DOI:10.1145/3517214

To view the accompanying Technical Perspective, visit doi.acm.org/10.1145/3517213

FANG: Leveraging Social Context for Fake News Detection Using Graph Representation

By Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan

Abstract

We propose Factual News Graph (FANG), a novel graphical social context representation and learning framework for fake news detection. Unlike previous contextual models that have targeted performance, our focus is on representation learning. Compared to transductive models, FANG is scalable in training as it does not have to maintain the social entities involved in the propagation of other news and is efficient at inference time, without the need to reprocess the entire graph. Our experimental results show that FANG is better at capturing the social context into a high-fidelity representation, compared to recent graphical and nongraphical models. In particular, FANG yields significant improvements for the task of fake news detection and is robust in the case of limited training data. We further demonstrate that the representations learned by FANG generalize to related tasks, such as predicting the factuality of reporting of a news medium.

1. INTRODUCTION

Social media have emerged as an important source of information for many worldwide. Unfortunately, not all information they publish is true. During critical events such as political elections or pandemic outbreaks, disinformation with malicious intent,[21] commonly known as "fake news," can disturb social behavior, public fairness, and rationality. Many sites and social media have devoted efforts to identify disinformation. For example, Facebook encourages users to report noncredible posts and employs professional fact checkers to expose the news in question. Manual fact-checking is also used by fact-checking websites such as Snopes, FactCheck, PolitiFact, and Full Fact. In order to scale with the increasing amount of information, automated news verification systems consider external knowledge databases as evidence.[23] Evidence-based approaches achieve high accuracy and offer potential explainability, but they also take considerable human effort. Moreover, fact-checking approaches for textual claims based on textual evidence are not easily applicable to claims about images or videos.

Recent work has taken a different tack, by exploring the contextual features of the news-dissemination process. They observed distinctive engagement patterns when social users face fake versus factual news.[6,13] For example, the fake news as shown in Table 1 had many engagements shortly after its publication. These are mainly verbatim recirculations with negative sentiment of the original post explained by the typically appalling content of fake news. After that short time window, we see denial posts questioning the validity of the news, and the stance distribution stabilizes afterwards with virtually no support. In contrast, the real news example in Table 1 leads to moderate engagement, mainly comprised of supportive posts with neutral sentiment that stabilize quickly. Such temporal shifts in user perception serve as important signals to distinguish fake from real news.

Previous work proposed partial representations of social context with (*i*) news, sources, and users as major entities and (*ii*) stances, friendship, and publication as major interactions.[5,16,17,22] However, they did not put much emphasis on the quality of the representation, on modeling the entities and their interactions, and on minimally supervised settings.

Naturally, the social context of news dissemination can be represented as a heterogeneous network where nodes and edges represent the social entities and the interactions between them, respectively. Network representations have several advantages over some existing Euclidean-based methods[11, 18] in terms of structural modeling capability for several phenomena such as echo chambers of users or polarized networks of news media. Graphical models also allow entities to exchange information, via (*i*) homogeneous edges, that is, user–user relationship, source–source citations; (*ii*) heterogeneous edges, that is, user–news stance expression, source–news publication; as well as (*iii*) high-order proximity (such as, between users who consistently support or deny certain sources, as illustrated in Figure 1). This allows the representation of heterogeneous entities to be dependent, leveraging not only fake news detection but also related tasks such as malicious user detection and source factuality prediction. Here, we focus on improving contextual fake news detection by enhancing the representations of social entities.

Our contributions can be summarized as follows:

The original version of this paper was published in *Proceedings of the 29th ACM Intern. Conf. on Information and Knowledge Management,* Oct. 2020.

Table 1. Engagement of social media users with respect to fake and real news articles.

News title (label)	Time	# Posts	S	D	C	R	Noticeable responses
Virginia Republican Wants Schools To Check Children's Genitals Before Using Bathroom **(Fake)**	3 h	38	0.00	0.03	0.19	0.78	"DISGUSED SO TRASNPHOBIC," "FOR GODS SAKE GET REAL GOP," "You cant make this up folks"
	3h–6h	21	0.00	0.10	0.10	0.80	"Ok This cant be real," "WTF IS THIS BS," "Rediculous RT"
	6 h+	31	0.00	0.10	0.14	0.76	"Cant make this up," "how is this real," "small government," "GOP Cray Cray Occupy Democrats"
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 **(Real)**	3 h	9	0.56	0.00	0.00	0.44	"#StopGunViolence," "guns r the problem"
	3 h+	36	0.50	0.00	0.11	0.39	"Some 1.15 million people have been killed by firearms in the United States since Lennon was gunned down," "#StopGunViolence"

Column 2 shows the time since publication, and columns 4–7 show the distribution of stances (S: Support, D: Deny, C: Comment, and R: Report).

(1) We propose a novel graph representation that models all major social actors and their interactions (see Figure 1).
(2) We propose the Factual News Graph (FANG), an inductive graph learning framework that effectively captures social structure and engagement patterns, thus improving representation quality.
(3) We report significant improvement in fake news detection when using FANG, and we further show that our model is robust in the case of limited training data.
(4) We show that the representations learned by FANG generalize to related tasks such as predicting the factuality of reporting of a news medium.
(5) We demonstrate FANG's explainability thanks to the attention mechanism of its recurrent aggregator.

2. RELATED WORK

2.1. Contextual fake news detection

Previous work on contextual fake news detection can be categorized based on the approach used to represent and learn its social context.

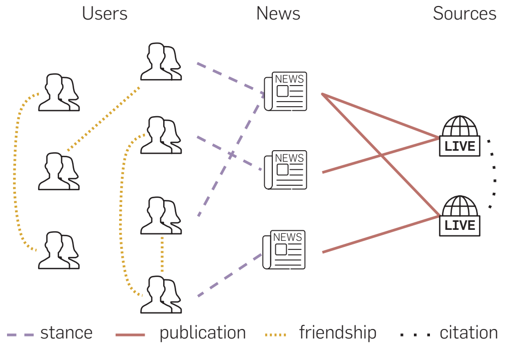
Euclidean approaches represent the social context as a flat vector or a matrix of real numbers. They typically learn a Euclidean transformation of the social entity features that best approximates the fake news prediction.[16]

However, given our formulation of social context as a heterogeneous network, Euclidean representations are less expressive. Although pioneering work used user attributes such as demographics, news preferences, and social features, for example, the number of followers and friends,[21] such work did not capture the user interaction landscape, that is, what kind of social figures they follow, which news topics they favor or oppose, and so forth. Moreover, in terms of FANG's graphical representation, node variables are no longer constrained by the independent and identically distributed assumption, and thus they can reinforce each other's representation via edge interactions.

Having acknowledged the above limitations, researchers have started exploring non-Euclidean or *geometric approaches*. In particular, they generalized the idea of using the social context when modeling a target user or the news source network and by developing representations that capture structural features about the entity.

The *Capture, Score, and Integrate* (CSI) model[18] used linear dimensionality reduction on the user cosharing adjacency matrix and combined it with news engagement features obtained from a recurrent neural network (RNN).

Figure 1. Graph representation of social context.



The *Tri-Relationship Fake News* (TriFN) detection framework[22]—although similar to our approach—neither differentiated user engagements in terms of stance and temporal patterns nor modeled source–source citations. Also, matrix decomposition approaches, such as CSI,[18] can be expensive in terms of graph node counts and ineffective for modeling high-order proximity.

Other work on citation source network,[9] propagation network,[14] and rumor detection[2] proposed models optimized solely for the objective of fake news detection, without accounting for representation quality, and therefore they are not robust to limited training data and cannot be generalized to other downstream tasks, as we show in Section 5.

2.2. Graph Neural Networks (GNNs)

GNNs have successfully generalized deep learning methods to model complex relationships and interdependencies on graphs and manifolds. Graph Convolutional Networks (GCNs) are among the first methods that effectively approximate convolutional filters.[7] However, GCNs impose a substantial memory footprint in storing the entire adjacency matrix. They are also not easily adaptable to our heterogeneous graph, where nodes and edges with different labels exhibit different information propagation patterns. Furthermore, GCNs do not guarantee generalizable representations and are transductive, requiring the inferred nodes to be present at training time. This is especially challenging for contextual fake news detection or general social network analysis, as their structure is constantly evolving.

With these considerations in mind, we build our work on GraphSage, which can generate embeddings by sampling and aggregates features from a node's local neighborhood.[4] GraphSage offers substantial flexibility in defining the information propagation pattern with parameterized random walks and recurrent aggregators. It is well-suited for representation learning with an unsupervised node proximity loss and generalizes well in minimal supervision settings. Moreover, it uses a dynamic inductive algorithm that allows the creation of unseen nodes and edges at inference time.

3. METHODOLOGY

3.1. Fake news detection using social context

Let us first define the social context graph G with its entities and interactions as shown in Figure 1:

(1) $A = \{a_1, a_2, \ldots\}$ is the list of **news articles** in question, where each a_i $(i = 1, 2, \ldots)$ is modeled as a feature vector $\boldsymbol{x}_a$.
(2) $S = \{s_1, s_2, \ldots\}$ is the list of **news sources**, where each source s_j $(j = 1, 2, \ldots)$ has published at least one article in A and is modeled as a feature vector $\boldsymbol{x}_s$.
(3) $U = \{u_1, u_2, \ldots\}$ is the list of **social users**, where each user u_k $(k = 1, 2, \ldots)$ has engaged in spreading an article in A or is connected with another user; u_k is modeled as a feature vector $\boldsymbol{x}_u$.
(4) $E = \{e_1, e_2, \ldots\}$ is the list of **interactions**, where each $e = \{v_1, v_2, t, x_e\}$ is modeled as a relation between two entities $v_1, v_2 \in A \cup S \cup U$ at time t; t is absent in time-insensitive interactions. The interaction type of e is given as a label x_e.

Table 2 summarizes the characteristics of different types of interactions, both homogeneous and heterogeneous. Stance is a special type of interaction, as it is not only characterized by edge labels and source/destination nodes but also by temporality as shown in the examples in Table 1. Recent work has highlighted the importance of incorporating temporality not only for fake news detection[18] but also for modeling online information dissemination.

We can now formally define our task as follows:

Definition 3.1. *Context-based fake news detection*: Given a social context graph $G = (A, S, U, E)$ 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 to predict whether a news article $a \in A$ is fake or real, in other words, $F_C : a \rightarrow \{0, 1\}$ such that,

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

3.2. Graph construction from social context

News articles. Textual[22] and visual[24] features have been widely used to model news article contents, by feature extraction, unsupervised semantics encoding, or learned representation. We use unsupervised textual representations as they are relatively efficient to construct and optimize. For each article $a \in A$, we construct a TF.IDF[19] vector from the text body of the article. We enrich the representation of news by weighting the pretrained embeddings from GloVe[15] of each word by its TF.IDF score, forming a semantic vector. Finally, we concatenate the TF.IDF and the semantic vector to form the news article feature vector $\boldsymbol{x}_a$.

News sources. We focus on characterizing news media sources using the textual content of their websites.[9] Similar to article representations, for each source s, we construct the source feature vector $\boldsymbol{x}_s$ as the concatenation of its TF.IDF vector and its semantic vector derived from the words in the *Homepage* and the *About Us* section, as some fake news websites openly declare their content to be satirical or sarcastic.

Social users. Online users have been studied extensively as the main propagator of fake news and rumors in social media. Shu et al.[22] conducted feature analysis of user profiles 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 U.S. constitution, free enterprise, strong military and Donald Trump #maga*" strongly indicates the user's political bias and suggests the tendency to promote certain narratives. We construct the user vector x_u as a concatenation of a TF.IDF vector and a semantic vector derived from the textual description in the user profile.

Social interactions. For each pair of social actors $(v_i, v_j) \in A \cup S \cup U$, we add an edge $e = \{v_i, v_j, t, x_e\}$ to the list of social interactions E if they are linked via interaction type x_e. Specifically, for the *followership* interaction, we examine whether user u_i follows user u_j; for the *publication* interaction, we check whether news article a_i was published by source s_j; for the *citation* interaction, we examine whether the *Homepage* of source s_i contains a hyperlink to source s_j. In the case of time-sensitive interactions, that is, *publication* and *stance*, we record their relative timestamp with respect to the article's earliest time of publication.

Stance detection. The task of characterizing the

Table 2. Interactions in FANG's social context network.

Interaction	Linking entities	Link type	Description	Temporal
Followership	User–user	Unweighted, undirected	Whether a user follows another user on social media	No
Citation	Source–source	Unweighted, undirected	Whether sources refers to another source via a hyperlink	No
Publication	Source–news	Unweighted, undirected	Whether the source published the target news	Yes
Stance	User–news	Multilabel, undirected	The stance of the user with respect to the news	Yes

viewpoint of a text with respect to another one is known as *stance detection*. In the context of fake news detection, we are interested in the stance of a user reply with respect to the title of a news article in question. We consider four stances: support with neutral sentiment or *neutral support*, support with negative sentiment or *negative support*, *deny*, and *report*.

We classify a post as verbatim reporting of the news article if it matches the article title after cleaning the text from emojis, punctuation, stop words, and URLs. We train a stance detector to classify the remaining posts as *support* or *deny* using our own dataset for stance detection between social media posts and news articles, which contains 2527 labeled source–target sentence pairs from 31 news events. For each event with a reference headline, the annotators were given a list of related headlines and posts, and they labeled whether each related headline or post supports or denies the claim made by the reference headline. Aside from the *reference headline–related headline* or the *headline–related post* sentence pairs, we further made second-order inferences for *related headline–related post* sentence pairs. If such a pair expressed a similar stance with respect to the reference headline, we inferred a *support* stance for the *related headline–related post*, and *deny* otherwise. Table 3 shows statistics about the dataset. The interannotator agreement is substantial, with a Cohen's Kappa of 0.78. We fine-tuned a RoBERTa-large transformer[10] on this data, achieving *Accuracy* of 0.8857, F_1 *score* of 0.8379, *Precision* of 0.8365, and *Recall* of 0.8395.

To further subclassify *support* posts into such with neutral and with negative sentiment, we fine-tuned a RoBERTa-large-based sentiment classifier on the Yelp ReviewPolarity dataset.[a] Altogether, the stance of a user-article engagement e is given as $stance(e)$.

3.3. Factual News Graph (FANG) framework

We now describe our FANG learning framework on the social context graph described in Section 3.2. Figure 2 shows an overview of our FANG model. Although optimizing for the fake news detection objective, FANG also learns generalizable representations for the social entities. This is achieved by optimizing three concurrent losses: (*i*) unsupervised *Proximity Loss*, (*ii*) self-supervised *Stance Loss*, and (*iii*) supervised *Fake News Detection Loss*.

Representation learning. We first discuss how FANG derives the representation of each social entity. Previous representation learning frameworks such as node2vec[3] computed a node embedding by sampling its neighborhood, as defined by the graph structure, and then optimizing for the proximity loss, similar to word2vec. These methods use the neighborhood structure only, and they are suitable when the auxiliary node features are unavailable or incomplete, that is, when optimizing for each entity's structural representation separately. Recently, GraphSage[4] was proposed to overcome this limitation by allowing auxiliary node features to be used jointly with proximity sampling as part of the representation learning.

Let $GraphSage(\cdot)$ be GraphSage's node-encoding function. Thus, we can now obtain the structural representation $\boldsymbol{z}_u \in \mathbb{R}^d$ of any user and source node as $\boldsymbol{z}_r = GraphSage(r)$, where d is the structural embedding dimension. For news nodes, we further enrich their structural representation with user engagement temporality, which we showed to be distinctive for fake news detection in Section 1. This can be formulated as learning an aggregation function $F(a, U)$ that maps a news a in question, and its engaged users U to a temporal representation $\boldsymbol{v}_a^{temp}$ that captures a's engagement pattern. Therefore, the aggregating model (that is, the aggregator) has to be time-sensitive. RNNs fulfill this requirement: specifically, the Bidirectional LSTM (Bi-LSTM) with attention can capture long-term dependencies in the information sequence in both the forward and backward directions.[12] By examining the model's attention, we learn which social profiles influence the decision, thus mimicking human analytic capability.

Our proposed LSTM input is a user–article engagement sequence $\{e_1, e_2, \ldots, e_{|U|}\}$. Let $meta(e_i) \in \mathbb{R}^l = (time(e_i), stance(e_i))$ be the concatenation of e_i's elapsed time since the news publication and a one-hot stance vector. Each engagement e_i has its representation $\boldsymbol{x}_{e_i} = (\boldsymbol{z}_{U_i}, meta\ (e_i))$, where $\boldsymbol{z}_{U_i} = GraphSage\,(U_i)$.

A Bi-LSTM encodes the engagement sequence and outputs two sequences of hidden states: (*i*) a forward one, $H^f = \boldsymbol{h}_1^f, \boldsymbol{h}_2^f, \ldots, \boldsymbol{h}_n^f$, which starts from the beginning of the engagement sequence, and (*ii*) a backward one,

Figure 2. Overview of our FANG framework.

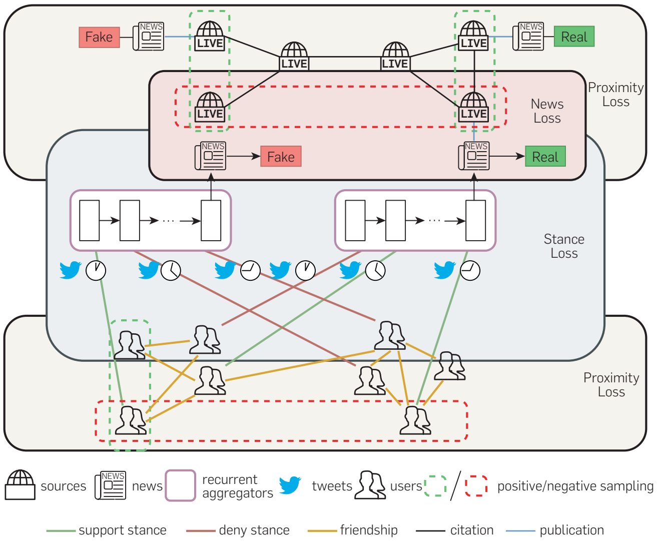


Table 3. Statistics about our stance-annotated dataset.

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

[a] https://www.kaggle.com/irustandi/yelp-review-polarity, last accessed 18 May 2021.

$H^b = \boldsymbol{h}_1^b, \boldsymbol{h}_2^b, \ldots, \boldsymbol{h}_n^b$, which starts from the end of the engagement sequence.

Let w_i be the attention weight paid by our Bi-LSTM encoder to the forward ($\boldsymbol{h}_i^f$) and to the backward ($\boldsymbol{h}_i^b$) hidden states. This attention should be derived from the similarity of the hidden state and the news features, that is, 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 follows:

$$w_i = \frac{exp(\boldsymbol{z}_a \mathbf{M}_e \boldsymbol{h}_i + meta(e_j)\mathbf{M}_m)}{\sum_{j=1}^{n} exp(\boldsymbol{z}_a \mathbf{M}_e \boldsymbol{h}_j + meta(e_j)\mathbf{M}_m)}. \tag{1}$$

where l is the meta dimension, e is the encoder dimension, and $\mathbf{M}_e \in \mathbb{R}^{d\times e}$ and $\mathbf{M}_m \in \mathbb{R}^{l\times 1}$ are the optimizable projection matrices for engagement and the meta features, which are shared across all engagements. We use w_i to compute the forward and the backward weighted feature vectors as $\boldsymbol{h}^f = \sum_i^n w_i \boldsymbol{h}_i^f$ and $\boldsymbol{h}^b = \sum_i^n w_i \boldsymbol{h}_i^b$, respectively.

Finally, we concatenate the forward and backward representation vectors to obtain the overall temporal representation $\boldsymbol{v}_a^{temp} \in \mathbb{R}^{2e}$ for article a. By explicitly setting $2e = d$, we can then combine the temporal and the structural representations as $\boldsymbol{z}_a = \boldsymbol{v}_a^{temp} + GraphSage(a)$.

Unsupervised proximity loss. We derive the *Proximity Loss* from the hypothesis that closely connected social entities often behave similarly. This is motivated by the echo chamber phenomenon, where social entities tend to interact with other entities of common interest to reinforce and to promote their narratives. This echo chamber phenomenon encompasses intercited news media sources publishing news of similar content or factuality, as well as social friends expressing similar stance with respect to news article(s) of similar content. Therefore, FANG should assign such nearby entities to a set of proximal vectors in the embedding space. From our observation that social entities are highly polarized, we also hypothesize that loosely connected social entities often behave differently. Thus, we want FANG to enforce that the representations of these disparate entities are distinctive.

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 (a) between sources and news or (b) between news, we divide the social context graph into two subgraphs, namely *news–source subgraph* and *user subgraph*. Within each subgraph G', we formulate the following *Proximity Loss* function:

$$\mathcal{L}_{prox.} = -\sum_{u\in G'} \sum_{r_p \in P_r} log\,(\sigma(\boldsymbol{z}_r^{\top} \boldsymbol{z}_{r_p})) + q \cdot \sum_{r_n \in N_r} log\,(\sigma(-\boldsymbol{z}_r^{\top} \boldsymbol{z}_{r_n})), \tag{2}$$

where $z_r \in \mathbb{R}^d$ is the representation of entity r, P_r is the set of nearby nodes or *positive set* of r, N_r is the set of disparate nodes or *negative set* of r, and q is a weighting factor. P_r is obtained using our fixed-length random walk, and N_r is derived using negative sampling.[4]

Self-supervised stance loss. We also propose an analogous hypothesis for the user–news interaction, in terms of stance. If a user expresses a stance with respect to a news article, their respective representations should be close. For each stance c, we first learn a user projection function $\alpha_c(u) = \mathbf{A}_c z_u$ and a news article projection function $\beta_c(a) = \mathbf{B}_c z_a$ that map a node representation of $\mathbb{R}^d$ to a representation in the stance space c of $\mathbb{R}^{d_c}$. Given a user u and a news article a, we compute their similarity score in the stance space c as $\alpha(u)^{\top}\beta(a)$. If u expresses stance c with respect to a, we maximize this score, and we minimize it otherwise. This is the stance classification objective, optimized using the *Stance Loss*:

$$\mathcal{L}_{stance} = -\sum_{u,\,a} \sum_{c} y_{u,\,a,\,c}\, log(f(u,\,a,\,c)), \tag{3}$$

where $f(u, a, c) = \text{softmax}(\alpha_c(u)^{\top}\beta_c(a))$ and

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

Supervised fake news loss. We directly optimize the main learning objective of fake news detection via the supervised *Fake News Loss*. In order to predict whether an article a is false, we obtain its contextual representation as the concatenation of its representation and the structural representation of its source, that is, $\boldsymbol{v}_a = (\boldsymbol{z}_a, \boldsymbol{z}_s)$.

This contextual representation is then input into a fully connected layer whose outputs are computed as $o_a = \mathbf{W}\boldsymbol{v}_a + b$, where $\boldsymbol{W} \in \mathbb{R}^{2d\times 1}$ and $b \in \mathbb{R}$ are the weights and the 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 the cross-entropy–based *Fake News Loss* $\mathcal{L}_{news}$, which we define as follows:

$$\mathcal{L}_{news} = \frac{1}{T}\sum_{a}\left\{y_a \cdot log(\sigma(o_a)) + (1 - y_a) \cdot log(1 - \sigma(o_a))\right\}, \tag{4}$$

where T is the batch size, $y_a = 0$ if a is fake, and 1 otherwise.

We define the total loss by linearly combining these three component losses: $\mathcal{L}_{total} = \mathcal{L}_{prox.} + \mathcal{L}_{stance} + \mathcal{L}_{news}$.

4. EXPERIMENTS

We conducted our experiments on a Twitter dataset collected by related work on rumor classification[8, 13] and fake news detection.[20] For each article, we collected its source, a list of engaged users, and their tweets if they were not already available in the previous dataset. This dataset also includes Twitter profile description and the list of Twitter profiles of the users that a given target user follows. We further crawled additional data about media sources, such as the content of their *Homepage* and their *About us* page, together with their frequently cited sources on their *Homepage*.

The truth value of the articles—namely, whether they are fake or real news—is based on two fact-checking websites: Snopes and PolitiFact. We release the source code of FANG

Table 4. Statistics about our dataset.

Fake	448	Publications/source	2.38	Cites/source	8.38
Real	606	Engagements/news	71.9	Friends/user	58.25
Sources	442	Neu. support/news	19.07	Deny/news	5.27
Users	54461	Neg. support/news	10.83	Report/news	36.73

and the stance detection dataset.[b] Table 4 shows some statistics about our dataset.

4.1. Fake news detection results

We benchmark the performance of FANG on fake news detection against several competitive models: (*i*) a content-only model, (*ii*) a Euclidean contextual model, and (*iii*) another graph learning model.

In order to compare our FANG model with the content-only model, we used a Support Vector Machine (SVM) model on TF.IDF feature vectors constructed from the news content (see Section 3.2). We also compared to a Euclidean model, CSI,[18] a fundamental yet effective recurrent encoder that aggregates the user features, the news content, and the user–news engagements. We reimplement the CSI model with source features by concatenating the overall score for the users and the article representation with our formulated source description to obtain the result vector for CSI's integrated module mentioned in the original paper. Lastly, we compared against the GCN graph learning framework.[7] First, we represented each of k social interactions in a separated adjacency matrix. We then concatenated GCN's output on k adjacency matrices as the final representation of each node, before passing the representation through a linear layer for classification.

We also studied the importance of modeling temporality by experimenting on two variants of CSI and FANG: (*i*) temporally insensitive CSI(-*t*) and FANG(-*t*) without *time*(*e*) in the engagement *e*'s representation $\boldsymbol{x}_e$, and (*ii*) time-sensitive CSI and FANG with *time*(*e*). Table 5 shows the macroscopic results. As an evaluation measure, we use the area under the Receiver Operating Characteristic curve (AUC ROC; hereafter, just AUC).

All context-aware models, that is, CSI(-*t*), CSI, GCN, FANG(-*t*), and FANG improve over the context-unaware baseline by 0.1153 absolute with CSI(-*t*) and by 0.1993 absolute with FANG in terms of AUC score. This shows that considering the social context is helpful for fake news detection. We further observe that both time-sensitive CSI and FANG improve over their time-insensitive variants, CSI(-*t*) and FANG(-*t*) by 0.0233 and 0.0339, respectively. These results demonstrate the importance of modeling the temporality of news spreading. Finally, the two graph-based models, FANG(-*t*) and GCN, perform consistently better than the Euclidean CSI(-*t*) by 0.0501 and 0.0386, respectively: this demonstrates the effectiveness of our social graph representation. Overall, we can conclude that our FANG model outperforms the other context-aware, temporally-aware, and graph-based models.

5. DISCUSSION

We now answer the following research questions (RQs) to better understand FANG's performance under different scenarios:

- RQ1: Does FANG work well with limited training data?
- RQ2: Does FANG differentiate between fake and real news based on their characteristic patterns in temporal engagement?
- RQ3: How effective is FANG's representation learning?

5.1. Limited training data (RQ1)

To address RQ1, we conducted the experiments described in Section 4.1 using different sizes of the training dataset. We observed consistent improvements over the baselines under both limited and sufficient data conditions. Figure 3 (left) further visualizes the experimental results. We can see that FANG consistently outperforms the two baselines for all training sizes: 10%, 30%, 50%, 70%, and 90% of the data. In terms of AUC score at decreasing training size, among the graph-based models, GCN's performance drops by 16.22% from 0.7064 at 90% to 0.5918 at 10%, whereas FANG's performance drops by 11.11% from 0.7518 at 90% to 0.6683 at 10%. We further observe that CSI's performance drops the least by only 7.93% from 0.6911 at 90% of the training data, to 0.6363 at 10% of the data. Another result from an ablated baseline, FANG(-*s*), where we removed the stance loss, highlights the importance of this self-supervised objective. At 90% of the training data, the relative underperforming margin of FANG(-*s*) compared to FANG is only 1.42% in terms of AUC score. However, this relative margin increases as the availability of training data decreases, to at most 6.39% at 30% of the training data. Overall, the experimental results emphasize our model's effectiveness even under scenarios with limited training data compared to the ablated version. This confirms a positive answer for RQ1.

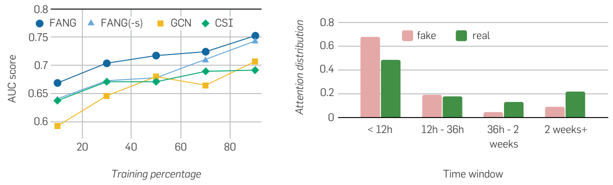
5.2. Engagement temporality study (RQ2)

To address RQ2 and to verify whether our model makes its decisions based on the distinctive temporal patterns between fake and real news, we examined FANG's attention mechanism. We accumulated the attention weights

Table 5. Comparison between FANG and baseline models on fake news detection, evaluated with AUC score.

Model	Contextual	Temporal	Graphical	AUC
Feature SVM				0.5525
CSI(-*t*) (without *time*(*e*))	✓			0.6678
CSI	✓	✓		0.6911
GCN	✓		✓	0.7064
FANG(-*t*) (without *time*(*e*))	✓		✓	0.7179
FANG	✓	✓	✓	**0.7518**

Figure 3. FANG's performance against baselines (AUC score) for varying training data sizes (left), and attention distribution across time windows for fake versus real news (right).



[b] http://github.com/nguyenvanhoang7398/FANG

produced by FANG within each time window and then compared them across time windows. Figure 3 (right) shows the attention distribution over time for fake and for real news.

We can see that, for fake news, FANG pays 68.08% of its attention to the user engagement that occurred in the first 12 h after a news article has been published. Its attention then sharply decreases to 18.83% for the next 24 h, then to 4.14% from 36 h to 2 weeks after publication, and finally to approximately 9.04% from the second week onward. However, for real news, FANG places only 48.01% of its attention on the first 12 h, which then decreases to 17.59% and to 12.85% in the time windows of 12–36 h, and 36 h to 2 weeks, respectively. We also observe that FANG maintains 21.53% attention even after 2 weeks.

Our model's characteristics are consistent with the general observation that the appalling nature of fake news generates the most engagements within a short period of time after its publication. Therefore, it is reasonable that the model places much emphasis on these crucial engagements. On the other hand, genuine news attracts fewer engagements, but it is circulated for a longer period of time, which explains FANG's persistent attention even after 2 weeks since the publication. Overall, the temporality study here highlights the transparency of our model's decision, largely thanks to the incorporated attention mechanism.

5.3. Representation learning (RQ3)

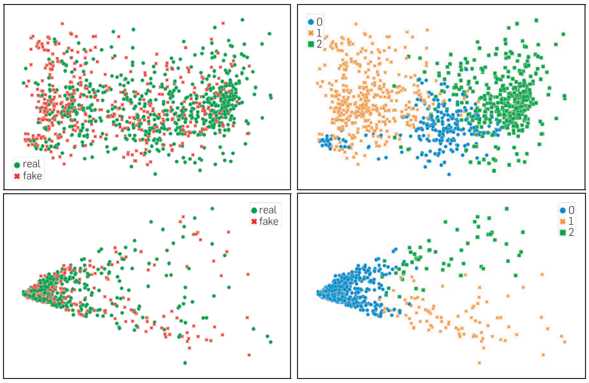
In the intrinsic evaluation, we verify how generalizable the minimally supervised news representations are for the fake news detection task. We first optimize both GCN and FANG on 30% of the training data to obtain news representations. We then cluster these representations using an unsupervised clustering algorithm, OPTICS.[1] The higher the homogeneity score, the more likely the news articles of the same factuality label (i.e., fake or real) should be close to each other, which yields higher quality representation.

In the extrinsic evaluation, we verify how generalizable the supervised source representations are for a new task: source factuality prediction. We first train FANG on 90% of the training data to obtain all source s representations as $z_s = GraphSage(s)$, and the total representation as $v_s = (z_s, x_s, \sum_{a \in publish(s)} x_a)$, where x_s, $publish(s)$, and x_a denote the source s content representation, the list of all articles published by s, and their content representations.

We propose two baseline representations that do not consider the content of the source s, $v'_s = (z_s, x_s)$. Finally, we train two separate SVM models for v_s and v'_s on the source factuality dataset, consisting of 129 sources of high factuality and 103 sources of low factuality, obtained from Media Bias/Fact Check[c] and PolitiFact.[d]

For intrinsic evaluation, the Principal Component Analysis (PCA) plot of labeled FANG representation (see Figure 4, top left) shows moderate collocation for the groups of fake and real news, whereas the PCA plot of labeled GCN representation (see Figure 4, bottom left) shows little collocation within either the fake or the real news groups. Quantitatively, FANG's OPTICS clusters (as shown in Figure 4, top right) achieve a homogeneity score of 0.051 based on news factuality labels, compared to a homogeneity score of 0.0006 for the GCN OPTICS clusters. This intrinsic evaluation demonstrates FANG's strong representation closeness within both the fake and the real news groups, indicating that FANG yields improved representations over another fully supervised graph neural framework.

Figure 4. 2D PCA plot of FANG's representations with factuality labels (top left) and OPTICS clustering labels (top right), and GCN's news representations with factuality labels (bottom left) and OPTICS clustering labels (bottom right).



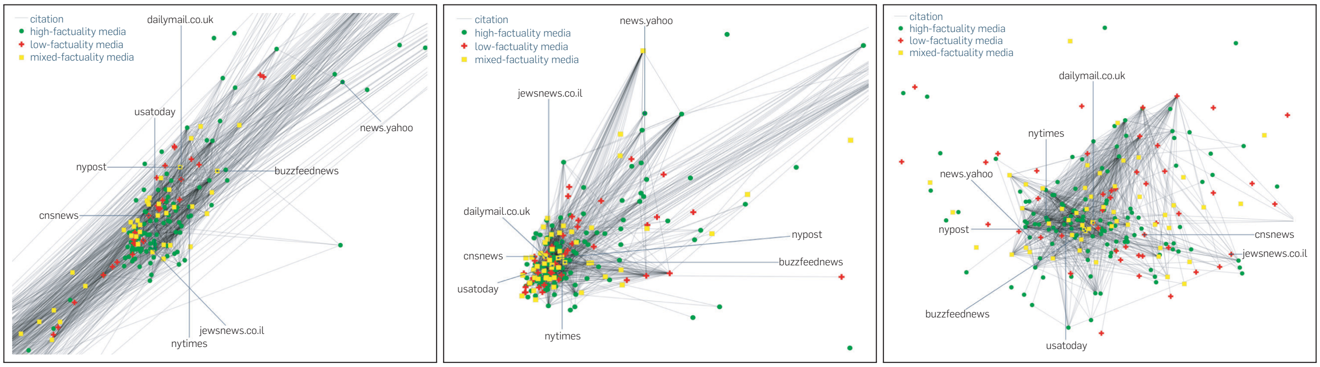
For the extrinsic evaluation on downstream source factuality classification, our context-aware model achieves an AUC score of 0.8049 (versus 0.5842 for the baseline). We further examined the FANG representations for sources to explain this 0.2207 absolute improvement. Figure 5 shows the source representations obtained from the textual features, GCN, and FANG with their factuality labels, that is, *high*, *mixed*, *low*, and the *citation* relationship. In the left subfigure, we can observe that the textual features are insufficient to differentiate the factuality of media, as a fake news site such as *cnsnews* could mimic factual media in terms of web design and news content.

However, the citation between a low-factuality website and high-factuality sites would not be as high, and it is effectively used by the two graph learning frameworks: GCN and (especially) FANG. Yet, GCN fails to differentiate low-factuality sites with higher citations, such as *jewsnews.co.il* and *cnsnews*, from high-factuality sites. On the other hand, sources such as *news.yahoo* despite being textually different, as shown in Figure 5 (left), should still cluster with other credible media for their high intercitation frequency. FANG, with much more emphasis on contextual representation learning, makes these sources more distinguishable. Its representation space gives us a glance into the landscape of news media, where there is a large central cluster of high-factuality intercited sources such as *nytimes*, *washingtonpost*, and *news.yahoo*. At the periphery lie less connected media outlets, inclusive of

[c] http://www.mediabiasfactcheck.com

[d] http://politifact.com

Figure 5. Plots for source representations using textual features (left), GCN (middle), and FANG (right) with factuality labels.



both high- and low-factuality ones.

We also see cases where all models failed to differentiate mixed-factuality media, such as *buzzfeednews* and *nypost*, which have high citation counts with high-factuality media. Overall, the results from intrinsic and extrinsic evaluation, as well as the observations, confirm RQ3 on the improvement of FANG's representation learning.

5.4. Scalable inductiveness

FANG overcomes the transductive limitation of previous approaches, although inferring the credibility of unseen nodes. MVDAM[9] has to randomly initialize an embedding and to optimize it iteratively using node2vec[3] for any unseen node, whereas FANG directly infers the embedding with its learned feature aggregator.

Other graphical approaches using matrix factorization[22] or graph convolutional layers[2, 14] learn parameters whose dimensionality is fixed to the network size N and can be as expensive as $O(N^3)^2$ in terms of inference time. FANG infers the embeddings of unseen nodes without the adjacency matrix, and its inference time only depends on the neighborhood size of the unseen nodes.

5.5. Limitations

We note that the entity and the interaction features are constructed before passing to FANG, and thus errors from upstream tasks, such as textual encoding or stance detection, propagate to FANG. Future work can address this in an end-to-end framework, where textual encoding and stance detection can be jointly optimized.

Another limitation is that the dataset for contextual fake news detection can quickly become obsolete as hyperlinks and social media traces at the time of publication might no longer be retrievable.

6. CONCLUSION AND FUTURE WORK

We have demonstrated the importance of modeling the social context for the task of fake news detection. We further proposed FANG, a graph learning framework that enhances representation quality by capturing the rich social interactions between users, articles, and media, thereby improving both fake news detection and source factuality prediction. We have demonstrated the efficiency of FANG with limited training data and its capability of capturing distinctive temporal patterns between fake and real news with a highly explainable attention mechanism. In future work, we plan more analysis of the representations of social users. We further plan to apply multitask learning to jointly address the tasks of fake news detection, source factuality prediction, and echo chamber discovery. C

References

1. Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J. OPTICS: Ordering Points to Identify the Clustering Structure. *ACM SIGMOD Rec. 28*, 2 (1999), 49–60.
2. Dong, M., Zheng, B., Hung, N.Q.V., Su, H., Li, G. Multiple rumor source detection with graph convolutional networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 2019, 569–578. DOI: https://doi.org/10.1145/3357384.3357994.
3. Grover, A., Leskovec, J. Node2vec: scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 2016, 855–864. DOI: https://doi.org/10.1145/2939672.2939754.
4. Hamilton, W.L., Ying, R., Leskovec, J. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 2017, 1025–1035.
5. Jin, Z., Cao, J., Jiang, Y.-G., Zhang, Y. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining, ICDM 2014* (Shenzhen, China, December 14-17, 2014). R. Kumar, H. Toivonen, J. Pei, J.Z. Huang, X. Wu, eds. 2014, 230–239. DOI: 10.1109/ICDM.2014.91.
6. Jin, Z., Cao, J., Zhang, Y., Luo, J. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, Phoenix, Arizona, 2016, 2972–2978.
7. Kipf T.N., Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017* (Toulon, France, April 24–26, 2017). Conference Track Proceedings. Opgehaal van, 2017. https://openreview.net/forum?id=SJU4ayYgl
8. Kochkina, E., Liakata, M., Zubiaga, A. PHEME dataset for Rumour Detection and Veracity Classification (Version 1). figshare. 2018. DOI: https://doi.org/10.6084/m9.figshare.6392078.v1
9. Kulkarni, V., Ye, J., Skiena, S., Wang, W.Y. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, 3518–3527. DOI: 10.18653/v1/D18-1388.
10. Liu, Y., Ott, M., Goyal, N., Jingfei D., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 (2019).
11. Liu, Y., Wu, Y.-F.B. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Article 44, 2018, 354–361.
12. Luong, T., Pham, H., Manning, C.D. Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural

Language Processing, 2015, 1412–1421. DOI: 10.18653/v1/D15-1166.

13. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.-F., Cha, M. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2016, 3818–3824.
14. Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M.M. Fake news Detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673 (2019).
15. Pennington, J., Socher, R., Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1532–1543. DOI: 10.3115/v1/D14-1162.
16. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 2016, 2173–2178. DOI: https://doi.org/10.1145/2983323.2983661.
17. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G. Where the truth lies: explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, 1003–1012. DOI: https://doi.org/10.1145/3041021.3055133.
18. Ruchansky, N., Seo, S., Liu, Y. CSI: a hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 2017, 797–806. DOI: https://doi.org/10.1145/3132847.3132877.
19. Salton, G., McGill, M.J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA, 1986.
20. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H. FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. *Big Data 8*, 3 (2020), 171–188.
21. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsletter 19*, 1 (2017), 22–36.
22. Shu, K., Wang, S., Liu, H. Beyond news contents: the role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 2019, 312–320. DOI: https://doi.org/10.1145/3289600.3290994.
23. Thorne, J., Vlachos, A. Automated fact checking: task formulations, methods and future directions. In E.M. Bender, L. Derczynski, P. Isabelle, eds. *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018* (Santa Fe, New Mexico, USA, August 20-26, 2018). 2018, 3346–3359. Opgehaal van. https://aclanthology.org/C18-1283/
24. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2018, 849–857. DOI: https://doi.org/10.1145/3219819.3219903.

Van-Hoang Nguyen ([vhnguyen]@u.nus.edu), National University of Singapore, Singapore.

Kazunari Sugiyama ([kaz.sugiyama]@i.kyoto-u.ac.jp), Kyoto University, Kyoto, Japan.

Preslav Nakov ([pnakov]@hbku.edu.qa), Qatar Computing Research Institute, HBKU, Doha, Qatar.

Min-Yen Kan ([kanmy]@comp.nus.edu.sg), National University of Singapore, Singapore.