Exploiting Tri-Relationship for Fake News Detection

Article · December 2017

CITATIONS
105

READS
6,652

3 authors:

Kai Shu
Illinois Institute of Technology
130 PUBLICATIONS **5,543** CITATIONS
SEE PROFILE

Suhang Wang
Pennsylvania State University
172 PUBLICATIONS **9,031** CITATIONS
SEE PROFILE

Huan Liu
Arizona State University
831 PUBLICATIONS **61,039** CITATIONS
SEE PROFILE

Exploiting Tri-Relationship for Fake News Detection

Kai Shu, Suhang Wang and Huan Liu
Computer Science and Engineering, Arizona State University, Tempe, 85281, USA
{kai.shu, suhang.wang, huan.liu}@asu.edu

Abstract

Social media for news consumption is becoming popular nowadays. The low cost, easy access and rapid information dissemination of social media bring benefits for people to seek out news timely. However, it also causes the widespread of fake news, i.e., low-quality news pieces that are intentionally fabricated. The fake news brings about several negative effects on individual consumers, news ecosystem, and even society trust. Previous fake news detection methods mainly focus on news contents for deception classification or claim fact-checking. Recent Social and Psychology studies show potential importance to utilize social media data: 1) Confirmation bias effect reveals that consumers prefer to believe information that confirms their existing stances; 2) Echo chamber effect suggests that people tend to follow likeminded users and form segregated communities on social media. Even though users social engagements towards news on social media provide abundant auxiliary information for better detecting fake news, but existing work exploiting social engagements is rather limited. In this paper, we explore the correlations of publisher bias, news stance, and relevant user engagements simultaneously, and propose a Tri-Relationship Fake News detection framework (TriFN). We also provide two comprehensive real-world fake news datasets to facilitate fake news research. Experiments on these datasets demonstrate the effectiveness of the proposed approach.

Introduction

People nowadays tend to seek out and consume news from social media rather than traditional news organizations. For example, 62% of U.S. adults get news on social media in 2016, while in 2012, only 49 percent reported seeing news on social media[1]. However, social media for news consumption is a double-edged sword. The quality of news on social media is much lower than traditional news organizations. Large volumes of "fake news", i.e., those news articles with intentionally false information, are produced online for a variety of purposes, such as financial and political gain (Klein and Wueller 2017; Allcott and Gentzkow 2017).

Fake news can have detrimental effects on individuals and the society. First, people may be misled by fake news and accept false beliefs (Nyhan and Reifler 2010; Paul and Matthews 2016). Second, fake news could change the way people respond to true news[2]. Third, the widespread of fake news could break the trustworthiness of entire news ecosystem. Thus, it is important to detect fake news on social media. Fake news is intentionally written to mislead consumers, which makes it nontrivial to detect simply based on news content. Thus, it is necessary to explore auxiliary information to improve detection. For example, several style-based approaches try to capture the deceptive manipulators originated from the particular writing style of fake news (Rubin and Lukoianova 2015; Potthast et al. 2017). In addition, previous approaches try to aggregate users' responses from relevant social engagements to infer the veracity of original news (Castillo, Mendoza, and Poblete 2011; Gupta, Zhao, and Han 2012).

The news ecosystem on social media involves three basic entities, i.e., news publisher, news and social media users. Figure 1 gives an illustration of such ecosystem. In Figure 1, p_1, p_2 and p_3 are news publishers who publish news $a_1, \ldots, a_4$ and $u_1, \ldots, u_6$ are users who have engaged in posting these news. In addition, users with similar interests can also form social links. The tri-relationship among publisher, news, and social engagements contains additional information to help detect fake news.

First, sociallogical studies on journalism have theorized the correlation between the partisan bias of publisher and news contents veracity (Gentzkow, Shapiro, and Stone 2014; Entman 2007), where partisan means the perceived bias of the publisher in the selection of how news is reported and covered. For example, in Figure 1, for p_1 with extreme left partisan bias and p_2 with extreme right partisan bias, to support their own partisan, they have high degree to report fake news, such as a_1 and a_3; while for a mainstream publisher p_3 that has least partisan bias, she has lower degree to manipulate original news events, and is more likely to write true news a_4. Thus, exploiting publisher partisan information can bring additional benefits to predict fake news.

Second, mining user engagements on social media towards the news also help fake news detection. Different users have different credibility levels on social media, and

[1]http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

[2]https://www.nytimes.com/2016/11/28/opinion/fake-news-and-the-internet-shell-game.html?

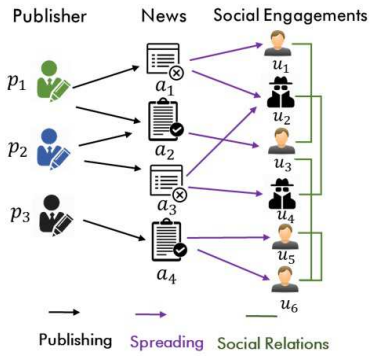


Figure 1: Tri-relationship among publishers, news pieces, and social media users for news dissemination ecosystem.

the credibility score - which means "the quality of being trustworthy" (Abbasi and Liu 2013) - has a strong indication of whether the user is more likely to engage fake news or not. Those less credible users, such as malicious accounts or normal users who are vulnerable to fake news, are more likely to spread fake news. For example, u_2 and u_4 are users with low credibility scores, and they tend to spread fake news more than other highly credible users. In addition, users tend to form relationships with like-minded people. For example, user u_5 and u_6 are friends on social media, so they tend to engage those news that confirm their own views, such as a_4.

Publisher partisan information can bridge the publisher-news relationship, while social engagements can capture the news-user relationship. In other words, they provide complementary information that has potential to improve fake news prediction. Thus, it's important to integrate these two components and model the tri-relationship simultaneously.

In this paper, we study the novel problem of exploiting tri-relationship for fake news detection. In essence, we need to address the following challenges (1) how to mathematically model the tri-relationship to extract news feature representations; and (2) how to take the advantage of tri-relationship learning for fake news detection. In an attempt to address these challenges, we propose a novel framework *TriFN* that captures the *Tri*-relationship for *F*ake *N*ews detection. The main contributions are:

- We provide a principled way to model tri-relationship among publisher, news, and relevant user engagements simultaneously;
- We propose a novel framework *TriFN* that exploits tri-relationship for fake news prediction; and
- We evaluate the effectiveness of the proposed framework for fake news detection through extensive experiments on newly collected real-world datasets.

Problem Statement

Even though fake news has been existed for long time, there is no agreed definition. In this paper, we follow the definition of fake news that is widely used in recent research (Shu et al. 2017; Zubiaga et al. 2017; Allcott and Gentzkow 2017), which has been shown to be able to 1) provide theoretical and practical values for fake news topic; and 2) eliminate the ambiguities between fake news and related concepts.

DEFINITION 1 (FAKE NEWS) *Fake news is a news article that is intentionally and verifiably false.*

Let $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ be the set of n news articles and $\mathcal{U} = \{u_1, u_2, ..., u_m\}$ be the set of m users on social media engaging the news spreading process. We denote $\mathbf{X} \in \mathbb{R}^{n \times t}$ as the news feature matrix. Users can become friends with other users and we use $\mathbf{A} \in \{0, 1\}^{m \times m}$ to denote the user-user adjacency matrix. On social media sites, users can easily share, comment and discuss about the news pieces. This kind of *social engagements* provide auxiliary information for fake news detection. We denote the social news engagement matrix as $\mathbf{W} \in \{0, 1\}^{m \times n}$, where $\mathbf{W}_{ij} = 1$ indicate that user u_i has engaged in the spreading process of the news piece a_j; otherwise $\mathbf{W}_{ij} = 0$. It's worth mentioning that we focus on those engagements that show that users agree with the news. For example, For example, we only utilize those users that directly post the news, or repost the news without adding comments. More details will introduced in Section . We also denote $\mathcal{P} = \{p_1, p_2, ..., p_l\}$ as the set of l news publishers. In addition, we denote $\mathbf{B} \in \mathbb{R}^{l \times n}$ as the publisher-news relation matrix, and $\mathbf{B}_{kj} = 1$ means news publisher p_k publishes the news article a_j; otherwise $\mathbf{B}_{kj} = 0$. We assume that the partisan labels of some publishers are given and available. We define $\mathbf{o} \in \{-1, 0, 1\}^{l \times 1}$ as the partisan label vectors, where -1, 0, 1 represents left-, neutral-, and right-partisan bias.

Similar to previous research (Shu et al. 2017), we treat fake news detection problem as a binary classification problem. In other words, each news piece can be true or fake, and we use $\mathbf{y} = \{\mathbf{y}_1; \mathbf{y}_2; ...; \mathbf{y}_n\} \in \mathbb{R}^{n \times 1}$ to represent the labels, and $\mathbf{y}_j = 1$ means news piece a_j is fake news; $\mathbf{y}_j = -1$ means true news. With the notations given above, the problem is formally defined as,

Given news article feature matrix $\mathbf{X}$*, user adjacency matrix* $\mathbf{A}$*, user social engagement matrix* $\mathbf{W}$*, publisher-news publishing matrix* $\mathbf{B}$*, publisher partisan label vector* $\mathbf{o}$*, and partial labeled news vector* $\mathbf{y}_L$*, we aim to predict remaining unlabeled news label vector* $\mathbf{y}_U$*.*

A Tri-Relationship Embedding Framework

In this section, we propose a semi-supervised detection framework by exploiting tri-relationship. The idea of modeling tri-relationship is demonstrated in Figure 2. Specifically, we first introduce the news latent feature embedding from news content, and then show how to model user social engagements and publisher partisan separately; At last, we integrate the components to model tri-relationship and provide a semi-supervised detection framework.

A Basic Model for News Content Embedding

The inherent manipulators of fake news can be reflected in the news content. Thus, it's important to extract basic feature representation from news text. Recently, it has been

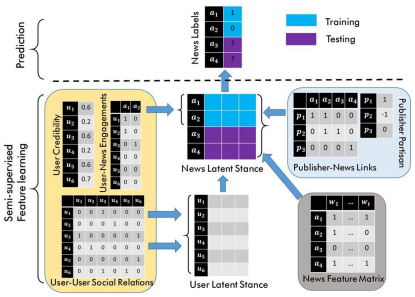


Figure 2: The tri-relationship embedding framework.

shown that nonnegative matrix factorization (NMF) algorithms are very practical and popular to learn document representations (Xu, Liu, and Gong 2003; Shahnaz et al. 2006; Pauca et al. 2004) . It tries to project the document-word matrix to a joint latent semantic factor space with low dimensionality, such that the document-word relations are modeled as inner product in the space. Specifically, giving the news-word matrix $\mathbf{X} \in \mathbb{R}_+^{n\times t}$, NMF methods try to find two nonnegative matrices $\mathbf{D} \in \mathbb{R}_+^{n\times d}$ and $\mathbf{V} \in \mathbb{R}_+^{t\times d}$ by solving the following optimization problem,

$$\min_{\mathbf{D},\mathbf{V}\geq 0} \| \mathbf{X} - \mathbf{D}\mathbf{V}^T \|_F^2 + \lambda(\|\mathbf{D}\|_F^2 + \|\mathbf{V}\|_F^2) \qquad (1)$$

where d is the dimension of the latent topic space. In addition, $\mathbf{D}$ and $\mathbf{V}$ are the nonnegative matrices indicating low-dimensional representations of news and words. Note that we denote $\mathbf{D} = [\mathbf{D}_L; \mathbf{D}_U]$, where $\mathbf{D}_L \in \mathbb{R}^{r\times d}$ is the news latent feature matrix for labeled news; while $\mathbf{D}_U \in \mathbb{R}^{(n-r)\times d}$ is the news latent feature matrix for unlabeled news. The term $\lambda(\|\mathbf{D}\|_F^2 + \|\mathbf{V}\|_F^2)$ is introduced to avoid over-fitting. With the basic model for news latent representation, next we introduce our solution to model i) the relationship between news and user social engagements, and ii) the relationship between news and publisher partisans.

News-User Social Engagements Embedding

The social engagements of users towards news articles have added value to guide the learning process of news latent features. Specifically, as shown in the yellow block in Figure 2, we explore i) user-user relations that are used to learn the basic user latent features; and ii) user-news engagement relations that encoding the correlations between user credibilities and news features guided by news veracity labels.

Basic User Feature Representation. On social media, people tend to form relationship with like-minded friends, rather than those users who have opposing preferences and interests. Thus, users that are connected are more likely to share similar latent interests towards news pieces. We use nonnegative matrix factorization method to learn the user latent representations (Tang, Aggarwal, and Liu 2016; Wang et al. 2017). Specifically, giving user-user adjacency matrix $\mathbf{A} \in \{0,1\}^{m\times m}$, we learn nonnegative matrix $\mathbf{U} \in \mathbb{R}_+^{m\times d}$ by solving the following optimization problem,

$$\min_{\mathbf{U},\mathbf{T}\geq 0} \|\mathbf{Y} \odot (\mathbf{A} - \mathbf{U}\mathbf{T}\mathbf{U}^T)\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{T}\|_F^2) \qquad (2)$$

where $\mathbf{U}$ is the user latent matrix, $\mathbf{T} \in \mathbb{R}_+^{d\times d}$ is the user-user correlation matrix, and $\mathbf{Y} \in \mathbb{R}^{m\times m}$ controls the contribution of $\mathbf{A}$. Since only positive samples are given in $\mathbf{A}$, we first set $\mathbf{Y} = sign(\mathbf{A})$ and then perform negative sampling and generate the same number of unobserved links and set weights as 0. $\lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{T}\|_F^2)$ is to avoid over-fitting.

Capturing relations of User Engagements and News The user engagements of news on social media has potential to provide rich auxiliary information to help detection fake news. However, users can express rather different and diverse opinions towards the news when spreading it, such as agree, against, neutral. In this paper, we focus on those engagements that *agree with* the news which can be directly implied in user actions. For example, we only utilize those users that directly post the news, or repost the news without adding comments. Those users that have different opinions are usually unavailable and needed to be inferred.

To model the user engagements, we consider the inherent relationship between the the credibilities of users and their posted/shared news pieces. Intuitively, we assume that users with low credibilities are more likely to spread fake news, while users with high credibilities are less likely to spread fake news. For example, low credibility users could be that 1) users that aim to spreading the diffusion scope of fake news; or 2) users that are susceptible to fake news. We adopt the existing method in (Abbasi and Liu 2013) to measure user credibility scores, which is one of the practical approaches. The basic idea in (Abbasi and Liu 2013) is that less credible users are more likely to coordinate with each other and form big clusters, while more credible users are likely to from small clusters. Thus, basically, the credibility scores are measured through the following major steps: 1) detect and cluster coordinate users based on user similarities; 2) weight each cluster based on the cluster size. Note that for our fake news detection task, we do not assume that credibility directly provided but infer the credibility score from widely available data, such as user-generated contents (Abbasi and Liu 2013).

Each user has a credibility score and we use $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_m\}$ to denote the credibility score vector, where a larger $\mathbf{c}_i \in [0,1]$ indicates that user u_i has a higher credibility. Since the latent features for low-credibility users are close to fake news latent features, while those of high-credibility users are close to true news latent features, we solve the following optimize problem,

$$\begin{aligned} \min \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{r} \mathbf{W}_{ij}\mathbf{c}_i(1 - \frac{1+\mathbf{y}_{Lj}}{2})||\mathbf{U}_i - \mathbf{D}_{L_j}||_2^2}_{\text{True news}} \\ + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{r} \mathbf{W}_{ij}(1 - \mathbf{c}_i)(\frac{1+\mathbf{y}_{Lj}}{2})||\mathbf{U}_i - \mathbf{D}_{L_j}||_2^2}_{\text{Fake news}} \end{aligned} \qquad (3)$$

where $\mathbf{y}_L \in \mathbb{R}^{r\times 1}$ is the the label vector of all partial labeled news. We consider two situations: i) for true news, i.e., $\mathbf{y}_{Lj} = -1$, we ensure that the latent features of high-credibility users are close to the true news latent features; ii) for fake news, i.e., $\mathbf{y}_{Lj} = 1$, we ensure that the latent features of low-credibility users are close to the fake news latent features. For simplicity, Eqn 3 can be rewritten as,

$$\min \sum_{i=1}^{m}\sum_{j=1}^{r} \mathbf{G}_{ij}||\mathbf{U}_i - \mathbf{D}_{L_j}||_2^2 \tag{4}$$

where $\mathbf{G}_{ij} = \mathbf{W}_{ij}(\mathbf{c}_i(1 - \frac{1+\mathbf{y}_{Lj}}{2}) + (1-\mathbf{c}_i)(\frac{1+\mathbf{y}_{Lj}}{2}))$. If we denote a new matrix $\mathbf{H} = [\mathbf{U}; \mathbf{D}_L] \in \mathbb{R}^{(m+r)\times d}$, we can also rewrite Eqn. 4 as a matrix form as below,

$$\begin{aligned} &\min \sum_{i=1}^{m}\sum_{j=1}^{r} \mathbf{G}_{ij}||\mathbf{H}_i - \mathbf{H}_j||_2^2 \\ \Leftrightarrow &\min \sum_{i,j=1}^{m+r} \mathbf{F}_{ij}||\mathbf{H}_i - \mathbf{H}_j||_2^2 \Leftrightarrow \min \ \mathrm{tr}(\mathbf{H}^T\mathbf{L}\mathbf{H}) \end{aligned} \tag{5}$$

where $\mathbf{L} = \mathbf{S} - \mathbf{F}$ is the Laplacian matrix and $\mathbf{S}$ is a diagonal matrix with diagonal element $S_{ii} = \sum_{j=1}^{m+r}\mathbf{F}_{ij}$. $\mathbf{F} \in \mathbb{R}^{(m+r)\times(m+r)}$ is computed as follows,

$$\mathbf{F}_{ij} = \begin{cases} 0, & i,j \in [1,m] \text{ or } i,j \in [m+1, m+r] \\ G_{i(j-m)}, & i \in [1,m], j \in [m+1, m+r] \\ G_{(i-m)j}, & i \in [m+1, m+r], j \in [1,m] \end{cases} \tag{6}$$

News Publisher Partisan Modeling

In real world, the partisan preference of news publisher are usually not explicitly available. We obtain the list of publishers' partisan labels from a well-known media bias fact-checking websites MBFC[3]. The partisan labels are checked with a principled methodology that ensures the reliability and objectivity of the partisan annotations. The labels are categorized as five categories: "left", "left-Center","least-biased","Right-Center" and "Right". To further ensure the accuracy of the labels, we only consider those news publishers with the annotations ["left","least-biased", "Right"], and rewrite the corresponding labels as [-1,0,1]. Thus, we can construct a partisan label vectors for news publishers as $\mathbf{o}$.

Fake news is often written to convey opinions or claims that support the partisan of the news publisher. Thus, a good news representation should be good at predicting the partisan of its publisher. This can be used to guide the learning process of news representation. As depicted in the blue block area in Figure 2, the basic idea is to utilize publisher partisan labels vector $\mathbf{o} \in \mathbb{R}^{l\times 1}$ and publisher-news matrix $\mathbf{B} \in \mathbb{R}^{l\times n}$ to optimize the news feature representation learning. Specifically, we optimization following objective,

$$\min \ \| \bar{\mathbf{B}}\mathbf{D}\mathbf{Q} - \mathbf{o}\|_2^2 + \lambda\|\mathbf{Q}\|_2^2 \tag{7}$$

where we assume that the latent feature of news publisher can be represented by the features of all the news it published, i.e., $\bar{\mathbf{B}}\mathbf{D}$. $\bar{\mathbf{B}}$ is the normalized user-news publishing relation matrix, i.e., $\bar{\mathbf{B}}_{kj} = \frac{\mathbf{B}_{kj}}{\sum_{j=1}^{n}\mathbf{B}_{kj}}$. $\mathbf{Q} \in \mathbb{R}^{d\times 1}$ is the weighting matrix that maps news publishers' latent features to corresponding partisan label vector $\mathbf{o}$. Note that we only consider those news publishers that have been fact-checked and have partisan labels in this regularization term.

[3]https://mediabiasfactcheck.com/

Proposed Framework - TriFN

We have introduced how we can learn news latent features by modeling different aspects of the tri-relationship. We further employ a semi-supervised linear classifier term to further guide the learning process of news latent features as,

$$\min \ \| \mathbf{D}_L\mathbf{P} - \mathbf{y}_L\|_2^2 + \lambda\|\mathbf{P}\|_2^2 \tag{8}$$

where $\mathbf{P} \in \mathbb{R}^{d\times 1}$ is the weighting matrix that maps news latent features to fake news labels. With all previous components, TriFN solve the following optimization problem,

$$\begin{aligned} \min_{\theta} \ & \underbrace{\|\mathbf{X} - \mathbf{D}\mathbf{V}^T\|_F^2}_{\text{News Feature Learning}} + \underbrace{\alpha\|\mathbf{Y}\odot(\mathbf{A} - \mathbf{U}\mathbf{T}\mathbf{U}^T)\|_F^2}_{\text{User-User Relation Modeling}} \\ & + \underbrace{\beta\mathrm{tr}(\mathbf{H}^T\mathbf{L}\mathbf{H})}_{\text{User-News Engagement Modeling}} \\ & + \underbrace{\gamma\|\bar{\mathbf{B}}\mathbf{D}\mathbf{Q} - \mathbf{o}\|_2^2}_{\text{News Publisher Partisan Modeling}} + \underbrace{\eta\|\mathbf{D}_L\mathbf{P} - \mathbf{y}_L\|_2^2}_{\text{Fake News Prediction}} \\ & + \lambda(\|\mathbf{D}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{U}\|_F^2 + \|\mathbf{T}\|_F^2 + \|\mathbf{P}\|_2^2 + \|\mathbf{Q}\|_2^2) \\ & s.t. \ \mathbf{D}, \mathbf{U}, \mathbf{V}, \mathbf{T} \geq 0 \end{aligned} \tag{9}$$

where the first term models the basic news latent features from news contents; and the second and third terms incorporate the user social engagement relationship; and the fourth term models publisher-news relationship. The last term incorporate semi-supervised classifier for news prediction. Therefore, this model provides a principled way to model tri-relationship for fake news prediction.

An Optimization Algorithm

In this section, we present the detail optimization process for the prosed framework TriFN. Note that if we update the variables jointly, the objective function in Eq. 9 is not convex. Thus, we propose to use alternating least squares to update the variables separately. For simplicity, we user $\mathcal{L}$ to denote the objective function in Eq. 9. Next, we introduce the updating rules for each variable in details.

Update $\mathbf{D}$ Let Ψ_D be the Lagrange multiplier for constraint $\mathbf{D} \geq 0$, the Lagrange function related to $\mathbf{D}$ is,

$$\begin{aligned} \min_{\mathbf{D}} & \|\mathbf{X} - \mathbf{D}\mathbf{V}^T\|_F^2 + \beta\mathrm{tr}(\mathbf{H}^T\mathbf{L}\mathbf{H}) + \gamma\|\bar{\mathbf{B}}\mathbf{D}\mathbf{Q} - \mathbf{o}\|_2^2 \\ & + \eta\|\mathbf{D}_L\mathbf{P} - \mathbf{y}_L\|_2^2 + \lambda\|\mathbf{D}\|_F^2 - \mathrm{tr}(\Psi_D\mathbf{D}^T) \end{aligned} \tag{10}$$

Note that $\mathbf{D} = [\mathbf{D}_L; \mathbf{D}_U]$ and $\mathbf{H} = [\mathbf{U}; \mathbf{D}_L]$. We rewrite $\mathbf{L} = [\mathbf{L}_{11}, \mathbf{L}_{12}; \mathbf{L}_{21}, \mathbf{L}_{22}]$. $\mathbf{X} = [\mathbf{X}_L, \mathbf{X}_U]$ and take the par-

tial derivative of $\mathcal{L}$ with respect to $\mathbf{D}_L$ and $\mathbf{D}_U$ separately,

$$\begin{aligned}\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \mathbf{D}_L} &= (\mathbf{D}_L\mathbf{V}^T - \mathbf{X}_L)\mathbf{V} + \beta\mathbf{L}_{21}\mathbf{U} + \beta\mathbf{L}_{22}\mathbf{D}_L \\ &+\gamma\bar{\mathbf{B}}_L^T(\bar{\mathbf{B}}_L\mathbf{D}_L\mathbf{Q} - \mathbf{o})\mathbf{Q}^T + \eta(\mathbf{D}_L\mathbf{P} - \mathbf{y}_L)\mathbf{P}^T \\ &+\lambda\mathbf{D}_L - \Psi_D \end{aligned} \tag{11}$$

$$\begin{aligned}\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \mathbf{D}_U} &= (\mathbf{D}_U\mathbf{V}^T - \mathbf{X}_U)\mathbf{V} + \lambda\mathbf{D}_U \\ &+ \gamma\bar{\mathbf{B}}_U^T(\bar{\mathbf{B}}_U\mathbf{D}_U\mathbf{Q} - \mathbf{o})\mathbf{Q}^T - \Psi_D \end{aligned} \tag{12}$$

Thus, the updating derivatives of $\mathcal{L}$ w.r.t. $\mathbf{D}$ is,

$$\begin{aligned}\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \mathbf{D}} &= (\mathbf{D}\mathbf{V}^T - \mathbf{X})\mathbf{V} + \lambda\mathbf{D} + \gamma\bar{\mathbf{B}}^T(\bar{\mathbf{B}}\mathbf{D}\mathbf{Q} - \mathbf{o})\mathbf{Q}^T \\ &+\big[\beta\mathbf{L}_{21}\mathbf{U} + \beta\mathbf{L}_{22}\mathbf{D}_L + \eta(\mathbf{D}_L\mathbf{P} - \mathbf{y}_L)\mathbf{P}^T; \mathbf{0}\big] - \Psi_D \end{aligned} \tag{13}$$

Due to KKT conditions (Boyd and Vandenberghe 2004) $\Psi_D(i,j)\mathbf{D}_{ij} = 0$, we set $\frac{\partial \mathcal{L}}{\partial \mathbf{D}} = 0$ and have,

$$\mathbf{D}_{ij} \leftarrow \mathbf{D}_{ij}\sqrt{\frac{\hat{\mathbf{D}}(i,j)}{\tilde{\mathbf{D}}(i,j)}} \tag{14}$$

where $\hat{\mathbf{D}}$ and $\tilde{\mathbf{D}}$ are defined as follows,

$$\begin{aligned}\hat{\mathbf{D}} &= \mathbf{X}\mathbf{V} + \gamma\big(\bar{\mathbf{B}}^T\mathbf{o}\mathbf{Q}^T\big)^+ + \gamma\big(\bar{\mathbf{B}}^T\bar{\mathbf{B}}\mathbf{D}\mathbf{Q}\mathbf{Q}^T\big)^- \\ &+ \big[\eta\big(\mathbf{D}_L\mathbf{P}\mathbf{P}^T\big)^- + \eta\big(\mathbf{y}_L\mathbf{P}^T\big)^+ + \beta(\mathbf{L}_{21}\mathbf{U})^- \\ &+ \beta(\mathbf{L}_{22}\mathbf{D}_L)^-; \mathbf{0}\big] \\ \tilde{\mathbf{D}} &= \mathbf{D}\mathbf{V}^T\mathbf{V} + \lambda\mathbf{D} + \gamma\big(\bar{\mathbf{B}}^T\bar{\mathbf{B}}\mathbf{D}\mathbf{Q}\mathbf{Q}^T\big)^+ + \gamma\big(\bar{\mathbf{B}}^T\mathbf{o}\mathbf{Q}^T\big)^- \\ &+ \big[\beta(\mathbf{L}_{21}\mathbf{U})^+ + \beta(\mathbf{L}_{22}\mathbf{D}_L)^+ + \eta\big(\mathbf{D}_L\mathbf{P}\mathbf{P}^T\big)^+ \\ &+ \eta\big(\mathbf{y}_L\mathbf{P}^T\big)^-; \mathbf{0}\big] \end{aligned} \tag{15}$$

where for any matrix $\mathbf{X}$, $(\mathbf{X})^+$ and $(\mathbf{X})^-$ denote the positive and negative parts of $\mathbf{X}$, respectively. Specifically, we have $(\mathbf{X})^+ = \frac{ABS(\mathbf{X})+\mathbf{X}}{2}$ and $(\mathbf{X})^- = \frac{ABS(\mathbf{X})-\mathbf{X}}{2}$, $ABS(\mathbf{X})$ is the matrix with the absolute value of elements in $\mathbf{X}$.

Update U, V and T The partial derivative of the Lagrange objective function w.r.t. $\mathbf{U}$ is as follows,

$$\begin{aligned}\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \mathbf{U}} &= \alpha(\mathbf{Y} \odot (\mathbf{U}\mathbf{T}\mathbf{U}^T - \mathbf{A}))\mathbf{U}\mathbf{T}^T \\ &+ \alpha(\mathbf{Y} \odot (\mathbf{U}\mathbf{T}\mathbf{U}^T - \mathbf{A}))^T\mathbf{U}\mathbf{T} \\ &+ \lambda\mathbf{U} - \Psi_U + \beta(\mathbf{L}_{11}\mathbf{U} + \mathbf{L}_{12}\mathbf{D}_L) \end{aligned} \tag{16}$$

So the updating rule is as follows,

$$\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij}\sqrt{\frac{[\hat{\mathbf{U}}](i,j)}{[\tilde{\mathbf{U}}](i,j)}} \tag{17}$$

where $\hat{\mathbf{U}}$ and $\tilde{\mathbf{U}}$ are defined as follows,

$$\begin{aligned}\hat{\mathbf{U}} &= \alpha(\mathbf{Y} \odot \mathbf{A})\mathbf{U}\mathbf{T}^T + \alpha(\mathbf{Y} \odot \mathbf{A})^T\mathbf{U}\mathbf{T} \\ &+ \beta(\mathbf{L}_{11}\mathbf{U})^- + \beta(\mathbf{L}_{12}\mathbf{D}_L)^- \\ \tilde{\mathbf{U}} &= \alpha(\mathbf{Y} \odot \mathbf{U}\mathbf{T}\mathbf{U}^T)\mathbf{U}\mathbf{T}^T + \alpha(\mathbf{Y} \odot \mathbf{U}\mathbf{T}\mathbf{U}^T)^T\mathbf{U}\mathbf{T} + \lambda\mathbf{U} \\ &+ \beta(\mathbf{L}_{11}\mathbf{U})^+ + \beta(\mathbf{L}_{12}\mathbf{D}_L)^+ \end{aligned} \tag{18}$$

Algorithm 1 The optimization process of TriFN framework

Require: $\mathbf{X}, \mathbf{A}, \mathbf{B}, \mathbf{W}, \mathbf{Y}, \mathbf{o}, \mathbf{y}_L, \alpha, \beta, \gamma, \lambda, \eta$
Ensure: $\mathbf{y}_U$
1: Randomly initialize $\mathbf{U}, \mathbf{V}, \mathbf{T}, \mathbf{D}, \mathbf{P}, \mathbf{Q}$
2: Precompute Laplacian matrix $\mathbf{L}$
3: **repeat**
4: Update $\mathbf{D}$ with Eqn 14
5: Update $\mathbf{U}$ with Eqn 18
6: Update $\mathbf{V}$ with Eqn 20
7: Update $\mathbf{T}$ with Eqn 22
8: Update $\mathbf{P}$,$\mathbf{Q}$ with Eqn 23
9: **until** convergence
10: Calculate $\mathbf{y}_U = \text{Sign}(\mathbf{D}_U\mathbf{P})$

The partial derivatives of the Lagrange objective w.r.t $\mathbf{V}$ is,

$$\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = (\mathbf{D}\mathbf{V}^T - \mathbf{X})^T\mathbf{D} + \lambda\mathbf{V} - \Psi_V \tag{19}$$

So the updating rule is as follows,

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij}\sqrt{\frac{[\mathbf{X}^T\mathbf{D}](i,j)}{[\mathbf{V}\mathbf{D}^T\mathbf{D} + \lambda\mathbf{V}](i,j)}} \tag{20}$$

The partial derivative of the Lagrange objective w.r.t $\mathbf{T}$ is,

$$\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \mathbf{T}} = \alpha\mathbf{U}^T(\mathbf{Y} \odot (\mathbf{U}\mathbf{T}\mathbf{U}^T - \mathbf{A}))\mathbf{U} + \lambda\mathbf{T} - \Psi_T \tag{21}$$

So the updating rule is as follows,

$$\mathbf{T}_{ij} \leftarrow \mathbf{T}_{ij}\sqrt{\frac{[\alpha\mathbf{U}^T(\mathbf{Y} \odot \mathbf{A})\mathbf{U}](i,j)}{[\alpha\mathbf{U}^T(\mathbf{Y} \odot \mathbf{U}\mathbf{T}\mathbf{U}^T)\mathbf{U} + \lambda\mathbf{T}](i,j)}} \tag{22}$$

Update P and Q Optimization w.r.t $\mathbf{P}$ and $\mathbf{Q}$ are essentially least square problem. By setting $\frac{\partial \mathcal{L}}{\partial \mathbf{P}} = 0$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{Q}} = 0$, the closed from solution of $\mathbf{P}$ and $\mathbf{Q}$ as follow,

$$\begin{aligned}\mathbf{P} &= (\eta\mathbf{D}_L^T\mathbf{D}_L + \lambda\mathbf{I})^{-1}\eta\mathbf{D}_L^T\mathbf{y}_L \\ \mathbf{Q} &= (\gamma\mathbf{D}^T\bar{\mathbf{B}}^T\bar{\mathbf{B}}\mathbf{D} + \lambda\mathbf{I})^{-1}\gamma\mathbf{D}^T\bar{\mathbf{B}}^T\mathbf{o} \end{aligned} \tag{23}$$

Optimization Algorithm of TriFN

In this section, we present the details to optimize TriFN in Algorithm 1. We first randomly initialize $\mathbf{U}, \mathbf{V}, \mathbf{T}, \mathbf{D}, \mathbf{P}, \mathbf{Q}$ in line 1, and construct the Laplacian matrix $\mathbf{L}$ in line 2. Then we repeatedly update related parameters through Line 4 to Line 7 until convergence. Finally, we predict the labels of unlabeled news $\mathbf{y}_U$ in line 10. The convergence of Algorithm 1 is guaranteed because the objective function is nonnegative and in each iteration it will monotonically decrease the objective value, and finally it will converge to an optimal point (Lee and Seung 2001).

The main computation cost comes from the fine-tuning variables for Algorithm 1. In each iteration, the time complexity for computing $\mathbf{D}$ is $\mathcal{O}(nd + nld^2 + rd + rm + n^2)$. Similarly, the computation cost for $\mathbf{V}$ is approximately $\mathcal{O}(tnd)$, for $\mathbf{U}$ is $\mathcal{O}(m^4d^3 + md)$, for $\mathbf{T}$ is about $\mathcal{O}(m^4d^3 + m^2d^2)$. To update $\mathbf{P}$ and $\mathbf{Q}$, the costs are approximately $\mathcal{O}(d^3 + d^2 + dr)$ and $\mathcal{O}(d^2ln + d^3 + dl)$.

Experiments

In this section, we present the experiments to evaluate the effectiveness of the proposed TriFN framework. Specifically, we aim to answer the following research questions:

- Is TriFN able to improve fake news classification performance by modeling publisher partisan and user engagements simultaneously?
- How effective are publisher partisan bias modeling and user engagement learning, respectively, in improving the fake news detection performance of TriFN?
- How can proposed method can handle early fake news detection when limited user engagement are provided?

We begin by introducing the datasets and experimental settings. Then we illustrate the performance of TriFN, followed by the parameter sensitivity analysis.

Datasets

Online news can be collected from different sources, such as news agency homepages, search engines, and social media sites. However, manually determining the veracity of news is a challenging task, usually requiring annotations with domain expertise who performs careful analysis of claims and additional evidence, context, and reports from authoritative sources. There are no agreed upon benchmark datasets for the fake news detection problem. Some publicly available datasets include: BuzzFeedNews[4], LIAR (Wang 2017), BS Detector[5], CREDBANK (Mitra and Gilbert 2015). BuzzFeedNews only contains headlines and text for each news piece. LIAR includes mostly short statements, which may not be fake news because the speakers may not be news publishers. The ground truth labels for BS Detector data are generated from a software rather than fact-checking from expert journalists, so any model trained on this data is really learning the parameters of BS Detector. Finally, CREDBANK include social engagements for specific topics, without specific news pieces and publisher information.

We create two comprehensive fake news datasets[6], which both contain publishers, news contents and social engagements information. The ground truth labels are collected from journalist experts from BuzzFeed and well-recognized fact-checking website PolitiFact[7]. For BuzeeFeed news, it comprises a complete news headlines in Facebook. We further enrich the data by crawling the news contents of those Facebook web links. The related social media posts are collected from Twitter using API by searching the headlines of news. Similar to previous setting, we treat fake news as those news with original annotation as mostly false and mixture of true and false (Potthast et al. 2017). For PolitiFact, the list of fake news articles are provided and corresponding news content can be crawled as well. Similar techniques can be applied to get related social media posts for PolitiFact.

[4]https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data

[5]https://github.com/bs-detector/bs-detector

[6]The dataset will be publicly available.

[7]http://www.politifact.com/subjects/fake-news/

Table 1: The statistics of datasets

Platform	BuzzFeed	PolitiFact
# Candidate news	182	240
# True news	91	120
# Fake news	91	120
# Users	15,257	23,865
# Engagements	25,240	37,259
# Social Links	634,750	574,744
# Publisher	9	91

The publishers' partisan labels are collected from a well-known media bias fact-checking websites MBFC[8]. Note that we balance the number of fake news and true news, so that we avoid that trivial solution (e.g., classifying all news as the major class labels) to achieve high performance and for fair performance comparison. The details are shown in Table 1.

Experimental settings

To evaluate the performance of fake news detection algorithms, we use the following metrics, which are commonly used to evaluate classifiers in related areas: $Accuracy = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$, $Precision = \frac{|TP|}{|TP|+|FP|}$, $Recall = \frac{|TP|}{|TP|+|FN|}$, and $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$, where TP, FP, TN, FN represent true positive, false positive, true negative and false negative, respectively. We compare the proposed framework with several state-of-the-art fake news detection methods, described as follows.

- **RST** (Rubin, Conroy, and Chen 2015): RST extracts news style-based features by combines the vector space model and rhetorical structure theory. The SVM classifier is applied for classification.
- **LIWC** (Pennebaker et al. 2015): LIWC is widely used to extract the lexicons falling into psycholinguistic categories. It's based on a large sets of words that represent psycholinguistic processes, summary categories, and part-of-speech categories. This method can capture the deception features from a psychology perspective.
- **Castillo** (Castillo, Mendoza, and Poblete 2011): This method predicts news veracity using social engagements. The features are extracted from user profiles and friendship network. To ensure fair comparison, we also add the credibility score of users inferred in Sec as an additional feature.
- **RST+Castillo**: This method combine features from RST and Castillo, and consider both news contents and user social engagements.
- **LIWC+Castillo**: This method combine features from LIWC and Castillo, and consider both news contents and user social engagements.

Note that for fair comparison and demonstration, we choose baselines that 1) only consider **news contents**,

[8]https://mediabiasfactcheck.com/

Table 2: Performance comparison for fake news detection

Datasets	Metric	RST	LIWC	Castillo	RST+Castillo	LIWC+Castillo	TriFN
BuzzFeed	Accuracy	0.610 ± 0.023	0.655 ± 0.075	0.747 ± 0.061	0.758 ± 0.030	0.791 ± 0.036	$\mathbf{0.864 \pm 0.026}$
	Precision	0.602 ± 0.066	0.683 ± 0.065	0.735 ± 0.080	0.795 ± 0.060	0.825 ± 0.061	$\mathbf{0.849 \pm 0.040}$
	Recall	0.561 ± 0.057	0.628 ± 0.021	0.783 ± 0.048	0.784 ± 0.074	0.834 ± 0.094	$\mathbf{0.893 \pm 0.013}$
	F1	0.555 ± 0.057	0.623 ± 0.066	0.756 ± 0.051	0.789 ± 0.056	0.802 ± 0.023	$\mathbf{0.870 \pm 0.019}$
PolitiFact	Accuracy	0.571 ± 0.039	0.637 ± 0.021	0.779 ± 0.025	0.812 ± 0.026	0.821 ± 0.052	$\mathbf{0.878 \pm 0.020}$
	Precision	0.595 ± 0.032	0.621 ± 0.025	0.777 ± 0.051	0.823 ± 0.040	0.856 ± 0.071	$\mathbf{0.867 \pm 0.034}$
	Recall	0.533 ± 0.031	0.667 ± 0.091	0.791 ± 0.026	0.792 ± 0.026	0.767 ± 0.120	$\mathbf{0.893 \pm 0.023}$
	F1	0.544 ± 0.042	0.615 ± 0.044	0.783 ± 0.015	0.793 ± 0.032	0.813 ± 0.070	$\mathbf{0.880 \pm 0.017}$

such as RST, LIWC, TriFNC; 2) only consider **social engagements**, such as Castillo; and 3) consider both **news content and social engagements**, such as RST+Castillo, LIWC+Castillo. There are different variants of TriFN, i.e., TriFN\S, TriFN\P, TriFN\SP. The number in the brackets are the number of features extracted. Moreover, we apply other different learning algorithms, such as decision tree, naive Bayes. We find out that SVM generally performs the best, so we use SVM to perform prediction on all baseline methods. The results are reported with 5-fold cross validation. The details are shown in Table 3.

Table 3: Summary of the detection methods for comparison

Method	**News Content**	**Social Engagements**	**Publisher Partisan**
RST (*28*)	✓		
LIWC (*93*)	✓		
Castillo (*10*)		✓	
RST+Castillo (*38*)	✓	✓	
LIWC+Castillo (*103*)	✓	✓	
TriFN\P	✓	✓	
TriFN\S	✓		✓
TriFN\PS	✓		
TriFN	✓	✓	✓

Performance Comparison

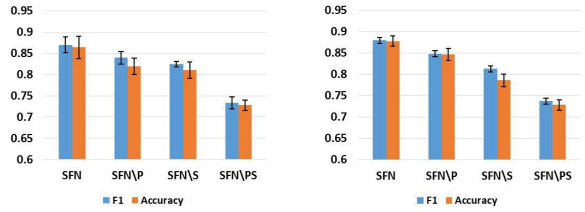
In this subsection, we evaluate the effectiveness of the proposed framework TriFN in terms of fake news classification. The comparison results are shown in Table 2. Note that we determine model parameters with cross-validation strategy, and we repeat the generating process of training/test set for three times and the average performance is reported. We first perform cross validation on parameters $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$, and choose those parameters that achieves best performance, i.e., $\lambda = 0.1$. We also choose latent dimension $d = 10$ for easy parameter tuning, and focus on the parameters that contribute the tri-relationship modeling components. The parameters for TriFN are set as $\{\alpha = 1e-4, \beta = 1e-5, \gamma = 1, \eta = 1\}$ for BuzzFeed and $\{\alpha = 1e-5, \beta = 1e-4, \gamma = 10, \eta = 1\}$ for PolitiFact. The parameter sensitivity analysis will be discussed in following section. Based on Table 2 and Figure 3, we make the following key observations:

- For news content based methods RST and LIWC, we can see that LIWC performs better than RST, indicating that LIWC can better capture the deceptiveness in text. The good results of LIWC demonstrate that fake news pieces are very different from real news in terms of choosing words that can reveal psychometrics characteristics.
- For methods based on news content and social engagements (i.e., RST+Castillo, LIWC+Castillo, TriFN\P), we can see TriFN\P performs the best. It indicates the effectiveness of modeling the latent news features and the correlation between user credibilities and news veracity. For example, TriFN\P achieves relative improvement of $1.70\%, 4.69\%$ on BuzzFeed, and $2.31\%, 4.35\%$ on PolitiFact, comparing with LIWC+Castillo in terms of $Accuracy$ and $F1$ score.
- Generally, methods using both news contents and social engagements perform better than those methods purely based on news contents (i.e., RST, LIWC), and those methods only based on social engagements (i.e., Castillo). This indicates that exploiting both news contents and corresponding social engagements is important.
- We can see that TriFN consistently outperforms the other two baselines that also exploit news contents and social engagements, in terms of all evaluation metrics on both datasets. For example, TriFN achieves average relative improvement of $9.23\%, 8.48\%$ on BuzzFeed and $6.94\%, 8.24\%$ on PolitiFact, comparing with LIWC+Castillo in terms of $Accuracy$ and $F1$ score. It supports the importance to model tri-relationship of publisher-news and news-user to better predict fake news.

User Engagements and Publisher Partisan Impact

In previous section, we observe that TriFN framework improves the classification results significantly. In addition to news contents, we also captures social engagements and publisher partisan. Now, we investigate the effects of these components by defining the variants of TriFN:

- TriFN\P - We eliminate the effect of publisher partisan modeling part $\gamma\|\bar{\mathbf{B}}\mathbf{D}\mathbf{Q} - \mathbf{o}\|_2^2$ by setting $\gamma = 0$.
- TriFN\S - We eliminate the effects of user social engagements components $\alpha\|\mathbf{Y} \odot (\mathbf{A} - \mathbf{U}\mathbf{U}^T)\|_F^2 + \beta\text{tr}(\mathbf{H}^T\mathbf{L}\mathbf{H})$ by setting $\alpha, \beta = 0$.
- TriFN\PS - We eliminate the effects of both publisher partisan and user social engagements, by setting $\alpha, \beta, \gamma = 0$. The model only consider news content embedding.



(a) BuzzFeed (b) PolitiFact

Figure 3: User engagements and publisher partisan impact.

The parameters in all the variants are determined with cross-validation and the performance comparison is shown in Figure 3, we have following observations:

- When we eliminate the effect of social engagements component , the performance of TriFN\S degrades in comparison with TriFN. For example, the performance reduces 5.2% and 6.1% in terms of F1 and Accuracy metrics on BuzzFeed, 7.6% and 10.6% on PolitiFact. The results suggest that social engagements in TriFN is important.
- We have similar observations for TriFN\P when we eliminate the effect of publisher partisan component. The results suggest the importance to consider publisher partisan in TriFN.
- When we eliminate both components in TriFN\PS, the results are further reduced compared to TriFN\S and TriFN\P. It also suggests that these components are complementary to each other.

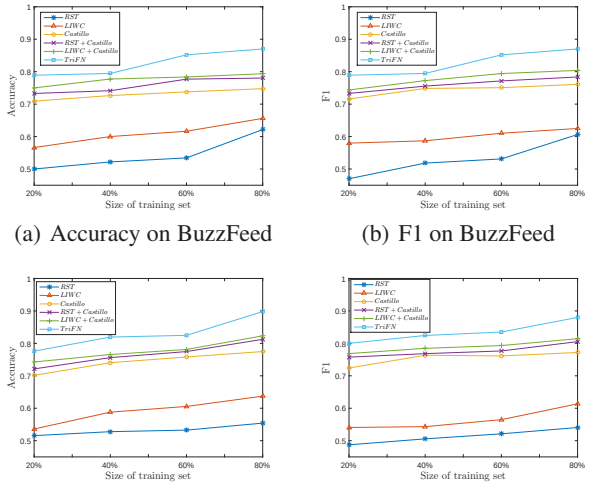
Through the component analysis of TriFN, we conclude that (i) both components can contribute to the performance improvement of TriFN; (ii) it's necessary to model both news contents and social engagements because they contain complementary information.

Impact of Training Data Size

We further investigate whether larger amounts of training data can improve the identification of fake news. We plot the learning curves with respect to different training data size, as shown in Figure 4. For TriFN, we fix other parameters as mentioned in last section when we change the training ratio. By plotting these learning curves, we can see that 1) Generally, the detection performance tends to increase with the increasing of training ratio for all compared methods on both datasets; 2) For different training size setting, the proposed TriFN consistently outperforms other baseline methods.

Model Parameter Analysis

The proposed TriFN has four important parameters. The first two are α and β, which control the contributions from social relationship and user-news engagements. γ controls the contribution of publisher partisan and η controls the contribution of semi-supervised classifier. We first fix $\{\alpha = 1e-4, \beta = 1e-5\}$ and $\{\alpha = 1e-5, \beta = 1e-4\}$ for BuzzFeed and PolitiFact, respectively. Then we vary η as $\{1, 10, 20, 50, 100\}$ and γ in $\{1, 10, 20, 30, 100\}$. The performance variations are depicted in Figure 5. We can see i) when η increases from 0, eliminating the impact of semi-supervised classification term, to 1, the performance

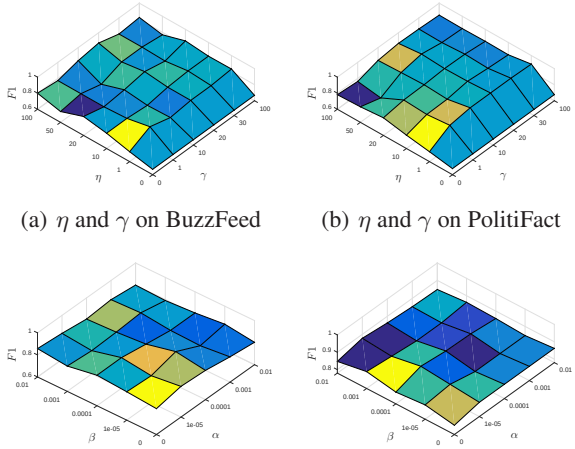


(a) Accuracy on BuzzFeed (b) F1 on BuzzFeed

(c) Accuracy on PolitiFact (d) F1 on PolitiFact

Figure 4: The learning curve on BuzzFeed and PolitiFact.

increase dramatically in both datasets. These results support the importance to combine semi-supervised classifier to feature learning; ii) generally, the increase of γ will increase the performance in a certain region, $\gamma \in [1, 50]$ and $\eta \in [1, 50]$ for both datasets, which easy the process for parameter setting. Next, we fix $\{\gamma = 1, \eta = 1\}$ and $\{\gamma = 10, \eta = 1\}$ for BuzzFeed and PolitiFact, respectively. Then we vary $\alpha, \beta \in [0, 1e-5, 1e-4, 1e-3, 0.001, 0.01]$. We can see that i) when α and β increase from 0, which eliminate the social engagements, to $1e-5$, the performance increases relatively, which again support the importance of social engagements; ii) The performance tends to increase first and then decrease, and it's relatively stable in $[1e-5, 1e-3]$.



(a) η and γ on BuzzFeed (b) η and γ on PolitiFact

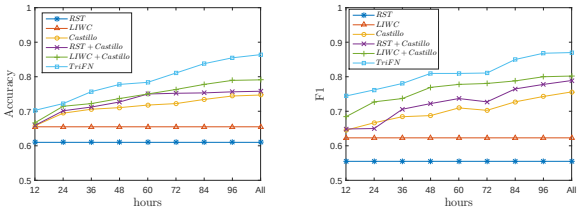
(c) α and β on BuzzFeed (d) α and β on PolitiFact

Figure 5: Model parameter analysis.

Early Fake News Detection

In real world scenario, early detection of fake news is very desirable to restrict the dissemination scope of fake news and prevent the future propagation on social media. Early fake

news detection aims to give early alert of fake news, by only considering limited social engagements within a specific range of time delay of original news posted. Specifically, we change the delay time in $[12, 24, 36, 48, 60, 72, 84, 96]$ hours. From Figure 6, we can see that: 1) generally, the detection performance is getting better when the delay time increase for those methods using social engagements information, which indicates that more social engagements on social media provide additional information for fake news detection; 2) The proposed TriFN always achieve best performance, which shows that the importance of modeling user-user relation and news-user relations to capture effective feature representations; and 3) Even in the very early stage after fake news has been published, TriFn can already achieve good performance. For example, TriFN can achieve F1 score more than 80% within 48 hours on both datasets.



(a) Accuracy on BuzzFeed (b) F1 on BuzzFeed

(c) Accuracy on PolitiFact (d) F1 on PolitiFact

Figure 6: The performance of early fake news detection on BuzzFeed and PolitiFact in terms of Accuracy and F1.

Related Work

Fake news detection methods generally focus on using *news contents* and *social contexts* (Shu et al. 2017).

For news content based approaches, features are extracted as linguistic-based and visual-based. Linguistic-based features aim to capture specific writing styles and sensational headlines that commonly occur in fake news content (Potthast et al. 2017; Afroz, Brennan, and Greenstadt 2012), such as lexical features and syntactic features. Visual-based features try to identify fake images (Gupta et al. 2013) that are intentionally created or capturing specific characteristics for images in fake news (Jin et al. 2017). News content based models include i) knowledge-based: using external sources to fact-checking claims in news content (Magdy and Wanas 2010; Wu et al. 2014), and 2) style-based: capturing the manipulators in writing style, such as deception (Feng, Banerjee, and Choi 2012; Rubin and Lukoianova 2015) and non-objectivity (Potthast et al. 2017).

For social context based approaches, the features include user-based, post-based and network-based. User-based features from user profiles to measure their characteristics and credibility (Castillo, Mendoza, and Poblete 2011; Kwon et al. 2013). Post-based features represent users' social response in term of stance (Jin et al. 2016), topics (Ma et al. 2015), or credibility (Castillo, Mendoza, and Poblete 2011). Network-based features are extracted by constructing specific networks, such as diffusion network (Kwon et al. 2013), co-occurrence network (Ruchansky, Seo, and Liu 2017), etc. Social context models basically include stance-based and propagation-based. Stance-based models utilize users' opinions towards the news to infer news veracity (Jin et al. 2016; Tacchini et al. 2017). Propagation-based models assume that the credibility of news is highly related to the credibilities of relevant social media posts, which several propagation methods can be applied (Jin et al. 2014; 2016; Gupta, Zhao, and Han 2012). It's worth mentioning that we can not directly compare the propagation-based approaches, because we assume we only have user actions, e.g., posting the news or not. In this case, the propagation signals inferred from text are the same and thus become ineffective.

In this paper, we are to our best knowledge the first to classify fake news by learning the effective news features through the tri-relationship embedding among publishers, news contents, and social engagements.

Conclusion and Future Work

Due to the inherent relationship among publisher, news and social engagements during news dissemination process on social media, we propose a novel framework TriFN to model tri-relationship for fake news detection. TriFN can extract effective features from news publisher and user engagements separately, as well as capture the interrelationship simultaneously. Experimental results on real world fake news datasets demonstrate the effectiveness of the proposed framework and importance of tri-relationship for fake news prediction. It's worth mentioning TriFN can achieve good detection performance in early stage of news dissemination.

There are several interesting future directions. First, it's worth to explore effective features for early fake news detection, as fake news usually evolves very fast on social media; Second, how to extract features to model fake news intention from psychology's perspective needs investigation. At last, how to identify low quality or even malicious users spreading fake news is important for fake news intervention.

References

Abbasi, M. A., and Liu, H. 2013. Measuring user credibility in social media. In *SBP*, 441–448. Springer.

Afroz, S.; Brennan, M.; and Greenstadt, R. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Security and Privacy (SP), 2012 IEEE Symposium on*, 461–475. IEEE.

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

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

Entman, R. M. 2007. Framing bias: Media in the distribution of power. *Journal of communication* 57(1):163–173.

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

Gentzkow, M.; Shapiro, J. M.; and Stone, D. F. 2014. Media bias in the marketplace: Theory. Technical report, National Bureau of Economic Research.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, 729–736. ACM.

Gupta, M.; Zhao, P.; and Han, J. 2012. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 153–164. SIAM.

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

Jin, Z.; Cao, J.; Zhang, Y.; and Luo, J. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI*, 2972–2978.

Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2017. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia* 19(3):598–608.

Klein, D. O., and Wueller, J. R. 2017. Fake news: A legal perspective.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Prominent features of rumor propagation in online social media. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 1103–1108. IEEE.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 556–562.

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

Magdy, A., and Wanas, N. 2010. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 103–110. ACM.

Mitra, T., and Gilbert, E. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, 258–267.

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

Pauca, V. P.; Shahnaz, F.; Berry, M. W.; and Plemmons, R. J. 2004. Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, 452–456. SIAM.

Paul, C., and Matthews, M. 2016. The russian firehose of falsehood propaganda model. *RAND Corporation*.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of liwc2015. Technical report.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Rubin, V. L., and Lukoianova, T. 2015. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology* 66(5):905–917.

Rubin, V. L.; Conroy, N.; and Chen, Y. 2015. Towards news verification: Deception detection methods for news discourse. In *Hawaii International Conference on System Sciences*.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. Csi: A hybrid deep model for fake news. *arXiv preprint arXiv:1703.06959*.

Shahnaz, F.; Berry, M. W.; Pauca, V. P.; and Plemmons, R. J. 2006. Document clustering using nonnegative matrix factorization. *Information Processing & Management* 42(2):373–386.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *KDD exploration newsletter*.

Tacchini, E.; Ballarin, G.; Della Vedova, M. L.; Moret, S.; and de Alfaro, L. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

Tang, J.; Aggarwal, C.; and Liu, H. 2016. Node classification in signed social networks. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 54–62. SIAM.

Wang, S.; Aggarwal, C.; Tang, J.; and Liu, H. 2017. Attribute signed network embedding. In *Proceedings of the 26th ACM International on Conference on Information and Knowledge Management*, 115–124. ACM.

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

Wu, Y.; Agarwal, P. K.; Li, C.; Yang, J.; and Yu, C. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment* 7(7):589–600.

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 267–273. ACM.

Zubiaga, A.; Wang, B.; Liakata, M.; and Procter, R. 2017. Stance classification of social media users in independence movements. *arXiv preprint arXiv:1702.08388*.