The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks

Yang Liu, Yi-Fang Brook Wu
New Jersey Institute of Technology
323 Dr Martin Luther King Jr Blvd
Newark, NJ

Abstract

In the midst of today's pervasive influence of social media, automatically detecting fake news is drawing significant attention from both the academic communities and the general public. Existing detection approaches rely on machine learning algorithms with a variety of news characteristics to detect fake news. However, such approaches have a major limitation on detecting fake news early, i.e., the information required for detecting fake news is often unavailable or inadequate at the early stage of news propagation. As a result, the accuracy of early detection of fake news is low. To address this limitation, in this paper, we propose a novel model for early detection of fake news on social media through classifying news propagation paths. We first model the propagation path of each news story as a multivariate time series in which each tuple is a numerical vector representing characteristics of a user who engaged in spreading the news. Then, we build a time series classifier that incorporates both recurrent and convolutional networks which capture the global and local variations of user characteristics along the propagation path respectively, to detect fake news. Experimental results on three real-world datasets demonstrate that our proposed model can detect fake news with accuracy 85% and 92% on Twitter and Sina Weibo respectively in 5 minutes after it starts to spread, which is significantly faster than state-of-the-art baselines.

Introduction

With the ever-increasing popularity of social media sites, user-generated messages can quickly reach a broad audience. Thus, social media has become an ideal place for fake news propagation. Fake news reaching a broad audience can cause elevated societal harm and economic damages and can also manipulate the outcome of political events. For example, during 2016 U.S. presidental election, the most discussed fake news stories tended to favor Donald Trump over Hillary Clinton (Silverman 2016). Thus, some commentators have suggested that Donald Trump would not have been elected president were it not for the influence of fake news (Allcott and Gentzkow 2017). Therefore, detecting fake news circulated on social media early in its propagation before it reaches a broad audience is highly desirable and socially beneficial.

Existing studies on automatically detecting fake news utilize machine learning algorithms that incorporate a variety of news characteristics on social media environments, e.g., text content, user characteristics, user comments, and propagation paths/trees or networks. A straightforward approach is to detect fake news based on its text content (Castillo, Mendoza, and Poblete 2011; Qazvinian et al. 2011; Takahashi and Igata 2012; Gupta et al. 2014; Popat 2017). However, these approaches have the following limitations. First, messages on popular social media sites, e.g., *Twitter*[1] and *Sina Weibo*[2] are short. Thus, the linguistic features extracted from them are often inadequate for machine learning algorithms to make accurate predictions. Second, these approaches can not be used to detect fake news that contains no text content but only a photo or a video. Another track of existing studies detect fake news through the characteristics of source users, i.e., users who first tweet the concerned news story on social media (Castillo, Mendoza, and Poblete 2011; Yang et al. 2012). However, these approaches ignore the characteristics of news spreaders, i.e., users who retweet the concerned news story, which can also be a discriminate clue about the truthfulness of the news story.

Recent studies have explored using temporal-linguistic features extracted from user comments (Zhao, Resnick, and Mei 2015; Ma et al. 2016; Ma, Gao, and Wong 2017) or temporal-structural features extracted from propagation paths/trees or networks (Jin et al. 2013; Wu, Yang, and Zhu 2015; Ma, Gao, and Wong 2017; Kwon, Cha, and Jung 2017) to detect fake news. Although these approaches are more effective at fake news detection than preliminary approaches that only adopt text content or source user characteristics, they have a significant limitation on efficiency, since temporal-linguistic and temporal-structural features are often unavailable or inadequate in the early stage of news propagation. For example, we observed that in the early stage of news propagation, most social media users tend to retweet the news story without adding any comment, and most users directly retweet the source tweet instead of retweeting someone else's retweet. As a result, both the temporal-linguistic and temporal-structural features are shallow which leads to a low accuracy of early detec-

[1] https://twitter.com
[2] https://weibo.com

tion of fake news. Another disadvantage of using temporal-linguistic features to detect fake news is that early user comments are easy to be manipulated. Fake news spreaders can provide fake comments on the fake news they spread. Compared to user comments, user characteristics are harder to be manipulated. Thus, we aim to design a new approach that can efficiently detect fake news based on user characteristics. There are also existing approaches that detect fake news based on a combination of different types of features (Castillo, Mendoza, and Poblete 2011; Yang et al. 2012; Sun et al. 2013; Ma et al. 2015; Kwon, Cha, and Jung 2017). One major limitation of these approaches is that they do not investigate which type of feature plays the most important role in detecting fake news, and if one or several types of feature is unavailable or inadequate in the early stage of news propagation, whether the effectiveness of these approaches will be affected.

To address the above-mentioned limitations of existing approaches, in this paper, we propose a novel approach for early detection of fake news on social media by classifying news propagation paths. We first model the propagation path of each news story as a multivariate time series, in which each tuple denotes the characteristics of a user who engaged in propagating the news. Then, we build a time series classifier with both recurrent and convolutional networks to predict whether a given news story is fake. Recurrent and convolutional networks can learn global and local variations of user characteristics respectively, which are discriminate clues for fake news detection. The main contributions of this paper can be summarized as follows:

- We are the first to model the propagation path of a news story on social media as a multivariate time series, e.g., a sequence of user characteristics, and the first to detect fake news through propagation path classification with a combination of recurrent and convolutional networks.
- We are the first to focus on improving the efficiency of early fake news detection while retaining comparable effectiveness as baseline approaches. Experimental results on three real-world datasets demonstrate that the proposed model can significantly improve the efficiency while slightly improving the effectiveness of early detection of fake news.
- The proposed model is more generalizable and robust in early detection of fake news since it only relies on common user characteristics which are more available, reliable and robust in the early stage of news propagation than linguistic and structural features widely-used by state-of-the-art approaches.

Related Works

Recent years, fake news (or rumor, misinformation) detection on social media has gained particular attention in the literature. A major track of existing studies aims at developing machine learning-based classifiers to automatically determine whether a news story spreading in a social media environment is fake based on a variety of news characteristics. A few early studies try to detect fake news based on linguistic features extracted from the text content of news stories. Castillo et al. (Castillo, Mendoza, and Poblete 2011) utilize a comprehensive set of linguistic features such as special characters, emoticon symbols, sentiment positive/negative words, hashtags, etc., to classify a news story as fake or true. Beyond those preliminary features, lexicon patterns and part-of-speech tags are explored in (Qazvinian et al. 2011). Named entities and clue keywords are adopted in (Takahashi and Igata 2012). Swear words and pronouns are examined in (Gupta et al. 2014). Language stylistic features, e.g., assertive verbs and factive verbs, are investigated in (Popat 2017). Besides text content, characteristics of source users have also been explored by several studies. Castillo et al. (Castillo, Mendoza, and Poblete 2011) utilize a set of user characteristics on Twitter, e.g., number of followers, number of friends, registration age to detect fake news. Yang et al. (Yang et al. 2012) explore a similar set of user characteristics on *Sina Weibo*, the most popular social media site in China.

A group of recent approaches utilizes temporal-linguistic features extracted from a sequence of user comments to detect fake news. Zhao et al. (Zhao, Resnick, and Mei 2015) detect fake news based on inquiry phrases from user comments. Ma et al. (Ma et al. 2016) utilize recurrent neural networks that capture temporal-linguistic features from a sequence of user comments to detect fake news. As an extension of Ma et al.'s approach, Chen et al. (Chen et al. 2017) incorporate a soft-attention mechanism into the recurrent neural networks to pool out distinct temporal-linguistic features with a particular focus.

Another group of recent approaches detects fake news based on temporal-structure features extracted from the propagation paths/trees or networks of news stories in social networks. Jin et al. (Jin et al. 2013) utilize epidemiological models to characterize information cascades in Twitter resulting from both true news and fake news. Wu et al. (Wu, Yang, and Zhu 2015) propose a graph kernel-based SVM classifier that learns high-order propagation patterns to detect fake news. Sampson et al. (Sampson et al. 2016) utilize implicit linkages between conversation fragments about a news story to predict its truthfulness. Ma et al. (Ma, Gao, and Wong 2017) propose a graph kernel-based SVM classifier which captures high-order patterns differentiating different types of fake news by evaluating the similarities between their propagation tree structures.

There are also hybrid approaches that combine different types of features to detect fake news. Castillo et al. (Castillo, Mendoza, and Poblete 2011) combine content-based, user-based, and propagation-based features to detect fake news. Yang et al. (Yang et al. 2012) combine content-based, user-based, location-based and client-based features. Sun et al. (Sun et al. 2013) combine content-based, user-based, and multimedia-based features. Ma et al. (Ma et al. 2015) combine the temporal variations of content-based, user-based, and diffusion-based features along the propagation timeline of news stories. Kwon et al. (Kwon, Cha, and Jung 2017) combine user, linguistic, structural and temporal features to detect fake news over varying time windows.

The problem of sequence/time series classification has been widely explored in the literature (Lesh, Zaki, and Ogihara 1999; Deshpande and Karypis 2002; Saigo et al. 2004;

Dong and Pei 2007). A recent study (Lee and Dernoncourt 2016) adopts recurrent and convolutional networks for sequential short-text classification. Inspired by this approach, in this paper we utilize a combination of recurrent and convolutional networks to classify news propagation paths to detect fake news.

Method

Problem Statement

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ be a set of news stories, $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$ be a set of social media users. Each user $u_j \in \mathcal{U}$ is associated with a *user vector* $\mathbf{x}_j \in \mathbb{R}^d$, which represents the characteristics of the user. We define the *propagation path* of a given news story a_i as a *variable-length multivariate time series* $\mathcal{P}(a_i) = \langle \ldots, (\mathbf{x}_j, t), \ldots \rangle$, in which each tuple $(\mathbf{x}_j, t)$ denotes that user u_j tweets/retweets the news story a_i at time t. In this paper, we set the time of a source tweet being posted to 0. Thus, $t > 0$ refers to the time of a retweet being posted. Each news story a_i is associated with a label $L(a_i)$ that reflects its truthfulness. Each label $L(a_i) \in \{0, 1\}^r$. When $r = 1$, $L(a_i) = 0$ denotes the news story a_i is true, and $L(a_i) = 1$ denotes a_i is fake. When $r > 1$, the label $L(a_i)$ is a categorical variable that reflects multiple levels of the truthfulness of the news story a_i, e.g., true, fake, or unverified, etc. Our goal is to design a model f that can predict the label of a given news story a_i based on its propagation path $\mathcal{P}(a_i)$, i.e., $\hat{L}(a_i) = f\big(\mathcal{P}(a_i)\big)$.

Since we aim to detect fake news as early as possible after it starts to spread, our model should be able to make predictions based on only a partial propagation path observed in the early stage of news propagation. We define the *partial propagation path* of a given news story a_i as $\mathcal{P}(a_i, T) = \langle (\mathbf{x}_j, t < T) \rangle$, where T is a *detection deadline* after which all the observed data cannot be used in detecting fake news. We call the task of predicting the truthfulness of news stories given partial propagation paths as *early detection of fake news*. In this case, we aim to design a model f_T that predicts the label of a given news story a_i based on its partial propagation path, i.e., $\hat{L}(a_i) = f_T\big(\mathcal{P}(a_i, T)\big)$.

The Proposed Model

The proposed fake news detection model consists of four major components, i.e., propagation path construction and transformation, RNN-based propagation path representation, CNN-based propagation path representation, and propagation path classification, which are integrated together to detect fake news at the early stage of its propagation. Fig. 1 shows the architecture of the proposed model. Next, we will introduce each of the major components.

Propagation Path Construction and Transformation Given a news story propagating on social media, we first construct its propagation path by first identifying the users who engaged in propagating the news. Then, its propagation path denoted as a *variable-length multivariate time series* $\mathcal{P}(a_i) = \langle \ldots, (\mathbf{x}_j, t), \ldots \rangle$ is constructed by extracting user characteristics from relevant user profiles. After $\mathcal{P}(a_i)$ is obtained, we transform it into a *fixed-length multivariate sequence*, denoted as $\mathcal{S}(a_i) = \langle \mathbf{x}_1, \ldots, \mathbf{x}_n \rangle$, where n is the length of the sequence. If there are more than n tuples in $\mathcal{P}(a_i)$, then $\mathcal{P}(a_i)$ will be truncated so that only the first n tuples will appear in $\mathcal{S}(a_i)$; If $\mathcal{P}(a_i)$ contains less than n tuples, then we randomly oversample tuples in $\mathcal{P}(a_i)$ to ensure the final length of $\mathcal{S}(a_i)$ equals n. Algorithm 1 shows the procedure of transforming a variable-length multivariate time series into a fixed-length multivariate sequence.

Algorithm 1 Algorithm for transforming a variable-length time series into a fixed-length sequence

Input: A variable-length time series $\mathcal{P}(a_i) = \langle \ldots, (\mathbf{x}_j, t), \ldots \rangle$, the length of the output fixed-length sequence n
Output: A fixed-length sequence $\mathcal{S}(a_i) = \langle \mathbf{x}_1, \ldots, \mathbf{x}_n \rangle$

```
if |P(a_i)| ≥ n then
    for (x_j, t) ∈ P(a_i)[1 : n] do
        S(a_i) ← S(a_i) ∪ ⟨x_j⟩
    end for
else
    d ← n − |P(a_i)|, s ← 0
    for (x_j, t) ∈ P(a_i) do
        c_1 ← UniformRandReal(0, 1)
        if c_1 > 0.5 then
            c_2 ← UniformRandInt(1, d − s)
        else
            c_2 ← 1
        end if
        s ← s + c_2
        for i ∈ [c_2] do
            S(a_i) ← S(a_i) ∪ ⟨x_j⟩
        end for
    end for
    if d − s > 0 then
        for i ∈ [d − s] do
            S(a_i) ← S(a_i) ∪ S(a_i)[s]
        end for
    end if
end if
```

RNN-Based Propagation Path Representation We utilize a variant of RNN called *Gated Recurrent Unit (GRU)* (Chung et al. 2014) to learn a vector representation for each transformed propagation path, i.e., $\mathcal{S}(a_i)$. For the t^{th} user vector in $\mathcal{S}(a_i)$, i.e., $\mathbf{x}_t$, a GRU unit takes as input $\mathbf{x}_t, \mathbf{h}_{t-1}$ and produces $\mathbf{h}_t$ as output according to the following formulas:

$$\begin{aligned} \mathbf{z}_t &= \sigma(U_z \mathbf{x}_t + W_z \mathbf{h}_{t-1}) \\ \mathbf{r}_t &= \sigma(U_r \mathbf{x}_t + W_r \mathbf{h}_{t-1}) \\ \tilde{\mathbf{h}}_t &= \tanh(U_h \mathbf{x}_t + \mathbf{h}_{t-1} \odot W_h \mathbf{r}_t) \\ \mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \end{aligned} \tag{1}$$

where $U_z, U_r, U_h \in \mathbb{R}^{m \times d}, W_z, W_r, W_h \in \mathbb{R}^{m \times m}$ are weight matrices, d is the dimension of the user vector $\mathbf{x}_t$, m is the output dimension of the GRU units. The symbols $\sigma(\cdot)$ and $\tanh(\cdot)$ denote the element-wise sigmoid and hyper-

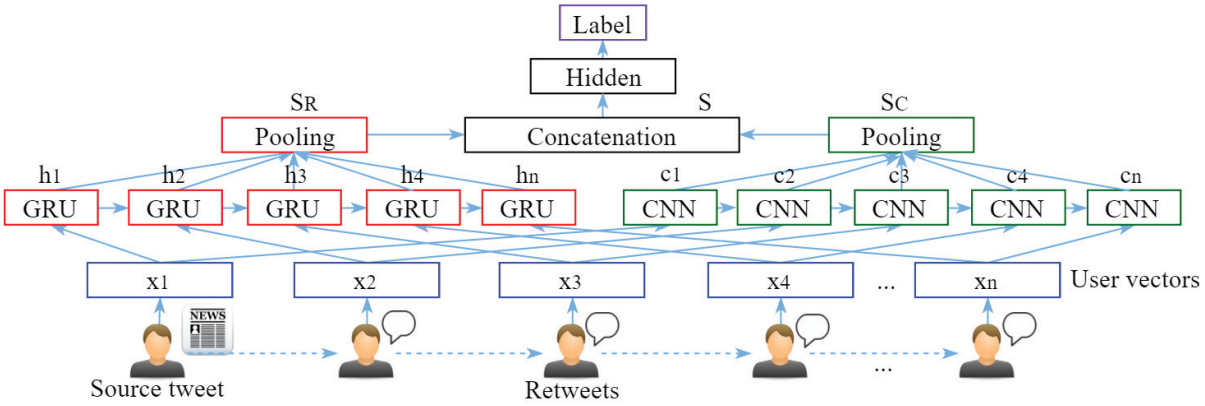


Figure 1: The architecture of the proposed fake news detection model

bolic tangent functions respectively, $\odot$ denotes the element-wise vector multiplication operation. $\mathbf{h}_0 = \mathbf{0}$. We then apply mean pooling to reduce the sequence of output vectors $\langle \mathbf{h}_1, \ldots, \mathbf{h}_n \rangle$ produced by GRU units into a single vector $\mathbf{s}_R = \frac{1}{n} \sum_{t=1}^{n} \mathbf{h}_t$, which is the final vector representation of $\mathcal{S}(a_i)$ that encodes the global variation of user characteristics.

CNN-Based Propagation Path Representation We also use convolutional networks (CNN) to learn another vector representation for each $\mathcal{S}(a_i)$. We first apply a 1-D convolution on h consecutive user vectors, i.e., $\langle \mathbf{x}_t, \ldots, \mathbf{x}_{t+h-1} \rangle$ with a *filter* $W_f \in \mathbb{R}^{h \times m}$ of height h, to produce a *scalar feature* $c_t \in \mathbb{R}$ according to the following formula:

$$c_t = \text{ReLU}(W_f \cdot X_{t:t+h-1} + b_f) \tag{2}$$

where $X_{t:t+h-1} \in \mathbb{R}^{h \times m}$ is the matrix whose i^{th} row is $\mathbf{x}_i$ and $b_f \in \mathbb{R}$ is a bias. The symbol $\text{ReLU}(\cdot)$ refer to the element-wise rectified linear unit function. We perform the same convolution operation with k filters to produce a multivariate feature vector $\mathbf{c}_t \in \mathbb{R}^k$. By repeating the same convolution operations for each window of h consecutive user vectors, we obtain a sequence of multivariate feature vectors, i.e., $\langle \mathbf{c}_1, \ldots, \mathbf{c}_{n-h+1} \rangle$. Then, we apply mean pooling to produce a final vector representation of $\mathcal{S}(a_i)$, i.e., $\mathbf{s}_C = \frac{1}{n} \sum_{t=1}^{n-h+1} \mathbf{c}_t$ that encodes the local variation of user characteristics.

Propagation Path Classification After $\mathbf{s}_R \in \mathbb{R}^m, \mathbf{s}_C \in \mathbb{R}^k$ are obtained through RNNs and CNNs respectively, they are concatenated into a single vector that represents the transformed propagation path, i.e. $\mathbf{s} \in \mathbb{R}^{m+k}$ by the following formula:

$$\mathbf{s} = \text{Concatenate}(\mathbf{s}_R, \mathbf{s}_C) \tag{3}$$

which is then fed into a multi-layer feedforward neural network that finally predicts the class label for the corresponding propagation path by the following formulas:

$$\begin{aligned} \mathbf{l}_j &= \text{ReLU}(W_j \mathbf{l}_{j-1} + \mathbf{b}_j), \ \forall j \in [q] \\ \mathbf{z} &= \text{Softmax}(\mathbf{l}_q) \end{aligned} \tag{4}$$

Table 1: Statistics of the datasets

Statistic	*Weibo*	*Twitter15*	*Twitter16*
# news stories	4664	1490	818
# true news	2351	374	205
# fake news	2313	370	205
# unverified news	0	374	203
# debunking of fake news	0	372	205
# users	2,746,818	276,663	173,487
# posts	3,805,656	331,612	204,820

where q is the number of hidden layers, $\mathbf{l}_j \in \mathbb{R}^{v_j}$ is the output of the j^{th} hidden layer ($\mathbf{l}_0 = \mathbf{s}$), v_j is the output dimension for the j^{th} hidden layer, $W_j \in \mathbb{R}^{v_j \times v_{j-1}}, \mathbf{b}_j \in \mathbb{R}^{v_j}$ are the weight matrix and bias for the j^{th} hidden layer respectively, $\mathbf{z} \in \mathbb{R}^r$ is the final output that represents the probability distribution over the set of r classes for the corresponding propagation path.

Experiments

Datasets

We evaluate the proposed model on three real-word data collections: *Weibo* (Ma et al. 2016), *Twitter15* (Ma, Gao, and Wong 2017) and *Twitter16* (Ma, Gao, and Wong 2017), which were originally collected from the most popular social media site in China and the U.S. respectively. In each dataset, source tweets are regarded as news stories. Propagation paths are constructed from the retweets of source tweets. The *Weibo* dataset only contains binary labels, i.e., "fake" and "true". The *Twitter15* and *Twitter16* dataset each contains four different labels, i.e., "fake", "true", "unverified", and "debunking of fake". Note that the label "debunking of fake" denotes a news story that tells people that a certain news story is fake. Table 1 shows the statistics of the three datasets. Following the same settings in the original papers that introduce these datasets, we randomly divide each dataset into three folds. We hold 10% of the dataset for validating our model and split the rest with ratio 3:1 for

Table 2: A summary of user characteristics in the three experimental datasets

No.	Feature	Type
1	LENGTH OF USER DESCRIPTION	Integer
2	LENGTH OF USERNAME	Integer
3	FOLLOWERS COUNT	Integer
4	FRIENDS COUNT	Integer
5	STAUTSES COUNT	Integer
6	REGISTRATION AGE	Integer
7	IS VERIFIED	Binary
8	IS GEO ENABLED	Binary

training and testing respectively. Since the original datasets do not include user information, we crawled all the related user profiles via *Twitter API*[3]. From the crawled user profiles, we extract eight common user characteristics that are available on both Twitter and Weibo, which are summarized in Table 2, to model news propagation paths as multivariate time series.

Baseline Models

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

- DTC (Castillo, Mendoza, and Poblete 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 a combination of news characteristics.
- SVM-TS (Ma et al. 2015) An SVM model that utilizes time-series to model the variation of news characteristics.
- DTR (Zhao, Resnick, and Mei 2015) A decision-tree-based ranking method for detecting fake news through enquiry phrases.
- GRU (Ma et al. 2016) An RNN-based model that learns temporal-linguistic patterns from user comments.
- RFC (Kwon, Cha, and Jung 2017) A random forest classifier that utilizes user, linguistic and structure characteristics.
- PTK (Ma, Gao, and Wong 2017) An SVM classifier with a propagation tree kernel that detects fake news by learning temporal-structure patterns from propagation trees.

We denote our proposed model as "PPC" (Propagation Path Classification), also as "PPC_RNN+CNN". We also implement two reduced version of the proposed model which only utilizes RNNs or CNNs respectively, denoted as "PPC_RNN" and "PPC_CNN" respectively.

Model configuration

We implemented our proposed model is using *Keras*[4]. The model is trained to minimize the binary/categorical loss function of predicting the class label of news stories in the training set. The weights and bias are updated using stochastic gradient descent with the Adadelta update rule (Zeiler 2012). Dropout (Srivastava et al. 2014) is applied on hidden layers above the concatenation layer to avoid overfitting. We set the number of training epochs to be 200. Early stopping is applied when the validation loss saturates for 10 epochs. The network structure are and hyperparameters are set based on the performance of our model on the validation set, which are show in Table 3.

Table 3: Model configuration

Hyperparameter	Choice	Experimental Range
GRU output dim	32	8 - 64
CNN # filters	32	8 - 64
CNN filter height	3	1 - 10
Dropout rate	0.5	0 - 1

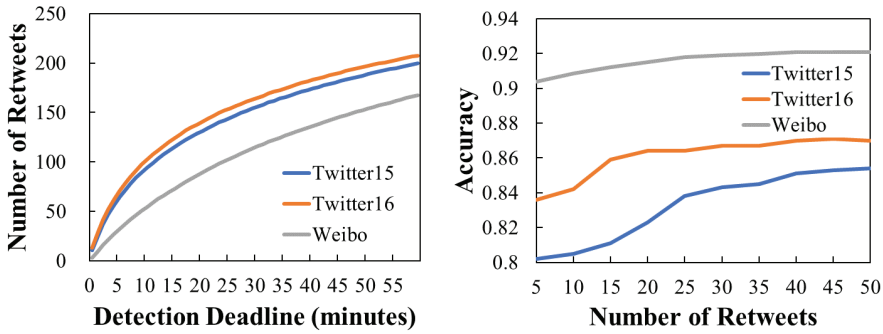
Note that the sequence length n used in Algorithm 1, which is also the number of source tweets plus the number of retweets we need to observe in a news propagation path to detect fake news, is related to both the effectiveness and efficiency of early detection of fake news. A longer sequence length might improve the effectiveness of fake news detection since more data will be observed. However the efficiency will be affected since it requires a longer time to observe a longer propagation path than a shorter one. On the other hand, a shorter sequence length improves the efficiency of early detection of fake news since we need less amount of data to make prediction. However, the effectiveness might be affected in this case. Therefore, we need to balance the trade-off between effectiveness and efficiency by choosing the most appropriate sequence length. Fig. 2 shows the speed of news propagation on social media and the speed of fake news detection conducted by our proposed model with both recurrent and convolutional networks. Fig. 2-(b) shows that the accuracy of our proposed model in detecting fake news saturates when the required number of retweets, i.e., the sequence length, is above 40 in the *Twitter15* and *Twitter16* datasets, and above 30 in the *Weibo* dataset respectively. Fig. 2-(a) shows that it requires about 5 minutes to observe 40 retweets in the *Twitter15* and *Twitter16* datasets and 30 retweets in the *Weibo* dataset. Therefore, when we observe more than 40 retweets on Twitter and more than 30 retweets on Weibo, our proposed model can detect fake news with accuracy around 92% and 85% on Twitter and Weibo respectively within five minutes after it starts to spread.

Results and Discussion

Table 4 and 5 show the performance of the proposed model and that of baseline models in the task of fake news detection on Twitter and Weibo dataset respectively. For most of the baseline models, their performance saturates when the detection deadline is above 24 hours. Therefore, to make a fair comparison, we set the detection deadline to be 24 hours here. We can find that the proposed models, i.e., PPC_RNN, PPC_CNN, and PPC_RNN+CNN outperform the baseline models. Among them, PPC_RNN+CNN performs the best. It achieves 84.2%, 86.3%, 92.1% accuracy on *Twitter15*,

[3]https://dev.twitter.com/rest/public

[4]https://keras.io/



(a) propagation speed

(b) detection speed

Figure 2: News propagation speed and fake news detection speed

Twitter16, and *Weibo* dataset respectively. Based on these results we can find that when observing relatively complete propagation paths, the proposed model outperforms the baseline models slightly in terms of effectiveness.

In the previous studies that introduce the peer models, a detection deadline of 24 hours is considered to be early. However, in this paper we aim to detect fake news as early as possible so that its harmful effects can be minimized. Therefore, we carefully investigate the performance of all the models in detecting fake news in less than 24 hours after it starts to spread. Fig. 3 shows the results of early detection of fake news. Among all the baseline models, we select three recent ones that focus on early detection of fake news, namely, DTR, GRU, and PTK. DTR and GRU rely on linguistic features extracted from user comments, while PTK relies on both linguistic and structural features extracted from propagation trees. We can find that when the detection deadline is less than 24 hours, the performance of the baseline models decreases significantly, while the performance of the proposed model is not affected since it only requires the first five minutes' data to make accurate predictions. Among the three baseline models, DTR yields the worst performance because the number of inquiry posts is usually very small in the early stage of news propagation. PTK yields a better performance than GRU because it utilizes temporal-structural features besides of temporal-linguistic features that are utilized by GRU.

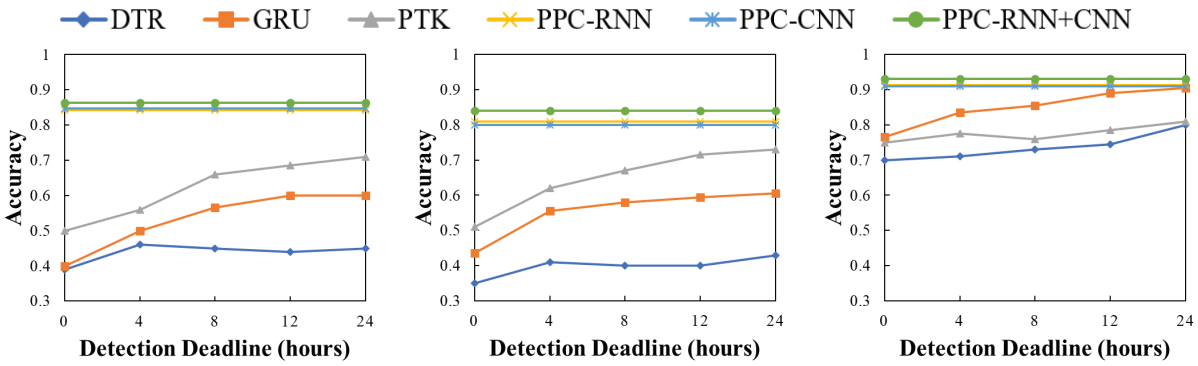
As pointed out by a recent study (Kwon, Cha, and Jung 2017), structural and temporal features are more useful for long-term detection of fake news but are often unavailable in the early stage of news propagation thus are less reliable for early detection of fake news; In contrast, user and linguistic features are more available in the early stage of news propagation thus are more reliable for early detection of fake news. Based on our empirical observations, we also found that linguistic features are less available than user characteristics in the very beginning of news propagation, e.g., in the first five minutes. Therefore, we assume that our model is more efficient at early detection of fake news than baseline models since it only relies on user characteristics. Experimental results on three real-world datasets demonstrate that the proposed model can significantly improve detection efficiency while slightly improving effectiveness of early detection of fake news. We also find that the two reduced models that only incorporates RNNs or CNNs respectively yield a similar accuracy which is higher than the accuracy of baseline models but lower than the accuracy of the complete proposed model that combines RNNs and CNNs. This demonstrates that both recurrent networks and convolutional networks can capture the global and local variations of user characteristics respectively. However, it is necessary to combine them to capture both the global and local variations of user characteristics to achieve the best performance of early detection of fake news.

Table 5: Fake news detection results on *Weibo* dataset when the detection deadline is 24 hours ("F":Fake News; "T": True News)

Method	Class	Acc.	Prec.	Recall	F_1
DTC	F	0.831	0.847	0.815	0.831
	T		0.815	0.847	0.830
SVM-RBF	F	0.818	0.822	0.812	0.817
	T		0.815	0.824	0.819
SVM-TS	F	0.857	0.839	0.885	0.861
	T		0.878	0.830	0.857
DTR	F	0.732	0.738	0.715	0.726
	T		0.726	0.749	0.737
GRU	F	0.910	0.876	0.956	0.914
	T		0.952	0.864	0.906
RFC	F	0.849	0.786	0.959	0.864
	T		0.947	0.739	0.830
PPC_RNN	F	0.912	0.878	0.958	0.916
	T		0.944	0.866	0.908
PPC_CNN	F	0.919	0.889	0.958	0.922
	T		0.946	0.880	0.916
PPC_ RNN+CNN	F	**0.921**	**0.896**	**0.962**	**0.923**
	T		**0.949**	**0.889**	**0.918**

Conclusion and Future Work

In this paper we propose a novel model for early detection of fake news on social media through classifying news propagation paths with both recurrent and convolutional net-



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

Figure 3: Results of early detection of fake news

Table 4: Fake news detection results on *Twitter15* and *Twitter16* datasets when the detection deadline is 24 hours ("T":True News; "F": Fake News; "U": Unverified News; "D": Debunking of Fake News)

Twitter15					
Method	Acc.	T	F	U	D
		F_1	F_1	F_1	F_1
DTC	0.454	0.733	0.355	0.317	0.415
SVM-RBF	0.318	0.455	0.037	0.218	0.225
SVM-TS	0.544	0.796	0.472	0.404	0.483
DTR	0.409	0.501	0.311	0.364	0.473
GRU	0.646	0.792	0.574	0.608	0.592
RFC	0.565	0.810	0.422	0.401	0.543
PTK	0.750	0.804	0.698	0.765	0.733
PPC_RNN	0.811	0.759	0.842	0.765	0.787
PPC_CNN	0.803	0.737	0.835	0.751	0.775
PPC_RNN+CNN	**0.842**	**0.811**	**0.875**	**0.790**	**0.818**

Twitter16					
Method	Acc.	T	F	U	D
		F_1	F_1	F_1	F_1
DTC	0.465	0.643	0.393	0.419	0.403
SVM-RBF	0.321	0.423	0.085	0.419	0.037
SVM-TS	0.574	0.755	0.420	0.571	0.526
DTR	0.414	0.394	0.273	0.630	0.344
GRU	0.633	0.772	0.489	0.686	0.593
RFC	0.585	0.752	0.415	0.547	0.563
PTK	0.732	0.740	0.709	0.836	0.686
PPC_RNN	0.842	0.809	0.865	0.836	0.839
PPC_CNN	0.847	0.812	0.871	0.833	0.841
PPC_RNN+CNN	**0.863**	**0.820**	**0.898**	**0.837**	**0.843**

works. After modeling the new propagation paths as multivariate time series of user characteristics, we apply recurrent and convolutional networks to capture both global and local variations of user characteristics along propagation paths to detect fake news. Experimental results on three real-word datasets demonstrate that our proposed model outperforms state-of-the-art fake news detection approaches in terms of both effectiveness and efficiency. Since our model only relies on common user characteristics which are more available, reliable and robust than complex features such as linguistic or structural features that are widely-used in state-of-the-art baseline approaches, it can detect fake news significantly faster than state-of-the-art baselines, e.g., in five minutes after the fake news starts to spread. Since we have found that user characteristics are discriminate indicators of the truthfulness of news stories, in the future, we plan to investigate whether user characteristics can help us identifying users who are easy to believe and spread fake news, and which features affect users' tendency to spread fake news most significantly, which are important problems in the prevention and debunking of fake news. Moreover, we will also incorporate semi-supervised learning and PU-learning (Li and Liu 2005) techniques into our model to deal with massive unlabeled news stories on social media.

References

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2):211–236.

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.

Chen, T.; Wu, L.; Li, X.; Zhang, J.; Yin, H.; and Wang, Y. 2017. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. *arXiv preprint arXiv:1704.05973*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Deshpande, M., and Karypis, G. 2002. Evaluation of techniques for classifying biological sequences. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 417–431. Springer.

Dong, G., and Pei, J. 2007. *Sequence data mining*, volume 33. Springer Science & Business Media.

Gupta, A.; Kumaraguru, P.; Castillo, C.; and Meier, P. 2014.

Tweetcred: Real-time credibility assessment of content on twitter. In *Proceedings of the International Conference on Social Informatics*, 228–243. Springer.

Jin, F.; Dougherty, E.; Saraf, P.; Cao, Y.; and Ramakrishnan, N. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, 8:1–8:9. ACM.

Kwon, S.; Cha, M.; and Jung, K. 2017. Rumor detection over varying time windows. *PloS one* 12(1):e0168344.

Lee, J. Y., and Dernoncourt, F. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Lesh, N.; Zaki, M. J.; and Ogihara, M. 1999. Mining features for sequence classification. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 342–346. ACM.

Li, X.-L., and Liu, B. 2005. Learning from positive and unlabeled examples with different data distributions. *Machine Learning: ECML 2005* 218–229.

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.

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

Ma, J.; Gao, W.; and Wong, K.-F. 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, 708–717.

Popat, K. 2017. Assessing the credibility of claims on the web. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 735–739. International World Wide Web Conferences Steering Committee.

Qazvinian, V.; Rosengren, E.; Radev, D.; and Mei, Q. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1589–1599. Association for Computational Linguistics.

Saigo, H.; Vert, J.-P.; Ueda, N.; and Akutsu, T. 2004. Protein homology detection using string alignment kernels. *Bioinformatics* 20(11):1682–1689.

Sampson, J.; Morstatter, F.; Wu, L.; and Liu, H. 2016. Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2377–2382. ACM.

Silverman, C. 2016. This analysis shows how fake election news stories outperformed real news on facebook. *BuzzFeed*.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15(1):1929–1958.

Sun, S.; Liu, H.; He, J.; and Du, X. 2013. Detecting event rumors on sina weibo automatically. In *Web Technologies and Applications*. Springer. 120–131.

Takahashi, T., and Igata, N. 2012. Rumor detection on twitter. In *Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligence Systems*, 452–457. IEEE.

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

Yang, F.; Liu, Y.; Yu, X.; and Yang, M. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 13:1–13:7. ACM.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhao, Z.; Resnick, P.; and Mei, Q. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, 1395–1405. International World Wide Web Conferences Steering Committee.