ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/335481117

Unsupervised Fake News Detection on Social Media: A Generative Approach

Article *in* Proceedings of the AAAI Conference on Artificial Intelligence · July 2019

DOI: 10.1609/aaai.v33i01.33015644

CITATIONS
65

READS
2,095

6 authors, including:

Kai Shu
Illinois Institute of Technology
105 PUBLICATIONS **4,104** CITATIONS
SEE PROFILE

Suhang Wang
Pennsylvania State University
150 PUBLICATIONS **6,815** CITATIONS
SEE PROFILE

Huan Liu
Arizona State University
791 PUBLICATIONS **54,492** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Feature engineering for outlier detection View project

Project Non-IID Outlier Detection View project

All content following this page was uploaded by Kai Shu on 14 September 2019.

The user has requested enhancement of the downloaded file.

Unsupervised Fake News Detection on Social Media: A Generative Approach

Shuo Yang,[†‡] Kai Shu,[‡] Suhang Wang,[§] Renjie Gu,[†] Fan Wu,[†] Huan Liu[‡]
[†]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[‡]School of Computing, Informatics and Decision Systems Engineering, Arizona State University, USA
[§]College of Information Sciences and Technology, Pennsylvania State University, USA
[†]{wnmmxy, grj165, wu-fan}@sjtu.edu.cn; [‡]{kai.shu, huan.liu}@asu.edu; [§]szw494@psu.edu

Abstract

Social media has become one of the main channels for people to access and consume news, due to the rapidness and low cost of news dissemination on it. However, such properties of social media also make it a hotbed of fake news dissemination, bringing negative impacts on both individuals and society. Therefore, detecting fake news has become a crucial problem attracting tremendous research effort. Most existing methods of fake news detection are supervised, which require an extensive amount of time and labor to build a reliably annotated dataset. In search of an alternative, in this paper, we investigate if we could detect fake news in an *unsupervised* manner. We treat truths of news and users' credibility as latent random variables, and exploit users' engagements on social media to identify their opinions towards the authenticity of news. We leverage a Bayesian network model to capture the conditional dependencies among the truths of news, the users' opinions, and the users' credibility. To solve the inference problem, we propose an efficient collapsed Gibbs sampling approach to infer the truths of news and the users' credibility without any labelled data. Experiment results on two datasets show that the proposed method significantly outperforms the compared unsupervised methods.

1 Introduction

The continuous growth of social media has provided users with more convenient ways to access news than ever before. According to Pew Research Center (Shearer and Gottfried 2017), about two-thirds of U.S. adults got news from social media in 2017. As people continue to benefit from the convenience and easy accessibility of social media, they also expose themselves to certain noisy and inaccurate information spread on social media, especially *fake news*, which consists of articles intentionally written to convey false information for a variety of purposes such as financial or political manipulation (Shu et al. 2017). For example, one of the most famous fake news is: *"Pope Francis shocks world, endorses Donald Trump for president, releases statement."* This news was extremely popular and has gained over 960,000 user engagements on Facebook[1]. The wide spread of fake news could inflict damages on social media platforms and also cause serious impacts on both individuals and society. Thus, detecting and mitigating fake news has become a crucial problem in recent social media studies.

Existing work on fake news detection is mostly based on supervised methods. They aim to build a classification model considering different sets of features including news content (Wang 2017), user profiles (Castillo, Mendoza, and Poblete 2011), message propagation (Wu and Liu 2018), and social contexts (Ma et al. 2015). Though they have shown some promising results, these supervised methods suffer from a critical limitation, *i.e.*, they require a reliably pre-annotated dataset to train a classification model. However, obtaining a large number of annotations is time-consuming and labor-intensive, as the process needs careful checking of news contents as well as other additional evidence such as authoritative reports. Leveraging a crowdsourcing approach to obtain annotations could alleviate the burden of expert checking, but the quality of annotations may suffer (Kim et al. 2018). As fake news is intentionally written to mislead readers, individual human workers alone may not have the domain expertise to differentiate real news and fake news (Bond Jr and DePaulo 2006).

In search of an alternative to supervised methods, we consider detecting fake news in an *unsupervised* manner. Our key idea is to extract users' opinions on the news by exploiting the auxiliary information of the users' engagements with the news tweets on social media, and aggregate their opinions in a well-designed unsupervised way to generate our estimation results. We observe that as news propagates, users engage differently on social media, such as publishing a news tweet, liking, forwarding, or replying to a news tweet. This information can, on a certain level, reflect the users' opinions on the news. For example, Figure 1 shows two news tweet examples regarding the aforementioned news. According to the users' tweet contexts, we can see that the user in Figure 1(a) disagreed with the authenticity of the news, which may indicate the user's high credibility in identifying fake news. On the other hand, it appears that the user in Figure 1(b) falsely believed the news or intentionally spread the fake news, implying the user's deficiency in the ability to identify fake news. Besides, as for other users who engaged in the tweets, it is likely that the users who liked/retweeted the first tweet also doubted the news, while

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html

(a) Doubting the authenticity of the news

(b) Agreeing to the authenticity of the news

Figure 1: News Tweet Examples

those who liked/retweeted the second tweet may also be deceived by the news. The users' opinions towards the news can also be discovered by examining their replies to the news tweets (Pang and Lee 2008).

Based on the intuition, we aim to exploit the users' opinions on news revealed by their engagement behaviors on social media to identify the authenticity of the news. However, a major challenge is that the social engagement information of social media users, as well as the extracted user opinions, are usually conflicting and unreliable, as the users usually have heterogeneous credibility in identifying fake news. In addition, as fake news is usually carefully written with the intent to mislead readers, it is very likely that the majority of the users' opinions are unreliable. Thus, a simple majority voting or averaging scheme may fail. One possible alternative is to employ truth discovery algorithms (Li et al. 2016), which are proposed to tackle conflicting information provided by multiple data sources. However, truth discovery algorithms only work on a simple source-item model, which can be represented as a bipartite graph and each edge of the graph denotes the data of each source-item pair. As the relationships among news, tweets, and users on social media form more complicated topologies (Jin et al. 2014), existing truth discovery algorithms may not be applicable.

In this work, we study the problem of unsupervised fake news detection with unreliable social engagements. In an attempt to address the challenges of this problem, we propose an unsupervised framework, namely UFD. It first extracts the users' opinions on the news by analyzing their engagements on social media, and builds a Bayesian probability graphical model capturing the complete generative process of the truths of news and the users' opinions. An efficient collapsed Gibbs sampling approach is proposed to detect fake news and estimate the users' credibility simultaneously. Experiments on two real-world datasets demonstrate the effectiveness of our proposed methods. The major contributions of this work are listed as follows.

- We investigate the problem of unsupervised fake news detection on social media by exploiting the users' unreliable social engagement information.
- We propose an unsupervised learning framework, UFD, which utilizes a probabilistic graphical model to model the truths of news and the users' credibility. An efficient collapsed Gibbs sampling approach is proposed to solve the inference problem.
- We conduct experiments on two real-world social media datasets, and the experiment results demonstrate the effectiveness of the proposed framework for fake news detection on social media.

2 Related Work

The problem of fake news detection has become an emerging topic in recent social media studies. Existing fake news detection approaches generally fall into two categories: using *news contents* and using *social contexts* (Shu et al. 2017).

For news content-based approaches, linguistic features or visual features are extracted. Linguistic features, such as lexical and syntactic features, capture specific writing styles and sensational headlines that commonly occur in fake news contents (Potthast et al. 2017), while visual features are used to identify fake images that are intentionally created or to capture specific characteristics for images in fake news (Gupta et al. 2013). Models that exploit the news contents-based features can be classified into (1) knowledge-based: using external sources to check the authenticity of claims in news contents (Magdy and Wanas 2010; Wu et al. 2014), and (2) style-based: capturing the manipulation in writing style, such as deception (Rubin and Lukoianova 2015) and non-objectivity (Potthast et al. 2017).

As for social context-based methods, they incorporate features from user profiles, post contents, and social networks. User profiles can be used to measure the users' characteristics and credibility (Castillo, Mendoza, and Poblete 2011). Features extracted from the users' posts represent the users' social responses, such as stances (Jin et al. 2016). Network features are extracted by constructing specific social networks, such as diffusion networks (Kwon et al. 2013) or co-occurrence networks (Ruchansky, Seo, and Liu 2017). The social context models can be categorized as either stance-based or propagation-based. Stance-based models utilize the users' opinions towards the news to infer news veracity (Jin et al. 2016), while propagation-based models apply propagation methods to model unique patterns of information dissemination (Jin et al. 2016; Wu, Yang, and Zhu 2015).

The aforementioned methods are all supervised approaches which mainly focus on extracting effective features, and use them to build supervised learning frameworks. In contrast, in this paper, we strive to address the problem of fake news detection in an unsupervised manner by exploiting the user engagement information. The key idea is the user credibility estimation, which was not considered by existing fake news detection methods.

3 Problem Model

In this section, we present details of the proposed framework UFD. We first introduce the hierarchical social engagement model, then present the problem details, and finally formalize the problem into a Bayesian network.

3.1 Hierarchical User Engagement

Definition 1 (Fake News)**.** *Fake news is a news report that is verifiably false.*

After a news is published, a large number of users may engage in its propagation over online social networks. The users may create tweets regarding the news, or engage with (*i.e.*, like, retweet, reply to) other users' tweets. Similar to (Jin et al. 2016), we define a *news tweet* as follows.

Definition 2 (News Tweet)**.** *A news tweet is a news message posted by a user on social media along with its social contexts.*

Figure 2 presents an overview of the hierarchical user engagement model in social media. Specifically, for each news in the news corpus, a number of news tweets can be observed and collected on social media platforms (*e.g.*, using Twitter's advanced search API with the title of the news). The collected information of each news tweet contains the contents of the tweet (*i.e.*, a news title, a link to the original article, a picture, and the user's own text content) and the corresponding second-level user engagements (such as likes, retweets, and replies). Besides, the profiles of the tweet poster and the users who engaged in the tweet can also be collected.

Note that among a large number of tweets regarding a news on social media, tweets posted by well-known verified users, so-called "big-V", can attract great attention with many likes, retweets, and replies, whereas tweets published by most of the unverified and unpopular users may not receive much attention[2]. Based on this observation, we divide the social media users into two groups: *verified users* and *unverified users*, where the user verification information can be easily obtained from their user profiles. Then, in preparing our data, we only consider the tweets created by verified users and the related social engagements (like, retweet, and reply) of the unverified users.

The benefits of this are three-fold. First, the long-tail phenomenon of social media data can be alleviated. Since there are a large number of unverified users' tweets, which do not have many social engagements, considering these tweets may introduce a lot of noise to our data without helping us identify fake news. Second, by classifying the users into verified users and unverified users, an implicit assumption is imposed that verified users, who may have large influences and high social status, may have higher credibility in differentiating between fake news and real news. The third benefit is the simplification of our model. As the users' behaviors on social media are complicated, incomplete, and noisy, a perfect characterization of the users' behaviors is intractable. By concentrating on a small portion of social media data, we can simplify our follow-up problem model and reduce the complexity of our problem formulation.

3.2 Problem Model

Suppose the set of news is denoted by $\mathcal{N}$, and the sets of verified and unverified users are denoted by $\mathcal{M}$ and $\mathcal{K}$, respectively. For each given news $i \in \mathcal{N}$, we collect all the verified users' tweets on this news. Let $\mathcal{M}_i \subseteq \mathcal{M}$ denote the set of verified users who published tweets for the news. Then, for the tweet of each verified user $j \in \mathcal{M}_i$, we collect the unverified users' social engagements. Let $\mathcal{K}_{i,j} \subseteq \mathcal{K}$ denote the set of unverified users who engaged in the tweet.

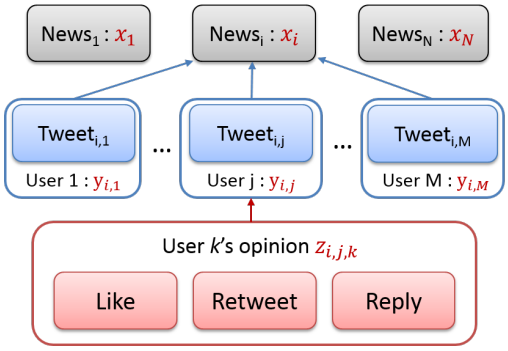


Figure 2: Hierarchical User Engagement Model

For each given news i, we use a latent random variable $x_i \in \{0, 1\}$ to denote its truth, *i.e.*, fake news ($x_i = 0$) or true news ($x_i = 1$). To infer whether a news piece is fake or not, we need to extract the users' *opinions* on the news from their engagement behaviors.

Definition 3 (User Opinion)**.** *A user's opinion on a news report is the user's implicitly expressed viewpoint towards the authenticity of the news.*

For each verified user $j \in \mathcal{M}_i$, we let $y_{i,j} \in \{0, 1\}$ denote the user's opinion on the news, *i.e.*, $y_{i,j}$ is 1 if the user thinks the news is real; and 0 otherwise. Several heuristics can be applied to extract $y_{i,j}$. Let News$_i$ and Tweet$_{i,j}$ denote the news content and the user j's own text content of the tweet, respectively. Then, $y_{i,j}$ can be defined as the sentiment of Tweet$_{i,j}$ (Gilbert 2014), or if the opinion of Tweet$_{i,j}$ is non-conflicting to that of News$_i$ (Dave, Lawrence, and Pennock 2003; Trabelsi and Zaiane 2014).

For verified user j's tweet on news i, many unverified users may like, retweet, or reply to the tweet. Let $z_{i,j,k} \in \{0, 1\}$ denote the opinion of the unverified user $k \in \mathcal{K}_{i,j}$. We assume that if the user k liked or retweeted[3] the tweet, then it implies that k agreed to the opinion of the tweet. If the user k replied to the tweet, then its opinion can be extracted by employing off-the-shelf sentiment analysis (Gilbert 2014) or conflicting opinion mining techniques (Dave, Lawrence, and Pennock 2003; Trabelsi and Zaiane 2014). It is common that an unverified user may conduct multiple engagements in a tweet (*e.g.*, liked and also replied to the tweet). In this case, the user's opinion $z_{i,j,k}$ is obtained using majority voting.

3.3 Probabilistic Graphical Model

Given the definitions of x_i, $y_{i,j}$, and $z_{i,j,k}$, we now present our unsupervised fake news detection framework (UFD). Figure 3 shows the probabilistic graphical structure of our model. Each node in the graph represents a random variable

[2]https://www.clickz.com/your-long-tail-influencers/39598/

[3]Twitter treats forwarding w/o comments as retweeting, while forwarding w/ comments is treated as publishing a new tweet.

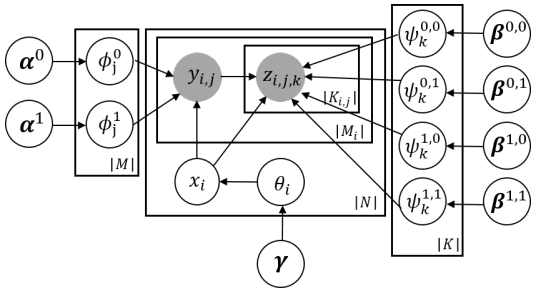


Figure 3: The Probabilistic Graphical Model

or a prior parameter, where darker nodes and white nodes indicate observed or latent variables, respectively.

1. News. For each news i, x_i is generated from a Bernoulli distribution with parameter θ_i:

$$x_i \sim \text{Bernoulli}(\theta_i)$$

The prior probability of θ_i is generated from a Beta distribution with hyperparameter $\boldsymbol{\gamma} = (\gamma_1, \gamma_0)$ as follows:

$$\theta_i \sim \text{Beta}(\gamma_1, \gamma_0)$$

where γ_1 is the prior true count and γ_0 is the prior fake count. If we do not have a strong belief in practice, we can assign a uniform prior indicating that each news has an equal probability of being true or fake.

2. Verified User. For each verified user j, its credibility in fake news identification is modelled with two variables ϕ_j^1 and ϕ_j^0. Specifically, ϕ_j^1 represent its *sensitivity* (true positive rate) and ϕ_j^0 its 1-*specificity* (false positive rate), *i.e.*,

$$\phi_j^1 := p(y_{i,j} = 1 | x_i = 1)$$
$$\phi_j^0 := p(y_{i,j} = 1 | x_i = 0)$$

These two parameters denote the probability that the user j thinks a news piece is real given the truth estimation of the news is true and fake, respectively. We generate the sensitivity of each user from a Beta distribution with hyperparameter $\boldsymbol{\alpha}^1 = (\alpha_1^1, \alpha_0^1)$. Here, α_1^1 is the prior true positive count, and α_0^1 is the prior false negative count:

$$\phi_j^1 \sim \text{Beta}(\alpha_1^1, \alpha_0^1)$$

The 1-specificity is generated from another Beta distribution with hyperparameter $\boldsymbol{\alpha}^0 = (\alpha_1^0, \alpha_0^0)$ as follows:

$$\phi_j^0 \sim \text{Beta}(\alpha_1^0, \alpha_0^0)$$

where α_1^0 is the prior false positive count and α_0^0 is the prior true negative count.

Given ϕ_j^1 and ϕ_j^0, we can see that the opinion of each verified user j in the news i is generated from a Bernoulli distribution with parameter $\phi_j^{x_i}$, *i.e.*,

$$y_{i,j} \sim \text{Bernoulli}(\phi_j^{x_i})$$

3. Unverified User. Different from the verified users, as the unverified users engage in the verified users' tweets, their opinions are likely to be influenced by the news itself and the verified users' opinions. Based on this observation, for each unverified user $k \in \mathcal{K}$, the following four variables are adopted to model its credibility:

$$\psi_k^{0,0} := p(z_{i,j,k} = 1 | x_i = 0, y_{i,j} = 0)$$
$$\psi_k^{0,1} := p(z_{i,j,k} = 1 | x_i = 0, y_{i,j} = 1)$$
$$\psi_k^{1,0} := p(z_{i,j,k} = 1 | x_i = 1, y_{i,j} = 0)$$
$$\psi_k^{1,1} := p(z_{i,j,k} = 1 | x_i = 1, y_{i,j} = 1)$$

where for each pair of $(u, v) \in \{0, 1\}^2$, $\psi_k^{u,v}$ represents the probability that the unverified user k thinks the news is true under the condition that the truth estimation of the news is u and the verified user's opinion is v. For each $\psi_k^{u,v}$, it is generated from a beta distribution with hyperparameter $\boldsymbol{\beta^{u,v}}$:

$$\psi_k^{u,v} \sim \text{Beta}(\beta_1^{u,v}, \beta_0^{u,v}).$$

Given the truth estimation of news x_i, and the verified user's opinion $y_{i,j}$, we generate the unverified user's opinion from a Bernoulli distribution with parameter $\psi_k^{x_i, y_{i,j}}$, *i.e.*,

$$z_{i,j,k} \sim \text{Bernoulli}(\psi_k^{x_i, y_{i,j}})$$

3.4 Problem Formulation

Our objective is to find instances of the latent truth variables that maximize the joint probability, *i.e.*, get the *maximum a posterior* (MAP) estimate for $\boldsymbol{x}$:

$$\hat{\boldsymbol{x}}_{\text{MAP}} = \arg\max_{\boldsymbol{x}} \iiint p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\Phi}\, \mathrm{d}\boldsymbol{\Psi} \quad (1)$$

where for simplicity of presentation, we use $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ to denote $\{\boldsymbol{\phi}^0, \boldsymbol{\phi}^1\}$ and $\{\boldsymbol{\psi}^{0,0}, \boldsymbol{\psi}^{0,1}, \boldsymbol{\psi}^{1,0}, \boldsymbol{\psi}^{1,1}\}$, respectively.

However, an exact inference on the posterior distribution may result in an exponential complexity. In the next section, we will propose an efficient inference algorithm.

4 Fake News Detection Algorithm

In this section, we propose an efficient collapsed Gibbs sampling algorithm to estimate the truths of news and the users' credibility simultaneously.

4.1 Gibbs Sampling

To deal with the infeasibility of exact inference, we turn to Gibbs sampling approach, which is a widely-used MCMC method to approximate a multivariate distribution when direct sampling is intractable (Robert and Casella 2013). Due to the conjugacy of exponential families, unknown parameters $\boldsymbol{\theta}, \Phi, \Psi$ can be integrated out in the sampling process. Thus, we only need to iteratively sample the truth of each news based on the following conditional distribution:

$$p(x_i = s | \boldsymbol{x}_{-i}, \boldsymbol{y}, \boldsymbol{z}), \quad (2)$$

where $s \in \{0, 1\}$ and $\boldsymbol{x}_{-i}$ denotes the truths estimations of all the news except i.

4.2 Update Rule

Using Bayes rule, Equation 2 can be rewritten as follow:

$$p(x_i = s|\boldsymbol{x}_{-i}, \boldsymbol{y}, \boldsymbol{z}) \propto p(x_i = s|\boldsymbol{x}_{-i})\, p(\boldsymbol{y}_{i,*}, \boldsymbol{z}_{i,*,*}|x_i = s, \boldsymbol{y}_{-i,*}, \boldsymbol{z}_{-i,*,*}), \quad (3)$$

where $\boldsymbol{y}_{i,*}$ denotes all the verified users' opinions regarding news i, and $\boldsymbol{z}_{i,*,*}$ denotes all the unverified users' opinions regarding news i.

Note that in Equation 3, the first term is the prior and the second term is the likelihood. We first examine the first term:

$$\begin{aligned} & p(x_i = s|\boldsymbol{x}_{-i}) \\ &= \int p(x_i = s, \theta_i|\boldsymbol{x}_{-i})\mathrm{d}\theta_i = \int p(x_i = s|\theta_i)p(\theta_i|\boldsymbol{x}_{-i})\mathrm{d}\theta_i \\ &= \frac{1}{\mathbf{B}(\gamma_1, \gamma_0)} \int (\theta_i)^s (1-\theta_i)^{1-s} (\theta_i)^{\gamma_1 - 1} (1-\theta_i)^{\gamma_0 - 1} \mathrm{d}\theta_i \\ &= \frac{1}{\mathbf{B}(\gamma_1, \gamma_0)} \int (\theta_i)^{\gamma_1 + s - 1} (1-\theta_i)^{\gamma_0 + (1-s) - 1} \mathrm{d}\theta_i \\ &= \frac{\mathbf{B}(\gamma_1 + s, \gamma_0 + 1 - s)}{\mathbf{B}(\gamma_1, \gamma_0)} = \frac{\gamma_t}{\gamma_1 + \gamma_0} \propto \gamma_s, \end{aligned} \quad (4)$$

where $\mathbf{B}()$ is the Beta function.

As for the second term in Equation 3, we have:

$$\begin{aligned} & p(\boldsymbol{y}_{i,*}, \boldsymbol{z}_{i,*,*}|x_i = s, \boldsymbol{y}_{-i,*}, \boldsymbol{z}_{-i,*,*}) \\ &= \prod_{j \in \mathcal{M}_i} \Big(p(y_{i,j}|x_i = s, \boldsymbol{y}_{-i,j}) \prod_{k \in \mathcal{K}_{i,j}} p(z_{i,j,k}|x_i = s, y_{i,j}, \boldsymbol{z}_{-i,j,k}) \Big) \end{aligned} \quad (5)$$

For the inner term of Equation(5), we have:

$$\begin{aligned} & p(z_{i,j,k}|x_i = s, y_{i,j}, \boldsymbol{z}_{-i,j,k}) \\ &= \int p(z_{i,j,k}|\psi_k^{s,y_{i,j}})\, p(\psi_k^{s,y_{i,j}}|\boldsymbol{z}_{-i,j,k})\, \mathrm{d}\psi_k^{s,y_{i,j}} \\ &\propto \frac{\beta_{z_{i,j,k}}^{s,y_{i,j}} + n_{k,-i,z_{i,j,k}}^{s,y_{i,j}}}{\beta_1^{s,y_{i,j}} + n_{k,-i,1}^{s,y_{i,j}} + \beta_0^{s,y_{i,j}} + n_{k,-i,0}^{s,y_{i,j}}} \end{aligned} \quad (6)$$

where $n_{k,-i,z_{i,j,k}}^{s,y_{i,j}}$ is the number of unverified user k's opinions with the value of $z_{i,j,k}$, when the referred news is not i, the truth estimation of the news i is s, and the opinion of the verified user's tweet it engaged with is $y_{i,j}$. The last step of Equation (6) is due to:

$$p(\psi_k^{s,y_{i,j}}|\boldsymbol{z}_{-i,j,k}) \sim Beta(\beta_1^{s,y_{i,j}} + n_{k,-i,1}^{s,y_{i,j}}, \beta_0^{s,y_{i,j}} + n_{k,-i,0}^{s,y_{i,j}})$$

For the outer term of Equation(5), we have:

$$\begin{aligned} & p(y_{i,j}|x_i = s, \boldsymbol{y}_{-i,j}) \\ &= \int p(y_{i,j}|\phi_j^s)\, p(\phi_j^s|\boldsymbol{y}_{-i,j})\mathrm{d}\phi_j^s \\ &\propto \frac{\alpha_{y_{i,j}}^s + m_{j,-i,y_{i,j}}^s}{\alpha_1^s + m_{j,-i,1}^s + \alpha_0^s + m_{j,-i,0}^s} \end{aligned} \quad (7)$$

where $m_{j,-i,y_{i,j}}^s$ is the number of verified user j's opinions whose values are $y_{i,j}$, when the referred news is not i and the truth estimation of the news is s. The last step of Equation (7) is due to:

$$p(\phi_j^s|\boldsymbol{y}_{-i,j}) \sim Beta(\alpha_1^s + m_{j,-i,1}^s, \alpha_0^s + m_{j,-i,0}^s)$$

Algorithm 1: Collapsed Gibbs Sampling

```
Randomly initialize x_i^(0) with 0 or 1, ∀i ∈ N;
Initialize counts m for ∀j ∈ M and n for ∀k ∈ K;
Sample record R ← ∅;
for t = 1 → iter_num do
    foreach news i ∈ N do
        Sample x_i^(t) using Equation (8);
        Update counts;
    if t > burn-in & t % thinning = 0 then
        R ← R ∪ {x^(t)};
return (1/|R|) Σ_{x^(t) ∈ R} x^(t);
```

Combining Equation (4), (6), and(7), we obtain the update rule of our collapsed Gibbs sampler:

$$\begin{aligned} & p(x_i = s|\boldsymbol{x}_{-i}, \boldsymbol{y}, \boldsymbol{z}) \\ &\propto \gamma_s \prod_{j \in \mathcal{M}_i} \left(\frac{\alpha_{y_{i,j}}^s + m_{j,-i,y_{i,j}}^s}{\alpha_1^s + m_{j,-i,1}^s + \alpha_0^s + m_{j,-i,0}^s} \times \right. \\ & \left. \prod_{k \in \mathcal{K}_{i,j}} \frac{\beta_{z_{i,j,k}}^{s,y_{i,j}} + n_{k,-i,z_{i,j,k}}^{s,y_{i,j}}}{\beta_1^{s,y_{i,j}} + n_{k,-i,1}^{s,y_{i,j}} + \beta_0^{s,y_{i,j}} + n_{k,-i,0}^{s,y_{i,j}}} \right) \end{aligned} \quad (8)$$

4.3 Fake News Detection Algorithm

Having obtained the update rule of collapsed Gibbs sampler. The fake news detection procedure is straightforward. Algorithm 1 shows the pseudo-code of the algorithm. We first randomly initialize the truth estimation of each news to either 0 or 1, and calculate the counts of each verified and unverified user based on the initial truth estimations. Then, we conduct the sampling process for a number of iterations. In each iteration, we sample the truth estimation of each news from its distribution conditioned on the current estimations of all the other news specified by Equation (8), and update the counts of each user accordingly.

Note that as with other MCMC algorithms, Gibbs sampler generates a Markov chain of samples that are correlated with nearby samples. As a result, samples from the beginning of the chain may not accurately represent the desired distribution, thus we discard the samples in the first few iterations (the *burn-in* period). Besides, a *thinning* technique is used to reduce correlations in the samples. In the end, we calculate the average values of the collected samples and round them up to 0 or 1 as the final estimations of the news.

4.4 User's Credibility

The user's credibility for identifying fake news can be readily obtained using the closed form solution, as the posterior probability is also a Beta distribution.

For each verified user $j \in \mathcal{M}$, we have its sensitivity and

1-specificity as follows:

$$\phi_j^1 = \frac{\mathbb{E}[m_{j,1}^1] + \alpha_1^1}{\mathbb{E}[m_{j,1}^1] + \alpha_1^1 + \mathbb{E}[m_{j,0}^1] + \alpha_0^1} \quad (9)$$

$$\phi_j^0 = \frac{\mathbb{E}[m_{j,1}^0] + \alpha_1^0}{\mathbb{E}[m_{j,1}^0] + \alpha_1^0 + \mathbb{E}[m_{j,0}^0] + \alpha_0^0} \quad (10)$$

where $\mathbb{E}[m_{j,y_{i,j}}^{x_i}]$ is the expected value of j's count where the truth estimation of news is x_i and j's opinion is $y_{i,j}$. It can be calculated using the average value of the $m_{j,y_{i,j}}^{x_i}$ records in the sampling process. For each unverified user, its sensitivity and 1-specificity can be calculated accordingly. The proposed method can also be easily adjusted to adapt streaming data scenarios by using the credibility learned on current stage as the prior for future data.

5 Experiment

In this section, we conduct experiments to evaluate the performance of our proposed method.

5.1 Dataset

In the experiment, we use two public datasets, *i.e.*, LIAR (Wang 2017) and BuzzFeed News[4] to evaluate the performance of our algorithm. LIAR is one of the largest fake news datasets, containing over 12,800 short news statements and labels collected from a fact-checking website *politifact.com*. BuzzFeed dataset contains 1,627 news articles related to the 2016 U.S. election from Facebook. We use Twitter's advanced search API with the titles of news to collect related news tweets. After eliminating duplicate news and filtering out the news with no verified user's tweets, we finally obtain 332 news for LIAR and 144 news for BuzzFeed. For each news tweet, the unverified users' engagements are also collected using web scraping. We observed that users tend to explicitly express negative sentiments (using words like "lie", "fake") when they think a news report is fake. Thus, we use the sentiments as their opinions. As for likes and retweets, we treat them as positive opinions. Note that if a user has very few engagement records, the user's credibility cannot be accurately estimated. Thus, we filter out the users who have less than 3 engagement records. Finally, the statistics of our datasets are shown in Table 1.

Table 1: The statistics of datasets

Datasets	LIAR	BuzzFeed
# News	332	144
# True news	182	67
# Fake news	150	77
# Tweets	2,589	1,007
# Verified users	550	243
# Unverified users	3,767	988
# Engagements	19,769	7,978
# Likes	5,713	1,277
# Retweets	10,434	2,365
# Replies	3,622	4,336

5.2 Experiment Setup

Performance Metric: We use the following metrics to evaluate the performance of our fake news detection algorithm: *accuracy*, *precision*, *recall*, and *F1-score*, which are widely-used to evaluate the performance of classification tasks.

Benchmark Algorithms: We compare our proposed algorithm with four unsupervised fake news detection benchmarks listed as follows. As there are no existing unsupervised methods taking the second-level user engagement information (like, retweet, reply) into consideration, the compared algorithms only utilize the first-level user engagement (*i.e.*, the opinions of the verified users) to generate estimations for the authenticity of each news. In contrast to the benchmarks, our proposed algorithm exploits the entire hierarchical user engagement information to identify fake news.

- Majority Voting: For each news, it outputs the most frequent verified user's opinion as the estimation result.
- TruthFinder (Yin, Han, and Philip 2008): It is an unsupervised learning method that iteratively calculates the truth estimation of each news based on the conflicting relationships among the verified users' tweets.
- LTM (Zhao et al. 2012): It is a graphical model-based truth discovery algorithm which considers the two-sided errors of each data contributor. However, it only works on a simple source-item model.
- CRH (Li et al. 2014): It is a general truth discovery framework that models the credibility of each user using a single unknown variable, representing the overall accuracy of the user's contributed data.

Parameter Settings: In the experiment, we set uniform priors for news count, *i.e.*, $\boldsymbol{r} = (5, 5)$ so that each news has an equal chance of being true or fake. We set prior for sensitivity as $\boldsymbol{\alpha}^0 = (7, 3)$ and prior for 1-specificity as $\boldsymbol{\alpha}^1 = (3, 7)$ to plug in the assumption that verified users are usually reliable and do not have high false positive or false negative rates. As for unverified users, for each pair of $(u, v) \in \{0, 1\}^2$, we set $\boldsymbol{\beta}^{u,v} = (1, 9)$ indicating the observation that most of the unverified users reveal positive opinions. As for the Gibbs sampling algorithm, the number of iterations is set to 100. The burn-in period and thinning are set to 20 and 4, respectively. Parameters for the benchmarks are set according to the suggestions of their papers.

5.3 Experiment Result

Performance Analysis: Table 2 and Table 3 show the experiment results on LIAR and BuzzFeed datasets, respectively. Precision, recall, and F1-score are measured on each news class to present complete characterizations of the algorithms. Several observations can be drawn. First, majority voting achieves the worst performance since it equally aggregates the users' opinions without considering the users' credibility information. Second, our proposed fake news detection algorithm UFD achieves the best performance in

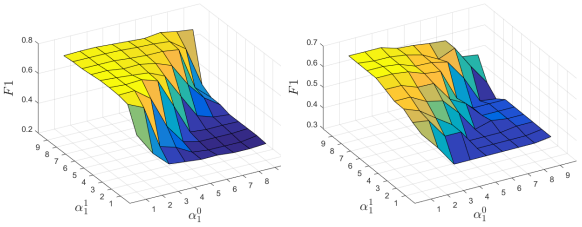
[4]https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/blob/master/data

Table 2: Performance comparison on LIAR dataset

Methods	Accuracy	True			Fake		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Majority Voting	0.586	0.624	0.628	0.626	0.539	0.534	0.537
TruthFinder	0.634	0.650	0.679	0.664	0.615	0.583	0.599
LTM	0.641	0.654	0.691	0.672	0.624	0.583	0.603
CRH	0.639	0.653	0.687	0.669	0.621	0.583	0.601
UFD	**0.759**	**0.766**	**0.783**	**0.774**	**0.750**	**0.732**	**0.741**

Table 3: Performance comparison on BuzzFeed dataset

Methods	Accuracy	True			Fake		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Majority Voting	0.556	0.532	0.373	0.439	0.567	0.714	0.632
TruthFinder	0.554	0.523	0.359	0.426	0.568	**0.720**	0.635
LTM	0.465	0.443	0.582	0.503	0.500	0.364	0.421
CRH	0.562	0.542	0.388	0.452	0.573	0.714	0.636
UFD	**0.679**	**0.667**	**0.714**	**0.690**	**0.692**	0.643	**0.668**



(a) LIAR (b) BuzzFeed

Figure 4: Hyperparameter Analysis ($\boldsymbol{\alpha}$)

LIAR dataset, outperforming the second best algorithm by 18.4% in terms of accuracy. In BuzzFeed dataset, UFD achieves the best performance except for recall on fake news class. Although majority voting, TruthFinder, and CRH achieve higher recall on fake news class, they have a high tendency classifying news as fake news, leading to poor performance in true news class. Thus, the experiment results validate the effectiveness of UFD. Comparing with the benchmarks that only exploits the information in news tweets, incorporating the second-level user engagements (likes, retweets, and replies) can dramatically improve the performance of fake news detection, as the number of second-level user engagements is usually much larger than the number of news tweets, providing further guidance for the truth inference procedure. Third, we can see that UFD performs better on LIAR dataset than BuzzFeed dataset, mainly due to the fact that the user engagements on BuzzFeed are sparser than LIAR.

We also conduct experiments to compare our algorithm with several supervised methods. It turns out that simple supervised classifier such as SVM and naive Bayes with n-gram do not achieve better performance than ours (with accuracy around 0.7), while recent advances, such as (Shu, Wang, and Liu 2017), could achieve accuracy over 0.8.

Table 4: Top accurate verified users on two datasets

User	Accuracy	Sensitivity	Specificity
amy_hollyfield	1.0	1.0	1.0
politico	0.909	0.833	1.0
loujacobson	0.84	0.842	0.833
dcexaminer	0.833	0.818	0.857
FoxNews	0.818	0.714	1.0

Impact of $\boldsymbol{\alpha}$ prior: To understand the prior for sensitivity $\boldsymbol{\alpha}^1$ and 1-specificity $\boldsymbol{\alpha}^0$, we vary α_1^1 and α_1^0 from 1 to 9, where α_0^1 and α_0^0 are set to $10 - \alpha_1^1$ and $10 - \alpha_1^0$, respectively. The F1-scores of our algorithm are presented in Figure 4. We can see that UFD works well with large α_1^1 (prior true positive count) and low α_1^0 (prior false positive count), and the performance decreases as α_1^1 decreases and α_1^0 increases. This is because imposing low true positive count or high false positive count will flip every truth estimation to enforce high likelihood leading to incorrect inferences.

User Credibility Estimation: Besides providing a truth estimation for each news, our algorithm also produces credibility estimations for each user. To give readers a taste of this part, Table 4 shows the top-5 credible verified users in the two datasets sorted according to accuracy. Among the top 5 users, amy_hollyfield is a news reporter of NBC7 and loujacobson is a senior correspondent for PolitiFact (a fact-checking website), while the other three users are well-known news agencies. These results are in line with people's expectation that professional news reporters and news agencies should have high expertise in identifying fake news.

6 Conclusion

In this paper, we consider the novel problem of unsupervised fake news detection. We extract the social media users' opinions from their hierarchy social engagement information. By treating the truths of news and the credibility of users are latent random variables, a probabilistic graphical

model is built to capture the complete generative spectrum. An efficient Gibbs sampling approach is proposed to estimate the news authenticity and the users' credibility simultaneously. We evaluate the proposed method on two real-world datasets, and the experiment results show that our proposed algorithm outperforms the unsupervised benchmarks.

As for future work, we plan to incorporate the features of news contents and user profiles into our current fake news detection model. In addition, building a semi-supervised learning framework to improve the performance of unsupervised model could also be an interesting research direction.

Acknowledgments

This work was supported in part by the National Key R&D Program of China 2018YFB1004703, in part by China NSF grant 61672348, 61672353, and 61472252, and in part by State Scholarship Fund of China Scholarship Council. The opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies or the government. This work is done when the first author was visiting Data Mining and Machine Learning lab in ASU.

References

Bond Jr, C. F., and DePaulo, B. M. 2006. Accuracy of deception judgments. *Personality and social psychology Review* 10(3):214–234.

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.

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, 519–528. ACM.

Gilbert, C. H. E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM)*.

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.

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

Kim, J.; Tabibian, B.; Oh, A.; Schölkopf, B.; and Gomez-Rodriguez, M. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*, 324–332. ACM.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Prominent features of rumor propagation in online social media. In *ICDM'13*, 1103–1108. IEEE.

Li, Q.; Li, Y.; Gao, J.; Zhao, B.; Fan, W.; and Han, J. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 1187–1198. ACM.

Li, Y.; Gao, J.; Meng, C.; Li, Q.; Su, L.; Zhao, B.; Fan, W.; and Han, J. 2016. A survey on truth discovery. *ACM Sigkdd Explorations Newsletter* 17(2):1–16.

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.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

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*.

Robert, C., and Casella, G. 2013. *Monte Carlo statistical methods*. Springer Science & Business Media.

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.

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

Shearer, E., and Gottfried, J. 2017. News use across social media platforms 2017. *Pew Research Center, Journalism and Media*.

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

Shu, K.; Wang, S.; and Liu, H. 2017. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*.

Trabelsi, A., and Zaiane, O. R. 2014. Mining contentious documents using an unsupervised topic model based approach. In *2014 IEEE International Conference on Data Mining (ICDM)*, 550–559. IEEE.

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

Wu, L., and Liu, H. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*, 637–645. ACM.

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.

Wu, K.; Yang, S.; and Zhu, K. Q. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering*, 651–662.

Yin, X.; Han, J.; and Philip, S. Y. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering* 20(6):796–808.

Zhao, B.; Rubinstein, B. I.; Gemmell, J.; and Han, J. 2012. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment* 5(6):550–561.

View publication stats