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: `https://github.com/nguyenvanhoang7398/GraphLearning`

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 news such as pandemic breakout, presidential elections, and so on, information factuality is an utmost concern in coordinating social behaviours and ensuring public fairness and rationality. Unfortunately, the tremendous volume and amazing speed of news propagation makes 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 fewer assumptions about the news content compensated by the large and readily available wisdom of the crowd.

In this work, we propose Factual News Graph (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. We observe consistent improvement across different data availability even when the amount of training data is highly limited. Our recurrent aggregator accounts for temporal propagation patterns and enhances explainability with attention mechanism. FANG is highly scalable in large graphs without concurrently storing all node features during training, and efficient in inference without re-processing the whole graph. Beyond fake news detection, our FANG also improves other social network analysis tasks such as assessing the factuality of journalism sources.

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. The Internet connects billions of users and is an influential medium to spread information. However, its starling speed is a challenge to any effort to verify information as scale, posing a great risk of unchecked falsehood. According to MIT research (Vosoughi, Roy, & Aral, 2018), falsehood reaches 1,500 people 6 times faster than true stories, and is 70 times more likely to be retweeted. In critical news such as pandemic breakout or political elections, false information with malicious intent (Shu, Sliva, Wang, Tang, & Liu, 2017), commonly known as “fake news”, disturbs social behaviors, public fairness and rationality. During the fight against COVID-19, the World Health Organization (WHO) addressed the *infodemic* caused by the life-costing fake news related to coronavirus infection and cures (Thomas, 2020). 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].

Fake news identification is a non-trivial task. Although the general public is confident in its ability to discriminate false information, recent studies have shown the opposite results.

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

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

More specifically, 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 recognizing fake news (Ng, 2018). As the intention behind fake news is to mislead readers, it is challenging to infer news veracity solely based on its content (Shu et al., 2017). Many sites and social media have devoted great effort in identifying false information. Facebook encourages users to report non-credible posts[3] and employs professional fact-checkers to expose questionable stories. This method is also used by fact-checking websites such as Snopes[4] or Politifact[5]. To scale 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).

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 could have spread and caused severe damage. Recent research direction in fake news detection takes another turn and explores various features of the news dissemination process. Our observations show that there is a distinctive engagement pattern of social users towards fake and real news. The fake news example in Table 1.1 had many engagements shortly 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 or neutral comments. On the other hand, the real news example in Table 1.2 invokes a moderate amount of engagements, mainly com-

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

[4] `snope.com`

[5] `politifact.com`

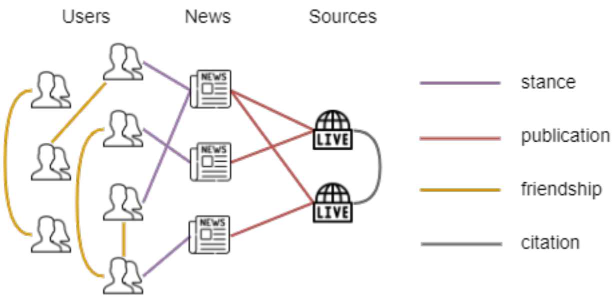


Figure 1.1: Graph representation of social context.

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

prised of supportive and neutral comments that stabilize quickly. Such temporal shifts in user perception serves as important signals in identifying false information. Many works have 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 so far a joint medium of all major social entities, *i.e.*, questionable news, media source, and social users, together with their interactions.

1.3 Our Approach

Social context of news dissemination can be inherently represented as a heterogeneous network whose nodes are the social entities and edges are the social interactions. A graph-based medium has several advantages over some existing Euclidean-based methods (Ruchansky, Seo, & Liu,

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

2017) (Liu & Wu, 2018) in terms of structural modeling capability for several phenomena like filtering bubbles of users or polarized network of news media. Network modelling also allows entities to exchange information, via heterogeneous 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, as illustrated in Figure 1.1. Previous works to learn such structure have employed spectral-based approaches (Shu et al., 2019a; Gupta et al., 2012), whose expensive matrix factorization does not account 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, or whether the approach is feasible given a limited annotated news for supervised learning.

We propose Factual News Graph (FANG), a graph learning framework for social context network in minimally supervised settings. FANG explores new information sources by looking at the news spreading pattern and temporality. FANG models the social actors — *i.e.*, news, sources, and users — under a common medium that accounts for their interaction, and utilizes recent advancements in Graph Neural Network (GNN) (Kipf & Welling, 2016; Grover & Leskovec, 2016). Given a joint representation space for social actors, we examine if the downstream node-level source factuality classification (SFC) can benefit from the embedding of the upstream fake news detection. FANG outperforms Euclidean baseline on fake news detection and achieves competitive performance given limited labels. We also analyze how the minimally supervised representations correlate with true labels, improving downstream analysis. The inductive nature of FANG is highly scalable in large graphs without concurrently storing all node features during training, and efficient in inference without re-processing the whole graph. Moreover, the prediction of FANG is easily explained thanks to the attention mechanism of its recurrent aggregator.

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 minimally supervised settings.

3. We conduct extensive experiments to assess the performance of FANG in fake news detection and its produced structural representations on downstream source factuality classification.

Chapter 2

Related Work

In this chapter, we review the existing work on contextual fake news detection and their formulation of news social context. Moreover, we survey recent advances in Graph Neural Network and form the premise of our proposed graph learning framework.

Existing works on contextual fake news detection can be categorized by their approach to representation and learning of social context. Specifically, Euclidean approaches represent social context in the Euclidean space as a flat vector or matrix of real number. These methods typically learns a Euclidean transformation of social entity's features, *i.e.,* the feature of social media who published and spread the news together with the news content itself, that best approximates the fake news prediction. The complexity of these transformation varies from the traditional shallow (as opposed to "deep") models, *i.e.,* Random Forest or Support Vector Machines (SVM) (Castillo, Mendoza, & Poblete, 2011; Yang, Liu, Yu, & Yang, 2012) to deep networks including Long-short Term Memory (Hochreiter & Schmidhuber, 1997) that models engagement temporality (Ruchansky et al., 2017). However, given our formulation of social context as an heterogeneous network of social entities and their interaction, Euclidean representations are less expressive of structural information (Bronstein, Bruna, LeCun, Szlam, & Vandergheynst, 2017). Although pioneering work employed individual attributes such as demographics, information preferences, social activeness, and structural features such as follower or friend counts (Ma, Gao, Wei, Lu, & Wong, 2015; Shu et al., 2017), they do not capture the user interaction landscape *i.e.,* what kind of social figures they follow, which news topics they favor or oppose.

Having acknowledged such limitations, research has started exploring non-Euclidean approaches. 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 on detecting falsehood have generalized the idea of social context by modelling an underlying user or news source network and dedicate a representation that captures an entity's structural features. The "Capture, Score, and Integrate" (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. The "tri-relationship embedding framework" (TriFN) (Shu et al., 2019a), which seems to be 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 potentially expensive in terms 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 to learn both local and global structural representations. Despite the incremental performance improvement in fake news detection, the authors neither considered a comprehensive social context graph nor examined the representation learning of social entities in a minimally supervised setting. Table 2.1 summarizes the comparison between aforementioned works on representation learning frameworks for social context. The social entities and social interactions are (1) users, (2) news, (3) sources, (4) user–user friendship, (5) user–news engagement, (6) source–news publication, (7) source–citation.

We also review the recent advancement of Graph Neural Network (GNN), the premise of our proposed framework. GNNs have successfully generalized Deep Learning methods to model complex relationships and interdependency on graphs and manifolds. Graph Convolution Network (GCN) was one of the first proposals. They effectively approximate the parameters of

Approach	Category	Modeled social entities & interactions	Temporality	Learning Framework	Structural Expressiveness
(Castillo et al., 2011; Ma et al., 2015; Yang et al., 2012)	Feature-based model	1., 2.	No	Euclidean-based learning	No
(Ruchansky et al., 2017)	Graph representation & Euclidean learning	1., 2., 4., 5.	Yes	Matrix factorization, Recurrent Neural Network	No
(Shu et al., 2019a)	End-to-end Graph	1., 2., 3., 4., 5., 6	No	Matrix factorization	No
(Kulkarni et al., 2018)	Graph representation & Euclidean learning	2., 3., 6., 7.	No	Graph Neural Network, Deep Network	No
Monti (2019) (Monti et al., 2019)	End-to-end Graph	1., 2., 4., 5.	No	Graph Neural Network	No
(Yuan et al., 2019)	End-to-end Graph	1., 2., 5.	No	Graph Neural Network, Multi-head attention	No
FANG (this work)	End2end Graph (supervised), Graph representation & Euclidean learning (unsupervised)	1., 2., 3., 4., 5., 6., 7.	Yes	Graph Neural Network	Yes

Table 2.1: Comparison between representation learning frameworks of social context

convolutional filters (Kipf & Welling, 2016). However, spectral-based GCN requires substantial memory footprint to store the entire adjacency matrix, and is not easily adapted to the heterogeneous graph setting, where nodes and edges of different labels have different information propagation patterns. Spatial-based Graph Attention Network (GAT) addressed these limitations with a batched aggregation strategy where information is selectively propagated from neighboring nodes (Veličković, Cucurull, Casanova, Romero, Lio, & Bengio, 2017). Both GCN and GAT work well on semi-supervised problems with moderate training data, but deteriorate sharply given few training examples. Moreover, both are transductive approaches, as they require inferred nodes to be present during training time, limiting the capability of online training. This is especially challenging in contextual fake new detection or general social network analysis as their structure is ever evolving. With these caveats in mind, we build our work upon GraphSage that generates embeddings by sampling and aggregating features from a node's local neighborhood (Hamilton, Ying, & Leskovec, 2017). GraphSage provides great flexibility in defining the information propagation pattern with parameterized random walks and recurrent aggregators. Furthermore, GraphSage is highly suitable for representation learning with unsupervised node proximity loss and can address our problem of minimally supervised fake news detection.

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 detail the methodology of FANG as well as its rationality.

3.1 Fake News Detection from Social Context

To ensure a consistent representation, let us first formally define what we mean by social context, then formulate the task of contextual fake news detection and introduce our research hypothesis. The fundamental entities and interactions of the social context C are defined 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 $\boldsymbol{x}_a$.

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 $\boldsymbol{x}_s$.

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 $\boldsymbol{x}_u$.

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, undirected	Following/follower relationship on mainstream social media	No
Citation	Source-source	Unweighted, undirected	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

Table 3.1: Interactions in social context.

describe 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.

Table 3.1 details the characteristics of each interaction, illustrated in Figure 1.1.

Publication and stance 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 use three stance labels, namely, *support*, *deny*, *report*. Two of them (*support*, *deny* are consistent with the major work on stance detection (Mohtarami, Baly, Glass, Nakov, Màrquez, & Moschitti, 2018). We assign the "report" stance label to a 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.

We refer to the definition of general Fake News Detection by (Shu et al., 2017) and formally define the task of context-based Fake News Detection:

Definition 3.1.1. *Context-Based Fake News Detection*: Given a social context $C(A, S, U, E)$ constructed from news articles A, news sources S, social users U, and social engagements E, Context-based Fake News Detection is defined as the binary classification task of predicting

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.

whether a news article $a \in A$ is fake or not, *i.e.*, $F_C : a \to \{0, 1\}$ such that,

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

Context-based Fake News Detection is considered as 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, improves the structural features of social actors A, S, U, resulting in more accurate fake news detection.*

3.2 Graph Construction from Social Context

We now detail our social context graph construction. We specify the selected features for each social entity (*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, 2018b; 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 take 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 exploit only textual features. For each article $a \in A$, we constructed a *tf.idf* (Spärck Jones, 2004) vector v_a^t from text body in the article. We enrich semantic representation of tokens by weighting them with pre-trained word embeddings obtained by GloVe (Pennington, Socher, & Manning, 2014) to construct 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 an 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 mainly focuses 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 semantically-sensitive vector v_s^s derived 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 works (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 Shu et al. (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 consisting of a *tf.idf* vector v_u^t and a semantic 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 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* and *stance*, we record their relative timestamp with respect to the article's earliest publication time.

Stance detection: The task of obtaining a viewpoint of a source text towards a target text is commonly known as *stance detection* (Küçük & Can, 2020). In our context of fake news detection, the target is the title of the questionable news while the source text is the user comments on when they decide to retweet or reply it. Popular stance detection dataset either do not explicitly describe the target text (Derczynski, Bontcheva, Liakata, Procter, Wong Sak Hoi, & Zubiaga, 2017), have a limited number of targets (Sobhani, Inkpen, & Zhu, 2017; Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016) or formulate the source and target text differently (Fake News Challenge[1]). As one of the frontiers, we have constructed a novel dataset for stance detection between tweets and news. We use this dataset to construct a classifier using a supervised learning framework to predict "support" and "deny" stances. This dataset contains 2,527 labeled source–target sentence pairs from 29 news events. For each event with a reference headline, annotators are given a list of related headlines and tweets. They label whether each related headline and tweet supports or denies the claim made by the reference headline. Besides the reference headline–related headline or headline–related tweet sentence pairs, we also made a second order inference for related headline–related tweet sentence pairs. If such pair expressed a similar stance towards the reference headline, the inferred stance for related headline–related tweet would be "support", and vice versa. Tables 3.3 and 3.4 show some examples annotation and the dataset statistics, respectively. The inter annotator agreement evaluated with Cohen's Kappa score is 0.7776, indicating a substantial agreement.

To choose the best stance classifier, we fine-tune pre-trained language models built upon large-scale transformers like BERT on our dataset. These are the models that have achieved state-of-the-art performance on many natural language understanding tasks (Wang, Singh,

[1] http://www.fakenewschallenge.org/

Event ID	Text	Type	Annotated Stance
greta-pay	Greta Thunberg tops annual list of highest-paid Activists!	reference headline	-
greta-pay	Greta Thunberg is the 'Highest Paid Activist'	related headline	support
greta-pay	No, Greta Thunberg not highest paid activist	related headline	deny
greta-pay	Can't speak for the rest of 'em, but as far as I know, Greta's just a schoolgirl and has no source of income.	related tweet	deny
greta-pay	The cover describes Greta Thunberg to be the highest paid activist in the world	related tweet	support
greta-pay	Can't speak for the rest of 'em, but as far as I know, Greta's just a schoolgirl and has no source of income.	related tweet	deny

Table 3.3: Some examples in the stance-annotated dataset.

	# News	# Samples	# Supports	# Denies
Train	29	2089	931	1158
Test	2	438	207	231

Table 3.4: An example in the stance-annotated dataset

Michael, Hill, Levy, & Bowman, 2018a). Roberta (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, & Stoyanov, 2019) is the best performing model, which achieves an *Accuracy* of 0.8857, *Macro* F_1 of 0.8379, *Precision* of 0.8365, and *Recall* of 0.8395, and is chosen as our stance classifier. The stance prediction of a user–article engagement e is given as $stance(e)$.

3.3 Learning Framework of Factual News Graph (FANG)

In this section, we describe our Factual News Graph learning framework (FANG) learning framework on the social context graph described in Section 3.2. Figure 3.1 illustrates the overview of our FANG. The social entities included are news aricles, media sources and social users.

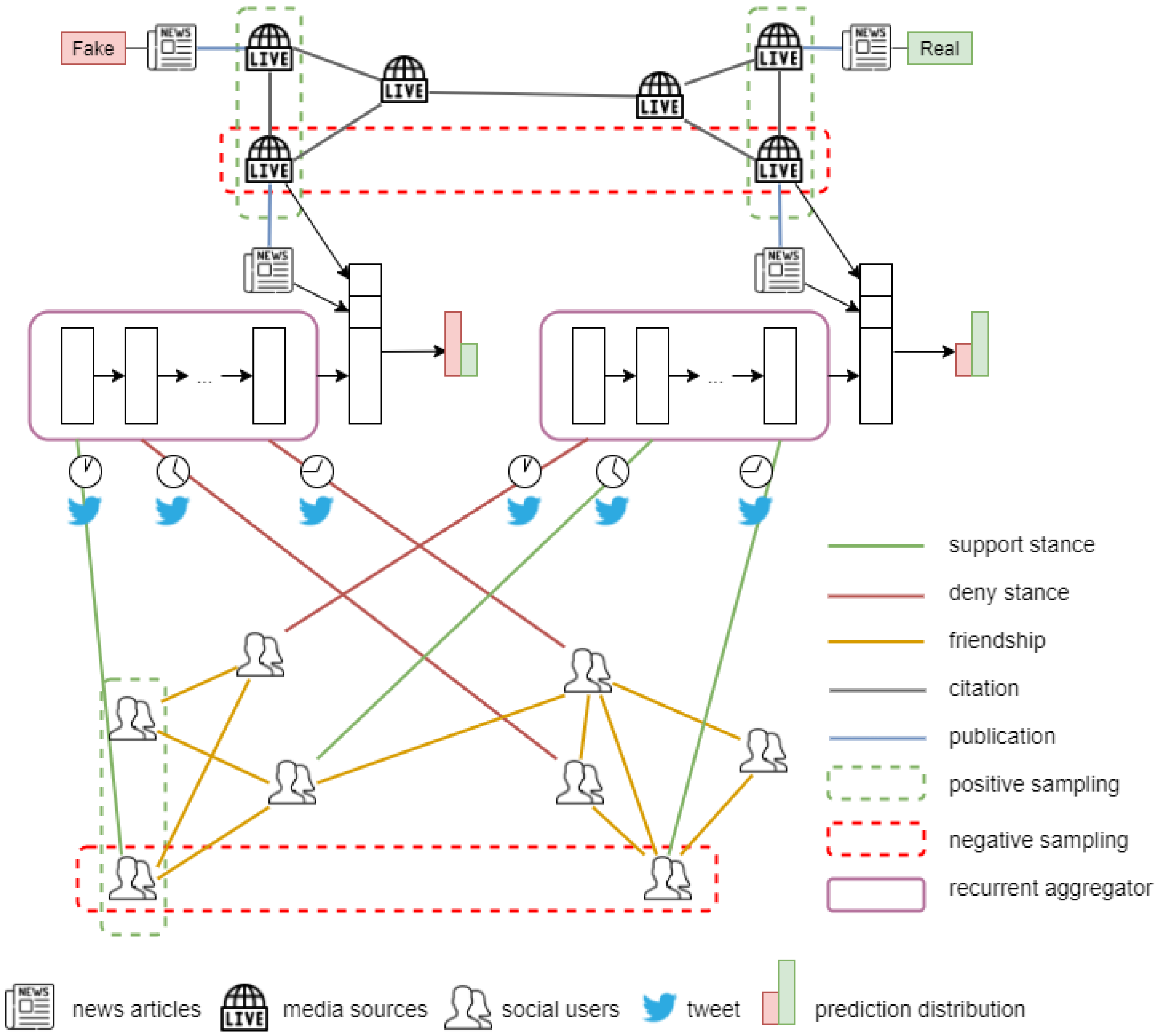


Figure 3.1: Overview of FANG framework.

Social representation of each entities are optimized on the basis of positive and negative sampling. News articles have its sources, its engaged user tweets and their stance aggregated using a recurrent model to predict for the news factuality. As discussed, FANG's motivation is not only to improve the performance of contextual fake news detection, but also to improve the quality of representation learning. That is, while optimizing for the fake news detection objective, FANG also learns meaningful representations of social entities being highly generalizable to downstream tasks. This is achieved by optimizing for three concurrent objectives: unsupervised *Proximity Loss*, self-supervised *Stance Loss*, and supervised *Fake News Detection*.

GraphSage Representation Learning. Before describing the details of each learning objective, we first discuss how FANG derives the representation of each social entities. Previous representation learning frameworks on graph including Deep Walk (Perozzi, Al-Rfou, & Skiena, 2014) and node2vec (Grover & Leskovec, 2016) compute a node embedding by sampling its neighbor and optimizing for the proximity loss. However, these methods are most effective when the node auxiliary features are not available or complete, therefore, separately optimize for each entity structural representation. A recent graph neural network, GraphSage (Hamilton et al., 2017), overcomes this limitation by leveraging node auxiliary features jointly with proximity sampling in representation learning. Let $GraphSage(.)$ be GraphSage's node encoding function, we obtained the structural representation $\boldsymbol{z}_u \in \mathbb{R}^d$ of node u as $\boldsymbol{z}_u = GraphSage(u)$.

Unsupervised Proximity Loss. We derive the Proximity Loss from our hypothesis that closely connected social entities often behave similarly. This hypothesis is motivated by the "echo chamber" phenomenon, where social entities tend to interact with other entities of common interest to reinforce and promote their narrative. The "echo chamber" phenomenon includes inter-cited media sources publishing news of similar content or factuality, and social friends expressing similar stance towards a piece of news or news articles of similar content. Therefore, FANG should assign such nearby entities to a set of closed vectors in the embedding space. We also hypothesize that loosely connected social entities often behave differently. We observe that social entities are highly polarized, especially on left-right politics — the topic most bombarded by false information (Boxell, Gentzkow, & Shapiro, 2017). FANG should enforce the representations of these disparate entities to be highly distinct.

The social interactions that define the above characteristics the most are user–user friendship, source–source citation and news–source publication. As these interactions are either between sources and news or between news, we divide the social context graph into two sub-graphs, *i.e.* news–source sub-graph and user sub-graph. Within each sub-graph G, we formulate the Proximity Loss function as

$$\mathcal{L}_{proximity} = -\sum_{u \in G} \sum_{u_p \in P_u} log(\sigma(\boldsymbol{z}_u^\top \boldsymbol{z}_{u_p})) + Q \cdot \sum_{u_n \in N_u} log(\sigma(-\boldsymbol{z}_u^\top \boldsymbol{z}_{u_n})) \tag{3.1}$$

where $z_u \in \mathbb{R}^d$ is the representation of entity u, P_u is the set of nearby nodes or *positive set* of u, and N_u is the set of disparate nodes or *negative set* of u. P_u is obtained my fixed length Random Walk, while N_u is obtained by negative sampling. Each $u_n \in N_u$ is sampled with the discrete probability density function $\forall v \in G \setminus P_u, P(u_n = v) = \frac{\boldsymbol{x}_u^\top \boldsymbol{x}_v}{Z}$ where $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ are the feature vectors of entity u and v, and Z is the probabilistic normalization factor. In other words, we choose the negative samples of an entity u based on the feature distance between u and the candidate sample v which is not in the positive samples. By minimizing the proximity loss described in Equation (3.1), we effectively force the representations of closed entities to be similar and the representations of disparate entities to be distinct.

Self-supervised Stance Loss. For the user–news interaction in terms of stance, we also put forward a analogous hypothesis. A user and news article pairs who express a stance should have their representation closed in that stance space. For each stance c, we first learn a user projection function $g_c^u(u) = A_c^u z_u$ and a news article projection function $g_c^a(a) = A_c^a z_a$ that maps a node representation of $\mathbb{R}^d$ to representation in stance c space of $\mathbb{R}^{d_c}$. Given a user u and a news article a, we compute their similarity score in stance c space as $g_u(u)^T g_a(a)$. If u actually expresses stance c towards a, we would like to maximize this similarity score, and vice versa, minimize this similarity score otherwise. We interpret this objective as a stance classification objective which can be optimized by the following Stance Loss.

$$\mathcal{L}_{stance} = -\sum_{u,a,c} y_{u,a,c} log(f(u,a,c)) \tag{3.2}$$

where $f(u,a,c)$ is defined as

$$f(u,a,c) = g_u(u)^T g_a(a) \tag{3.3}$$

and

$$y_{u,a,c} = \begin{cases} 1 & \text{if } u \text{ expresses stance } c \text{ over } a \\ 0 & \text{otherwise.} \end{cases}$$

Note that the stance label c of an engagement e between user u and article a is defined as $stance(e)$ using the stance classifier described in Section 3.2. By optimizing for the above stance edge classification objective, we effectively force the representations of an engaged user–news article via a stance to be closed in that stance space.

Supervised Fake News Loss. We directly optimize for the main learning objective of fake news detection via the supervised Fake News Loss. This can be formulated as learning an aggregation function $F(a, s, U)$ that maps a questionable news a, its source s and its engaged users U to a real number in $[0, 1]$, indicating the probability of a being trustworthy. As observed in Chaper 1, we hypothesize that there are distinctive temporal patterns between false and genuine information. Therefore, the aggregating model, *i.e.* the aggregator, has to be time-sensitive. Recurrent Neural Networks fulfilled such a requirement where the Bidirectional Long-short Term Memory (Bi-LSTM) can capture a long-term dependency in information sequence in forward and backward directions (Hochreiter & Schmidhuber, 1997). On top of the vanilla Bi-LSTM, we also incorporate an attention mechanism that focus on essential engagement during the encoding process. Attention mechanism is not only expected to improve our model quality but also its explanability (Luong, Pham, & Manning, 2015; Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017). By examining the model's attention, we would know which Twitter profiles influenced its decision, therefore leverage not only the machine but also human analytic capability. Our proposed LSTM's input is a user–article engagement sequence $\{e_1, e_2, \cdots, e_{|U|}\}$. Let $meta(e_i) \in \mathbb{R}^l = (time(e_i), stance(e_i))$ be the concatenation e_i stance and meta. Each engagement e_i has its representation x_{e_i} following form:

$$\boldsymbol{x}_{e_i} = (\boldsymbol{z}_{U_i}, meta(e_i)) \tag{3.4}$$

where $\boldsymbol{z}_{\boldsymbol{U_i}} = GraphSage(U_i)$, $time(e_i)$ and $stance(e_i)$ are the time since published and stance of engagement e_i respectively.

A BiLSTM encodes the engagement sequence and outputs two sequence of hidden states: a forward sequence, $H^f = \boldsymbol{h}_1^f, \boldsymbol{h}_2^f, \ldots, \boldsymbol{h}_n^f$ that starts from the beginning of the engagement sequence; and a backward sequence, $H^b = \boldsymbol{h}_1^b, \boldsymbol{h}_2^b, \ldots, \boldsymbol{h}_n^b$ that starts from the end of the engagement sequence. For many sequence encoding tasks, knowing both past (left) and future (right) contexts has proven to be effective (Dyer, Ballesteros, Ling, Matthews, & Smith, 2015). The states $\boldsymbol{h}_i^f$ and $\boldsymbol{h}_j^b$ in the forward and backward sequences are computed as follows:

$$\boldsymbol{h}_i^f = LSTM(\boldsymbol{h}_{i-1}^f, \boldsymbol{x}_{\boldsymbol{e_i}}),\ \boldsymbol{h}_j^b = LSTM(\boldsymbol{h}_{j+1}^b, \boldsymbol{x}_{\boldsymbol{e_j}}),$$

where e is the number of encoder units, and $\boldsymbol{h}_i^f, \boldsymbol{h}_j^b \in \mathbb{R}^e$ are the i^{th} and j^{th} hidden state vector of the forward (f) and backward (b) sequence. Let w_i be the attention weight paid by our BiLSTM encoder to the forward $(\boldsymbol{h}_i^f)$ and backward $(\boldsymbol{h}_i^b)$ hidden states. This attention should be derived from the similarity of the hidden state and the news features, *i.e.* how relevant the engaging users are to the discussed content, and the particular time and stance of the engagement. Therefore, we formulate the attention weight w_i as:

$$w_i = \frac{exp(\boldsymbol{z}_a \mathbf{A}^e \boldsymbol{h}_i + meta(e_i)\mathbf{A}^m)}{\sum_{j=1}^{n} exp(\boldsymbol{z}_a \mathbf{A}^e \boldsymbol{h}_j + meta(e_j)\mathbf{A}^m)}. \tag{3.5}$$

Let l, d, e are the meta dimension, structural embedding dimension and encoder dimension, $\mathbf{A}^e \in \mathbb{R}^{d \times e}$ and $\mathbf{A}^m \in \mathbb{R}^{l \times 1}$ are the optimizable projection matrices shared across all engagements. w_i is then used to compute forward and backward weighted feature vectors:

$$\boldsymbol{h}^f = \sum_{i}^{n} w_i \boldsymbol{h}_i^f, \quad \boldsymbol{h}^b = \sum_{i}^{n} w_i \boldsymbol{h}_i^b. \tag{3.6}$$

We concatenate the forward and backward vectors to obtain a single representation engagement $\boldsymbol{r}_a$ for article a, following previous bi-directional LSTM practice (Ma & Hovy, 2016).

To predict whether article a is false, we use a combination of the its structural representation $\boldsymbol{z}_a$, its engagement representation $\boldsymbol{r}_a$ and its source s's structural representation $\boldsymbol{z}_s$. The article's contextual representation is obtained as the combination of all these features:

$$\boldsymbol{v}_a = (\boldsymbol{z}_a, \boldsymbol{r}_a, \boldsymbol{z}_s) \tag{3.7}$$

This contextual representation is input into a fully connected layer whose outputs are computed as follows:

$$o_a = \mathbf{W}\boldsymbol{v}_a + b \tag{3.8}$$

where $\boldsymbol{W} \in \mathbb{R}^{2e \times 1}$ and $b \in \mathbb{R}$ are weights and biases of the layer. The output value $o_a \in \mathbb{R}$ is finally passed through a sigmoid activation function $\sigma(\cdot)$, and trained using cross-entropy Fake News Loss $\mathcal{L}_{news}$ defined as follows:

$$\mathcal{L}_{news} = \frac{1}{T} \sum_{a} \{y_a \cdot log(\sigma(o_a)) + (1 - y_a) \cdot log(1 - \sigma(o_a))\}, \tag{3.9}$$

where $y_a = 0$ if a is a fake article and 1 otherwise.

End-to-end Total Loss. We define the Total Loss as the sum of three losses, namely Proximity Loss, Stance Loss and Fake News Loss.

$$\mathcal{L}_{total} = \mathcal{L}_{proximity} + \mathcal{L}_{stance} + \mathcal{L}_{news} \tag{3.10}$$

The loss function is differentiable, thus trainable with the Adam optimizer (Kingma & Ba, 2014).

Chapter 4

Results & Discussions

In this section, we conduct experiments to evaluate the effectiveness of our proposed graph representation and learning framework. The goal of our experiments is to answer the following research questions (RQs):

- RQ1: Does FANG work better than content-only models, contextual Euclidean models and contextual graph models?
- RQ2: Does FANG work well with limited training data?
- RQ3: What is the rationale behind temporality awareness of our model?
- RQ4: How effective is FANG's representation learning?

4.1 Dataset and Experiment Settings

We conducted our experiments on a Twitter dataset collected by related-work in rumor classification (Ma, Gao, Mitra, Kwon, Jansen, Wong, & Cha, 2016) and fake news detection (Shu, Mahudeswaran, Wang, Lee, & Liu, 2018). For each article, its source and the list of users and their tweets discussing it were collected. Our dataset includes Twitter profile description and the list of Twitter profiles they follow. We crawl additional data of media sources, including the content of their homepage and "about-us" page, together with their frequently cited sources by collecting the hyperlink references in their homepage. The source truths of the articles, namely, whether they are false or genuine, are based on two fact-checking websites — Snopes[1] and Politifact[2]. We release the source code of FANG at `https://github.com/nguyenvanhoang7398/GraphLearning`. Table 4.1 shows some statistics on our dataset.

# Fake News	150
# Real News	151
# Sources	199
# Users	85394
Avg. citation / source	5.11
Avg. publication / source	1.51
Avg. friendship / user	85.78
Avg. support / news	99.14
Avg. deny / news	13.44
Avg. report / news	70.25
Avg. engagement / news	182.84

Table 4.1: Some statistics on our dataset.

4.2 FANG vs Content- and Graph-based Baselines (RQ1)

To address RQ1, we benchmark FANG performance in Fake News Detection against a content-only baseline, a Euclidean contextual baseline, and another graph learning baseline. To compare our FANG with the content-only model, we employ a Support Vector Machine model on TF-IDF feature vectors constructed from news content (see Section 3.2). In addition, to compare our approach with Euclidean models, we choose CSI (Ruchansky et al., 2017), a fundamental yet effective recurrent encoder that aggregate the user features, news content, and user–news engagements. We reimplement CSI with source features by concatenating the overall score for users, the article representation with our formulated source description to obtain the result vector in the Integrate module. We also verify the importance of temporality by conducting the

[1] `https://www.snopes.com`

[2] `https://www.politifact.com`

experiments on two variants for both CSI and FANG: (1) i-CSI and i-FANG without $time(e)$. in the engagement e's representation $\boldsymbol{x_{e_i}}$, and (2) CSI and FANG with $time(e)$. For the third confirmation on the superiority over other graph learning frameworks, we choose Graph Convolutional Network (Kipf & Welling, 2016). To keep consistency with the original work, 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 i^{th} social actor defined in Section 3.2. Let d_l be the size of any node hidden representation obtained at the l^{th} layer of GCN. We define the convolution operation in our graph consisting of multi-label edges as follows:

$$H_{l+1} = \|_{k=1}^{K} (\tilde{D}^{-\frac{1}{2}} \tilde{M}_k \tilde{D}^{-\frac{1}{2}} H_l \Theta_l), \tag{4.1}$$

where $H_l \in \mathbb{R}^{N \times d_l}$ is the hidden representation matrix of social actors at l^{th} layer, $\|$ is the concatenation operation, K is the number of relations, $\Theta \in \mathbb{R}^{d_l \times d_{l+1}}$ is the matrix of optimizable weights at l^{th} 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 $\mathcal{L}$ defined as follows:

$$\mathcal{L} = \frac{1}{|A_c|} \sum_{a}^{A_c} \boldsymbol{y}_a \cdot log(\sigma(\boldsymbol{o}_a)) + (1 - \boldsymbol{y}_a) \cdot log(1 - \sigma(\boldsymbol{o}_a)) \tag{4.2}$$

We use F_1 score–the harmonic mean of Precision and Recall to evaluate all models' performance in Fake News Detection. The macroscopic results are presented in Table 4.2.

All context-aware models, namely i-CSI, CSI, GCN, i-FANG, and FANG, imporves the context-unaware baseline by 7.44% with i-CSI and 27.34% with FANG in F_1 score. This consistent improvements verify that our hypothesis that considering social context improves the accuracy of fake news detection. Secondly, we also observe that both time-sensitive CSI and FANG improves their time-insensitive variants, i-CSI and i-FANG by 5.26% and 6.3% in F_1 score, respectively. The results verify our hypothesis that being aware to the temporality of news spreading pattern is beneficial in fake news detection. Finally, two graph-based models, i-FANG and GCN are consistently better than the Euclidean CSI by 5.61% and 13.6%, respec-

Systems	Context-aware	Temporality-aware	Graph-based	F_1 score
Feature-only SVM				0.6785
i-CSI	✓			0.7529
CSI	✓	✓		0.8055
GCN	✓		✓	0.8090
i-FANG	✓		✓	0.8889
FANG	✓	✓	✓	0.9519

Table 4.2: Comparison between FANG and baseline models in fake news detection evaluated with F_1 score.

tively, and justify the effectiveness of our social graph representation described in Section 3.2. Overall, compred with context-aware, temporality-aware, and graph-based, FANG achieves the best performance at macroscopic level for fake news detection, confirming RQ1.

4.3 Limited Training Data (RQ2)

To address RQ2, we conducted the experiments described in Section 4.2 given varying training data availability. These experiments confirm our model's consistent absolute performance and relative improvement over baseline models given both limited and sufficient data. The experiment results are represented in Table 4.3 and visualized in Figure 4.1.

Systems	F_1@0.1	F_1@0.3	F_1@0.5	F_1@0.7	F_1@0.9
CSI	0.743	0.7443	0.6851	0.7546	0.8055
GCN	0.6123	0.6866	0.6407	0.6523	0.8090
FANG	0.8274	0.8366	0.8010	0.8218	0.9519

Table 4.3: Performance of FANG against baselines by varying training availability evaluated with F_1 score. The last column on F_1@0.9 is identical to Table 4.2.

FANG consistently outperforms two baselines at all given training availability of 10%, 30%, 50%, 70% and 90%. Between graph-based models, GCN's performance drops by 24.31% from

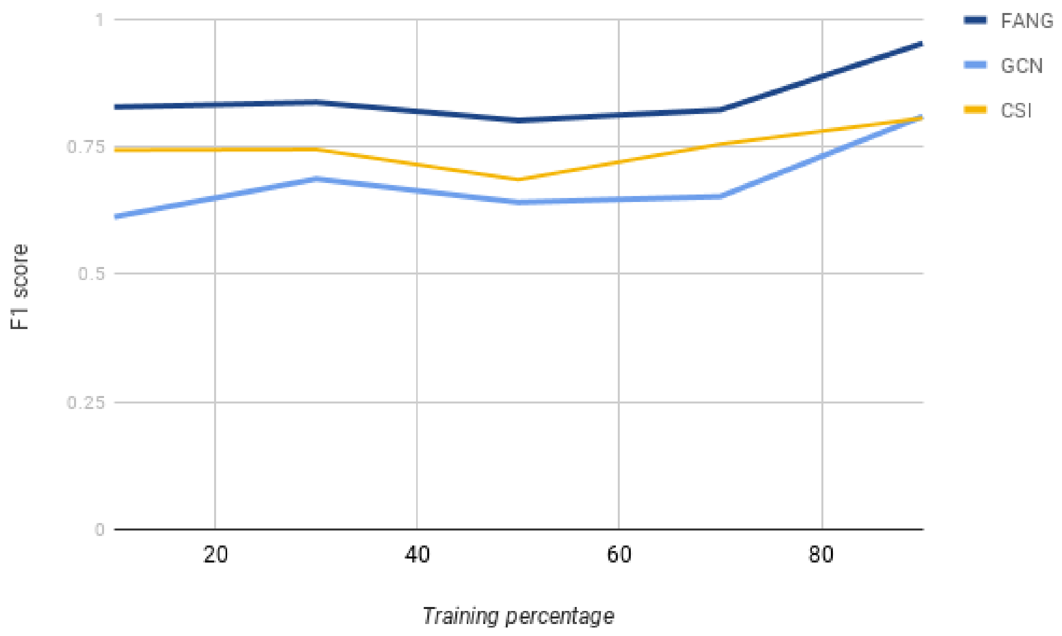


Figure 4.1: Experiment result plot

0.809 in F_1@0.9 to 0.6123 in F_1@0.1, while FANG's performance drops by 13.07% from 0.9519 in F_1@0.9 to 0.8274 in F_1@0.1. We also observe that CSI's performance drops least by only 7.76% from 0.8055 in F_1@0.9 to 0.743 in F_1@0.1. This high generalization can be explained by the simplicity of Euclidean models as they are built upon the "independent and identically distributed" assumption for random variables. Overall, the experimental results highlight our model's effectiveness even at low training availability compared to its GNN and Euclidean counterparts, which confirms RQ2.

4.4 Temporality Study (RQ3)

To address RQ3 and verify that our model makes its decision based on the distinctive temporal patterns between fake and real news, we examine FANG's attention. We accumulate the attention weights produced by FANG within each time window and compare them across time windows. Figure 4.2 shows such attention distribution over time with regard to fake and real news.

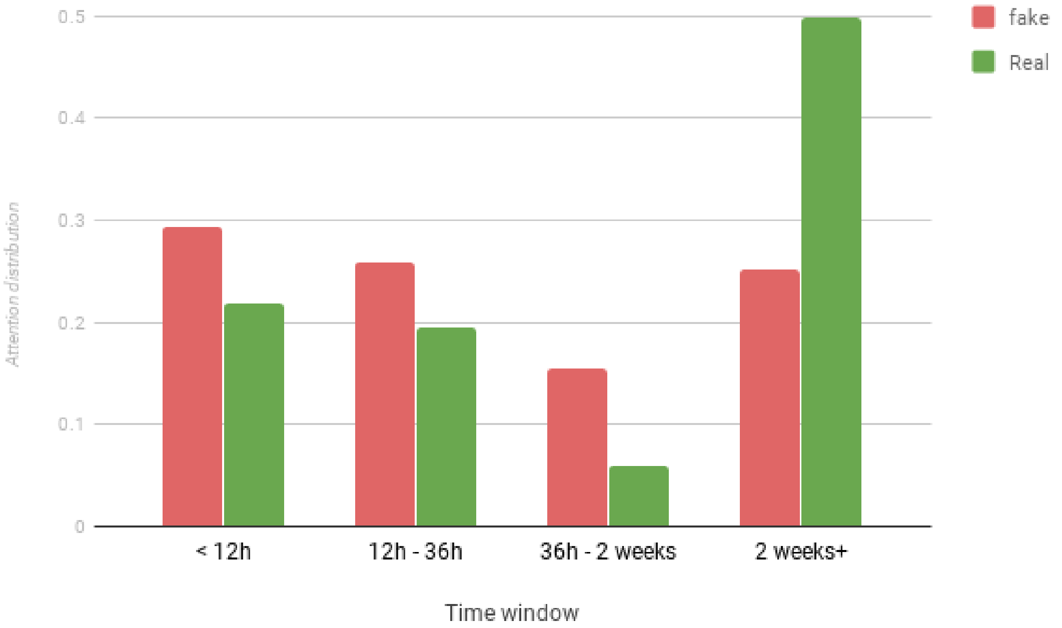


Figure 4.2: FANG's attention distribution across time windows with regards to fake and real news.

We observe that FANG pays almost 60% its attention on the user engagement occurred in the first 36 hours (approximately 30% on the first 12 hours and 26% on the 12 to 36 hours) after a news has been published to decide that it is fake. Its attention then sharply decreases to 15% for the time window of 36 hours to two weeks after publication, and approximately 25% from the second weeks onward. On the other hand, for the decided real news, FANG maintains approximately 40% of its attention on the first 36 hours, but a much longer attention of 50% after two weeks since publication. This characteristics of FANG is consistent with the general observation that the appalling nature of fake news generates the most engagements within a short time after its publication. Therefore, it is reasonable that the model places emphasis on these crucial engagements. On the other hand, genuine news attracts fewer engagements but is circulated for a longer period, thus explains FANG's persistent attention even after two weeks since publication. Overall, the temporality study highlights the transparency of our model decision, largely thanks to the incorporated attention mechanism. We also analyze how FANG's decisions arise from the different spreading patterns of fake and real news, which confirms RQ3.

4.5 Representation Learning (RQ4)

We verify the improved quality of our representations in both intrinsic and extrinsic evaluations.

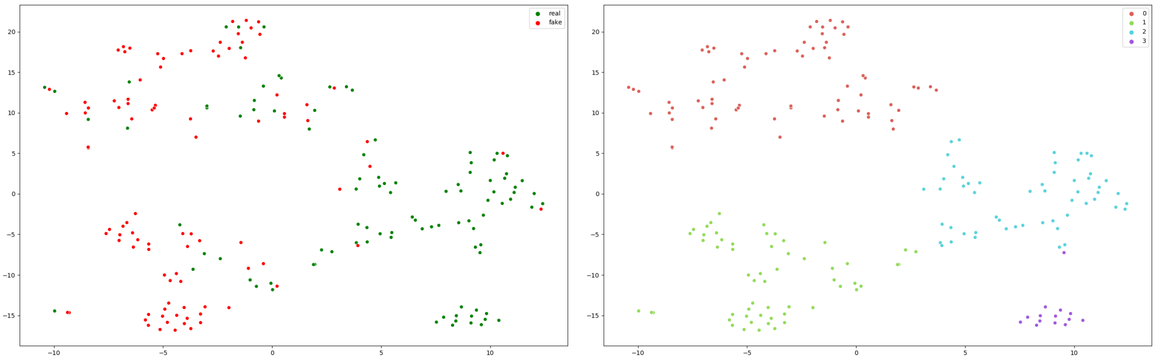


Figure 4.3: 2D t-SNE plot of FANG's representations with factuality news labels (left) and Mean shift clustering labels (right).

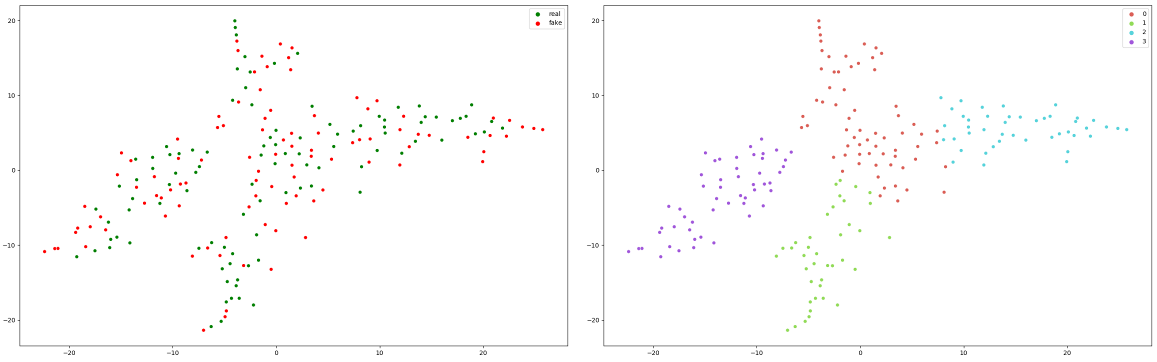


Figure 4.4: 2D t-SNE plot of GCN's representations with factuality news labels (left) and Mean shift clustering labels (right).

In the intrinsic evaluation, we verify how generalizable the minimally supervised news representations are in the intrinsic fake news detection task. More specifically, we first optimize FANG on 10% of training data to obtain all news representations. We then cluster these representations using Mean shift (Comaniciu & Meer, 2002), an unsupervised clustering algorithm, and measure the homogeneity score–the extent to which clusters contain a single class. The higher the homogeneity score is, the more likely news articles of similar factuality labels (*i.e.* fake or real) are closed to each other and the higher quality their representations are. We visualize FANG's representations in 2-dimension with factuality labels and Mean shift clustering labels (Figure 4.3's left and right, respectively). We also provide GCN's representations, which is a fully supervised approach in Figure 4.4, to benchmark FANG's representation quality against.

In the extrinsic evaluation, we verify how generalizable the sufficiently supervised source representation are in the extrinsic source factuality classification task. In specific, we first optimize FANG on 90% of training data to obtain all any source s's representation as $z_s = GraphSage(s)$, and then obtain source s's total representation r_s as follows:

$$v_s = (z_s, x_s, \sum_{a \in publish(s)} x_a) \tag{4.3}$$

where x_s, $publish(s)$, and x_a denote source s's content representation, the list of all articles published by s, and their content representations, respectively.

We propose two baseline representations that do not consider source s content, $v'_s = (z_s, x_s)$. Finally, we train two separate SVM models for v_s and v'_s on the source factuality dataset and record their performance shown in Tables 4.4 and 4.5. respectively.

	High factuality	Low factuality
Train	17	21
Test	22	18

Table 4.4: Some statistics of our source factuality dataset.

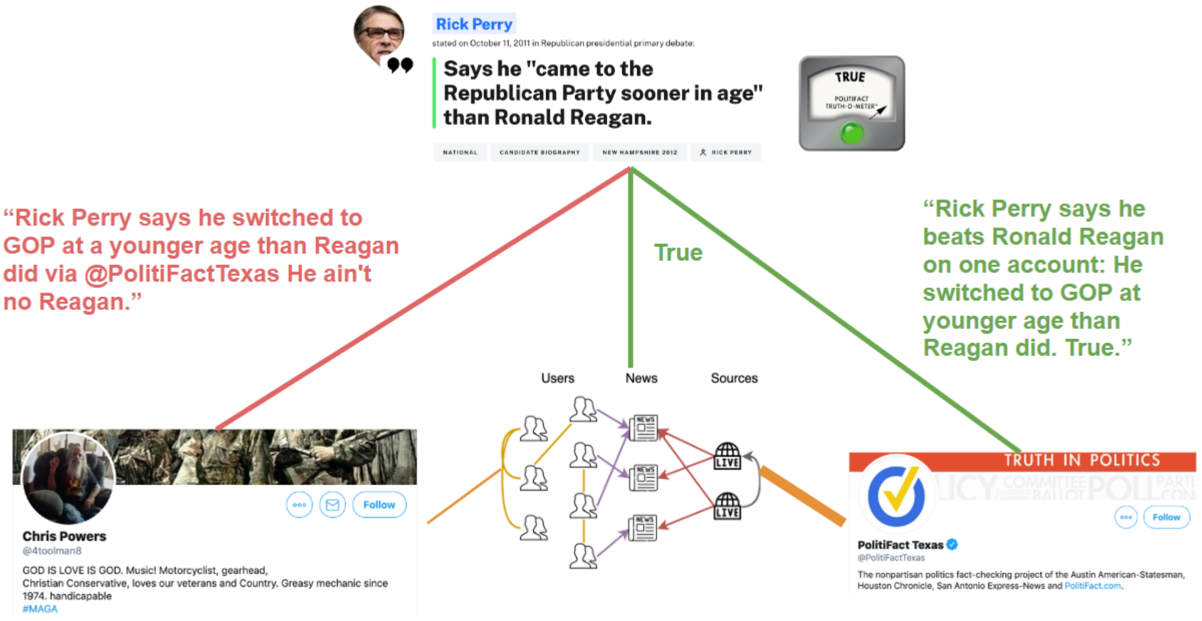
Systems	Context-aware	F_1 score
Baseline		0.5535
Proposed	✓	0.6350

Table 4.5: Performance of our proposed model versus baseline on source factuality classification in terms of F_1 score.

For instrinsic evaluation, the t-SNE plot of labeled FANG's (Figure 4.3's left) representation shows a moderate collocation within both groups of fake and real news, while the t-SNE plot of labeled GCN's representation (Figure 4.4's left) shows no significant collocation within any group of fake or real news. Quantitatively, FANG's Mean shift clusters as shown in (Figure 4.3's right) obtain a homogeneity score of 0.3027 based on news factuality labels, compared with 0.0135 homogeneity score obtained by GCN's Mean shift clusters. This intrinsic evaluation shows FANG's strong representation closeness within both fake and real news groups, indicating our model's improved representation learning over another fully supervised graph neural frameworks. For extrinsic evaluation on downstream source factuality classification, Table 4.5 shows that our proposed method trained using FANG's contextual vector improves the context-unaware baseline by 8.15% in F_1 score. This experiment, however simple, confirms the usefulness of FANG's contextual representation on a task other than fake news detection. Overall, results from intrinsic and extrinsic evaluation confirms RQ4 on the effectiveness of FANG's representation learning.

4.6 Microscopic Analysis

In this section, we examine our model's prediction on specific test examples in Figure 4.5 and Figure 4.6.



Rick Perry

stated on October 11, 2011 in Republican presidential primary debate:

Says he "came to the Republican Party sooner in age" than Ronald Reagan.

"Rick Perry says he switched to GOP at a younger age than Reagan did via @PolitiFactTexas He ain't no Reagan."

"Rick Perry says he beats Ronald Reagan on one account: He switched to GOP at younger age than Reagan did. True."

Chris Powers

@4toolman8

GOD IS LOVE IS GOD. Music! Motorcyclist, gearhead, Christian Conservative, loves our veterans and Country. Greasy mechanic since 1974. handicapable #MAGA

PolitiFact Texas

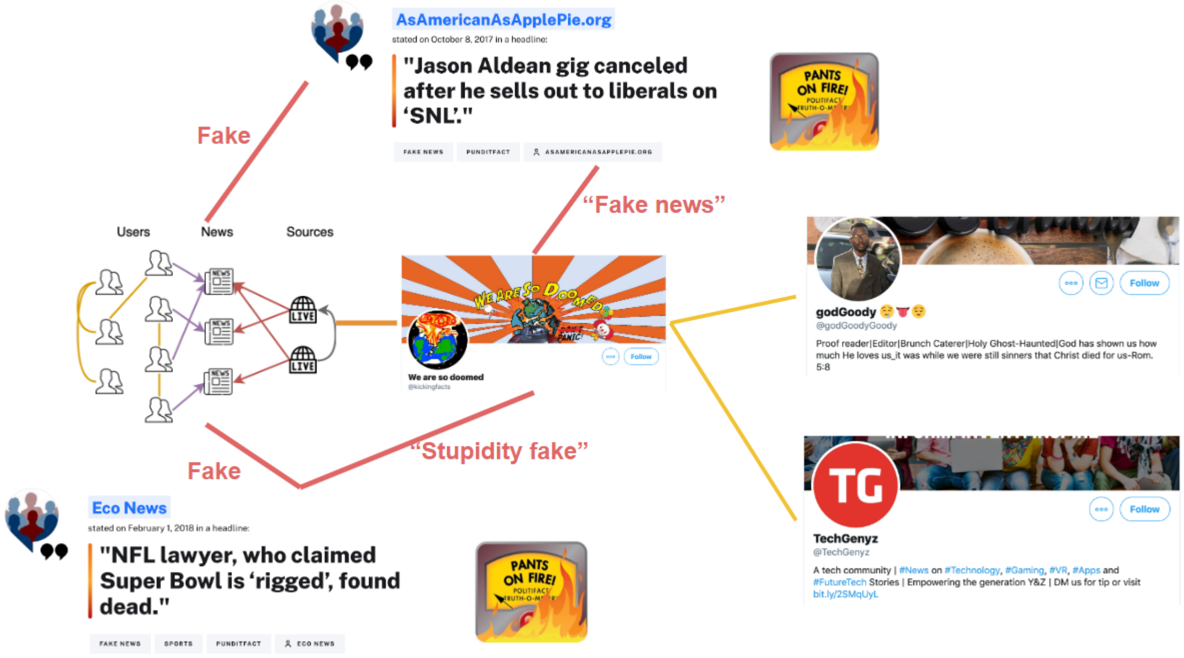
@PolitiFactTexas

The nonpartisan politics fact-checking project of the Austin American-Statesman, Houston Chronicle, San Antonio Express-News and PolitiFact.com.

Figure 4.5: A test example explaining FANG's decision.

In the first example (Figure 4.5) of a news titled "Ricky Perry says he "came to the Republican Party sooner in age" than Ronald Reagan.", FANG pays most attention to a tweet by

"Politifact Texas". This can be explained by the Twitter profile's description of a fact-checking organization, indicating a high reliability. On the other hand, another denying tweet from "Chris Powers" is not paid as much attention due to the insignificant description of its author's profile. Our model bases its prediction of the news article being true thanks to the support stance of the fact-checker, which is indeed the correct label.



"Jason Aldean gig canceled after he sells out to liberals on 'SNL'."

Proof reader|Editor|Brunch Caterer|Holy Ghost-Haunted|God has shown us how much He loves us_it was while we were still sinners that Christ died for us-Rom. 5:8

"NFL lawyer, who claimed Super Bowl is 'rigged', found dead."

A tech community | #News on #Technology, #Gaming, #VR, #Apps and #FutureTech Stories | Empowering the generation Y&Z | DM us for tip or visit bit.ly/2SMqUyL

Figure 4.6: Another test example explaining FANG's decision.

In the second example (Figure 4.6) of a news titled "Jason Aldean gig canceled after he sells out to liberals on "SNL"", FANG pays most attention to a tweet by "We are so doomed". Although this profile does not provide any description, it has a record of correctly denying the fake news "NFL lawyer, who claimed Super Bowl is "rigged", found dead.". Furthermore, the profiles that follow it, "TechGenyz" and "godGoody", have credible description of a tech community and a proof reader. This explains why our model bases its prediction of the news being fake thanks to the reliable denial, which is indeed the correct label.

Chapter 5

Conclusion

In this work, we have addressed the importance of modeling social context of news spreading as a graph. In addition to a novel and comprehensive graph representation, we have also proposed FANG, a graph learning framework based on GraphSage that outperforms Euclidean and graph neural baselines in fake news detection. FANG demonstrates its effectiveness even when the amount of training data is limited. Our proposed recurrent attentional aggregator enhances FANG's explanability and capability in capturing distinctive temporal patterns between fake and real news. FANG is also a successful representation learner as shown in both intrinsic evaluation in embedding space and extrinsic evaluation in downstream source factuality classification task.

In future work, we would like to conduct more analysis on the representations of social users. One major assumption the ideal representations should correlate to the social news-spreading phenomena including "echo chambers". Another aspect for expansion is a finer-grain formulation for certain social interactions that can be directed, such as follower-ship and retweets. And finally, with the rapid breakthrough of mainstream Graph Neural Network, our work can always be revised with the state-of-the-art.

References

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.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). *Is the internet causing political polarization? evidence from demographics* (Technical report). National Bureau of Economic Research.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 2017, 18–42.

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.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, *24*(5), May, 2002, 603–619.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., & Zubiaga, A. (2017). SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 69–76), Vancouver, Canada, August, 2017: Association for Computational Linguistics.

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., & Smith, N. A. (2015). Transition-Based Dependency Parsing with Stack Long Short-Term Memory. *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)* (pp. 334–343), 2015.

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.

Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *NIPS*, 2017.

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.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 12, 1997, 1735–80.

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. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, , 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.

Küçük, D., & Can, F. (2020). Stance detection: A survey. *CSUR*, 2020.

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.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)* (pp. 1412–1421), 2015.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 3818–3824), 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.

Ma, X., & Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)* (pp. 1064–1074), 2016.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 31–41), San Diego, California, June, 2016: Association for Computational Linguistics.

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.

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.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). Deepwalk: Online learning of social representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 701–710), 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.

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.

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.

Sobhani, P., Inkpen, D., & Zhu, X. (2017). A dataset for multi-target stance detection. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 551–557), Valencia, Spain, April, 2017: Association for Computational Linguistics.

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

Thomas, Z. (2020). Who says fake coronavirus claims causing 'infodemic'.

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* (pp. 5998–6008), 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, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018a). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355), Brussels, Belgium, November, 2018: Association for Computational Linguistics.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018b). 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.

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.

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.