Understanding Misinformation: The Tale of Fake News and Fake Reviews

by

Kenan Xiao

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 6, 2022

Keywords: Fake news detection, Fake Review Detection, Emotion analysis, Text Network Analysis, Deep learning

Approved by

Ashish Gupta, Professor of Department of Systems and Technology
Xiao Qin, Alumni Professor of Computer Science and Software Engineering
Gerry Dozier, Charles D. McCrary Eminent Chair Professor of Computer Science and Software Engineering
Shubhra (Santu) Karmaker, Assistant Professor of Computer Science and Software Engineering

Abstract

Misinformation has been long issues in the global communities because of the booming usage of social networks, online retail platforms and so on. The wide spreading of the massive amount of misinformation has recently become a global risk. Therefore, effective detection methods on misinformation is required to combat bad influence. In this dissertation study, we make the following three contributions by focusing on two types of misinformation detection, namely, fake news detection and fake review detection.

The first contribution of this study is the fake news engagement and propagation path framework or FNEPP, in which we devise a novel fake news detection technique from a social-context perspective. The widespread fake news on social media has boosted the demand for reliable fake news detection techniques. Such dissemination of fake news can influence public opinions, allowing unscrupulous parties to control the outcomes of public events such as elections. More recently, a growing number of methods for detecting fake news have been proposed. Most of these approaches, however, have significant limitations in terms of timely detection of fake news. To facilitate early detection of fake news, we propose FNEPP - a unique framework that explicitly combines multiple social context perspectives like news contents, user engagements, user characteristics, and the news propagation path. The FNEPP framework orchestrates two collaborative modules - the engagement module and the propagation path module - as composite features. The engagement module captures news contents and user engagements, whereas the propagation path module learns global and local patterns of user characteristics and news dissemination patterns. The experimental results driven by the two real-world datasets demonstrate the effectiveness and efficiency of the proposed FNEPP framework.

The second contribution of the dissertation lies in an emotion-aware fake review detection framework. Customers are increasingly relying on product reviews when making purchasing decisions. Fake reviews, on the other hand, obstruct the value of online reviews. Thus, automatic fake review detection is required. Previous research devoted most efforts on examining linguistic features, user behavior features, and other auxiliary features in fake review detection. Unfortunately, emotion aspects conveying in the reviews haven't yet been well explored. After delving in the effective emotion representations mined from review text, we design and implement the emotion-aware fake review detection framework anchored on ensemble learning. The empirical study on the two real-world datasets confirms our model's performance on fake review detection. To investigate how people perceive fake and real reviews differently in terms of emotion aspects, we prepare 200 real product reviews and 200 fake reviews, and random assign 20 reviews to each participant to determine the level of authenticity, credibility, and believability based on 1 - 100 scale. The results from an LIWC-22 emotion analysis intuitively demonstrate people's perception on fake reviews from the aspect of emotions.

The last contribution of the dissertation study is a two-tier text network analysis framework. As the global COVID-19 pandemic boosted the demand of online shopping, the number of online reviews increased dramatically on online shopping platforms. More often than not, customers have the tendency of referring to the product reviews before making buying decisions when products are not physically presented. Fake reviews are designed to influence buyers' purchasing decisions. Existing research devoted their efforts on designing automatic fake review detection systems; however, a text network analysis on fake reviews is missing. To close this technological gap, we construct a two-tier text network analysis framework guiding the investigation of the network-level characteristics and text characteristics of fake reviews. We conduct the extensive experiments driven by the Amazon product review dataset using Gephi. We unfold key findings on guiding the design of next-generation fake-review detection systems.

Acknowledgments

In this acknowledgement, I would like to express my sincere appreciation to my advisors, lab members and family members. I would not be able to finish the dissertation without any of your help!

First of all, I highly appreciate the help from my advisors, Dr. Ashish Gupta and Dr. Xiao Qin. Dr. Gupta spares his efforts in guiding me to formulate, desgin, and revise my research work. I can always find his help if I got stuck on some research obstacles. Under Dr. Qin's supervision, my research, academic writing skills, and multi-tasking ability got substantial improvements. Dr. Gupta and Dr. Qin have set a perfect role model on how to be professional.

Secondly, I am grateful for my wonderful committee members, Dr. Gerry Dozier and Dr. Shubhra (Santu) Karmaker. They provided me with a number of valuable suggestions, by which my dissertation can be substantially improved. I also would like to express my special appreciation to Dr. Le Chen for the willingness of serving as my university reader.

Thirdly, I would like to give high praise to Auburn University and the Department of Computer Science and Software Engineering. I am grateful for the opportunity of finishing my Ph.D.study at Auburn University. I wouldn't have achieved this without the spiritual and financial support from the War Eagle family.

Fourthly, I would express my appreciation for my lab-mates, Chengfei Wang and Xiaopu Peng. They always provided genuine advice and discussion of my research. They also maintained excellent atmosphere of the whole research group.

Last but not least, I received much spiritual support from my family members, my parents, my parents-in-law, and most importantly my wife. My wife, Yankun He, showed

great love and support when I was feeling down. She took such good care of me, and I have very few burdens on my shoulders.

This dissertation research is made possible by the support from the U.S. National Science Foundation under Grants IIS-1618669 and OAC-1642133.

To my parents , parents-in-law

and my dearest wife Yankun He

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation of Social-Context Based Fake News Detection

Nowadays, people prefer searching and consuming news via social media platforms rather than traditional news venues. According to a Pew Research Center survey conducted between August 31 and September 7, 2020, slightly over half of U.S. adults (53%) claim they read news from social media "often" or "sometimes".[1] Social media, of course, is a double-edged sword in terms of news consumption and distribution. Generally speaking, the quality of news written on social media is not on par with that of news published through traditional sources. Massive amounts of fake news as well as purposefully misleading information are crafted online for a variety of reasons, including financial and political benefits [1, 33].

Growing evidence indicate that fake news can impose negative impacts on both individuals and society. First of all, individuals may be duped by fake news and adopt wrong opinions [70, 74]. Second, fake news is intended to potentially alter people's reactions to legitimate news. Third, widespread dissemination of fake news has a potential to undermine the entire news ecosystem's credibility. As a result, it is crucial and demanding to swiftly identify fake news on social media. Fake news is purposefully designed to deceive readers and; therefore, it is non-trivial to detect fake news solely by scanning news content. In fact, concentrating on news content published on on social media becomes inadequate because news does not exist independently in the form of articles [81]. In order to develop effective and accurate fake news detection systems, we advocate for a diversity of supplementary information gleaned from social media to facilitate fake-news detection.

[1] `https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/`

A typical news propagation pattern is shown in Figure 1.1. To fully characterize the news ecosystem on social media, we propose to model fake news from four perspectives captures from social media data. The most intrinsic characteristic is the text of news articles. Content-based approaches (1) either determine if a news title coheres with its news or (2) measure quality of the writings. Efforts in automating text assessment have advanced machine learning algorithms that categorize news content as legitimate or fake based on hand-crafted and data-specific textual properties [21, 24, 51, 57, 58, 84]. The development of these cutting-edge detection schemes is challenging because the linguistic properties of fake news are still not fully unraveled. Furthermore, various types of fake news, topics, and media platforms have distinctive linguistic properties.

The second driving force behind this study is the user engagements that reflect responses from news engaged users on social media. According to specialists, fake news frequently contains biased and aggressive language that is designed to construct clickbaits or cause confusions [8, 83]. The New York Times, for example, reported individuals benefiting from the publication of online fake news; the more provocative, a higher response will result in bigger financial benefit [55].

The third aspect motivating our research lies in user characteristics. Spreaders of fake news can post misleading comments as fake news propagate. In comparison to user comments, user characteristics require strenuous effort to manipulate. Efforts in fake news detection by utilizing a series of user characteristics have been investigated in a handful of studies [6, 109, 93]. One notable weakness of those techniques is the lack of consideration of the most significant types of characteristic to detect fake news and whether or not one or more features are unavailable or insufficient in the early period of news dissemination impact the efficacy of these techniques.

The final intriguing aspect is the news dissemination path. A recent study suggests that fake news propagates differently from real news even at the early stages of spreading [120].

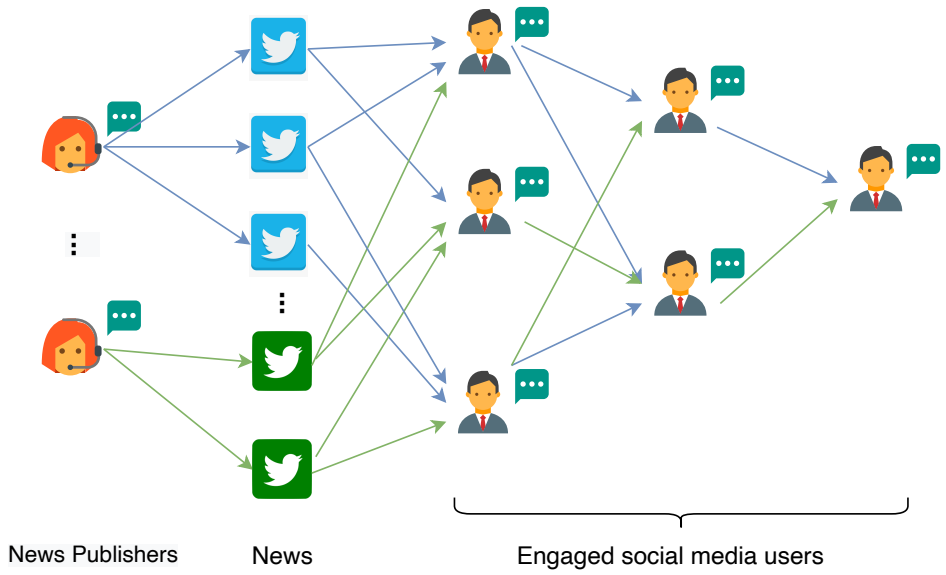

Figure 1.1: A Typical News Propagation Pattern on Social Media

Fake news propagates significantly further, faster, deeper, and broader than real news in various categories of information [102]. Thus, news dissemination patterns are valuable features in discerning fake news from legitimate ones. Recent research has studied characteristics taken from propagation paths or networks utilizing temporal-structure to detect false news (see, for example, [28, 107, 53, 34]). At an early stage of news propagation, however, these temporal-structural aspects are typically missing or inadequate.

An overarching goal of this study is to devise a generic and robust system to detect fake news posted on social media. Our work focuses on improving fake news detection systems for social media. Our novel technique is centered around modeling the four social-context perspectives of fake news, namely, fake news content, user engagements, user characteristics, and news propagation path. We frame two collaborative modules, the engagement module and the propagation path module, to accommodate the four social-context characteristics. The engagement module is designed as a Recurrent Neural Network (RNN), which takes in representations of news content and user engagement information. We construct the propagation path module as two cooperative Neural Networks (RNN and CNN), which receive user characteristics features and propagation path information. The two modules are integrated to optimize the fake news detection task jointly.

Experiments on two real-world datasets reveal that the *FNEPP* framework outperforms the existing models in terms of accuracy and other evaluation metrics thanks to concurrently modeling the four social-context perspectives of fake news. The results of early fake news detection also demonstrate that the proposed *FNEPP* framework has the advantage of accurately detecting fake news in the early stage of its dissemination.

In a nutshell, we offer the following three major contributions in this study

(1) We present a principled way for concurrently modeling the four perspectives of fake news posted on social media.

(2) We offer a unique framework *FNEPP* that seamlessly integrates the four characteristics of fake news ecosystems within two collaborative modules.

(3) The experiments driven by two real-world datasets confirm *FNEPP*'s effectiveness and efficiency while retaining the benefit of early detection of fake news.

1.2 Motivation of Emotion-Aware Fake Review Detection Framework

Online product reviews are crucial and inescapable aspects of the online business, since customers' buying decisions are significantly influenced by online reviews of products and services. According to the report from Spiegel Research Center, slightly over 90% of consumers turn to online reviews for suggestions before making purchases. As online shopping develops in popularity, fake online reviews appear almost everywhere in major online retailors, such as Amazon and eBay. A recent review transparency report from TripAdviosr stated that more than one million online reviews on the TripAdvisor website were detected to be fraudulent in 2018 [99]. Such predominance of fake reviews breaks the rules of the online business environment and misguide consumers when it comes to making purchasing decisions.

The example of demonstrating the harm and severeness of fake online reviews in online market is not unusual. Agencies, such as Federal Trade Commission (FTC), have been making huge efforts in identifying and reporting fake reviews cases within online marketing to protect consumers. The Federal Trade Commission (FTC) has issued warnings to hundreds of businesses regarding fake reviews and other misleading endorsements. For example, the Texas-based company, founded by CEO Sunday Riley, posted fake reviews on Sephora from 2015 to 2017, aiming to promote their products [2]. In 2018, one of China's most well-known tourism recommendation platforms, Mafengwo.com, was sued for generating fake reviews, which includes behaviors such as replicating reviews from competitors. The site later admitted to being implicated in the fake reviews problem. Recently, Amazon has filed a lawsuit against 'Fake Review Brokers,' who attempted to profit from the generation of fake

[2]https://www.cnn.com/2019/10/22/us/sunday-riley-fake-reviews-trnd/index.html

and deceptive customer reviews. Fake reviews and other forms of deceptive endorsements cheat consumers and undercut honest businesses.

As government agencies devote their efforts in identifying and reporting the fake reviews regarding online businesses, academic researchers and industry engineers have been developing robust auto-detection systems of fake online reviews. Over the last decade, significant advancements in the automated detection of fake reviews have been developed because of the major progress in natural language processing. Most developed fake review detection methods fall into two categories: supervised machine learning methods and unsupervised machine learning methods. Supervised machine learning type of fake review detection methods require a corpus of reviews (labelled with real or fake) is typically used for training and testing purposes. Linguistic features and behavioral features are commonly included in these methods [41, 56, 32, 31, 118, 43]. Owing to the fact that correctly labeled fake review datasets are extremely expensive to create, researchers have developed unsupervised machine learning models to identify fake reviews [36, 89, 16, 69, 40, 106, 44].

Previous research devoted most efforts on examining the linguistic features, user behavior features, and other auxiliary features in fake review detection. However, emotion aspects conveying in the reviews haven't yet been well explored. In marketing literature, brands and merchandises tend to utilize the emotion appeal to interest potential customers. Similar strategies are often adopted in fake product reviews. Evidence have shown that fake product reviews attempt to achieve the fake reviewers' objectives by evoking strong emotional feelings (fear, anger, passion, etc) rather than by a rational appeal. This could happen when two competitive merchandises hired people to write fake reviews against each other or write highly activated emotional fake reviews to exaggerate their products. Therefore, it is crucial to carefully examine how emotion aspects benefits the fake review detection task. Another interesting question we address in this paper is how people perceive fake and real reviews differently in terms of emotion aspects. Results will highlight the key factors of why

people believe in fake review in terms of emotion, and inspire making guidelines to avoid fake reviews.

We first systematically review the existing fake review detection methodologies and emotion models in the Section 2.3, 2.4, and 2.5. In Chapter 4, we first introduce the emotion representations of the text information of reviews inspired by the emotion models from three perspectives: emotion distribution, emotion intensity, and emotion dimensionality. We then demonstrate how these representations can be integrated into an emotion-aware fake review detection framework to help improve the performance of fake review detection. Next, we conducted experiments on two real-world fake review datasets, and made comparisons with state-of-the-art end-to-end fake review detection models. The results reveal the advantage of our emotion-aware fake review detection model in terms of accuracy and other evaluation metrics. Finally, we conducted a series of experiments on investigating how people perceive reviews in terms of emotions. Based on our findings, we provided guidelines and suggestions for combating fake product reviews.

1.3 Motivation of the Two-Tier Text Network Analysis Framework

Opinions are indeed a fundamental characteristic of humans: through fast growing technologies of Internet, people all over the world have much more access to other people's opinions including product reviews. Online merchandises, including Amazon and eBay, encourage buyers to post their genuine reviews of products recently purchased. These online shopping platforms promote the exchange of reviews and in turn increase public confidence in product reviews. On the flip side, competitions also comes into the picture; malicious merchandises will posit bad competitions by:

1) posting fake reviews by defaming the products or services of their competitors.

2) adding fake reviews by exaggerating the benefits and advantages of their products or services.

Fake reviews, also known as spammers, can cause severe problems, including financial issues of the merchandises. Additionally, businesses may lose their customers if fake reviews possess an unfair edge to their competitors. A large variety of news source reported the manipulative use of fake reviews. Samsung, for example, once hired fake review writers to post fake negative usage experience of HTC smartphones so as to defame the brand of HTC [3].

As the global COVID-19 pandemic boosted the demand of online shopping, the number of online reviews dramatically soars on online shopping platforms. Also, people have the tendency of referring to product reviews more often when making buying decisions because products are not often reachable. This situation raises higher chances for online shoppers to deal with fake reviews. By and large, two types of fake reviews appear often based on the sentiments of fake reviews.

1) Fake positive reviews, which usually exaggerate the good sides of products, aim to deceive buyers to buy the recommended products in the reviews. The goal of fake positive reviews is to boost product sales.

2) Fake negative reviews, which typically provide wrongly negative feedback of products, aim to persuade buyers to stay away from buying products.

Governments have produced significant efforts in reporting and preventing the generation of fake reviews, and the demonstration of the harm and severity of fake online reviews in the online businesses is not uncommon. To safeguard customers, agencies such as the Federal Trade Commission (FTC) have made tremendous efforts to uncover and disclose examples of fake reviews in online marketing. Hundreds of businesses have been warned by the Federal Trade Commission (FTC) about posting fake reviews and other deceptive recommendations. For instance, in order to market their products, the Texas-based company formed by CEO

[3]https://abcnews.go.com/Technology/samsung-fined-paying-people-criticize-htcs-products/story?id=20671547

Sunday Riley posted fake reviews on Sephora from 2015 to 2017 [4]. In 2018, one of China's most well-known tourism recommendation platforms, Mafengwo.com, one of China's most well-known tourism recommendation platforms, was charged in 2018 for manufacturing fake evaluations, including activities such as copying recommendations from rivals. The site later admitted to being implicated in the fake reviews problem. More recently, Amazon has recently filed a federal lawsuit against "Fake Review Brokers", who attempted to benefit from the creation of fake customer reviews.

In the academia, researchers devoted their efforts in designing automatic fake review detection systems to identify the fake reviews in timely manner. The majority of developed approaches to detecting fake reviews are classified into two groups: supervised machine learning methods [41, 56, 32, 31, 118, 43] and unsupervised machine learning methods [36, 89, 16, 69, 40, 106, 44]. Typically, supervised machine learning-based methods, which require labelled reviews for training and testing reasons, commonly examine the linguistic and behavioral characteristics. Researchers also proposed unsupervised machine learning models to identify fake reviews due to the high cost of creating correctly labeled fake review datasets.

As sophisticated detection methods have been devised to combat fake reviews, the text network analysis on fake reviews are missing from the radar. From the text network analysis perspectives, one may quantify and differ the network properties of fake and real reviews, which present crucial insights for the design and development of fake-review detection systems. Certain distinctive features observed from text network analysis can in turn benefit the development of such systems. One main property of networks is community: a network may include several communities with tightly connected nodes within each community. By examining the communities with each network graph, we are capable of grouping nodes with similar characteristics and examine the similarities within each community. In the example

[4]https://www.cnn.com/2019/10/22/us/sunday-riley-fake-reviews-trnd/index.html

of fake reviews, semantic features such as topic modeling, can be observed within each community, thereby making it possible to obtain good understanding of communities amid the detection of fake reviews.

In this part of the dissertation research, we first systematically review the existing fake review detection methodologies in the Section 2.4. The existing community detection algorithms are presented in Section 2.6. In Chapter 5, we introduce the text network analysis approaches, followed by the design of the two-tier text network analysis framework. Secondly, we conduct two-tier text network analysis on an Amazon dataset. The tier-1 analysis compares the network level characteristics between fake product reviews and real product reviews. The tier-2 analysis is in charge of comparing text characteristics of latent communities of fake-review networks and real-review networks. Per our findings, we provide distinctiveness features to help design the next-generation fake review detection systems.

1.4 Contributions of the Dissertation Research

This dissertation mainly focuses on understanding and combating the misinformation using natural language processing techniques. We investigated two types of misinformation, namely, fake news and fake reviews. For combating fake news, we proposed a social-context based fake news detection framework (FNEPP) that seamlessly combines four characteristics available in social context information. In terms of fake reviews, we cut from the angle of emotion analysis. We proposed an emotion-aware fake review detection framework (EmoAware), which utilizes the power of ensemble learning. Our third contribution is to explore the properties of real and fake reviews through the lens of network analysis. We examine the network properties that are embedded within the reviews. Network analytic characteristics of product fake and real reviews can provide unique insights into the structural makeup and link properties of important topics within the entire corpus.

Our organization of this section is as follows.

(1) Section 1.4.1 summarizes the contributions of our proposed fake news engagement and propagation path (FNEPP) framework from theoretical and experimental aspects.

(2) We first provide the motivation of our proposed EmoAware framework in Section 1.4.2. Next, we summarize the contribution of EmoAware framework theoretically and experimentally. We highlighted our contributions to guiding people to combating fake reviews.

(3) In Section 1.4.3, we first illustrate the motivation and intuition of why text network analysis on fake product reviews is beneficial for designing fake review detection systems. Next, we summarize the contributions of this study.

1.4.1 Contributions of Our FNEPP framework

Experiments on two real-world datasets reveal that the *FNEPP* framework outperforms the existing models in terms of accuracy and other evaluation metrics thanks to concurrently modeling the four social-context perspectives of fake news. The results of early fake news detection also demonstrate that the proposed *FNEPP* framework has the advantage of accurately detecting fake news in the early stage of its dissemination.

In a nutshell, we offer the following three major contributions in this study

(1) We present a principled way for concurrently modeling the four perspectives of fake news posted on social media.

(2) We offer a unique framework *FNEPP* that seamlessly integrates the four characteristics of fake news ecosystems within two collaborative modules.

(3) The experiments driven by two real-world datasets confirm *FNEPP*'s effectiveness and efficiency while retaining the benefit of early detection of fake news.

1.4.2 Contributions of Our EmoAware Framework

Previous research devoted most efforts on examining the linguistic features, user behavior features, and other auxiliary features in fake review detection. However, emotion aspects conveying in the reviews haven't yet been well explored. In marketing literature, brands and merchandises tend to utilize the emotion appeal to interest potential customers. Similar strategies are often adopted in fake product reviews. Evidence have shown that fake product reviews attempt to achieve the fake reviewers' objectives by evoking strong emotional feelings (fear, anger, passion, etc) rather than by a rational appeal. This could happen when two competitive merchandises hired people to write fake reviews against each other or write highly activated emotional fake reviews to exaggerate their products. Therefore, it is crucial to carefully to devleop a new system - called *EmoAware* - to examine how emotion aspects benefits the fake review detection task. Another interesting question we address in this part of the dissertation study is how people perceive fake and real reviews differently in terms of emotion aspects.

As the second part of this dissertation research, we first examine the motivation of fake review detection, especially exploring how emotion conveyed in the review text helps improve the performance of fake review detection. We provide the literature of the definition of "fake review". Existing approaches of fake review detection have been thoroughly identified. We compare our EmoAware framework with the state-of-the-art emotion models in the text mining literature, which motivates our proposed emotion representations in fake review detection. More importantly, we spearhead the development of EmoAware - an emotion-aware fake review detection framework inspired by ensemble learning methods. Three perspectives of modeling emotion conveyed in the review text are seamlessly integrated in the framework.

We carry out extensive experiments to glean the results highlighting the key factors of why people believe in fake review in terms of emotion, and inspire making guidelines to avoid fake reviews. More specifically, a series of experiments on two real-world datasets have demonstrated the effectiveness of our proposed model. Importantly, we conduct a

survey-based qualitative analysis, expecting to to evaluate how human perceive fake review differently compared with machine learning models.

We summarize the contributions made in the second part of this dissertation research.

(1) We present a principal way of representing emotion conveyed in review text processed by the state-of-the -art emotion models.

(2) We offer an emotion-aware fake review detection framework (EmoAware) that utilizes the power of ensemble learning. We systematically examine the role of three emotion features, namely, emotion distribution, emotion intensity, and emotion dimensionality.

(3) The experiments on the two real-world datasets demonstrate the effectiveness of our proposed EmoAware framework. The ablation study, in which various combinations of emotion features are constructed, show the benefits of embracing the emotion features in our framework.

(4) We conduct a survey-based qualitative research to identify how human perceive fake review differently in terms of emotion.

1.4.3 Contributions of the Two-Tier Text Network Analysis Framework

Previous research devoted their efforts in designing automatic fake review detection systems to identify the fake reviews in timely manner. The majority of developed methods for detecting fake reviews is classified into two groups: supervised machine learning methods [41, 56, 32, 31, 118, 43] and unsupervised machine learning methods [36, 89, 16, 69, 40, 106, 44]. As advanced detection approaches have been presented to combat fake reviews, there is a lack of text network analysis on fake reviews. From the standpoint of text network analysis, one may quantify and differentiate the network features of fake and authentic reviews, which can provide important insights for the design of false review detection systems. Certain distinctive features observed from text network analysis can in turn benefit the

development of such systems. A key characteristic of networks is the community, and a network may have multiple communities with densely interconnected nodes. By analyzing the communities within each network graph, we may group nodes with similar properties and investigate the similarities within each community. In the case of fake product reviews, we can observe semantic aspects inside each community, such as topic modeling, to gain a deeper understanding of the communities in fake reviews.

As the third part of this dissertation research, we first illustrate the motivation behind leveraging text network analysis to discover fake product reviews; we pay particular attention to explore how the findings of text network analysis benefit the design of fake-review detection systems. Next, we survey the literature on text network analysis, particularly in the realm of community detection algorithms. We perform a two-tier text network analysis on the Amazon fake product review dataset. More importantly, we demonstrate that results of the two-tier text network analysis are expect to identify the distinctive semantic features of fake reviews and real reviews.

We summarize the contributions made in the third part of this dissertation research.

(1) We illustrate the necessity and motivation of text network analysis on fake product reviews.

(2) We propose a two-tier text network analysis framework. The first tier analysis compares the network level characteristics between fake product reviews and real product reviews. The second tier analysis compares text characteristics of latent communities of fake-review networks and real-review networks.

(3) We perform the proposed two-tier text network analysis on Amazon product review dataset, and conclude our findings on guiding the design of fake review detection systems.

1.5 Dissertation Organization

The rest of this dissertation is organized as follows. The next chapter presents prior studies and related research works. In Chapter 3, we proposed the Fake News Engagement and Propagation Path (*(*FNEPP)) framework, and conducted experiments on two real-world datasets to demonstrate the effectiveness and efficiency of *FNEPP* framework.

In Chapter 4, we first presented the proposed Emotion-aware Fake Review Detection Framework (EmoAware). Next, we conducted empirical evaluations on Amazon dataset and OSF dataset to confirm the good performance of EmoAware framework. Lastly, we performed quantitative experiments to address how people perceive reviews in terms of emotions.

In Chapter 5, a two-tier text network analysis framework on product reviews is proposed. Then, we conduct the experiments on the Amazon product review dataset, and conclude our findings on guiding the design of fake review detection systems. Finally, Chapter 6 concludes the dissertation, and provide the direction of the future work.

Chapter 2

Literature Review

This chapter summarizes the theoretical backgrounds and related studies that are necessary for comprehending this dissertation. We systematically review the definition of fake news and fake reviews in the literature. We identify the existing two main types of fake news detection approaches, namely, content-based approaches and social-context based approaches. According to the types of machine learning models, we review supervised learning type of fake review detection, unsupervised learning type of fake review detection, and semi-supervised learning type of fake review detection methods. Lastly, we examine the existing emotion models in the text mining literature, which motivate us to design our emotion-aware fake review detection framework.

The organization of the rest of this chapter is as follows.

2.1 Fake News Detection Related Concepts and Phenomenon

2.1.1 The Definition of "Fake News"

The problem of fake news had endured since the printing press was developed in 1439 when news began to spread rapidly throughout the world [4]. Due to the rapidly growing prevalence of social media, fake news now can interact with a much larger target audience nowadays and causes severe harm to society. Academic communities and industries have conducted extensive research on fake news detection; however, there is no consensus on the definition of fake news. To systematically review the definitions, we begin by discussing and contrasting different definitions of fake news that have been employed in previous research. After that, we formally define the fake news, which will be used throughout the remainder of this dissertation.

Different from the realism of traditional news, fake news is intended to be seen as implausible [3]. Cohen *et al.* presented a generalized definition of fake news, which they defined as "everything from harmful stories to the political advertisement" [10]. Journalists often write news articles based on the web search and social media without actual verification [10], which eventually lead to misinformation and fake news. The research community widely adopts the definition of fake news as "a news story that is purposefully and verifiably incorrect and has the capability of misleading readers." According to the *authenticity* and *intent* of fake news, Shu *et al.* [92] defines fake news as "fake news is a news article that is intentionally and verifiably false" in a concise manner. The rapid growth of social media users and the diversity of social media platforms enable the fast-spreading of fake news, which creates huge advertising benefits. Klein *et al.* [33] define "fake news" as referring to the internet dissemination of deliberately or knowingly false claims of fact. Recall that our goal is to detect fake news via a social context perspective; we adopt the definition of fake news from previous research and define the fake news as follows.

Definition 2.1.1 *Fake News:* Fake news is defined as a news article that conveys verifiably false information and intentionally deceives audiences.

2.1.2 Why are people vulnerable to fake news?

In order to propose a better fake news detection model with the provision of sound theories, we investigate the reasons why people are vulnerable to fake news and have the tendency to spread false information.

Psychology Perspective: We surveyed why people are vulnerable to misinformation, such as fake news, from a psychological perspective. First of all, people are cognitive misers [98]. We favor simpler, more accessible solutions to issues over those requiring more thinking and effort; for instance, we don't put much effort into discerning the veracity of the news. Second, according to dual-process theory, people possess two ways of thinking: System 1, an automatic process that requires little effort; and System 2, an analytical process that

requires more effort [75]. Humans are in favor of the automatic process with little effort when thinking. This essentially poses the risk of misinformation. For instance, people may recall something but completely forget that it was discredited. Third, people tend to judge things based on heuristics because it is much simpler than complex analysis [61]. However, heuristics often lead to wrong judgments and conclusions. For example, in order to determine the reliability of a social media post, individuals could depend on a 'social endorsement heuristic,' which states that someone you trust has retweeted the message. Whatever level of confidence individuals have in that person, it is not an entirely trustworthy indication. Fourth, cognitive dissonance is the unpleasant sensation when one is confronted with the knowledge that opposes one's beliefs. This can cause individuals to discard trustworthy information in order to ease the cognitive dissonance [96]. Last, similar to cognitive dissonance, confirmation bias refers to the propensity to believe information supporting one's pre-existing views and disregard information that contradicts them [68]. Fake news publishers may utilize the conformation bias to design fake news that favors readers' beliefs.

Social Science Perspective: In the social science community, researchers devoted their efforts to understand why people are vulnerable to fake news. Social identity theory proposed by [97] illustrates that social acceptance and affirmation are critical components of an individual's identity and self-esteem. As fake news propagates within a group of members on social media, members may follow the majority opinion of the news because of the desire for social acceptance. Prospect theory [2] defines decision making as the process through which individuals make decisions in order to optimize relative benefits or avoid losses according to their present condition. Social media users may promote fake news propagation owing to increasing benefits or avoiding losses.

2.2 Existing Methodologies of Fake News Detection

Most existing techniques on fake news detection are categorized into two types, namely, content-based approaches (see Section 2.2.1) and social context-based approaches (see Section 2.2.2). A content-based approach aims to classify news based on the content of information to be verified, whereas a social context-based scheme utilizes rich secondary information user responses, user characteristics, and the pattern of news propagation through social media to identify fake news.

2.2.1 Content-based Fake News Detection

Fact-Checking

Fact-checking, which originated in journalism, is a process for determining the veracity of news by comparing the information derived from unverified news material (e.g., its claims or statements) to facts. The most reliable way of news verification relies on the domain experts, also known as *fact-checkers*, to discern the veracity of the news. The advantages of manual fact-checking lie in the flexibility of management and high accuracy; however, it is highly labor-intensive and inefficient with a larger amount of to-be-verified news. A growing number of fact-checking websites emerged to serve the public better. *PolitiFact* offers "the PolitiFact scorecard," which summarizes the legitimacy distribution among all claims on a certain topic. Figure 2.1 demonstrates a scoreboard on the topic of "Donald Trump". The scoreboard distribution indicates the credibility of a certain topic [114] and identifies the worthiness of verifying a news topic. The provision of fact-checking websites can support the development of fake news datasets. For instance, FakeNewsNet dataset[91] contains labeled news articles that are verified by the fact-checking websites *PolitiFact* and *GossipCop* respectively.

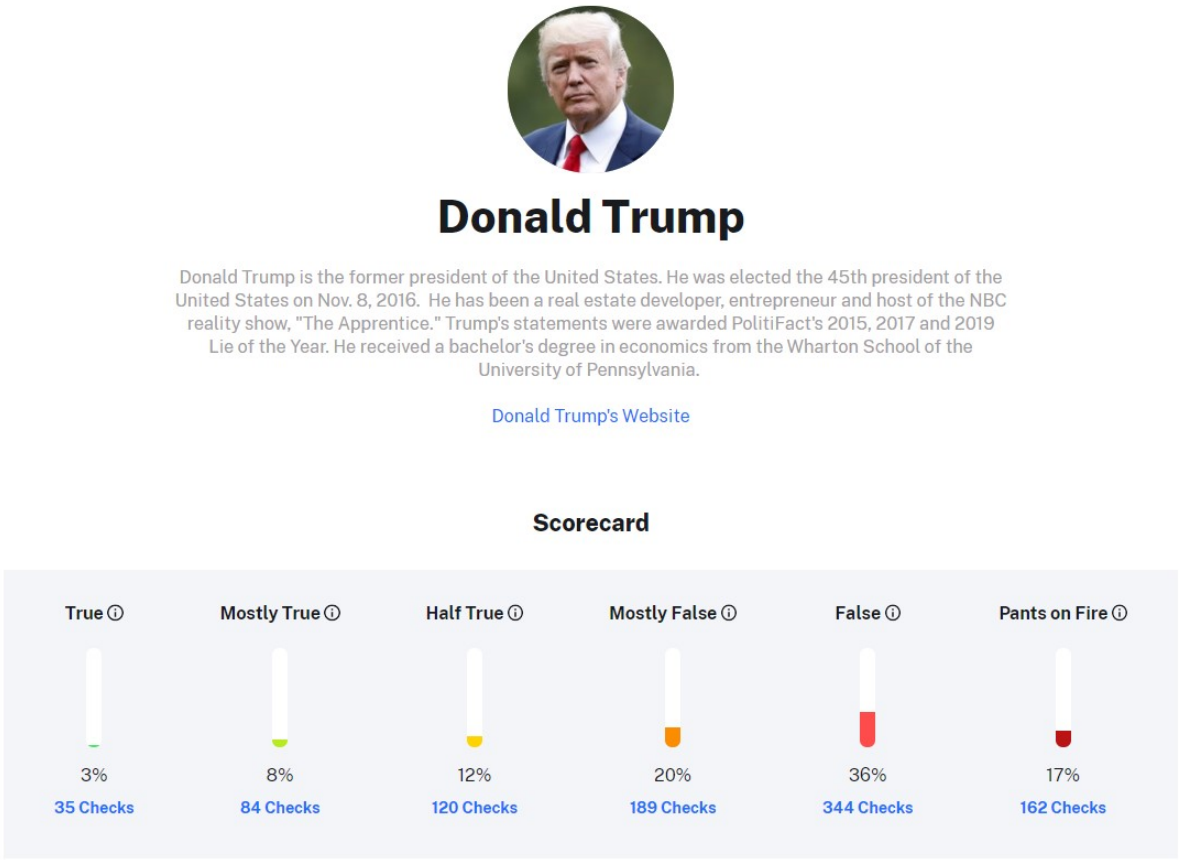

Figure 2.1: PolitiFact ScoreBoard on Topic "Donald Trump"

Cue and Feature-based Approaches

By developing a collection of linguistic cues that are instructive of content's truthfulness, cue and feature-based approaches are deployed to discern fake news from real news. Driscoll [18] applied the scientific content analysis (SCAN) scheme incorporating cues related to deception detection. While Driscoll's [18] examination indicated favorable findings in distinguishing real from fake comments using SCAN, other successive studies have proven SCAN to be inefficient according to more strict evaluations. Zhou *et al.* [121] constructed a cue set with 14 linguistic-based cues, which are effective for deception detection. Later research has examined more refined hand-crafted cue sets that are more specifically focused on the challenge of detecting fake news. Rubin *et al.* [85] evaluated a variety of textual features, including the frequency of punctuation marks and text sentiment. Zhao *et al.* [119] offered a variety of regular expressions to capture patterns of inquiry and correction in social media posts. Additionally, they incorporated platform-specific features such as the number of "hashtags" and "mentions" in Twitter postings and the ratio of inquiry or correction posts inside a cluster of posts with a high degree of textual similarity. The lack of generalization and task-specific traits, however, limits the usage of cue and feature-based approaches.

Linguistic Analysis-based Approaches

Unlike the aforementioned cue and feature-based approaches, linguistic analysis-based approaches require no task-specific, hand-engineered cue sets. The most powerful technique of linguistic analysis for detecting fake news is based on n-grams [62, 72, 73]. Mihalcea and Strapparava [62] intend to determine how the texts varied and if n-gram analysis was sufficient to distinguish falsehoods from the truth. Naive Bayes and Support Vector Machine (SVM) classifiers are trained using the term frequency vectors of n-grams in the texts. Part-of-Speech (POS) tags are generated by categorizing each word in a phrase by its syntactic function, such as nouns or adjectives. Multiple studies have discovered a strong correlation between the frequency distribution of POS tags and the style of POS tags the content

under consideration. Part-of-Speech (POS) tags were employed to extract the linguistic characteristics of fake news text [73]. Deeper syntactic features, such as features constructed from Probabilistic Context-Free Grammars (PCFG) [30] trees are explored in the context of fake news detection. Feng *et al.* [20] examined the use of Probabilistic Context-Free Grammars (PCFG) to encode deeper syntactic features for deception detection.

Deep Learning Approaches

Deep learning approaches have made substantial progress in text mining and comprehension, especially the ability to learn effective representations. Not surprisingly, existing deep learning approaches devised for fake news detection usually include convolutional neural networks (CNN) and recurrent neural networks (RNN) [105, 77, 79, 101]. Wang [105] advocated using convolutional neural networks (CNNs) to detect fake news based on its content. Qian *et al.* [77] showed that the proposed variants of CNNs, named Two-Level Convolutional Neural Network (TCNN), outperformed the linguistic analysis-based methods in detecting fake news and have the capability of handling long news articles.

2.2.2 Social Contextual-based Fake News Detection

Hand-crafted Features

Early attempts in fake news detection involve hand-engineered features, including propagation pattern features, temporal pattern features, and text-based and user-related features. For example, Castillo *et al.* [6] constructed a feature set to embrace user-based features, text-based features, propagation-based features and applied a decision tree model to classify fake news. Variants of the above features are comprised of other network-based features that are somewhat extended or tailored to an appropriate context, such as geographic locations [109] or temporal features [34]. Those approaches usually lack generality while demanding tedious human efforts.

Propagation Pattern Analysis

Research has been conducted in utilizing news propagation patterns and structures for fake news detection. Ma *et al.* [53] compared the similarity between propagation trees using tree kernels to detect fake news. A similar strategy was adopted in [107] with random walk graph kernel over propagation trees. Later research from Ma *et al.* [54] suggested extracting propagation characteristics from diffusion cascades using recursive neural networks, which are frequently utilized in syntactic and semantic parsing. Jin *et al.* [28] established a mathematical model named *SEIZ* to model a way of sharing news on social media among people.

Temporal Pattern Analysis

Discrepancies in the temporal dynamics of user engagements for news articles are beneficial for detecting false news. Previous work leverages recurrent neural networks to capture temporal patterns [86, 51]. For instance, Ruchansky *et al.* [86] partitioned a sequence of engagements into discrete time intervals with a desired level of abstraction. In another study, Ma *et al.* [51] proposed to sample engagements at regular intervals from the time series to capture temporal differences.

User Responses Analysis

User text responses and user analysis have been explored in the realm of fake news detection. User responses can be highly revealing in terms of discovering fake news. The textual response feature is represented using TF-IDF features as well as doc2vec word embeddings in [86]. Chen *et al.* [7] focused on the collected textual information by LSTM architecture coupled with an attention mechanism. Such a strategy allows for collecting typical fake news words and phrases and the visualization of which part of the text is indicative of truth or deception. When it comes to spreading fake news, news consumers might act as sources

or proponents of misinformation. As such, prior studies [50, 94, 66, 17] incorporated user features to enhance the overall performance of fake news detection systems.

2.3 The Definition of "Fake Review"

The definition of "fake review" has not been well studied in academic. Few research study has discussed a formal definition of "fake review". Zhang *et al.* [113] defined "fake review" as "deceptive reviews provided with an intention to mislead consumers in their purchase decision making, often by reviewers with little or no actual experience with the products or services being reviewed". We adopt the definition for the following reasons. First of all, the definition does not solely emphasize the misleading or imprecise information conveyed by the review, but focuses on the deceptive intention of the reviewers. Parties' who post fake reviews are intentional either to promote their own products or harm the reputations of competitors, and make more profits out of it. Secondly, one can easily separate "fake review" from other commercial operations according to this definition. In the example of "influencer marketing", internet celebrities would receive free product from different brands, and are paid to advertise the products to their followers [59]. Fake reviewers, however, pretend to be real users of the products and express their opinions of the products to achieve goals, such as promoting products or damage the reputation of them. Because of the negative consequences and inherent fallacy of fake reviews, posting them on online platforms should be legally prohibited and unacceptable. In the next subsection, we will review the state-of-the-art fake review detection methods.

2.4 Fake Review Detection

Fake reviews are becoming more widely recognized as a key source of concern for internet shoppers. Sellers tend to write positive fake reviews of their merchandise and write negative fake reviews to demote competitor's products. These behaviors result in influencing or misleading consumers' judgments and so as to boost their sales volume. Machine learning

algorithms have helped build automatic fake review detection systems because of consumers' limited ability of identifying fake reviews. Numerous research have been published in recent years about fake review detection. Existing research on detecting fake review is commonly divide into three types: supervised learning approaches, unsupervised learning approaches, and semi-supervised learning approaches. We will review existing fake review detection system according to this intrinsic methodologies.

2.4.1 Fake Review Detection by Supervised Learning

Supervised learning utilizes labeled datasets to train algorithms that accurately classify data or predict outcomes.It requires a corpus of reviews (labelled with real or fake) is typically used for training and testing purposes. We review the supervised learning types of fake review detection by two folds, namely traditional machine learning models and deep learning models.

Jindal and Liu [29] first studied the fake product review detection based on the similarity between reviews, because fake review writers are inclined to write duplicated reviews. Logistic Regression method are used to build the fake review detection model. Inspired by this effort, following research [45, 38] utilized cosine similarity among reviews to help identify fake review. Li *et al.* [38] employed supervised machine learning models such as SVM, Logistic Regression, and Naive Bayes to detect fake review. Lin *et al.* [46] proposed to utilize the Sparse Additive Generative Model, which is a variant of Bayesian generative model [19], to construct a model to detect fake reviews from various domains. A generalized additive model and topic modelling [13] are combined in the model. Features, such as unigram, part-of-speech (POS) tagging and linguistic query and word account (LIWC) are included in the model. Sedighi *et al.* [89] evaluated suitable features using classic feature selection methods, which can be enhanced by considering data correlation when selecting suitable features. A decision tree model was used to build fake review detection classifier. Besides the TF-IDF feature, Khurshid *et al.* [31] created the Content Feature set and Primal Feature set, and performed feature selection to identify key features. Various classifiers, including

Naive Bayes, Random Forest, JRip, and AdaBoost, were trained to discern fake reviews. Evaluation results demonstrated two main findings: 1) Primal Feature set plays an significant role in boosting the detection accuracy; 2) The model performed bad on imbalanced dataset. The continuous work from Khurushid *et al.* [32] proposed an ensemble learning model to boost the performance of fake review detection. The ensemble learning model consists of two tiers, where tier 1 includes three types of classifiers, namely DMNB, J48, and LibSVM. Tier 2 utilized a Logistic Regression meta classifier to calibrate the wrong predictions from tier 1 and produce more accurate results. Multiple feature selection methods, such as particle swarm optimization and cuckoo search, are employed to examine the feature space and select best sets. As ensemble learning methods tend to have better performance in fake review detection, another line of research from Mani *et al.* [56] reveals an ensemble learning model utilizing unigram and bigram features. The ensemble model trained Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM) classifiers, and employed stacking and voting strategies to ensemble three models and achieve better prediction accuracy. Li *et al.* [39] proposed a fake review detection model based on the Co-bursting phenomenon, which indicates that fake review creators tend to post fake reviews with a short and concentrated time period. Inspired by the temporal behavior of fake review creators, multi-hidden Markov model was utilized by examining the posting time of reviewers to detect the fake reviews. Mohawesh *et al.* [65] identified the concept drift problem in fake review detection, which indicates that the characteristics of the reviews arbitrarily change over time. Supervised machine learning methods were used to explore the impact of concept drifts. Experimental results demonstrated that the performance of traditional statistical machine learning models, such as SVM and Logistic Regression, dropped significantly in terms of accuracy.

As the significant advances in natural language processing (NLP) by deep learning approaches [63, 11], fake review detection tasks are also beneficial from the rapid progress of deep learning. Compared to traditional statistical machine learning methods, deep learning

approaches are capable of learning rich representations from text and capturing semantic meaning of text using word embeddings. From the deep neural network's structure perspectives, most existing work can be categorized into CNN-based [42, 118, 115, 43, 111] and RNN-based model [82, 103, 48, 27, 112, 15].

Li *et al.* [42] first utilizes CNN-based deep neural networks on fake review detection. Word2vec (Skip-gram) embeddings are fed into a sentence weighted neural network model as input. The architecture of the proposed model consists of two convolutional layers: the sentence layer, which generates a sentence composition, and the document layer, which transforms the sentence vector into a document vector. Zhao *et al.* [118] proposed a word order-preserving CNN model for fake review detection, which substitute the max pooling layer with word order reserving pooling layer when calculating word embeddings. Later, Zhang *et al.*[115] incorporates the context information by introducing a recurrent convolutional neural network, named DRI-RCNN. The convolutional layer aims to train the overall word embeddings of a given word, while the recurrent learns the context vector. Li *et al.* [43] investigated the cold start problem in fake review detection. The users' relationship and users' behavior are considered in the model. The proposed model structure consists of four layers: item embedding layers, rating embedding layers, review embedding networks, and user embedding layers. The results demonstrated that the users' social relationships can resolve the cold start problem in fake review detection to some extent. You *et al.* [111] considered the fake review detection problem as an outlier detection problem. The aspect rating of reviews are calculated by the lexicon-based approach. The local outlier factor algorithm was utilized to detect the fake review.

Recurrent neural network is adept at dealing with sequential data, such as text data and time series data. RNN-based model structure, such as LSTM, Bi-LSTM, GRU, and attention mechanism, has significant success in NLP tasks. Ren *et al.*[82] proposed to use gated recurrent neural network (GRU) along with attention mechanism to learn document representation, and utilized it as features to detect fake reviews. More recently, Wang *et*

al. [103] designed to use the long short-term memory recurrent neural network as the main part of the whole model. The entire model consists of three layers, namely input layer, an LSTM layer, and the output layer. As the single directional structure, such as LSTM and GRU, tends to overlook the backward context-dependency, Liu *et al.* [48] incorporated the Bi-LSTM model. Besides the Glove word embeddings features, part-of-sppech (POS) tagging and first-person pronoun features are fed into the model as input. In order to cope with the variation of length of reviews, Jain *et al.* [27] proposed to use multiple instance learning methods in detecting fake reviews. The CNN model was used to extract n-gram feature., whereas the GRN was used to discover semantic relationships among the retrieved features.

2.4.2 Fake Review Detection by Unsupervised Learning

Owing to the fact that correctly labeled fake review datasets are extremely expensive to create, researchers have developed unsupervised machine learning models to identify fake reviews [36, 89, 16, 69, 40, 106, 44].

Lau *et al.* [36] first introduced the Semantic Language Model (SLM) in fake review detection. They utilized the cosine similarity to determine if two reviews are similar enough to be fake reviews. Although SLM has shown its effectiveness in fake review detection, treating duplicated reviews as fake reviews remain questionable. Dong *et al.* [16] proposed a topic sentiment model, which includes four levels: document, topic, word, and sentiment. The LDA model was used to extract topic and sentiment features, and fed these features into SVM and random forest model. To achieve the probablistic distribution between words and topics, researchers utilized the Gibbs Sampling methods. Li *et al.* [40] developed a technique to identify a collection of fake reviews based on the subjects that were suggested by the users. The model is composed of three stages: first, they determined the groups and their corresponding topics. Next, K-means clustering was used to categorize the reviews into different groups. In the end, fake reviews group are classified based on bursting time

and duplicated content. To resolve the cold start problem in fake review detection, Wang *et al.* [106] introduced a deep neural network model that jointly learns the behavioral and textual information. The model detects the fake review from unlabeled data. Li *et al.* [44] also investigated the cold start problem in fake review detection. They proposed an unsupervised learning model that integrates user's behavior representation and user social relations. However, the proposed model did not outperform the state-of-the-art methods.

2.4.3 Fake Review Detection by Semi-supervised Learning

There are millions of product reviews available over the Internet. In order to utilize these unlabelled product review, semi-supervised learning approaches come into pictures.

The use of positive and unlabelled learning methods (PU) has achieved significant performance in text classification task. Inspired by the original PU learning, Ren *et al.* [108] proposed a variant of PU learning, named mixing population, to detect fake reviews. From the unlabelled dataset, some trustworthy negative cases were detected. Some demonstrative positive and negative examples were presented by the combination of Latent Dirichlet Allocation (LDA) and K-means clustering methods. According to the Dirichlet process mixture model, all fake reviews were divided into separate groups. Next, they used the individual nature and population nature to identify the group labels of fake reviews.

Deng et al. [12] investigated the use of content features and metadata features to design a PU semi-supervised learning model to detect fake reviews. They labelled the reviews based on the similarities. If two reviews are highly duplicated, they considered the reviews as fake reviews. The K-means clustering algorithm was adopted to classify the reviews by calculating the percentage of the fake review in each group. Each group was classified by the threshold value. The review was labelled positive if it has large distance from the trusted negative reviews

More recently, Yilmaz *et al.* [110] explored the use of textual content and reviewer items network features to propose a semi-supervised learning model (SPR2EP) to identify

fake reviews. Firstly, doc2vec method was adopted to produce the document embeddings. Secondly, the node2vec method was used to generate node embedding from the network data. The link of the network graph was generated by the reviewer item feature. As the reviewer writes a review about an item, the node2vec can learn the vector representation for items and reviewers. Finally, the logistic regression algorithm was utilized to detect fake reviews.

Wang *et al.* [104] proposed a model for detecting fake reviews that combines a number of features, such as review text features and reviewer features. Firstly, they explored whether the use of emotion can boost the performance. Secondly, they combined the training data with the extracted features, and continuously update them using the rolling decision-making approach. Lastly, they utilize multiple machine learning methods, such as SVM, Decision Tree, and Random Forest, to identify the fake reviews.

2.5 Emotion Models in Text Mining Literature

Emotion models determine how emotions are expressed. The models presume that emotions exist in different states, necessitating the need to discriminate between them. Various forms of expressing emotions are identified in [5], however discrete and dimensional emotion models are of critical importance to this study (DEMs and DiEMs, respectively).

Discrete emotion models

The discrete model of emotions considers classifying emotions into various categories or groups. Notable examples include:

- The Paul Ekman model, which classifies emotions into six fundamental types. The theory posits that there are six basic emotions that arise from discrete brain systems as a result of how an experiencer sees a given situation; therefore, emotions are independent. These fundamental emotions are happiness, sadness, anger, disgust, surprise,

and fear. Nonetheless, the combination of these emotions may produce additional complex emotions, such as remorse, shame, pride, lust, and so on.

- As Ekman, Robert Plutchik's approach postulates that there are a small number of core emotions, which occur in pairs of opposites and form complex emotions when combined. In addition to the six basic emotions proposed by Ekman, he identified eight such essential emotions, namely trust and anticipation. The eight contrasting emotions include happiness against sadness, trust against contempt, anger against fear, and surprise against anticipation. According to Plutchik, there are variable degrees of intensity for each emotion that come from how an experiencer interprets events.

- The Orthony, Clore, and Collins (OCC) model disagreed with Ekman and Plutchik's comparison of "fundamental emotions." However, they believed that emotions resulted from how individuals viewed events and that emotions differed in terms of their intensity. In addition to Ekman's eight basic emotions, they classified an additional sixteen feelings, including relief, envy, self-reproach, appreciation, shame, pity, disappointment, adoration, hope, fears-confirmed, sadness, gloating, like, and dislike.

Dimensional emotion models (DiEMs)

The dimensional model assumes that emotions are not independent and there is a relationship between them, hence necessitating the placement of emotions in a spatial realm. Thus, dimensional models place emotions on a dimensional space (unidimensional, i.e., 1-D, and multidimensional, i.e., 2-D and 3-D) illustrating how connected emotions are and, typically, reflecting the two fundamental behavioral states of good and bad. Both unidimensional and multidimensional DiEMs are modified by their relative frequencies (low to high). Uni-dimensional models are rarely employed although their essential principle penetrates most multidimensional models. This article provides additional information regarding multidimensional models for portraying emotions.

2.6 Community Detection Algorithms

The social network is comprised of nodes and edges, where nodes represent the entities and edges represent the relationship between these entities. In a network graph, some parts of the graph are tightly connected, while others are loosely connected or sparse. More often, the tight connected parts in the graph form a community. The community detection task is to identify the existing tightly connected parts within the network. Figure 2.2 demonstrates an example of community detection in a network graph. Nodes with same color represents a detected commn=unity.

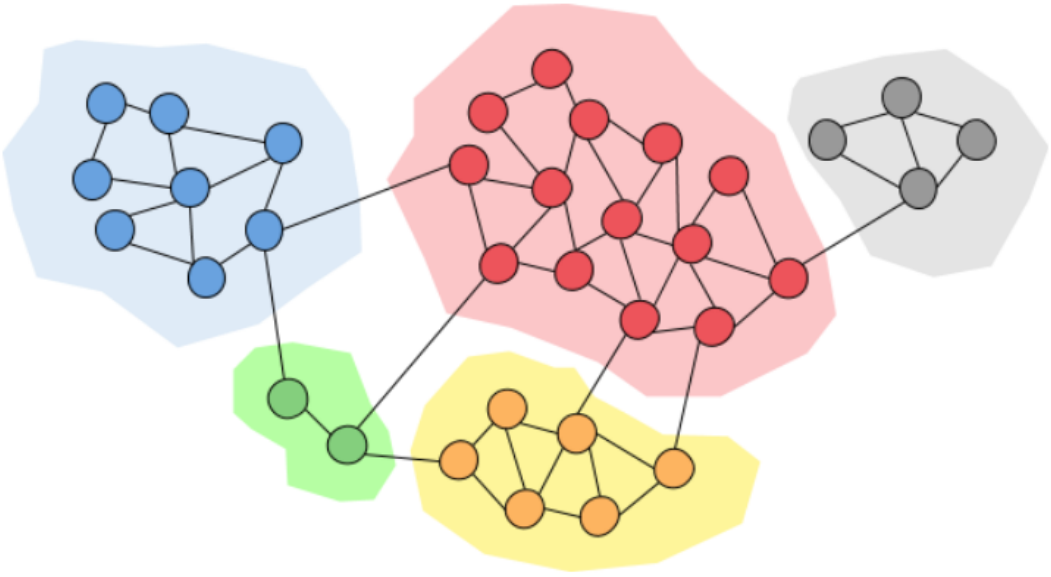

Figure 2.2: Example of Community Detection in A Network Graph

Most existing community detection algorithm can be categorized into two types, namely Girvan–Newman algorithms (GN) [22] and label propagation algorithms (LC) [71]. One of popular community detection algorithms are based on the modularity of the networks. The standard for evaluating the quality of communities is based on the value of modularity. Greater modularity indicates that the performance of community detection is better. Newman *et al.* first proposed a community detection algorithm based on the value of modularity [22]. Later work from Blondel *et al.* also adopts similar strategies. The Leuven

algorithm is achieved by continuously dividing the communities to raise the modularity. The aforementioned community detection algorithms have been well utilized in the research area.

In terms of the LC algorithm, the most applicable scenario is the non-overlapping community detection task [90]. The intuition of the LC algorithm is that the label of a node is determined by its neighboring nodes' most frequent labels. The advantages of the LC algorithm are two-folds. Firstly, the LC algorithm converges in a short period of time. Secondly, it can be applied to most scenarios because no prior parameters is required. One drawback of the LC algorithm is that we need to estimate the iteration times beforehand. Xie *et al.* proposed an variant of the LC algorithm, named LabelRank, that integrated the Markov random walk and the LC algorithm. The LabelRank algorithm sacrifices the computational complexity, but achieves better results in terms of accuracy and stability.

Even though the above methods can achieve relatively acceptable results, those methods are not capable of detecting overlapping communities. In order to detect overlapping communities within a network, researchers proposed multiple methodologies [117, 95, 47]. Zhang *et al.* [117] proposed the SAEC algorithm based on the idea of spectral clustering algorithm. The spectral clustering algorithm assumes that the weight of the link is negatively correlated with the distance of two nodes. The SAEC algorithm first caluclates the probability transfer matrix, and classifies the edges to different communities utilizing spectral clustering algorithm. The SAEC prompts the overlapping community detection.

One drawback of the SAEC algorithm is that the label space can be exploded exponentially. One possible solution is to explore the latent features. Pizzuti *et al.* [76] proposed the algorithm according to the network generative models for overlapping communities, which maximizes the probability of generating the network. The objective function aims to estimate the possible features of each node. The essence of this method is that initialization must first create the most significant node and its neighbors, followed by the node attributes,

until the algorithm's termination condition is met. However, there are two main drawbacks of this method.

- Firstly, the method is not stable because of excessive overlapping communities.
- Secondly, prior to presenting the final results of community detection, it must consume a substantial amount of space to store and manage the network topology.

To resolve the above issues, NRL-based community detection methods [100] are proposed.

2.7 Summary

In this chapter, we first systematically review the definition of fake news and fake review. We also identify and summarize existing methods in fake news detection and fake review detection.

For existing fake news detection methods, we review these approaches based on content-based approaches and social-context based approaches. We also identified these methods based on the types of features.

In terms of fake review detection, we review the existing methods based on the intrinsic types of learning methods, which supervised learning, unsupervised learning, and semi-supervised learning. We also explored the existing theories in emotion models. We review the emotion models based on the different represented aspects of emotions, namely, discrete emotion models and dimensional emotion models. These emotion models in turn lay solid foundations of designing our framework.

Lastly, to comprehend the existing community detection methods for network analysis, we reviewed these methods based on two main types, namely Girvan–Newman algorithms (GN) [22] and label propagation algorithms (LC) [71].

The next several chapters will discuss our proposed framework, experiments and key findings in the area of fake news detection, fake review detection, and text network analysis.

Chapter 3

Fake News Engagement and Propagation Path (*FNEPP*) Framework

The widespread fake news on social media has boosted the demand for reliable fake news detection techniques. Such dissemination of fake news can influence public opinions and society. More recently, a growing number of methods for detecting fake news have been proposed. However, most of these approaches have significant limitations in timely detection of fake news. To facilitate early detection of fake news, we propose in this Chapter a unique framework *FNEPP (Fake News Engagement and Propagation Path)* from a social context perspective, which explicitly combines news contents, user engagements, user characteristics, and the news propagation path as composite features of two collaborative modules. The engagement module captures news contents and user engagements, while the propagation path module learns global and local patterns of user characteristics and news dissemination patterns. Experimental results on two real-world datasets demonstrate the effectiveness and efficiency of the proposed *FNEPP* framework.

The rest of of this chapter is organized as follows. In section 3.1, we first recognize the challenges of fake news detection strategies, followed by illustrating the intuition and basic ideas of the design of *FNEPP* framework. The social-context based fake news detection task is also formalized using set notations in this section. In section 3, we systematically introduce the two collaborative modules of *FNEPP* framework separately , followed by how the integration of two modules works. In section 3.3, we conduct experiments on the two real-world datasets, and confirm the effectiveness and efficiency of the *FNEPP* framework. Lastly, we summarize this chapter in section 3.5.

3.1 Challenges, Basic Ideas, and Problem Statement

As the popularity of social media platforms, such as Twitter, grows up significantly these days, people consume daily news heavily relying on these social media platforms. Milicious parties tend to post fake news on social media platforms to deceive users so that they can achieve financial or political benefits. In this section, we first post the challenges and basic ideas of social-context based fake news detection in Section 3.1.1 and Section 3.1.2, respectively. Then, we formally define, in Section 3.1.3, the social-context based fake news using set notations, which lay out a foundation for our proposed *FNEPP* framework depicted in Section 3.2.

3.1.1 Challenges

Before introducing our basic ideas in the first part of the dissertation, let us emphasize an array of three challenges to be addressed in this chapter.

- Challenge 1. The quality of news written on social media.
- Challenge 2. The widespread dissemination of fake news.
- Challenge 3. It is non-trivial to detect fake news.

Challenge 1. The quality of news written on social media. It is a common practice for users to search news via social media platforms rather than traditional news venues. A Pew Research Center survey indicates that slightly over half of U.S. adults (53%) claim they read news from social media "often" or "sometimes". [1] We reckon that the quality of news written on social media is not on par with that of news published through traditional sources.

Challenge 2. The widespread dissemination of fake news. Past evidence shows that fake news impose negative impacts on both individuals and society. Individuals may be

[1] https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/

duped by fake news and adopt wrong opinions [70, 74], and fake news is intended to potentially alter people's reactions to legitimate news. Furthermore, widespread dissemination of fake news has a potential to undermine the entire news ecosystem's credibility. It is crucial and demanding to swiftly identify fake news on social media.

Challenge 3. It is non-trivial to detect fake news. Fake news is purposefully designed to deceive readers: it is non-trivial to detect fake news solely by scanning news content. In fact, concentrating on news content published on on social media becomes inadequate because news does not exist independently in the form of articles [81]. In order to develop effective and accurate fake news detection systems, we advocate for a diversity of supplementary information gleaned from social media to facilitate fake-news detection.

3.1.2 Basic Ideas

To fully characterize the news ecosystem on social media, we propose to model fake news from four perspectives captures from social media data. The most intrinsic characteristic is the text of news articles. Content-based approaches (1) either determine if a news title coheres with its news or (2) measure quality of the writings. Efforts in automating text assessment have advanced machine learning algorithms that categorize news content as legitimate or fake based on hand-crafted and data-specific textual properties [21, 24, 51, 57, 58, 84]. The development of these cutting-edge detection schemes is challenging because the linguistic properties of fake news are still not fully unraveled. Furthermore, various types of fake news, topics, and media platforms have distinctive linguistic properties.

The second driving force behind this study is the user engagements that reflect responses from news engaged users on social media. According to specialists, fake news frequently contains biased and aggressive language that is designed to construct clickbaits or cause confusions [8, 83]. The New York Times, for example, reported individuals benefiting from the publication of online fake news; the more provocative, a higher response will result in bigger financial benefit [55].

The third aspect motivating our research lies in user characteristics. Spreaders of fake news can post misleading comments as fake news propagate. In comparison to user comments, user characteristics require strenuous effort to manipulate. Efforts in fake news detection by utilizing a series of user characteristics have been investigated in a handful of studies [6, 109, 93]. One notable weakness of those techniques is the lack of consideration of the most significant types of characteristic to detect fake news and whether or not one or more features are unavailable or insufficient in the early period of news dissemination impact the efficacy of these techniques.

The final intriguing aspect is the news dissemination path. A recent study suggests that fake news propagates differently from real news even at the early stages of spreading [120]. Fake news propagates significantly further, faster, deeper, and broader than real news in various categories of information [102]. Thus, news dissemination patterns are valuable features in discerning fake news from legitimate ones. Recent research has studied characteristics taken from propagation paths or networks utilizing temporal-structure to detect false news (see, for example, [28, 107, 53, 34]).

Our work focuses on improving fake news detection systems for social media. Our novel technique is centered around modeling the four social-context perspectives of fake news, namely, fake news content, user engagements, user characteristics, and news propagation path. We frame two collaborative modules, the engagement module and the propagation path module, to accommodate the four social-context characteristics. The two modules are combined seamlessly as the Fake News Engagement and Propagation Path (*FNEPP*) framework. The engagement module is designed as a Recurrent Neural Network (RNN), which takes in representations of news content and user engagement information. We construct the propagation path module as two cooperative Neural Networks (RNN and CNN), which receive user characteristics features and propagation path information. The two modules are integrated to optimize the fake news detection task jointly.

3.1.3 Problem Formulation

This section introduces the set of notations and formalize the fake news detection task. We assume that a series of fake news interactions occur across a time interval $[0, T]$. Our goal is to detect fake news early after it starts to spread on social media. Therefore, we should promptly detect fake news within a short time period (T is a small value). In what follows, Our detection model consists of four vital sets, namely, article set A, user set U, engagement set E, and propagation path set P. In what follows, we formally define these four sets to pave a way for the problem statement of this study.

(1) $A = \{a_1, a_2, ..., a_i, ..., a_{|A|}\}$ is a set of news articles to be classified as fake or legitimate news.

(2) $U = \{u_1, u_2, ..., u_j, ..., u_{|U|}\}$ is a set of social media users, where each user u_j engaged in spread a news article in set A.

(3) $E = \{e_1, e_2, ..., e_k, ..., e_{|E|}\}$ is a set of engagements. Each e_k is essentially represented as a 3-tuple, (a_i, u_j, t), where user u_j retweets or comments about the news article a_i at time t.

(4) $P = \{p_{a_1}, p_{a_2}, ..., p_{a_i}, ..., p_{a_{|A|}}\}$ is a set of news propagation path. Each propagation path p_{a_i} is associated with news article a_i. Propagation path is naturally denoted as a multivariate time series $p_{a_i} = \{..., (\mathbf{x_{u_j}}, t_{u_j}), ...\}$, where $\mathbf{x_{u_j}}$ is a vector representation of user u_j who engages with news article a_i and $t_{u_j} \in [0, T]$.

With the above notation in place, we formally formalize the problem of detecting fake news from social context in *Definition 1.*

Definition 1. *Social context-based fake news detection.* Given a set of news articles A, a set of social media users U, a set of engagements E, and a set of news propagation path P, social context-based fake news detection is defined as a binary classification problem to

predict a label $\hat{y_{a_i}} \in \{0, 1\}$ for news article a_i, where $\hat{y_{a_i}} = 1$ indicates a_i is fake, while $\hat{y_{a_i}} = 0$ indicates a_i is legitimate.

Recall that our goal is to pinpoint fake news in an early stage of news dissemination. In the above formal problem statement, the performance of a fake news detection system is closely related to parameter T. As such, we undertake an empirical study to delve in the correlation between parameter T and detection performance. Please refer to Section 3.4 for the results with respect to early fake news detection.

3.2 Fake News Engagement and Propagation Path (*FNEPP*) Framework

In this section, we describe the details of our proposed framework, *FNEPP*. *FNEPP* mainly consists of two modules, namely engagement module and propagation path module, that collectively capture news contents, user engagements, user characteristics, and news propagation paths. The engagement module is dedicated to capturing the most efficient representations of user engagements and news articles. The propagation path module is responsible for capturing the news propagation path along with user characteristics. The details of the proposed framework are shown in Figure 3.1.The engagement module extracts a temporal representation of news articles using a Recurrent Neural Network (more accurately, an Long Short-Term Memory (LSTM) model). The user engagements are represented as vectors and fed into the LSTM to produce a final representation vector $\mathbf{e_{a_i}}$ for the engagement module. The propagation path module utilizes the vector representations of user characteristics to construct propagation paths as multivariate time series. Recurrent Neural Networks (more precisely, GRUs) and Convolutional Neural Networks(CNN) extract the global and local propagation patterns, respectively.

With the help of the engagement module, the model can extract a low-dimensional vector representation $\mathbf{e_{a_i}}$ of user engagements and news content for a particular news article. The propagation path module utilizes the first N engaged users' characteristics within the time interval $[0, T]$ to obtain a vector representation $\mathbf{p_{a_i}}$ that captures local and global

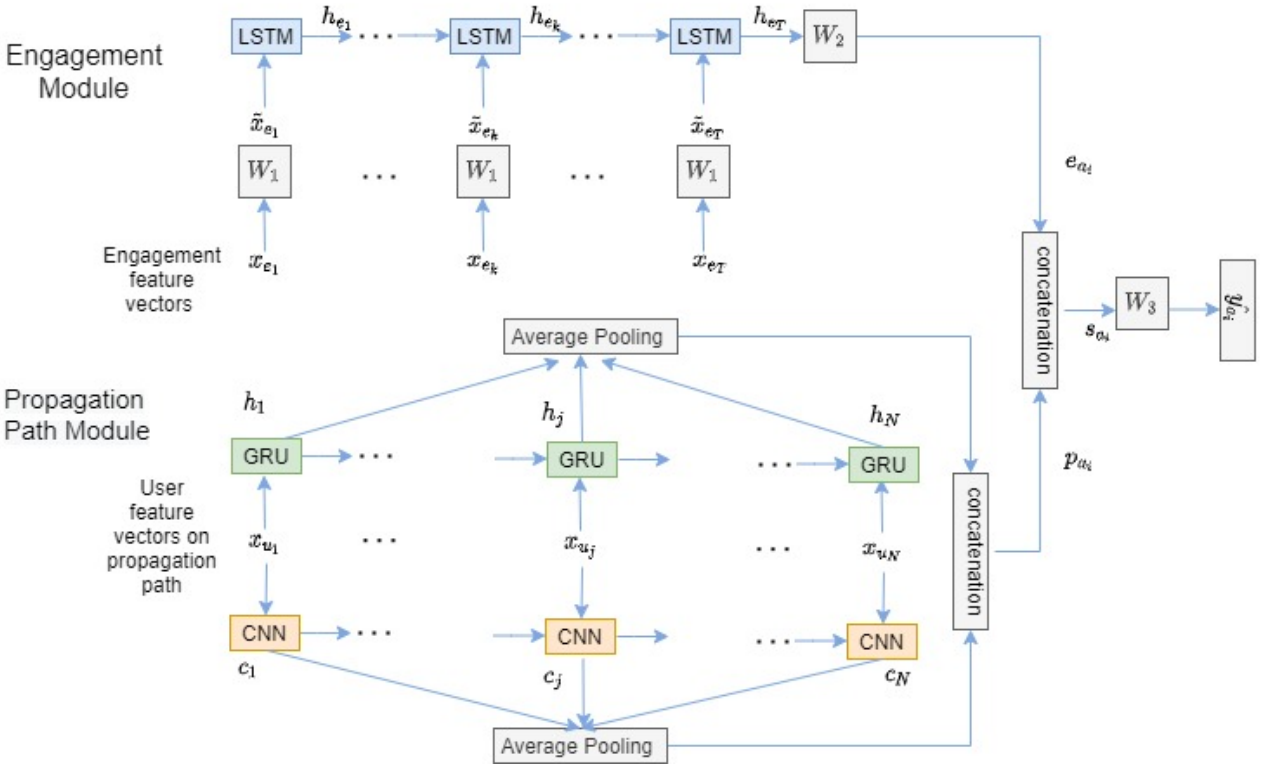

Figure 3.1: The Architecture of *FNEPP*

propagation patterns. By concatenating vectors obtained by two modules, a final prediction of the veracity of the news is achieved.

3.2.1 Engagement Module

The objective of the engagement module is to capture the pattern of users' temporal interactions with a news article a_i in accordance with the occurrence and distribution. Explicitly, the module can capture both the number of engaged users of a_i and the pattern of those interactions over time. Additionally, the textual data associated with the interactions, such as the content of users' retweets, are combined in the module.

We employ a Recurrent Neural Network (RNN) as the basis for the engagement module since RNNs can effectively combine diverse information sources and catch temporal patterns within data. Notably, we choose the LSTM model because of its capability of processing variable-length input and its tendency for capturing long-term dependencies [26]. The critical component of this module is the selection of engagement feature vector $\mathbf{x_{e_k}}$ shown in Figure 3.1, which serves as the input to the LSTM cell.

The engagement feature vector $\mathbf{x_{e_k}}$ essentially consists of four parts and can be represented as the following vector:

$$\mathbf{x_{e_k}} = (\mathbf{x_u}, \mathbf{x_a}, \Delta t, n) \tag{3.1}$$

The first part $\mathbf{x_u}$ aims to model the engaged users. We create a binary incidence matrix representing the news articles that a particular user has interacted with. The binary index matrix is high-dimensional and sparse because the number of social media users is much larger than the number of news spreading over social media. Therefore, we employ the Singular value decomposition (SVD) for a binary incidence matrix to obtain lower-dimensional representation for engaged users. The second part $\mathbf{x_a}$ is capable of capturing the text of each engagement. In order to prevent hand-crafted textual features, we apply the *doc2vec*

[37] embeddings on the text of each engagement. Since we want to capture the occurrence and distribution of engagements over time, we introduce two variables (1) the number of engagements n and (2) the time interval between two consecutive engagements Δt.

As shown in Figure 3.1, we add an embedding layer right after the raw input vector $\mathbf{x_{e_k}}$. Because the input features are constructed from different sources, it is not an advisable practice to feed the input vector $\mathbf{x_{e_k}}$ directly into the LSTM unit. The embedding layer is a fully connected layer, which transforms the raw input vector $\mathbf{x_{e_k}}$ to $\tilde{\mathbf{x}}_{\mathbf{e_k}}$ by the following formula:

$$\tilde{\mathbf{x}}_{\mathbf{e_k}} = tanh(W_1 \mathbf{x_{e_k}} + \mathbf{b_1}) \tag{3.2}$$

where W_1 is the fixed weight matrix and $\mathbf{b_1}$ is the fixed bias vector for all $\mathbf{x_{e_k}}$. The transformed vector $\tilde{\mathbf{x}}_{\mathbf{e_k}}$ is supplied into the LSTM as the input. The last hidden state vector $\mathbf{h_{e_T}}$ is fed into a fully connected layer to obtain the final vector representation $\mathbf{e_{a_i}}$ for news article a_i in the engagement module.

$$\mathbf{e_{a_i}} = tanh(W_2 \mathbf{h_{e_T}} + \mathbf{b_2}) \tag{3.3}$$

In summary, the engagement module encodes the engagement patterns by a lower-dimensional vector $\mathbf{e_{a_i}}$. The vector $\mathbf{e_{a_i}}$ captures the temporal pattern of user engagements for news article a_i in terms of the occurrence and distribution of engagements and all textual contents.

3.2.2 Propagation Path Module

The primary task of propagation path module is to assess each user through their profiles and other available information on social media and learn representations to discern bogus propagation patterns from real ones. The global representations of the propagation path are learned by an RNN-based sub-module, while the CNN-based sub-module aims to

extract local patterns of the propagation path. The global and local representations of the propagation path are integrated as the final representation for the propagation path module.

The propagation path for a particular news article a_i is naturally represented as a multivariate time series as follows:

$$p_{a_i} = \{..., (\mathbf{x_{u_j}}, t_{u_j}), ...\} \tag{3.4}$$

where $\mathbf{x_{u_j}}$ is a vector representation of user u_j who engaged with news article a_i and $t_{u_j} \in [0, T]$. $\mathbf{x_{u_j}}$ is constructed by extracting the user characteristics from their social media profile and relevant information. Further technical details on constructing $\mathbf{x_{u_j}}$ will be demonstrated in the Section 3.3.2.

We consider the interactions that happened within a time interval $[0, T]$ after the news article created on social media. The number of interactions on the propagation path may vary for different news articles. Therefore, to unify a fixed-length propagation path, we propose the following transformation. We assume that the length of the transformed propagation path is N.

- Case 1: If the length of p_{a_j} is not smaller than N, we keep the first N tuples of p_{a_j} as final propagation path sequence $\tilde{p_{a_j}}$.

- Case 2: If p_{a_j} contains less than N tuples, we randomly sample $(N - |p_{a_j}|)$ times and concatenate the sampled tuples to achieve the final propagation path sequence $\tilde{p_{a_j}}$ with length N.

The fixed-length propagation path is represented as follows:

$$\tilde{p_{a_j}} = \{(\mathbf{x_{u_1}}, t_1), ..., (\mathbf{x_{u_j}}, t_j), ..., (\mathbf{x_{u_N}}, t_N)\} \tag{3.5}$$

For our propagation path module, we only consider the relative time order of each user. Thus, we sort the user feature vectors in $\tilde{p_{a_j}}$ based on an ascending time order and omit the

time in the tuple afterward. We rewrite $\tilde{p_{a_j}}$ as follows:

$$\tilde{p_{a_j}} = \{\mathbf{x_1}, ..., \mathbf{x_n}, ..., \mathbf{x_N}\} \tag{3.6}$$

where $\tilde{p_{a_j}}$ is an ordered sequence according to the engagement time with news article a_j.

Local Propagation Path Representation: Convolutional Neural Networks (CNN) are particularly suited for capturing local variations and representations. We propose to use 1D CNN to learn a vector representation for each propagation path $\tilde{p_{a_j}}$. We assume the user feature vector $\mathbf{x_n} \in \mathbb{R}^l$. We stack the user feature vectors into a user feature matrix $X \in \mathbb{R}^{N \times l}$. 1D CNN is applied on h successive users with a filter $W_c \in \mathbb{R}_{h \times l}$. Each convolution operation produces a scalar feature c_n as follows:

$$c_n = ReLU(W_c X_{n:n+h-1} + b_c) \tag{3.7}$$

where $X_{n:n-h+1}$ is the subset of h consecutive user feature vectors. $b_c \in \mathbb{R}$ is a scalar bias term. We repeat the above convolution operations with m filters and obtain a feature vector $\mathbf{c_n} \in \mathbb{R}^m$. We apply the convolution procedure for all subsets of consecutive h user feature vectors and obtain a sequence of feature $\{\mathbf{c_1}, ...\mathbf{c_{N-h+1}}\}$. The average pooling is utilized to produce the final vector representation $\mathbf{p_C}$ for local propagation path representation.

$$\mathbf{p_C} = \frac{1}{N} \sum_{n=1}^{N-h+1} \mathbf{c_n} \tag{3.8}$$

Global Propagation Path Representation: In order to capture the global patterns of propagation path, we propose to utilize Gated Recurrent Unit (GRU) to learn vector representations for transformed propagation paths. A GRU unit takes a user feature vector $\mathbf{x_n}$ and produces the hidden state representation $\mathbf{h_n}$ based on the following formulations,

which is adopted from [9].

$$\begin{aligned}
\mathbf{z_n} &= \sigma(U_z \mathbf{x_n} + W_z \mathbf{h_{n-1}}) \\
\mathbf{r_n} &= \sigma(U_r \mathbf{x_n} + W_r \mathbf{h_{n-1}}) \\
\tilde{\mathbf{h}}_\mathbf{n} &= tanh(U_h \mathbf{x_n} + \mathbf{h_{n-1}} \odot W_h \mathbf{r_n}) \\
\mathbf{h_n} &= (1 - \mathbf{z_n}) \odot \mathbf{h_{n-1}} + \mathbf{z_n} \odot \tilde{\mathbf{h}}_\mathbf{n}
\end{aligned} \tag{3.9}$$

where $U_z, U_r, U_h, W_z, W_r, W_h$ are weight matrices, and $\odot$ denotes the element-wise vector multiplication. The detailed description of GRU model can be found in [9]. We apply average pooling over all hidden states produced by GRU and obtain our vector representation $\mathbf{p_R}$ for global propagation paths as follows:

$$\mathbf{p_R} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{h_n} \tag{3.10}$$

We concatenate the $\mathbf{p_C}, \mathbf{p_R}$ and obtain $\mathbf{p_{a_i}}$ as our final vector representation of propagation path module.

3.2.3 Integration

As described earlier, the engagement module combines the news article and user engagements to capture the engagement patterns while the propagation path module incorporates the user characteristics from social media and propagation path to discern the fake news dissemination patterns from real ones. On behalf of accommodating representations from two modules, we concatenate $\mathbf{e_{a_i}}, \mathbf{p_{a_i}}$ together as the vector $\mathbf{s_{a_i}}$. $\mathbf{s_{a_i}}$ is utilized as input to a fully connected layer to predict whether the news article a_i is fake or not.

$$\hat{y_{a_i}} = \sigma({W_3}^T \mathbf{s_{a_i}} + \mathbf{b_3}) \tag{3.11}$$

	PHEME	WEIBO
# Users	37,175	2,746,818
# News Stories	5,802	4,664
# Real News	3,830	2,351
# Fake News	1,972	2,313
Avg. Time Length / news	26	1,983

Table 3.1: Statistics of the Datasets

where W_3 is the weight matrix shown in Figure 3.1, $\mathbf{b_3}$ is the bias term. We apply the cross-entropy loss function for training our model.

The advantage of this integration is that it unifies the two modules to form a more accurate prediction. In addition, the model learns distinctive patterns between fake news and real news by jointly training the engagement module and propagation path module simultaneously.

3.3 Experimental Design

This section discusses the experimental design and setup to quantitatively show the efficacy of *FNEPP* on two real-world news datasets. We evaluate our approach's performance by comparing it to several baseline and state-of-the-art models in terms of accuracy, precision, recall rate, and $F1$-*score*.

3.3.1 Data

To make a fair comparison, we conduct the experiments on two real-world social media datasets that were also used in previous research, PHEME [122], and WEIBO [51]. The PHEME dataset and WEIBO dataset contain breaking news and each news associates with a set of user engagements. The profiles of the engaged users are also available in each dataset, which provides the convenience of constructing propagation paths for modeling purpose. A summary of key statistics is described in Table 3.1.

Features	Type
length of user name	integer
length of user description	integer
follower counts	integer
friends counts	integer
favorites counts	integer
status counts	integer
user verified status	boolean
geo enabled status	boolean

Table 3.2: User Characteristics for Constructing $\mathbf{x_{u_j}}$

3.3.2 Experimental Setup

Before explaining the major findings, we articulate the specific features within each dataset. Next, we introduce the hyperparameters for training our model. The alternative models, serving as competitors to our model, are briefly outlined at the end of this subsection.

Features: The engagement module essentially judiciously extracts an engagement vector for each news article $\mathbf{x_{e_k}} = (\mathbf{x_u}, \mathbf{x_a}, \Delta t, n)$. Feature vector $\mathbf{x_u}$ is constructed by the SVD decomposition with a rank 10 for the PHEME and WEIBO datasets. In order to apply *doc2vec* to obtain textual feature $\mathbf{x_a}$, we perform text segmentation on the WEIBO dataset. The embedding dimension is set to 100 for both datasets, which result in $\mathbf{x_a}$ with 100 dimensions. The dimension of the engagement vector $\mathbf{x_{e_k}}$ is 112. For the propagation path module, we construct user feature vector $\mathbf{x_{u_j}}$ for each engaged user with the following features listed in Table 3.2. It is note worthy that the user feature vectors are derived from the common user characteristics available in the two tested datasets.

The choice of the characteristics summarized Table 3.2 unravels the legitimacy of social media users to some extent. For instance, social disrupters tend to post and spread fake news via zombie accounts on social media. Characteristics like follower counts, friend counts, and user verified status can help in discerning potential zombie accounts.

Hyperparameters: For the propagation path module, the GRU units' output dimension is 32. The size of the CNN filter is set to three, and we employ 32 CNN filters to extract

local propagation path representations. For the engagement module, the hidden dimension of LSTM is set to 50. For the training purpose, we apply Adam optimizer with a learning rate 0.0001.

Comparison Models: We compare our proposed model against the following four alternative models found in the literature.

DTR: A decision-tree-based ranking algorithm identifies fake news using query terms. [119].

SVM-TS: A linear support vector machine classification model utilizes time series to simulate the temporal change of social context characteristics [52].

GRU: A rumor detection model advocates RNNs and GRU for long-term representation learning of relevant posts [51].

CSI: An effective recurrent encoder aggregates user features, news content, and user–news engagements [86].

3.4 Results and Analysis

Methods	Class	Accu.	Prec.	Rec.	F_1
DTR	R	0.562	0.549	0.704	0.617
	F		0.588	0.421	0.491
SVM-TS	R	0.651	0.642	0.686	0.663
	F		0.663	0.617	0.639
GRU	R	0.722	0.734	0.712	0.723
	F		0.722	0.733	0.728
CSI	R	0.742	0.743	0.728	0.736
	F		0.735	0.750	0.743
FNEPP	R	0.780	0.789	0.764	0.776
	F		0.771	0.794	0.783

Table 3.3: Comparison Results from the PHEME Dataset ("F: " fake news; "R": real news)

Overall Comparison. Tables 3.3 and 3.4 demonstrate the performance of all the compared models detecting fake news from the PHEME and WEIBO datasets. The results

Methods	Class	Accu.	Prec.	Rec.	F_1
DTR	R	0.732	0.726	0.749	0.737
	F		0.738	0.715	0.726
SVM-TS	R	0.857	0.878	0.830	0.857
	F		0.839	0.885	0.861
GRU	R	0.892	0.922	0.864	0.893
	F		0.876	0.926	0.901
CSI	R	0.905	0.895	0.907	0.901
	F		0.915	0.909	0.912
FNEPP	R	0.919	0.907	0.915	0.911
	F		0.928	0.921	0.925

Table 3.4: Comparison Results from the WEIBO Dataset ("F: " fake news; "R": real news)

indicate that our proposed *FNEPP* outperforms all the competitive models in almost every evaluation metric. For example, when it comes to the PHEME dataset, *FNEPP* boosts the accuracy of DTR, SVM-TS, GRU, and by 22%, 13%, 6%, and 3.8%, respectively. Similarly, *FNEPP* has a clear edge over the four alternative methods in terms of fake news detection accuracy and precision on the WEIBO dataset.

Detailed Analysis. DTR and SVM-TS deliver poor performances on fake news detection because both methods are solely reliant on hand-crafted features. DTR's subpar performance is attributed by the insufficient coverage of patterns described by regular expressions. SVM-TS performs relatively better than DTR thanks to the incorporation of temporal information. The poor detection accuracy of DTR and SVM-TS indicates that hand-crafted features are inadequate for encoding semantic information of news content. Unlike our *FNEPP*, DTR and SVM-TS fail to capture complex feature interactions - key players in fake news detection. GRU integrates the temporal linguistic features, thereby being superior to DTR and SVM-TS. CSI detects fake news from a social context perspective that combines the user engagements and their responses in the model. The performance of GRU suggests that deep learning models can learn semantic representations while enhancing feature interactions. CSI performs slightly better than GRU in the two datasets, implying

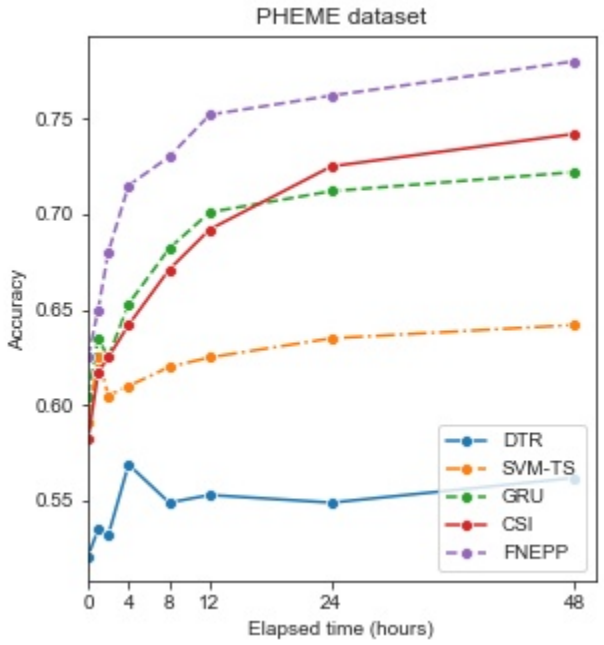

Figure 3.2: Results of Fake News Early Detection on PHEME Dataset

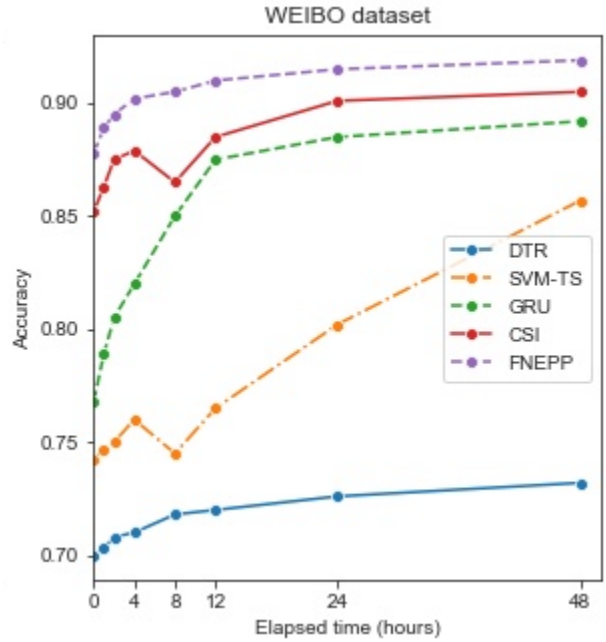

Figure 3.3: Results of Fake News Early Detection on WEIBO Dataset

that the detection accuracy is improved with the provision of valuable social context. Compared to CSI, *FNEPP* incorporates user characteristics along with propagation path, which significantly improves the accuracy of fake news detection. In a nutshell, our results confirm that our *FNEPP* boosts the overall performance of the fake news detection system by the virtue of modeling four perspectives of the social context.

Early Fake News Detection. One of the crucial aspects of fake news detection is to identify fake news in an early dissemination stage on social media. Early alerts essentially prevent further spreading of fake news harassing social media users' opinions. In order to evaluate the performance of early fake news detection, we compare multiple methods by varying time interval $[0, T]$. Accuracy performance is attained by progressively adding data up to a checkpoint T while maintaining the desired time interval. Figure 3.2 and 3.3 plot the accuracy of all the competitors as a function of parameter T. All the methods enjoy accuracy improvement over time. More importantly, our model exhibits a clear advantage over the alternatives at an early stage. Our model swiftly learns to detect fake news, achieving good performances using less than 8-hour data. In particular, the results on the WEIBO dataset reveal that *FNEPP* outperforms GRU, SVM-TS, and DTR using the 4-hour data. Not surprisingly, similar trends are observed from the PHEME dataset. After 8 hours of spreading fake news, *FNEPP* dramatically surpasses all the compared models. These results confirm the advantage of our model over the existing solutions in terms of early fake news detection.

3.5 Summary

To model fake news detection from a social context perspective, we developed in this chapter a novel fake news detection framework called *FNEPP* that seamlessly incorporates news contents, user engagements, user characteristics, and propagation paths using two cooperative modules.

We conducted extensive experiments driven by two real-world datasets to shed light on the effectiveness of *FNEPP*. We demonstrated the capability of capturing distinctive temporal patterns between fake and real news. The promising results unfold the high efficiency of *FNEPP* in the realm of detecting fake news on social media at an early stage.

Chapter 4

Emotion-Aware (EmoAware) Fake Review Detection Framework

Before the framework design, we observe that the existing studies devoted epic efforts on examining the linguistic features, user behavior features, and other auxiliary features in fake review detection techniques. Emotion aspects conveying in the reviews, unfortunately, haven't yet been well explored. In marketing literature, brands and merchandises tend to utilize the emotion appeal to interest potential customers. Similar strategies are often adopted in the realm of detecting fake product reviews. Evidence has shown that fake product reviews attempt to achieve the fake reviewers' objectives by evoking strong emotional feelings, including fear, anger, and passion, rather than by a rational appeal. This phenomenon could occur when two competitive merchandises hired people to write fake reviews against each other or write highly activated emotional fake reviews to exaggerate their products. In this chapter, we carefully develop a new system - called *EmoAware* - to examine a way of taking full benefits of emotion aspects to optimizing fake review detection techniques. Another interesting question we address in this part of the dissertation study is how people perceive fake and real reviews differently in terms of emotion aspects.

In the dissertation research articulated in this chapter, we first examine the motivation of fake review detection, especially exploring how emotion conveyed in the review text helps improve the performance of fake review detection. We provide the literature of the definition of "fake review". The existing approaches of fake review detection have been thoroughly and qualitatively compared: we place our EmoAware framework side by side with the state-of-the-art emotion models in the text mining literature, which motivates our proposed emotion representations in fake review detection. More importantly, we spearhead the development of EmoAware - an emotion-aware fake review detection framework inspired by ensemble

learning methods. Three perspectives of modeling emotion conveyed in the review text are seamlessly integrated in the framework.

We carry out extensive experiments to glean the results highlighting the key factors of why people believe in fake review in terms of emotion, and inspire making guidelines to avoid fake reviews. More specifically, a series of experiments on two real-world datasets demonstrate the effectiveness of our proposed model. Importantly, we conduct a survey-based qualitative analysis, expecting to evaluate how human perceive fake review differently compared with machine learning models.

The rest of this chapter is organized as follows. In section 4.1, we first present the challenges of fake review detection strategies, followed by illustrating the basic ideas and intuition of the design of the EmoAware framework. In Section 4.2, we systematically shed bright light on the emotion representations of reviews - such a representation approach is inspired by the discrete and dimensional emotion models. Section 4.3 elaborates on the EmoAware framework, in which each underpinning module is detailed. We conduct the empirical evaluations on the two real-world datasets in Section 4.4.3. we design quantitative experiments to explore how people perceive fake reviews differently in terms of emotions in section 4.6. Lastly, the summary of the chapter can be found in section 4.7.

4.1 Challenges and Basic Ideas

According to a report from Spiegel Research Center, slightly over 90% of consumers turn to online reviews for suggestions before making purchases. As people tend to rely on product reviews prior to making purchasing decisions, fake reviews have been manipulated and used to achieve the purposes of boosting seller's profit or harming the brands of competitors. In this section, we first post the challenges encountered in the development of fake review detection mechanisms in Section 4.1.1. Next, we illustrate the basic idea of our solution in Section 4.1.2, which summarizes the motivation and intuition of our proposed EmoAware framework depicted in Section 4.3.

4.1.1 Challenges

Before introducing our basic ideas proposed in the second part of the dissertation, we shed light on a series of challenges to be resolved in this chapter.

- Challenge 1. There is the lack of suitable fake product review datasets.
- Challenge 2. There is a growing need for appropriate emotion representations catered for product reviews.
- Challenge 3. Accurately detecting fake reviews is a non-trivial task.

Challenge 1. There is the lack of suitable fake product review datasets. There are millions of reviews available over the Internet, but very little amount of the review data are labelled as fake or real - lacking such ground truth becomes a daunting challenge to validate modern fake-review detection techniques. Accurate labelling fake product reviews, of course, is extremely labor-intensive and cost-inefficient. To address this major concern, we adopt the two standard fake product review datasets, namely the Amazon dataset and the OSF dataset in our experiments.

Challenge 2. There is a growing need for appropriate emotion representations catered for product reviews. Few studies have explored the roles of emotion in the arena of fake review detection. Prior research efforts considering emotion are mostly related to use the sentiments of review as auxiliary features to assist fake review detection systems. Inspired by the widely used emotion models in emotion distribution learning, we propose a thorough emotion representations of review text - the representations incorporate multifaceted features including emotion distribution, emotion intensity, and emotion dimensionality.

Challenge 3. Accurately detecting fake reviews is a non-trivial task. As fake reviews are designed purposefully to deceive readers, it is non-trivial to devise software systems to automatically discover fake reviews solely based on semantic meanings. Previous

research devoted epic efforts on examining the linguistic features, user behavior features, and other auxiliary features in fake review detection. Emotion aspects conveying in the reviews, which are still in an early development phase, haven't yet been well explored. In marketing literature, brands and merchandises tend to utilize the emotion appeal to interest potential customers, and similar strategies are often adopted in fake product reviews. Evidence has shown that fake product reviews attempt to achieve the fake reviewers' objectives by evoking strong emotional feelings - including fear, anger, and passion - rather than by a rational appeal [87]. Therefore, we investigate how much performance the emotion features can boost accuracy of fake review detection systems, the framework of which will be thoroughly investigated in this chapter.

4.1.2 Basic Ideas

Our proposed emotion-aware fake review detection framework is inspired by two methodologies, namely, emotion modeling and ensemble learning. Emotion models, determining how emotions are expressed, presume that emotions exist in different states, necessitating a need to discriminate between them. Various forms of expressing emotions are showcased in [5], but discrete and dimensional emotion models are of critical importance to this study. In what follows, we briefly introduce discrete emotion models (DEMs) and dimensional emotion models (DiEMs).

- Discrete emotion models or DEMs: A discrete model of emotions considers placing emotions into distinct classes or categories.

- Dimensional emotion models or DiEMs: A dimensional model presupposes that emotions are not independent: there exists a relation among emotions that ought to be represented in a spatial space.

In our EmoAware framework, we adopt both discrete emotion modeling and dimensional emotion modeling methods to thoroughly model the emotion aspects conveyed from product review text.

Ensemble learning models, taking full advantage of a variety of machine learning models, are able to approach the classification task from multiple perspectives that might be impossible to be addressed in an individual classifier. The advantages of ensemble learning models are threefold.

- Firstly, ensemble models, by and large, achieve better performance than those of individual models.
- Secondly, ensemble methods do not suffer from overfitting or underfitting thanks to suppressed bias and variance.
- Thirdly, ensemble models are usually stable and less noisy.

Given the above three impressive benefits, we opt for the ensemble learning method as a technological underpinning of the EmoAware framework - a novel design that utilizes the weighted average ensemble methods to obtain optimized final predictions.

4.2 Emotion Representations

To comprehend the emotional aspects of reviews and their impact on review consumers, we propose to model the emotions conveyed from reviews. We construct our emotion representations for the aforementioned three emotion sources from various aspects of emotions, including emotion distribution, emotional intensity, and emotion dimensionality.

4.2.1 Emotion Distribution

The emotions conveyed by a piece of text are often leveraged by several emotion-indicating words that are annotated in the emotion lexicons such as NRC Word-Emotion

Association Lexicon [64]. Each emotion-indicating word associates with one specific emotion. For example, "sad" expresses sadness, whereas the word "angry" indicates anger. By examining these words throughout the review text, we can extract the emotion distribution from the review text. The distribution can serve as one of the effective representations of the emotions expressed in the review text. For instance, Figure 4.1 demonstrates that the emotion distribution of a sentence when considering six different emotion categories. We refer to the "NRC Word-Emotion Association Lexicon", also known as "EmoLex", as our emotional lexicons [64]. EmoLex is a collection of English words associated with eight fundamental emotions, namely, anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. EmoLex also annotate the positive and negative emotion expressed by different words.

Given a piece of text, emotion distribution aims to examine the frequencies of each emotion category. To formalize the extraction of the emotion distributions, we made the following assumptions. We first model a piece of text as a sequence of meaningful word collections $\mathbb{W}$,

$$\mathbb{W} = \{w_1, w_2, ..., w_K\} \tag{4.1}$$

where K is the length of the text after preprocessing. We utilize the set $\mathbb{E}$ to represent the N different emotion categories. The set $\mathbb{E}$ can be written as:

$$\mathbb{E} = \{e_1, e_2, ..., e_N\} \tag{4.2}$$

For the case of "NRC Word-Emotion Association Lexicon", we utilize eight aforementioned fundamental emotions along with positive and negative sentiments. In the lexicon dictionaries, each type of emotion associates with a set of words that expresses the emotion. For instance, words such as "furious", "irate" belong to the emotion "anger". For a specific emotion type e_n, its associated emotional words can be represented as a set $\mathbb{V}_{e_n}$.

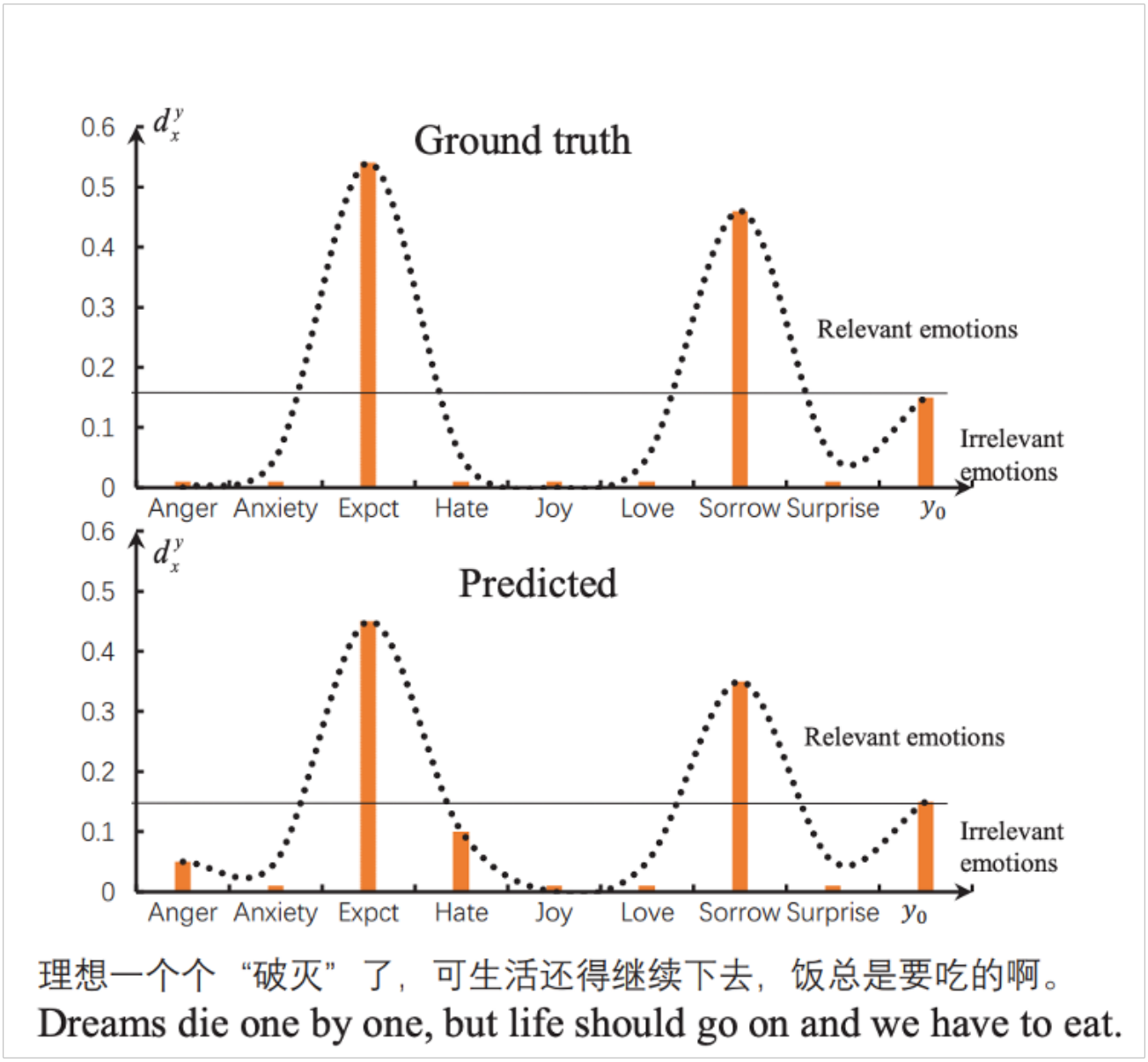

Figure 4.1: An Example of Emotion Distribution of A Sentence

$$\mathbb{V}_{e_n} = \{v_{e_n,1}, v_{e_n,2}, \ldots, v_{e_n,M}\} \tag{4.3}$$

M is the total number of words expressing the emotion e_n.

Given the above assumptions, we calculate the frequency distribution of a specific emotion e_n as follows:

$$emo_{dist}(\mathbb{W}, e_n) = \frac{\sum_{w_k \in \mathbb{W} \cap \mathbb{V}_{e_n}} count(w_k, \mathbb{W})}{K} \tag{4.4}$$

where $count(w_k, \mathbb{W})$ is the number of times that word w_k appears in the text $\mathbb{W}$.

After calculating the frequency distribution of all emotions in the set E, the whole emotion distribution of the text $\mathbb{W}$ is a vector whose entries are the frequency distribution of each emotion e_n.

$$emo_{dist}(\mathbb{W}, \mathbb{E}) = [emo_{dist}(\mathbb{W}, e_1), emo_{dist}(\mathbb{W}, e_2), \ldots, emo_{dist}(\mathbb{W}, e_N)]^\intercal \tag{4.5}$$

Note that $emo_{dist}(\mathbb{W}, \mathbb{E})$ represents the emotion distribution of the text information, and is distinctive and indicative feature of use when discerning the veracity of the reviews.

4.2.2 Emotion Intensity

We presented the distinctive emotion distribution representation within the text, which evaluate the frequencies of discrete emotion categories. However, the intensity of each emotion category is essential to be assessed. Words in the same emotion category may express different intensities [23, 116]. For example, the word "furious" is much stronger than the word "angry" when describing the "anger' emotion. In order to capture the full characteristics of emotions, we propose to model the emotion intensity feature. The process of extracting emotion intensity feature is similar to the emotion distribution extraction process. All the assumptions made in the subsection 4.2.1 still hold when extracting the emotion intensity feature. Given a set of emotion categories $\mathbb{E}$, the set of emotion words $\mathbb{V}_{e_n}$ for the given emotion e_n, and the text as word collections $\mathbb{W}$. We calculate the intensity for a specific emotion e_n of text $\mathbb{W}$ as follows:

$$emo_{int}(\mathbb{W}, e_n) = \sum_{w_k \in \mathbb{W} \cap \mathbb{V}_{e_n}} intensity(w_k, e_n) \tag{4.6}$$

The function $intensity(w_k, e_n)$ is a mapping from word w_k to its intensity of emotion category e_n. The emotion intensity distribution for the given text $\mathbb{W}$ is a vector whose entries are the intensity of each emotion category within the text $\mathbb{W}$.

$$emo_{int}(\mathbb{W}, \mathbb{E}) = [emo_{int}(\mathbb{W}, e_1), emo_{int}(\mathbb{W}, e_2), \ldots, emo_{int}(\mathbb{W}, e_n)]^\intercal \tag{4.7}$$

Note that $emo_{int}(\mathbb{W}, \mathbb{E})$ represents the emotion intensity distribution of the text information and is another indicative feature of use when discerning the veracity of reviews.

4.2.3 Emotion Dimensionality

Motivated by the PAD emotional state model [60], we examine the emotion states conveyed from the review text. PAD emotional state model is one of the dimensional emotion models, which utilizes three numerical dimensions, Pleasure, Arousal and Dominance to represent all emotions. We refer to the NRC Valence, Arousal, and Dominance (NRC-VAD) Lexicon. Each word is annotated with different level of pleasure, arousal, and dominance that are within $[0, 1]$ scale.

To formulate the calculation of emotion dimensionality of a given review text, we assume that the set of words $\mathbb{U} = \{u_1, u_2, ..., u_L\}$ contains words that associate with levels of pleasure, arousal, and dominance. The set of vectors $\mathbb{T} = \{\bar{t_{u_1}}, \bar{t_{u_2}}, ..., \bar{t_{u_L}}\}$ contains the level of pleasure, arousal, and dominance for a given word. Given the text $\mathbb{W}$, the emotion dimensionality feature can be represented as follows:

$$emo_{dim}(\mathbb{W}, \mathbb{U}) = \sum_{w_k \in \mathbb{W} \cap \mathbb{U}} \bar{t_{w_k}} \tag{4.8}$$

Note that $emo_{dim}(\mathbb{W}, \mathbb{U})$ represents the emotion dimensionality distribution of the text information and is another indicative feature of use when discerning the veracity of reviews.

4.3 Emotion-Aware Fake Review Detection Framework

In this section, we present a systematically way of integrating the three aforementioned emotion features into the fake review detection framework, which judiciously utilize the power of ensemble learning methods. Recall that ensemble learning schemes take advantage of a handful of machine learning models: the ensemble learning models are able to approach the classification task from many different perspectives that might be impossible for an individual classifier. Ensemble learning models demonstrate three salient strengths. Firstly, ensemble models can typically achieve better performance compared to individual models. Secondly, ensemble methods stay far away from overfitting or underfitting because of reduced bias and variance. Thirdly, ensemble models are usually stable and less noisy. Considering these advantages of ensemble learning methods, we advocate for ensemble learning as a center piece to construct design our detection framework, the high-level architecture of which is depicted in Figure 4.2.

The framework consists of the following four modules, which are sequentially running in a batch manner. This sequential design is quite scalable because we are positioned to build a four-stage pipeline to facilitate big data processing.

- The original review module.
- The feature extraction module.
- The base classifiers module.
- The result aggregation module.

Let us put these four core components under the microscope. The framework starts with examining original review text in the original reviews module. Given product review text, we perform preprocessing steps such as: stop-words removal, url deleting, punctuation marks removal. After the preprocessing phase is accomplished, the review text will be fed into the feature extraction module. During the second stage in the processing pipeline, we

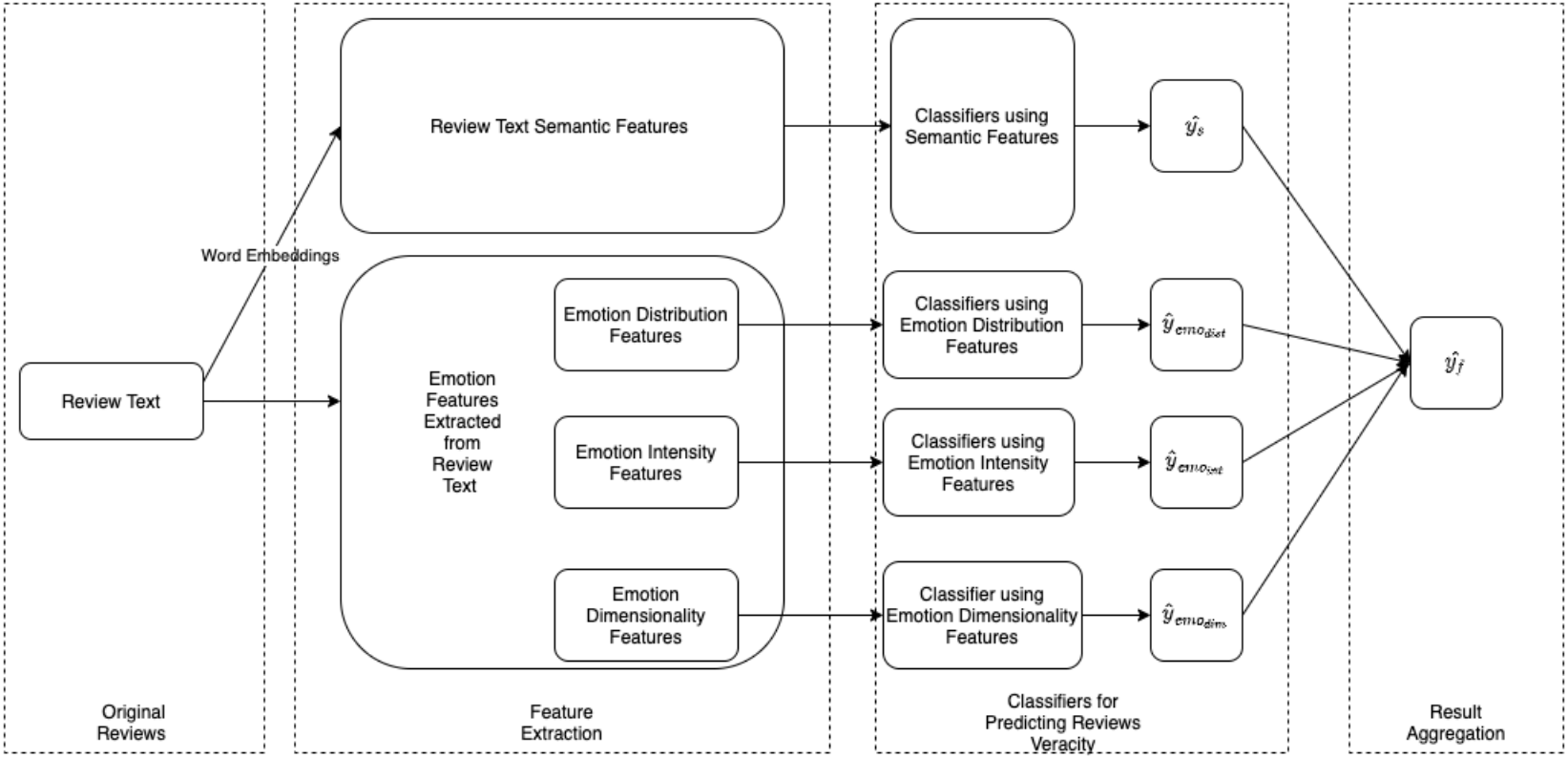

Figure 4.2: The High-level Architecture of the Emotion-Aware Fake news Detection Framework.

first extract the semantic features of review text by creating word embeddings. Emotion distribution features, emotion intensity features, and emotion dimensionality features are extracted based on the equations presented in Section 4.2.

After creating the semantic feature and three emotion features, we input these features to the base classifiers module. In this module, four classifiers are trained to predict the veracity of reviews. A classifier using semantic features solely is trained to predict the label of the review, $\hat{y}_s$. Similarly, the classifier using emotion distribution features predicts the label of the review as $\hat{y}_{emo_{dist}}$. The classifier anchored on emotion intensity features predict the label of the review as $\hat{y}_{emo_{int}}$. The classifier powered by emotion dimensionality features predicts the label of the review as $\hat{y}_{emo_{dim}}$.

As we collect the predicted label of the above base classifiers, we make use of weighted average ensemble to combine all the four predictions as a final prediction $\hat{y_f}$. To obtain $\hat{y_f}$, we first record the probabilities of the review being real from all four classifiers. We then perform a grid search on the training set to derive the best set of weights - a key part of the performance optimization process. Finally, we calculate the probability of the review being real by averaging the four probabilities. The final label $\hat{y_f}$ is obtained by the cut-off probability being set to 0.5.

4.4 Experimental Design

We first introduce the details and statistics of two real-world datasets, namely Amazon dataset and OSF dataset, in Section 4.4.1. Section 4.4.2 presents the baseline methods, followed by the evaluation metrics used in this study. Lastly, we illustrate the design of the empirical study in Section 4.4.3.

4.4.1 Datasets

In the experimental design, we are focusing on two benchmark datasets, namely, the Amazon dataset and the OSF dataset. Table 4.1 summarizes the descriptions of the two

datasets. The Amazon dataset was initially released and hosted on the Kaggle website - a popular data repository platform that renders a no-setup, customizable, Jupyter Notebooks environment. The reviews in the Amazon dataset are labelled as either real or fake. For each review, the data set includes a number of additional attributes, including ratings, verified purchases, product categories, product IDs, product titles, and review titles. The Amazon dataset consists of 21,000 reviews, which are evenly distributed across product categories.

Dataset	# of fake reviews	# of real reviews	Avg. review length	Fake/real avg. review length
Amazon	10,500	10,500	69.2	59.3/79.1
OSF	20,216	20,216	67.5	61.3/73.6

Table 4.1: Fake Review Datasets Used in This Study

Similar to the first dataset, the OSF dataset [87] contains a mixture of real and fake reviews. Given the publicly available Amazon Review Data (2018) dataset [67], the fake reviews were generated by GPT-2 model [78]. Therefore, the OSF dataset can be viewed as a review dataset that is comprised of fake news originated by a machine learning model: the OSF dataset is a well-balanced dataset with 20,216 real reviews and 20,216 fake reviews. Given these two intrinsically different datasets, we are capable of gauging the performance measures of the implemented machine learning models processing and handling human generated fake reviews accompanied by machine generated fake reviews.

4.4.2 Baseline Methods and Evaluation Metrics

Before diving into our major findings, we first bring forth the alternative models - four competitors to our EmoAware - and the three performance evaluation metrics. We select the state-of-the-art end-to-end deep learning models as the alternative models because we maintain fair comparisons. The state-of-the-art language model BERT (Bidirectional Encoder Representations from Transformers) has been proven to be effective and widely used in existing fake review detection systems [80]. Therefore, we choose BERT and its variants as the competitors to EmoAware - our solution that is expected to be a front runner. The alternative models implemented for comparison purpose are tabulated below.

- BERT: is a transformer-based pre-trained model proposed by the Google AI team [14]. The BERT model is pre-trained on text paragraphs containing 2.5 billion words and books corpus of English language, which is available on Wikipedia.

- RoBERTa: is named as t [49]. RoBERTa deploys the same strategy as that of the BERT model. While combined with better training methods, RoBERTa tends to outperform BERT in a selection of NLP tasks [49].

- DistilBERT: is a distilled version of the BERT pre-trained model. DsitilBERT - proposed by Sanh *et al.* [88] - is a lighter and faster version of BERT, utilizing fewer parameters and still preserving high performance.

- ALBERT: is named as A Lite BERT. This version of BERT model improves several drawbacks of original BERT model - a baseline solution that exhibits long training time and suffers from memory limitations. [35].

To test the performance of the emotion-aware fake review detection framework, we elect an array of standard evaluation metrics catering for binary classification tasks. The evaluation metrics used throughout this chapter are precision, accuracy, recall, and F1-score.

- Accuracy: This performance metric represents the accuracy of a model to be evaluated in our empirical study. The accuracy measure is calculated by dividing the number of correct predictions by the total number of predictions.

- Precision: Precision is the percentage of positive predictions from the total predicted positive data samples.

- Recall: This metric tests a model's capacity of finding positive samples, thereby showing how many times positive predictions were wrong.

- F1-score: F1-score - a mixed bag of measures - is a harmonic mean of precision and recall.

4.4.3 The Design of An Empirical Study

We design our empirical evaluations in two parts. The first part aims to systematically evaluate our proposed emotion-aware fake review detection framework using two real-world datasets. We compare our solution with the existing baseline schemes coupled with the state-of-the-art models in various evaluation metrics. The second part aims to explore how humans perceive fake reviews differently in terms of emotion. Therefore, we designed a survey-based study to let participants rate the credibility of reviews. Next, we conduct statistical tests to determine whether there are significant differences between human's perception of fake reviews and the machine learning model's perception of fake reviews.

4.5 Overall Performance and Robustness Comparisons

In this section, we first test the performance and robustness of our proposed emotion-aware fake review detection framework by varying the base models within the framework (see Section 4.5.1). Then, in Section 4.5.2, we compare our best set of model selections with those mentioned above in end-to-end alternative models. Finally, in section 4.5.3, we vary the combination of emotion features and present the results of the ablation study.

4.5.1 Models for Emotion-Aware Fake Review Detection Framework

Recall the framework shown in Figure 4.2. There are four classifiers presented in the model. The classifiers can be divided into two types: semantic feature classifiers and emotion feature classifiers. In order to test the performance and robustness of our proposed framework, we choose to vary the combination of two types of classifiers. The implemented classifiers are listed in Table 4.2.

Semantic Feature Classifiers	Emotion Feature Classifiers
LSTM	SVM
Bi-LSTM	MLP (Multilayer Perceptron)
BERT	

Table 4.2: Implemented Models for Emotion-Aware Fake Review Detection Framework

Figure 4.3 and Figure 4.4 present the overall comparison results on Amazon and OSF datasets. Accuracy, precision, recall, and F1-score are reported in the figures. The x-axis represents the various combinations of the semantic classifier and emotion classifiers. The y-axis shows the value of each evaluation metric on the scale of $[0, 1]$. For example, the rightmost bars in Figure 4.3 are the results of using BERT as the semantic feature classifier and MLP as the emotion feature classifiers.

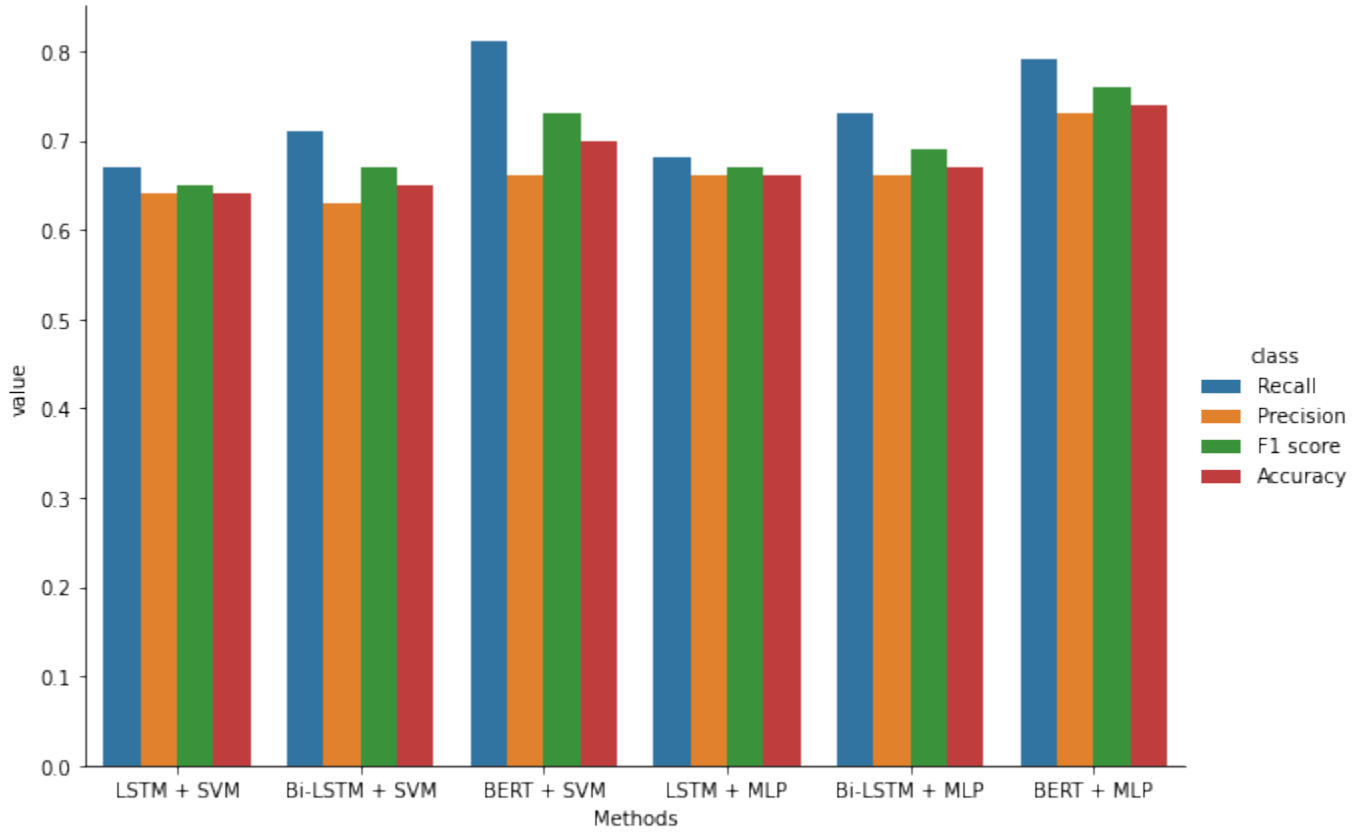

Figure 4.3: Overall Comparison on Amazon Dataset

From Figure 4.3, we found that when choosing BERT as the semantic feature classifier and MLP as the emotion feature classifiers achieves the best performance. The recall, accuracy, and precision are within a similar range for all combinations of classifiers. Utilizing BERT as the semantic classifier usually boosts the performance because of its capability to deal with contextual-based paragraphs. Using LSTM combined with other emotion features

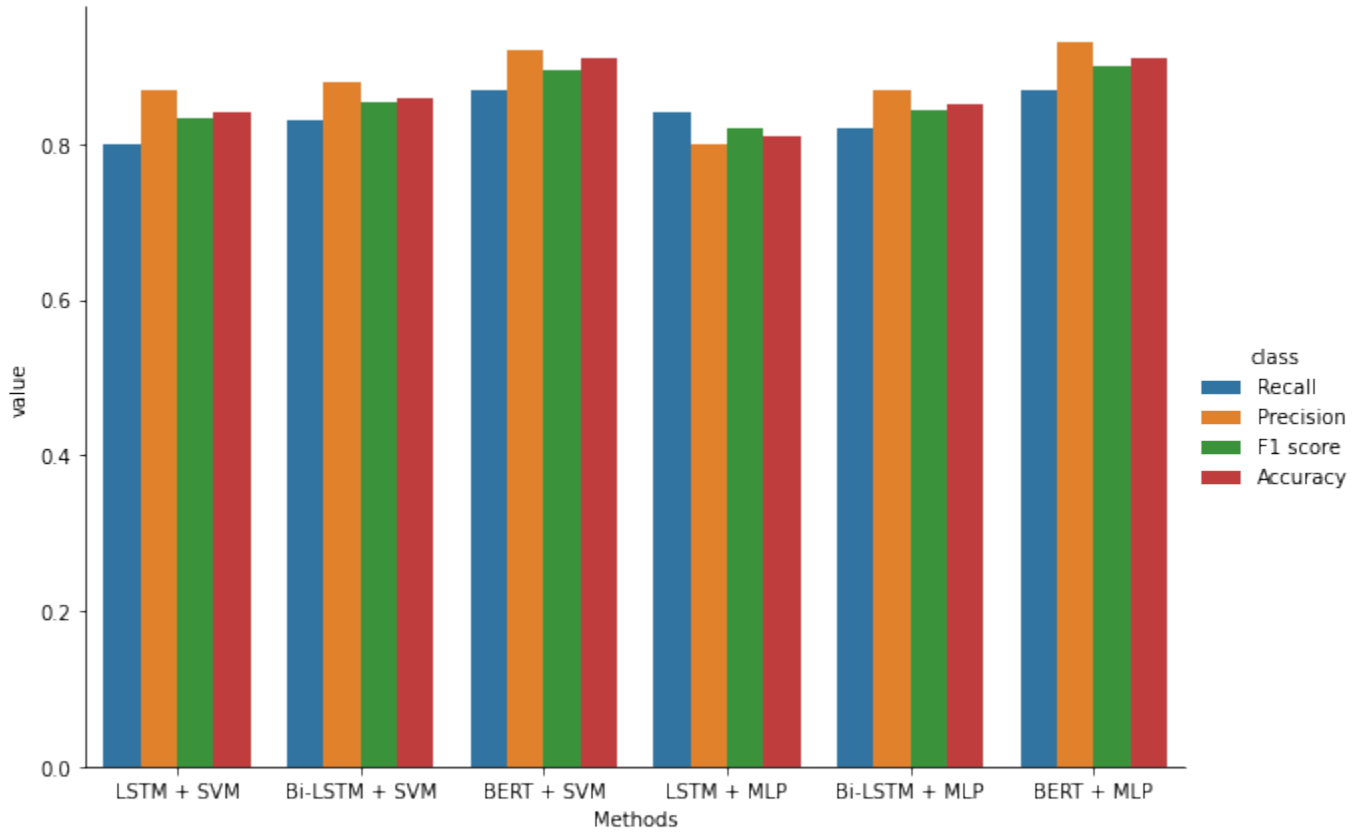

Figure 4.4: Overall Comparison on OSF Dataset

performs the worst because LSTM cannot deal with backward contextual dependencies. Similar patterns can be observed in Figure 4.4. However, we noticed that our emotion-aware fake review detection framework achieves better results in the OSF dataset. This phenomenon indicates that the machine learning model-generated fake reviews are more trivial to be detected.

4.5.2 End-to-end Model Comparison

In this section, we compare our proposed EmoAware framework with the alternative end-to-end model. We choose to utilize BERT as the semantic classifier and SVM as emotion feature classifiers. The results of the two datasets are shown in Table 4.3 and 4.4. Our proposed *EmoAware* framework outperforms all 4 BERT-based end-to-end models. Our EmoAware utilizes BERT as the semantic feature classifier and SVM as emotion feature

classifiers. For the Amazon dataset, it improves the accuracy of ALBERT, BERT, DistilBERT, and RoBERTa by 3%, 2.1%, 1.8%, and 1.3%, respectively. Similar patterns can be observed in the OSF dataset. For the OSF dataset, it boosts the accuracy of ALBERT, BERT, DistilBERT, and RoBERTa by 8.5%, 6.5%, 3.7%, and 4.8%, respectively. Besides the improvement of accuracy, other evaluation metrics have a significant promotion. Therefore, we confirm that our proposed *EmoAware* has the advantage of better performance in the fake review detection task.

Methods	Class	Accu.	Prec.	Rec.	F_1-Score
ALBERT	R	0.658	0.633	0.715	0.672
	F		0.626	0.724	0.671
BERT	R	0.672	0.643	0.735	0.686
	F		0.638	0.754	0.691
DistilBERT	R	0.687	0.663	0.732	0.696
	F		0.658	0.757	0.704
RoBERTa	R	0.692	0.668	0.747	0.705
	F		0.665	0.745	0.703
EmoAware	R	0.705	0.685	0.783	0.731
	F		0.662	0.810	0.729

Table 4.3: Amazon Dataset: EmoAware v.s. End-to-End Methods, Semantic Classifier: Bert; Emotion Classifiers: SVM ("F: " fake reviews; "R": real reviews)

Methods	Class	Accu.	Prec.	Rec.	F_1-Score
ALBERT	R	0.807	0.821	0.814	0.817
	F		0.793	0.810	0.801
BERT	R	0.825	0.835	0.819	0.827
	F		0.820	0.823	0.822
DistilBERT	R	0.873	0.853	0.879	0.866
	F		0.857	0.883	0.870
RoBERTa	R	0.862	0.849	0.871	0.860
	F		0.852	0.877	0.864
EmoAware	R	0.910	0.909	0.875	0.892
	F		0.921	0.870	0.894

Table 4.4: OSF dataset: EmoAware v.s. End-to-End Methods, Semantic Classifier: Bert; Emotion Classifiers: SVM ("F: " fake reviews; "R": real reviews)

Methods	Class	Accu.	Prec.	Rec.	F_1-Score
BERT	R	0.672	0.643	0.735	0.686
	F		0.638	0.754	0.691
BERT + EmoDist	R	0.676	0.647	0.737	0.689
	F		0.644	0.758	0.696
BERT + EmoInt	R	0.683	0.651	0.728	0.687
	F		0.653	0.752	0.699
BERT + EmoDim	R	0.691	0.668	0.761	0.711
	F		0.652	0.775	0.708
BERT + EmoDist + EmoInt	R	0.685	0.662	0.731	0.695
	F		0.645	0.755	0.696
BERT + EmoDist + EmoDim	R	0.689	0.643	0.752	0.693
	F		0.663	0.773	0.714
BERT + EmoInt + EmoDim	R	0.696	0.675	0.767	0.718
	F		0.653	0.789	0.715
EmoAware	R	0.705	0.685	0.783	0.731
	F		0.662	0.810	0.729

Table 4.5: Ablation Study on Amazon Dataset, Semantic Classifier: Bert; Emotion Classifiers: SVM (”F: ” fake reviews; ”R”: real reviews)

Methods	Class	Accu.	Prec.	Rec.	F_1-Score
BERT	R	0.825	0.835	0.819	0.827
	F		0.820	0.823	0.822
BERT + EmoDist	R	0.828	0.837	0.823	0.830
	F		0.822	0.818	0.820
BERT + EmoInt	R	0.827	0.832	0.822	0.827
	F		0.829	0.821	0.825
BERT + EmoDim	R	0.831	0.840	0.825	0.832
	F		0.829	0.832	0.830
BERT + EmoDist + EmoInt	R	0.847	0.853	0.856	0.854
	F		0.851	0.831	0.841
BERT + EmoDist + EmoDim	R	0.880	0.883	0.873	0.878
	F		0.892	0.871	0.881
BERT + EmoInt + EmoDim	R	0.874	0.882	0.854	0.868
	F		0.887	0.869	0.878
EmoAware	R	0.910	0.909	0.875	0.892
	F		0.921	0.870	0.894

Table 4.6: Ablation Study on OSF Dataset, Semantic Classifier: Bert; Emotion Classifiers: SVM (”F: ” fake news; ”R”: real news)

4.5.3 An Ablation Study

Our proposed *EmoAware* framework consists of three emotion feature classifiers. In order to identify which emotion feature representations and classifiers contribute most to the prediction, we conduct an ablation study on our proposed framework. In this study, we select BERT as the semantic feature classifier and SVM as the emotion feature classifiers. Recall that the whole framework utilizes emotion distribution features (EmoDist), emotion intensity features (EmoInt), and emotion dimensionality (EmoDim). We vary the presence of these features in the experiments. Table 4.5 and 4.6 demonstrate the results of the ablation study. *EmoAware* using all emotion features provides the best performance in terms of all four evaluation metrics. When using one emotion feature, the emotion dimensionality feature boosts the performance most. This phenomenon indicates that the emotion dimensionality feature has more distinctiveness between fake reviews and real reviews. When two emotion features are included in the model, the combination of emotion intensity and emotion dimensionality features provides the best performance. This confirms that emotion dimensionality and emotion intensity play significant roles in improving the performance of fake review detection.

4.6 Human Perception of Product Reviews Based on Emotions

One research question that particularly interests us is how people perceive fake reviews in terms of emotions. Two directions of human perception of fake reviews will be addressed in this section.

- How do people perceive fake reviews differently compared with real ones? Due to the scope of our dissertation, we are particularly interested in the emotion perspectives. For instance, what emotion categories of human-labeled fake reviews and human-labeled real reviews have distinctions?

- How do people perceive reviews differently compared with machine learning models' perception in terms of emotion?

Next, we will describe the design of quantitative experiments to address these research questions.

4.6.1 Data Collection

In order to make fair comparisons between human perception of fake reviews and machine learning models' perception of fake reviews, we randomly selected 100 fake reviews and 100 real reviews from the test set of the Amazon dataset. For the OSF dataset, we also randomly extracted 100 fake reviews and 100 real reviews from the test set of the OSF dataset. In total, we have 400 reviews in a well-balanced dataset.

We created a survey and distributed it to participants via Auburn Qualtrics. Each participant will be provided with 20 random product reviews, which include a mix of real and fake reviews. Participants will read carefully through each product review content and assess them based on the three criteria provided (i.e., credibility, authenticity, and believability). All three criteria reflect how believable these reviews appear to participants on a scale from 1 (Low) to 100 (High). To ensure the validity and reliability of the survey, we use three measurement items to describe the perceived credibility of provided product reviews. Figure 4.5 is one example question from a randomly selected survey.

We collected 306 valid survey responses. After deleting outliers (inconsistency of rating three items of the same questions), we have 298 valid responses. Since we randomly assigned 20 reviews to each survey participant via Auburn Qualtrics, two reviews from the Amazon dataset were not assigned to any survey participant. Thus, we have 398 reviews that have been rated by participants. Table 4.7 summarizes the descriptive statistics of the survey results.

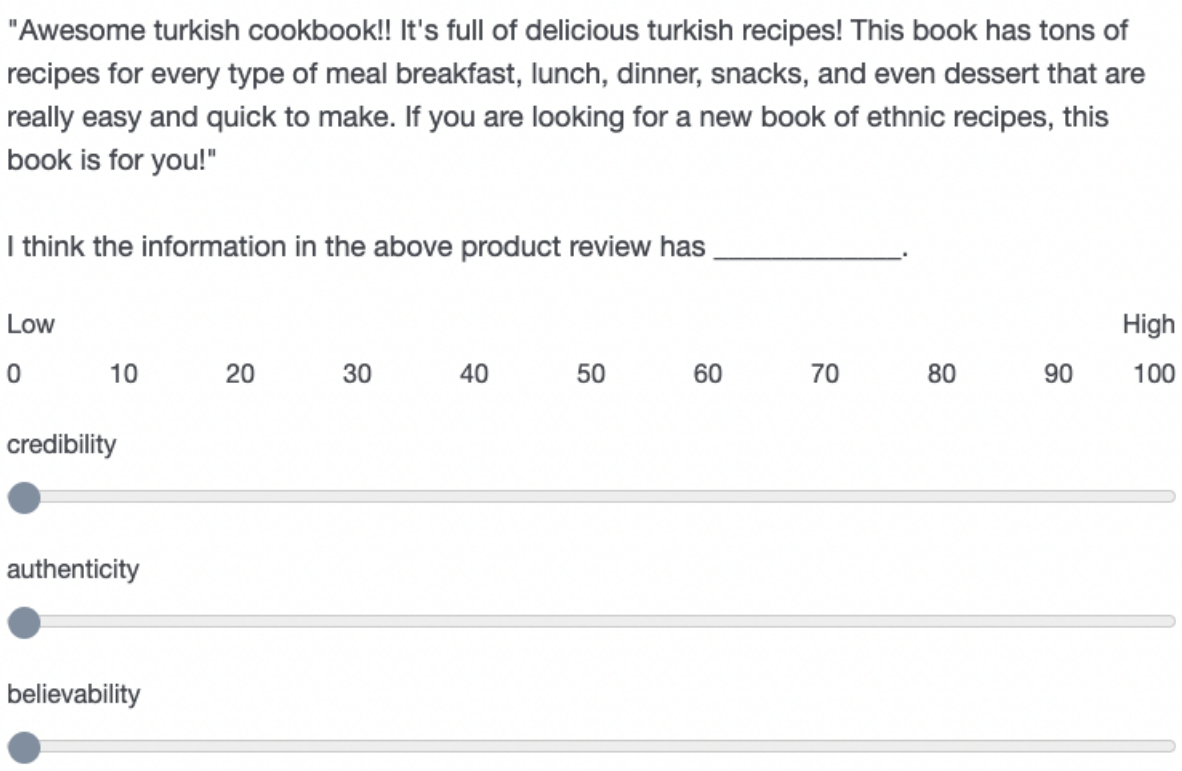

Figure 4.5: An Example of Survey Questions

Dataset	# of fake reviews used in the survey	# of real reviews used in the survey
Amazon	99	99
OSF	100	100

Table 4.7: Survey Data Information, 398 Valid Responses in Total.

4.6.2 Overall Performance

After collecting the survey data, we calculated the accuracy, precision, recall, and $F1$-Score of human judgment on reviews. For each survey participant, we averaged their ratings of credibility, authenticity, and believability of each rated review. Next, we took the average of each review's rating as the final rating. Since the rating scores are on the 0-100 scale, we divided the final rating of each review by 100 so that we obtain the human's perceived probabilities of reviews being real. We chose the threshold of 0.5 as the cut-off probability of the review being real. If the human perceived probability of a certain review is greater

than 0.5, we assume that human believes the review is a real one. Otherwise, it is a fake one. The whole procedure of generating human perceived labels is demonstrated in Figure 4.6.

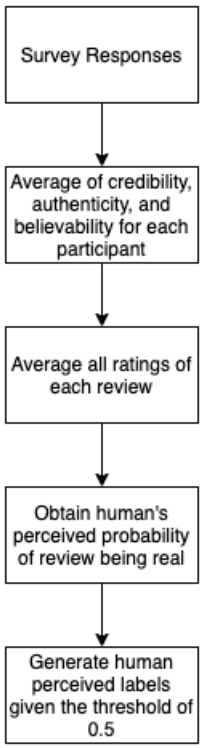

Figure 4.6: The Procedure of Generating Human Perceived Labels

After generating human perceived labels, we calculated the confusion matrix on the survey subset of the Amazon dataset and OSF dataset. Table 4.8 shows the confusion matrix for human judgment on survey subset of Amazon dataset. Participants correctly classify 107 out of 198 reviews. Note that 76 out of 99 fake reviews are correctly classified, and 68 real reviews have been wrongly classified as fake reviews. This phenomenon indicates that human shows less trust in reviews and are very conservative when referring to others' review. Similar results on the OSF dataset are shown in Table 4.9.136 out of 200 reviews are correctly classified by participants. Note that 73 out of 100 fake reviews are correctly classified, and 37 real reviews have been wrongly classified as fake reviews. We observe that people can discern the machine learning model generated fake reviews from real reviews more easily.

N = 198	Predicted Real	Predicted Fake	Support
Actual Review	31	68	99
Actual Fake	23	76	99
	54	144	

Table 4.8: Confusion Matrix for Human Judgement on Survey Subset of Amazon Dataset

N = 200	Predicted Real	Predicted Fake	Support
Actual Review	63	37	100
Actual Fake	27	73	100
	90	110	

Table 4.9: Confusion Matrix for Human Judgement on Survey Subset of OSF dataset

Besides the human evaluation of the two datasets, we also calculated the confusion matrix of our proposed EmoAware framework on two datasets. Table 4.10 shows the confusion matrix on the survey subset of the Amazon dataset. Compared to Table 4.8, EmoAware can correctly identify 25% more reviews than human judgment. EmoAware can also distinguish fake reviews from real reviews better than human-labeled. For the machine learning model generated fake reviews, our proposed EmoAware can easily detect fake reviews from real ones. Table 4.11 demonstrate the good performance of EmoAware on the OSF dataset. EmoAware achieved 98% of accuracy in identifying fake reviews. All these comparisons can confirm our motivation for proposing the emotion-aware fake review detection framework.

N = 198	Predicted Real	Predicted Fake	Support
Actual Review	58	41	99
Actual Fake	25	74	99
	83	115	

Table 4.10: Confusion Matrix for Machine Learning Model (EmoAware) on Survey Subset of Amazon Dataset

N = 200	Predicted Real	Predicted Fake	Support
Actual Review	100	0	100
Actual Fake	2	98	100
	102	98	

Table 4.11: Confusion Matrix for Machine Learning Model (EmoAware) on Survey Subset of OSF Dataset

4.6.3 Quantitative Analysis

To fully examine human's ability to predict the realness of reviews. We introduce the following two research questions.

- **RQ 1**: Is there a statistically significant difference between the human prediction of the realness of reviews and machine learning models' predictions regarding evaluation metrics, such as accuracy, precision, and recall?
- **RQ 2**: Is there a statistically significant difference in human prediction between the realness of human-generated reviews and machine learning model-generated fake reviews in terms of evaluation metrics?

To answer **RQ 1**, we define the predictions made by human judgements as group A, whereas the predictions made by machine learning models as group B. One of our research questions is to examine whether the observed proportion of evaluation metrics (accuracy, precision, recall) in group A have statistically significant difference compared to the observed proportion of evaluation metrics in group B?

In order to resolve this question, we propose to utilize two sample z test of proportions. Two sample z test of proportions is used to examine whether two populations differ significantly in terms of some particular characteristics. In other words, compare the fraction of two distinct populations that share a characteristic. It calculates the range of values that is likely to include the difference between the population proportions.

Evaluation Metrics	z-score	z-critical	p-value
Accuracy	-2.568	1.980	0.01
Precision	1.874	1.980	0.061
Recall	0.332	1.980	0.740

Table 4.12: Report of Two Sample z Test of Proportions of Group A and Group B on Amazon Dataset

Table 4.12 demonstrates the results of two sample z test of proportions on the Amazon dataset. From the second row of Table 12, we conclude that human judgment on the realness

of reviews has statistically significantly lower accuracy than the machine learning model's prediction in terms of the Amazon dataset ($p = 0.01, z = -2.568$). However, we did not find a statistically significant difference in precision and recall on the Amazon dataset.

Evaluation Metrics	z-score	z-critical	p-value
Accuracy	-8.352	1.980	6.727×10^{-17}
Precision	-6.332	1.980	2.418×10^{-10}
Recall	-5.021	1.980	5.15×10^{-7}

Table 4.13: Report of Two sample z Test of Proportions of Group A and Group B on OSF Dataset

Table 4.13 summarizes the results of two sample z test of proportions on OSF dataset. Table 4.13 shows that human judgment on the realness of reviews has statistically significantly lower accuracy, precision, and recall than the machine learning model's prediction in terms of the OSF dataset. The results of the two datasets agreed on the fact that human accuracy in predicting the realness of reviews is much lower than machine learning models.

In order to answer the **RQ 2**, we define the predictions made by human judgements on the Amazon dataset as group C, whereas the predictions made by human judgements on the OSF dataset group D. Our goal is to investigate whether the observed proportion of evaluation metrics (accuracy, precision, recall) in group C have statistically significant difference compared to the observed proportion of evaluation metrics in group D?

Evaluation Metrics	z-score	z-critical	p-value
Accuracy	-2.856	1.980	0.004
Precision	-2.179	1.980	0.029
Recall	0.613	1.980	0.540

Table 4.14: Report of Two Sample z Test of Proportions of Group C and Group D

From Table 4.14, we conclude that human judgment on human-generated reviews has statistically lower accuracy than on machine learning model-generated reviews ($p = 0.004, z = -2.856$). In terms of precision, human prediction on human-generated reviews demonstrates a statistically significant difference from machine learning model-generated reviews

($p = 0.029, z = -2.179$). We did not observe a statistically significant difference in the recall metric.

In summary, we performed the two sample z test of proportions to address the proposed two research questions at the beginning of this section. Results demonstrate two major findings:

- There is a statistically significant difference between human predictions and machine learning model predictions. Human prediction performance is often worse than machine learning models' performance in terms of accuracy, precision, and recall.
- Human prediction on human-generated reviews has statistically significantly lower accuracy and precision than on machine learning model-generated reviews.

4.6.4 LIWC-2022 Emotion Analysis

Another direction of research interests us is that how people perceive reviews in terms of emotional aspects. We investigate people's emotional aspects when consuming reviews from the following three perspectives.

- **RQ 3**: Which set of emotion categories and set of emotion dimensions can have distinctions between human perceived real reviews and fake reviews?
- **RQ 4**: Which set of emotion categories and set of emotion dimensions can have a significant difference between human perceived fake reviews and actual fake reviews?
- **RQ 5**: Which set of emotion categories and set of emotion dimensions can have a significant difference between human perceived fake reviews and machine learning model predicted fake reviews?

To answer the above research questions, we chose to perform the emotion analysis utilizing Linguistic Inquiry and Word Count (i.e., LIWC-22) text analysis software. LIWC-22 has been proven to be reliable software and has been widely adopted by previous studies.

LIWC-22 is regarded as the top standard for assessing content in the text, including individuals' cognitions and emotions (LIWC) [25]. The use of LIWC-22 allows us to establish the body of knowledge of fake reviews by emotion analysis.

LIWC-22 possesses a large dictionary that can analyze cognitive and emotional aspects. However, the dictionary does not contain detailed emotion categories and emotion dimensions, such as valence and arousal. In order to enrich the power of LIWC-22, we utilize the feature of customizing dictionaries. We first adopted the NRC Word-Emotion Association Lexicon (EmoLex), which is also used in the extraction of emotion distribution features. LIWC-22 loaded with EmoLex enables us to examine the detailed emotion categories conveyed from the text. We can examine eight emotion categories (i.e.anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) and two sentiments (i.e., positive and negative). Secondly, we utilized the NRC Valence, Arousal, and Dominance (NRC-VAD) Lexicon to investigate the difference in emotion dimensionality. LIWC-22 loaded with NRC-VAD lexicons allowed us to explore the difference in the levels of arousal, valence, and dominance embedded within the text.

The procedures of how LIWC-22 loaded with EmoLex works are as follows. LIWC-22 first received the text data, and the text was subsequently tokenized into single-word tokens. LIWC-22 then automatically counts the number of words associated with each of the eight emotion categories and computes the percentage of occurrence of eight emotion categories in each review. The eight emotion categories analyzed in this study are measured as the proportion of words that fall into each emotion category to the total number of words in the reviews. For instance, a value of 5 for the anger emotion of a review means that the proportion of words associated with anger is 5%. Thus, the ratio scale for the eight emotion categories has meaningful absolute zero points and equal intervals.

For the LIWC-22 loaded with NRC-VAD, the working procedures are the same as LIWC-22 loaded with EmoLex. Since the NRC-VAD lexicon annotates the intensity or level of valence, arousal, and dominance, the meaning of reported numbers by LIWC-22 is changed.

For example, a value of 2.7% for the arousal of a review means that the weighted intensity of arousal is 2.7%. Weights are the frequencies of words associated with arousal.

Firstly, we explore the emotional distinctiveness between fake and real reviews. We reported our results on the Amazon dataset. Table 4.15, Table 4.16, and Table 4.17 show the mean and standard deviation of the eight emotions and the Mann-Whitney U test results for comparing the emotion distribution of fake reviews and real reviews on the Amazon dataset. We selected the Mann-Whitney U test, which is a non-parametric test for comparing two independent samples. The advantage of the Mann-Whitney U test is that it does not require the assumptions of normality and equal variance. Also, it is extremely suitable for comparing two samples that are different in size.

EmoLex	Fake Review		Real Review		Mann-Whitney U test	
Emotions	Mean	SD	Mean	SD	Test Stat	p
anger	1.094	0.884	1.191	0.972	5367	0.886
anticipation	2.801	2.384	3.306	2.730	3306	0.050
disgust	0.790	0.613	1.047	0.871	4480	0.482
fear	1.285	0.794	1.107	0.695	4810	0.654
joy	2.862	2.268	2.184	1.881	3102	0.040
negative	2.107	1.514	2.930	2.424	2795	0.026
positive	5.583	4.091	4.821	3.636	2975	0.036
sadness	1.171	1.141	1.389	1.078	4635	0.576
surprise	1.411	1.056	1.273	0.949	5031.5	0.748
trust	2.914	2.278	3.434	2.443	3223	0.047

Table 4.15: LIWC-22 Loaded with Emolex; Mann-Whitney U Test on the LIWC-22 Results (Fake Reviews vs Real Reviews on Amazon Dataset)

Table 4.15 is the LIWC-22 analysis results that utilize the LIWC-22 loaded with EmoLex. We noticed that anticipation and trust have a significant difference between fake reviews and real reviews on the Amazon dataset. Positive sentiment and negative sentiment all have significant differences between fake reviews and real views. Table 4.16 summarizes the results from using LIWC-22's original library. It demonstrates that there is a significant difference in emotions between fake reviews and real ones. The results also agree with Table 4.15 that the anger and sadness do not have a significant difference.

LIWC-22 Dict	Fake Review		RealReview		Mann-Whitney U test	
Emotions	Mean	SD	Mean	SD	Test Stat	p
emotion	2.876	2.575	2.317	2.175	3127	0.042
emo_pos	2.684	2.489	1.766	1.434	2905	0.030
emo_neg	0.631	0.415	0.584	0.392	4952.5	0.900
emo_anx	0.018	0.016	0.022	0.011	5145	0.932
emo_anger	0.051	0.037	0.127	0.039	4714	0.842
emo_sad	0.177	0.028	0.123	0.051	4894	0.883

Table 4.16: LIWC-22's Original Dictionary; Mann-Whitney U test on the LIWC-22 Results (Fake Reviews vs Real Reviews on Amazon Dataset)

NRC-VAD	Fake Review		RealReview		Mann-Whitney U test	
Emotions	Mean	SD	Mean	SD	Test Stat	p
valence	0.751	0.376	0.480	0.397	2607	0.003
arousal	0.536	0.287	0.413	0.310	2782	0.025
dominance	0.459	0.222	0.343	0.268	2921	0.039

Table 4.17: LIWC-22 Loaded with NRC-VAD; Mann-Whitney U Test on the LIWC-22 Results (Fake Reviews vs Real Reviews on Amazon Dataset)

Table 4.17 presents the results of LIWC-22 analysis using the NRC-VAD dictionary. It examines the aspect of emotion dimensionality from the level of valence, arousal, and dominance. The results show that the levels of valence, arousal, and dominance all have a significant difference between real reviews and fake reviews. The findings indicate that modeling the fake review detection from emotional aspects is a decent direction.

After we examine the difference between fake reviews and real reviews in terms of emotions, we are particularly interested in how people perceive emotion differently between human-predicted fake reviews and human-predicted real reviews. The results from this exploration can indicate why people fall into fake reviews from emotional perspectives.

Table 4.18 demonstrates the LIWC-22 results that evaluate the emotion difference between human predicted fake reviews and human predicted real reviews. The results show that emotion categories such as anticipation and trust still have a significant difference between human-predicted fake reviews and human-predicted real reviews. Besides that, the emotion "joy" also has a significant difference between human-predicted fake reviews and

EmoLex	Fake Review		RealReview		Mann-Whitney U test	
Emotions	Mean	SD	Mean	SD	Test Stat	p
anger	1.448	1.275	1.028	0.771	3600	0.424
anticipation	3.678	3.125	2.694	2.500	3285	0.048
disgust	1.368	1.036	1.050	0.852	3559	0.361
fear	1.729	1.372	1.329	1.177	3461.5	0.237
joy	3.302	3.083	2.251	2.027	2905	0.030
negative	3.136	2.841	2.287	1.833	3128	0.042
positive	5.153	4.511	4.703	3.632	4150	0.168
sadness	1.507	1.068	1.185	0.842	4310.5	0.341
surprise	1.525	0.989	1.175	0.980	4231.5	0.241
trust	3.343	3.193	2.650	2.431	3350	0.050

Table 4.18: LIWC-22 Loaded with EmoLex; Mann-Whitney U test on the LIWC-22 Results (Human Predicted Fake Reviews vs Human Predicted Real Reviews on Amazon Dataset)

human-predicted real reviews. Table 4.19 shows the LIWC-22 analysis results, which also indicates that emotion, positive emotion, in particular, has a statistically significant difference between human predicted fake reviews and real ones.

LIWC-22 Dict	Fake Review		RealReview		Mann-Whitney U test	
Emotions	Mean	SD	Mean	SD	Test Stat	p
emotion	2.974	2.648	2.319	2.014	2982	0.039
emo_pos	2.749	2.291	1.673	1.214	2802	0.014
emo_neg	0.811	0.702	0.531	0.393	3727	0.656
emo_anx	0.000	0.000	0.027	0.015	3969	0.952
emo_anger	0.234	0.109	0.035	0.021	3781.5	0.770
emo_sad	0.192	0.113	0.134	0.119	3910	0.823

Table 4.19: LIWC-22's Original Dictionary; Mann-Whitney U Test on the LIWC-22 Results (Human Predicted Fake Reviews vs Human Predicted Real Reviews on Amazon Dataset)

NRC-VAD	Fake Review		RealReview		Mann-Whitney U test	
Emotions	Mean	SD	Mean	SD	Test Stat	p
valence	0.816	0.360	0.396	0.382	2507	0.000
arousal	0.598	0.284	0.354	0.291	2442	0.000
dominance	0.531	0.248	0.252	0.233	2427	0.000

Table 4.20: LIWC-22 Loaded with NRC-VAD; Mann-Whitney U Test on the LIWC-22 Results (Human Predicted Fake Reviews vs Human Predicted Real Reviews on Amazon Dataset)

Through the lens of NRC-VAD lexicons, we are able to examine the differences between human-predicted fake reviews and real ones in terms of the level of valence, arousal, and

dominance. From Table 4.20, we observe that all three emotion dimensions have a statistically significant difference between human predicted real reviews and fake reviews. The results confirm that the emotion representations proposed in our EmoAware framework are capable of boosting the performance of fake review detection.

4.7 Summary

In this chapter, we first identify the challenges of the fake review detection task. Our idea and intuition of the EmoAware framework are illustrated. Secondly, we introduce the emotion representations of the text information of reviews inspired by the emotion models from three perspectives: emotion distribution, emotion intensity, and emotion dimensionality. Thirdly, we then demonstrate how these representations can be integrated into an emotion-aware fake review detection framework (EmoAware) to help improve the performance of fake review detection. Experiments using two real-world datasets demonstrates the effectiveness of our proposed framework. To explore how humans perceive fake reviews in terms of emotions, we present a systematical quantitative study to resolve this question.

Chapter 5

Two-Tier Text Network Analysis Framework

Before the two-tier text network analysis framework design, we observe that significant efforts have been made by governments to report and prohibit the creation of fake reviews. The demonstration of the harm and severity of fake online reviews in the online businesses is not uncommon. In an effort to protect consumers, institutions such as the Federal Trade Commission (FTC) have made major efforts to identify and reveal cases of fake reviews in online marketing. Similar efforts are often made in the realm of detecting fake product reviews. The majority of developed methods for detecting fake reviews is classified into two groups: supervised machine learning methods [41, 56, 32, 31, 118, 43] and unsupervised machine learning methods [36, 89, 16, 69, 40, 106, 44]. As advanced detection approaches have been presented to tackle fake reviews, there is a lack of text network analysis on fake reviews. From the angle of text network analysis, one may analyze and differentiate the network characteristics of fake and genuine reviews, which can provide vital insights for proposing effective fake review detection systems. In this chapter, we carefully develop a two-tier text network analysis framework to examine the difference of network features and textual features between fake and real reviews. More importantly, we conclude our findings on guiding the design of fake review detection systems.

In the dissertation research articulated in this chapter, we start with a comprehensive review of the existing fake review detection methodologies as well as community detection algorithms. Next, we present the challenges and basic ideas of our proposed two-tier text network analysis. The detailed design of the two-tier text network analysis is discussed, followed by conducting the experiments on the Amazon product review dataset. The tier-1 analysis compares the network-level characteristics between fake product reviews and real

product reviews. The tier-2 analysis compares text characteristics of latent communities of fake-review networks and real-review networks. Our findings give rise to distinct features - vital underpinnings that furnish the design of fake review detection systems.

The rest of this chapter is organized as follows. In Section 5.1, we present the challenges of this study, followed by illustrating the basic ideas and intuition of the design of the two-tier text network analysis framework. In Section 5.2, we systematically shed bright light on the advantages of text network analysis. Section 5.3 elaborates on the two-tier text network analysis framework. We conduct the empirical evaluations using Amazon dataset in Section 5.4. Lastly, the summary of the chapter can be found in Section 5.5.

5.1 Challenges and Basic Ideas

Since the worldwide COVID-19 epidemic enhanced the demand for online purchasing, the quantity of online reviews on shopping websites surged substantially. In addition, consumers have an inclination to consult in product reviews more frequently while making purchasing judgments because items are frequently inaccessible. This condition increases the likelihood that online buyers may encounter fake reviews. In this section, we first identify the challenges encountered in the development of two-tier text network analysis on product reviews in Section 5.1.1. Next, we illustrate the basic idea of our solution in Section 5.1.2, which summarizes the motivation and intuition of our proposed two-tier text network analysis framework depicted in Section 5.3.

5.1.1 Challenges

Before discussing the basic ideas offered in the third part of the dissertation, we illuminate a number of challenges to be addressed in this chapter.

- Challenge 1. The lack of suitable fake product review datasets.
- Challenge 2. The proper design of text network analysis on product reviews.

Challenge 1. The lack of suitable fake product review datasets. There are plenty of reviews available on the Internet, but very little of the review data are labeled as fake or real; without such ground truth, comparison between fake and real reviews from the lens of text network analysis is a daunting task. Obviously, accurately labeling fake product reviews is incredibly labor-intensive and expensive. In order to address this major concern, our studies utilize the standard available Amazon fake product review datasets.

Challenge 2. The proper design of text network analysis on product reviews. Few studies examined the difference of fake product reviews and real ones through the lens of text network analysis. The major challenge is to design the procedures of text network analysis that helps make sense of the distinctiveness between fake and real product reviews in terms of network level characteristics. We propose a two-tier text network analysis framework that systematically investigates the network level differences and textual level differences of fake and real product reviews.

5.1.2 Basic Ideas

The text network analysis is one way for encoding the associations between words in a text and building a network of the connected words. The method assumes that language and knowledge may be described as networks of words and their connections. Our intuition of the design of two-tier text network analysis lies in two folds. Firstly, we investigate the distinctiveness of fake and real product reviews from the semantic of sentence level. The intuition design of the first tier of text networks is as follows.

- **Node**: A node represents a product review.
- **Edge**: An edge is a weighted edge, which represents the similarities between two connected nodes.

We construct two aforementioned text networks utilizing the fake product reviews and real product reviews. Comparisons are made through the network level characteristics. To

examine the modules and their possible hierarchical organization, we perform the community detection on two text networks.

Secondly, to better understand the networks from word levels, we construct another set of networks within each community. The structure of the second tier of text networks is presented as follows.

- **Node**: A node represents the bi-grams extracted from product reviews.
- **Edge**: An edge represents the co-occurrence of two connected nodes (bi-grams) in the same review or in the highly similar reviews within the same community.

Through the words level comparison, we are capable of understanding the content and making sense of each community.

5.2 Text Network Analysis

Let us first introduce the background knowledge of general network analysis. The major concentration of network analysis is investigating interactions among nodes and a generated network structure. Consequently, research applying network analysis aims to discover, measure, and comprehend the influence of the strength of the linkages among nodes. The location of nodes and the patterns of densely linked clusters within a network are also essential inspections in text network analysis. For instance, the weight of an edge can represent the strength of communication between two friends in a social network. We infer some social, economic, and technological implications from other measures provided by the network. For example, the degree of centrality of a node measures the number of linkages connecting to other nodes.

As an essential feature revealed from networks, communities are gauged as dense subnetworks or clusters. Identifying communities enable researchers or policymakers to comprehend how a network coheres: the community might indicate important social aspects such

as social capital and trust, polarization, and homogeneity. More recently, community detection techniques have become a popular and practical research area centered around network analysis and graph mining.

The text network analysis aims to uncover human-generated text structures, meanings, and biases. Nodes in text network analysis are words or sentences, whereas edges represent nodes' associations. The central words frequently appear in the text with many other words (higher degree) or with other particular central words (higher betweenness); as a result, the words express crucial meanings. Similarly, when words are formed into the same community, the words sometimes convey similar meanings. This structure enables researchers to discover the predominant themes and topics conveyed in the text. For example, Figure 5.1 presents a text network. Each node represents a word, and words with the same color are classified into the same community. We are able to make sense of the main topic from the central words. For instance, the purple community conveys a message of the journey of America.

The salient advantage of the text network analysis lies in the following three aspects.

- **Determine the relative placement of categories in a broader discourse.** A network of words forms a continuous map in which overlapping word clusters are related. Researchers may determine the relative placement of categories within a large picture. Modeling words within a network, one may associate qualitative and quantitative approaches by virtue of applying text network analysis.

- **Automatically classify words into categories based on their co-occurrences within the text.** Most existing techniques require pre-defining a list of words within each category when detecting themes. In contrast, the text network analysis automatically classifies words into categories based on the co-occurrences within the text.

- **Flexible to revise the network structure based on various research questions.** Researchers can adjust network structures according to various research questions. For

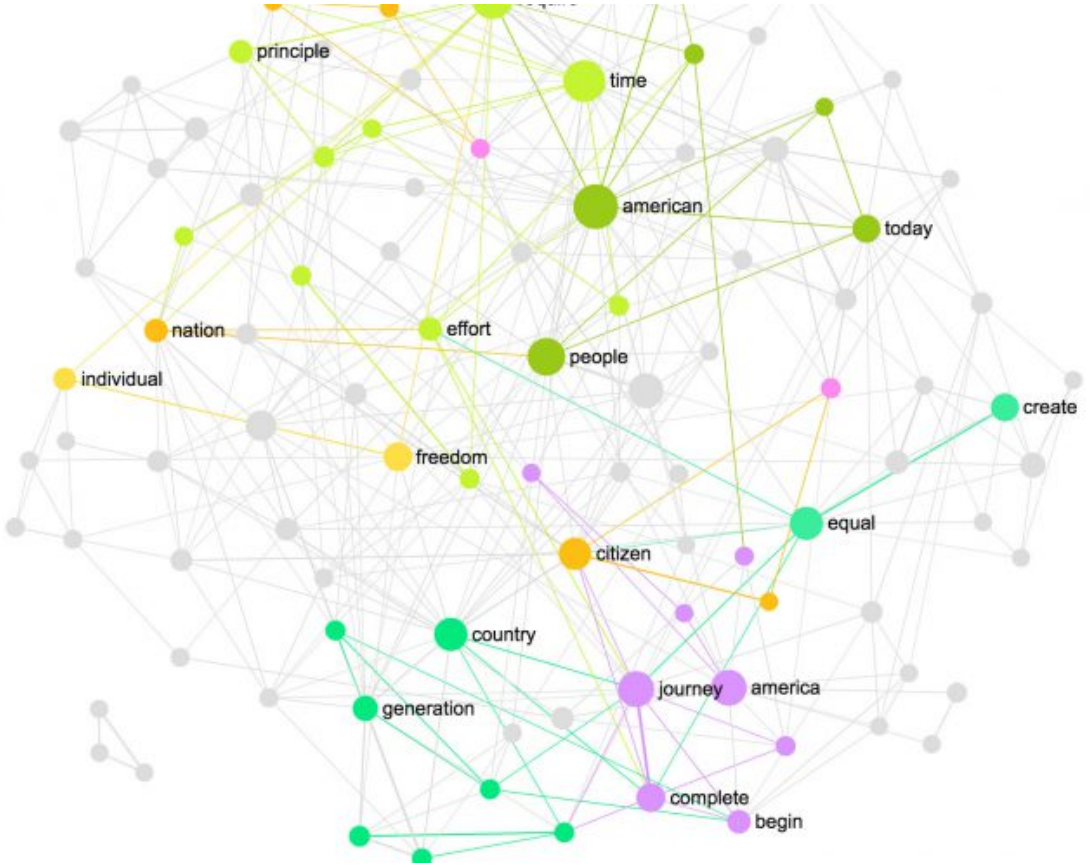

Figure 5.1: An Example of Text Network Analysis

example, one can change the definition of edges while examining the text's multiple properties.

Given the above justifications of deploying text network analysis, we advocate for this method as a core piece in our proposed detection framework. In the next section, we are primed to design a two-tier text network analysis framework - a novel technique that investigates the characteristics between fake product reviews and real ones at the network level as well as the textual level.

5.3 A Two-Tier Text Network Analysis Framework

In this section, we delineate our proposed two-tier text network analysis framework in detail. Figure 5.2 depicts the framework tier-wisely.

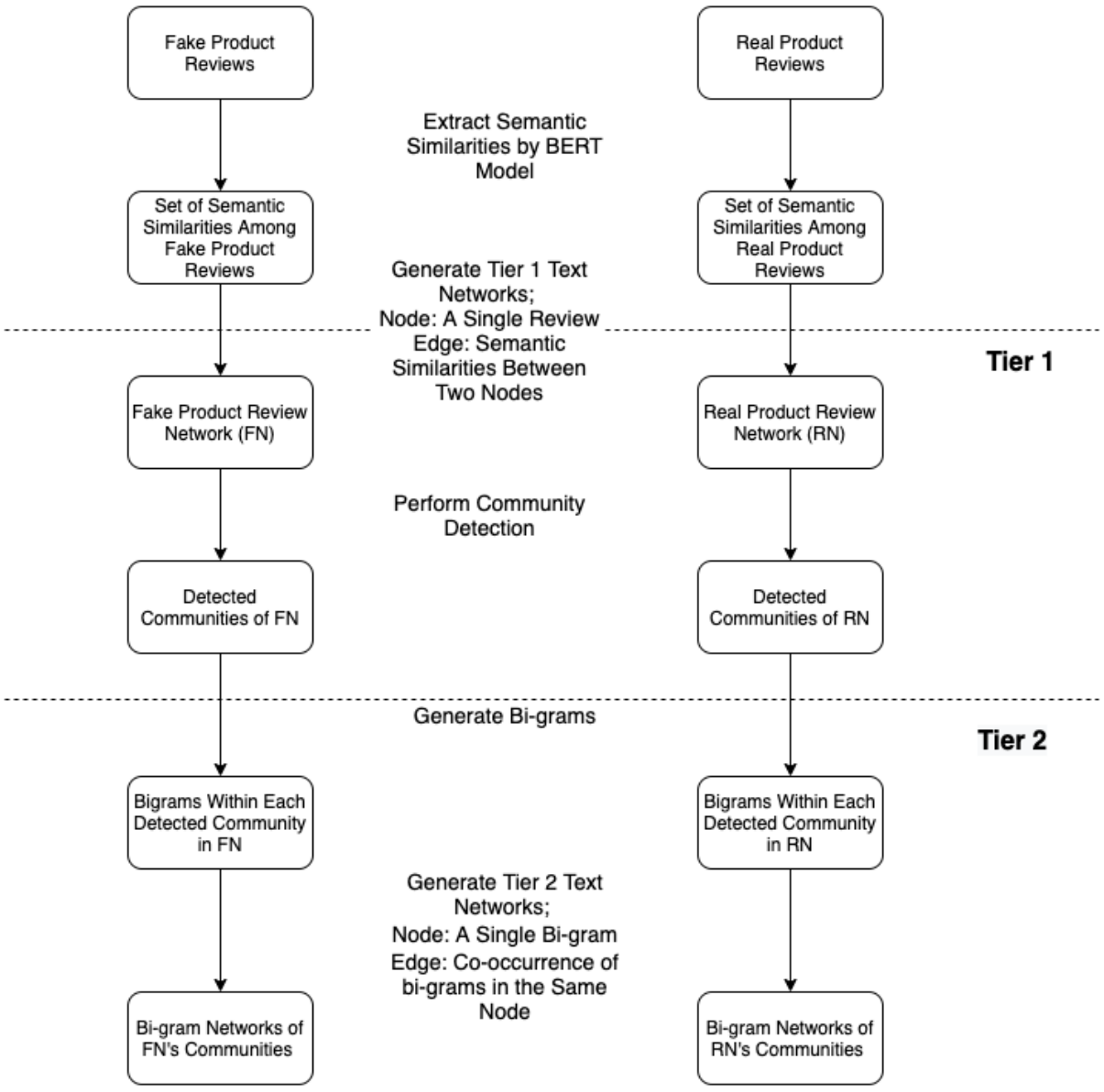

Figure 5.2: The High-level Architecture of the Two-tier Text Network Analysis Framework.

The framework starts with a corpus of fake reviews and a corpus of real reviews. The tier-1 text networks aim to analyze differences between fake and real reviews from the perspective of review semantic similarity. We examine the following research question **RQ 1** in tier 1.

- **RQ 1.** What are the differences within network-level characteristics between fake product reviews and legitimate ones?

Given the fake product reviews and real product reviews, we extract - as an initial phase - the pairwise semantic similarities utilizing the BERT model. After obtaining the set of semantic similarities among fake product reviews, we generate the tier 1 text network, namely, the fake product review network (FN). The node in FN represents a single fake review, whereas an edge between two nodes denotes the semantic similarities between two nodes. In order to decrease the number of edges, we elect the similarity threshold as 0.5. If the semantic similarity between two nodes is smaller than 0.5, we consider that the two nodes do not share significant similarities. Thus, no connected edge exists between the two nodes. We perform the same procedures to forge a real product review network (RN). After creating the two networks - FN and RN, we utilize the community detection algorithms to identify communities in existing networks. Now, we have our tier 1 network structures, and comparisons of network-level characteristics can be made at this point.

The Tier-2 analysis is in charge of comparing the latent communities within each text network. We examine the research question **RQ 2** below in tier 2.

- **RQ 2.** At the word level, what distinct characteristics can be observed within each community of two networks?

When it comes to the tier-2 analysis, we first generate the bi-grams within each detected community of two networks. Next, we construct the tier 2 text networks. A node represents a single bi-gram, while the edge between two nodes is created when the two bi-grams appear in the same review within the same community. After building multiple bi-grams networks, we are capable of examining the distinct characteristics within the community.

In this section, we systematically introduce our proposed two-tier text network analysis framework. We conduct experiments to gauge the effectiveness of the proposed framework in the next section.

5.4 Experimental Study

We first introduce the details and statistics of the Amazon dataset in Section 5.4.1. Next, we present the results of tier 1 analysis in Section 5.4.2 and the tier 2 analysis in Section 5.4.3.

5.4.1 Dataset

The experimental design is centered around the Amazon dataset. Table 5.1 summarizes the distinctive traits of the Amazon dataset, which was initially published and hosted on the Kaggle website - a popular data repository platform that offers a no-installation, customized Jupyter Notebooks environment. The reviews in the Amazon dataset are categorized as either real or fake. A number of other features, including ratings, confirmed purchases, product categories, product IDs, product titles, and review titles, are included for each review in the dataset. The Amazon dataset contains 21,000 reviews that are dispersed evenly across product categories.

Dataset	# of fake reviews	# of real reviews	Avg. review length	Fake/real avg. review length
Amazon	10,500	10,500	69.2	59.3/79.1

Table 5.1: Amazon Fake Review Dataset.

Because of the time-expensiveness in computing the pairwise similarities, we randomly select 1,000 fake product reviews and 1,000 real product reviews from the Amazon dataset. The 1,000 fake reviews and the 1,000 authentic reviews serve as the fake product review corpus and real product review corpus in the two-tier text network analysis.

5.4.2 Tier-1 Analysis Results

Recall that (see also Section 5.3) the Tier-1 analysis aims to explore the comparisons of network-level of characteristics between fake reviews and legitimate reviews. We first generate the FN and RN networks based on the description in Section 5.3. We report

the six crucial network parameters: Average Degree, Average Weighted Degree, Average Betweenness Centrality, Closeness Centrality, # of Communities, and Eccentricity.

Table 5.2 tabulates the network parameters of *FN* and *RN*.

Network	Avg. Degree	Avg. Weighted Degree	Betweenness centrality
Fake Review Network	562	338	231
Real Review Network	537	321	219

Network	Closeness Centrality	Eccentricity	# of Communities
Fake Review Network	0.707	2.63	3
Real Review Network	0.695	2.694	2

Table 5.2: Comparison of Fake and Real Reviews Networks on Network Characteristics.

The average degree of a network resembles the average number of connected nodes from a single node in the network. The average degree of the FN network is larger than RN (562 for fake reviews, 537 for real reviews), which may suggest that fake reviews share higher semantic similarities compared to real reviews. Similar patterns are observed in the average weighted degree. The average weighted degree of the FN network is higher than RN (338 for fake reviews, 321 for real reviews), implying that the similarity of fake reviews is stronger than real reviews.

Betweenness centrality quantifies the frequency at which a specific node appears on the shortest path between two other nodes. We discover that the betweenness centrality is higher for fake reviews (avg. 231 for fake reviews vs. avg. 219 for real reviews), which indicates that fake reviews have more tendency to associate with the other fake reviews.

Closeness centrality estimates the average shortest geodesic distance (number of unique edges) between a given node and every other node in a network. We observe very little difference in the closeness centrality of the two networks (avg. 0.707 for fake reviews vs. avg. 0.695 for real reviews).

The average eccentricity represents the distance of a given node to the furthest node within the same network. The finding unfolds that the value of average eccentricity is larger in authentic reviews. This phenomenon shows that the real review network overall is more

disconnected than fake reviews; the fake reviews are more cohesive and share more similar semantic meanings.

Community detection is performed on both networks. The results unveil that the fake review network consists of three communities, whereas the real review network possesses two communities. However, one of the communities in the fake review network has very few number of nodes. Therefore, we neglect the small community in the fake review network while performing the tier-2 analysis.

5.4.3 Tier-2 Analysis Results

Recall that (see also Section 5.3) the tier-2 analysis is responsible for investigating the properties of each detected community by creating a second level of bi-grams networks. The edge between two nodes is formed when two bi-grams appear in the same review within the same community. After constructing many bi-grams networks, we are primed to examine the community's unique characteristics.

From the tier-2 analysis, we generate four unique bi-grams networks based on the communities in the fake and real review networks. Surprisingly, we do not locate any significant discrepancies in different communities based on the bi-grams networks. For this reason, we only present the comparisons of bi-gram networks between real and fake reviews.

We examine the network through the lens of variation in degree range. Figure 5.3 and Figure 5.4 demonstrate the high-degree nodes of bi-gram networks. We notice that the high-degree nodes convey similar meanings between real and fake reviews. Mostly, these nodes express the overall impression of the products. For example, the bi-gram "highly recommended" represents the users' opinions of a product. This phenomenon confirms our intuitions and speculations because customers write reviews by first illustrating the overall ratings of products.

Next, we investigate the medium-degree nodes within the bi-gram networks. Figure 5.5 illustrates an example of the medium-degree range of bi-grams networks of fake reviews. The

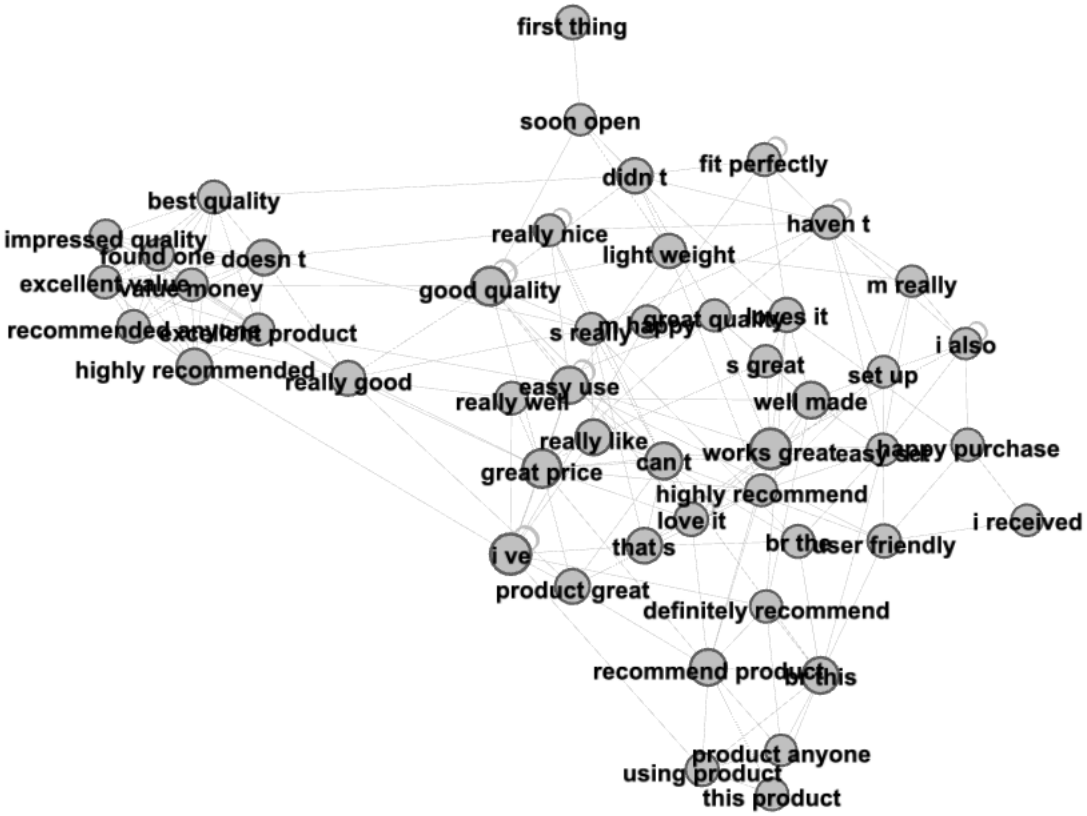

Figure 5.3: An Example of High Degree Range of Bi-grams Networks of Fake Reviews.

results reveal that the medium-degree nodes talk about the product in more detail, such as the user experience and detailed reviews. Similar patterns are observed in the bi-grams networks of genuine reviews. Comparing the bi-grams of the two figures, we conclude that unlike the legitimate reviews, the fake product reviews tend to use more emotion-activating words to assess the products. For example, words such as "love," "perfectly," and "enjoy" appear more frequently in the bi-grams networks of fake reviews. This group of experiments on medium-degree nodes reflects the emotion analysis study in this dissertation.

We overlook the low-degree nodes in the bi-grams networks because the generality of low-degree nodes is too low to interpret and compare among fake reviews and legitimate ones.

Figure 5.4: An Example of High Degree Range of Bi-grams Networks of Real Reviews.

5.5 Summary

In this chapter, we first discussed the challenges and basic ideas of our proposed two-tier text network analysis. After demonstrating the construction of a two-tier text network analysis, we elaborated on the experimental results derived from the Amazon product review dataset. The tier-1 analysis compares the network-level characteristics between fake product reviews and real product reviews. The tier-2 analysis compares text characteristics of latent communities of fake-review networks and real-review networks. Per our findings, we pinpointed distinctiveness features to facilitate the design of modern fake-review detection systems.

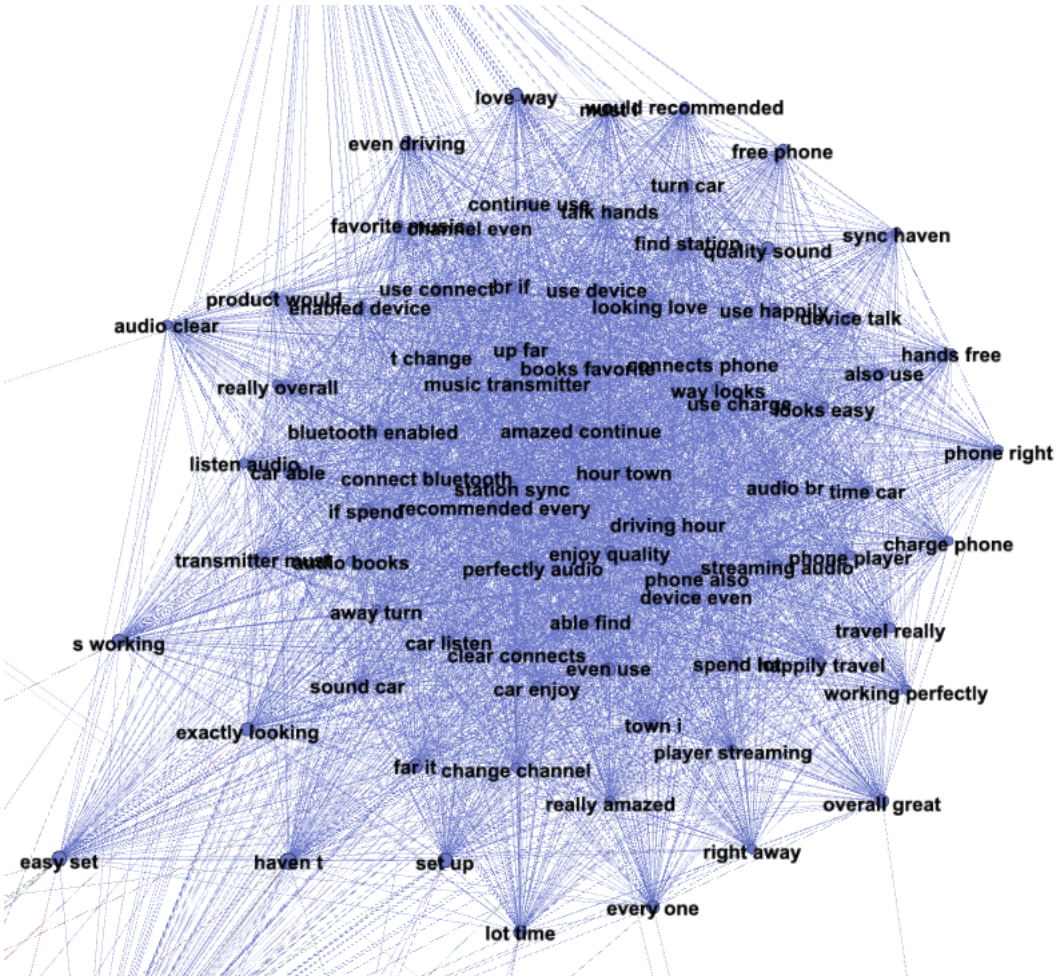

Figure 5.5: An Example of Medium Degree Range of Bi-grams Networks of Fake Reviews.

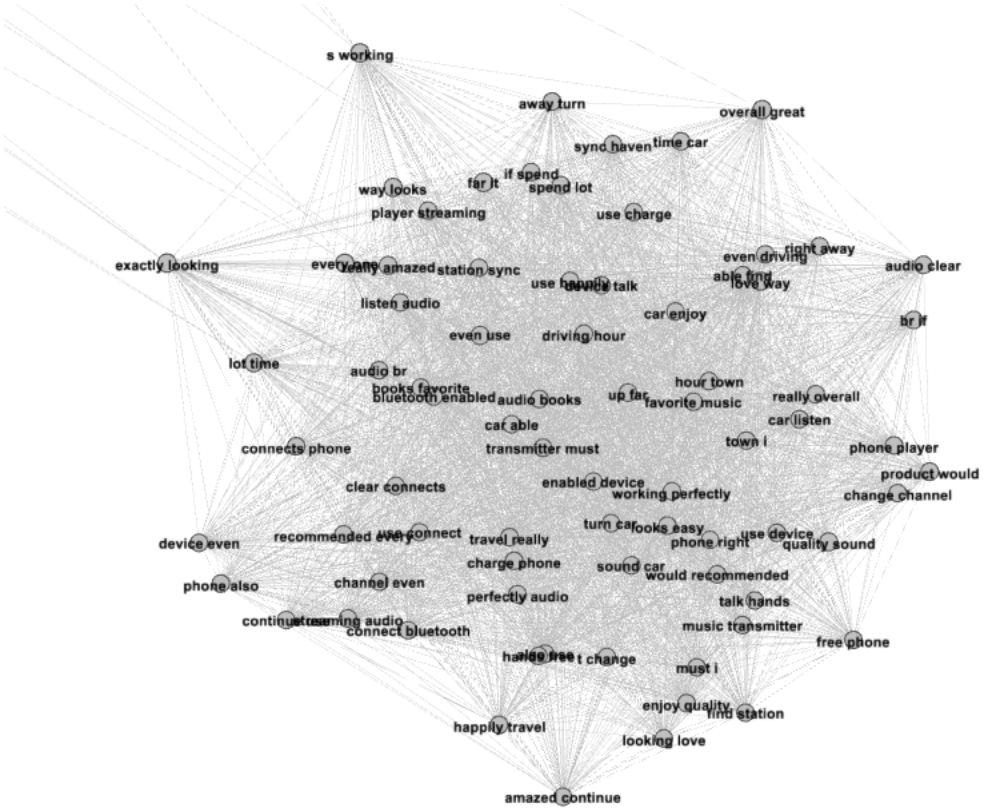

Figure 5.6: An Example of Medium Degree Range of Bi-grams Networks of Real Reviews.

Chapter 6

Conclusion and Future Work

In this chapter, we systematically state the main contributions made in this dissertation study. Then, we identify the possible directions for future studies.

6.1 Main Contributions

6.1.1 The Fake News Engagement and Propagation Path (FNEPP) Framework

To model fake news detection from a social context perspective, we developed a novel fake news detection framework called *FNEPP* that seamlessly incorporates news contents, user engagements, user characteristics, and propagation paths using two cooperative modules. We conducted extensive experiments driven by two real-world datasets to shed light on the effectiveness of *FNEPP*. We demonstrated the capability of capturing distinctive temporal patterns between fake and real news. The promising results unfold the high efficiency of *FNEPP* in the realm of detecting fake news on social media at an early stage.

From the perspective of social-context-based approaches, we are in a good position to tackle a raft of challenges that are still open to the research community. Specifically, one intriguing future direction is to analyze a rich set of representations of social media users. Given a massive amount of available social media data, we plan to utilize multi-task learning to jointly optimize fake news detection tasks, stance detection, and prediction of source authenticity.

6.1.2 The Emotion-Aware (EmoAware) Fake Review Detection Framework

With a surge in fake product reviews, the issue of detecting fake reviewers is of increasing importance to firms and their customers. In the second part of the dissertation study, we

examined the motivation of fake review detection, especially exploring how emotion conveyed in the review text helps improve the performance of fake review detection. We provided the literature on the definition of "fake review." Existing approaches to fake review detection have been thoroughly identified. We also surveyed the state-of-the-art emotion models in the text mining literature, which motivates our proposed emotion representations in fake review detection. We proposed an emotion-aware fake review detection framework inspired by ensemble learning methods. Three perspectives of modeling emotion conveyed in the review text are seamlessly integrated into the framework. A series of experiments on two real-world datasets have demonstrated the effectiveness of our proposed model. In order to evaluate how humans perceive fake reviews differently compared with machine learning models, we conducted a survey-based qualitative analysis.

6.1.3 The Two-Tier Text Network Analysis Framework

In the third part of this dissertation research, we illustrated the motivation for text network analysis on fake product reviews, focusing on how the findings of text network analysis contribute to the design of systems for detecting fake reviews. Following this, we presented the concepts of text network analysis and review the community discovery algorithms. The Amazon fake product reviews dataset is utilized to perform a two-tier text network analysis. The first tier analysis compares the network-level characteristics between fake product reviews and real product reviews. The second tier analysis compares text characteristics of latent communities of fake-review networks and real-review networks. More importantly, the results of the two-tiered text network analysis reveals the distinct network and semantic characteristics of fake and real reviews.

6.2 Future Research Studies

6.2.1 Future Directions for the Fake News Engagement and Propagation Path (*FNEPP*) Framework

Recall that in the first part of this dissertation study (see also Chapter 3, we focused on data sourced from social media platforms such as Twitter and WEIBO, which are predominantly unmoderated and free-flowing. Future studies can focus on moderated fake news data, such as in fake news sites that exist in edited or semi-edited forms. The size of the dataset utilized in the study is relatively small. Future studies will benefit from increased dataset size. Also, emotions, which could potentially play a key role in fake news detection, have not been studied and are one of the modeling limitations. Future studies could investigate the significance of emotional signals in fake news detection. This study predominantly focuses on misinformation detection. However, broader management and governance approaches need to be taken for better control and mitigation of misinformation. Future studies can look at this problem from a disaster control perspective.

6.2.2 Future Directions for the EmoAware Fake Review Detection Framework

As labeled-fake-review datasets are expensive to acquire, a drawback of our EmoAware framework is that the framework belongs to supervised learning methods. In a future project, we will design unsupervised or semi-supervised learning models for fake review detection tasks. Our results of exploring the role of emotion in fake-review detection are expected to help us design those types of novel models.

6.2.3 Future Directions for the Two-Tier Text Network Analysis Framework

Our proposed two-tier framework is capable of examining network-level comparisons and community-level comparisons. Besides the high-level characteristics, we intend to observe if generated networks exhibit certain network properties. For instance, we will be interested

in studying the "small world" phenomenon in both generate networks. Another direction along this line is to integrate topic modeling in the process of network analysis. The topic modeling methods are primed to offer us the meanings of a given dataset or clustered communities, thereby being potentially beneficial in designing network-analysis frameworks for misinformation detection.

Bibliography

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

[2] S. E. Asch. Effects of group pressure upon the modification and distortion of judgments. In *Documents of gestalt psychology*, pages 222–236. University of California Press, 1961.

[3] M. Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication research*, 41(3):430–454, 2014.

[4] P. Biyani, K. Tsioutsiouliklis, and J. Blackmer. " 8 amazing secrets for getting more clicks": detecting clickbaits in news streams using article informality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[5] J. C. Borod et al. *The neuropsychology of emotion*. Oxford University Press, 2000.

[6] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.

[7] T. Chen, X. Li, H. Yin, and J. Zhang. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 40–52. Springer, 2018.

[8] Y. Chen, N. J. Conroy, and V. 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.

[9] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[10] M. Cohen. Fake news and manipulated data, the new gdpr, and the future of information. *Business Information Review*, 34(2):81–85, 2017.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[12] H. Deng, L. Zhao, N. Luo, Y. Liu, G. Guo, X. Wang, Z. Tan, S. Wang, and F. Zhou. Semi-supervised learning based fake review detection. In *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, pages 1278–1280. IEEE, 2017.

[13] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological bulletin*, 129(1):74, 2003.

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

[15] N. Dhamani, P. Azunre, J. L. Gleason, C. Corcoran, G. Honke, S. Kramer, and J. Morgan. Using deep networks and transfer learning to address disinformation. *arXiv preprint arXiv:1905.10412*, 2019.

[16] L.-y. Dong, S.-j. Ji, C.-j. Zhang, Q. Zhang, D. W. Chiu, L.-q. Qiu, and D. Li. An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews. *Expert Systems with Applications*, 114:210–223, 2018.

[17] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun. User preference-aware fake news detection. *arXiv preprint arXiv:2104.12259*, 2021.

[18] L. N. Driscoll. A validity assessment of written statements from suspects in criminal investigations using the scan technique. *Police Stud.: Int'l Rev. Police Dev.*, 17:77, 1994.

[19] J. Eisenstein, A. Ahmed, and E. P. Xing. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048, 2011.

[20] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, 2012.

[21] W. Ferreira and A. Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168, 2016.

[22] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[23] C. Guo, J. Cao, X. Zhang, K. Shu, and M. Yu. Exploiting emotions for fake news detection on social media. *arXiv preprint arXiv:1903.01728*, 2019.

[24] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

[25] A. Gupta, H. Li, A. Farnoush, and W. Jiang. Understanding patterns of covid infodemic: A systematic and pragmatic approach to curb fake news. *Journal of business research*, 140:670–683, 2022.

[26] M. Hüsken and P. Stagge. Recurrent neural networks for time series classification. *Neurocomputing*, 50:223–235, 2003.

[27] N. Jain, A. Kumar, S. Singh, C. Singh, and S. Tripathi. Deceptive reviews detection using deep learning techniques. In *International Conference on Applications of Natural Language to Information Systems*, pages 79–91. Springer, 2019.

[28] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. 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.

[29] N. Jindal and B. Liu. Analyzing and detecting review spam. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 547–552. IEEE, 2007.

[30] M. Johnson. Pcfg models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632, 1998.

[31] F. Khurshid, Y. Zhu, C. W. Yohannese, and M. Iqbal. Recital of supervised learning on review spam detection: An empirical analysis. In *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 1–6. IEEE, 2017.

[32] F. Khurshid, Y. Zhu, X. Zhuang, M. Ahmad, and M. Ahmad. Enactment of ensemble learning for review spam detection on selected features. *International Journal of Computational Intelligence Systems*, 12(1):387, 2018.

[33] D. Klein and J. Wueller. Fake news: A legal perspective. *Journal of Internet Law (Apr. 2017)*, 2017.

[34] S. Kwon, M. Cha, and K. Jung. Rumor detection over varying time windows. *PloS one*, 12(1):e0168344, 2017.

[35] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[36] R. Y. Lau, S. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li. Text mining and probabilistic language modeling for online review spam detection. *ACM Transactions on Management Information Systems (TMIS)*, 2(4):1–30, 2012.

[37] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[38] F. H. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. In *Twenty-second international joint conference on artificial intelligence*, 2011.

[39] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao. Bimodal distribution and co-bursting in review spam detection. In *Proceedings of the 26th international conference on world wide web*, pages 1063–1072, 2017.

[40] J. Li, P. Lv, W. Xiao, L. Yang, and P. Zhang. Exploring groups of opinion spam using sentiment analysis guided by nominated topics. *Expert Systems with Applications*, 171:114585, 2021.

[41] L. Li, B. Qin, W. Ren, and T. Liu. Document representation and feature combination for deceptive spam review detection. *Neurocomputing*, 254:33–41, 2017.

[42] L. Li, W. Ren, B. Qin, and T. Liu. Learning document representation for deceptive opinion spam detection. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 393–404. Springer, 2015.

[43] Q. Li, Q. Wu, C. Zhu, J. Zhang, and W. Zhao. An inferable representation learning for fraud review detection with cold-start problem. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[44] Q. Li, Q. Wu, C. Zhu, J. Zhang, and W. Zhao. Unsupervised user behavior representation for fraud review detection with cold-start problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 222–236. Springer, 2019.

[45] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948, 2010.

[46] Y. Lin, T. Zhu, X. Wang, J. Zhang, and A. Zhou. Towards online review spam detection. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 341–342, 2014.

[47] H. Liu, L. Fen, J. Jian, and L. Chen. Overlapping community discovery algorithm based on hierarchical agglomerative clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(03):1850008, 2018.

[48] W. Liu, W. Jing, and Y. Li. Incorporating feature representation into bilstm for deceptive review detection. *Computing*, 102(3):701–715, 2020.

[49] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[50] Y. Liu and Y.-F. 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.

[51] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.

[52] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. 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.

[53] J. Ma, W. Gao, and K.-F. Wong. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics, 2017.

[54] J. Ma, W. Gao, and K.-F. Wong. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics, 2018.

[55] S. Maheshwari. How fake news goes viral: A case study. *The New York Times*, 20, 2016.

[56] S. Mani, S. Kumari, A. Jain, and P. Kumar. Spam review detection using ensemble machine learning. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 198–209. Springer, 2018.

[57] B. Markines, C. Cattuto, and F. Menczer. Social spam detection. In *Proceedings of the 5th international workshop on adversarial information retrieval on the web*, pages 41–48, 2009.

[58] D. M. Markowitz and J. T. Hancock. Linguistic traces of a scientific fraud: The case of diederik stapel. *PloS one*, 9(8):e105937, 2014.

[59] J. M. Martínez Otero. Fake reviews on online platforms: perspectives from the us, uk and eu legislations. *SN Social Sciences*, 1(7):1–30, 2021.

[60] A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 1995.

[61] M. J. Metzger and A. J. Flanagin. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics*, 59:210–220, 2013.

[62] R. Mihalcea and C. 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.

[63] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[64] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.

[65] R. Mohawesh, S. Tran, R. Ollington, and S. Xu. Analysis of concept drift in fake reviews detection. *Expert Systems with Applications*, 169:114318, 2021.

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

[67] J. Ni, J. Li, and J. McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, 2019.

[68] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[69] S. Noekhah, N. binti Salim, and N. H. Zakaria. Opinion spam detection: Using multi-iterative graph-based model. *Information Processing & Management*, 57(1):102140, 2020.

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

[71] T. Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social networks*, 35(2):159–167, 2013.

[72] M. Ott, C. Cardie, and J. T. Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 497–501, 2013.

[73] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

[74] C. Paul and M. Matthews. The russian "firehose of falsehood" propaganda model. *Rand Corporation*, pages 2–7, 2016.

[75] G. Pennycook. A perspective on the theoretical foundation of dual process models. *Dual process theory*, 2:34, 2017.

[76] C. Pizzuti. Evolutionary computation for community detection in networks: A review. *IEEE Transactions on Evolutionary Computation*, 22(3):464–483, 2017.

[77] F. Qian, C. Gong, K. Sharma, and Y. Liu. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI*, volume 18, pages 3834–3840, 2018.

[78] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[79] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. 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.

[80] D. Refaeli and P. Hajek. Detecting fake online reviews using fine-tuned bert. In *2021 5th International Conference on E-Business and Internet*, pages 76–80, 2021.

[81] Y. Ren, B. Wang, J. Zhang, and Y. Chang. Adversarial active learning based heterogeneous graph neural network for fake news detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 452–461. IEEE, 2020.

[82] Y. Ren and Y. Zhang. Deceptive opinion spam detection using neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 140–150, 2016.

[83] V. L. Rubin. Deception detection and rumor debunking for social media. *The SAGE Handbook of Social Media Research Methods*, pages 342–364, 2017.

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

[85] V. L. Rubin, N. Conroy, Y. Chen, and S. 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.

[86] N. Ruchansky, S. Seo, and Y. 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.

[87] J. Salminen, C. Kandpal, A. M. Kamel, S.-g. Jung, and B. J. Jansen. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771, 2022.

[88] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[89] Z. Sedighi, H. Ebrahimpour-Komleh, and A. Bagheri. Rlosd: Representation learning based opinion spam detection. In *2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)*, pages 74–80. IEEE, 2017.

[90] M. SHI, Y. ZHOU, and Y. XING. Community detection by label propagation with leaderrank method. *Journal of Computer Applications*, 35(2):448, 2015.

[91] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. 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.

[92] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[93] K. Shu, S. Wang, and H. Liu. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435. IEEE, 2018.

[94] K. Shu, S. Wang, and H. Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320, 2019.

[95] H. SUN and Y. LI. Overlapping community discovering algorithm based on latent features. *Journal of Computer Applications*, 35(12):3477, 2015.

[96] M. Taddicken and L. Wolff. 'fake news' in science communication: Emotions and strategies of coping with dissonance online. *Media and Communication*, 8(1):206–217, 2020.

[97] H. Tajfel, J. C. Turner, W. G. Austin, and S. Worchel. An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56(65):9780203505984–16, 1979.

[98] D. Taraborelli. How the web is changing the way we trust. *Current issues in computing and philosophy*, pages 194–204, 2008.

[99] TripAdvisor. 2019 trip advisor review transparency report. 2019.

[100] C. Tu, X. Zeng, H. Wang, Z. Zhang, Z. Liu, M. Sun, B. Zhang, and L. Lin. A unified framework for community detection and network representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1051–1065, 2018.

[101] S. Volkova, K. Shaffer, J. Y. Jang, and N. 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.

[102] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[103] C.-C. Wang, M.-Y. Day, C.-C. Chen, and J.-W. Liou. Detecting spamming reviews using long short-term memory recurrent neural network framework. In *Proceedings of the 2nd International Conference on E-commerce, E-Business and E-Government*, pages 16–20, 2018.

[104] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao. Fake review detection based on multiple feature fusion and rolling collaborative training. *IEEE Access*, 8:182625–182639, 2020.

[105] W. Y. Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[106] X. Wang, K. Liu, and J. Zhao. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 366–376, 2017.

[107] K. Wu, S. Yang, and K. Q. Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE, 2015.

[108] R. Yafeng, J. Donghong, Z. Hongbin, and Y. Lan. Deceptive reviews detection based on positive and unlabeled learning. *Journal of Computer Research and Development*, 52(3):639–648, 2015.

[109] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7, 2012.

[110] C. M. Yilmaz and A. O. Durahim. Spr2ep: A semi-supervised spam review detection framework. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 306–313. IEEE, 2018.

[111] L. You, Q. Peng, Z. Xiong, D. He, M. Qiu, and X. Zhang. Integrating aspect analysis and local outlier factor for intelligent review spam detection. *Future Generation Computer Systems*, 102:163–172, 2020.

[112] Z.-Y. Zeng, J.-J. Lin, M.-S. Chen, M.-H. Chen, Y.-Q. Lan, and J.-L. Liu. A review structure based ensemble model for deceptive review spam. *Information*, 10(7):243, 2019.

[113] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic. What online reviewer behaviors really matter? effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems*, 33(2):456–481, 2016.

[114] J. Zhang, L. Cui, Y. Fu, and F. B. Gouza. Fake news detection with deep diffusive network model. *arXiv preprint arXiv:1805.08751*, 2018.

[115] W. Zhang, Y. Du, T. Yoshida, and Q. Wang. Dri-rcnn: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management*, 54(4):576–592, 2018.

[116] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, and K. Shu. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476, 2021.

[117] Z. Zhang, Z. Zhang, W. Yang, and X. Wu. An overlapped community partition algorithm based on line graph. In *International Conference on Web-Age Information Management*, pages 277–281. Springer, 2013.

[118] S. Zhao, Z. Xu, L. Liu, M. Guo, and J. Yun. Towards accurate deceptive opinions detection based on word order-preserving cnn. *Mathematical Problems in Engineering*, 2018, 2018.

[119] Z. Zhao, P. Resnick, and Q. 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.

[120] Z. Zhao, J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, D. Li, J. Wu, and S. Havlin. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9(1):7, 2020.

[121] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106, 2004.

[122] A. Zubiaga, M. Liakata, and R. Procter. Exploiting context for rumour detection in social media. In *International Conference on Social Informatics*, pages 109–123. Springer, 2017.