Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users Based on Weakly Supervised Learning

Chunyuan Yuan[1,2], **Qianwen Ma**[1,2], **Wei Zhou**[1,*], **Jizhong Han**[2] **and Songlin Hu**[1,2,*]
[1] Institute of Information Engineering, Chinese Academy of Sciences
[1] School of Cyber Security, University of Chinese Academy of Sciences
{yuanchunyuan,maqianwen,zhouwei,hanjizhong,husonglin}@iie.ac.cn

Abstract

The dissemination of fake news significantly affects personal reputation and public trust. Recently, fake news detection has attracted tremendous attention, and previous studies mainly focused on finding clues from news content or diffusion path. However, the required features of previous models are often unavailable or insufficient in early detection scenarios, resulting in poor performance. Thus, early fake news detection remains a tough challenge. Intuitively, the news from trusted and authoritative sources or shared by many users with a good reputation is more reliable than other news. Using the credibility of publishers and users as prior weakly supervised information, we can quickly locate fake news in massive news and detect them in the early stages of dissemination.

In this paper, we propose a novel **S**tructure-aware **M**ulti-head **A**ttention **N**etwork (SMAN), which combines the news content, publishing, and reposting relations of publishers and users, to jointly optimize the fake news detection and credibility prediction tasks. In this way, we can explicitly exploit the credibility of publishers and users for early fake news detection. We conducted experiments on three real-world datasets, and the results show that SMAN can detect fake news in 4 hours with an accuracy of over 91%, which is much faster than the state-of-the-art models.

1 Introduction

The widespread dissemination of fake news has lead to a significant influence on personal fame, public trust, and security. For example, spreading misinformation, such as "Asians are more vulnerable to novel coronavirus" [1] about COVID-19 has very serious repercussions, making people ignore the harmfulness of the virus and directly affecting public health. Research has shown that misinformation spreads faster, farther, deeper, and more widely than true information (Vosoughi et al., 2018). Therefore, fake news detection on social media has attracted tremendous attention recently in both research and industrial fields.

Early research on fake news detection mainly focused on the design of effective features from various sources, including textual content, user profiling data, and news diffusion patterns. Linguistic features, such as writing styles and sensational headlines (Kwon et al., 2013), lexical and syntactic analysis (Potthast et al., 2017), have been explored to separate fake news from true news. Apart from linguistic features, some studies also proposed a series of user-based features (Castillo et al., 2011; Shu et al., 2018), and temporal features (Kwon et al., 2013) about the news diffusion. However, these feature-based methods are very time-consuming, biased, and require a lot of labor to design. Besides, these features are easily manipulated by users.

To solve the above problems, many recent studies (Ma et al., 2016; Yu et al., 2017; Guo et al., 2018; Shu et al., 2019; Yuan et al., 2019) apply various neural networks to automatically learn high-level representations for fake news detection. For example, recurrent neural network (RNN) (Ma et al., 2016),

* Corresponding author.

[1] https://www.thestar.com.my/news/regional/2020/03/11/myth-busters-10-common-rumours-about-covid-19

Proceedings of the 28th International Conference on Computational Linguistics, pages 5444–5454
Barcelona, Spain (Online), December 8-13, 2020

convolutional neural network (CNN) (Yu et al., 2017), matrix factorization (Shu et al., 2019) and graph neural network (Yuan et al., 2019) are applied to learn the representation of content and diffusion graph of news. These methods only apply more types of information for fake news detection, but paying little attention to early detection. Moreover, these models can only detect fake news in consideration of all or a fixed proportion of repost information, while in practice they cannot detect fake news in the early stage of news propagation (Song et al., 2018). Some studies (Liu and Wu, 2018; Song et al., 2018; Zhou et al., 2019) explore to detect fake news early by relying on a minimum number of posts. The main limitation of these methods is that they ignore the importance of publishers' and users' credibility for the early detection of fake news.

When we humans see a piece of breaking news, we firstly may use common sense to judge whether there are factual errors in it. At the same time, we will also consider the reputation of the publishers and reposted users. People tend to believe the news from a trusted and authoritative source or the news shared by lots of users with a good reputation. If the publisher is reliable, we tend to believe this news. On the other hand, if the news is reposted by many low-reputation users in a short period, it may be that some spammers tried to heat up on the news (Chen and Chen, 2015; Vosoughi et al., 2018), resulting in lower credibility of the news.

Inspired by the above observation, we explicitly take the credibility of publishers and users as supervised information, and model fake news detection as a multi-task classification task. We can annotate a small part of publishers and users by their historical publishing and reposting behaviors. Although the credibility of publishers and users does not always provide correct information, they are necessary complementary supervised information for fake news detection. To make the credibility information generalized to other unannotated users, we construct a heterogeneous graph to build the connections of publishers, news, and users. Through a graph-based encoding algorithm, every node in the graph will be influenced by the credibility of publishers and users.

In this paper, we address the following challenges: (1) How to fully encode the heterogeneous graph structure and news content; and (2) How to explicitly utilize the credibility of publishers and users for facilitating early detection of fake news. To tackle the above challenges, we propose a novel structure-aware multi-head attention network for early detection of fake news. Firstly, we design a structure-aware multi-head attention module to learn the structure of the publishing graph and produce the publisher representations for the credibility prediction of publishers. Then, we apply the structure-aware multi-head attention module to encode the diffusion graph of the news among users and generate user representations for the credibility prediction of users. Finally, we apply a convolutional neural network to map the news text from word embedding to semantic space and utilize the fusion attention module to combine the news, publisher, and user representations for early fake news detection.

The contributions of this paper can be summarized as follows:

- We propose a novel strategy that explicitly takes the credibility of publishers and users as weakly supervised information for facilitating early detection of fake news.
- We provide a principled way to jointly utilize the credibility of publishers and users, and the heterogeneous graph for credibility prediction and fake news detection.
- We conduct extensive experiments on three real-world datasets. Experimental results show that our model achieves significant improvement over state-of-the-art models on both fake news detection and early detection tasks.

2 Related Work

2.1 Feature-based Methods

Early studies in fake news detection concentrate on designing some good features for separating fake news from true news. These features are mainly extracted from text content or users' profile information. Linguistic patterns, such as special characters and keywords (Castillo et al., 2011), writing styles and sensational headlines (Kwon et al., 2013), lexical and syntactic features (Feng et al., 2012; Potthast et al.,

2017), temporal-linguistic features (Ma et al., 2015; Zhao et al., 2015a), have been explored to detect fake news. Apart from linguistic features, some studies also proposed a series of user-based features (Castillo et al., 2011; Yang et al., 2012), e.g. the number of fans, registration age, and genders (Castillo et al., 2011) to find clues for fake news detection.

However, the language used in social media is highly informal and ungrammatical, which makes traditional natural language processing techniques hard to effectively learn semantic information from news content. Second, designing effective functions is often time-consuming and relies heavily on expert knowledge in specific fields. There are some features are often unavailable or inadequate in the early stage of news propagation.

2.2 Deep Learning Methods

Recurrent neural network (RNN) (Ma et al., 2016), convolutional neural network (CNN) (Yu et al., 2017) and graph neural network (Yuan et al., 2019) have been imported to learn the representations from news content or diffusion graph. Some studies also combine news content and users' response, such as conflicting viewpoints (Jin et al., 2016), topics (Guo et al., 2018), or stance (Bhatt et al., 2018; Li et al., 2019), to find clues by neural networks for fake news detection. These methods only apply more types of information for fake news detection, but paying little attention to early detection.

Recently, some studies (Liu and Wu, 2018; Song et al., 2018; Shu et al., 2019; Zhou et al., 2019) have proposed some methods to detect fake news at the early stage of propagation. However, these methods ignored the importance of publishers' and users' credibility for the early detection of fake news. Different from these studies, our method explicitly takes the credibility of publishers and users as weakly supervised information for facilitating fake news detection. We propose a novel deep learning model to simultaneously optimize the fake news detection task and users' credibility prediction task.

3 Problem Formulation

Let $\mathcal{N} = \{m_1, m_2, \dots m_{|\mathcal{N}|}\}$ be the set of news. Each news m_j has one publisher at least and K users $\{R_1, R_2, \dots, R_K\}$ to repost it at most. The publisher-news relations form a publishing graph $\mathcal{G}(V_p, E)$. The publisher-user relations form a diffusion graph $\mathcal{G}(V_u, E)$. In the diffusion graph of news, we regard users who repost the news as the neighbor nodes of the publisher. We use $|P|$, $|N|$, and $|U|$ to denote the amount of publishers, news, and users respectively.

For fake news detection task, our target is to learn a function $p(c|m_j, \mathcal{P}, \mathcal{N}, \mathcal{U}; \theta_3)$ to predict whether a piece of news is fake or not. c is class label of the news and θ_3 represents all parameters of the model.

In this paper, we design a credibility prediction subtask to explicitly utilize the users' or publishers' credibility information for fake news detection. For credibility prediction task, our goal is to learn a function $p(c|\mathcal{G}(V_p, E), \mathcal{P}; \theta_1)$ or $p(c|\mathcal{G}(V_u, E), \mathcal{U}; \theta_2)$ to predict the credit scores of publishers or users by publishing graph or diffusion graph.

4 The Proposed Framework

The proposed framework consists of three major components: (1) publisher credibility prediction; (2) user credibility prediction; and (3) fake news classification. Figure 1 illustrates the architecture of the proposed model.

4.1 Publisher Credibility Prediction

In recent years, the multi-head attention mechanism (Vaswani et al., 2017) shows the superior ability to learn the semantic representations of documents in the natural language process, which inspires us to extend it to learn node representations for graph representation learning. In this paper, we extend the Multi-head Attention (Vaswani et al., 2017) as a structure-aware multi-head attention module to encode the structure of the graph and learn the node representation from the publishing graph.

The structure-aware multi-head attention module has three input items: the query item, the key item and the value item, namely $Q \in \mathbb{R}^{n_q \times d}$, $K \in \mathbb{R}^{n_k \times d}$, and $V \in \mathbb{R}^{n_v \times d}$ respectively, where n_q, n_k, and n_v denote the number of nodes in each item, and d is the dimensionality of the node embedding. The

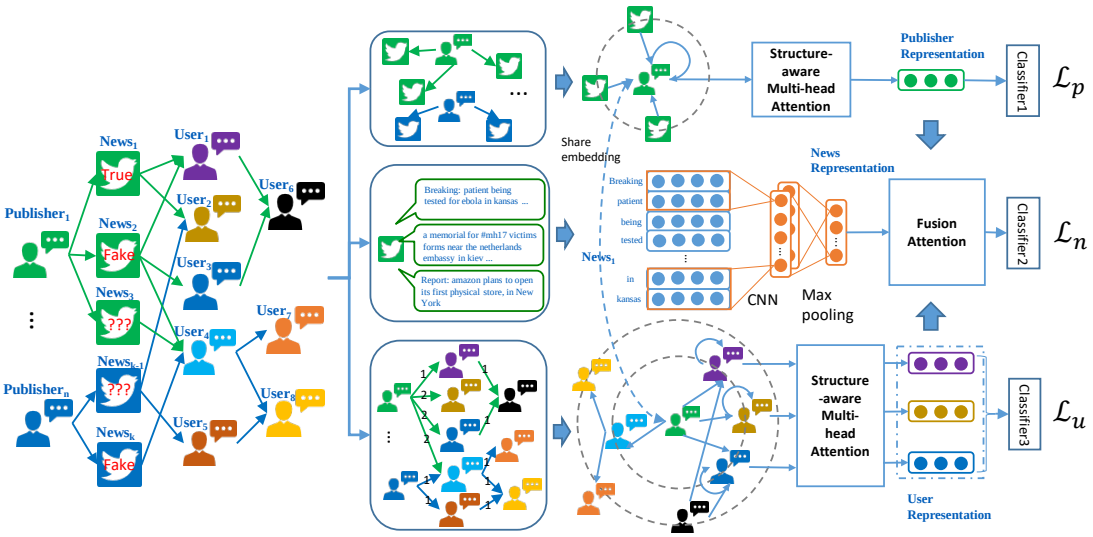


Figure 1: The architecture of the proposed fake news detection model.

attention module first takes each node in the query to attend to all nodes in the key item via a dot-product attention unit. But in fact, it is impossible for each node to establish connections with all nodes in the social graph. Thus, we encode the adjacent relations of the graph structure into the attention module. The adjacency matrix $\mathbf{A}^{pn} \in \mathbb{R}^{|P|\times|N|}$, whose element $\mathbf{A}^{pn}_{ij}$ denotes that publisher i deliver a piece of news j. Finally, we apply those attention weights upon the value item:

$$\mathbf{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})_h = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_h\mathbf{K}^T}{\sqrt{d}} \odot \mathbf{D}^{\mathbf{p}-\frac{1}{2}}\mathbf{A}^{pn}\mathbf{D}^{\mathbf{n}-\frac{1}{2}}\right)\mathbf{V}\,, \tag{1}$$

where $\mathbf{W}_h \in \mathbb{R}^{d\times d}$ is a transformation matrix. $\mathbf{D}^p_{ii} = \sum_j \mathbf{A}^{pn}_{ij}$ and $\mathbf{D}^n_{jj} = \sum_i \mathbf{A}^{pn}_{ij}$ are diagonal matrices, which are applied to normalize the adjacency matrix $\mathbf{A}^{pn}$. $\odot$ denotes element-wise product.

The entries of $\mathbf{V}$ are then linearly combined with the weights to form a new representation of $\mathbf{Q}$. In this way, the structure-aware attention module can capture relations across query nodes and key nodes, and further use the relations to aggregate embeddings in the query to produce new node representations. We usually let $\mathbf{K} = \mathbf{V}$. Therefore, every node in $\mathbf{Q}$ is represented by its most similar nodes in $\mathbf{V}$.

For each head of attention captures relations among $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ from one aspect, we expand one head attention to multi-head schema: $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are dispensed to h heads. Specifically, $\forall h \in [1, H]$ the output of head h is given by following formulation:

$$\mathbf{Z}_h = \mathbf{Attention}(\mathbf{P},\mathbf{N},\mathbf{N})_h\,, h \in [1, H] \tag{2}$$

where $\mathbf{P} \in \mathbb{R}^{|P|\times d}$ is the publishers' embeddings and $\mathbf{N} \in \mathbb{R}^{|N|\times d}$ is the news embeddings. H is the amount of heads in attention module. Every publisher and news is transformed into a d-dimensional embedding by their id and the vector is initialized by normal distribution (Glorot and Bengio, 2010).

Then, the output features of multi-head attention are concatenated together and a fully-connected layer is applied to transform it as final output, which is formalized as:

$$\tilde{\mathbf{P}} = \mathbf{ELU}\left([\mathbf{Z}_1;\mathbf{Z}_2;\ldots;\mathbf{Z}_H]\,\mathbf{W}_o\right) + \mathbf{P}\,, \tag{3}$$

where $\mathbf{W}_o \in \mathbb{R}^{Hd\times d}$ is a linear transformation matrix and $\mathbf{ELU}(x)$ is an activation function.

We obtain publishers' representations $\tilde{\mathbf{P}} \in \mathbb{R}^{|P|\times d}$ after above procedure. Finally, we use these features to predict the publishers' credibility, which can be formulated as follows:

$$p_i(c|\mathcal{G}(V_p, E), \mathcal{P}; \theta_1) = \mathbf{softmax}(\tilde{\mathbf{P}}_i\mathbf{W}_p + \mathbf{b}_p)\,, \tag{4}$$

where $\mathbf{b}_p$ is a bias term, and $\mathbf{W_p} \in \mathbb{R}^{d\times|c|}$ and $|c|$ is the total levels of credibility. The credit scores have three levels ($|c| = 3$): "unreliable", "uncertain", and "reliable". The annotation of credibility will

be introduced in Section 5.1. Finally, the publisher credibility prediction task can be transformed into a classification task.

We apply the cross-entropy loss as the optimization objective:

$$\mathcal{L}_p = -\sum_{i=1}^{|P|} y_i^{(p)} \log p_i(c|\mathcal{G}(V_p, E), \mathcal{P}; \theta_1) + \frac{\lambda}{2}||\theta_1||_2^2 \,, \tag{5}$$

where $y_i^{(p)}$ is the true credibility of publisher i and θ_1 denotes all parameters need to be trained in this subtask. We apply ℓ_2 regularization on all parameters of the model to overcome overfitting problem. λ is a regularization factor.

4.2 User Credibility Prediction

Same as publisher credibility prediction task, we apply user credibility as weakly supervised information to facilitate fake news detection. Firstly, we construct the diffusion graph of news $\mathcal{G}(V_u, E)$, which records how news propagated from publishers to other users. The nodes V_u of the graph belongs to the user set and the edges denote the diffusion traces.

Suppose that every news will be reposted by K different users at most. We use matrix $\mathbf{R} \in \mathbb{R}^{|U|\times K}$ to denote the user ids who had reposted the news before. '0' is padded at the start of the matrix $\mathbf{R}$ when the amount of reposted users is less than K. We still apply structure-aware multi-head attention to learn the user node representation from the diffusion graph. The attention unit is defined as follows:

$$\mathbf{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})_h = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_h\mathbf{K}^T}{\sqrt{d}} \odot \mathbf{D}^{\mathbf{u}-\frac{1}{2}}\mathbf{A}^{uu}\mathbf{D}^{\mathbf{u}-\frac{1}{2}}\right)\mathbf{V} \,, \tag{6}$$

where $\mathbf{W}_h \in \mathbb{R}^{d\times d}$ is a transformation matrix. $\mathbf{D}_{ii}^u = \sum_j \mathbf{A}_{ij}^{uu}$ is a diagonal matrix, which is used to normalize the adjacency matrix $\mathbf{A}^{uu}$. The complete computation process is shown in Algorithm 1.

Algorithm 1: The diffusion graph encoding algorithm.

Input:
1. Adjacency matrix $\mathbf{A}^{uu}$ of the diffusion graph $\mathcal{G}(V_u, E)$;
2. User embeddings $\mathbf{U} \in \mathbb{R}^{|U|\times d}$;
3. $\mathbf{Lookup}(\cdot)$ that can extract user embedding from $\mathbf{U}$ by user id.
4. Weight matrices $\mathbf{W}_o \in \mathbb{R}^{Hd\times d}$, $\mathbf{W}_h \in \mathbb{R}^{d\times d}$ and $h \in [1, 2, \ldots, H]$;
5. User ids matrix $R \in \mathbb{R}^{|U|\times K}$.

Output: User representations $\tilde{\mathbf{R}}$.

for $j \in [1, 2, \ldots, K]$ **do**
  for $h \in [1, 2, \ldots, H]$ **do**
    $\mathbf{R}_j = \mathbf{Lookup}(R_j)$;
    Calculate $\mathbf{Z}_h = \mathbf{Attention}(\mathbf{R}_j, \mathbf{U}, \mathbf{U})_h$ by Equation (6) ;
  end
  $\tilde{\mathbf{R}}_j = \mathbf{ELU}\left([\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_H]\,\mathbf{W}_o\right) + \mathbf{R}_j$
end
return $\tilde{\mathbf{R}} = [\tilde{\mathbf{R}}_1; \tilde{\mathbf{R}}_2; \ldots; \tilde{\mathbf{R}}_K]$

To learn abundant representations from different reposting relations, we extend structure-aware attention to employ a multi-head paradigm. Specifically, H independent attention units execute the transformation of Equation 6, and then their features are concatenated, resulting in the user representations.

Finally, we use these users' representations $\tilde{\mathbf{R}} \in \mathbb{R}^{|U|\times K\times d}$ to predict the users' credibility scores, which can be formulated as follows:

$$p_{ij}(c|\mathcal{G}(V_u, E), \mathcal{U}; \theta_2) = \mathbf{softmax}(\tilde{\mathbf{R}}_{ij}\mathbf{W}_r + \mathbf{b}_r) \,, \tag{7}$$

where $i \in [1, \ldots, |U|]$ and $j \in [1, \ldots, K]$. $\mathbf{W}_r \in \mathbb{R}^{d\times|c|}$ is a trainable matrix and $|c|$ is the levels of credibility. $\mathbf{b}_r \in \mathbb{R}^{|c|}$ is a bias term.

The credit scores of users are annotated in the same way as the credit scores of publishers. We apply the cross-entropy loss as the optimization function:

$$\mathcal{L}_u = -\sum_{i=1}^{|U|}\sum_{j=1}^{K} y_{ij}^{(u)} \log p_{ij}(c|\mathcal{G}(V_u, E), \mathcal{U}; \theta_2) + \frac{\lambda}{2}||\theta_2||_2^2 \,, \tag{8}$$

where $y_{ij}^{(u)}$ is the credibility of user u_{ij} and θ_2 denotes all parameters needed to be trained in this subtask.

4.3 Fake News Classification

For the fake news classification, we combine news with publishing and diffusion graph to more comprehensively capture the differences in the content and diffusion mode of true and false news.

4.3.1 News Content Representation

There have been many natural language processing models that can be used to learn the text representation from word sequence embeddings, such as CNN (Kim, 2014; Kalchbrenner et al., 2014) and RNN (Tai et al., 2015; Yang et al., 2016). For a fair comparison, we also apply CNN (Kim, 2014) as the basic component to learn the representation of news, which is the same as paper (Yuan et al., 2019).

4.3.2 Fusion Attention Unit

After content encoding, we have obtained news content representation $\mathbf{m}_j \in \mathbb{R}^{3d}$ for news m_j from word embeddings by CNN. Then, we will introduce how to fuse the publisher, user, and content representations for classification.

Firstly, we find publisher id p_i from the publishing and diffusion graph by news id m_j. Then, we look up publisher representation $\tilde{\mathbf{P}}_i \in \mathbb{R}^d$ from all publisher representations table $\tilde{\mathbf{P}}$ by publisher id p_i. And by the same way, we look up user representations $\tilde{\mathbf{R}}_i \in \mathbb{R}^{K\times d}$ from all user representations table $\tilde{\mathbf{R}}$ by publisher id p_i. $\tilde{\mathbf{R}}_i$ denotes K different users who had reposted the news m_j.

We aggregate the reposted user embeddings $\tilde{\mathbf{R}}_i \in \mathbb{R}^{K\times d}$ by an attention module:

$$\mathbf{R}' = \sum_{k=1}^{K} \alpha_k \tilde{\mathbf{R}}_k \,, \quad \alpha = \textbf{softmax}(\mathbf{N}_j \tilde{\mathbf{R}}_i^T) \,, \tag{9}$$

where $\mathbf{N}_j \in \mathbb{R}^{1\times d}$ is the embedding of news m_j looked up from the news embeddings table $\mathbf{N}$.

Then, we fuse the publisher representation and user combined representation by a heuristic method:

$$\tilde{\mathbf{m}}_j = [\tilde{\mathbf{P}}; \mathbf{R}'; \tilde{\mathbf{P}} \odot \mathbf{R}'; \tilde{\mathbf{P}} - \mathbf{R}']\mathbf{W}_F + \mathbf{b}_F \,, \tag{10}$$

where $\mathbf{W}_F \in \mathbb{R}^{4d\times d}$ is transformation matrix and $\mathbf{b}_F \in \mathbb{R}^d$ is a bias term.

News content representation captures the semantic difference between fake and true news. $\tilde{\mathbf{m}}_j$ captures the differences between fake and true news from the diffusion graph. Both representations are important for fake news detection, thus they are concatenated as final features. A fully-connected layer is applied to project the final representation into the target space of classes probability:

$$p(c|m_j, \mathcal{P}, \mathcal{N}, \mathcal{U}; \theta_3) = \textbf{softmax}([\mathbf{m}_j; \tilde{\mathbf{m}}_j]\mathbf{W}_m + b) \,, \tag{11}$$

where $\mathbf{W}_m \in \mathbb{R}^{4d\times|c|}$ is a transformation matrix and $b \in \mathbb{R}$ is a bias term.

Finally, the cross-entropy loss is used as the optimization objective function for fake news detection:

$$\mathcal{L}_n = -\sum_{j=1}^{|N|} y_j^{(n)} \log p(c|m_j, \mathcal{P}, \mathcal{N}, \mathcal{U}; \theta_3) + \frac{\lambda}{2}||\theta_3||_2^2 \,, \tag{12}$$

where $y_j^{(n)}$ is the gold class probability of news m_j.

For simultaneously optimize the credibility prediction task and fake news detection task, we combine all these optimization objective as follows:

$$\mathcal{L}\left(c|\mathcal{G}(V_p, E), \mathcal{G}(V_u, E), \mathcal{N}; \theta\right) = \mathcal{L}_p + \mathcal{L}_u + \mathcal{L}_n \,, \tag{13}$$

where $\theta = \{\theta_1, \theta_2, \theta_3\}$ represents all parameters of the model SMAN.

5 Experiments

In this section, we introduce the experiments to evaluate the effectiveness of SMAN. Specifically, we aim to answer the following evaluation questions:

- EQ1: Can SMAN improve fake news classification performance by jointly optimizing the fake news detection task and publishers' and users' credibility prediction task?
- EQ2: How effective are publishers' and users' credibility prediction tasks, respectively, in improving the detection performance of SMAN?
- EQ3: Can SMAN improve the performance of fake news early detection task?

5.1 Datasets

We evaluate SMAN on three real-world datasets: Twitter15 (Ma et al., 2017), Twitter16 (Ma et al., 2017), and Weibo (Ma et al., 2016). Table 1 shows the statistics of the three datasets.

Table 1: Dataset statistics. The label "true news" denotes a microblog that debunks the fake news.

Statistic	# news	# non-fake news (NR)	# fake news (FR)	# unverified news (UR)	# true news (TR)	# users	# retweets
Twitter15	1490	374	370	374	372	276,663	331,612
Twitter16	818	205	205	203	205	173,487	204,820
Weibo	4664	2351	2313	0	0	2,746,818	3,805,656

For a fair comparison, we use the train, validation, and test set that is split by (Yuan et al., 2019), where 10% samples as the validation dataset, and split the rest for training and test set with a ratio of 3:1.

The credit scores of publishers and users in these three datasets is annotated according to the training set. In this paper, we have defined three levels of credibility for publishers and users: (1) "0" means "reliable" (the publisher has never delivered fake or unverified news); (2) "1" means "uncertain" (the publisher not only delivers true news, but also publishes false news); (3) "2" means "unreliable" (publishers always publish false news and unverified news, but never publish true news).

5.2 Baseline Models

We compare our model with a series of fake news detection methods as follows:

(1) Feature-based methods: **DTC** (Castillo et al., 2011): A decision tree-based model that utilizes a combination of news characteristics. **SVM-RBF** (Yang et al., 2012): An SVM model with RBF kernel that utilize the news features. **SVM-TS** (Ma et al., 2015): An SVM model that utilizes time-series to model the variation of news characteristics. **DTR** (Zhao et al., 2015b): A decision-tree-based method for detecting fake news through enquiry phrases. **RFC** (Kwon et al., 2017): A random forest classifier that utilizes user, linguistic and structure features. **cPTK** (Ma et al., 2017): An SVM classifier with a propagation tree kernel that detects fake news by learning temporal-structure patterns.

(2) Deep Learning methods: **GRU** (Ma et al., 2016): A RNN-based model that learns temporal-linguistic patterns from user comments. **RvNN** (Ma et al., 2018): A bottom-up and a top-down tree-structured model based on recursive neural networks for fake news detection on Twitter. **PPC** (Liu and Wu, 2018): A model that detects fake news through propagation path classification with a combination of recurrent and convolutional networks. **GLAN** (Yuan et al., 2019): A model that jointly encodes the local semantic and global structural of the diffusion graph.

5.3 Evaluation Metrics and Parameter Settings

Same as previous studies (Liu and Wu, 2018; Ma et al., 2018; Yuan et al., 2019), we also adopt accuracy, precision, recall and F1 score as the evaluation metrics. The parameters of SMAN are updated by Adam algorithm (Reddi et al., 2018) with default parameters. All word embeddings of the model are initialized with the 300-dimensional word vectors, which is released by (Yuan et al., 2019). The convolutional kernel size is set to (3, 4, 5) with 100 kernels for each kind of size. The number of heads in structure-aware multi-head attention H is chosen from $\{1, 2, 3, \ldots, 11, 12\}$ and is set to 7. The λ in Equation (5), (8), (12) is chosen from $\{1e^{-8}, 1e^{-7}, \ldots, 1e^{-2}\}$ and is set to $1e^{-6}$. The source code will be released in the future.

5.4 Results and Analysis

To answer EQ1, we compare SMAN with baselines introduced in Section 5.2 for fake news classification. The experimental results of all baseline methods are shown in Table 2, 3, and 4. For fair comparison, the performance of baselines is directly cited from previous studies (Ma et al., 2018; Liu and Wu, 2018; Yuan et al., 2019). The **GLAN** model is the state-of-the-art method when submitting this paper.

Table 2: Experimental results on Twitter15 dataset.

Method	Accuracy	NR	FR	TR	UR
		F_1	F_1	F_1	F_1
DTR	0.409	0.501	0.311	0.364	0.473
DTC	0.454	0.733	0.355	0.317	0.415
RFC	0.565	0.810	0.422	0.401	0.543
SVM-RBF	0.318	0.455	0.037	0.218	0.225
SVM-TS	0.544	0.796	0.472	0.404	0.483
cPTK	0.750	0.804	0.698	0.765	0.733
GRU	0.646	0.792	0.574	0.608	0.592
RvNN	0.723	0.682	0.758	0.821	0.654
PPC	0.842	0.811	0.875	0.818	0.790
GLAN	0.905	**0.924**	0.917	0.852	0.927
SMAN	**0.929**	0.922	**0.945**	**0.915**	**0.933**

Table 3: Experimental results on Twitter16 dataset.

Method	Accuracy	NR	FR	TR	UR
		F_1	F_1	F_1	F_1
DTR	0.414	0.394	0.273	0.630	0.344
DTC	0.465	0.643	0.393	0.419	0.403
RFC	0.585	0.752	0.415	0.547	0.563
SVM-RBF	0.321	0.423	0.085	0.419	0.037
SVM-TS	0.574	0.755	0.420	0.571	0.526
cPTK	0.732	0.740	0.709	0.836	0.686
GRU	0.633	0.772	0.489	0.686	0.593
RvNN	0.737	0.662	0.743	0.835	0.708
PPC	0.863	0.820	0.898	0.843	0.837
GLAN	0.902	0.921	0.869	0.847	0.968
SMAN	**0.935**	**0.946**	**0.920**	**0.894**	**0.979**

Table 4: Fake news detection results on Weibo dataset.

Method	Acc	NR			FR		
		Precision	Recall	F_1	Precision	Recall	F_1
DTR	0.732	0.726	0.749	0.737	0.738	0.715	0.726
DTC	0.831	0.815	0.847	0.830	0.847	0.815	0.831
RFC	0.849	0.947	0.739	0.830	0.786	0.959	0.864
SVM-RBF	0.818	0.815	0.824	0.819	0.822	0.812	0.817
SVM-TS	0.857	0.878	0.830	0.857	0.839	0.885	0.861
GRU	0.910	**0.952**	0.864	0.906	0.876	0.956	0.914
PPC	0.921	0.949	0.889	0.918	0.896	**0.962**	0.923
GLAN	0.946	0.949	0.943	0.946	0.943	0.948	0.945
SMAN	**0.951**	0.937	**0.967**	**0.952**	**0.967**	0.936	**0.951**

We bold the best performance of each column in all tables. From the tables, we can observe that:

(1) Methods based on manually designed features (DTR, DTC, RFC, SVM-RBF, cPTK, and SVM-TS) have poorer performance. It indicates: 1) hand-crafted features cannot effectively encode semantic information of news content; 2) these methods cannot perform deep feature interaction; thus unable to fully learn the difference between fake and true news.

(2) Deep learning methods (GRU, RvNN, PPC, and GLAN) significantly outperform conventional classifiers that using manually designed features. This observation indicates deep learning models can learn better semantic representations and perform better feature interactions. We can also observe that GLAN is more effective than RvNN and PPC because it can deeply integrates local semantic and global diffusion structure for fake news detection.

(3) SMAN achieves significant improvement compared with GLAN. Different from GLAN, SMAN not only optimizes the fake news detection task but also tries to predict the credibility of publishers and users. The results show that the credibility of publishers and users is critical for learning the differences between fake and true news.

5.5 Ablation Study

To answer EQ2, we further perform some ablation studies over the different modules of SMAN. The experimental results are presented in Table 5.

Table 5: The ablation study results on the Twitter15, Twitter16, and Weibo datasets.

Models	Twitter15 Accuracy	Twitter16 Accuracy	Weibo Accuracy
SMAN_{base}	0.929	0.935	0.951
w/o Publisher Credibility (PC)	0.887	0.913	0.930
w/o User Credibility (UC)	0.905	0.880	0.938
w/o Publisher and User Credibility (PUC)	0.863	0.851	0.911

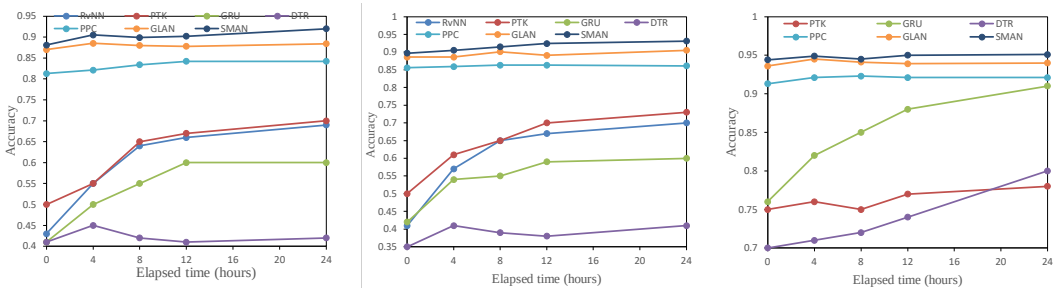
We first evaluate the impact brought by the publishers' credibility prediction subtask. We can observe that the performance drops a lot without PC. The publishers' credibility prediction subtask can exploit publishing relations between publishers and corresponding news to transfer the influence of publishers' credibility to news credibility, thus facilitating the fake news detection. The ablation results also prove it is very important to explicitly encode the credibility of publishers.

Then, we analyze the influence of the user credibility prediction subtask. We can observe that the absence of UC also causes significant performances to decline on all datasets. Intuitively speaking, if a piece of news is reposted by many low-reputation users, its credibility will indeed be greatly reduced. Same as PC task, the users' credibility also can be transferred to news credibility by diffusion graph, and thereby it can improve the detection performance.

Finally, we also find that the performance is much lower than the complete model SMAN after removing both publisher and user credibility prediction subtasks, which further proves that both tasks provide complementary information to each other. Thus, it is essential to jointly optimize the fake news detection and credibility prediction tasks.

5.6 Early Detection

For fake news detection task, one of the most essential targets is to detect fake news as soon as possible to intervene in time (Zhao et al., 2015b). To answer EQ3, we compared different methods of different time delays, and the performance is evaluated by the accuracy obtained when we incrementally add data up to the checkpoint given the targeted time delay.



(a) Twitter15 (b) Twitter16 (c) Weibo

Figure 2: Results of early fake news detection on the Twitter15, Twitter16 and Weibo dataset.

By changing the time delays, the accuracy of several competitive models is shown in Figure 2. In 0 to 4 hours, SMAN significantly outperforms the tree-based methods and feature-based methods and achieves better performance over the state of the art method, indicating the superior early detection performance of SMAN. Particularly, SMAN achieves about 91% accuracy on Twitter15 and Twitter16 datasets, and 95% accuracy on Weibo within 4 hours, which is much faster than most of the baselines.

After 8 hours, our model significantly surpasses the state of the art method. We can see that using more reposting relations will make the construction of the diffusion graph more complete and make the influence of credibility more easily transfer from publishers and users to news representations. Overall, the experimental results show that SMAN can not only improve the detection performance but also significantly reduce the time required for detection.

6 Conclusion

This paper proposes a novel structure-aware multi-attention network, which combines news content, the heterogeneous graphs among publishers and users, and jointly optimizes the task of false news detection and user credibility prediction for early fake news detection. Different from most existing research extracting hand-crafted features or deep learning methods, we explicitly treat the credibility of publishers and users as a kind of weakly supervised information for facilitating fake news detection. Extensive experiments conducted on three real-world datasets show that the proposed model can significantly surpass other state-of-the-art models on both fake news classification and early detection task.

Acknowledgements

We thank the anonymous reviewers for their feedback. This research is supported in part by the National Key Research and Development Program of China under Grant 2018YFC0806900.

References

Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM.

Yu-Ren Chen and Hsin-Hsi Chen. 2015. Opinion spam detection in web forum: a real case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 173–183.

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

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 943–951.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Thirtieth AAAI conference on artificial intelligence*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 655–665.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. *2013 IEEE 13th International Conference on Data Mining*.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PLOS ONE*, 12, 01.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179.

Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, pages 3818–3824.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. In *International Conference on Learning Representations*.

Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435. IEEE.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 312–320.

Changhe Song, Cunchao Tu, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Ced: Credible early detection of social media rumors. *arXiv preprint arXiv:1811.04175*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS '12*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *The 19th IEEE International Conference on Data Mining*. IEEE.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015a. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*, pages 1395–1405.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015b. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, WWW'15, pages 1395–1405. International World Wide Web Conferences Steering Committee.

Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1614–1623.