu, Dumais, Awadallah, & Liu, 2020)	Google Scholar/ Scopus/IEEE	IEEE Intelligent Systems	Journal	IEEE	14	2020
	(J. F. George et al., 2018)	Google Scholar/ EBSCOhost	MIS Quarterly	IS Journal	MISQ	43	2018
	(Konstantinovskiy et al., 2021)	Google Scholar/ Scopus/ ACM	Digital Threats: Research and Practice	Journal	ACM New York	76	2021
	(Hassan et al., 2017)	Google Scholar/ ACM	Proceedings of the VLDB Endowment	Conference	VLDB Endowment	203	2017
	(Gencheva et al., 2017)	Google Scholar/ Scopus	RANLP 2017	Conference	ACL (Association for Computational Linguistics)	77	2017
	(Ratkiewicz et al., 2011)	Google Scholar/ Scopus/ACM	*Proceedings of the 20th International Conference Companion on World Wide Web* (WWW)	Conference	ACM	519	2011
	(Ciampaglia et al., 2015)	Google Scholar/ Scopus	PLOS one	Journal	PLOS	494	2015

	(Shi & Weninger, 2016)	Google Scholar/ Scopus/ACM	Proceedings of the 25th International Conference Companion on World Wide Web (WWW)	Conference	ACM	80	2016
	(Yuan et al., 2021)	Google Scholar/ Scopus /ScienceDirect	DSS (Decision Support Systems)	IS Journal	Elsevier	14	2021
	(K. K. Kumar & Geethakumari, 2014)	Google Scholar/ Scopus	Human-centric Computing and Information Sciences	Journal (CS)	SpringerOpen	240	2014
	(Gupta, Lamba, Kumaraguru, et al., 2013)	Google Scholar/ Scopus/ACM	Proceedings of the 22nd International Conference on World Wide Web (WWW)	Conference	ACM	683	2013
	(Lim, 2020)	Google Scholar/ Scopus /ScienceDirect	Journal of Academic Librarianship	Journal (SS/ Education)	Elsevier	22	2020
	(Biyani et al., 2016)	Google Scholar	AAAI conference on artificial intelligence	Conference		167	2016
Combat (Mitigation/Remedy)	*(King et al., 2021)*	Google Scholar/ Scopus /EBSCOhost	Journal of Management Information Systems (JMIS)	Journal (IS)	Taylor & Francis	1	2021
	(Tripathy et al., 2010)	Google Scholar/ Scopus/ACM	ACM international conference on Information and knowledge management	Conference	ACM	157	2010
	(N. P. Nguyen et al., 2012)	Google Scholar/ Scopus/ACM	ACM Web Science Conference	Conference	N/A	254	2012

	(Budak et al., 2011)	Google Scholar/ Scopus/ACM	Proceedings of the 20th International Conference on World Wide Web	Conference	N/A	895	2011
	(L. Yang et al., 2020)	Google Scholar/ Scopus /ScienceDirect	Information Sciences	Journal (IS)	Elsevier	90	2020
	(Tong et al., 2017)	Google Scholar/ Scopus	IEEE Transactions on Network Science and Engineering	Journal (CS)	IEEE	112	2017
	(H. Zhang et al., 2015)	Google Scholar/ Scopus	International Conference on Computational Social Networks	Conference	Springer	55	2015
	(He et al., 2015)	Google Scholar/ Scopus/IEEE	IEEE 35Th international conference on distributed computing systems	Conference	IEEE	115	2015
	(Tambuscio et al., 2015)	Google Scholar/ Scopus/ACM	Proceedings of the 20th International Conference on World Wide Web	Conference	ACM	164	2015
	(Figl et al., 2019)	Google Scholar/ Scopus	International Conference on Information Systems (ICIS)	Conference (IS)	Association for Information Systems (AIS)	9	2019
	(Fan et al., 2013)	Google Scholar/ Scopus/IEEE	IEEE 33Th international conference on distributed computing systems	Conference	IEEE	129	2013
	(Wang et al., 2017)	Google Scholar/ Scopus/IEEE	IEEE Transactions on Knowledge and Data Engineering	Journal (CS, IS)	IEEE	136	2017

	(Kimura et al., 2009)	Google Scholar/ Scopus/ACM	ACM Transactions on Knowledge Discovery from Data (TKDD)	Journal (CS)	ACM	235	2009
	(Lin et al., 2019)	Google Scholar/ Scopus/IEEE	IEEE Transactions on Vehicular Technology	Journal	IEEE	45	2019
	(Shrivastava et al., 2020)	Google Scholar/ Scopus/IEEE	IEEE Transactions on Computational Social Systems	Journal	IEEE	51	2020
	(Amoruso et al., 2020)	Google Scholar/ Scopus/ACM	Journal of Artificial Intelligence Research	CS Journal	AI Access Foundation	47	2020
	(Nyhan & Reifler, 2010)	Google Scholar/ Scopus/JSTOR	Political Behavior	Journal (Politics)	Springer	2755	2010
	(Farajtabar et al., 2017)	Google Scholar/ Scopus	International Conference on Machine Learning	Conference	N/A	163	2017
	(Gimpel et al., 2021)	Google Scholar/ Scopus	JMIS (Journal of Managament Information Systems)	IS Journal	Taylor & Francis	23	2021
	(Garrett & Poulsen, 2019)	Google Scholar/ Scopus /EBSCOhost	Journal of Computer-Mediated Communication	Journal	Oxford University Press	37	2019
	(A. Kim & Dennis, 2019)	Google Scholar/ Scopus /EBSCOhost	MIS Quarterly	IS Journal	MISQ	186	2019
	(A. Kim et al., 2019a)	Google Scholar/ Scopus	JMIS (Journal of Management Information Systems)	IS Journal	Taylor & Francis	164	2019
	(P. L. Moravec et al., 2019)	Google Scholar/ Scopus	MIS Quarterly	IS Journal	MISQ	173	2019

	(P. L. Moravec et al., 2022)	Google Scholar/ Scopus/INFORMS	Information Systems Research	IS Journal	INFORMS	6	2022
	(P. Moravec et al., 2020)	Google Scholar/ Scopus/INFORMS	Information Systems Research	IS Journal	INFORMS	50	2020
	(Pennycook, Bear, et al., 2020)	Google Scholar/ Scopus/INFORMS	Management Science	IS Journal	INFORMS	338	2020
	(Pennycook & Rand, 2019)	Google Scholar/ Scopus/JSTOR	Proceedings of the National Academy of Sciences	Journal (Multidisciplinary)	National Acad Sciences	448	2019
	(Pennycook, McPhetres, et al., 2020)	Google Scholar/ Scopus	Psychological science	Journal	Sage Publications	1143	2020