B.Comp. Dissertation

Leveraging Social Context in Fake News Detection with Graph Representation

By

Nguyen Van Hoang

Department of Computer Science

School of Computing

National University of Singapore

2019/20

B.Comp. Dissertation

Leveraging Social Context in Fake News Detection with Graph Representation

By

Nguyen Van Hoang

Department of Computer Science

School of Computing

National University of Singapore

2019/20

Project No: H0791800
Advisor: A/Prof. Min-Yen Kan, Dr. Kazunari Sugiyama
Deliverables:
Report: 1 Volume
Source Code: 1 DVD

Abstract

The popularity of social media in Web 2.0 has changed how the public shares and consumes news. The rapid dissemination of information allows both genuine and wrong news to reach millions of audiences within hours, impacting public opinions and decisions. In critical events such as pandemic responses or presidential elections, information factuality is an upmost concern in coordinating social behaviours and ensuring public fairness and rationality. Unfortunately, the tremendous volume and startling speed of news propagation renders manual fact-checking methods ineffective. Automatic approaches are mainly categorized into content-based models, which rely on the subject matter and its existing knowledge, and context-based models, which utilize social media perception towards the questionable news. While the former arguably gives explainable predictions, the latter makes few assumptions about the news content by utilizing the large and readily available wisdom of the crowd.

In this work, we propose Fake News Graph or FANG – a novel graph-based social context representation and learning framework for fake news detection. We discuss its superiority in modeling social context as a graph compared with both content-based and Euclidean context-based baselines. Our approaches not only improves the macro performance of fake news detection but also obtains quality representations that generalize well to down-stream social network analysis. We observe consistent improvement across different data availability especially when given a limited training amount. Our recurrent aggregator accounts for temporal propagation patterns and enhances explainability with attention mechanism.

Subject Descriptors:
C5 Computer System Implementation
G2.2 Graph Algorithms

Keywords:
Problem, algorithm, implementation

Implementation Software and Hardware:
Solaris 10, g++ 3.3, Tcl/Tk 8.4.7

Acknowledgement

I am greatly thankful to have the guidance of Professor Min-Yen Kan, Dr. Kazunari Sugiyama, and recently Dr. Preslav Nakov. I would also like to extend my appreciation to Pan Liangming and Animesh Prasad, who are currently PhD candidates at NUS, for giving valuable research suggestions, as well as my fellows from NUS's Web IR / NLP Group (WING). I also like to thank Toshiki Tomihira, who was visiting WING as a research intern when I first started researching fake news, for many constructive discussion and collaboration.

List of Figures

List of Tables

Table of Contents

Chapter 1

Introduction

1.1 Background

There are two sides of Web 2.0. Social media connects billions of Internet users and is a conductive medium to spread information. The current speed of news dissemination is a great challenge to any effort to verify it as scale, and creates a risk of leaving falsehood unchecked. Fake news, as defined by its malicious intent (Shu, Sliva, Wang, Tang, & Liu, 2017), can be weaponized to distort public perception. Pizzagate conspiracy theory[1], although debunked later, went viral during the 2016 United States presidential election, wrongfully tarnished the reputation of some candidates and benefited others. The popularity of the term "fake news" gave itself the title "word of the year" by American Dialect Society[2].

Recent research by MIT (Vosoughi, Roy, & Aral, 2018) confirms that fake news reaches 1,500 people 6 times faster than true stores. and is 70 more likely to be retweeted. To make matter worse, modern recommendation systems and social networks personalize user news feed and suggest connections based on preference for existing narratives, such as vaccination (Ludolph, Allam, & Schulz, 2016). These groups of online users form echo chambers (Quattrociocchi, Scala, & Sunstein, 2016), and amplify the propagation of news by confirming their bias. Fake news identification is a non-trivial task. Although general public are confident in their ability to

[1] https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory

[2] https://time.com/5091268/fake-news-word-of-the-year/

discriminate false information, recent studies have shown the exact opposite. 39% of Americans claim to be "very confident" and an extra 45% to be "somewhat confident" in recognizing fake news, however, 75% of them view a false story as accurate despite having seen it (Edkins, 2016). Another study conducted in Singapore shows that 90% of the subjects falsely believe in at least one out of five fake headlines, even though four out of five Singaporeans claim that they can confidently spot fake news (Ng, 2018). As fake news is written with the intention to mislead readers, inferring news veracity solely based on its content is challenging (Shu et al., 2017). Many sites and social media have devoted great effort in identifying false information. Facebook encourages user to report non-credible posts [3], them it employs expert analysis to expose or confirm questionable stories. This method is also used by fact-checking websites such as Snopes[4] or Politifact[5]. With the ever-increasing amount of information, automated news verification systems consider external knowledge databases as evidences (Hassan, Zhang, Arslan, Caraballo, Jimenez, Gawsane, Hasan, Joseph, Kulkarni, Nayak, & others, 2017) (Thorne & Vlachos, 2017) to verify controversial claims.

1.2 The Problem

Fact-checking or content-based models achieve high accuracy, but often takes human resources and time to collect and process sufficient evidences, during which false information might have spread and caused severe damage. Recent research direction in fake news detection (FND) takes another turn and explores various features of the news dissemination process. Our observations show that there is a distinctive engagement pattern of social user towards fake and real news. The fake news example in Table 1.1 has a large number of engagements within a short amount of time right after its publication. These engagements are mainly verbatim reports or negative comments explained by the typically appalling contents of fake news. After that short window we begin to see denial comments questioning the validity of the news, whereas the negative comments dwindles. The stance distribution stabilizes afterwards with no support

[3] https://www.facebook.com/help/572838089565953?helpref=faq_content

[4] snope.com

[5] politifact.com

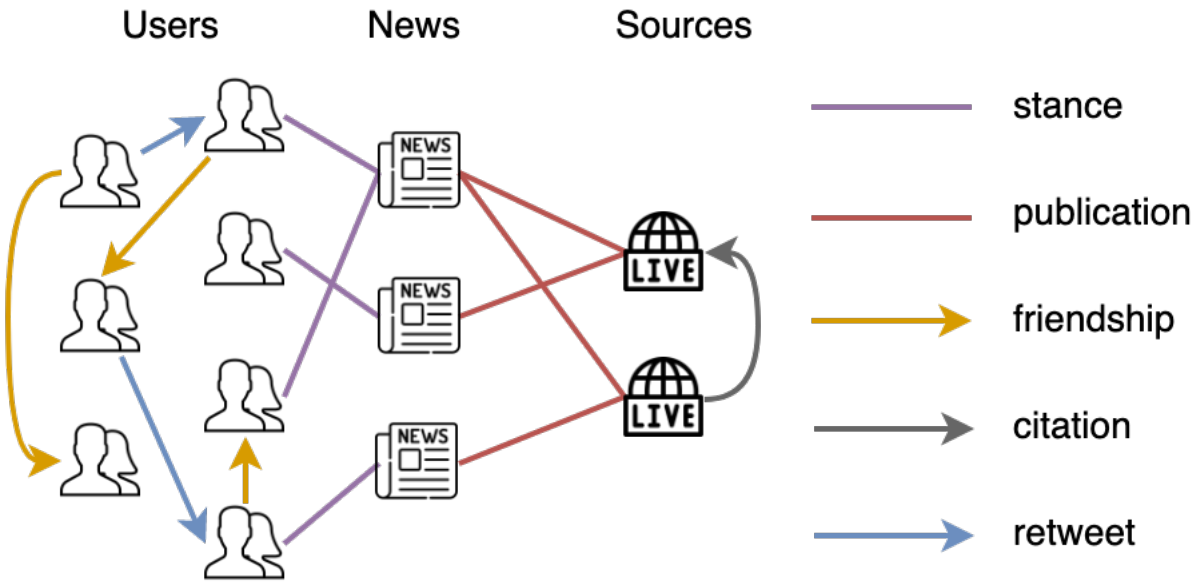

Figure 1.1: Graph Representation of Social Context

or neutral comments. On the other hand, the real news example in Table 1.2 invokes moderate engagements mainly comprised of supportive and neutral comments, and the distribution of stance stabilizes quickly. Such temporal shift in user perception serves as an important signal in identifying false information. Many work has partly proposed a joint-representation of social context features as a network of social actors, whose links model various interactions of either stance expression between social users and news or following/follower relationships within social users (Jin, Cao, Jiang, & Zhang, 2014) (Gupta, Zhao, & Han, 2012)(Popat, Mukherjee, Strötgen, & Weikum, 2017) (Shu, Wang, & Liu, 2019a). However, no works have comprehensively proposed a joint of medium of all major social entities, i.e., questionable news, media source, and social users, and their interactions.

1.3 Our Solution

Social context of news dissemination is inherently represented as a heterogeneous graph whose nodes are the social entities and their auxiliary features, and whose edges are the social interactions. A graph medium has several advantages over some existing Euclidean-based methods (Ruchansky, Seo, & Liu, 2017) (Liu & Wu, 2018) in terms of structural modeling capabil-

News title	Elapsed time - Total tweets	Stance distribution - Support / Deny / Comment_neutral / Comment_negative / Report	Noticeable responses
Virginia Republican Wants Schools To Check Children's Genitals Before Using Bathroom	3h - 38	.00 / .03 / .03 / .16 / .78	"DISGUSED SO TRASNPHOBIC", "FOR GODS SAKE GET REAL GOP", "You cant make this up folks"
	(3h - 6h) - 21	.00 / .10 / .05 / .05 / .80	"Ok This cant be real", "WTF IS THIS BS", "Rediculous RT"
	> 6h - 31	.00 / .10 / .07 / .07 / .76	"Cant make this shit up", "how is this real", "small government", "GOP Cray Cray Occupy Democrats"

Table 1.1: Temporal engagement of social users towards fake news

News title	Elapsed time - Total tweets	Stance distribution - Support / Deny / Comment_neutral / Comment_negative / Report	Noticeable responses
1,100,000 people have been killed by guns in the U.S.A. since John Lennon was shot and killed on December 8, 1980	3h - 9	.56 / .00 / .00 / .00 / .44	"#StopGunViolence", "guns r the problem"
	> 3h - 36	.50 / .00 / .11 / .00 / .39	"Some 1.15 million people have been killed by firearms in the United States since Lennon was gunned down", "#StopGunViolence"

Table 1.2: Temporal engagement of social users towards real news

ity for several phenomena like filtering bubbles of users or polarized network of news media. Network modelling also allows entities to exchange information, via homogeneous edges, i.e., user-user relationship, source-source citations, heterogeneous edges, i.e., user-news stance expression, source-news publication, and high order proximity, i.e., between users who consistently support or deny certain sources, which are illustrated in Figure 1.1. Previous attempts to learn such structure have employed spectral-based approaches (Shu et al., 2019a) (Gupta et al., 2012), whose expensive matrix factorization lacks consideration for higher order proximity, and whose representations are highly specific to the discussed task. Furthermore, there has been no in-depth analysis of the learned embeddings and their correlation to underlying communities.

We propose Fake News Graph (FANG), a graph learning framework for the proposed network in both supervised and unsupervised settings. FANG explores new information sources by looking at the context of news spreading pattern and temporality. FANG models the social actors (news, sources, users) under a common medium that accounts for their interaction, and uti-

lizes recent advancements in Graph Neural Network (GNN) (Kipf & Welling, 2016a) (Grover & Leskovec, 2016). Given a joint representation space of social actors, we examine if separate node classification tasks, i.e., troll detection for users, FND for news, bias/factuality detection for sources, can benefit from a supervised multi-task learning objective, or unsupervised pre-trained context-aware embedding. FANG outperforms Euclidean baseline on the task of FND and achieves competitive performance given limited labels. We also analyze how the learned structural representations of social actors correlate with their underlying communities such as echo chambers of users or polarized networks of media, and how robust they are to various downstream tasks.

Our major contributions are summarized as follows:

1. We provide a novel graph representation for the social context of news dissemination, which models all major social actors and interactions.
2. We propose FANG, a novel graph learning framework that explores the unique structures of social context graph in both supervised and unsupervised settings.
3. We conduct extensive experiments to assess the performance of FANG in FND and its produced structural representations on downstream classification tasks of social entities.

1.4 Structure of Our Report

This report is organized as follows:

- In Chapter 1, we introduce the problem of fake news, motivate the exploration of social context and present our solution of a novel graph representation and learning framework.
- In Chapter 2, we review the related literature in the domain of FND and graph learning frameworks.
- In Chapter 3, we formalize the problems of representing social context and detecting fake news. We describe the proposed methodology to address the problems in details.
- In Chapter 4, we evaluate our proposed solution by analyzing the experimental results.

- In Chapter 5, we conclude the report with a discussion on future research directions.

As our work is ongoing, some aspects of the problem have not been discussed yet and will be revisited in later reports.

Chapter 2

Related Work

We review existing works on representation learning frameworks of social context, specifically euclidean-based and network-based, which are more relevant to our research. We also discuss the recent advancement in Geometric Deep Learning framework, and how they can be applied to our social context graph. Here, we highlight the advantages and disadvantages of different methods, and motivate the choice of our solution.

Euclidean-based approaches seek to incorporate social context by extracting features of media or users who spread the news and combining those with content-based features (Castillo, Mendoza, & Poblete, 2011) (Yang, Liu, Yu, & Yang, 2012). User features at individual-level (Shu et al., 2017) consider both attributes such as demographics, information preferences, social activeness, and network structure such as follower or friend counts, which do not reveal much about each user's neighborhood other than his centrality. Even at group-level, user features are merely the aggregation of individual-level features like "average number of users" (Ma, Gao, Wei, Lu, & Wong, 2015). Such representations are relatively easy to construct and analyzed but completely lack the modelling capability for users from a common community. Furthermore, users can engage to a piece of news in different ways by expressing different stances or sentiments via various temporal patterns, which was not considered in these early works.

Networks in which nodes constantly exchange and propagate information such as trust (Kamvar, Schlosser, & Garcia-Molina, 2003) or any auxiliary attributes (Liao, He, Zhang, & Chua, 2018) have been widely discussed. Recent works in detection falsehood has generalize the idea

to social context by modelling an underlying user or news source network and dedicate a representation that captures an entity's structural features. CSI (Ruchansky et al., 2017) used linear dimensionality reduction on user co-sharing adjacency matrix and combine it with the news engagement feature obtained from a recurrent neural network. TriFN (Shu et al., 2019a), which might seem similar to our proposal, neither differentiated user engagements in term of stance and temporal patterns, nor modeled source-source citation. Furthermore, such matrix decomposition approaches, including CSI (Ruchansky et al., 2017), are potential expensive in term of graph node counts and ineffective in modeling high-order proximity. Other works on citation source network (Kulkarni, Ye, Skiena, & Wang, 2018), propagation network (Monti, Frasca, Eynard, Mannion, & Bronstein, 2019), rumor detection (Yuan, Ma, Zhou, Han, & Hu, 2019) utilized recent advances in graph neural network and multi-head attention attention to learn both local and global structural representation. However, the authors neither considered a comprehensive social context graph, nor presented any study for the embedding expressiveness for underlying social structures such as echo chambers or polarized media networks, probably due to an incomplete graph representation.

Recent work in Graph Neural Network (GNN) have successfully generalize Deep Learning methods to learn both supervised and unsupervised representation of objects with complex relationships and interdependency. GCN (Kipf & Welling, 2016a) effectively approximate the parameters of convolutional filters in large graph. Other unsupervised approaches such as Node2Vec (Grover & Leskovec, 2016) generalize distributed sampling to construct structural representation of graph nodes. Such spatial-based approaches using GNN have major advantages over spectral-based matrix factorization in term of modeling capacity for high-order proximity, edge-aware feature propagation (Veličković, Cucurull, Casanova, Romero, Lio, & Bengio, 2017) and scalability in number of graph nodes.

Table 2.1 summarizes the comparison between the discussed works on representation learning frameworks of social context.

Approach	Category	Modeled social entities & interactions	Temporal	Learning Framework	Structural Expressiveness
Castillo (2011) (Castillo et al., 2011), Yang (2012), Ma (2015) (Ma et al., 2015) (Yang et al., 2012)	Feature-based model	Users, News	No	Euclidean-based learning	No
CSI (Ruchansky et al., 2017)	Graph representation & Euclidean learning	User, News, User-user followership, User-news interaction	Yes	Matrix factorization, Recurrent Neural Network	No
TriFN (Shu et al., 2019a)	End-to-end Graph	User, News, Sources, User-user followership, User-news interaction, Source-news publication	No	Matrix factorization	No
Kulkarni (2018) (Kulkarni et al., 2018)	Graph representation & Euclidean learning	News, Sources, Source-source citation, Source-news publication	No	GNN, Deep Network	No
Monti (2019) (Monti et al., 2019)	End-to-end Graph	Users, News, User-user followership, User-news interaction	No	GNN	No
GLAN (Yuan et al., 2019)	End-to-end Graph	Users, News, User-news interaction	No	GNN, Multi-head attention	No
FANG (this work)	End2end Graph (supervised), Graph representation & Euclidean learning (unsupervised)	User, News, Sources, User-user followership, User-news interaction, Source-news publication, Source-source citation	Yes	GNN	Yes

Table 2.1: Comparison between representation learning frameworks of social context

Chapter 3

Problem and Algorithm

In this chapter, we first formulate the problem of fake news detection, and our research hypothesis. We then discuss the construction of social context graph, including feature extraction of each social entities and edge interaction labelling. We will describe the methodology of FANG in details as well as the intuition behind it.

3.1 Fake News Detection from Social Context

To ensure a consistent representation, let us first define social context, formulate the task of context-FND and formalize our research hypothesis. The fundamental entities and interactions of the social context C are described as follows:

1. Let $A = \{a_1, a_2, ...\}$ be the list of **news articles** that are being propagated through the social media. a is defined by a feature vector x_a.

2. Let $S = \{s_1, s_2, ...\}$ be the list of **news sources** where each source s_j has published at least one article in A. s is defined by a feature vector x_s.

3. Let $U = \{u_1, u_2, ...\}$ be the list of **social users** where each user has engaged in spreading any article in A or has a connection with any user in U. u is defined by a feature vector x_u.

4. Let $E = \{e_1, e_2, ...\}$ be the list of interactions where each interaction $e = \{v_1, v_2, t, x_e\}$

Interaction	Linking entities	Linking type	Description	Time-sensitive
Followership	User-user	Unweighted, directed	Following/follower relationship on mainstream social media	No
Citation	Source-source	Unweighted, directed	The percentage of reference hyperlink between one media source to another	No
Publication	Source-news	Unweighted, undirected	The relationship between a media source and its published articles	Yes
Stance	User-news	Multi-label, undirected	The perception of social users towards a news article	Yes
Retweet	User-user	Unweighted, directed	The propagation of information from one user towards another	Yes

Table 3.1: Interactions in social context

describes an interaction between two entities $v_1, v_2 \in A \cap S \cap U$ at time t, where t can be absent in interactions that are time-insensitive. The interaction type of e is defined as the label x_e.

The characteristics of each interaction is further described in Table 3.1 and illustrated in Figure 1.1.

Publication, stance and retweets are special types of interaction as it is not only characterized by its spatial features, i.e., edge labels, source/destination nodes, but also by its temporal features, i.e., when the interactions happen. Recent works have highlighted the importance of incorporating temporality in modeling social context engagement, not only in fake news detection (Ruchansky et al., 2017) (Ma et al., 2015), but also in modeling online information dissemination (He, Gao, Kan, Liu, & Sugiyama, 2014). In this work, we are using six stance labels, namely *support*, *deny*, *negative_comment*, *neutral_comment*, *unrelated*, *report*. Four of our stance labels, *support*, *deny*, *comment*, *unrelated* are consistent with the recent work on stance detection (Mohtarami, Baly, Glass, Nakov, Màrquez, & Moschitti, 2018). Based on our observation of the distinguished sentiment of engagement towards fake and real news in Chapter 1, we classify *comment* further into *negative_comment* and *neutral_comment*. We assign the "report" stance label to an user-news engagement when the user simply propagates the news article without expressing any opinion. Altogether, stances are used to characterize news articles by their perceived public opinions, as well as social users by their view on different journalism content. Table 3.2 shows examples for different stances.

News title	Tweet	Stance
US Representatives Agree To Illicit UN Gun Control Plans	More proof we should pull out of the UN and throw them out of the US Pot us Real Donald Trump US Representative	support
	This can't be right	deny
	Oh my giddy aunt	neutral_comment
Pence Michelle Obama Is The Most Vulgar First Lady We've Ever Had	How dare you sir that's our First Lady respect Pence Michelle Obama Is The Most Vulgar First Lady We've Ever Had	negative_comment
	RT Janice GW Pence Michelle Obama Is The Most Vulgar First Lady We've Ever Had USA Newsflash	report
	Lawrence run with this PLEASE	unrelated

Table 3.2: Examples of user stances towards news articles

We refer to the definition of general FND by Kaishu (Shu et al., 2017) and formally define the task of context-based FND.

Definition 3.1.1. *Context-Based FND*: Given a social context $C(A, S, U, E)$ constructed from news articles A, news sources S, social users U, and social interactions E, Context-based FND is defined as the binary classification task of predicting whether a news article $a \in A$ is fake or not, i.e., $F_C : a \rightarrow \{0, 1\}$ such that,

$$F_C(a) = \begin{cases} 0 & \text{if } a \text{ is a fake article} \\ 1 & \text{otherwise} \end{cases}$$

Context-based FND is considered a semi-supervised learning problem where we train our classifier on the partially labelled articles to approximate F_C and predict whether the unlabeled articles are fake or not.

Hypothesis 3.1.1. *Given a social context $C(A, S, U, E)$, we hypothesize that our graph representation and learning framework, i.e., FANG, improve the structural features of social actors A, S, U, resulting in more accurate fake news detection.*

3.2 Graph Construction from Social Context

In this section, we detail our social context graph construction. We specify the selected features for each social entities, i.e., x_a, x_s, x_u, and how we obtain the labels for each social interaction, i.e., x_e.

News Articles: News content is a major component in detecting the authenticity of the news itself. Textual (Castillo et al., 2011) (Yang et al., 2012) (Shu et al., 2019a) (Popat, Mukherjee, Strötgen, & Weikum, 2018) and visual (Wang, Ma, Jin, Yuan, Xun, Jha, Su, & Gao, 2018) (Khattar, Goud, Gupta, & Varma, 2019) have been widely used to model news article contents, either by feature extraction, unsupervised semantics encoding, or learned representation. In this work, we are looking at unsupervised semantics encoding as it is relatively efficient to construct and optimize. We are aware of non-end-to-end limitation and would like to leave this as an open research question.

In the scope of this report, we consider only textual features. For each article $a \in A$, we constructed the tf-idf (Spärck Jones, 2004) vector v_a^t from text body in the article. We increase the semantics expressiveness of token-based representation by weighting them with pre-trained word embeddings obtained from GloVe (Pennington, Socher, & Manning, 2014) to obtain v_a^s. The news article feature vector x_a is created by concatenating v_a^t and v_a^s.

News Sources: Characteristics of media sources that publishes the questionable news have been widely adopted as a essential indicator of the news trustworthy (Baly, Karadzhov, Alexandrov, Glass, & Nakov, 2018) (Kulkarni et al., 2018). Commonly utilized features include journalism topics, lexicon-derived bias, url structure and social network trace (Baly et al., 2018). This report focuses mainly on characterizing media sources by their reporting textual content. For each source s, similar to article representations, we constructed the source feature vector x_s as the concatenation of a bag-of-word tf-idf vector v_s^t and a semantics-sensitive vector v_s^s derive from the "homepage" and "about-us" directory. A portion of fake news spreading websites give a "disclaimer" of being a satirical or sarcastic media in the their "about-us" — a helpful signal for the journalism quality.

Social Users: Online users have been studied extensively as the major propagator of fake news and rumors in social media. As discussed in chapter 2, previous work (Castillo et al., 2011) (Yang et al., 2012) utilized attributes such as demographics, information preferences, social activeness, and network structure such as follower or friend counts. A recent work by Kai Shu (Shu, Zhou, Wang, Zafarani, & Liu, 2019b) conducted a feature analysis on user profile and

pointed out the importance of signals derived from profile description and timeline content. A text description such as "American mom fed up with anti american leftists and corruption. I believe in US constitution, free enterprise, strong military and Donald Trump #maga" strongly indicates the user political bias and suggest the tendency to promote certain narratives. We calculate the user vector x_u as the concatenation of a pair of tf-idf vector v_u^t and semantics vector v_u^s derived from the user profile text description.

Social interactions: For every social actor pairs $(v_i, v_j) \in A \cap S \cap U$, we add an edge $e = \{v_i, v_j, t, x_e\}$ to the list of social interactions E if they interact via interaction type x_e. Specifically, for followership, we examine if user u_i follows user u_j on social media; for publication, we examine if news a_i was published by source s_j; for retweet, we examine if user u_i has ever retweeted user u_j; for citation, we examine if the homepage of source s_i contains any hyperlink to source s_j. In the case of time-sensitive interactions, i.e. *publication*, *retweet* and *stance*, we record their relative timestamp with respect to the article's earliest publication time.

Stance detection: The task of obtaining the correct type for each *stance* interaction is formulated as *stance detection*. Given the large number of user-news interactions and the limited amount of annotation in current social context data, we train a stance classification model. Although the labeled data for tweet stance classification is limited, the abundance of labeled data for stance classification on general news (Hanselowski, PVS, Schiller, Caspelherr, Chaudhuri, Meyer, & Gurevych, 2018) and tweet pair classification (Xu, Callison-Burch, & Dolan, 2015) inspires us to conduct Transfer Learning (TL) (Pan & Yang, 2009). The TL architecture is described in figure 3.1. Notice that this is a cross-corpus, i.e. from general news to tweets, and cross-task, i.e. from tweet paraphrase identification to stance detection, TL.

The first model is trained on tweet corpus with the objective to detect if two input tweets are paraphrases, while the second model is trained on news article corpus with the objective to detect the stance of an article's body towards its headline (Hanselowski et al., 2018). The final stance detection model is trained on a limited amount of annotated tweet pairs with stance labels from FakeNewsNet (Shu, Mahudeswaran, Wang, Lee, & Liu, 2018) of 2110 samples of 474 supports, 298 denies, 387 comments and 951 unrelated tweets. This model is trained with using

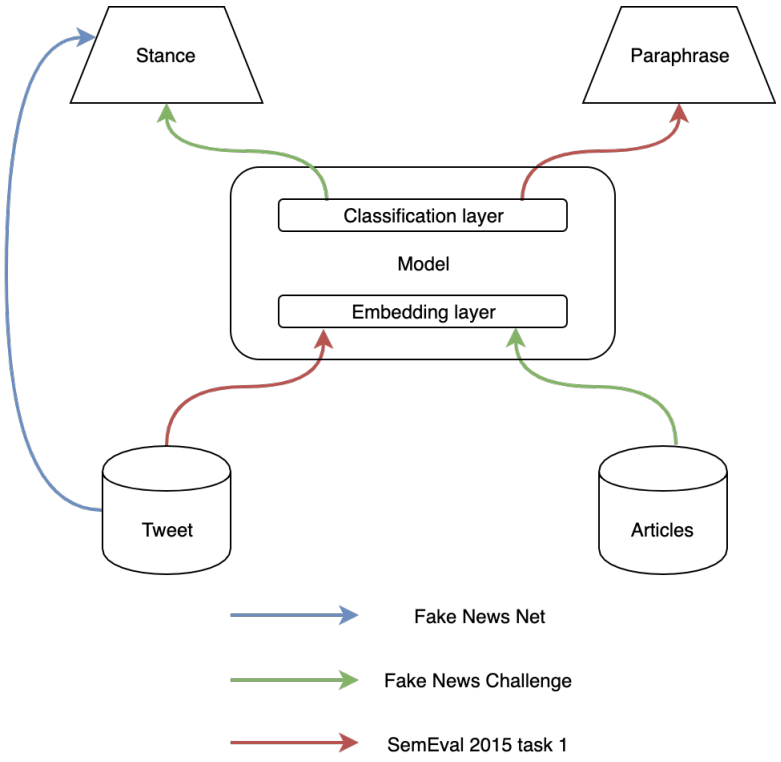

Figure 3.1: Cross-corpus, cross-task TL for stance detection

the embedding layer initialized with the first model's pretrained weights and the classification layer initialized with the second model's pretrained weights. We choose BiDAF (Seo, Kembhavi, Farhadi, & Hajishirzi, 2016) for our classifier as it is a light-weight deep learning model to learn the mutually attentive representations of two input sentences. Experiment results in table 3.3 show that our TL BiDAF reached 0.9 macro F1 and manage to improves state-of-the-art pretrained BERT-based (Devlin, Chang, Lee, & Toutanova, 2019) language models by 31.8% macro F1 error rate. For each user-news interaction $e = (u_i, a_j, t, x_e)$ at time t, we input the tweet text and news article's title into our stance detection model to obtain the stance label x_e. We also finetune a pre-trained Roberta model (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, & Stoyanov, 2019) on Stanford Sentiment Treebank[1] to obtain a sentiment classifier and differentiate *neutral* and *negative* comments.

[1] `https://nlp.stanford.edu/sentiment/treebank.html`

Model	Accuracy	Precision	Recall	Macro F1
BiDAF	0.8657	0.8558	0.8454	0.8465
Bert-base-uncased (Devlin et al., 2019)	0.8458	0.8153	0.8248	0.8194
XLNet-base-cased (Yang, Dai, Yang, Carbonell, Salakhutdinov, & Le, 2019)	0.8308	0.7986	0.8004	0.7976
XLM (Lample & Conneau, 2019)	0.8109	0.7806	0.7835	0.7817
Roberta-base (Liu et al., 2019)	0.8806	0.8545	0.8678	0.8588
TL BiDAF	**0.9154**	**0.9059**	**0.9032**	**0.9037**

Table 3.3: Stance Detection Results

3.3 FANG

In this section, we discuss the our current research progress on Time-insensitive supervised FANG — the proposed Graph Learning Framework in supervised learning without considering temporality. We will discuss the potential research for time-sensitive supervised FANG, as well as unsupervised FANG in section 5.2.

Time-insensitive Graph Neural Network baseline: Inspired by the recent advancement of Graph Convolutional Networks (GCN) (Kipf & Welling, 2016a), we propose a baseline GCN-FANG which is a direct application of GCN to our proposed graph representation. Let the total number of social entities to be $N = |A| + |S| + |U|$. Each relation r_k amongst our ten relations (four non-stance relations and six stances) is represented as a binary adjacency matrix $M_k \in \mathbb{R}^{N \times N}$ such that

$$M_k^{(ij)} = \begin{cases} 1 & \text{if their is } r_k \text{ relation between the } i\text{th and the } j\text{th social actor} \\ 0 & \text{otherwise} \end{cases}$$

where $M_k^{(ij)}$ is the element at ith row and jth column of M_k.

In consistent with the original paper, we define $X \in \mathbb{R}^{N \times d_1}$ to be the input node feature matrix where $X^{(i)} \in \mathbb{R}^{d_1}$ is the feature vector of ith social actor defined in section 3.2. Let d_l be the size of any node hidden representation obtained at the lth layer of GCN-FANG. We define the convolution operation in our graph consisting of multi-label edges as

$$H_{l+1} = \|_{k=1}^{K} (\tilde{D}^{-\frac{1}{2}} \tilde{M}_k \tilde{D}^{-\frac{1}{2}} H_l \Theta_l) \tag{3.1}$$

where $H_l \in \mathbb{R}^{N \times d_l}$ is the hidden representation matrix of social actors at lth layer, $\|$ is

the concatenation operation, $K = 10$ is the number of relations, $\Theta \in \mathbb{R}^{d_l \times d_{l+1}}$ is the matrix of optimizable weights at lth layer, and $\tilde{D}^{-\frac{1}{2}}\tilde{M}_k\tilde{D}^{-\frac{1}{2}}$ is the computation to obtain normalized graph Laplacian matrix. Notice that $H_1 = X$ is the input feature matrix. After L layers, we obtain the output vector as the final hidden representation of each article a, or $\boldsymbol{o_a} = H_L^{(\bar{a})}$ where $\bar{a}$ is the index of a. $\boldsymbol{o_a}$ is passed through a softmax activation function σ, and all learnable weights are trained using cross-entropy loss as defined by $\mathcal{L}$:

$$\mathcal{L} = \frac{1}{|A_c|}\sum_{a}^{A_c} \boldsymbol{y}_a \cdot log(\sigma(\boldsymbol{o}_a)) + (1 - \boldsymbol{y}_a) \cdot log(1 - \sigma(\boldsymbol{o}_a)) \tag{3.2}$$

where A_c is a subset of labeled news articles where the label of each article a is denoted as a one-hot encoded $\boldsymbol{y}_a \in \{0, 1\}^2$. The loss function is differentiable, thus trainable with Adam (Kingma & Ba, 2014).

Chapter 4

Evaluation

In this section, we conducted experiments to evaluate the effectiveness of our proposed Graph Representation and Learning Framework. We will find answers to the following research questions:

1. Do the proposed Graph Representation and Learning Framework work better than a euclidean representation baseline?
2. Do the proposed Graph Representation and Learning Framework work well given limited training data?
3. Do the proposed Graph Representation and Learning Framework given missing auxiliary features?
4. Do the context-based models work better given temporal features?

4.1 Experiment Settings

Baselines: To answer the first, third and fourth research question, we choose CSI (Ruchansky et al., 2017) as our euclidean-based representation and learning baseline. We construct three variants

- t-CSI where we retain the original model

	FakeNewsNet	Controlled
#Users	345440	2558
#Sources	566	35
#News	1056 (432 fake, 624 real)	40 (20 fake, 20 real)
#Citation density		3%
#Following density		0.003%
#Stance density (support + deny)		2.5%

Table 4.1: Dataset statistics

- CSI where we remove the timestamp of social engagement
- f-CSI where we concatenate the user feature vector and source user vector obtained in section 3.2 to CSI's news article hidden representations

For our proposed GCN-FANG model, we also obtain two corresponding variants, GCN-FANG where we remove node features and f-GCN-FANG where we retain node features.

Dataset: The main dataset for this research is FakeNewsNet (Shu et al., 2018), which contains the attributes and interactions of social actors. As the full dataset from Twitter is currently under retrieval, we extract a much smaller controlled dataset for experimentation which statistics are described in table 4.1. The high sparsity of relation adjacency matrix provides both a limitation and opportunity for optimization.

To answer the second question, we conducted experiments with different ratio of train/test and measured the classification metric of *macro F1 score* on three baselines and two variants of our proposal.

4.2 Results & Discussion

The experiment results are described in table 4.2 and visualized in figure 4.1 and figure 4.2.

Overall improvement: We observe a consistent improvement from modeling social context as our proposed network structure. Both f-GCN-FANG and GCN-FANG outperform its euclidean-based CSI counterpart on almost all settings with different amount of data. The average improvement margins of F1 score are 10.42% and 17% respectively. This can be explained by the richer interactions and structural modeling at both local and global from graph-based approaches.

	f-GCN-FANG	GCN-FANG	f-CSI	CSI	t-CSI
0	1	1	1	1	1
0.1	1	1	1	0.6667	1
0.2	1	0.873	0.8889	0.6667	0.9091
0.3	1	0.8286	0.9091	0.6667	0.8
0.4	0.875	0.873	0.75	0.7692	0.7826
0.5	0.8	0.7494	0.9	0.5556	0.9474
0.6	0.873	0.703	0.8333	0.5455	0.75
0.7	0.9161	0.5942	0.75	0.5263	0.7333
0.8	0.8245	0.7163	0.6923	0.6667	0.7027
0.9	0.8381	0.6364	0.4653	0.381	0.6667
1	0	0	0	0	0

Table 4.2: Experiment results

Limited training data: Both the baselines and our proposed models perform worse when training data is limited. However, if we observe the performance when the validation ratios are 0.2 and 0.9, the gain for F1 score from FANG is much larger given the limited data, which are 11.11% and 37.28% respectively. This proves the robustness of our learning framework especially with limited training quantity.

Limited or absent auxiliary features: Both the baselines and our proposed models perform worse when node features are missing. However, the improvement margin for F1 score is greater between GCN-FANG and CSI, indicating that our proposed Graph Representation and Learning Framework excel at detecting fake news when social actors' attributes are absent or limited. The experiment results also emphasize the importance of incorporating auxiliary social features together with structural features to obtain the best performance.

Temporality improvement: Figure 4.2 highlights the improvement from incorporating temporal features to CSI baseline, which gives t-CSI a consistent and average improvement margin for F1 score of 20.53%. This complies with our observation on the distinguished temporal engagement patterns between false and real news that can be utilized for fake news detection.

Overall, the experiment results confirm our research questions on the overall improvement, improvement given limited training data, improvement given absent auxiliary features and improvement given temporality.

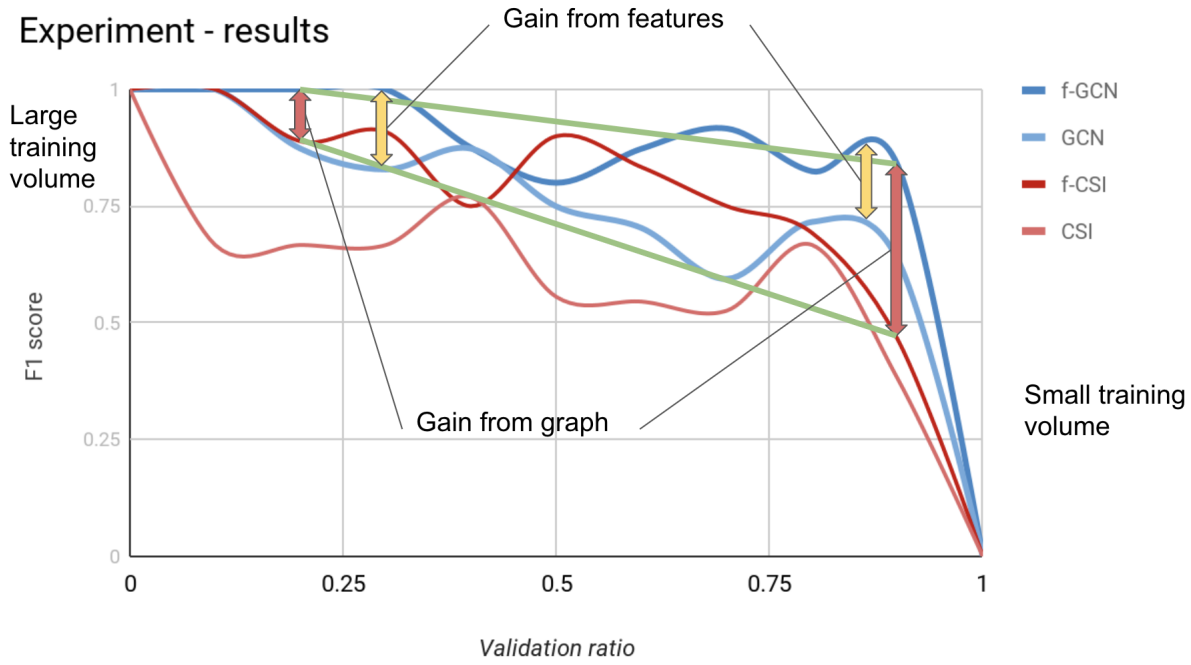

Figure 4.1: Experiment result plot

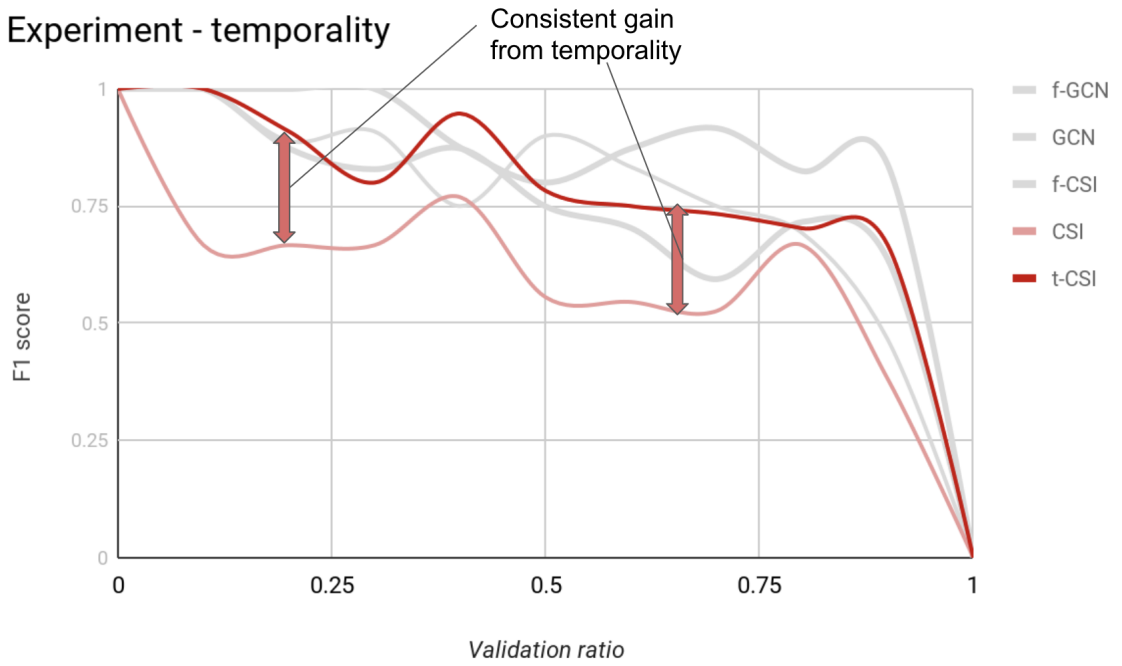

Figure 4.2: Temporal experiment result plot

Chapter 5

Conclusion

5.1 Contributions

In this research report, we have addressed the importance of modeling social context of information propagation as graph. In addition to proposing a novel and comprehensive graph representation, we also described GCN-FANG, a graph learning framework based on GCN that shows consistent improvement over euclidean baseline in the task of fake news detection. We highlighted our greater gain in supervised learning settings of limited auxiliary features and limit training quantity, as well as the importance of incorporating temporal features in learning to represent social context.

5.2 Future Work

This research is yet to complete. Before concluding this report, we would like to discuss several research direction that help us understand more about the manifold of social context and how we can derive intuitively representations of social entities that benefit various downstream tasks.

Unsupervised or self-supervised FANG: As mentioned in section 3.3, we are looking forward to extending our current graph learning framework to unsupervised setting. The objective is to unsupervisedly capture a context-aware representation for each social entity that strongly correlates with underlying social communities such as echo chambers or polarized

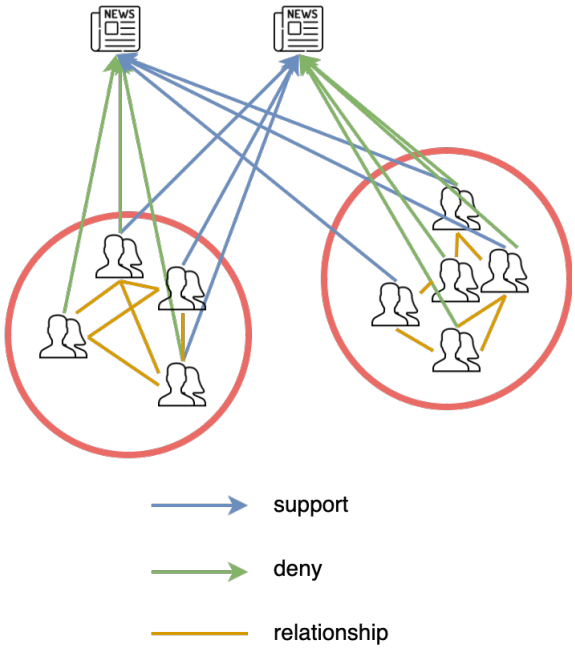

Figure 5.1: Underlying stance-based community

news network. One potential approach is an adaptation of neighborhood sampling scheme like Node2Vec (Grover & Leskovec, 2016), where the specific sampling strategy, i.e. which edge to traverse, walk length, context size, is different between social entities. A GNN-based auto-encoding approach, such as VGAE (Kipf & Welling, 2016b) can also be considered. We are also inspired by the recent breakthrough of pretrained language model with self-supervised objective (Devlin et al., 2019). A research direction is to structurally represent a node by predicting for readily available links in our networks such as stances. Any unsupervised representation strategy for users that is stance-oriented can help us capture groups of users who share a common narrative, as illustrated in figure 5.1, which is an indication of echo chamber effect. Such unsupervised representation framework can also be time-sensitive and captures those who often engage in spreading a news article after a certain amount of time since publication.

Mutli-task and multi-domain generalization: Our graph provides a common medium for joint representations of heterogeneous social actors. Although we are examining fake news detection, any node classification task can benefit from our representation and learning framework, such as source bias and factuality prediction (Baly et al., 2018) or troll user detection (Atanasov, Morales, & Nakov, 2019), either in a supervised setting with multi-task objective, or in an un-

supervised setting by directly consuming our pretrained representations. We also realize that social context is a shared aspect in both fake news detection and rumor detection (Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018) and are looking forward to generalizing our approach to the other domain.

References

Atanasov, A., Morales, G. D. F., & Nakov, P. (2019). Predicting the role of political trolls in social media. *arXiv preprint arXiv:1910.02001*, , 2019.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3528–3539), 2018.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th international conference on World wide web* (pp. 675–684), ACM, 2011.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186), 2019.

Edkins, B. (2016). Americans believe they can detect fake news. studies show they can't.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855–864), ACM, 2016.

Gupta, M., Zhao, P., & Han, J. (2012). Evaluating event credibility on twitter. *Proceedings of the 2012 SIAM International Conference on Data Mining* (pp. 153–164), SIAM, 2012.

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180*, , 2018.

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

He, X., Gao, M., Kan, M.-Y., Liu, Y., & Sugiyama, K. (2014). Predicting the popularity of web 2.0 items based on user comments. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 233–242), ACM, 2014.

Jin, Z., Cao, J., Jiang, Y.-G., & Zhang, Y. (2014). News credibility evaluation on microblog with a hierarchical propagation model. *2014 IEEE International Conference on Data Mining* (pp. 230–239), IEEE, 2014.

Kamvar, S. D., Schlosser, M. T., & Garcia-Molina, H. (2003). The eigentrust algorithm for reputation management in p2p networks. *Proceedings of the 12th international conference on World Wide Web* (pp. 640–651), ACM, 2003.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). Mvae: Multimodal variational autoencoder for fake news detection. *The World Wide Web Conference* (pp. 2915–2921), ACM, 2019.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, , 2014.

Kipf, T. N., & Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, , 2016.

Kipf, T. N., & Welling, M. (2016b). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, , 2016.

Kulkarni, V., Ye, J., Skiena, S., & Wang, W. Y. (2018). Multi-view models for political ideology detection of news articles. *arXiv preprint arXiv:1809.03485*, , 2018.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, , 2019.

Liao, L., He, X., Zhang, H., & Chua, T.-S. (2018). Attributed social network embedding. *IEEE Transactions on Knowledge and Data Engineering*, *30*(12), 2018, 2257–2270.

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

Liu, Y., & Wu, Y.-F. B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ludolph, R., Allam, A., & Schulz, P. J. (2016). Manipulating google's knowledge graph box to counter biased information processing during an online search on vaccination: Application of a technological debiasing strategy. *Journal of Medical Internet Research*, *18*(6), 2016.

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

Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., & Moschitti, A. (2018). Automatic stance detection using end-to-end memory networks. *arXiv preprint arXiv:1804.07581*, , 2018.

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, , 2019.

Ng, H. (2018). 4 in 5 singaporeans confident in spotting fake news but 90 per cent wrong when put to the test: Survey.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 2009, 1345–1359.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543), 2014.

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1003–1012), International World Wide Web Conferences Steering Committee, 2017.

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2018). Credeye: A credibility lens for analyzing and explaining misinformation. *Companion Proceedings of the The Web Conference 2018* (pp. 155–158), International World Wide Web Conferences Steering Committee, 2018.

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo chambers on facebook. *Available at SSRN 2795110*, , 2016.

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

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, , 2016.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, , 2018.

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

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

Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2019b). The role of user profile for fake news detection. *arXiv preprint arXiv:1904.13355*, , 2019.

Spärck Jones, K. (2004). Idf term weighting and ir research lessons. *Journal of documentation*, *60*(5), 2004, 521–523.

Thorne, J., & Vlachos, A. (2017). An extensible framework for verification of numerical claims. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 37–40), Association for Computational Linguistics, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*, , 2017.

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

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857), ACM, 2018.

Xu, W., Callison-Burch, C., & Dolan, B. (2015). SemEval-2015 task 1: Paraphrase and semantic similarity in twitter (PIT). *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 1–11), Denver, Colorado, June, 2015: Association for Computational Linguistics.

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on sina weibo. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics* (p. 13), ACM, 2012.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, , 2019.

Yuan, C., Ma, Q., Zhou, W., Han, J., & Hu, S. (2019). Jointly embedding the local and global relations of heterogeneous graph for rumor detection. *arXiv preprint arXiv:1909.04465*, , 2019.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, *51*(2), 2018, 32.