Doctoral Dissertation

Towards Keeping up with Fake News in the Social Media Ecosystem

Taichi Murayama

March 16, 2022

Graduate School of Science and Technology
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Taichi Murayama

Thesis Committee:

Professor Eiji Aramaki	(Supervisor)
Professor Satoshi Nakamura	(Co-supervisor)
Visiting Professor Junichiro Yoshimoto	(Co-supervisor)
Associate Professor Shoko Wakamiya	(Co-supervisor)
Dr. Takeshi Sakaki	(Hotto Link Inc.)

Towards Keeping up with Fake News in the Social Media Ecosystem*

Taichi Murayama

Abstract

Fake news has caused significant damage to various fields of society, such as the economy, politics, disasters, and social events. In this dissertation, we address the challenges related to fake news, e.g., such as understanding how fake news is spread on social media, the construction of a non-English dataset for fake news research, etc. We try to tackle issues related to fake news by aiming to keep up with them well together in the social media ecosystem in the following three steps. First, we try to understand how fake news is spread in social media in Chapter 2. We propose the novel modeling method by utilizing the point process to describe fake news spreading on Twitter. Chapter 3 examines whether fake news detection from social media based on our observations and understanding of the spreading of fake news gained from Chapter 2. Specifically, we verify whether temporal information in the spread of fake news is effective in the detection. Then, we investigate the useful attempts to counter fake news in real society, especially in Japanese society, by constructing the Japanese fake news dataset and fake news collection system, which bridge the gap between research findings and the application in Chapter 4. Finally, in Chapter 5, we summarize our research and discuss issues that we should address in the future. This dissertation is the link between the fundamental understanding of fake news and the practical application to the actual society.

Keywords: fake news, fact-checking, computational social science, dataset, time-series modeling

*Doctoral Dissertation, Graduate School of Science and Technology,
Nara Institute of Science and Technology, March 16, 2022.

Contents

List of Figures

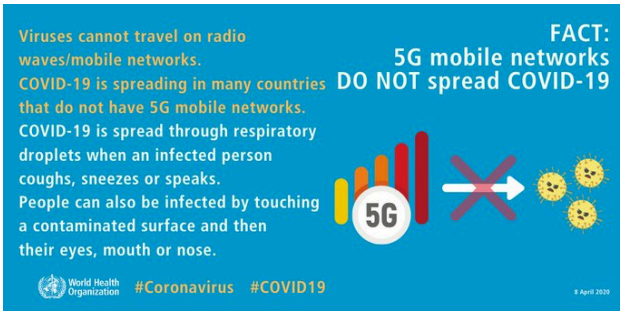
1. Introduction

1.1. Background

Fake news has caused significant damage to various fields of society. For example, in the stock market, a false report of the United Airlines parent company's bankruptcy in 2008 caused the company's stock price to drop by 76% in a few minutes; it closed 11% lower than the previous day's closing, with a negative effect persisting for more than six days [3]. During the 2016 U.S. presidential election, 529 different low-credibility statements [4] were spread on Twitter, and 25% of the news outlets linked to tweets, either fake or extremely biased, supporting Trump or Clinton, have potentially influenced the election [5]. It may cause significant effects on real events, for example 'Pizzagate" on Reddit led to shooting [6]. In addition to stock markets and political events, the situation is the same for public health. Fake news on infectious diseases, such as Ebola [7], yellow fever [8], and Zika [9], appears to be spreading on the internet. In particular, since 2020, COVID-19 pandemic has fueled the spread of news from unreliable sources [10, 11]; the number of English-language fact-checks increases by more than 900% from January to March 2020 [12] (Figure **??**). Fake news has a great impact on other events in various countries, such as Brexit in Europe [13, 14], salt panic in China [15], deadly violence between the two groups in Ethiopia [16], and natural disasters such as the Great East Japan Earthquake in 2011 [17, 18], and the Chile earthquake in 2010 [19].

However, fake news is not a new phenomenon, it has become a bigger problem owing to social media. It is easy to share impressive news with many people with the rise of social media such as Facebook, Twitter, and Weibo compared to traditional news media such as newspapers and television [20]. The Pew Research Center reports that half of U.S. adults mainly receive news from social media [21].

(a) Explosions in the White House and Barack Obama were injured, according to a hacked Twitter account named Associated Press.



(b) WHO combats the spread of misinformation through a series of communication campaigns. It debunks fake news, such as 5G mobile networks spreads COVID-19 through Facebook and other social media.

Figure 1.1.: Examples of fake news: (a) On April 23, 2013, a hacked Twitter account named Associated Press posted fake news claiming two explosions occurred in the White House and Barack Obama was injured. Although the White House and Associated Press assured the public minutes later, the report was not true; the fast diffusion to millions of users had caused severe social panic, resulting in a loss of $136.5 billion in the stock market [1]. (b) In the first three months of 2020, nearly 6,000 people worldwide were hospitalized because of coronavirus misinformation [2]. During this period, approximately 800 people may have died because of misinformation related to COVID-19. To try to control the COVID-19 infodemic, WHO has teamed up with the governments of various nations to create and distribute content to combat the spread of misinformation through a series of communication campaigns.

Social media provides an ideal environment for communication and information acquisition as well as encourages users to share information without distance barriers among individuals. In addition, the existence of an echo chamber effect becomes the mechanism for accelerating fake news dissemination [19, 22, 23]. In particular, younger and older people often believe fake news, which is an incentive

to create, publish, and spread fake news for substantial potential political and economic benefits [24]. The prevalence of fake news on social media has the potential to damage the trustworthiness of online journalism and can cause widespread panic and pose a major problem in the aforementioned examples. Therefore, it has become increasingly important for policymakers and social media operators to detect, prevent, and suppress the creation of fake news and encourage users to protect themselves from fake news.

However, countering fake news in the real world is difficult because we still do not completely understand the diffusion styles of fake news and fake news with diversity in terms of topics and media platforms attempt to distort the truth with diverse linguistic styles. We address the challenges related to fake news in the following steps for aiming to keep up with fake news in the social media ecosystem. First, we try to understand how fake news spreads on social media through the modeling method, particularly on Twitter (Chapter 2). Then, based on our observations and understanding of the spread of fake news, we examine whether it is possible to detect fake news on social media (Chapter 3). Finally, we examine useful measures to counter fake news in real society, particularly in Japan, by bridging the gap between research findings and applications (Chapter 4). Our study is the link between the fundamental understanding of fake news and its practical application in real life. An overview is shown in Figure 1.2. Before providing a summary of our work in Section 1.4, we describe the definition of "fake news" in Section 1.2 and related concepts of fake news in Section 1.3.

1.2. Definition of Fake News

Claire Wardle, the co-founder and leader of the First Draft [25], announced that the term fake news is woefully inadequate to describe the issues, and distinguished between three types of information content problems: misinformation, disinformation, and malinformation [26]. Thus, it is not easy to construct a general definition of "fake news" for the current diverse circumstances.

"Fake news is false news" is a broad definition of fake news [27]. Similarly, D.M. Lazer et al. [28] described "fake news is fabricated information that mimics news media content in form but not in organizational process or intent." This

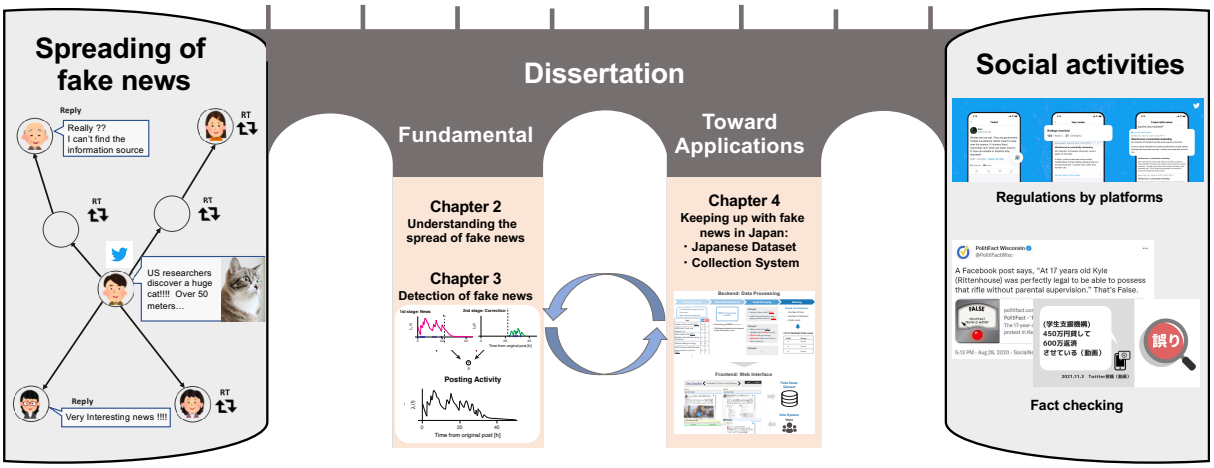


Figure 1.2.: Overview of this dissertation. This dissertation aims to bridge the gap between fake news events and social activities such as fact-checking by explaining how to address the fake news problem in the social media ecosystem in three steps: fake news modeling, fake news detection, and challenges with social implementation. The first two attempts to understand fake news. Based on the findings from the former, the latter is an attempt to solve the fake news problem in the real world, particularly in Japan.

broad definition emphasizes only information authenticity and does not consider information intention. This allows us to cover different types of fake news based on their purpose or intent, such as satire and parody [29]. There are few studies [30–32] leveraging the definition.

Most research emphasizes "intention" in the definition of fake news as a narrow definition. H. Allcott and K. Shu et al. [20,33] defined fake news as "a news article that is **intentionally** and verifiably false." X. Zhang et al. [34] described "fake news refers to all kinds of false stories or news that are mainly published and distributed on the Internet, in order to **purposely** mislead, be fool or lure readers for financial, political or other gains." Other studies [35–37] have also emphasized intention in the definition of fake news. However, we consider that most studies related to fake news do not completely follow this definition. Researchers use fake news datasets labeled as fake or not based on the judgement of fact-checking sites, most of which do not consider the intention of the creator, because it is difficult to

Concept	Authenticity	Intention
Fake news (broad) [27]	Non-factual	Undefined
Fake news (narrow) [20, 33]	Non-factual	Mislead
Misinformation	Non-factual	Undefined
Disinformation	Non-factual	Mislead
Rumor	Undefined	Undefined
Hoaxes & Satire	Non-factual	Entertain

Table 1.1.: Comparison between concepts related to fake news

determine whether each fake news item was created with the dishonest intention of misleading readers. We refer the reader to [38] for a more detailed discussion of the range of meanings. A summary of fake news and related concepts is provided in Table 1.1.

Therefore, the definition of the phrase "fake news" is ambiguous, and there is some criticism of this ambiguity. For example, the British government decided that the phrase "fake news" would no longer be used in official documents because it is a poorly defined and misleading term that conflates a variety of false information [39]. Claire Wardle, the co-founder and leader of the First Draft, announced that the phrase "fake news" is woefully inadequate to describe related issues and distinguishes among three types of information content problems: misinformation, disinformation, and malinformation [26]. Disinformation is related to the intention of users to create and share content, whereas malinformation is associated with the harmfulness of the information to society.

1.3. Related Concepts of Fake News

In the context of conveying false information, the term "fake news" is closely related to several concepts: satire, rumor, clickbait, and so on. There are salient differences among them in terms of degrees of contexts of usage and functions of serving different propagation purposes.

- **Misinformation** is false information that is inaccurate or misleading in

a macro aspect [28, 40]. It spreads unintentionally because of honest mistakes [41] or knowledge updated without the purpose of misleading.

- **Disinformation** is false information that misleads others intentionally for some purpose (e.g., to deceive people [42], to promote biased agenda [43]), which is close to a narrow definition of fake news. In contrast to misinformation, it spreads because of the deliberate attempt to deceive or mislead [41]. The survey by B. Guo et al. [44] divided false news into two categories: misinformation and disinformation. The term "deception" is sometimes considered similar to disinformation [27].

- **Rumor** is unverified and relevant information being circulated and it could later be confirmed as true, false, or left unconfirmed [45,46]. It could spread from one user to another. Before the term "fake news" became popular, the classification task of determining whether news is true or not was called "rumor detection," and there had been a lot of research on it. Studies on rumor are surveyed in [47].

- **Hoaxes** is deliberately fabricated information made to masquerade as truth [42]. It often causes serious material damage to victims because it includes relatively complex and large-scale fabrications [29, 48].

- **Satire**, which contains a lot of irony and humor, is written with the purpose of entertaining or criticizing the readers [49]. These news articles are frequently published on some sites, such as SatireWire.com [50] and The Onion [51]. It could be harmful if satire news, ignoring context, is shared by many people. "Parody" is a similar concept to satire, but with the difference that parody uses non-factual information to inject humor [38].

- **Hyperpartisan** news is extremely one-sided or biased news in the political context [37]. Biased does not imply fake; however, some papers [52, 53] report that it has a high potential for being false in parts of hyperpartisan news and that false information spreads widely in the alt-right community, such as 4chan's /pol/ board [54] and Gab [55].

- **Propaganda** is a form of persuasion that attempts to influence the emotions, attitudes, opinions, and actions of specified target audiences for polit-

ical, ideological, and religious purposes through the controlled transmission of one-sided messages [56, 57]. Recently, it has often been used to influence election results and opinions in a political context.

- **Spam** is fabricated information that ranges from self-promotions to false announcements of the products. On review sites, spam provides positive reviews to unfairly promote them or unjustified negative reviews to competing products to damage their reputations [58]. In the social ecosystem, spam targets users to disseminate malware and commercial spam messages and promote affiliate websites [59].

- **Clickbait** is a story with eye-catching headlines that is intended to attract traffic and benefit from advertising revenue [43,60]. Because the discrepancy between content and headline is a main component of clickbait, it is one of the least severe types of false information.

The list of these terms is based on [27, 44, 61] and is extended to build upon the existing literature. Similar to the term "fake news", there are no formal definitions for these terms. In addition, these terms should not be treated as exhaustive representations of the false information ecosystem (e.g., half-truth [62] and factoid [63]).

1.4. Outline

This dissertation aims to bridge the gap between research and real life application by explaining how to address the fake news problem in the social media ecosystem based on three steps: fake news modeling, fake news detection, and struggles with social implementation. The remainder of this dissertation is organized as follows: In Chapter 2, we propose an effective modeling method for fake news diffusion on Twitter. First, we explain the modeling method using a point process, assuming that fake news bursts twice. We then discuss the experiment results to evaluate the effectiveness of our model in predicting the number of posts related to fake news in the future. In Section 3, we propose a fake news detection model based on the insights gained in Section 2. The model utilizes temporal feature of posts on social media that have not been utilized before and achieves high accuracy

compared to other models. In Chapter 4, we seek solutions to the fake news problem in the real world, particularly in Japan. We first introduce the Japanese fake news dataset to promote fake news research in Japan, and then introduce a fake news collection system from Twitter to help with fact-checking. Following the dataset expansion, our system may utilize fake news characteristics and the fake news detection system described in Chapters 2 and 3. In Chapter 5, we conclude the dissertation and suggest future research directions.

2. Modeling the Spread of Fake News on Twitter

2.1. Background

In this chapter, we investigate how fake news spreads on Twitter. This subject is relevant to an important research question in social science: how does unreliable information or rumors spread in society? It also has practical implications for detecting and mitigating fake news [31,33]. Previous studies have mainly focused on the path that fake news items travel as they spread on social networks [30,64], which clarified the structural aspects of the spread. However, little is known about the temporal and dynamic aspects of how fake news spreads online.

Here, we focus on Twitter and assume that fake news spreads through a two-stage process. In the first stage, a fake news item spreads as an ordinary news story. The second stage occurs after a correction time, when most users realize the falsity of the news story. Then, the information regarding that falsehood spreads as another news story. We formulate this assumption by extending the time-dependent Hawkes process (TiDeH) [65], which is a state-of-the-art point process model for predicting re-sharing dynamics on Twitter. To validate the proposed model, we compiled two datasets of fake news items on Twitter.

Our research is similar to previous studies on predicting the future popularity of online content [66, 67]. A standard approach for predicting popularity is to apply a machine learning framework such that the prediction problem can be formulated as a classification [68,69] or regression [70] task. Another approach to the prediction problem is to develop a temporal model and fit the model parameters by using a training dataset. This approach consists of two types of models: time-series and point process models. The time-series model describes the num-

ber of posts in a fixed window. For example, Matsubara et al. [71] proposed SpikeM to reproduce temporal activities in blogs, Google Trends, and Twitter. In addition, Proskurnia et al. [72] proposed a time-series model that considers a promotion effect (e.g., promotion through social media and the front page of the petition site) to predict the popularity dynamics of an online petition. A point process model describes posted times in a probabilistic manner by incorporating the self-exciting nature of information spreading [73, 74]. Point process models have also driven theoretical studies on the effect of network structure and event times on diffusion dynamics [75]. Various point process models have been proposed for predicting the final number of re-shares [74, 76] and their temporal patterns [65] on social media. Furthermore, these models have been applied to interpret endogenous and exogenous shocks to activities on YouTube [77] and Twitter [78]. To the best of our knowledge, the proposed model is the first to incorporate a two-stage process, which is an essential characteristic of the spread of fake news. Although some studies [79] proposed a model for the spread of fake news, they focused on modeling the qualitative aspects and did not evaluate the prediction performance using a real dataset.

Our research is also useful for fake news detection study. Numerous attempts have been made to automatically detect fake news and rumors [31, 33]. Typically, fake news is detected based on the textual content. Hassan et al. [80] extracted multiple categories of features from sentences and applied a support vector machine classifier to detect fake news. Rashkin et al. [81] developed a long short-term memory (LSTM) neural network model for fact-checking news. The temporal information of a cascade, for example, the timing of posts and re-shares triggered by a news story, may improve fake news detection performance. Kwon et al. [82] showed that temporal information improves rumor classification performance. Temporal information has also been shown to improve fake news detection performance [83], rumor stance classification [84], misinformation source identification [85], and detection of fake retweeting accounts [86]. A deep neural network model [83] can also incorporate temporal information to improve the fake news detection performance. However, a limitation of the neural network model is that it can utilize only part of the temporal information and cannot handle cascades with many user responses. The proposed model parameters can be used

as a compact representation of the temporal information, which helps overcome this limitation.

Our contribution in this chapter is summarized as follows:

- We propose a simple point process model based on the assumption that fake news spreads as a two-stage process.
- We evaluate the predictive performance of the proposed model, which demonstrates its effectiveness.
- We conduct a text mining analysis to validate the assumptions of the proposed model.

2.2. Method: Modeling the Information Spread of Fake News

We developed a point process model to describe the dynamics of the spread of a fake news item. A schematic of the proposed model is shown in Figure 4.11. The proposed model is based on the following two assumptions.

- Users do not know the falsity of a news item in the early stage. The fake news spreads as an ordinary news story (Figure 4.11: 1st stage).
- Users recognize the falsity of the news item around a correction time t_c. The information that the original news is fake spreads as another news story (Figure 4.11: 2nd stage).

In other words, the proposed model assumes that the spread of a fake news item consists of two cascades: 1) a cascade of the original news story and 2) a cascade asserting the falsity of the news story. In this study, we used the term *cascade*, meaning tweets or retweets triggered by a piece of information. To describe each cascade, we use the time-dependent Hawkes process model, which properly considers the circadian nature of users and the aging of information.

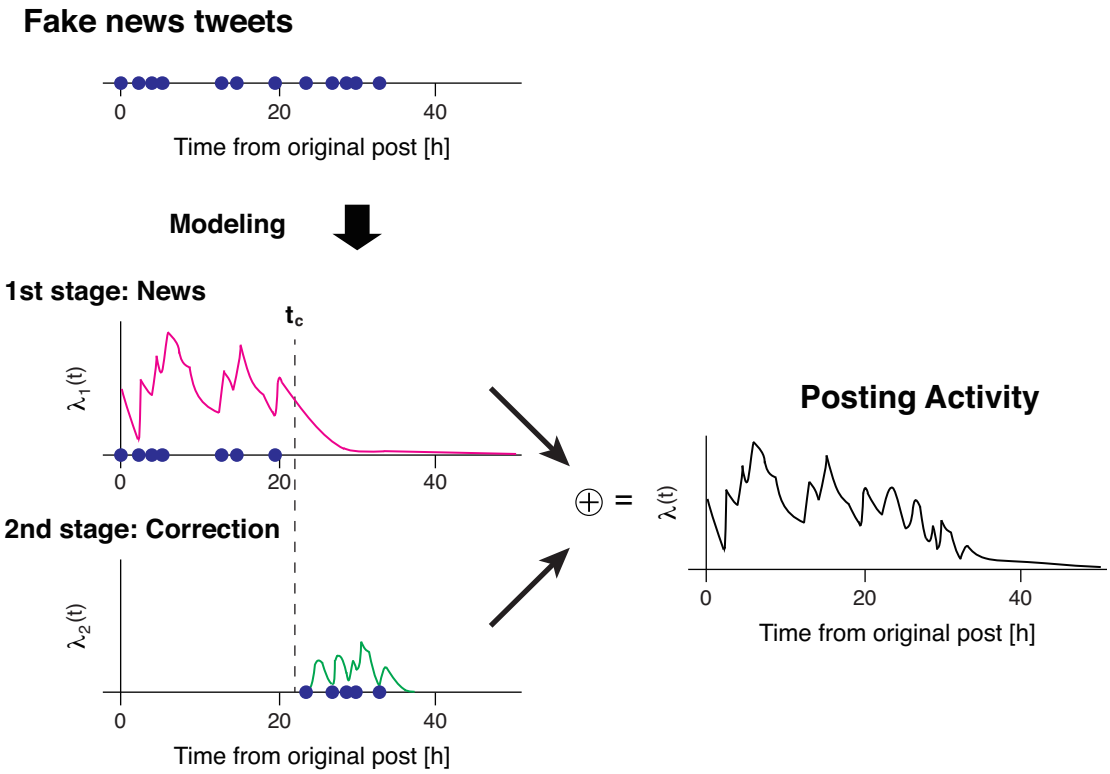


Figure 2.1.: Schematic of the proposed model. We propose a model that describes how posts or re-shares related to fake news items spread on social media (fake news tweets). Blue circles represent the time stamps of the tweets. The proposed model assumes that information spread is described as a two-stage process. Initially, a fake news item spreads as a novel news story (1st stage). Following a correction time t_c, Twitter users recognize the falsity of a news item. Then, the information that the original news item was false spreads as another news story (2nd stage). The posting activity related to fake news $\lambda(t)$ (right: black) is specified by the summation of the activities of the two stages (left: magenta and green).

2.2.1. Time-Dependent Hawkes process (TiDeH): Model of a Single Cascade

We describe a point-process model of a single cascade as information spreading triggered by a news story on social media. The time-dependent Hawkes process (TiDeH) [65] for modeling a single cascade is designed to represent two characteristics: the rhythm of daily activities and the decay of people's interests. The model is based on point process models [87], where the probability of obtaining a post or reshare in a small time interval $[t, t + \Delta t]$ is written as $\lambda(t)\Delta t$, where $\lambda(t)$ is the instantaneous rate of the cascade, that is, the intensity function. The intensity function of TiDeH model depends on the previous posts in the following manner:

$$\lambda_{\mathrm{TiDeH}}(t) = p(t)h(t), \tag{2.1}$$

and the memory function $h(t)$ is defined as follows:

$$h(t) = \sum_{i:t_i<t} d_i \phi(t - t_i), \tag{2.2}$$

where $p(t)$ is the infection rate, t_i is the time of the ith post, and d_i is the number of followers of the ith post. The infection rate $p(t)$ incorporates two main properties in the cascade: the circadian rhythm and decay owing to the aging of the information as follows:

$$p(t) = a\left\{1 - r\sin\left(\frac{2\pi}{T_m}(t + \theta_0)\right)\right\}\mathrm{e}^{-(t-t_0)/\tau},$$

where the time of the original post is assumed to be $t_0 = 0$, and $T_m = 24$ hours is the period of oscillation. Each parameter plays a role as follows: a represents the intensity of the news item on the spreading, r represents the relative amplitude of daily activity, θ_0 represents the phase of the oscillation in daily activity, and τ represents the time constant of decay to describe the behavior of people losing interest. Memory kernel $\phi(t)$ represents the probability distribution of the reaction time of a follower. A heavy-tailed distribution was adopted for the memory kernel [65, 74]

$$\phi(s) = \begin{cases} c_0 & (0 \leqq s \leqq s_0) \\ c_0(s/s_0)^{-(1+\gamma)} & (\text{Otherwise}) \end{cases}$$

The parameters were set as $c_0 = 6.94 \times 10^{-4}$ (/seconds), $s_0 = 300$ seconds, and $\gamma = 0.242$.

2.2.2. Proposed Model of the Spread of Fake News

We formulate a point process model for the spread of a fake new item. We assume that the spread consists of two cascades: one owing to the original news item and the other owing to the correction of the news item. The activity of the fake news cascade can be expressed as the sum of the two cascades using TiDeH. We modeled the cascade from 0 to t_c with the first term and the cascade from t_c to t_max with the second term.

$$\lambda_{\mathrm{prop}}(t) = p_1(t)h_1(t) + p_2(t)h_2(t). \tag{2.3}$$

The first term $p_1(t)h_1(t)$ represents the cascade rate caused by the original news item.

$$p_1(t) = a_1 \left\{1 + r \sin\left(\frac{2\pi}{T_m}(t + \theta_0)\right)\right\} \mathrm{e}^{-t/\tau_1},\ h_1(t) = \sum_{i:t_i<\min(t,t_c)} d_i \phi(t - t_i), \tag{2.4}$$

where a_1 represents the impact of the original news item on the spreading, τ_1 is the decay time constant, $\min(t, t_c)$ represents the smaller of the two values (t or t_c), and t_c is the correction point of the fake news item.

The second term $p_2(t)h_2(t)$ represents the cascade induced by the correction.

$$p_2(t) = a_2 \left\{1 + r \sin\left(\frac{2\pi}{T_m}(t + \theta_0)\right)\right\} \mathrm{e}^{-(t-t_c)/\tau_2},\ h_2(t) = \sum_{i:t_c<t_i<t} d_i \phi(t - t_i), \tag{2.5}$$

where a_2 represents the impact of the falsity of the news on the spreading and τ_2 is the decay time constant. We assumed that the circadian parameters of $p_2(t)$ are identical to those of $p_1(t)$. Mathematically, the proposed model includes TiDeH as a special case. We consider the proposed model that satisfies the following conditions

$$\tilde{a} = a_1 = a_2 e^{-t_c/\tilde{\tau}}, \quad \tilde{\tau} = \tau_1 = \tau_2. \tag{2.6}$$

The proposed model is equivalent to TiDeH (with parameters $a = \tilde{a}$ and $\tau = \tilde{\tau}$) by substituting Eq. 2.6 into Eqs 2.3, 2.4, and 2.5.

2.3. Parameter Fitting

Here, we describe the procedure for fitting the parameters from the event time series (e.g., tweeted times). Seven parameters, $\{a_1, \tau_1; a_2, \tau_2; r, \theta_0; t_c\}$, were determined by maximizing the log-likelihood function:

$$l = \sum_i \log \lambda(t_i) - \int_0^{T_{\rm obs}} \lambda(s) ds, \tag{2.7}$$

where t_i is the ith tweeted time and $\lambda(t)$ is the intensity obtained by Eq. 2.3, and $T_{\rm obs}$ is the observation time. We first fixed the correction point t_c, and the other parameters were optimized using the Newton method [88], provided by Scipy [89], within a range of $12 < \tau_1, \tau_2 < 2T_{\rm obs}$ (hours). The correction point is separately optimized using Brent's method [87] within the range of $0.1T_{\rm obs} < t_c < 0.9T_{\rm obs}$.

We validated the fitting procedure by applying synthetic data generated by the proposed model (Eq. 2.3). Figure 2.2 shows the dependence of the estimation accuracy on the observation time $T_{\rm obs}$. To evaluate accuracy, we calculated the median and interquartile ranges of the estimates from 100 trials. The estimation error decreases as the observation time increases. This result suggests that this fitting procedure can reliably estimate the parameters for sufficiently long observations (≥ 36 hours). The medians of the absolute relative errors obtained from 36 hours of synthetic data were 18%, 11%, 38%, 38%, and 10% for a_1, τ_1, a_2, τ_2, and t_c, respectively. The estimation accuracy of the second cascade parameters (a_2, τ_2) was worse than that of the first cascade parameters (a_1, τ_1). This seems to be caused by insufficient observed data. The first cascade parameters were estimated from the entire dataset, whereas the second cascade parameters were estimated from the observation data after correction point t_c. Moreover, the model parameters are not identifiable [90, 91] when $a_1 = a_2 e^{-t_c/\tau_2}$ and $\tau_1 = \tau_2$. Because the proposed model is equivalent to TiDeH ($a_2 = 0$, $t_c \geq T_{\rm obs}$) in this case, other parameter sets can also reproduce the observed data. Figure 2.3 shows that the fitting procedure can estimate the parameters accurately except for the non-identifiable domain.

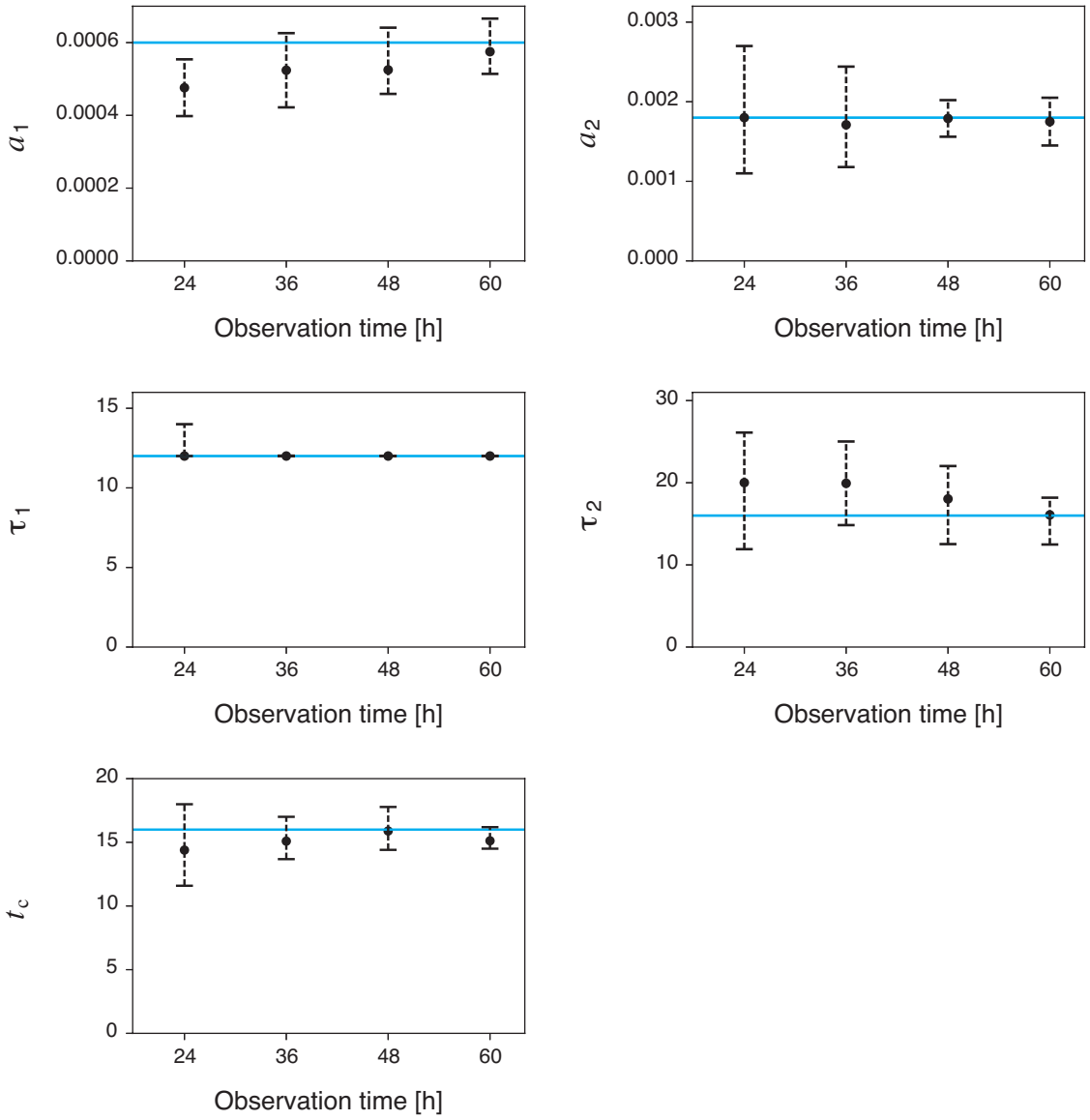


Figure 2.2.: Dependence of the estimation accuracy of parameters $\{a_1, \tau_1; a_2, \tau_2; t_c\}$ on observation time. Black circles and error bars represent the median and interquartile range of estimates obtained from 100 synthetic data. Cyan lines indicate the true value: $a_1 = 0.0006$, $a_2 = 0.0018$, $\tau_1 = 12$, $\tau_2 = 16$, and $t_c = 16$.

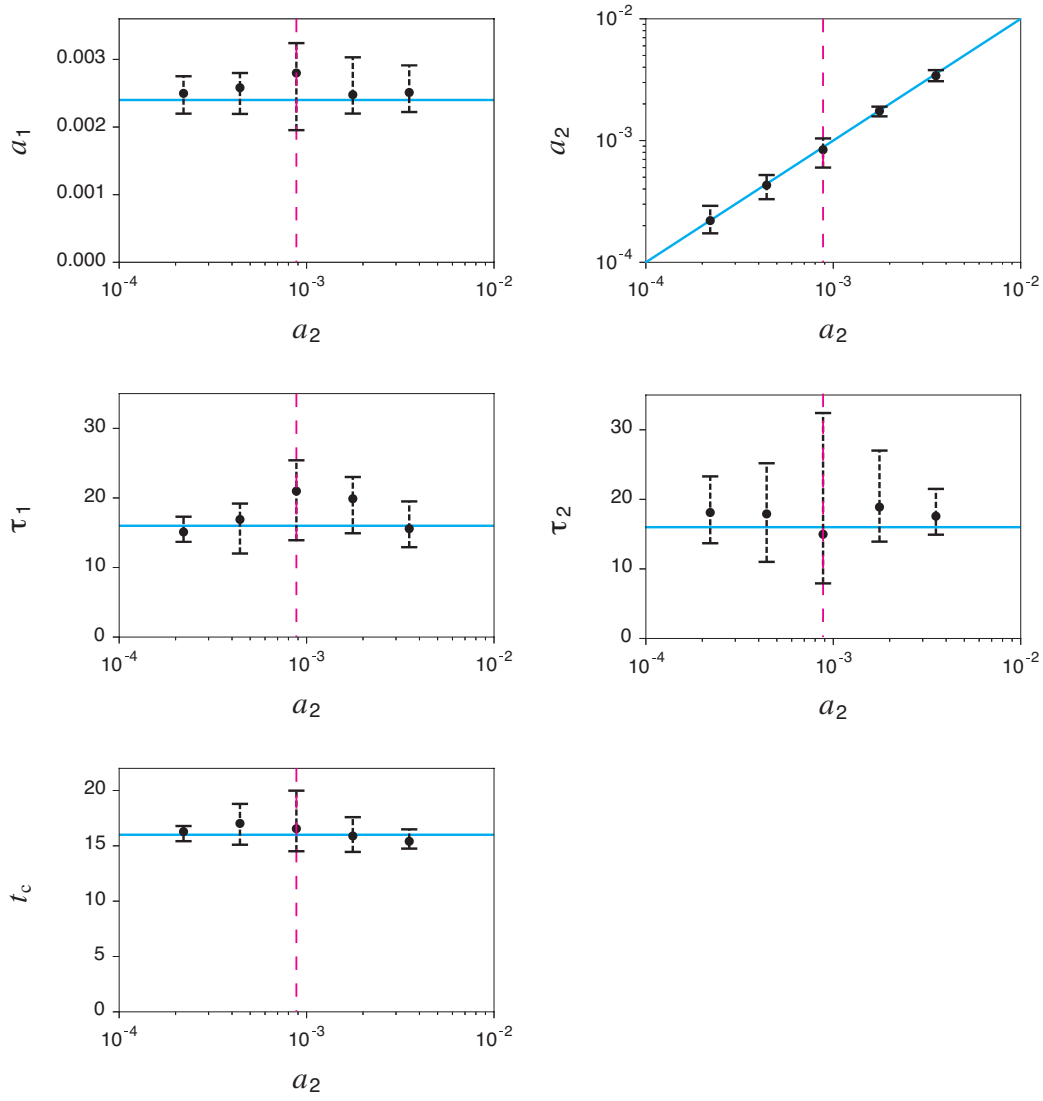


Figure 2.3.: Estimation accuracy of parameters around the non-identifiable domain. Black circles and error bars represent the median and interquartile range of estimates obtained from 100 synthetic data. Dashed magenta lines represent the non-identifiable domain satisfying $a_2 = a_1 e^{-t_c/\tau_2}$. Cyan lines indicate the true value: $a_1 = 0.0024$, $\tau_1 = \tau_2 = 16$, $t_c = 16$, and a_2 is changed between 2.2×10^{-4} and 3.5×10^{-3}.

2.4. Dataset

We evaluated the proposed model and examined the correction point for fake news based on two datasets of the spread of fake news items. Datasets of the spread of fake news based on retweets of the original news post [92,93] are publicly available. However, rather than simply retweeting, information-sharing of fake news can be complex. To cover information spread in detail, we manually compiled two datasets of fake news items spread on Twitter. In our dataset, 61% and 20% of the tweets are retweets of original posts in the Recent Fake News dataset and the 2011 Tohoku Earthquake and Tsunami dataset, respectively.

Table 2.1.: Recent Fake News (RFN): Details of 6 US fake news items

News No.	Title	Date	No. Posts	$T_{\max}$
a. Abolish	America came along as the first country to end (slavery) within 150 years.[a]	2019-03-21	1159	36
b. Notredame	A video clip from the Notre Dame cathedral fire shows a man walking alone in a tower of the church "dressed in Muslim garb."	2019-04-16	1641	132
c. Islamic	Did Ilhan Omar hold 'Secret Fundraisers' with 'Islamic Groups Tied to Terror'?	2019-03-27	10811	130
d. Lionhunter	Was a trophy hunter eaten alive by lions after he killed 3 baboon families?	2019-03-25	25071	88
e. Newzealand	Did New Zealand take Fox News or Sky News off the air in response to mosque shooting coverage?	2019-03-25	11711	88
f. Sonictrans	Will the animated character of Sonic the Hedgehog be transgender in a new film?	2019-05-06	2319	132

[a]Verbatim quote from Katie Pavlich on Politifact.com, March 19, 2019.

Table 2.2.: 2011 Tohoku Earthquake and Tsunami (Tohoku): Details of 19 Japanese fake news items

News No.	Title	Date	No. Posts	$T_{\max}$
a. Saveenergy	Large-scale power saving required even in the Kansai region.	2011-03-12	2846	174
b. EscapeTokyo	The bureaucracy in the Ministry of Defense says "You should escape from Tokyo"	2011-03-18	1056	92
c. Isodin	Isodin is effective against radiation.	2011-03-12	2421	118
d. Seaweed	Seaweed is effective against radiation.	2011-03-12	1798	118
e. Blog	The blog "I want you to know what a nuclear plant is."	2011-03-13	501	170
f. Hutaba	Officials in Hutaba hospital left patients behind and fled.	2011-03-17	1525	118
g. Remark1	Former chief cabinet secretary Sengoku's remark in Tokushima was inappropriate.	2011-03-13	638	170
h. Remark2	Former prime minister Hatoyama remarked "We cannot live within a 200-kilometer radius of the nuclear power plant."	2011-03-16	955	120
i. Visit	Chief Cabinet Secretary Edano visits Korea a few days after the earthquake.	2011-03-15	1973	168
j. Regulation	Ms. Renho proposes to regulate convenience stores to save energy.	2011-03-12	7561	156
k. Rescue	Ms. Tsujimoto protests US military's rescue activities.	2011-03-16	1887	144
l. Taiwan	Taiwan's aid is rejected by the Japanese government.	2011-03-12	2736	156
m. School seismic	Budget for school seismic retrofitting was cut by the project screening.	2011-03-12	1044	174
n. Debt	South Korea asks Japan to borrow money. Moreover, Japan agrees to this.	2011-03-16	399	174
o. Sanjyo	Sanjo Junior High School stopped functioning due to international students.	2011-03-17	379	162
p. Fujitv	Japanese TV company Fuji donated to UNICEF Japan.	2011-03-16	885	124
q. Cartoonist	Japanese cartoonist Mr.Oda donated 1.5 billion yen.	2011-03-12	2546	171
r. Starvation	An infant in Ibaraki died of starvation.	2011-03-16	2025	144
s. Turkey	Turkey donates 10 billion yen for Japan.	2011-03-12	2380	158

2.4.1. Recent Fake News (RFN)

We collected the spread of ten fake news items from two fact-checking sites, Politifact.com [94] and Snopes.com [95], between March and May 2019. PolitiFact is an independent, non-partisan site for online fact-checking, mainly for US political news and politician statements. Snopes.com, one of the first online fact-checking websites, handles political, social, and topical issues. Using the Twitter API,

tweets that were highly relevant to fake news stories were crawled based on keywords and URLs. We selected six fake news stories based on two conditions: 1) the number of posts must be greater than 300 and 2) the observation period must be longer than 36 hours (Indicated by the experiments conducted on the synthetic data, as shown in Figure 2.2). A summary of the collected fake news stories is presented in Table 2.1.

2.4.2. Fake News on the 2011 Tohoku Earthquake and Tsunami (Tohoku)

Numerous fake news stories emerged after the 2011 earthquake off the Pacific coast of Tohoku [96, 97]. We collected tweets posted in Japanese from March 12 to March 24, 2011, using sample streams from the Twitter API. There were 17,079,963 tweets in total. We first identified 80 fake news items based on a fake news verification article [98] and obtained the keywords and related URLs for the news items. Then, we extracted the highly relevant tweets from the fake news. Finally, we selected 19 fake news stories using the same conditions as those used in the RFN dataset. A summary of the collected fake news items is presented in Table 2.2.

2.5. Experimental Evaluation

To evaluate the proposed model, we considered the following prediction task: For the spread of fake news items, we observed a tweet sequence $\{t_i, d_i\}$ up to time T_{obs} from the original post ($t_0 = 0$), where t_i is the ith tweeted time, d_i is the number of followers of the ith tweeting person, and T_{obs} represents the duration of the observation. We then predicted the time series of the cumulative number of posts related to the fake news item during the test period $[T_{\text{obs}}, T_{\text{max}}]$, where T_{max} is the end of the period. In this section, we describe the experimental setup and the proposed prediction procedure and compare the performance of the proposed method with state-of-the-art approaches.

2.5.1. Setup

The total time interval $[0, T_{\max}]$ was divided into the training and test periods. The training period was set to the first half of the total period $[0, 0.5T_{\max}]$ and the test period was the remaining period $[0.5T_{\max}, T_{\max}]$. The prediction performance was evaluated based on the mean and median absolute errors between the actual time series and its predictions.

$$\text{Mean Absolute Error} = \frac{1}{n_b}\sum_{k=1}^{n_b}|\hat{N}_k - N_k|,$$

$$\text{Median Absolute Error} = \text{Median}(|\hat{N}_k - N_k|) \quad (k = 1, 2, \cdots n_b),$$

where $\hat{N}_k$ and N_k are the predicted and actual cumulative numbers of tweets in the k-th bin $[(k-1)\Delta + T_{\text{obs}}, k\Delta + T_{\text{obs}}]$, respectively, n_b is the number of bins, and $\Delta = 1$ hour is the bin width.

2.5.2. Prediction Procedure based on the Proposed Model

First, we fitted the model parameters using the maximum likelihood method from observational data (see Section 4). Second, we calculated the intensity function $\hat{\lambda}(t)$ during the prediction period $t \in [T_{\text{obs}}, T_{\max}]$:

$$\hat{\lambda}_{\text{prop}}(t) = \hat{\lambda}_1(t) + \hat{\lambda}_2(t) \tag{2.8}$$

with

$$\hat{\lambda}_1(t) = p_1(t)\sum_{i:t_i<t_c} d_i\phi(t - t_i), \tag{2.9}$$

where $\hat{\lambda}_1(t)$ and $\hat{\lambda}_2(t)$ are the intensities of the first and second cascades, respectively. The intensity due to the original news item $\hat{\lambda}_1(t)$ is calculated using the fitted parameters $\{a_1, \tau_1; r, \theta_0\}$ and observations $\{t_i, d_i\}$ before the inferred correction time t_c. The number of followers was fixed as 1 ($d_i = 1$) for the Tohoku dataset because follower information was not available in the data. The intensity owing to the correction $\hat{\lambda}_2(t)$ is provided by the solution of the integral equation

$$\hat{\lambda}_2(t) = f(t) + d_p p_2(t)\int_{T_{\text{obs}}}^{t}\hat{\lambda}_2(s)\phi(t - s)ds, \tag{2.10}$$

where

$$f(t) = p_2(t) \sum_{i:t_c<t_i<T_{\text{obs}}} d_i \phi(t - t_i),$$

and d_p is the average number of followers during the observation period.

2.5.3. Compared Methods

We evaluated the prediction performance of the proposed model and compared it with three baseline methods: linear regression (LR) [70], reinforced Poisson process (RPP) [99] and TiDeH [65]. Details of the LR and RPP methods are summarized as below:

2.5.3.1. Linear Regression (LR)

Linear regression is applied to the logarithm of the cumulative number of posts up to time t as follows:

$$\log R_t = \alpha_t + \log R(T_{\text{obs}}) + \sigma_t \xi_t,$$

where R_t is the cumulative number of posts at prediction time t, $R(T_{\text{obs}})$ is the cumulative number of posts at observation time T_{obs}, and ξ_t represents the Gaussian random variable with zero mean and unit variance. The parameters $\{\alpha_t, \sigma_t^2\}$ are estimated using the maximum likelihood method from the training data, where the tweet sequence for the entire period is available. The cumulative number of posts is predicted by the unbiased estimator as follows:

$$\hat{R}_t = R(T_{\text{obs}}) \exp(\hat{\alpha}_t + \hat{\sigma}_t^2/2),$$

where $\hat{R}_t$ is the prediction of the cumulative number and $\hat{\alpha}_t$ and $\hat{\sigma}_t^2$ are the fitted parameters.

2.5.3.2. Reinforced Poisson Process (RPP)

RPP is a point process model similar to TiDeH, where the instantaneous function is written as follows:

$$\lambda(t) = c f_\gamma(t) r_\alpha(R(t)),$$

where $f_\gamma(t) = t^{-\gamma}$ describes the aging effect and $r_\alpha(R) = \epsilon + \frac{1-e^{-\alpha(R+1)}}{1-e^{-\alpha}}$ is a reinforcement mechanism associated with the multiplicative nature of the spreading. The model parameters $\{c, \gamma, \alpha\}$ are determined using the maximum likelihood method. The cumulative number of posts was evaluated based on the expectation of the RPP model, which is as follows:

$$\frac{dR}{dt} = \lambda(t)$$

This can be solved analytically as follows:

$$R(t) = (\log(1 + e^x) - x - \log\tilde{\epsilon} - \alpha)/\alpha,$$

with

$$x(t) = \frac{\tilde{\epsilon}c\alpha(T_{\rm obs}^{1-\gamma} - t^{1-\gamma})}{(1-\gamma)(1-e^{-\alpha})} - (R(T_{\rm obs}) + 1)\alpha - \log(\tilde{\epsilon} - e^{-\alpha(R(T_{\rm obs})+1)}),$$

and $\tilde{\epsilon} = 1 + \epsilon(1 - e^{-\alpha})$. This expression is used to predict the cumulative number.

2.5.4. Prediction Results

Figure 2.4 shows three examples of the time series of the cumulative number of posts related to fake news items and their prediction results. The proposed method (Fig. 2.4: magenta) follows the actual time series more accurately than the baselines. The proposed method reproduces the slowing-down effect on the posting activity, whereas the baseline models often overestimate the number of posts.

Next, we examined the distribution of the proposed model's parameters. The spreading effect of the falsity of news item a_2 was weaker than that of the news story itself a_1 for most fake news items (67% and 79 % in the RFN and Tohoku datasets, respectively). The result can be attributed to the fact that the news story itself is more surprising to users than the falsity of the news. The decay time constant of the first cascade τ_1 was approximately 40 (hours) in both datasets: the median (interquartile range) was 35 (22−92) hours and 40 (19−54) hours for the RFN and Tohoku datasets, respectively. The time constant of the second cascade, τ_2, was widely distributed in both datasets, consistent with the results observed in the synthetic data (Figure 2.2). The correction point t_c tends to be

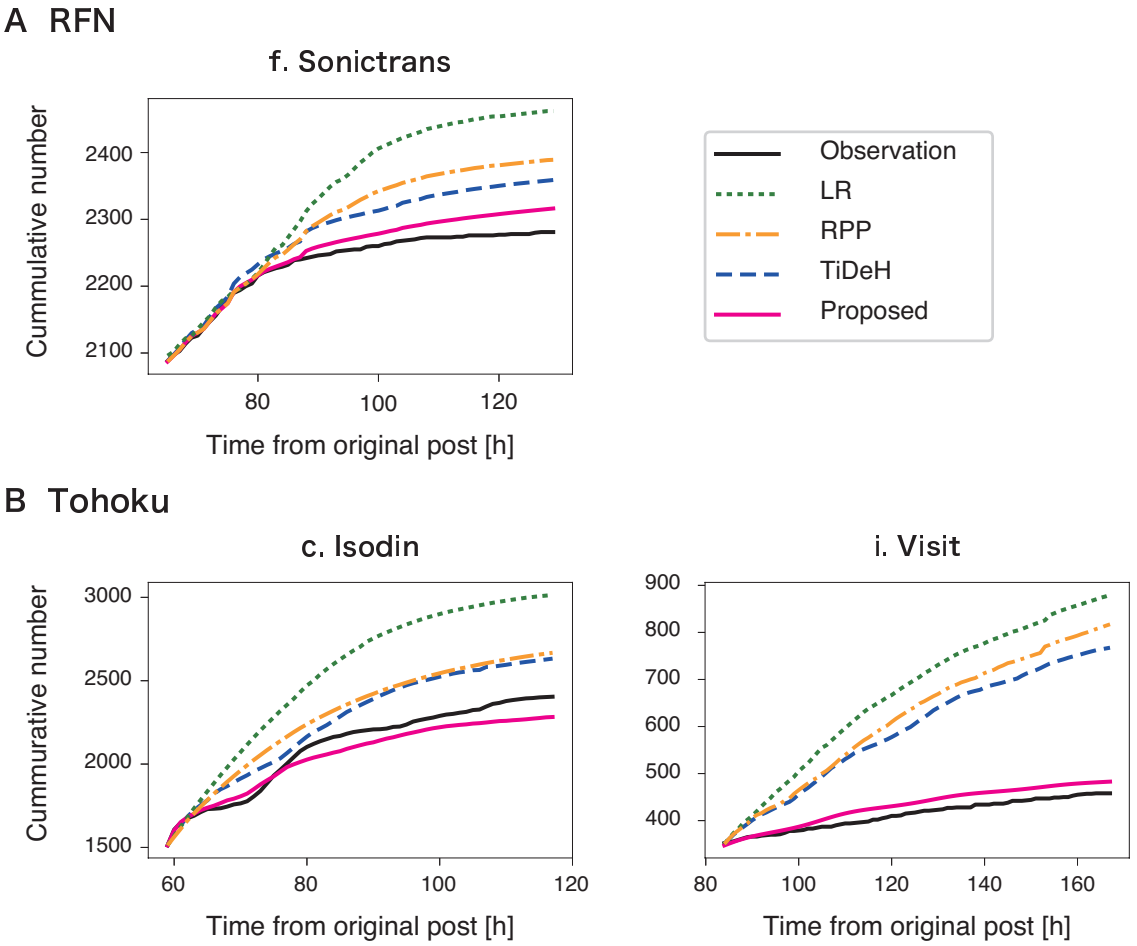


Figure 2.4.: Predicting the time series of the cumulative number of posts related to a fake news item. Prediction results from (A) RFN and (B) Tohoku datasets are shown. The green, orange, and blue dashed lines represent the prediction results for the baseline models, LR, RPP, and TiDeH, respectively. The black and magenta lines represent the observations and their prediction results of the proposed model.

around 30˜40 hours after the original post: 32 (21−54) hours and 37 (31−61) hours for the RFN and Tohoku datasets, respectively. A previous study [100] reported that fact-checking sites detect fake news 10–20 h after the original post. The result implies that Twitter users recognize the falsity of a fake news item 10−20 hours after the initial report by fact-checking sites.

Table 2.3.: Prediction performance on the two datasets: mean and median absolute errors per hour. The best results are shown in bold for each case.

Datasets	RFN		Tohoku	
Metric	Mean	Median	Mean	Median
LR	88.3	5.08	13.9	4.51
RPP	61.8	3.12	8.23	2.30
TiDeH	54.2	1.89	4.12	1.99
Proposed	**36.9**	**1.37**	**2.40**	**1.80**

Finally, we evaluated the prediction performance using two fake news datasets (Table 2.3). Table 2.3 demonstrates that the proposed method outperforms the baseline methods in both datasets and metrics. A comparison of the mean error for the proposed model and TiDeH suggests that the two-stage spreading mechanism reduces the mean error by 32 % and 42 % for the RFN and Tohoku datasets, respectively. Consistent with previous studies [65, 74], the methods based on the point process model (the proposed method, TiDeH, and RPP) outperform the linear regression (LR) method. The proposed model performed best for most fake news items (100% and 89% in the RFN and Tohoku datasets, respectively). However, TiDeH outperformed the proposed model for the other datasets (8%), the proposed model outperformed the other baselines (RPP and LR). Furthermore, we evaluated the goodness-of-fit of the model using the Akaike's information criterion (AIC) [101]. A comparison of the AIC values implies that the proposed model achieves a better fit than TiDeH for most fake news items (100% and 89% in the RFN and Tohoku datasets, respectively). These results suggest that fake news occasionally spreads in a single cascade rather than in two. This might happen when users already know the falsity of the news in advance (e.g., on April Fool's Day) or when they are not interested in the falsity of the news. In summary, these results prove that the proposed method is effective in predicting the spread of fake news posts on Twitter.

2.6. Inferring the Correction Point

We demonstrated that the proposed method outperformed existing methods in predicting the evolution of the spread of a fake news item. The proposed model assumes that Twitter users understand the falsity of news around the correction point t_c. In this section, we examine the validity of this assumption through text mining.

First, we compared the frequency of fake words with the inferred correction time, t_c (Figure 2.5). The fake word frequency is regarded as the number of tweets with fake words (e.g., false rumors, fake, not true, and not real) in each hour. The spread of fake news items in the RFN dataset contained fewer "fake" words than those in the Tohoku dataset. In the RFN dataset, 29 and 277 fake words in the tweets of b. Notredome and f. Sonictrans, and in the Tohoku dataset during the observation period (150 hours), 1,752, 1,616, 1,723, and 1,930 fake words in the tweets of a. Saveenergy, l. Taiwan, q. Cartoonist, and s. Turkey, respectively. This is because most tweets in the RFN dataset are retweets of the original posts. We observed that fake words were posted around correction points. The peak of the fake word frequency was close to the correction point for Taiwan and Cartoonist in the Tohoku dataset (Figure 2.5).

Next, we compared the word cloud before and after the correction point t_c. Figure 2.6 demonstrates an example of a fake news item spreading "Turkey" in the Tohoku dataset. The fake news story is about the huge financial support (10 billion yen) from Turkey to Japan. The word cloud before the correction point implies that this fake news item has spread because Turkey is considered a pro-Japanese country. The term "False rumor" frequently appears after the correction point. The word "Taiwan" also appears after the correction point, and is related to another fake news story about Taiwan. These results suggest that Twitter users became aware of the falsity of news after the correction point, which supports the key assumption of the proposed model.

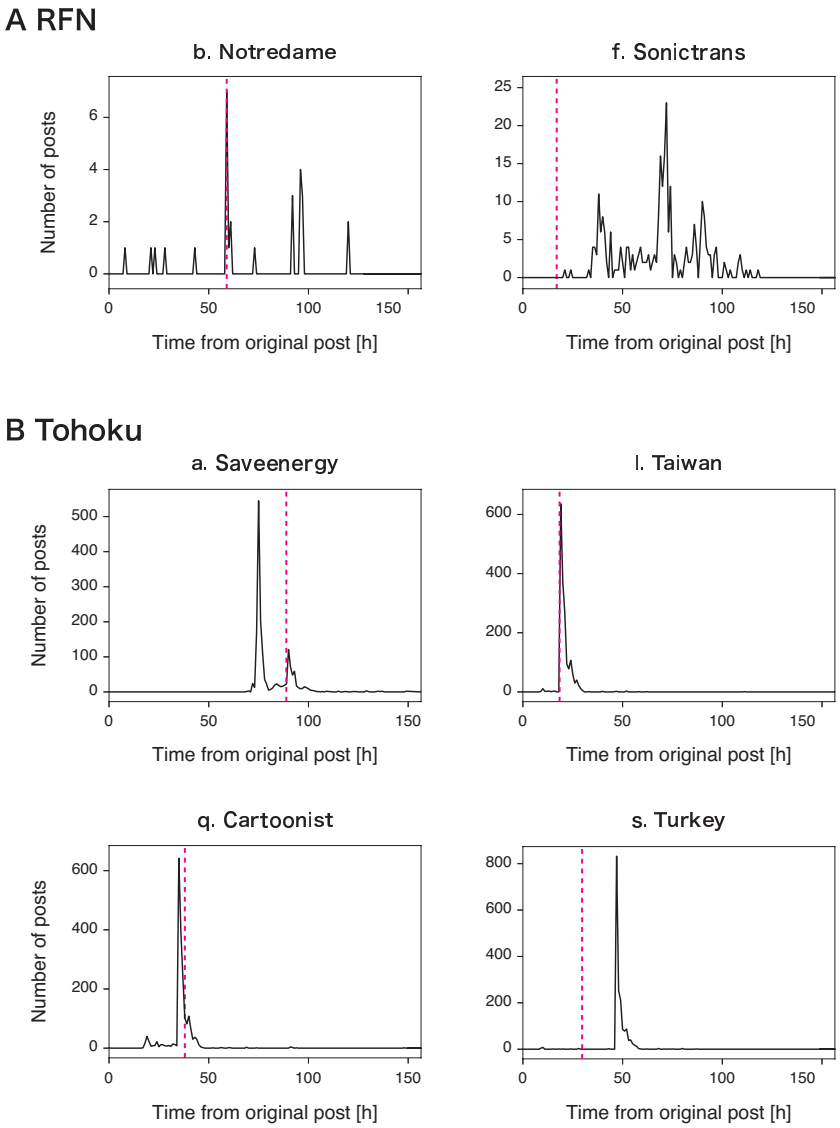


Figure 2.5.: Time series of fake word frequency for fake news items: (A) RFN and (B) Tohoku datasets. In each panel, the black line represents the time series of the "fake" word count per hour for the tweets related to the fake news item and the magenta vertical lines represent the correction point t_c.

Figure 2.6.: Example of a word cloud before (left) and after (right) correction point t_c. Each cloud shows the top ten most frequent words in the fake news story (Turkey in the Tohoku dataset).

2.7. Conclusion

In this section, we propose a point process model for predicting the future evolution of the spread of fake news on Twitter (i.e., tweets and re-tweets related to a fake news story). The proposed model describes the fake news spread as a two-stage process. First, a fake news item spreads as an ordinary news story. Then, users recognize the falsity of the news story and spread it as another news story. We validated this model by compiling two datasets of fake news items spread on Twitter. We have shown that the proposed model outperforms state-of-the-art methods in accurately predicting the spread of fake news items. Moreover, the proposed model can infer the correction point of a news story. Our text-mining-based results indicate that Twitter users recognize the falsity of the news story around the inferred correction time.

There are several interesting directions for future work. The first is to investigate the cascades that exhibit multiple bursts. While most fake news cascades exhibit the two-stage spreading pattern, this pattern can also be observed associated with cascades in general. A previous study [102] found that cascades of image memes on Facebook consist of multiple popularity bursts and argued that content virality is the primary driver of cascade recurrence. Our study implies that a change in the perception of content can be another driver. Additional research is needed to determine whether this hypothesis explains cascading re-

currence better than content virality.

The second direction would be to extend the proposed model. We simply assumed a two-stage process for the spread of a fake news item. This could be extended to describe the spread of fake news in more detail. For example, we can consider multiple types of tweets or hidden variables to incorporate a soft switch from the first to the second stage. Although our model is for short-term fake news, which has not been spread for longer than a week, there are also fake news items that has been prevalent for a long period of time, such as COVID-19 vaccines are dangerous to human health [103]. Our model could be extended to handle such cases. Another direction would be to apply the proposed model to practical problems, such as fake news detection and mitigation. We believe that the proposed model significantly contributes to the modeling of the spread of fake news. This is also beneficial for the extraction of a compact representation of temporal information related to the spread of a fake news item.

The third direction is to investigate how the model parameters change depending on the content of the fake news and the users that spread it. We have not investigated the point owing to a lack of sufficient data. It is important to clarify certain points, such as "How does the diffusion differ between political and entertainment content?", and "Is fake news spread by authenticated users different from that spread by general users?"

3. Fake News Detection using Temporal Features Extracted via Point Process

3.1. Background

This chapter proposes a fake news detection model that leverages the temporal characteristics of social media posts. Our model is based on the idea that temporal movements of social media posts are useful for detecting fake news. Our findings in Chapter 2 reinforce the idea. Existing studies [104, 105] have also investigated whether temporal features are effective in detecting fake news. The time series of posts referring to fake news exhibited movements that differed from those of real news. Nevertheless, few studies have considered the amount of attention fake news attracts over a period.

We propose a fake news detection model that leverages the attention to news changing over time, which is calculated using a self-exciting point process from the post-publication time and the likelihood of people reading the post (determined by the number of followers). In this study, we designated the attention to the news as an "infectiousness value" because it can be measured based on the probability of re-sharing of the information by each new user. The infectiousness value can be regarded as an index of public interest in the news, and it normally decreases over time for real news. Conversely, our underlying finding in Chapter 2 is that the infectiousness value of fake news upsurges twice: the first upsurge results from the original news (including false information), and the second upsurge results from news items for which people doubt or rectify the false information.

The infectiousness value of information is more robust than that of existing

features, which depend on fake news propagators. For example, the text features of early users can easily be manipulated by providing fake comments for diffusion. User features and user-article relationships are being transformed by the regulation of platforms and account suspensions. Propagation paths/trees are difficult to manipulate, but they are expensive to obtain. Infectiousness values are also difficult to manipulate because they are calculated from a series of posts and not by early movement. The number of followers and post-publication time, which are used to calculate the infectiousness values, can be easily obtained.

The proposed fake news detection model leverages three features: combining existing features, texts, and users with an attention-based mechanism, and implementing the infectiousness value. As preliminary research, we investigated whether temporal features can distinguish real news from fake news to validate their effectiveness. Then, experiments were conducted to demonstrate that each module, such as the temporal features, is useful for detecting fake news.

The contributions of this study are as follows:

- We elucidate the differences in infectiousness values associated with real and fake news and consider the differences for fake news detection using a point process.
- We propose a new multi-modal method that combines text and user features with infectiousness values.
- We show the effectiveness of the proposed model for fake news detection on social media posts through experimental procedures.

3.2. Related Work

The "fake news detection task," which assesses the truthfulness of a certain piece of news from news content or social media posts, has been performed by many researchers to save working hours and automate the process. In recent years, with the development of deep learning models, many models have been proposed to achieve high detection performance. In this section, we briefly discuss existing studies that are closely related to our study in terms of the features used. In recent years, with the development of deep learning models, many models have

been proposed to achieve high detection performance. In this section, we briefly discuss existing studies that are closely related to our study in terms of the features used. A classical fake news detection model learns the textual style of fake news and then classifies them as fake based on input data, such as information on the news content or social media posts, comprising the news [82, 106–108]. An initial study [109] used various linguistic features such as special characters, sentiment words, and emojis to detect fake news. Another study [110] used bag-of-words, the presence of URLs, and hashtags; then, a support vector machine (SVM) was used to detect false claims. Recent studies used deep learning models to capture temporal—linguistic features. Ma et al. [93] used recurrent neural networks (RNNs) to capture temporal—linguistic features from the bag-of-words of user posts. Another study of Ma et al. [111] used RNNs based on the text in a reply tree. Zhang et al. [112] proposed a detection model comprising two parts: a claim encoder and a reply encoder. Their concatenation determines the posterior belief of the claim veracity. Other examples include convolutional neural networks [113], hierarchical attention networks [114], and neural network models using discourse-level structures [115].

Moreover, several methods have been examined to detect fake news using the characteristics of the users who posted the information. Previous studies [109, 116, 117] used various models based on user characteristics such as the number of followers, number of friends, and registered age. Recently, the relationship between news articles and users has been used to determine news credibility, assuming that if two articles have a strong relationship as determined by the number of users who re-shared them, they are likely to share the same label [118]. Other studies have employed detection methods based on propagation paths/trees or networks of posts on social media. Jin et al. [119] used epidemiological models for capturing and characterizing information cascades. Ma et al. [120] proposed a graph kernel-based SVM classifier that calculates the similarity between propagation tree structures.

Multi-modal approaches combine features of different types to detect fake news. For example, Ruchansky et al. [83] combined text and user behavior, whereas Wang et al. [121] combined text and visual features extracted from posts on social media. Our model effectively combines text and user features using contextual

intermodal attention [122] to determine the relationship between a user and the post content.

In a method of fake news detection using temporal features similar to the proposed method, Lukasik et al. [84] demonstrated the importance of using post-temporal information for rumor stance classification. Kwon et al. [123] used SpikeM [71] to mathematically capture the time series behavior of information for long-term rumor detection, in addition to using other features (e.g., linguistic, user, and network). In this study, we demonstrated that temporal features are useful for short-term fake news detection. The proposed multimodal framework utilizes linguistic, user, and temporal features that are easy to obtain to capture the characteristics of fake news.

3.3. Preliminary Research

Figure 2.5 in Chapter 2 shows the characteristics of posts about fake news. In this section, we further compare the time series of posts and the infectiousness values converted to the time series between fake news and real (not fake) news Then, we validate the contribution of temporal features in posts on social media to determine whether the news is fake or real.

Figure 3.1 presents several time series of fake and real news in the US, with details presented in Table 3.1. Each news item has two time series: the upper one indicates the number of posts per hour, and the lower one indicates the infectiousness values calculated using the self-exciting point process described in Section 3.4, which represents the probability of re-sharing. The time series of the number of posts about true news shows a large upsurge in a few hours; however, it decays quickly over time. Several upsurges in the time series of fake news are based on the assumption "the attention by the post of question or denial to the news causes the second upsurge," which is proven in Chapter 2. Most time series of fake news posts show a second upsurge after about a day, aside from the first upsurge in the figure of the number of posts. These time series exhibit unstable behavior in the infectiousness values of fake news compared to those of real news. A previous study [123] indicated that the time series of rumors has multiple upsurges in long-term observations (56-day), unlike those of non-

Table 3.1.: Descriptions of the collected news.

News No.	Content	Date
Fake-1: Foxnews	New Zealand took Fox News/Sky News off the air to mosque shooting coverage.	2019-03-25
Fake-2: Sonic	The animated character of Sonic the Hedgehog ill be transgender in a new film.	2019-03-25
Fake-3: Notredame	A man dressed in Muslim garb walked in a Notre Dame tower during the fire.	2019-04-13
Real-1: BritishIS	British man who fought against IS guilty of terrorism.	2019-10-24
Real-2: Pengagon	Pentagon official overseeing Ukraine testifies in impeachment inquiry after GOP delay.	2019-10-24
Real-3: Hunterbiden	Examining Hunter Biden's legal work in Romania.	2019-10-25

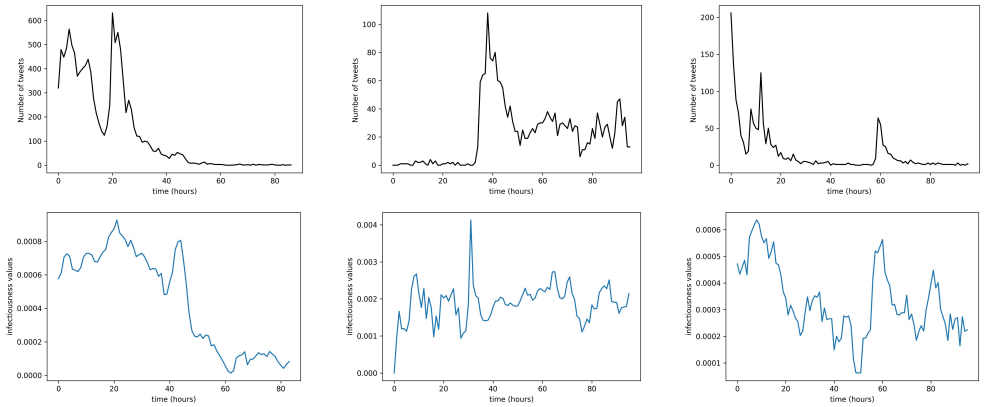
rumors. By contrast, our results demonstrated that the time series of fake news has multiple upsurges in short-term observations (4-day), unlike real news.

3.4. Fake News Detection Model using Temporal Features

As mentioned previously, although temporal features are useful, fake news detection using temporal features alone cannot achieve sufficient performance. For this reason, we proposed a novel multi-model method to detect fake news from many posts on social media. The proposed model effectively combines linguistic and user features using Attention. It then combines them with temporal features. The overall model architecture is shown in Figure 3.2.

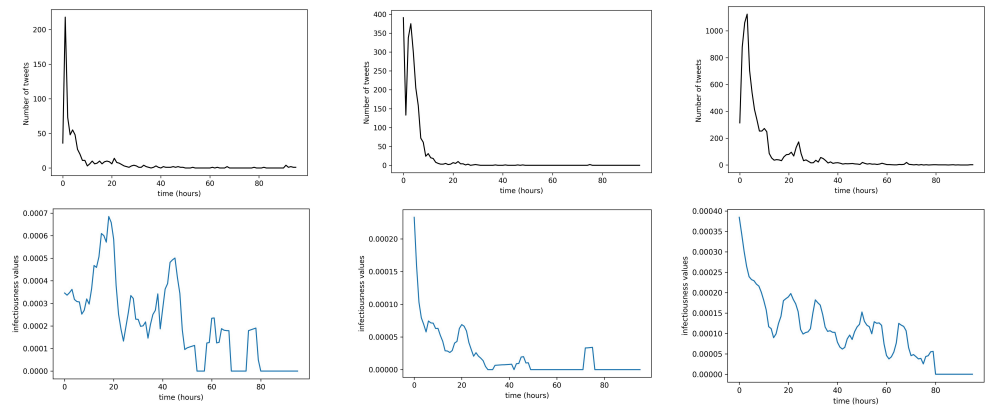
3.4.1. Problem Statement

The task of fake news detection is the prediction of the news label (true or fake), based on news stories, including posts on social media. We assume A_i is a set



1. Foxnews 2. Sonic 3. Notredame

Fake



1. BritishIS 2. Pentagon 3. Hunterbiden

Real

Figure 3.1.: Time series of posts about fake/real news in the US. Each news item has two time series extracted from the 96-hour observation period (the X-axis represents hours), with the upper showing the number of posts per hour (the Y-axis represents the number of posts), and the lower indicating the infectiousness values calculated using the self-exciting point process (the Y-axis represents the infectiousness values). Similar phenomena can also be observed for time series of infectiousness values. Details of these news items are described in Table 3.1.

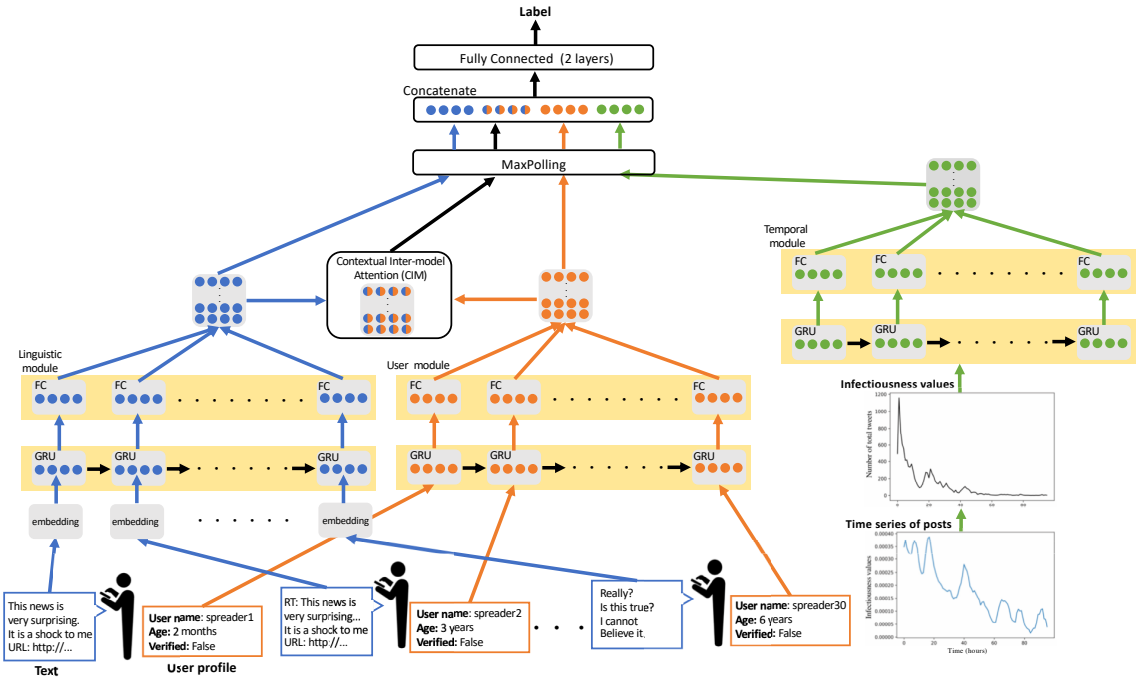


Figure 3.2.: Architecture of the proposed fake news detection model. GRUs have been used to learn the latent representations of linguistic, user, and temporal features. Additionally, a pairwise contextual inter-model attention mechanism (CIM) has caused combinations of linguistic and user features. Finally, the model predicts the news label by concatenating these features.

of news stories consisting of N_i posts, and $A_i = \{a_1, a_2, ..., a_{N_i}\}$. Each post $a_t = (\mathbf{l}_t, \mathbf{u}_t)$ consists of two features: linguistic feature $\mathbf{l}_t$ and user feature $\mathbf{u}_t$. Temporal features of a news story i are represented as $\mathbf{s}_i$. In addition, each news item A_i is associated with a label $L(A_i)$, which has categorical variables $\{0, 1\}^\tau$. We aim to learn a fake news detection function $f : f(A_i, \mathbf{s}_i) \rightarrow L(A_i)$ that maximizes prediction accuracy.

3.4.2. Model Structure

The model comprises various components. Linguistic, User, and Temporal modules convert inputs to latent features. Contextual intermodel attention combines

Table 3.2.: Major notations

Notation	Definition or Description
A_i	i_{th} news story
a_t	t_{th} post of news story
$\mathbf{l}_t$	linguistic feature of t_{th} post
$\mathbf{u}_t$	user feature of t_{th} post
$\mathbf{s}_i$	temporal features of i_{th} news story
$\mathbf{s}^h$	Infectiousness values at each point
$\mathbf{l}_t^e$	I^l-dimensional post embedding of t_{th} post
$\tilde{\mathbf{h}}_t^*$	the hidden state of t_{th} post through GRU in each module
$\mathbf{h}_t^*$	the hidden states of t_{th} post through FC in each module
$\mathbf{h}^{max\_*}$	each hidden states through MaxPooling
$\mathbf{z}$	Final output representing the class probability
H^*	each module output consisting of a sequence $[\mathbf{h}_t^*]$
T^*	Number of sequence lengths in each module
E^*	Number of dimensions about hidden states $\mathbf{h}_t^*$ in each module
E^{con}	Number of dimensions about $\mathbf{f}^1$

latent features generated by linguistic and user modules with attention. Finally, the Classification module outputs the prediction label.

3.4.2.1. Linguistic Module

We first converted the raw text of each post a_t to the linguistic feature $\mathbf{l}_t$ for the interpretation of the model. Thus, we used the tf-idf values of the vocabulary terms for each post. We used the top-K vocabularies according to their tf-idf values. Therefore, we converted the post to linguistic feature $\mathbf{l}_t \in \mathbb{R}^K$, which is a K-dimensional vector. Linguistic feature $\mathbf{l}_t$ created from the post is high and sparse dimensional. Therefore, we convert the vector $\mathbf{l}_t$ into low-dimensional representation. Rather than using pretrained vectors based on external collections, we learn the embedding matrix using our model.

$$\mathbf{l}_t^e = Embedding(\mathbf{l}_t), \tag{3.1}$$

where $\mathbf{l}_t^e \in \mathbb{R}^{I^l}$ denotes the I^l-dimensional post embedding vector of $\mathbf{l}_t$.

From each post embedding, $L_i^e = [\mathbf{l}_1^e, \mathbf{l}_2^e, ..., \mathbf{l}_{T^l}^e]$ be a sequence of embedding posts, we extract latent linguistic features to use gated recurrent units [124] (GRU). GRUs based on RNNs can capture long-term dependency to learn temporal—linguistic features from early posts on social media. GRU considers $\mathbf{l}_t^e$ and $\tilde{\mathbf{h}}_{t-1}$ as input and produces $\tilde{\mathbf{h}}_t$ as the output, the respective formulas of which are described below:

$$\begin{aligned} \mathbf{z}_t^l &= \sigma\left(U_z^l \mathbf{l}_t^e + W_z^l \tilde{\mathbf{h}}_{t-1}^l\right), \\ \mathbf{r}_t^l &= \sigma\left(U_r^l \mathbf{l}_t^e + W_r^l \tilde{\mathbf{h}}_{t-1}^l\right), \\ \mathbf{f}_t^l &= tanh\left(U_h^l \mathbf{l}_t^e + \tilde{\mathbf{h}}_{t-1}^l \odot W_h^l \mathbf{r}_t\right), \\ \tilde{\mathbf{h}}_t^l &= \left(1 - \mathbf{z}_t^l\right) \odot \tilde{\mathbf{h}}_{t-1}^l + \mathbf{z}_t^l \odot \mathbf{f}_t^l, \end{aligned} \tag{3.2}$$

where $\mathbf{z}_t, \mathbf{r}_t$ represent the reset and update gates, respectively, at time t. In addition, $U_z^l, U_r^l, U_h^l \in \mathbb{R}^{I^l \times E^l}$, $W_z^l, W_r^l, and W_h^l \in \mathbb{R}^{E^I \times E^l}$ are parameters for the respective gates. E^l denotes the output dimension of GRU. We present equation (2) as shown below:

$$\tilde{\mathbf{h}}_t^l = GRU\left(\mathbf{l}_t^e\right), \quad t \in \left\{1, ..., T^l\right\}. \tag{3.3}$$

Then, the hidden state $\tilde{\mathbf{h}}_t^l$ of the GRU is applied by the fully connected layer (FC), resulting in $\mathbf{h}_t^l \in \mathbb{R}^{E^l}$, where E^l is the number of outputs in the FC layer.

$$\mathbf{h}_t^l = FC(\tilde{\mathbf{h}}_t^l) \tag{3.4}$$

3.4.2.2. User Module

We used eight common characteristics extracted from user profiles of social media as user features, similar to [117]. These eight characteristics are listed in Table 3.3.

We represent the eight common features on post a_t as $\mathbf{u}_t \in \mathbb{R}^{I^u}$. As with linguistic features, we used GRUs to capture long-term dependency and FC for user features, as follows:

$$\begin{aligned} \tilde{\mathbf{h}}_t^u &= GRU(\mathbf{u}_t), \quad t \in \{1, ..., T^u\} \\ \mathbf{h}_t^u &= FC(\tilde{\mathbf{h}}_t^u) \end{aligned} \tag{3.5}$$

Table 3.3.: List of user feature extracted user-profiles on social media

Characteristics	Type
Length of user description	Integer
Length of user name	Integer
No. of followers	Integer
No. of follows	Integer
No. of posts	Integer
Registration age	Integer
Verified	binary
Geo enabled	binary

3.4.2.3. Temporal Module

In the previous section, we described the differences between the appearance times of posts on true and fake news. To capture the potential components of these behaviors, we convert the time series of posts to infectiousness values, which represent the re-share probability and decrease as the news becomes stale, using the self-exciting point process model (designated as SEISMIC) [74]. SEISMIC, based on the Hawkes process [125], calculates the infectiousness values p_t at time t using the number of posts R_t at time t and the intensity λ_t, which is calculated as presented below:

$$\lambda_t = p_t \sum_{t_i \leq t, i \geq 0} n_i \phi\left(t - t_i\right), \quad t \geq t_0. \tag{3.6}$$

$$\phi\left(s\right) = \begin{cases} c & if \quad 0 < s \leq s_0, \\ c(s/s_0)^{-(1+\theta)} & if \quad s > s_0, \end{cases} \tag{3.7}$$

where n_i represents the number of people accessing the news (the number of followers). In addition, $\phi(\cdot)$ denotes the memory kernel, which quantifies the delay between a post arriving at a user and the user re-sharing it. These parameters were estimated by [74]: s_0 is 5 min, θ is 0.242, and $c = 6.27 \times 10^{-4}$. This process is called *self-exciting*, because each previous observation i contributes to the intensity λ_t.

The estimation of p_t to vary over time depends on a sequence of one-sided

kernels $K_t(s)$, which up-weights the most recent posts and down-weighs older posts.

$$p_t = \frac{\sum_{i=1}^{R_t} K_t(t - t_i)}{\sum_{i=0}^{R_t} n_i \int_{t_i}^{t} K_t(t - s)\, \phi(s - t_i)\, ds} \tag{3.8}$$

$$K_t(s) = \max\left\{1 - \frac{2s}{t}, 0\right\}, \quad s > 0. \tag{3.9}$$

The aforementioned equations are used to calculate the infectiousness values p_t at each point from the time and number of followers for each post. As described here, $\mathbf{s}_i$ is defined as $\{..., (time_t, follower_t), ...\}, t \in \{1, ..., N\}$, where $time_t$ represents the time elapsed since the first post. Subsequently, $\mathbf{s}_i$ is converted to the infectiousness values $\mathbf{s}^h = \left(\mathbf{s}_1^h, \mathbf{s}_2^h, ..., \mathbf{s}_{T^s}^h\right)$ at each point. With the linguistic and user features, we utilize GRUs and FC for the temporal features, as explained below:

$$\begin{aligned} \mathbf{s}^h &= ConvertInfectiousness(\mathbf{s}_i) \\ \tilde{\mathbf{h}}_t^s &= GRU(\mathbf{s}_t^h), \quad t \in \{1, ..., T^s\} \\ \mathbf{h}_t^s &= FC(\tilde{\mathbf{h}}_t^s) \end{aligned} \tag{3.10}$$

3.4.2.4. Contextual Inter-model Attention

A post comprises linguistic features and user features, which often have mutual interdependence. In addition, GRUs are unable to capture characteristics of their interdependence. To capture their interdependence characteristics, we used a pairwise contextual inter-model attention mechanism (designated as CIM) [122], using each latent representation by GRUs.

We compute the attention between the output of the linguistic features $H^l = [\mathbf{h}_1^l, \mathbf{h}_2^l, ..., \mathbf{h}_{T^l}^l] \in \mathbb{R}^{T^l \times E^l}$ and that of user features $H^u = [\mathbf{h}_1^u, \mathbf{h}_2^u, ..., \mathbf{h}_{T^u}^u] \in \mathbb{R}^{T^u \times E^u}$ to leverage the contextual information related to each post for fake news detection, where E^l and E^u are the same, as well as T^l and T^u are the same. First, a pair of matching matrices M_1 and $M_2 \in \mathbb{R}^{T^l \times T^u}$ is computed as presented as follows:

$$M_1 = H^l \cdot H^{u\mathsf{T}}; \quad M_2 = H^u \cdot H^{l\mathsf{T}} \tag{3.11}$$

Furthermore, we obtain the probability distribution scores $N_1, N_2 \in \mathbb{R}^{T^l \times T^u}$ over the respective matching matrices M_1 and M_2 to compute the attention weights on

the contextual posts using a softmax function. We then compute the modality-wise attentive representations.

$$\begin{aligned} N_1(i,j) &= \frac{e^{M_1(i,j)}}{\sum_{k=1}^{T^l} e^{M_1(i,k)}}, \quad for\ i,j = 1,...,T^l, \\ N_2(i,j) &= \frac{e^{M_2(i,j)}}{\sum_{k=1}^{T^l} e^{M_2(i,k)}}, \quad for\ i,j = 1,...,T^l, \\ O_1 = N_1 \cdot H^u, &\quad O_2 = N_2 \cdot H^l \end{aligned} \tag{3.12}$$

Finally, we compute the element-wise matrix multiplication (3.13) for the attention to the important components. We then concatenate them to obtain attention representations between H^l and H^u.

$$\begin{aligned} A_1 = O_1 \odot H^l, &\quad A_2 = O_2 \odot H^u \\ H^{ul} = concat\left[A_1, A_2\right] &\in \mathbb{R}^{T^l \times 2E^l} \end{aligned} \tag{3.13}$$

3.4.2.5. Classification Module

After obtaining each feature through each module, we apply them to MaxPooling and concatenate each feature into a single vector $\mathbf{f}^1 \in \mathbb{R}^{E^l+E^u+2E^l+E^s}$.

$$\begin{aligned} \mathbf{h}^{max\_l} &= MaxPooling\left(H^l\right) \\ \mathbf{h}^{max\_u} &= MaxPooling\left(H^u\right) \\ \mathbf{h}^{max\_s} &= MaxPooling\left(H^s\right) \\ \mathbf{h}^{max\_ul} &= MaxPooling\left(H^{ul}\right) \\ \mathbf{f}^1 = concat[\mathbf{h}^{max\_l}, \mathbf{h}^{max\_u}, \mathbf{h}^{max\_ul}, \mathbf{h}^{max\_s}] & \end{aligned} \tag{3.14}$$

To predict the class label for each news item, we use fully connected layers (FC) with an activation function such as $ReLU$, which are two layers, to identify the complex relations between the respective features. The final output $\mathbf{z} \in \mathbb{R}^\tau$ represents the probability distribution over a set of τ classes using the $Softmax$ function.

$$\begin{aligned} \mathbf{f}^2 &= ReLU\left(FC\left(\mathbf{f}^1\right)\right) \\ \mathbf{z} &= Softmax\left(FC\left(\mathbf{f}^2\right)\right) \end{aligned} \tag{3.15}$$

Table 3.4.: Summary of datasets

Statistics	Weibo	Twitter15	Twitter16
No. of all news	4664	1479	813
No. of true news	2351	371	204
No. of fake news	2313	363	205
No. of unverified news	-	373	205
No. of debunking	-	372	199
No. of posts for training	2973	942	517
No. of posts for validation	525	167	97
No. of posts for test	1166	370	204

3.5. Experiments

3.5.1. Datasets

For the experimental evaluation, we used three publicly available datasets: Weibo, released by [93], Twitter15, and Twitter16, released by [120]. Each dataset of posts related to fake news was collected from the most popular social media sites, Weibo in China and Twitter in the US. The Weibo dataset was annotated with one of the two labels: "true" or "fake." Twitter datasets are annotated with one of four class labels: "true," "fake," "unverified," or "debunking of fake." Table 3.4 presents a summary of the datasets.

For the experiments, we divided each dataset into training, validation, and testing sets. First, we split each dataset in a ratio of 3:1 for the training and test sets. We then hold 15% of the training set for the validation set.

3.5.2. Comparative Models

We conducted comprehensive comparisons between our models and some baselines for fake news detection tasks.

The comparative models are presented below:

- **SVM-TS** [126]: A linear SVM classifier that uses time-series to model the

variation of social context features. This model also uses diffusion-based features such as the average number of re-shares, in addition to linguistic and user features.

- **CSI** [83]: CSI is a hybrid deep-learning model that uses information from user texts, responses, and behaviors. This model calculates the source characteristics based on user behavior and classifies an article as fake.
- **GRU-2** [93]: GRU-2 is equipped with two GRU hidden layers and an embedding layer following the input layer for learning rumor representations by modeling the sequential structure of relevant posts over time.
- **PPC** [117]: PPC is a time series classifier that incorporates both recurrent and convolutional networks that respectively capture user characteristics along the propagation path.
- **Ours (w/o CIM)**: It is a model removing the contextual inter-model attention module from our model for validating the effectiveness of CIM.
- **Ours (w/o time)**: This model used only the two features for learning the model. It uses linguistic and user features to validate the effectiveness of temporal features.
- **Ours (freq)**: This model replaces infectiousness values with the number of posts in each period for validating the effectiveness of the module converting to infectiousness values.

3.5.3. Experimental Settings

Our model was trained to minimize the binary/categorical loss function to predict the class label of each news item in the training set. During training, all the model parameters were updated using gradient-based methods following the Adadelta update rule [127]. In addition, dropout [128], for which the value was set to 0.5, was applied to the hidden layers $\tilde{\mathbf{h}}_t^*, \mathbf{h}_t^*, \mathbf{f}^1$, and $\mathbf{f}^2$ to avoid overfitting. The number of training epochs was set to 500. Early stopping was applied as the validation loss saturated for 10 epochs.

The network structure and hyper-parameters were set based on the validation set and previous studies [93, 117]. We set 5000 vocabularies as the top-K based on the tf-idf values to input the linguistic module. We converted these tf-idf values to embedding vectors, the dimension of which I^t was 100. We set eight as I^u based on Table 3.3. The sequence lengths of the GRUs for the linguistic and user features, T^l and T^u, were chosen as above 30 in the Weibo dataset and above 40 in the Twitter15 and Twitter16 dataset, based on results of a previous study [117]. In the case study, most time series of the number of fake news posts showed a second upsurge approximately one day after the initial post. Therefore, the sequence length of GRUs for the temporal features T^s was 47, composed of infectiousness values for the first two days (these hourly values were obtained from the initial post).

The output size of each GRU (E^l, E^u, and E^s) was selected from (16, 32, 64, and 128), and the hidden dimension of output FC $\mathbf{f}^2$ was selected from (E^{con}, $\frac{E^{con}}{2}$, $\frac{E^{con}}{4}$, and $\frac{E^{con}}{8}$) in the validation period, where E^{con} is the size of $\mathbf{f}^1$ and is the same as $(E^l + E^u + 2E^l + E^s)$.

We used accuracy and the F1-measure as metrics to evaluate the model capabilities. Classification tasks such as fake news detection are commonly evaluated by their accuracy. The F1-measure also enables the accuracy to address class imbalance. We used the accuracy over all categories and the F1-measure for each class to evaluate model performance. The equation for the accuracy and F1-measure is shown below:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FP + TN + FN}, \\ F_1 &= \frac{2Recall \cdot Precision}{Recall + Precision}, \\ Recall &= \frac{TP}{TP + FN}, \\ Precision &= \frac{TP}{TP + FP}, \end{aligned} \tag{3.16}$$

where TP is true positive, FP is false positive, TN is true negative, and FN is false negative.

Table 3.5.: Fake news detection results on each dataset

	Weibo			Twitter15					Twitter16				
Model	Acc.	F_1		Acc.	F_1				Acc.	F_1			
		T	F		T	F	U	D		T	F	U	D
SVM-TS	0.827	0.831	0.837	0.599	0.772	0.598	0.608	0.544	0.574	0.743	0.488	0.551	0.549
CSI	0.780	0.750	0.803	0.556	0.601	0.631	0.550	0.530	0.507	0.552	0.511	0.475	0.443
GRU-2	0.876	0.872	0.879	0.794	0.822	0.815	0.849	0.697	0.750	0.761	0.750	**0.771**	0.723
PPC	0.914	0.912	0.917	0.806	0.748	0.840	0.807	0.730	0.778	0.803	0.760	0.711	0.767
Ours (w/o CIM)	0.920	0.922	0.917	0.814	0.807	0.813	**0.870**	0.745	0.791	0.850	0.782	0.747	0.791
Ours (w/o time)	0.912	0.913	0.910	0.814	0.857	0.806	0.868	0.677	0.791	0.864	0.829	0.717	0.776
Ours (freq)	0.921	0.931	0.908	0.807	0.872	0.815	0.828	0.660	0.805	0.864	0.801	0.740	0.699
Ours	**0.937**	**0.937**	**0.936**	**0.831**	**0.880**	**0.850**	0.833	**0.758**	**0.819**	**0.870**	**0.831**	0.739	**0.841**

3.6. Results and Discussion

Results are presented in Table 3.5. Our model outperformed most baselines, indicating the advantages of our multi-model method and temporal features. One baseline based on hand-crafted features, **SVM-TS**, is a better model because it combines various features, including linguistic, user, and temporal features. In contrast, **CSI** showed low accuracy. The model calculates a user relation score from the training data and then detects fake news from the test data using the scores of users who appear in both the training and test data. We infer that few users appeared in both the training and test data that were used for the experiments, resulting in low accuracy. Most deep learning-based models, such as **Ours**, **GRU-2**, and **PPC**, outperformed feature engineering-based models, such as **SVM-TS**. Deep neural networks help to learn better hidden representations of people's responses to news on social media for fake news detection. The results show that **GRU-2** and **PPC**, used linguistic and user features, respectively, to capture complex hidden features indicative of their responses, achieved both high accuracy and a high F1-measure.

To validate the effectiveness of each module, we also conducted experiments with models that excluded CIM and temporal features from the proposed model. Compared with the proposed model excluding the CIM module **Ours (w/o CIM)**, the proposed model **Ours** achieved higher accuracy and F1-measures on all datasets, except for unverified label data. This result demonstrates that **Ours (w/o CIM)** is insufficient for learning the hidden representations of the user and linguistic features differently. In addition, inter-dependencies between

the linguistic and user features were useful in detecting whether a message was fake or not because one post consisted of two features. Compared to **Ours (w/o time)**, the proposed model **Ours** achieved higher scores, except for the unverified label data in the Twitter15 dataset. In a previous study [123], the time series of rumors was useful for detecting rumors in long-term observations (56 days). However, these results support our claim that temporal features can be useful for early fake news detection (2 days). **Our (freq)** model has replaced infectiousness values with the number of posts in each period for the validation of the conversion to infectiousness values. Its accuracy was slightly higher than that of **Ours (w/o time)** for the Weibo and Twitter16 datasets when adding the number of posts. Simultaneously, the degree of increased accuracy was not much higher than that of the proposed model **Ours**. This result shows that conversion to infectiousness values is useful for capturing latent information from the temporal features of the fake news detection problem.

Ours performs the best for most measures and datasets, with demonstrating its effectiveness, except for unverified label data. Specifically, our model achieved the highest accuracy (0.937) for the Weibo test subset, highest accuracy (0.831) for the Twitter15 test subset, and the highest accuracy (0.819) for the Twitter16 test subset. Additionally, our model achieved the highest performance in terms of the F1 score for the true, fake, and debunking label data. However, our model did not produce good results for classifying unverified labels and was not significantly different from the other models. Presumably, judging ambiguous labels such as unverified, not true, and fake is difficult, even when adding temporal features.

We evaluated the details of the contributions of the temporal features. To examine these contributions, we compared the proposed models with varying time frames to obtain temporal features from zero (without temporal features) to six days. Figure 3.3 shows the accuracies in three datasets, and Table 3.6 shows the mixing matrix of each result with varying time frames in the Weibo dataset. The accuracy of the proposed model improves gradually as the time frame lengthens. However, the performance of the proposed model remained unchanged for more than three days of that time frame. Specifically, the model using five days of the Weibo dataset achieved an accuracy of 0.939. Those using four days of Twitter15 and Twitter16 datasets achieved accuracies of 0.867 and

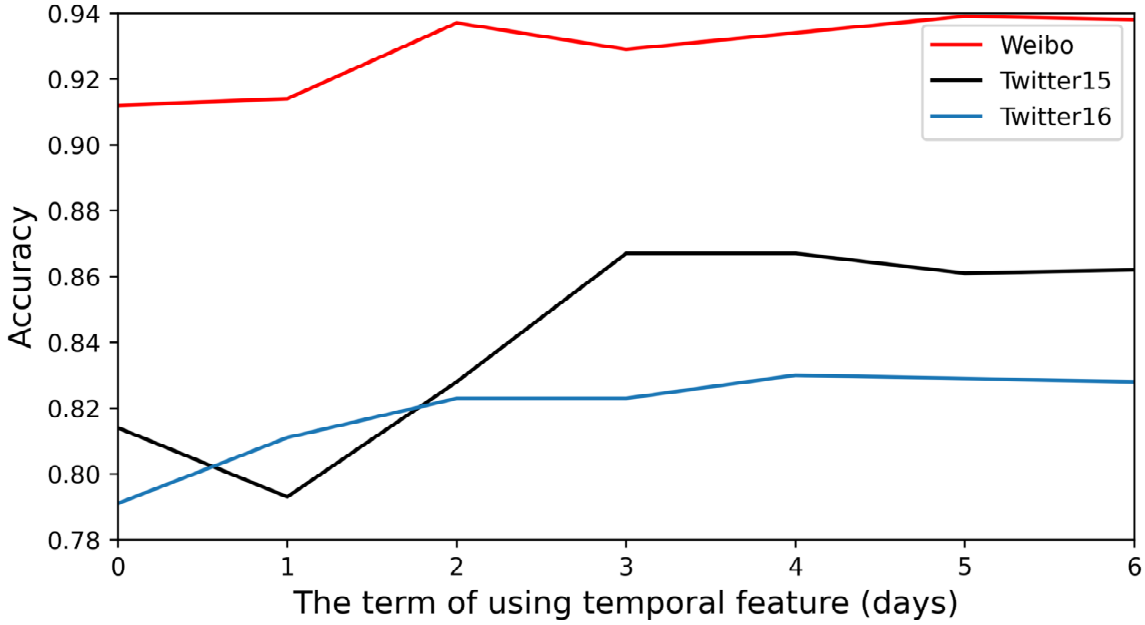


Figure 3.3.: Accuracy of the proposed model with temporal features obtained from varying time frames of each dataset: the X-axis represents the time frames ranging from 0 without the temporal features to 6 days; whereas, the Y-axis represents the accuracy. When a longer time-frame is used, it appears that more accuracy is achieved.

0.830, respectively. Although we set the time frame as the first two days in the experimental settings, the results show that approximately four or five days would be an appropriate period of time to obtain the temporal features for fake news detection.

3.7. Conclusion

We conclude by emphasizing the following points in this chapter.

1. We ascertained differences in time series behaviors between true and fake news from short-term observations.

2. We proposed a new multi-model combining text and user features and using infectiousness values.

Table 3.6.: Mixing matrix with varying time frames in the Weibo dataset

Time frames	-	one day	two days	three days	four days	five days	six days
Accuracy	0.913	0.916	0.937	0.923	0.930	0.939	0.938
TP (%)	46.0	46.1	46.9	46.3	46.7	47.2	47.1
TN (%)	45.4	45.5	46.7	46.0	46.2	46.7	46.7
FP (%)	4.1	3.9	3.2	3.8	3.3	2.9	3.0
FN (%)	4.5	4.5	3.2	4.0	3.7	3.2	3.2

3. The proposed model empirically shown to be effective for the fake news detection problem in the experiments.

However, it remains uncertain whether the temporal features are useful for ambiguous data such as debunking labels. Future studies must examine how temporal features can be flexibly used for such ambiguous data.

4. Towards Countermeasures against the Problem of Fake News in Japanese Society

4.1. Background

This chapter describes our approaches to countering fake news spread on social media in Japanese society. Recently, social applications based on fake news research have become active owing to their growing need. For example, NewsGuard [129] and Hoaxy [100] were developed as applications for tracking and visualizing fake news in English posts on social media and websites. Particularly, in US society, such applications are widespread among many people. This is because of the existence of rich resources, including active fact-checking sites such as Politifact [94] and Snopes [95], research on English resources such as datasets and competitions, and the findings from studies on the construction of fake news detection models and the analysis of fake news ecosystems based on these resources.

However, applying these research results to countries where English is not the native language, such as Japan, is difficult because of the following two issues. The first issue is that there are few fact-checking organizations in the country. In the US, for example, there are a wide variety of other fact-checking organizations, including Buzzfeed [130], GossipCop [131], Poynter [132], and Factcheck.org [133], in addition to Politifact and Snopes. By contrast, in Japan, the Fact Check Initiative Japan [134] is only an active fact-checking organization that collaborates with multiple news media. The number of news stories verified is also very different between the two countries. Politifact, one of the US fact-checking sites, has

verified 151 news stories in one month (October 2021), whereas the Fact Check Initiative Japan has verified 279 news stories in two years (from September 2019 to September 2021). The difference in the number of fake news stories spread in each country is one factor; however, this situation particularly indicates that the demand and effort for fact-checking in Japan are quite primitive. This first issue causes another issue: there are few non-English (Japanese) resources available for fake news detection. The labeling of the veracity of news articles and social media posts in many existing fake news detection datasets depends on the judgments of the existing fact-checking organizations. Fewer fact-checked articles cause fewer samples in the dataset or more annotation effort in dataset construction. Therefore, there is still no Japanese fake news detection dataset. These issues are the reasons why the construction of a fake news detection system in practice is difficult.

This chapter introduces two trials of the fake news problem in Japan: the construction of a Japanese fake news dataset and the fake news collection system. The first trial is the construction of the first fake news dataset in Japanese. The construction does not simply follow the same procedures used in the existing English fake news datasets. We propose an annotation scheme that adds novel perspectives based on our findings, which result from our comprehensive survey of existing fake news detection datasets. This is a useful scheme for building fake news datasets, not only for Japanese but also for other languages. The second trial is the construction of a fake news collection system from social media posts without relying on fake news detection datasets. Many publicly available and non-manual fake news detection systems mainly consist of claim matching, which verifies whether a suspicious post has already been mentioned in a fact-checked article or a trained fake news detection model. However, it is difficult to build a collection system in the context of Japanese language resources, where there are few fact-checked articles and no dataset for training a fake news detection model. Therefore, we developed a fake news collection system to utilize users, called "guardian," who indicated the possibility of false news stories by reply messages on social media. We discuss our fake news annotation scheme for Japanese fake news dataset construction in Section 4.2 and building a fake news collection system in Section 4.3.

4.2. Fake News Annotation Scheme

4.2.1. Motivations

Researchers have been working on tasks such as fake news detection to combat social problems caused by the spread of fake news. The important task of fake news detection aims to classify whether the spreading news content is false based on news articles and social media posts related to it. Additionally, many fake news datasets have been constructed as resources to facilitate this task, such as FakeNewsNet [92], Twitter16 [120], and CoAID [135]. These existing studies on fake news detection and the corresponding dataset construction have focused nearly exclusively on the factuality aspect of the news – **Can we fully understand "fake news" and various events it causes based on these datasets provided factuality labels?** This is primarily the motivation behind our work. To promote an understanding of fake news, we consider it necessary to provide not only factual information, but also information from various perspectives, such as the intention of the false news disseminator, the harmfulness of the news to our society, and the target of the news.

We propose a novel annotation scheme to capture the various perspectives of false news, which is based on our investigations into the definition of "fake news" and existing fake news detection datasets. We annotated each news story and its social media posts using the following points: (1) factuality; (2) intention of the disseminator; (3) target; (4) method of reporting the target; (5) purpose; (6) potential harm to society; and (7) types of harm. These annotations from various perspectives are useful for facilitating an in-depth understanding of fake news, which is a complex phenomenon. For example, it would be interesting to consider how replies change when the disseminator knows whether the news is false. The annotations also provide significant value for real-world applications, such as building a fake news detection system that reveals the potential dangers of false information for journalists, fact-checkers, policymakers, and government entities.

We then constructed a Japanese fake news dataset according to the annotation scheme. To the best of our knowledge, this is the first such attempt in Japan. The construction of this dataset will facilitate our understanding of how fake

news spread in Japan. In the future, we plan to apply this method to other fake news datasets in English and other languages. Applying our annotation scheme to fake news in multiple countries and comparing the results is expected to enable further detailed analysis of fake news.

4.2.2. Issues in Existing Fake News Detection Datasets

Many datasets have been constructed for the fake news detection task, which assesses the truthfulness of a certain piece of news from news content or social media posts. We examined 51 fake news detection datasets and identified four issues that needed to be resolved. Our examination of existing fake news detection datasets is described in Appendix A.1.

We identified four issues that need to be resolved.

Intention] Even though many studies adopt a narrow definition of fake news that considers the intent of the disseminators, all datasets have labels based on the broad definition that focus only on the factual aspects of each news item, not on the intention. This situation implies a divergence between the definition of fake news in technological development and its original narrow definition of fake news. We consider that most fake news detection models built on existing datasets should be called "false information detection models." Additionally, news created with malicious intent is more persuasive than that created without such aims, and malicious users typically participate in the propagation of false news to enhance its visibility on social media [136]. Therefore, it is necessary to annotate the intentions of news disseminators to develop a highly explainable detection model.

Harmfulness to society Fake news may have a greater or lesser detrimental effect on society. For example, parody news that is clearly false is less harmful to society; however, false news about elections or COVID-19 vaccines is very harmful owing to its strong influence on people's decision-making. This perspective is not reflected in most existing datasets. A dataset called COVID-Alam [137] annotated each COVID-19 fake news item with its degree of harm to society. We consider that it would be useful for the decision on the priority of fact-checking to make a detailed annotation of various

types of news, not only COVID-19 news. For example, it is important to consider which aspects of society can be harmed by the news and the amount of impact.

Languages The linguistic characteristics and diffusion patterns of fake news vary according to country and language. However, the language included in most fake news datasets is English, and they primarily focus on US society. This is because although there is a growing awareness that fact-checking is an important action worldwide, there are still only a few fact-checking organizations with an adequate workforce, which forms the basis for dataset construction, in countries other than the US. However, fake news detection datasets in languages other than English have also increased owing to the global infodemic caused by COVID-19. Of the 51 datasets that we examined, 11 included languages other than English, and eight of them were COVID-19 datasets. The construction of a non-English fake news dataset that targets various topics leads to the analysis of fake news across languages and the identification of unique non-variant characteristics that are independent of language.

Labels A total of 33 datasets out of 51 are assigned a binary label, fake or real, because binary classification makes machine learning models easier to apply. Other datasets have fine-grained labels, typically more than two labels; however, the criteria vary across datasets based on the rating provided by fact-checking sites; for example, Politifact has six labels (true, mostly true, half true, mostly false, false, and pants on fire) and Snopes primarily has five labels (true, mostly true, mixture, mostly false, and false). Such variation in the categorization criterion in each dataset confuses the dataset users. Furthermore, related to the aforementioned issues of "intention" and "harmfulness to society," a fine-grained and consistent annotation scheme is required to develop a more general and robust fake news detection model.

4.2.3. Annotation Scheme

4.2.3.1. Instructions

We present an annotation scheme developed through careful discussion and insights gained from an examination of existing datasets. This section describes the key questions in our annotation. Q1–Q5 are aimed at constructing a more fine-grained labeling compared to the binary labeling in existing datasets or the rating provided by fact-checking sites. These questions primarily cover "intention" and " label " issues in existing datasets. Q6 and Q7, which are extensions of COVID-Alam [137] applied to general news, attempt to identify the harmful effects on society, related to the second issue "Harmfulness to society". We asked the annotators to answer these questions based on fact-checking articles and original texts. The reclassification of false news using a shared annotation scheme with fine-grained labeling can achieve a common framework for understanding false news that is independent of the rating of various fact-checking sites. This is also useful in building detection models that are highly interpretive.

Q1: What rating does the fact-checking site assign to the news? This is a very simple question, and can be answered by simply searching for the corresponding fact-checking site. This also plays a role in removing inappropriate annotators. Annotators generally choose a rating between true and false, and the options vary depending on the fact-checking site. If the annotators select true or half-true, this implies that they automatically skip subsequent questions (Q2–Q7) that are asked only about false news.

Q2-1: Does the news disseminator know that the news is false? This question asks for a subjective judgment. It covers "intention," which is an issue in existing fake news detection datasets. We asked the annotators to determine whether the spread of fake news was intentional and classified their responses into four categories based on their observations of fact-checking articles and original social media posts. If they select yes, the news can be considered fake news following the narrow definition; note that we call them "disinformation" based on previous research [27]. However, we cannot definitively classify news as fake because it may be satire or parody. If they select no, the disseminator does not intend to spread false news, which we call "misinformation." Moreover, these

decision branches are differentiated according to the degree of the annotator's belief, which is the distinction between "definitely" and "probably." Such labeling of the intention may reveal the difference in users' behavior for each type of false information, such as the type of false information people spread without knowing that it is false. This is also important, irrespective of the study's use of a broad or narrow definition. The possible answers to Q2-1 are as follows:

1. *Yes, the news disseminator definitely knows that the news is false (Disinformation)*
2. *Yes, the news disseminator probably knows that the news is false (Disinformation)*
3. *No, the news disseminator probably does not know the news is false (Misinformation)*
4. *No, the news disseminator definitely does not know that the news is false (Misinformation)*

In addition, we set the following questions regarding the type of news depending on the selection of Q2-1:

Q2-2A: If yes (disinformation), how was the news created? This question is designed to annotate how intentionally disseminated news is created. As a result of our detailed discussion and analysis of a previous study [61], we observed that each intentionally spread news story can be classified based on one of the four categories: fabricated content, manipulated image, manipulated text, and false context. First, these news stories can be categorized as either completely created news or news created by falsifying the original resources. We call the former "fabricated content." The latter can be divided into three classes depending on the object of falsification: "manipulated image" refers to content that has been manipulated for an image or video, "manipulated text" refers to content that has been manipulated for news text or social media messages related to the news, and "false context" refers to content that is shared with false contextual information despite the content being genuine.

1. *Fabricated content*
2. *Manipulated image*
3. *Manipulated text*
4. *False context*

Q2-2B: If no (misinformation), how does the disseminator misunderstand the news? We label why the disseminator has spread the false news with no intention. This is an important annotation for us to consider that prevents the future spread of false news. Similar to Q2-2A, we observed that the reasons for spreading false news with no intention can be classified into three categories: trusting other sources, inadequate understanding, and misleading. The first category refers to trusting information from other sources. This frequently occurs when non-native English speakers mistranslate English articles and research papers or trust false information originally disseminated in English. The second category refers to inadequate understanding and uncertain assumptions made by the disseminator. This may be caused by the disseminator not having thoroughly read the news. The final category refers to the case in which the disseminator may adequately understand the news, but insufficiently convey it to the reader; that is, it refers to representing information in a misleading way.

1. *Trusting other sources*
2. *Inadequate understanding*
3. *Misleading*

Q3: At whom or what is the false news targeted? The main target of false news, namely the target that is primarily affected by the fact that the news is false, is useful information for news clustering and retrieval. The task of identifying such information has not yet been completed. However, we believe that it will be an important task to promote the understanding of fake news in the future. To enable the application of information-extraction techniques, we provide the annotators with the following instructions: "Extract the targets that are primarily affected by the fact that the news is false from the claim sentences on fact-checking sites or the original social media post about the false news (multiple extractions possible)."

Q4: Does the news flatter or denigrate the target? We annotated the stance of the news towards the target, that is, flattery or denigration. Even within the category of false news, the reader's impression of good behavior news, such as donations, is very different from that of bad behavior news, such as criminal acts, even though the target has not actually performed either act. This annotation provides important information for understanding the impact of fake news on

society, particularly for analyzing the impact of fake news on polarization. The annotations are as follows:

1. *Flattery*
2. *Denigration*
3. *Neither / No such intention*

Q5: What is the purpose of the false news? Just as some news media lean towards liberal or conservative views and report the news accordingly, some false news stories are fabricated with the intention of spreading the disseminator's own theory; for example, the COVID-19 vaccine is dangerous to human health. Although the purpose of some false news items cannot be inferred, we set the following categories for false news purposes: The first category is satire or parody news for entertaining or criticizing readers [49]. These false news stories are not commonly referred to as fake news. The second is partisan news, which is extremely one-sided or biased news with a political context. Biasing in itself does not mean that the news is fake; however, some studies [52,53] report that partisan news is highly likely to be false. This annotation is important to understand the relationship between partisan news and false news. The third is propaganda, a form of persuasion that attempts to influence the emotions, attitudes, opinions, and actions of specified target audiences for ideological, religious, and other purposes [56]. Propaganda may also include political purposes in general; however, we instructed the annotators to categorize propaganda with political purposes in the partisan category to aid distinction. This question is expected to clarify the relationship between false news and the following categories.

1. *Satire / Parody*
2. *Partisan*
3. *Propaganda*
4. *No purpose / Unknown*

Q6: To what extent is the news harmful to society? This is a particularly subjective question. The purpose is to identify news stories that can negatively affect society, including specific people and companies. Specifically, we asked the annotators to indicate their degree of harm to society on a real scale of 0–5. A score of 0 indicates that the news poses no harm to society, such as satire or parody news. A score of 5 indicates that the news is definitely harmful to society.

To obtain the annotators' answers, which did not vary greatly, we asked them to label the degree of harm using a combination of two perspectives: how much truth is in the text description and how much damage may be caused by believing the news.

Q7: What types of harm can the news cause? This question helps us to understand what types of harm the news causes or has the potential to cause. We set up seven categories of harm that fake news can cause, and added the option "not sure" for cases in which a decision cannot be made. The categories are described below. Some news stories may be aligned with more than one category; however, we asked the annotators to choose the category that they considered the most appropriate.

1. *Harmless (e.g., Satire / Parody)*
2. *Confusion and anxiety about society*
3. *Threat to honor and trust in people and companies*
4. *Threat to correct understanding of politics and social events*
5. *Health*
6. *Prejudice against country and race*
7. *Conspiracy Theory*
8. *Not sure*

4.2.4. Japanese Fake News Dataset

4.2.4.1. Original Data

To construct the Japanese fake news dataset following our annotation scheme, we first collected verified articles published in Fact Check Initiative Japan [134]. We targeted the news that was spread on Twitter via these verified articles and further searched for posts or news articles that triggered the spread of false information. We asked the annotators to annotate 307 news stories that were verified by the Fact Check Initiative Japan between July 2019 and October 2021. Examples of these annotations are shown in Figures 4.1 and 4.2.

Annotation

Q1: What rate does the fact-checking site label to the news?	False
Q2-1: Does the news disseminator know that the news is false?	1. Yes, the news disseminator definitely know the news is false
Q2-2A: If Yes, how was the news created?	2. Manipulated image
Q3: At Who or what is the false news targeted?	立憲民主党，枝野幸男
Q4: Does the news flatter or denigrate the target?	2. Denigration
Q5: What the purpose of the false news?	2. Partisan
Q6: To what extent is the news harmful to society?	4
Q7: What types of harm do the news have?	3. Threat honor and trust in people and company

Figure 4.1.: Original tweet and the corresponding fact-checking article of its target for our annotation are shown on the left. The labeled information from our annotation is shown on the right-hand side. The targeted content is a video of a party debate attached by a social media influencer on Twitter. It is stated in the fact checker's judgment that this video creates a bad impression of the opposition leader (Mr. Edano) because it omits parts of the debate.

4.2.4.2. Annotation

As a pilot annotation, four annotators independently annotated 20 examples and attempted to resolve cases of disagreement in a meeting. Based on the discussions, the annotation scheme and guidelines have been refined. Finally, we asked three annotators to answer the question introduced in Section 4.2.3.1, regarding 307 verified news stories, by checking the verification articles, triggered posts, and news articles. In the annotation process, we calculated the inter-annotator agreement using Fleiss kappa. The Fleiss kappa was generally high for each question. For example, it was greater than 0.8 for Q2-1 and greater

Annotation

Q1: What rate does the fact-checking site label to the news?	Inaccurate
Q2-1: Does the news disseminator know that the news is false?	4. No, the news disseminator definitely know the news is false
Q2-2B: If No, how does the disseminator misunderstand the news?	3. Misleading
Q3: At Who or what is the false news targeted?	牡蠣
Q4: Does the news flatter or denigrate the target?	3. Neither/No such intention
Q5: What the purpose of the false news?	4. No purpose/Unknown
Q6: To what extent is the news harmful to society?	4
Q7: What types of harm do the news have?	5. Health

Figure 4.2.: Targeted content describes how to eat oysters for the prevention of food poisoning. The fact checker notes that the method prescribed for eating does not reduce the likelihood of food poisoning.

than 0.7 for Q4. Also, even Q7, where eight options existed, it was greater than 0.6. By contrast, for Q2-2A and Q2-2B, which were subjective questions, it was approximately 0.5. Note that kappa values of 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.0 correspond to fair, moderate, substantial, and perfect agreement, respectively [138]. The annotation agreement varies depending on the label of the question, particularly Q7, regarding the types of harm. For example, there was a high rate of agreement for labels 5. "Health" and 6. "Prejudice against national and racial," whereas there was a low rate of agreement for labels of 3. "Threat to honor and trust in people, company, and good," which is frequently confused with the labels of 4. "Threat to correct understanding of politics and social events."

4.2.4.3. Data Statistics

Table 4.1 shows relevant statistics on the annotations. Q1 shows the distribution of the fact-checking judgment for each news story. Most articles selected by Japanese fact-checking organizations for verification are false stories. Thus, the selection of articles by Japan's fact-checking organization was biased (only five news stories were true). For Q2-1, the labels "disinformation" and "misinformation" were applied to 13% and 87% of news stories, respectively. In most cases, the disseminator was unaware that the news was false. The class distribution of Q2-2A, for which only the news stories labeled as disinformation were annotated, was relatively balanced. In Q2-2B for misinformation, "inadequate understanding" accounts for approximately half of the annotations. For Q4, which asks whether the news flattens or denigrates the target, the distribution is skewed toward "denigration" in 60% of the news stories. This suggests that most false news is written to discredit people. For Q5, which asks the purpose of the false news, most news stories are labeled as "no purpose / unknown." Propaganda and partisan false news were identified in approximately 20% of news stories each. For Q6, the extent to which the news was harmful to society, the annotators chose average scales of 1–2 and 2–3 for many false news stories from a range of 0–5. For Q7, which asks what types of harm the news has, the majority of news stories are labeled as "threat to honor and trust in people and companies" (36%). Most news stories labeled as "health" are related to COVID-19. The "harmless" and "conspiracy theory" labels only constitute a small percent of new stories. Our fine-grained annotations can be a useful tool for understanding false news trends in a target country.

In addition to the annotation results, we collected posts and related context information on 186 news stories from Twitter, which triggered the spread of false information. The data collected from Twitter included 471,446 tweets (2,534 tweets per news story), 277,106 users (1,489 users per news story), and 17,401 conversations (93 conversations per news story). We have published these annotation results, collected tweet IDs, fact-checked articles, and other related information in `https://zenodo.org/record/5831617`.

(a) Disinformation (Labeled 1 or 2) (b) Misinformation (Labeled 3 or 4)

Figure 4.3.: Word cloud for "Q2-1: Does the news disseminator know that the news is false?"

4.2.5. Analysis of the Japanese Fake News Dataset

Our dataset includes multi-dimensional information related to news content and social context. We provide some preliminary quantitative analyses to illustrate the characteristics of the dataset. In general, we analyze news stories from the perspectives of true and fake. However, our dataset includes few news stories labeled as true; therefore, this section primarily focuses on news stories from the perspectives of misinformation and disinformation obtained from Q2-1. Note that the analysis does not cover most fake news in Japan, but only news verified by fact-checking organizations. Thus, a bias in the news stories may exist.

4.2.5.1. Tweet Contents

We aim to understand what news story topics are spread in each category. Therefore, we created a word cloud from the content of each tweet that spread most extensively on Twitter according to each news story. From Figure 4.3, we can observe the contents of the disinformation and misinformation in news stories based on the labels used for Q2-1. The word "日本 (Japan)" is prominent in both word clouds. In particular, in Figure 4.3(a), the word cloud for news stories labeled disinformation contains the words "コロナ (coronavirus)" and "ワクチン (vaccine)," which are related to COVID-19. However, the word cloud for news stories labeled misinformation in Figure 4.3(b) contains the words "中国 (China)," "バイデン

(a) 1. Trusting other sources (b) 2. Inadequate understanding (c) 3. Misleading

Figure 4.4.: Word cloud for “Q2-2B: If no (misinformation), how does the disseminator misunderstand the news?”

(a) 1. Flattery (b) 2. Denigration

Figure 4.5.: Word cloud for “Q4: Does the news flatter or denigrate the target?”

(Mr. Biden),” and “トランプ (Mr. Trump),” which are related to names of foreign people and countries. We also explored the contents from the categories of misinformation, trusting other sources, inadequate understanding (Figure 4.4(b)), and misleading (Figure 4.4(c)) based on the label in Q2-2B. Figure 4.4(a), word cloud about news stories labeled trusting other sources, shows the words “コロナ (coronavirus),” “バイデン (Mr. Biden),” “トランプ (Mr. Trump),” and “票 (vote)” in large print. This suggests that many Japanese users trusted and spread news stories about foreign events such as the US presidential election and COVID-19. Figure 4.4(b), word cloud about news stories labeled inadequate understanding,

shows the words “バイデン (Mr. Biden)” and “中国 (China)” Similar to trusting other sources, they may spread content without fully understanding foreign events. Figure 4.4(c), word cloud about news stories misleading, shows the words “中國 (China)” and “台灣 (Taiwan)” expressed in Chinese related to events in East Asia.

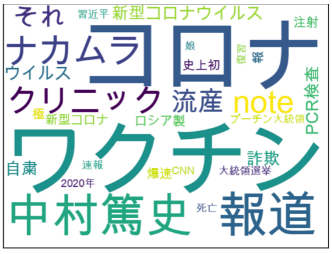
Figure 4.5 are the word clouds about news stories based on Q4, which asks whether the news flatters or denigrates the target. Figure 4.5(a), word cloud about news stories labeled flattery, shows the words “大統領選 (the US presidential election)”, “投票 (election)” and “郵便 (a part word of mail-in-ballot election)” about the US presidential election. The US presidential election was a hot topic in Japan, where a lot of false news was spread. This suggests that there were false news stories that flattered Trump’s movement. Figure 4.5(b), word cloud about news stories labeled denigration, shows the words “中国 (China)” and “中國 (China in Chinese)”. This suggests that there have been many false news stories denigrating Japan’s neighbors, such as China.

Figure 4.6 shows the word clouds for news stories based on Q7, which asks what types of harm the news can cause. Figure 4.6(a), which exhibits a word cloud on news stories labeled as confusion and anxiety about society, has the words “感染 (infection),” “死亡 (death),” and “副作用 (side effects),” which describe anxiety about COVID-19 and its vaccines. Figure 4.6(b), which displays a word cloud on news stories labeled as a threat to honor and trust in people and companies, includes the words “大阪市 (Osaka city, a regional city in Japan).” This suggests that false news stories about local elections and the Osaka government can be attributed to this category. Figure 4.6(c), which shows a word cloud on news stories labeled as a threat to the correct understanding of politics and social events, includes the words “中国 (China),” “バイデン (Mr. Biden),” and “トランプ (Mr. Trump).” This indicates that news stories that promote a false understanding of foreign events in China and the US are spreading. Figure 4.6(d), which displays a word cloud on news stories labeled health, includes the words “コロナ (coronavirus),” “ワクチン (vaccine),” and “クリニック (clinic),” which are related to COVID-19 events. Figure 4.6(e), which exhibits a word cloud on news stories labeled as prejudice against country and race, has the words “中国 (China),” “中國 (China in Chinese),” and “中国人 (Chinese person).” This suggests that many

(a) 2. Confusion and anxiety about society

(b) 3. Threat to honor and trust in people and companies

(c) 4. Threat to the correct understanding of politics and social events



(d) 5. Health

(e) 6. Prejudice against national and racial

(f) 7. Conspiracy theory

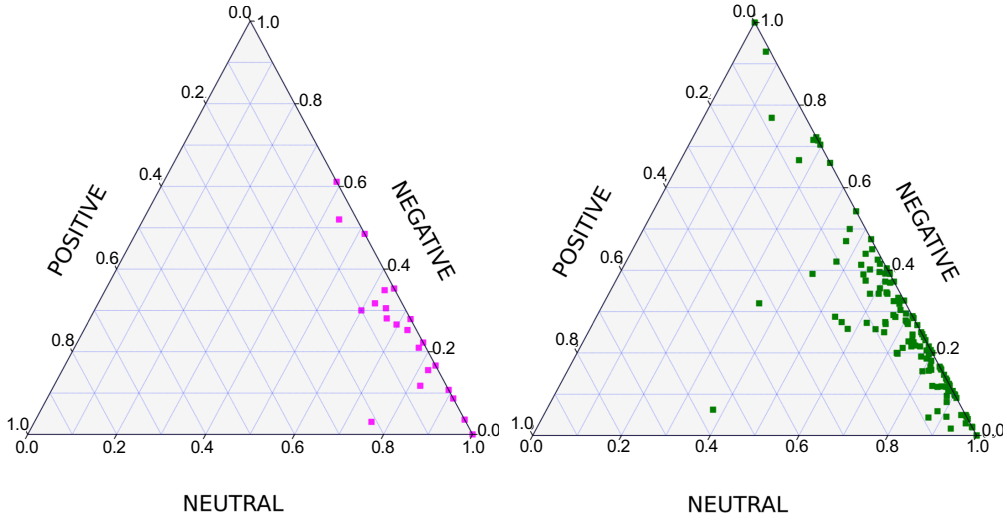
Figure 4.6.: Word cloud for “Q7: What types of harm can the news cause?”

false news items may cause prejudice against China. Figure 4.6(f), which shows a word cloud on news stories labeled conspiracy theory, includes the words “ビッグ (big)” and “発表 (announcement).” It seems that these words are often used when people want to spread conspiracy theories.

4.2.5.2. Sentiment of Responses

People express their emotions or opinions regarding false news through social media posts, such as skeptical opinions and sensational reactions. These features are important signals in the study of false news in general [32, 139].

We performed sentiment analysis on replies to user posts that spread false news using the sentiment classification API in Amazon comprehend [140], which leverages a pretraining language model. This API classifies emotions from the input

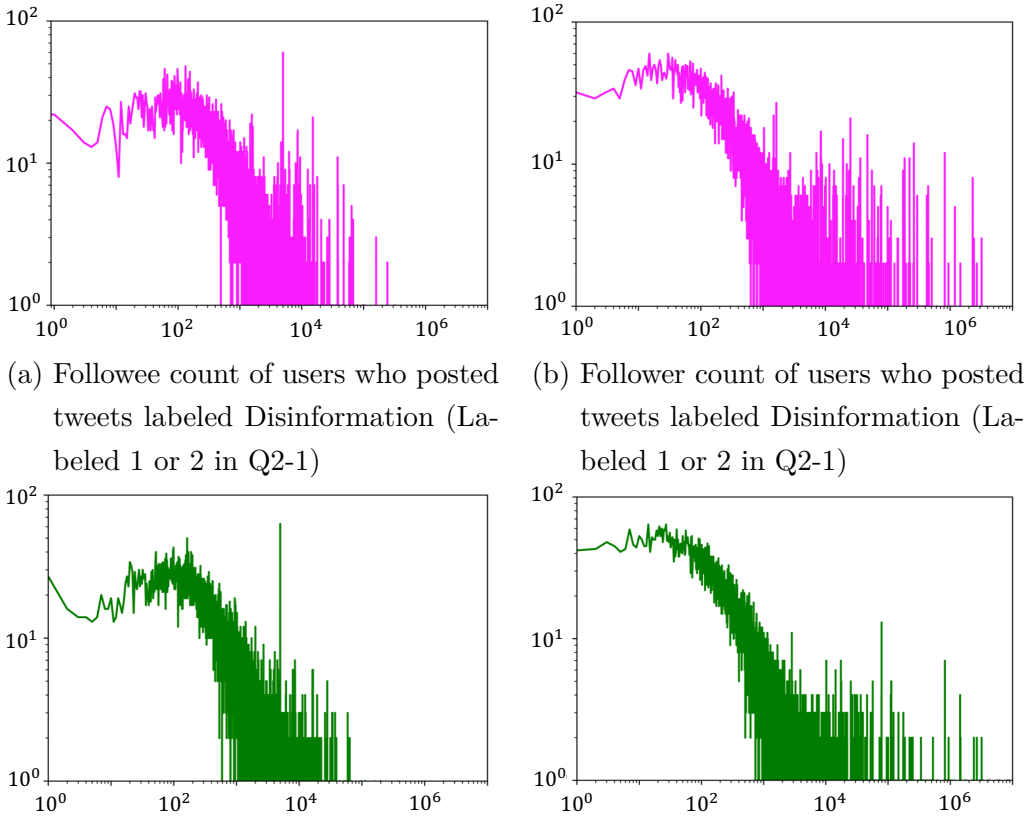


(a) Sentiment about disinformation (Labeled 1 or 2)

(b) Sentiment about misinformation (Labeled 3 or 4)

Figure 4.7.: Ternary plot of the ratio of the positive, neutral, and negative sentiment refers to tweets related to news labeled as disinformation and misinformation in "Q2-1: Does the news disseminator know that the news is false?"

text into four categories: positive, negative, neutral, or mixed. Figure 4.7 shows the relationship between the positive, neutral, and negative replies to news stories from the perspectives of disinformation and misinformation obtained from Q2-1. It represents the ratio of sentiments (positive, negative, or neutral), which are predicted based on all replies to the related tweets of each news story. The ternary plots of both disinformation and misinformation show that most replies to each news item were neutral instead of emotional responses. An analysis of the emotional replies shows that although some news stories labeled as misinformation had a high ratio of positive replies, most news stories had more negative replies. It is suggested that false news is likely to cause negative emotions, regardless of misinformation and disinformation.



(a) Followee count of users who posted tweets labeled Disinformation (Labeled 1 or 2 in Q2-1)

(b) Follower count of users who posted tweets labeled Disinformation (Labeled 1 or 2 in Q2-1)

(c) Followee count of users who posted tweets labeled Misinformation (Labeled 3 or 4 in Q2-1)

(d) Follower count of users who posted tweets labeled Misinformation (Labeled 3 or 4 in Q2-1)

Figure 4.8.: Distribution of the follower and followee count related to tweets labeled as disinformation or misinformation. The X-axis represents the follower/folowee count and the Y-axis represents the number of users.

4.2.5.3. User Profiles

We aim to analyze the users who spread false information. False news dissemination processes and user information are effective for fake news detection and for understanding the formation of an echo-chamber cycle [141], as mentioned in Chapter 3.

Figure 4.8 shows the distribution of the count of followers and followees of 20,000 users, randomly selected from users who posted news stories labeled as disinformation or misinformation. Users who spread disinformation often have

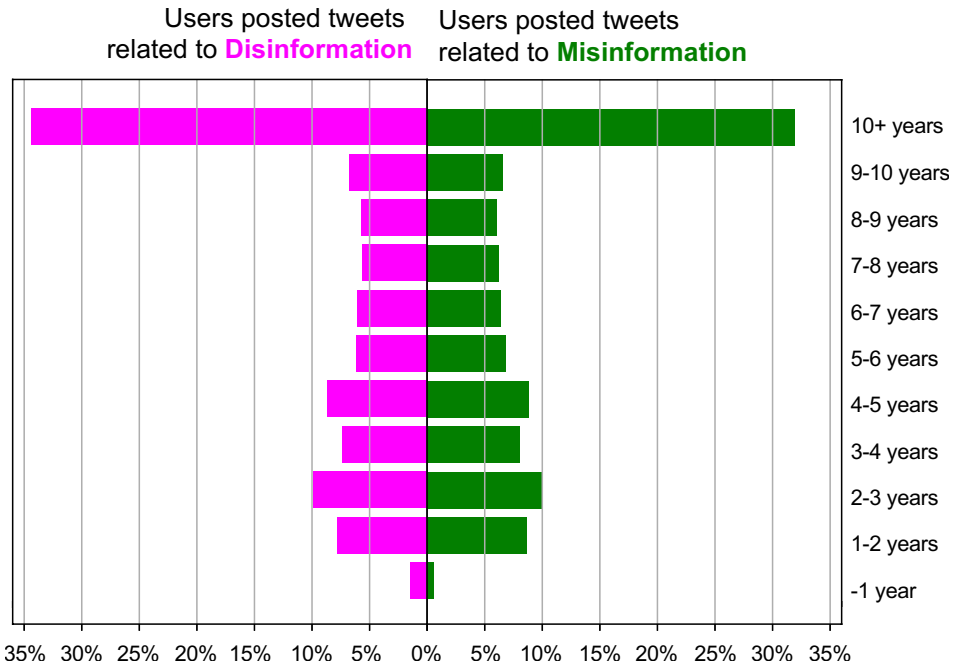


Figure 4.9.: Distribution of the time that elapsed since the user account creation date from two perspectives: disinformation and misinformation.

more followers than those who spread misinformation. The follower and followee counts of users generally follow a power-law distribution, which is commonly observed in social network structures. There is a spike of approximately 5,000 in the followee count distribution for both owing to Twitter restrictions [142].

Figure 4.9 shows the distribution of the time that has elapsed since each user created their account. The distributions for disinformation and misinformation are similar. However, when compared with reports on the distribution of users who spread false news in the US [92], our results exhibit two features. One is that few users have created accounts less than a year ago. Another feature is that users who had been using Twitter for more than ten years accounted for a large portion of disinformation and misinformation disseminators. We believe that these characteristics are because of the fact that social media "bot accounts" are less active in Japan than in the US.

Finally, we investigated the ratio between "bot accounts" and human users that were involved in tweets related to misinformation and disinformation. We

randomly selected 10,000 users from each category and performed bot detection using the Botometer API [143]. As a result, the ratio of "bot accounts" to human users is similar in the two categories: approximately 8% for disinformation and 6% for misinformation. However, a comparison of reports on the ratio of bot users that spread false news in the US [92] and Japan shows that there are fewer bot users in Japan. Specifically, approximately 22% of users that disseminate false news are bots in the US, whereas the corresponding percentage for Japanese users is less than 10%.

4.3. Japanese Fake News Collection System

4.3.1. Motivation

Fact-checking organizations verify suspicious news stories using domain experts to combat the growing amount of fake news on social media. However, such verification is highly reliable, it is burdened by time-consuming and labor-intensive tasks and requires several days from the start of the spread of the news to the time of verification. Because fake information on social media spreads rapidly and widely, it is necessary to detect the spread at an early stage.

As an assistant to the fact-checking organizations, some online tracking systems, that automatically detect and collect fake news by machine learning methods have been developed, such as NewsGuard [129] and Hoaxy [100]. Although these tracking systems play a crucial role in gathering fake news, they have been developed primarily for English. This is attributed to the fact that most of the resources, such as fact-checking articles and fake news detection datasets, which are utilized to build the tracking tools, are in English. In other words, it is difficult to build systems that target any language that do not contain fact-checked articles or detection datasets.

We then pay attention to the "guardian," who performs the fact-checking intervention toward posts of uncertain truth, to address the aforementioned issues. Figure 4.10 shows an example of fact-checking activity by guardian. There is no certainty that Guardians ' posts are true, but they have the potential to find false news faster and more than fact-checking organizations. We developed a fake

Figure 4.10.: Example of fact-checking activity by guardian.

news collection system, "fake guardian, " which uses guardian as a social sensor to collect information that may be false on Twitter, targeting Japanese tweets. Because our system can be developed without resources, such as fact-checking articles, it can be applied to other low-resource languages.

4.3.2. Related Work

4.3.2.1. Fake Tracking Tools and Systems

Several tools and systems have been developed to track the movement of fake news to instantly check for its veracity and investigate its dissemination. Hoaxy [100] is a framework for collecting and tracking fact-checking information and related misinformation. Users can search for topics in which they are interested and check the diffusion visualization of the respective topics. The FakeNewsTracker [144] is a system for fake news data collection, detection, and visualization on social media. These systems collect verified fake news sources from fact-checking organizations and find posts that match these sources. In addition, these systems are utilized to develop fake news datasets: Hoaxy dataset [145], which has been accumulated using Hoaxy, consists of retweeted messages with links to either fact-checking or misinformation articles, FakeNewsNet [92], constructed using FakeNewsTracker, contains various types of information such as news content, and spatio-temporal and social contexts.

There are other representative systems and tools as follows: NewsVerify [146], a real-time news certification system, starts to track news after user input and detects the credibility of events from Sina Weibo. NewsGuard [129] provides the credibility and transparency of news website scores, which sum to 100, by assessing a group of trained journalists. The credibility of the sites visited is informed by the extensions provided for major browsers. TweetCred [147] is a real-time web-based system that assesses the credibility of content on Twitter and is available as a browser extension. The system provides a score of credibility for each tweet based on previously generated classifiers, and it validates this score by asking for user feedback. B.S. Detector [148], which is also provided as a browser extension tool, searches all links on a given web page for references to unreliable sources, checking against a manually compiled list of domains, OpenSources [149]. Truthy [150], Rumorlens [151], and Twitter trails [152] are tools for detecting and displaying the diffusion of misinformation based on a semi-automatic approach, where users can explore the propagation of posts with an interactive dashboard. ClaimBuster [153] searches for suspicious posts and then explores whether it has already been fact-checked or searches for external knowledge bases to verify

whether it is correct. XFake [154] visualizes the attribute information of false news detected by a trained fake news detection model to facilitate interpretation. FotoForensics [155] is a tool for image manipulation detection that analyzes the distribution of the image compression levels.

4.3.2.2. Guardian

Recently, "fact-checking intervention," in which social media users cite fact-checking websites and reply to fake news spreaders, has become a useful strategy to mitigate the spread of fake news [156]. It is helpful for verification during the fact-checking process. This finding encourages research toward the "guardian," where a user performs fact-checking intervention. For example, Aniko et al. [157] analyzed the acts of snooping, friends, followers, the person followed, or strangers on Twitter. Nguyen et al. [158, 159] developed a fact-checking URL recommendation model to encourage guardians to engage more in fact-checking activities. Our proposed system is an extension of Zhao et al.'s findings [156] into a practical system. They investigated the possibility of collecting rumors spread on social media using inquiry phrases such as *"Really?"*.

4.3.3. Fake Guardian: Japanese Fake News Collection System

We first present an overview of the proposed system. Then, we introduce details of the respective components in our system. This system is publicly available online in Japanese at `https://aoi.naist.jp/fakeguardians`.

4.3.3.1. Overview

We describe our system from the backend to the frontend. Figure 4.11 presents an overall picture of the system framework. The backend extracts fake news from guardian tweets in four steps: **Crawling**, **Noise-removal**, **Grouping**, and **Ranking**. The frontend displays daily fake news and receives feedback through the voting function from users on whether guardians' point to fake news is correct or not.

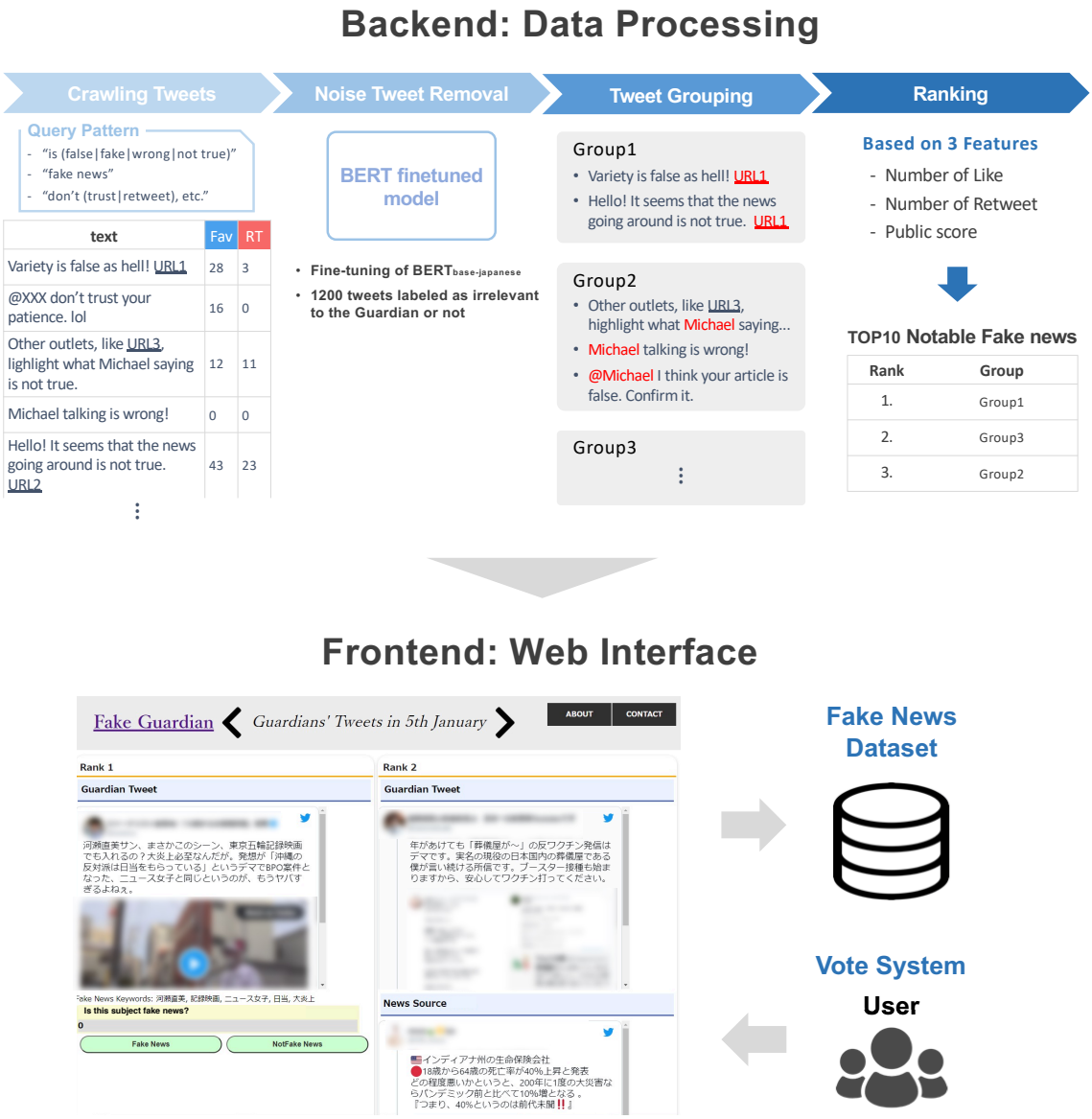


Figure 4.11.: Overview of Fake Guardian: Our system extracts noteworthy guardian tweets against fake news from Twitter on the backend and shows the processed data on the frontend.

4.3.3.2. Backend

We organize and rank the crawling data for ease of checking. This step, in turn, has four steps: crawling, noise removal, grouping, and ranking. Processing all

collected tweets is time-consuming. Therefore, we only use tweets that have more than three shares. These steps are applied every day on one day of tweets.

- **Crawling** We need to find the debunking patterns used by guardians as search keywords before collecting guardians' tweets. The usage patterns were selected based on our observations of discussions related to fake news on Twitter. Our selected patterns are as below: "は (FAKEWORDS)", "(FAKEWORDS) です", "(FAKEWORDS) である", "という (FAKEWORDS)", and "(信じ|拡散し) ない", where FAKEWORDS indicates each of these words: デマ, フェイク, 間違い, 不正確, 誤報, 虚偽, 事実無根. The crawling is executed continuously and the collected tweets are saved in our database.

- **Noise Removal**

We collected tweets containing certain keywords, which included those not generated by guardians. Our keywords are selected to achieve a low false-negative rate, which means collecting as many of the guardians' tweets as possible to avoid missing them. By contrast, the false positive rate is high, which means that irrelevant and noisy tweets are collected. Therefore, we applied the trained model to remove irrelevant and noisy tweets from the collected tweets. The model was constructed by fine-tuning the pretrained model, $\text{BERT}_{\text{base-Japanese}}$ [160] with 10 epochs using 1,200 tweets as training data, which consists of posts collected by our selected patterns between September 2019 and March 2020 and are labeled as irrelevant to guardians. Specifically, 660 tweets were labeled as irrelevant and 540 tweets were labeled as relevant. The constructed model was evaluated using 300 tweets (165 tweets and 135 tweets), collected and labeled in the same way as the training data, and yielded an accuracy metrics of 0.87. Our system aims to collect guardian tweets with a low false positive rate by applying the constructed model.

- **Grouping**

This step was designed to gather tweets referring to the same event cluster

within the same group. It is difficult to apply supervised machine-learning-based methods for grouping because the types of tweets vary every day. We then executed a simple and robust rule-based grouping method using the extracted suspicious event phrases and other features, such as the URL. The rules of grouping are presented below:

1. Set tweets with the same URL into the same group
2. Set tweets replying to the same tweet into the same group
3. Calculate the distance between each tweet in the previous step, using the word mover's distance (WMD) [161]. Set tweets that have fewer than threshold τ into the same group

To calculate WMD, we use word vectors from [162]. The threshold τ was set as 0.25.

- **Ranking**

In this step, we aim to rank each news story generated from the aforementioned process, allowing people to pay attention to it. Our ranking method is inspired by an unsupervised method of [163]. The method ranks each news story according to several features scores. Each of the features captures one aspect of whether the news story is noteworthy. After this calculation, the method finally outputs the ranking of noteworthy news stories according to the high average rank across all features.

Our system chooses three features for ranking: **number of likes**, **number of retweets**, and **public score**, which calculates the percentage of followers among retweet users. We assume that the news story is more noteworthy if the first two features, the number of likes and the number of retweets, are larger. And, the smaller the public score, the more noteworthy the news story. This means that the news story is more noteworthy when it spreads to people other than the spreader's followers.

4.3.3.3. Frontend

We designed an interface that displays the processed data in the backend to enable users to check the daily ranked fake news events. Additionally, we attached a voting function that asked users whether a guardian tweet indicated fake news. This function collects user feedback to create a more sophisticated system. An example is depicted in Figure 4.12.

The top 10 news stories are displayed in the proposed system. Keywords, Guardian tweet, and Voting system are shown for each story. Keywords describe the characteristic words to help understand the kind of news story. Guardian tweet shows the tweet to indicate that the news story may be false and potentially false information such as tweet content and URL that it refers to, which our system obtained through crawling. In addition, we introduced a voting system that enables users to vote on whether each event story is fake. We aim to collect feedback from users to enable this system to be more sophisticated. The system clearly shows each news story has a strong possibility of being fake using this structure.

4.3.4. Effectiveness of Our System

Confirming whether a collected guardian tweet has identified fake news is important for validating the effectiveness of the proposed system. We asked two annotators to label 122 Japanese tweets from November 1, 2021, to November 14, 2021, from the following viewpoints.

1. Do the guardian tweets point to possible falsity in the news story?

2. Are the subjects of the collected event truly fake?

Both questions comprised binary options: yes or no.

The results confirmed a substantial level of agreement; Cohen's kappa score was 0.77 for (a) and 0.70 for (b). For the tweets on which the two annotators did not agree, a third annotator (the author) labeled the tweet. The results of (a) indicate that 77% of the collected tweets indicated possible falsity in the news stories. This suggests that the selected patterns and noise-tweet removal in our system are functional. The results in (b) also indicated that approximately 52%

Figure 4.12.: Example of top news stories on November, 10th, 2021. The frontend in our system shows the ranking of the noteworthy news stories to users.

of the collected tweets were truly false. These results suggest that our system can collect a large amount of fake news that social media users pay attention to, even though the amount collected depends on the quality of the guardian.

4.4. Conclusion

We conclude by emphasizing the following points in this chapter.

- We identified issues that need to be resolved in dataset construction by the exhaustive survey of existing fake news detection datasets.
- We proposed a novel annotation scheme to capture the news from various perspectives, not only factuality, and developed the first Japanese fake news dataset using the annotation scheme.
- We developed the Japanese fake news collection system "Fake Guardian."

Our fake news collection system has the potential to be used for the extension of fake news datasets. In the future, after collecting sufficient samples for training, it can combine the findings of fake news characteristics with the fake news detection system presented in Chapters 2 and 3.

Table 4.1.: Distribution of the Japanese fake news dataset.

Q1: What rating does the fact-checking site attribute to the news?	307
True	1
Half-True	4
Inaccurate	50
Misleading	52
False	153
Pants on Fire	16
Unknown Evidence	30
Suspended Judgment	1
Q2-1: Does the news disseminator know that the news is false?	301
1. Yes, the news disseminator definitely knows that the news is false. (Disinformation)	20
2. Yes, the news disseminator probably knows that the news is false. (Disinformation)	19
3. No, the news disseminator probably does not know that the news is false. (Misinformation)	155
4. No, the news disseminator definitely does not know that the news is false. (Misinformation)	107
Q2-2A: If yes, how was the news created?	39
1. Fabricated content	15
2. Manipulated image	12
3. Manipulated text	6
4. False context	6
Q2-2B: If no, how does the disseminator misunderstand the news?	262
1. Trusting other sources	61
2. Inadequate understanding	131
3. Misleading	70
Q4: Does the news flatter or denigrate the target?	301
1. Flattery	25
2. Denigration	181
3. Neither / No such intention	95
Q5: What is the purpose of the false news?	301
1. Satire / Parody	6
2. Partisan	70
3. Propaganda	67
4. No purpose / Unknown	158
Q6: To what extent is the news harmful to society? (average)	301
$0 \sim 1$ (including 1)	17
$1 \sim 2$ (including 2)	128
$2 \sim 3$ (including 3)	112
$3 \sim 4$ (including 4)	41
$4 \sim 5$ (including 5)	3
Q7: What types of harm can the news cause?	301
1. Harmless (e.g., Satire / Parody)	6
2. Confusion and anxiety about society	41
3. Threat to honor and trust in people and companies	109
4. Threat to correct understanding of politics and social events	63
5. Health	29
6. Prejudice against country and race	42
7. Conspiracy theory	11
8. Not sure	0

5. Conclusion

In this chapter, we summarize our research results and their broader impacts, and discuss promising research directions.

5.1. Summary

In this dissertation, we addressed the challenges related to fake news for aiming to keep up with fake news in the social media ecosystem, following the steps from the foundation to application. We studied three research tasks: (1) modeling the spread of fake news on Twitter; (2) developing a fake news detection model utilizing temporal features; and (3) developing a Japanese fake news dataset and fake news collection system from Twitter.

In Chapter 2, we propose a modeling method that considers the spread of a fake news item as a two-stage process: initially, fake news spreads as a piece of ordinary news; then, when most users start recognizing the falsity of a news item, it spreads as another news story. We validated this model using two datasets of fake news items spread on Twitter. We showed that the proposed model is superior to current state-of-the-art methods in accurately predicting the evolution of the spread of fake news items. Moreover, text analysis suggested that our model appropriately infers the correction time, that is, the moment when Twitter users start realizing the falsity of the news item. Our model can contribute to understanding the dynamics of fake news spread on social media.

In Chapter 3, we propose a fake news detection model combined with a point process algorithm to utilize temporal features generated from social media posts based on the findings that the diffusion and temporal patterns of fake news are different from those of real news in Chapter 2. Furthermore, we proposed a novel multi-modal attention-based method, which includes linguistic and user features

alongside temporal features, for detecting fake news. The results obtained from three public datasets indicate that the proposed model outperforms the existing methods and demonstrates the effectiveness of temporal features for fake news detection.

In Chapter 4, we introduce the construction of a Japanese fake news dataset and fake news collection system. In the construction of the Japanese fake news dataset, we first examined existing fake news datasets to identify issues to be addressed. We proposed an annotation scheme to consider the harmfulness to society and the intention of the disseminator to solve these issues, and developed the first Japanese fake news dataset based on the scheme. In the construction of the fake news collection system from social media posts, we mainly focused on "guardian," who indicate the possibility of falsity in the news story by their reply messages on social media, and attempted to search for fake news without relying on fake news detection datasets and fact-checking articles. The system is designed to be easy to use and can extract fake news, which has not been explored by fact-checking organizations. Our system has the potential to utilize fake news characteristics and the fake news detection system presented in Chapters 2 and 3.

5.2. Future Work

Fake news research is still at an early stage and will continue to be actively developed because of the high social demand. We present some promising research directions as follows:

- **Fake news detection model with high interpretability** Many fake news detection models have been developed with the aim of improving accuracy by incorporating new architectures such as attention networks and transformers. In addition to model development, they also achieved improved accuracy by utilizing various contextual information sources, such as network information, news resources, and text styles. In Chapter 3, we show that temporal features are useful for fake news detection. However, even if a fake news detection model determines that a news item is fake, the question of whether people will believe it without evidence and explanation

arises. To convey convincing evidence for fact-checking results, we believe it is important to show not only whether the news is fake, but also the following points: from what knowledge does it determine the news is fake (use of knowledge graph); from what sources is the news manipulated; and what information do we need to understand that the news is false. The proposal of a novel model and dataset construction are important tasks for achieving highly interpretable outputs [164].

- **Bias of Fake news dataset** Reducing the bias in datasets used for fake news research is also a key issue, as increased awareness of the issue of fairness in AI has caused [165]. Some studies [166,167] investigated whether the inference of the models learned in the FNC_dataset [168] and FEVER [169] datasets for a fact-verification task are biased. They find the bias of the words in trained models; for example, most of the attention weights by the model are assigned to noun phrases, and they propose a mitigation strategy. In addition to the bias of words, there may be other biases that have not yet been confirmed in datasets related to fake news, such as author bias [170], annotator bias [171], gender bias [172], and racial bias [173]. In particular, it is important to consider political bias, that is, whether the dataset includes more data for a particular partisanship [174] because of the strong relationship between fake news and hyperpartisan news [175]. It is important to consider the existence of these biases and their mitigation.

- **Mitigation of fake news** Fake news detection is not enough to curb the spread of fake news. We need to consider methods other than detection, such as discovering disseminators and proactively blocking target users, to mitigate the effects of fake news. In addition, estimating the potential population affected by fake news is useful for decision-making regarding mitigation strategies. For example, Castillo et al. [109] estimated the number of users potentially affected by online news. Farajtabar et al. [176] proposed a representation method for the network activities of fake news using multivariate hawk processes (MHPs) with self and mutual excitations. They examined how to achieve the control incentives spontaneous events for the mitigation. Wang et al. [177] proposed the diffusion methodology,

which combats the spread of false news by proactively diffusing fact-checked information. Such strategies to mitigate the effects of fake news are being explored; however, they are still in their early stages.

- **Understanding of fake news in countries other than the US** There are many studies on how fake news spreads and what impact it has. Most of these are targeted at US events in the social media ecosystem, such as US elections [178, 179] and shooting events [180]. These studies find that many fake news items are spread by certain news sites. However, in Japan, more than half of the false news is spread from personal accounts on Twitter, as shown by the Japanese fake news dataset created in Chapter 4. This indicates that the spread and creation of fake news varies greatly from country to country. Clarifying this point will enable us to implement countermeasures against fake news that are appropriate for each country.

References

[1] Patti Domm. False rumor of explosion at white house causes stocks to briefly plunge; ap confirms its twitter feed was hacked. https://www.cnbc.com/id/100646197. [accessed on November 9th, 2021].

[2] WHO. Fighting misinformation in the time of covid-19, one click at a time. https://www.who.int/news-room/feature-stories/detail/fighting-misinformation-in-the-time-of-covid-19-one-click-at-a-time. [accessed on November 9th, 2021].

[3] Carlos Carvalho, Nicholas Klagge, and Emanuel Moench. The persistent effects of a false news shock. *Journal of Empirical Finance*, 18(4):597–615, 2011.

[4] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. Detection and analysis of 2016 us presidential election related rumors on twitter. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 14–24. Springer, 2017.

[5] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14, 2019.

[6] Mike Wendling. The saga of 'pizzagate': The fake story that shows how conspiracy theories spread. `https://www.bbc.com/news/blogs-trending-38156985`, 2016. [accessed on November 9th, 2021].

[7] Sunday Oluwafemi Oyeyemi, Elia Gabarron, and Rolf Wynn. Ebola, twitter, and misinformation: a dangerous combination? *Bmj*, 349, 2014.

[8] Yeimer Ortiz-Martínez and Luisa F Jiménez-Arcia. Yellow fever outbreaks and twitter: Rumors and misinformation. *American journal of infection control*, 45(7):816–817, 2017.

[9] Michele Miller, Tanvi Banerjee, Roopteja Muppalla, William Romine, and Amit Sheth. What are people tweeting about zika? an exploratory study concerning its symptoms, treatment, transmission, and prevention. *JMIR public health and surveillance*, 3(2):e7157, 2017.

[10] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22:100104, 2021.

[11] Chayakrit Krittanawong, Bharat Narasimhan, Hafeez Ul Hassan Virk, Harish Narasimhan, Joshua Hahn, Zhen Wang, and WH Wilson Tang. Misinformation dissemination in twitter in the covid-19 era. *The American journal of medicine*, 133(12):1367, 2020.

[12] J Scott Brennen, Felix Simon, Philip N Howard, and Rasmus Kleis Nielsen. Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 7(3):1, 2020.

[13] Adam Kucharski. Study epidemiology of fake news. *Nature*, 540(7634):525–525, 2016.

[14] Marco T Bastos and Dan Mercea. The brexit botnet and user-generated hyperpartisan news. *Social science computer review*, 37(1):38–54, 2019.

[15] The Gurdian. Chinese panic-buy salt over japan nuclear threat. `https://www.theguardian.com/world/2011/mar/17/chinese-panic-buy-salt-japan`, 2011. [accessed on November 9th, 2021].

[16] Ashley Lime. A year in fake news in africa. `https://www.bbc.com/news/world-africa-46127868`, 2018. [accessed on November 9th, 2021].

[17] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. Rumor diffusion and convergence during the 3.11 earthquake: a twitter case study. *PLoS one*, 10(4):e0121443, 2015.

[18] Takako Hashimoto, David Lawrence Shepard, Tetsuji Kuboyama, Kilho Shin, Ryota Kobayashi, and Takeaki Uno. Analyzing temporal patterns of topic diversity using graph clustering. *The Journal of Supercomputing*, 77(5):4375–4388, 2021.

[19] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79, 2010.

[20] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

[21] Elisa Shearer. More than eight-in-ten americans get news from digital devices. https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/, 2021. [accessed on November 9th, 2021].

[22] Kathleen Hall Jamieson and Joseph N Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

[23] Olivia Solan. Facebook ' s failure: did fake news and polarized politics get trump elected? https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories, 2016. [accessed on November 9th, 2021].

[24] Brett Edkins. Americans believe they can detect fake news. studies show they can't. https://www.forbes.com/sites/brettedkins/2016/12/20/americans-believe-they-can-detect-fake-news-studies-show-they-cant, 2016. [accessed on November 9th, 2021].

[25] First draft. https://firstdraftnews.org/. [accessed on November 9th, 2021].

[26] 'f*** news' should be replaced by these words, claire wardle says. https://money.cnn.com/2017/11/03/media/claire-wardle-fake-news-reliable-sources-podcast/index.html, 2017. [accessed on November 9th, 2021].

[27] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

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

[29] Victoria L Rubin, Yimin Chen, and Nadia K Conroy. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

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

[31] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[32] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[33] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

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

[35] Eni Mustafaraj and Panagiotis Takis Metaxas. The fake news spreading plague: was it preventable? In *Proceedings of the 2017 ACM on web science conference*, pages 235–239, 2017.

[36] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4, 2015.

[37] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, 2018.

[38] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.

[39] Newsweek. British government bans the phrase 'fake news'. https://www.newsweek.com/british-government-bans-phrase-fake-news-1182784, 2018. [accessed on November 9th, 2021].

[40] Adam Kucharski. Post-truth: Study epidemiology of fake news. *Nature*, 540:525–525, 12 2016.

[41] Peter Hernon. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government information quarterly*, 12(2):133–139, 1995.

[42] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602, 2016.

[43] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, 2017.

[44] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36, 2020.

[45] Cody Buntain and Jennifer Golbeck. Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215. IEEE, 2017.

[46] Warren A Peterson and Noel P Gist. Rumor and public opinion. *American Journal of Sociology*, 57(2):159–167, 1951.

[47] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

[48] Jan Harold Brunvand. *American folklore: An encyclopedia*, volume 1551. Routledge, 2006.

[49] John Brummette, Marcia DiStaso, Michail Vafeiadis, and Marcus Messner. Read all about it: The politicization of "fake news" on twitter. *Journalism & Mass Communication Quarterly*, 95(2):497–517, 2018.

[50] Satirewire. `https://www.satirewire.com/`. [accessed on November 9th, 2021].

[51] The onion. `https://www.theonion.com/`. [accessed on November 9th, 2021].

[52] Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. Kek, cucks, and god emperor trump: A measurement study

of 4chan ' s politically incorrect forum and its effects on the web. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):92–101, 2017.

[53] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014, 2018.

[54] 4chan. `https://boards.4chan.org`. [accessed on November 9th, 2021].

[55] Gab. `https://gab.com/`. [accessed on November 9th, 2021].

[56] Garth S Jowett and Victoria O'donnell. *Propaganda & persuasion*. Sage publications, 2018.

[57] Cristian Lumezanu, Nick Feamster, and Hans Klein. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[58] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230, 2008.

[59] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442, 2010.

[60] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19, 2015.

[61] Claire Wardle. Fake news. it ' s complicated. https://firstdraftnews.org/articles/fake-news-complicated/. [accessed on November 9th, 2021].

[62] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 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, 2019.

[63] Zhijing Wu, Mark Sanderson, B Barla Cambazoglu, W Bruce Croft, and Falk Scholer. Providing direct answers in search results: A study of user behavior. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1635–1644, 2020.

[64] Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9(1):1–14, 2020.

[65] Ryota Kobayashi and Renaud Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[66] Ugur Kursuncu, Manas Gaur, Usha Lokala, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. Predictive analysis on twitter: Techniques and applications. In *Emerging research challenges and opportunities in computational social network analysis and mining*, pages 67–104. Springer, 2019.

[67] Alexandru Tatar, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1):1–20, 2014.

[68] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936, 2014.

[69] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[70] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[71] Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 6–14, 2012.

[72] Julia Proskurnia, Przemyslaw Grabowicz, Ryota Kobayashi, Carlos Castillo, Philippe Cudré-Mauroux, and Karl Aberer. Predicting the success of online petitions leveraging multidimensional time-series. In *Proceedings of the 26th International Conference on World Wide Web*, pages 755–764, 2017.

[73] Naoki Masuda, Taro Takaguchi, Nobuo Sato, and Kazuo Yano. Self-exciting point process modeling of conversation event sequences. In *Temporal networks*, pages 245–264. Springer, 2013.

[74] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1513–1522, 2015.

[75] Jean-Charles Delvenne, Renaud Lambiotte, and Luis EC Rocha. Diffusion on networked systems is a question of time or structure. *Nature communications*, 6(1):1–10, 2015.

[76] Alexey N Medvedev, Jean-Charles Delvenne, and Renaud Lambiotte. Modelling structure and predicting dynamics of discussion threads in online boards. *Journal of Complex Networks*, 7(1):67–82, 2019.

[77] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be hip: Hawkes intensity processes for social media popularity. In *Proceedings of the 26th International Conference on World Wide Web*, pages 735–744, 2017.

[78] Kazuki Fujita, Alexey Medvedev, Shinsuke Koyama, Renaud Lambiotte, and Shigeru Shinomoto. Identifying exogenous and endogenous activity in social media. *Physical Review E*, 98(5):052304, 2018.

[79] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, 13(9):e0203958, 2018.

[80] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812, 2017.

[81] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.

[82] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE, 2013.

[83] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

[84] Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398, 2016.

[85] Mehrdad Farajtabar, Manuel Gomez Rodriguez, Mohammad Zamani, Nan Du, Hongyuan Zha, and Le Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In *Artificial Intelligence and Statistics*, pages 232–240. PMLR, 2015.

[86] Hridoy Sankar Dutta, Vishal Raj Dutta, Aditya Adhikary, and Tanmoy Chakraborty. Hawkeseye: Detecting fake retweeters using hawkes process and topic modeling. *IEEE Transactions on Information Forensics and Security*, 15:2667–2678, 2020.

[87] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

[88] Stephen G Nash. Newton-type minimization via the lanczos method. *SIAM Journal on Numerical Analysis*, 21(4):770–788, 1984.

[89] Scipy. `https://www.scipy.org/`. [accessed on November 9th, 2021].

[90] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[91] Camille Gontier and Jean-Pascal Pfister. Identifiability of a binomial synapse. *Frontiers in computational neuroscience*, 14:86, 2020.

[92] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 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, 2020.

[93] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2016, pages 3818–3824, 2016.

[94] Politifact. `https://www.politifact.com/`. [accessed on November 9th, 2021].

[95] Snopes.com. `https://www.snopes.com/`. [accessed on November 9th, 2021].

[96] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. Rumor diffusion and convergence during the 3.11 earthquake: a twitter case study. *PLoS one*, 10(4):e0121443, 2015.

[97] Takako Hashimoto, David Lawrence Shepard, Tetsuji Kuboyama, Kilho Shin, Ryota Kobayashi, and Takeaki Uno. Analyzing temporal patterns of topic diversity using graph clustering. *The Journal of Supercomputing*, 77(5):4375–4388, 2021.

[98] The social psychology of panic revealed by categorizing 80 post-disaster hoaxes. `https://blogos.com/article/2530/`. [accessed on November 9th, 2021].

[99] Shuai Gao, Jun Ma, and Zhumin Chen. Modeling and predicting retweeting dynamics on microblogging platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 107–116, 2015.

[100] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750, 2016.

[101] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[102] Justin Cheng, Lada A Adamic, Jon M Kleinberg, and Jure Leskovec. Do cascades recur? In *Proceedings of the 25th international conference on world wide web*, pages 671–681, 2016.

[103] Rachel Schraer. Covid vaccine: Fertility and miscarriage claims fact-checked. `https://www.bbc.com/news/health-57552527`, 2021. [accessed on November 9th, 2021].

[104] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637, 2020.

[105] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9, 2018.

[106] Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 309–312, 2009.

[107] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[108] Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17, 2016.

[109] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.

[110] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, 2017.

[111] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, 2018.

[112] Qiang Zhang, Aldo Lipani, Shangsong Liang, and Emine Yilmaz. Reply-aided detection of misinformation via bayesian deep learning. In *The world wide web conference*, pages 2333–2343, 2019.

[113] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. A convolutional approach for misinformation identification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3901–3907, 2017.

[114] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, 2019.

[115] Hamid Karimi and Jiliang Tang. Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442, 2019.

[116] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7, 2012.

[117] Yang Liu and Yi-Fang Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[118] Duc Minh Nguyen, Tien Huu Do, Robert Calderbank, and Nikos Deligiannis. Fake news detection using deep markov random fields. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1391–1400, 2019.

[119] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis*, pages 1–9, 2013.

[120] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, 2017.

[121] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[122] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3454–3466, 2018.

[123] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. Rumor detection over varying time windows. *PloS one*, 12(1):e0168344, 2017.

[124] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

[125] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[126] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1751–1754, 2015.

[127] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[128] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks

from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[129] Newsguard. `https://www.newsguardtech.com/`. [accessed on November 9th, 2021].

[130] Buzzfeed. `https://www.buzzfeed.com/`. [accessed on November 9th, 2021].

[131] Suggest. `https://www.suggest.com/`. [accessed on November 9th, 2021].

[132] Poynter. `https://www.poynter.org/`. [accessed on November 9th, 2021].

[133] Factcheck.org. `https://www.factcheck.org/`. [accessed on November 9th, 2021].

[134] Fact check initiative japan. `https://fij.info/`. [accessed on November 9th, 2021].

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

[136] Harvey Leibenstein. Bandwagon, snob, and veblen effects in the theory of consumers' demand. *The quarterly journal of economics*, 64(2):183–207, 1950.

[137] Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 913–922, 2021.

[138] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[139] Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of*

the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1589–1599, 2011.

[140] Amazon Comprehend. https://docs.aws.amazon.com/comprehend/index.html. [accessed on November 23th, 2021].

[141] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1):1–12, 2016.

[142] About following on twitter. https://help.twitter.com/en/using-twitter/twitter-follow-limit. [accessed on November 23th, 2021].

[143] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proc. of WWW*, pages 273–274, 2016.

[144] Kai Shu, Deepak Mahudeswaran, and Huan Liu. Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, 25(1):60–71, 2019.

[145] Pik-Mai Hui, Chengcheng Shao, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. The hoaxy misinformation and fact-checking diffusion network. In *Proc. of International AAAI Conference on Web and Social Media*, pages 528–530, 2018.

[146] Xing Zhou, Juan Cao, Zhiwei Jin, Fei Xie, Yu Su, Dafeng Chu, Xuehui Cao, and Junqiang Zhang. Real-time news certification system on sina weibo. In *Proc. of the International Conference on World Wide Web*, pages 983–988, 2015.

[147] Tweetcred. `https://chrome.google.com/webstore/detail/tweetcred/fbokljinlogeihdnkikeeneiankdgikg?hl=en`. [accessed on November 9th, 2021].

[148] B.s. detector. `https://github.com/selfagency/bs-detector`. [accessed on November 9th, 2021].

[149] Opensources.co. http://www.opensources.co/. The link has expired.

[150] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252, 2011.

[151] Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*, volume 5, 2014.

[152] Panagiotis Takas Metaxas, Samantha Finn, and Eni Mustafaraj. Using twittertrails. com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 69–72, 2015.

[153] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2017.

[154] Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji, and Xia Hu. Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference*, pages 3600–3604, 2019.

[155] Fotoforensics. http://fotoforensics.com/. [accessed on November 9th, 2021].

[156] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*, pages 1395–1405, 2015.

[157] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. Get back! you don ' t know me like that: The social mediation of fact checking interventions in twitter conversations. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[158] Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 275–284, 2018.

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

[160] Hugging face: cl-tohoku/bert-base-japanese. https://huggingface.co/cl-tohoku/bert-base-japanese. [accessed on November 9th, 2021].

[161] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *Proc. of International conference on machine learning*, pages 957–966, 2015.

[162] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proc. of the International Conference on Language Resources and Evaluation*, 2018.

[163] Goran Glavaš and Sanja Štajner. Simplifying lexical simplification: Do we need simplified corpora? In *Proc. of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, 2015.

[164] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, 2020.

[165] Fairness in ai 2021. https://2021.ai/fairness-in-ai/. [accessed on November 9th, 2021].

[166] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, 2019.

[167] Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, 2019.

[168] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, 2018.

[169] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, 2018.

[170] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proc. of NAACL*, pages 602–608, 2019.

[171] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv:1701.08118*, 2017.

[172] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *Proc. of SocInfo*, pages 405–415, 2017.

[173] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proc. of ACL*, pages 1668–1678, 2019.

[174] Maximilian Wich, Jan Bauer, and Georg Groh. Impact of politically biased data on hate speech classification. In *Proc. of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, 2020.

[175] Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. *Buzzfeed News*, 20, 2016.

[176] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention. In *International Conference on Machine Learning*, pages 1097–1106. PMLR, 2017.

[177] Ke Wang, Waheeb Yaqub, Abdallah Lakhdari, and Basem Suleiman. Combating fake news by empowering fact-checked news spread via topology-based interventions. *arXiv preprint arXiv:2107.05016*, 2021.

[178] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14, 2019.

[179] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

[180] Kate Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[181] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, 2014.

[182] Benjamin D Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses similar, repetitive content in text body, more similar to satire than real news. In *Proceedings of the 2nd International Workshop on News and Public Opinion at ICWSM*, 2017.

[183] Craig Silverman. This analysis shows how viral fake election news stories outperformed real news on facebook. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook, 2016. [accessed on November 9th, 2021].

[184] Melissa Zimdars. False, misleading, clickbait-y, and satirical “news” sources. https://d279m997dpfwgl.cloudfront.net/wp/2016/11/Resource-False-Misleading-Clickbait-y-and-Satirical-“News”-Sources-1.pdf, 2016.

[185] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer, 2017.

[186] William Yang Wang. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, 2017.

[187] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, 2018.

[188] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.

[189] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, 2018.

[190] UTK Machine Learning Club. Fake news: Build a system to identify unreliable news articles. https://www.kaggle.com/c/fake-news, 2018.

[191] Fatemeh Torabi Asr and Maite Taboada. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):2053951719843310, 2019.

[192] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. Fa-kes: A fake news dataset around the syrian war. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 573–582, 2019.

[193] Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876, 2019.

[194] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108. Association for Computational Linguistics, 2019.

[195] Archita Pathak and Rohini Srihari. BREAKING! presenting fake news corpus for automated fact checking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 357–362, 2019.

[196] Ria Gandhi. Getting real with fake news. https://towardsdatascience.com/getting-real-with-fake-news-d4bc033eb38a, 2020.

[197] Gautam Kishore Shahi and Durgesh Nandini. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*, 2020.

[198] Kdd 2020 truefact workshop: Making a credible web for tomorrow: Shared task 2. https://www.microsoft.com/en-us/research/event/kdd-2020-

truefact-workshop-making-a-credible-web-for-tomorrow/#!shared-tasks, 2020.

[199] Kaggle: Your machine learning and data science community. [accessed on November 9th, 2021].

[200] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

[201] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.

[202] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, 2017.

[203] Giovanni C Santia and Jake Ryland Williams. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, 2018.

[204] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. In *2nd Workshop on Data Science for Social Good, SoGood 2017*, pages 1–15. CEUR-WS, 2017.

[205] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2):e0118093, 2015.

[206] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816, 2017.

[207] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, 2018.

[208] Shan Jiang and Christo Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.

[209] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, 2019.

[210] Nguyen Thanh Tam, Matthias Weidlich, Bolong Zheng, Hongzhi Yin, Nguyen Quoc Viet Hung, and Bela Stantic. From anomaly detection to rumour detection using data streams of social platforms. *Proceedings of the VLDB Endowment*, 12(9):1016–1029, 2019.

[211] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 516–523, 2020.

[212] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, page 1165–1174, 2020.

[213] Julio CS Reis, Philipe Melo, Kiran Garimella, Jussara M Almeida, Dean Eckles, and Fabrício Benevenuto. A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proceedings of*

the International AAAI Conference on Web and Social Media, volume 14, pages 903–908, 2020.

[214] Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157, 2020.

[215] Vinay Setty and Erlend Rekve. Truth be told: Fake news detection using user reactions on reddit. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 3325–3328, 2020.

[216] Enyan Dai, Yiwei Sun, and Suhang Wang. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862, 2020.

[217] Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. Drink bleach or do what now? covid-hera: A dataset for risk-informed health decision making in the presence of covid19 misinformation. *arXiv preprint arXiv:2010.08743*, 2020.

[218] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 72–81, 2021.

[219] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*, 2020.

[220] Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. *arXiv preprint arXiv:2011.03327*, 2020.

[221] Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. No rumours please! a multi-indic-lingual approach for covid fake-tweet detection. *arXiv preprint arXiv:2010.06906*, 2020.

[222] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. Covid-19-fakes: A twitter (arabic/english) dataset for detecting misleading information on covid-19. In *International Conference on Intelligent Networking and Collaborative Systems*, pages 256–268. Springer, 2020.

[223] Chen Yang, Xinyi Zhou, and Reza Zafarani. Checked: Chinese covid-19 fake news dataset. *Social Network Analysis and Mining*, 11(1):1–8, 2021.

[224] Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. A covid-19 rumor dataset. *Frontiers in Psychology*, 12, 2021.

[225] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1867–1870, 2015.

[226] Tre Tomaszewski, Alex Morales, Ismini Lourentzou, Rachel Caskey, Bing Liu, Alan Schwartz, Jessie Chin, et al. Identifying false human papillomavirus (hpv) vaccine information and corresponding risk perceptions from twitter: Advanced predictive models. *Journal of medical Internet research*, 23(9):e30451, 2021.

[227] Oberiri Destiny Apuke and Bahiyah Omar. Fake news and covid-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, 56:101475, 2021.

[228] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. Arcov-19: The first arabic covid-19 twitter dataset with propagation networks. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 82–91, 2021.

A. Appendices

A.1. Dataset of Fake News Detection

This appendix section introduces the fake news detection dataset from the following two perspectives:

1. **News articles**: We introduce datasets that are utilized to detect fake news mainly from the body of the news article. The style of each news article is an important feature for detection.

2. **Social Media Posts**: We introduce datasets that are utilized to detect fake news mainly from social media posts related to each news. User and network information in social media, in addition to text in social media posts, are important features.

A.1.1. News articles

Politifact14 [181] is one of the initial datasets developed for fake news detection. The paper introducing Politifact14 is also the first to suggest "fact checking" and assesses the truthfulness of news publishers' statements. The main element of the dataset, with 221 samples, was a statement (also called a headline). Its label has a five-point scale: true, mostly true, half-true, mostly false, and false. Horne et al. [182] constructed two types of datasets to analyze the differences between the styles of three news items: fake, real, and satire. One is **Buzzfeed_political**, whose main topic is the 2016 US presidential election, constructed from Buzzfeed's 2016 article on fake election news on Facebook [183]. The dataset has 36 real

Table A.1.: Summary of datasets of fake news detection on news articles

Dataset	Instances	Labels	Topic domain	Raters	Language	Year
Politifact14 [181]	221 headlines	5	Politics, Society	Fact-checking sites (Politifact, Channel 4)	English	2014
Buzzfeed_political [182]	71 articles	2	the 2016 US election	Buzzfeed page [183]	English	2017
Random_political [182]	225 articles	3	Politics	List of Zimdars [184]	English	2017
Ahmed2017 [185]	25,200 articles	2	News in 2016	Fact-checking site (Politifact)	English	2017
LIAR [186, 187]	12,836 claims	6	-	Fact-checking site (Politifact)	English	2017
TSHP-17_politifact [188]	10,483 statements	6	-	Fact-checking site (Politifact)	English	2017
FakeNewsAMT [189]	480 articles	2	Sports, Business, Entertainment, Politics, Technology, Education	Generating fake news by Crowdsourcing	English	2018
Celebrity [189]	500 articles	2	Celebrity	Fact-checking site (GossipCop)	English	2018
Kaggle_UTK [190]	25,104 articles	2	-	-	English	2018
MisInfoText_Buzzfeed [191]	1413 articles	4	-	Fact-checking site (Buzzfeed)	English	2019
MisInfoText_Snopes [191]	312 articles	5	-	Fact-checking site (Snopes)	English	2019
FA-KES [192]	804 articles	2	Syrian War	Expert annotators	English	2019
Spanish-v1 [193]	971 articles	2	Science, Sport, Politics, Society, Environment, International	Fact-checking sites (VerificadoMX, Maldito Bulo, Caza Hoax)	Spanish	2019
fauxtography [194]	1,233 articles	2	-	Fact-checking site (Snopes)	English	2019
Breaking! [195]	679 articles	3	2016 US election	BS Detector	English	2019
TDS2020 [196]	46,700 articles	2	-	News Sites (BreitBart, The Onion, InfoWars)	English	2020
FakeCovid [197]	12,805 articles	2–18	COVID-19	Fact-checking sites (Snopes, Poynter)	40 languages	2020
TrueFact_FND [198]	6,236 articles	2	-	-	English	2020
Spanish-v2 [193]	572 articles	2	Science, Sport, Politics, Society, Environment, International	Fact-checking sites (VerificadoMX, Maldito Bulo, Caza Hoax)	Spanish	2021

news stories and 35 fake news stories. The other is **Random_political**, the theme of which is political news, constructed from a list of Zimdars [184]. It comprises 75 real news stories, 75 fake news stories, and 75 satire news stories. Because fake news on political topics is published more frequently than on other topics, a dataset focusing on political events, such as Buzzfeed_political and Random_political, is constructed. For example, **TSHP-17_politifact** [188] is a dataset comprising individual statements made by public political figures that have been labeled according to the ratings used by Politifact (between True and Pants-on-fire (six classes)). **Breaking!** [195] is a dataset comprising news during and before the 2016 US presidential election, used to implement a classification model based on linguistic features. Articles in the dataset were divided into three categories: false, partial truth, and opinions, and they also labeled the questionability of each news item.

Ahmed2017 [185] consists mainly of news from 2016. The number of samples in the dataset was 12,600 real articles obtained from Reuters.com and 12,600 articles determined to be fake based on Politifacts. **LIAR** [186] is a large-scale dataset in fake news detection. It is annotated using six fine-grained labels (true, mostly true, half true, mostly false, false, and pants on fire) and comprises 12,836 short statement claims from 2007 to 2016, along with information regarding the speaker, a label of credibility, the subject, the context of the statement, and others, related to each claim in the form of metadata. Alhindi et al. extended the LIAR dataset by automatically extracting the justification for each claim that people provided in the fact-checking articles associated with the claim [187]. Pérez-Rosas et al. [189] introduced two datasets covering seven different news domains for fake news detection and exploratory analysis to identify linguistic differences between fake and legitimate news content. In **FakeNewsAMT**, one dataset of [189], legitimate news, totaling 240, was obtained from a variety of mainstream news websites such as ABC News, CNN, USA Today, The New York Times, Fox News, Bloomberg, and CNET, among others. They did not utilize fake news spread on the Internet as fake news in the dataset. They asked crowd workers at Amazon Mechanical Turk to generate 240 fake news stories based on legitimate news to cover a variety of news domains. In **Celebrity**, a dataset from [189], legitimate news is obtained from online magazines such as Entertainment Weekly, People Magazine, and RadarOnline, among other tabloid and entertainment-oriented publications. In addition, fake news is obtained based on the ratings provided by GossipCop. It is composed of 250 legitimate news and 250 fake news stories. Torabi Asr et al. [191] developed two datasets of news article texts labeled by fact-checking websites to address the lack of data with reliable labels. Each sample of **MisInfoText_Buzzfeed**, one dataset from [191], was crawled based on the labeling followed by Buzzfeed. It comprises 1,090 mostly true news articles, 170 mixtures of true and false news, 64 mostly false news articles, and 56 articles containing no factual content. Another dataset is **MisInfoText_Snopes**, developed based on verified articles in Snopes. The collected articles were assigned to Snopes' labels, such as fully true, mostly true, mixture of true and false, mostly false, and fully false.

Kaggle [199], an online community of data scientists and machine learning

practitioners, has become a platform for publishing fake news datasets, despite the issue that the method of creating datasets is not clear from research papers. **Kaggle_UTK** [190] provided by Kaggle is a dataset for classifying reliable and unreliable news articles. **TDS2020** [196] is a dataset that combines the Kaggle dataset comprising articles from fake news resources such as Breitbart, The Onion, Infowars, as well as some mainstream web articles from CNN, BCC, The Guardian, and others, as real news, to enhance the fake news detection model. It comprises 24,194 fake news articles and 22,506 true news articles published after 2015. **TrueFact_FND** [198], which is hosted on Kaggle, is a dataset prepared for one of the shared tasks in the KDD 2020 TrueFact Workshop: making a credible web for tomorrow. The competition, which aims to build a high-quality fake news detection model, prepares an original dataset.

In addition to the aforementioned datasets, various other datasets have been developed for fake news detection tasks based on news article content. **FAKES** [192] is a dataset constructed to focus on the specific nature of news reporting on war incidents, particularly, the Syrian War. They used keywords relevant and specific to each of the events of the war and then built a corpus of 804 news articles from news sites extoling various political positions, such as Reuters, Etilaf, SANA, Al Arabiya, the Lebanese National News Agency, and Sputnik. They labeled 426 true articles and 376 fake articles, based on whether the VDC database contains records of casualties during the Syrian conflict and whether the extraction of casualty information from each news article by crowdsourcing is in agreement. Posadas-Durá et al. [193] developed fake news detection datasets, which are collections of Spanish news compiled from several resources on the web: **Spanish-v1** and **Spanish-v2**. A total of 1,223 news articles in the dataset were tagged with only two classes, true or fake, with regard to fact-checking sites for a community of Hispanic origins, such as VerificadoMX, Maldito Bulo, and Caza Hoax. Zlatkova et al. proposed a novel dataset called **fauxtography** [194] that focuses on the relationship between fake news and images. Each sample was provided as an image-claim pair for a new task to predict the factuality of a claim with respect to an image. Image-claim pairs tagged as fake are gathered from a special section for image-related fact-checking, called fauxtography in Snopes. It comprises 641 false pairs and 592 true pairs. Shahi et al. proposed the first mul-

tilingual cross-domain dataset of 5,182 fact-checked news articles for COVID-19, collected from January 4, 2020, to May 15, 2020, called **FakeCovid** [197]. They collected fact-checked articles and classified the truthfulness rate of each article, following the judgment of 92 different fact-checking websites. The dataset, which includes posts in 40 languages from 105 countries, was utilized as a baseline for CheckThat! Task 3 in CLEF2021. The summary of these datasets is shown in Table A.1.

A.1.2. Social media posts

Fake news detection on social media has become an important task because the development of social media has increased the number of users receiving fake news [20]. In addition, we can develop a fake news detection model with high quality by leveraging textual information as well as contextual information such as user information and network information on social media. These backgrounds activate dataset construction for fake news detection from social media. The summary of these datasets is shown in Table A.2.

The performance comparison of the fake news detection model from social media posts frequently leverages specific datasets: Twitter15, Twitter16, Twitter-ma, PHEME, and FakeNewsNet. Ma et al. constructed **Twitter15** and **Twitter16**, which are the most standard datasets for fake news detection [120] based on [93] and [225]. They find source tweets that are highly retweeted or replies related to each news item and gather all propagation threads on Twitter. In addition, they assigned four labels to the source tweets by referring to the labels of the events they originated from: true rumors, false rumors, non-rumors, and unverified rumors. **Twitter-ma** [93] is a dataset for the task of classifying rumors or non-rumors, and is composed of reported events during March-December 2015 by Snopes. For the dataset construction, they crawled threads, which comprised 498 rumors and 494 non-rumor news pieces from Twitter. **PHEME** [201], is a dataset collected and assigned three labels, and includes 330 threads related to nine different breaking threads, such as Prince to play in Toronto, the Ottawa shooting, and Ferguson unrest. The dataset contains conversations on Twitter

Table A.2.: Summary of datasets of fake news detection on social media posts

Dataset	Instances	Labels	Topic Domain	Raters	Platform	Language	Year
MediaEval_Dataset [200]	15,629 posts	2	-	-	Twitter, Facebook, Blog Post	English	2015
PHEME [201]	330 threads	3	Society, Politics	Crowdsourcing	Twitter	English	2016
Twitter-ma [93]	992 threads	2	-	Fact-checking site (Snopes)	Twitter	English	2016
RUMDECT [93]	4,664 threads	2	-	Sina community management	Weibo	Chinese	2016
RumorEval2017 [202]	297 threads	3	-	PHEME [201]	Twitter	English	2016
Twitter15 [120]	1,478 threads	4	-	Fact-checking sites (Snopes, emergent)	Twitter	English	2017
Twitter16 [120]	818 threads	4	-	Fact-checking sites (Snopes, emergent)	Twitter	English	2017
BuzzFace [203]	2,263 threads	4	Politics	Buzzfeed [130]	Facebook	English	2017
Some-like-it-hoax [204]	15,500 posts	2	Science	[205]	Facebook	English	2017
Media_Weibo [206]	9,528 posts	2	-	Sina community management	Weibo	Chinese	2017
PHEME-update [207]	6,425 threads	3	Society, Politics	PHEME [201]	Twitter	English	2018
FakeNewsNet [92]	23,921 news	2	Politics, Celebrity	Fact-checking sites (Politifact, GossipCop)	Twitter	English	2018
Jiang2018 [208]	5,303 posts	5	-	Fact-checking sites (Politcact, Snopes)	Twitter, Youtube, Facebook	English	2018
RumorEval2019 [209]	446 threads	3	Natural disaster	Fact-checking sites (Politcact, Snopes) (Politcact, Snopes)	Twitter, Reddit	English	2019
Rumor-anomaly [210]	1,022 threads	6	Politics, Fraud & Scam, Crime, Science, etc.	Fact-checking site (Snopes)	Twitter	English	2019
WeChat_Dataset [211]	4,180 news	2	-	WeChat	WeChat	English	2020
Fang [212]	1,054 threads	2	-	PHEME [201], Twitter-ma [?], FakeNewsNet [92]	Twitter	English	2020
WhatsApp [213]	3,083 images	2	Brazilian elections, Indian elections	Fact-checking sites (aosfatos.org, boomlive.in, e-farsas, etc.)	WhatsApp	-	2020
Fakeddit [214]	1,063,106 posts	2,3,6	-	Expert annotators	Reddit	English	2020
Reddit_comments [215]	12,597 claims	2	-	Fact-checkiing sites (Snopes, Politifact, emergent)	Reddit	English	2020
HealthStory [216]	1,690 threads	2	Health	HealthNewsReview	Twitter	English	2020
HealthRelease [216]	606 threads	2	Health	HealthNewsReview	Twitter	English	2020
CoAID [135]	4,251 threads	2	COVID-19	Fact-checking sites (Politifact, FactCheck.org, etc.)	Twitter	English	2020
COVID-HeRA [217]	61,286 posts	5	COVID-19	CoAID [135], Expert annotators	Twitter	English	2020
ArCOV19-Rumors [218]	162 threads	2	COVID-19	Fact-checking sites (Fatabyyano, Misbar)	Twitter	Arabic	2020
MM-COVID [219]	11,173 threads	2	COVID-19	Fact-checking sites (Snopes, Poynter)	Twitter	English, Spanish, Portuguese, Hindi, French, Italian	2020
Constraint [220]	10,700 posts	2	COVID-19	Fact-checking sites (Politifact, Snopes)	Twitter	English	2020
Indic-covid [221]	1,438 posts	2	COVID-19	Expert annotators	Twitter	Bengali, Hindi	2020
COVID-19-FAKES [222]	3,047,255 posts	2	COVID-19	WHO, UN, UNICEF	Twitter	Arabic, English	2020
CHECKED [223]	2,104 threads	2	COVID-19	Sina community management	Weibo	Chinese	2021
COVID-Alam [137]	722 tweets	5	COVID-19	Expert annotators	Twitter	English, Arabic	2021
COVID-RUMOR [224]	2,705 posts	2	COVID-19	Fact-checking sites (Snopes, Politifact, Boomlive)	Twitter, Websites	English	2021

initiated by a rumor tweet. Each tweet was grouped by story, and then annotated

based on whether the stories were confirmed to be true or false, or unverified if they could not be validated during the collection period. **PHEME-update** [207] is the extended version of the PHEME dataset [201]. This dataset contains three levels of annotation. First, each thread is annotated as either a rumor or non-rumor. Second, rumors are labeled as either true, false, or unverified. Third, each tweet is annotated for stance classification through crowdsourcing. The number of rumors in the dataset was 2,402, with 1,067 true, 638 false, and 697 unverified rumors. **FakeNewsNet** [92] is a dataset for the fake news detection task that contains a rich social media context and has been used in many studies. The dataset contains articles and related tweets fact-checked by PolitiFact or GossipCop. They retrieve 467,000 tweets in the PolitiFact dataset, and 1.25 million in GossipCop, and label them as either real or fake. Its strengths include the availability of rich social context information such as original posts, responses, and re-shared posts, user profiles, followers/followees, and social networks. However, there is the issue of a significant amount of time required to gather tweets using the Twitter API, owing to the volume of the information.

In addition to these well-known datasets, various researchers have created other datasets that are suitable for the proposed model or add new information. Jiang et al. proposed a dataset of 5,303 social media posts with 2,615,373 comments from multiple social media platforms, such as Facebook, Twitter, and YouTube, called **Jiang2018**, to analyze the linguistic difference between posts related to true and false articles [208]. They use 5-way scaling to rate each post following the judgment criteria of fact-checking sites: true, mostly true, half true, mostly false, and false. **Rumor-anomaly** [210] is a dataset of large social network information, including posts on Twitter, used for showing the effectiveness of their approach to detecting rumors at the network level, following a graph-based scan approach. It comprises four million tweets, three million users, 28,893 hashtags, and 305,115 linked articles, revolving around 1,022 rumors from May 1, 2017, to November 1, 2017, which contain several rumors related to the Las Vegas shooting and information published by the US administration. Each sample is annotated following the Snopes rating. **Fang** [212] is a dataset that is a combination of three datasets; PHEME [201], Twitter-ma [93] and FakeNewsNet [92]. A dataset that included stance information was used to evaluate the performance of their

model.

Certain competitions related to fake news detection are held, and these organizers sometimes prepare a novel dataset to improve fake news detection techniques in social media posts. **MediaEval_Dataset** [200] is utilized in the competition "verifying multimedia use" in MediaEval2015 and MeidaEval2016, which is a benchmarking initiative dedicated to evaluating new algorithms for multimedia access and retrieval and attracts participants interested in multimodal approaches to multimedia. The dataset was a set of fake and real social media posts mainly shared on Twitter to create a classifier for posts containing multimedia. The strength of this dataset is that it contains many posts with images and videos. **RumorEval2017** [202] and **RumorEval2019** [209] are utilized in the workshop "RumorEval" in SemEval, which evaluates semantic analysis systems for exploring the nature of meaning in language. RumorEval performs two tasks: stance classification toward rumors and veracity classification, using these datasets, RumorEval2017 and RumorEval2019, which are comprised of sourced posts with replies. They annotated four labels (support, denial, query, or comment) for stance classification and three labels (true, false, or unverified) for veracity classification.

However, nearly all datasets consist of posts and comments on Twitter owing to the convenience of Twitter API, some datasets are mainly comprised of posts on other social media platforms, such as Sina Weibo, Facebook, Reddit, and WhatsApp. **RUMDECT** [93] and **Media_Weibo** [206] are a dataset collected from one of China's social media platforms, Sina Weibo. Sina Weibo posts in these datasets are classified based on the judgment of Sina community management, which examines doubtful posts reported by users and verifies them as false or real based on users' reputations. RUMDECT consists of 2,313 rumors and 2,351 non-rumors that constitute a rumor detection model. Media_Weibo aims to develop a multimedia dataset that includes images, similar to Medieval_Dataset, and comprises original post texts, attached images, and available social contexts, including rumor and non-rumor sources. **BuzzFace** [203] and **Some-like-it-hoax** [204] are datasets collected from posts on Facebook. BuzzFace is based on the BuzzFeed dataset [130], which consists of 2,282 articles, along with several Facebook features (e.g., number of likes) and an assigned veracity rating.

They crawled comments and reactions related to articles in the BuzzFeed dataset on Facebook and obtained more than 1.6 million comments using the Facebook Graph API. Each article in BuzzFace has four categories: no factual content, a mixture of true and false, mostly true, and mostly false. Some-like-it-hoax consists of 15,500 Facebook posts in the scientific field and 909,236 users, to verify whether the classification of hoaxes as real or fake is possible based on user reactions. **Fakeddit** [214] and **Reddit_comments** [215] are datasets collected from threads on Reddit. Fakeddit is a large multimodal dataset consisting of over 1 million submissions from 22 different subreddits and multiple categories of fake news from March 19, 2008, to October 24, 2019. The dataset includes the submission title and image, comments, and various submission metadata, including the score, upvote-to-downvote ratio, and number of comments. It is classified based on fine-grained fake news categorizations: 2-way (fake and true), 3-way (fake, a mixture of fake and true, and true), and 6-way (true, satire/parody, misleading content, imposter content, false connection, and manipulated content). Setty et al. proposed a Reddit-based fake news detection dataset, Reddit_comments, comprising 12,597 threads with over 662,000 comments. Wang et al. constructed the **WeChat_Dataset** [211] to test whether they can leverage user reports as weak supervision for fake news detection. The dataset included a large collection of news articles published via WeChat official accounts and associated user reports. WhatsApp [213] is a dataset focusing on the spread of fake news in two events: the 2018 Brazilian elections and the 2019 Indian elections on WhatsApp, where misinformation campaigns have been used. The dataset mainly consisted of fact-checked images labeled as misinformation or not-misinformation, which were searched from the WhatsApp dataset using a perceptual hashing approach.

The COVID-19 pandemic has caused people to notice that fake news on health topics, such as the human papillomavirus (HPV) vaccine [226] and the COVID-19 vaccine [227], can have a social impact on people. Therefore, the construction of the fake news detection dataset, focusing on health-related topics, mainly COVID-19, is gaining speed. Dai et al. proposed two datasets: each sample in **HealthStory** is reported by news media such as Reuters Health, and each sample in **HealthRelease** is from various institutes, including universities, research centers, and companies [216]. These samples were annotated based on whether each

news item was fake, as per an evaluation by experts on HealthNewsReview.org. The dataset included wide-ranging context-based information related to news, such as user profiles, user networks, and retweets, for analyses. **CoAID** [135], a general fake news detection dataset related to COVID-19 from social media posts, includes 4,251 news items and 296,000 related user engagements, ranging from December 1, 2019 to September 1, 2020. **COVID-HeRA** [217], the extension of the CoAID dataset, has been constructed to flag unreliable posts based on the potential risk and severity of the statements and understand the impact of COVID-19 misinformation in health-related decision-making. The dataset is classified into five categories: real news/claims, not severe, possibly severe misinformation, highly severe misinformation, and refutes/rebuts misinformation. **Constraint** [220], which is used in the CONSTRAINT 2021 shared task, consists of social media posts on Twitter to identify whether they contain real or fake information. **COVID-RUMOR** [224] is a COVID-19 rumor dataset used for the study of sentiment analysis and other rumor classification tasks, including stance verification of COVID-19 rumors. A total of 6,834 samples, including 4,129 articles and 2,705 tweets, were annotated with sentiment and stance labels in addition to veracity labels (true, false, and unverified).

These datasets mainly comprise English posts, whereas some datasets for other languages have been constructed for the social impact of COVID-19 as a global event. **MM-COVID** [219] is a multilingual and multimodal dataset including 3,981 pieces of fake news content and 7,192 pieces of true news content from English, Spanish, Portuguese, Hindi, French, and Italian, verified by Snopes and Poynter. **COVID-Alam** [137] is also a multilingual dataset covering Arabic and English. Expert annotators labeled each post in the dataset in detail regarding five questions, including "To what extent does the tweet appear to contain false information?" "Will the tweet's claim have an impact on or be of interest to the general public?" **ArCOV19-Rumors** [218], which is the extension of an Arabic Twitter dataset ArCOV-19 [228], includes 162 verified claims and relevant tweets to these claims. The labeling rule in ArCOV19-Rumors aims to support two kinds of misinformation detection problems on Twitter: claim-level verification, which is a two-class (fake or not) classification task for each claim and all corresponding relevant tweets; and tweet-level verification, which is also a two-class

classification task given a tweet with its propagation network, such as reply and re-share information.

COVID-19-FAKES [222] is also a COVID-19 dataset including Arabic tweets, consisting of 3,047,255 posts collected using certain keywords and labeled as real or misleading. **Indic-covid** [221] is an Indian dataset that collects Hindi and Bengali tweets to detect fake news in the early stages of the COVID-19 pandemic from social media. **CHECKED** [223] is the first Chinese dataset on COVID-19 misinformation. The dataset provides 2,104 verified microblogs from December 2019 to August 2020 with rich context information, including 1,868,175 reposts, 1,185,702 comments, and 56,852,736 likes that prove the spread and reaction to these verified microblogs on Weibo.