Fake News Detection using Temporal Features Extracted via Point Process

Taichi Murayama Shoko Wakamiya, Eiji Aramaki
Nara Institute of Science and Technology
murayama, wakamiya, aramaki@is.naist.jp

Abstract

Many people use social networking services (SNSs) to easily access various news. There are numerous ways to obtain and share "fake news," which are news carrying false information. To address fake news, several studies have been conducted for detecting fake news by using SNS-extracted features. In this study, we attempt to use temporal features generated from SNS posts by using a point process algorithm to identify fake news from real news. Temporal features in fake news detection have the advantage of robustness over existing features because it has minimal dependence on fake news propagators. Further, we propose a novel multi-modal attention-based method, which includes linguistic and user features alongside temporal features, for detecting fake news from SNS posts. Results obtained from three public datasets indicate that the proposed model achieves better performance compared to existing methods and demonstrate the effectiveness of temporal features for fake news detection.

Introduction

Social networking services (SNSs), such as Facebook and Twitter, provide many people with instant and convenient access to news. However, SNSs constitute an effective platform for obtaining and sharing news that are not carefully fact-checked and may include false or uncertain information, called "fake news." (Sharma et al. 2019) define "fake news" as "a news article or message published and propagated through media, carrying false information regardless the means and motives behind it." In our paper, the same definition is used.

The wide spread of fake news cannot only harm social media platforms but society in general. For example, during the US 2016 presidential election, fake news favoring different candidate were shared more than 37 million times on SNSs and strongly affected the election results (Budak 2019; Bovet and Makse 2019). Consequently, the unprecedented growth of fake news reflects a strong need for detecting and mitigating fake news circulation (Lazer et al. 2018). To confront these societal challenges, websites such as Snopes.com[1] and PolitiFact.com[2] track and debunk rumors and manually assess rumor credibility based on evidence. These fact-checking sites are expensive to operate legitimately and require a considerable amount of time to validate and publish the credibility of a rumor. Contrary to fact-checking websites, existing work on fake news detection mainly applies machine-learning methods based on various characteristics of SNSs, e.g., text content (Ma et al. 2016), user characteristics (Liu and Wu 2018) and propagation paths/trees (Wu, Yang, and Q. Zhu 2015).

In addition to existing features, it is assumed that the temporal movements of SNS posts are also useful for detecting fake news (Kai et al. 2020). Recent research (Shao et al. 2018) showed that social bots influence the initial spread of fake news. Time series of posts referring to fake news exhibit different movement from those of real news. Nevertheless, few studies have considered the amount of attention fake news attract over time.

This study proposes a fake news detection model that takes advantage of the attention to news changing over time, i.e., the temporal features. The attention is calculated using a self-exciting point process from the post publication time and the likelihood of people reading the post (determined by the number followers). In this study, we designate the attention to the news as an "infectiousness value" because it can be measured based on the probability of re-share of the information by each new user. The infectiousness value can be regarded as an index of the public interest in the news and, for real news, it normally decreases over time. Conversely, our underlying assumption is that the infectiousness value of fake news upsurges twice: the first upsurge results from the original news (including the false information), and the second results from news items for which people doubt or correct the false information.

The infectiousness value of the information is more robust than that of existing features, which depend on fake news propagators. For example, text features of early users can be easily manipulated by providing fake comments for diffusion. User features and user-article relationship are being transformed by the regulation of platforms and account sus-

[1] https://www.snopes.com/
[2] https://www.politifact.com/

pension. Propagation paths/trees are difficult to manipulate but it is expensive to obtain them. Infectiousness values are also difficult to manipulate because the values are calculated from a series of posts, not by early movement. Furthermore, the number of followers and post publication time, which are used for calculating the infectiousness values, can be easily obtained.

The proposed fake news detection model leverages three features: combing existing features, texts, and users with an Attention-based mechanism and implementing the infectiousness value. As preliminary research, we investigate whether temporal features can distinguish real news from fake news to validate their effectiveness. Then, experiments are carried out to demonstrate that each module, such as the temporal features, is useful for detecting fake news.

The contributions of this study are as follows. (1) We elucidate the differences of infectiousness values associated with real and fake news and consider the differences for fake news detection using a point process. (2) We propose a new multi-modal method that combines text and user features with infectiousness values. (3) We show the effectiveness of the proposed model for fake news detection on SNSs through experimental procedures.

Related Work

Early studies attempted to detect fake news based on linguistic features extracted from texts (Castillo, Mendoza, and Poblete 2011; Derczynski et al. 2017). Recent studies used deep learning models to capture temporal–linguistic features. (Ma et al. 2016) used recurrent neural networks (RNNs), which capture temporal–linguistic features from a bag-of-words of user posts. (Ma, Gao, and Wong 2018) used recursive neural networks based on the texts of a reply tree. Further examples include convolutional neural networks (Yu et al. 2017), hierarchical attention networks (Ma et al. 2019), and neural-network models using discourse-level structures (Karimi and Tang 2019).

Moreover, several methods were examined for detecting fake news using the characteristics of users who post the information. In fact, (Castillo, Mendoza, and Poblete 2011; Yang et al. 2012; Liu and Wu 2018) used various models based on user characteristics, such as the number of followers, number of friends, and registered age. Recently, the relationship between news articles and users is used to determine news credibility assuming that if a strong relation exists between two articles as determined by the number of users who re-shared them, the two articles are likely to share the same label (Nguyen et al. 2019).

Other studies employ detection methods based on propagation paths/trees or networks of posts on SNS. (Ma, Gao, and Wong 2017) proposed a graph-kernel-based support vector machine (SVM) classifier that calculates the similarity between propagation tree structures.

Multi-modal approaches combine features of different types to detect fake news. For example, (Ruchansky, Seo, and Liu 2017) combined texts and user behavior, while (Wang et al. 2018) combined texts and visual features extracted from SNS posts. Our model effectively combines

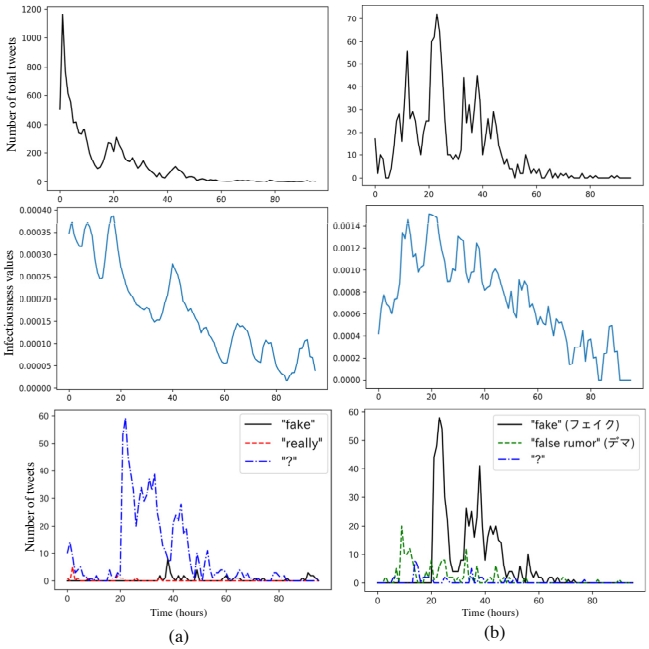

Figure 1: Time series of two fake news items regarding: (a) Islamic news in the U.S. and (b) Okinawa's news in Japan extracted from Twitter API for a 96-hour observation period. The time series show the number of total tweets in the upper panels, the infectiousness values calculated using a self-exciting point process in the middle panels, and the number of tweets which include words used for doubts and denials ("fake," "really" and "?" as words for the U.S. and "fake," "false rumor" and "?" as words of Japan), in the bottom panels. The result of (a) indicates few explicit words such as "fake," but the question mark which represents doubt appeared many times in the same timing as the second upsurge around 20 hours. The result of (b) shows that explicit words indicating news as fake/false, appeared after the first upsurge at around 22 hours.

text and user features using contextual inter-modal attention (Ghosal et al. 2018) to catch the relationship between a user and a post content.

In a method of fake news detection using temporal features similar to the proposed method, (Lukasik et al. 2016) demonstrated the importance of using post temporal information for rumor stance classification. (Kwon, Cha, and Jung 2017) used SpikeM (Matsubara et al. 2012) to mathematically capture the time series behavior of information for long-term rumor detection, in addition to using other features (e.g., linguistic, user, network). In this study, we demonstrate that temporal features are also useful for short-term fake news detection. The proposed multi-modal framework utilizes linguistic, user, and temporal features, which are easy to obtain, to capture the characteristics of fake news.

Preliminary Research

We validated the contribution of temporal features in SNS posts to judge whether the news are fake or real (not fake). Using Twitter API [3], we obtained real and fake news items

[3] https://developer.twitter.com

published in 2019 in the U.S. and Japan. Additionally, for the U.S. news, we collected posts about fake news from the URLs and keywords, as extracted from Snopes.com and PolitiFact.com articles. Because Japan has no major fact-checking websites, for the Japan news, we collected posts about fake news from major media, public organizations, and companies denied in Japan. We also collected real news from the URLs of news articles by major media.

Figure 1 presents the time series of the two fake news examples in the U.S. (a) and Japan (b). Each news item has three time series. The upper one indicates the number of tweets on each hour and the middle one indicates the infectiousness values calculated using the self-exciting point process described in the "Fake News Detection Model" section, which represents the probability of re-share. It is thought that the time series of the number of tweets about real news shows a large upsurge in a few hours but decays quickly over time (Zhao et al. 2015). Contrarily, the time series of the number of tweets about fake news (see upper panels) shows a second upsurge after approximately a day following the unstable behavior in the infectiousness value of fake news. These behaviors are observed in other fake news and other countries. An earlier study (Kwon, Cha, and Jung 2017) indicated that the time series of rumors have multiple upsurges during long-term observation periods (56 days), unlike those of non-rumors. In contrast, our results demonstrate that time series of fake news have multiple upsurges in short-term observations (4 days), unlike those of real news. The time series graphs and a description of the collected news are presented in the URL [4].

We assumed that the multiple upsurges in the time series of fake news are caused by the attention received by posts questioning or denying the news. To test this hypothesis, we examined whether the second upsurge coincides with the increase of posts expressing doubt and denial. As shown in the bottom panels of Figure 1, posts expressing doubt and denial appeared multiple times after the first upsurge within 48 hours. For example, the result of (a) indicates few explicit words such as "fake," but the question mark which represents doubt appeared many times in the same timing as the second upsurge around 20 hours. The result of (b) shows that explicit words indicating news as fake/false, appeared around 22 hours. These results support our assumption and are mostly in agreement with a previous study (Shao et al. 2016), which indicated a characteristic time lag between fake news and fact-checking. Additionally, we have inferred that the multiple upsurges related to fake news are caused by renewed public interest because the meaning of news changes after questioning or denial (Fig. 1). The differences between the time series of fake and real news suggest that temporal features, which are more difficult to manipulate than others, can be useful for detecting fake news.

Fake News Detection Model

Although temporal features are useful, fake news detection using temporal features alone cannot achieve sufficient performance. Consequently, we propose a novel multi-modal method to detect fake news from many SNS posts. The proposed model effectively combines linguistic and user features using an Attention module and then implements temporal features. The overall model architecture is presented in Figure 2.

[4]https://docs.google.com/document/d/193Xv0AqmHB1F-UuaRuXpZOeMtfjNMnNrmTBUTjkoFIw/edit?usp=sharing

Table 1: Major notations

Notation	Definition
A_i	i_{th} news story
a_t	t_{th} post of news story A_i
$\mathbf{l}_t$	linguistic feature of t_{th} post
$\mathbf{u}_t$	user feature of t_{th} post
$\mathbf{s}_i$	temporal features of i_{th} news story
$\mathbf{s}^h$	infectiousness values at each point
$\mathbf{l}_t^e$	I^l-dimensional post embedding of t_{th} post
$\tilde{\mathbf{h}}_t^*, \mathbf{h}_t^*$	hidden state of the t_{th} post through GRU and FC in each module
$\mathbf{h}^{max\_*}$	each hidden states through MaxPooling
$\mathbf{z}$	the final output representing class probability
H^*	each module output consisting of a sequence $[\mathbf{h}_t^*]$
T^*	Number of sequence lengths of each module
E^*	Number of dimensions about hidden states $\mathbf{h}_t^*$ in each module

Problem Statement

The task of fake news detection is the prediction of the news label (real or fake), given the SNS posts related to the news. Let A_i be a news story consisting of N_i posts; $A_i = \{a_1, a_2, ..., a_{N_i}\}$. Each post $a_t = (\mathbf{l}_t, \mathbf{u}_t)$ consists of a linguistic feature $\mathbf{l}_t$ and a user feature $\mathbf{u}_t$. The temporal features of a news story are represented as $\mathbf{s}_i$. Additionally, each news story A_i is associated with a label $L(A_i)$, which has categorical variables $\{0, 1\}^\tau$. We aim to learn a fake news detection function $f : f(A_i, \mathbf{s}_i) \rightarrow L(A_i)$ that maximizes the prediction accuracy.

Model Structure

The model comprises various components. The linguistic, user, and temporal modules convert inputs to latent features. The contextual inter-modal attention module combines the latent features generated by the linguistic and user modules with attention. Finally, the classification module outputs the prediction. Table 1 represents the major notations.

Linguistic module We first converted the raw text of each post a_t to the linguistic feature $\mathbf{l}_t$ for model interpretation. Then, we used the tf-idf values of the vocabulary terms of each post. We used the top-K vocabularies according to their tf-idf values. Therefore, for each post, we extracted the linguistic feature $\mathbf{l}_t \in \mathbb{R}^K$, which is a K-dimensional vector. The linguistic feature $\mathbf{l}_t$ created from the post corresponds to sparse high-dimensional data. Therefore, we convert the vector $\mathbf{l}_t$ into a low-dimensional representation. Instead of using pre-trained vectors based on external collections, we learn the embedding matrix through our model;

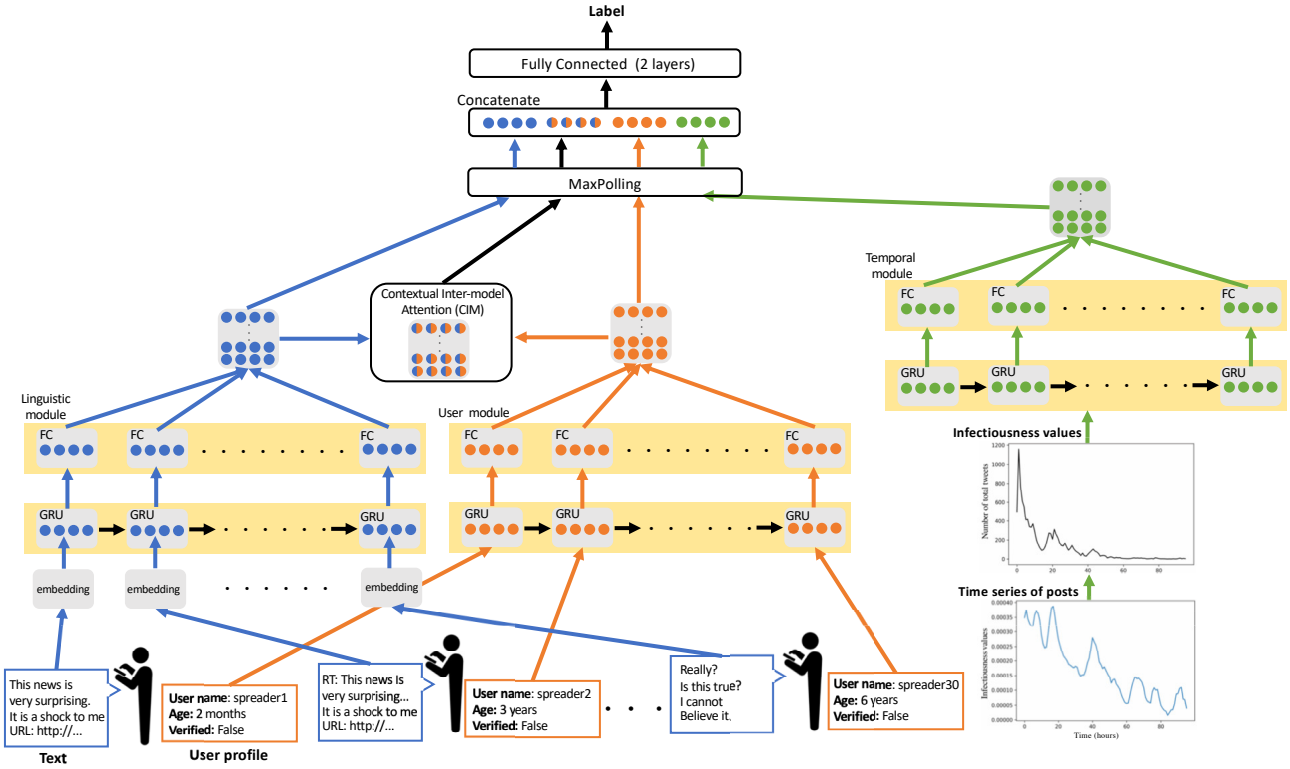

Figure 2: Architecture of the proposed fake news detection model. GRUs are used to learn the latent representations of linguistic, user, and temporal features. Then, CIM combines the linguistic and user features. Finally, the label of news is predicted by concatenating these features.

$\mathbf{l}_t^e = Embedding(\mathbf{l}_t)$, where $\mathbf{l}_t^e \in \mathbb{R}^{I^l}$ denotes the I^l-dimensional post embedding vector of $\mathbf{l}_t$.

From each post consisting of a sequence of embedded posts $L_i^e = [\mathbf{l}_1^e, \mathbf{l}_2^e, ..., \mathbf{l}_{T^l}^e]$, we extract the latent linguistic features to use gated recurrent units (Cho et al. 2014) (GRUs). Actually, GRUs based on an RNN can capture long-term dependencies to learn the temporal–linguistic features of early posts on SNS. A GRU takes $\mathbf{l}_t^e$ and $\tilde{\mathbf{h}}_{t-1}$ as input and produces $\tilde{\mathbf{h}}_t$ as output. The respective formulas are described below:

$$\mathbf{z}_t^l = \sigma\left(U_z^l \mathbf{l}_t^e + W_z^l \tilde{\mathbf{h}}_{t-1}^l\right), \ \mathbf{f}_t^l = tanh\left(U_h^l \mathbf{l}_t^e + \tilde{\mathbf{h}}_{t-1}^l \odot W_h^l \mathbf{r}_t\right)$$
$$\mathbf{r}_t^l = \sigma\left(U_r^l \mathbf{l}_t^e + W_r^l \tilde{\mathbf{h}}_{t-1}^l\right), \ \tilde{\mathbf{h}}_t^l = \left(1 - \mathbf{z}_t^l\right) \odot \tilde{\mathbf{h}}_{t-1}^l + \mathbf{z}_t^l \odot \mathbf{f}_t^l$$

where $\mathbf{z}_t^l$ and $\mathbf{r}_t^l$ represent the reset and update gate at time t, respectively. Furthermore, $U_z^l, U_r^l, U_h^l \in \mathbb{R}^{I^l \times E^l}$, $W_z^l, W_r^l, W_h^l \in \mathbb{R}^{E^l \times E^l}$ are parameters used for the respective gates. E^l denotes the output dimension of the GRU. Then, the hidden state $\tilde{\mathbf{h}}_t^l$ of the GRU is applied by the fully connected (FC) layer, resulting in $\mathbf{h}_t^l \in \mathbb{R}^{E^l}$, as shown below:

$$\tilde{\mathbf{h}}_t^l = GRU\left(\mathbf{l}_t^e\right), \quad \mathbf{h}_t^l = FC(\tilde{\mathbf{h}}_t^l), \quad t \in \{1, ..., T^l\} \tag{1}$$

User module We used eight common characteristics extracted from SNS user profiles as the user features; the length of user description, length of user name, number of followers, number of follows, number of posts, registration age, and whether verified mark and geo information are attached to the account. These are similar to (Liu and Wu 2018). The eight common features for a post a_t are represented by $\mathbf{u}_t \in \mathbb{R}^{I^u}$. As with the linguistic features, we use GRUs to capture long-term dependencies and FC layers for the user features, as shown below:

$$\tilde{\mathbf{h}}_t^u = GRU(\mathbf{u}_t), \quad \mathbf{h}_t^u = FC(\tilde{\mathbf{h}}_t^u), \quad t \in \{1, ..., T^u\} \tag{2}$$

Temporal module In the previous section, we described the differences between the appearance time of posts about real and fake news. To capture the potential components of this behavior, we convert the time series of posts to infectiousness values, which represent the re-share probability and drop as the news gets stale, via a self-exciting point process model (designated as SEISMIC) (Zhao et al. 2015). SEISMIC, based on the Hawkes process (Hawkes 1971), calculates the infectiousness value s_t^h at time t using the number of posts R_t until time t and the intensity λ_t. s_t^h is the input of the GRUs in the temporal module.

$$\lambda_t = s_t^h \sum_{t_i \le t, i \ge 0} n_i \phi\left(t - t_i\right), \quad t \ge t_0. \tag{3}$$

$$\phi\left(s\right) = \begin{cases} c & if \quad 0 < s \le s_0, \\ c(s/s_0)^{-(1+\theta)} & if \quad s > s_0, \end{cases} \tag{4}$$

where n_i represents the number of people accessing the news (number of followers). Additionally, $\phi(\cdot)$ denotes the

memory kernel, which quantifies the delay between the arrival and re-share of a post by a user. These parameters are estimated by (Zhao et al. 2015): s_0 is 5 min, θ is 0.242, and $c = 6.27 \times 10^{-4}$. This process is designated as *self-exciting* because each previous observation i contributes to the intensity λ_t.

The estimation of the temporal variance of s_t^h relies on a sequence of one-sided kernels $K_t(s)$, which up-weights the most recent posts and down-weights older posts. These one-sided kernels keep the estimator s_t^h close to the ever-changing real values.

$$s_t^h = \frac{\sum_{i=1}^{R_t} K_t(t - t_i)}{\sum_{i=0}^{R_t} n_i \int_{t_i}^{t} K_t(t-s)\,\phi(s - t_i)\,ds} \tag{5}$$

$$K_t(s) = \max\left\{1 - \frac{2s}{t}, 0\right\}, \quad s > 0. \tag{6}$$

Eqs. (5) and (6) are used to calculate the infectiousness values s_t^h from the publication time and number of followers of each post up to time t. As described herein, $\mathbf{s}_i = \{..., (time_t, follower_t), ...\}, t \in \{1, ..., N\}$ is the input of the function to convert to the infectiousness values, where $time_t$ represents the time elapsed from the first post. Then, $\mathbf{s}_i$ is converted to the infectiousness values $\mathbf{s}^h = \left(s_{t_1}^h, s_{t_2}^h, ..., s_{t_{T^s}}^h\right)$ at each point, e.g., every hour. As with the linguistic and user features, we utilize the GRUs and FC layers for the temporal features that are converted from every post information $\mathbf{s}_i$, as explained below:

$$\tilde{\mathbf{h}}_t^s = GRU(s_t^h), \quad \mathbf{h}_t^s = FC(\tilde{\mathbf{h}}_t^s) \quad t \in \{1, ..., T^s\} \tag{7}$$

Contextual Inter-model Attention Each post comprises linguistic and user features, which often have mutual interdependence. However, GRUs are unable to capture characteristics of their interdependence. Therefore, we used a pairwise contextual inter-modal attention mechanism (designated as CIM) (Ghosal et al. 2018), using the latent representations generated by the GRUs.

We compute the attention between the output of the linguistic features $H^l = [\mathbf{h}_1^l, \mathbf{h}_2^l, ..., \mathbf{h}_{T^l}^l] \in \mathbb{R}^{T^l \times E^l}$ and that of user features $H^u = [\mathbf{h}_1^u, \mathbf{h}_2^u, ..., \mathbf{h}_{T^u}^u] \in \mathbb{R}^{T^u \times E^u}$ to leverage the contextual information related to each post to detect fake news, where $E^l \equiv E^u$ and $T^l \equiv T^u$. First, a pair of matching matrices $M_1, M_2 \in \mathbb{R}^{T^l \times T^u}$ are computed as $M_1 = H^l \cdot H^{u\mathsf{T}}, M_2 = H^u \cdot H^{l\mathsf{T}}$.

Furthermore, we obtained the probability distribution scores $N_1, N_2 \in \mathbb{R}^{T^l \times T^u}$ over the respective matching matrices M_1 and M_2 to compute the attention weights on contextual posts using a softmax function. Then, we computed the modality-wise attentive representations.

$$\begin{aligned} N_1(i,j) &= \frac{e^{M_1(i,j)}}{\sum_{k=1}^{T^l} e^{M_1(i,k)}}, \quad for\ i, j = 1, ..., T^l \\ N_2(i,j) &= \frac{e^{M_2(i,j)}}{\sum_{k=1}^{T^l} e^{M_2(i,k)}}, \quad for\ i, j = 1, ..., T^l \\ O_1 &= N_1 \cdot H^u, \quad O_2 = N_2 \cdot H^l \end{aligned} \tag{8}$$

Finally, we computed the element-wise matrix multiplication for the attention to the important components. Then, we concatenated the calculation values A_1 and A_2 to obtain the attention representations between H^l and H^u.

$$\begin{aligned} A_1 &= O_1 \odot H^l, \quad A_2 = O_2 \odot H^u \\ H^{ul} &= concat\,[A_1, A_2] \in \mathbb{R}^{T^l \times 2E^l} \end{aligned} \tag{9}$$

Classification module After obtaining the features through the modules, we applied them to MaxPooling and concatenated each feature into a single vector $\mathbf{f}^1 \in \mathbb{R}^{E^l + E^u + 2E^l + E^s}$,

$$\mathbf{f}^1 = concat[\mathbf{h}^{max\_l}, \mathbf{h}^{max\_u}, \mathbf{h}^{max\_ul}, \mathbf{h}^{max\_s}] \tag{10}$$

where $\mathbf{h}^{max\_*}$ indicates hidden states H^* through MaxPooling, i.e., $\mathbf{h}^{max\_l} = MaxPooling\left(H^l\right)$.

For predicting the class label for each news item, we used FC layers with an activation function, such as $ReLU$ that consists of two layers, to identify the complex relations between the respective features. The final output $\mathbf{z} \in \mathbb{R}^\tau$ represents the probability distribution over the set of τ classes through the softmax function.

$$\mathbf{f}^2 = ReLU\left(FC\left(\mathbf{f}^1\right)\right), \quad \mathbf{z} = Softmax\left(FC\left(\mathbf{f}^2\right)\right) \tag{11}$$

Experimental Procedure

Datasets

To experimentally evaluate our model, we used three publicly available datasets: Weibo released by (Ma et al. 2016), and Twitter15 and Twitter16 released by (Ma, Gao, and Wong 2017). Each dataset of posts related to fake news was collected from the most popular social media platforms, i.e., Weibo [5] in China and Twitter [6] in the U.S. The Weibo dataset is annotated with one of two class labels: "true" or "fake." The Twitter datasets are annotated with one of four class labels: "true," "fake," "unverified" or "debunking of fake." Table 2 presents a summary of the datasets. It should be noted that the dataset size is smaller at the time of release because some SNS stories and posts cannot be acquired owing to changes in disclosure statements and post deletion.

For the experiments, we divided each dataset into training, validation, and test sets. Each dataset was split following a ratio of 3:1 for acquiring the training and test sets, respectively. A 15% of the training set was held for the validation set.

Comparative Methods

We made comparisons between the proposed model and the following existing baseline methods of fake news detection.

- **SVM-TS** (Ma et al. 2015): A linear SVM classifier that uses time-series to model the variation of social context features. This model also uses diffusion-based features, such as the average number of re-shares, in addition to linguistic and user features.

[5] https://www.weibo.com

[6] https://twitter.com

Table 2: Summary of datasets

Dataset	Weibo	Twitter15	Twitter16
No. of true news	2351	371	204
No. of fake news	2313	363	205
No. of unverified news	-	373	205
No. of debunking	-	372	199
No. of training posts	2973	942	517
No. of validation posts	525	167	97
No. of test posts	1166	370	204

- **CSI** (Ruchansky, Seo, and Liu 2017): CSI is a hybrid deep-learning model that uses information from user texts, responses, and behaviors. This model calculates the source characteristic based on the user behavior, and classifies an article as fake or not.
- **GRU-2** (Ma et al. 2016): GRU-2 is equipped with two GRU hidden layers and an embedding layer following the input layer for learning rumor representations by modeling the sequential structure of relevant posts over time.
- **PPC** (Liu and Wu 2018): PPC is a time series classifier that incorporates both recurrent and convolutional networks, which respectively capture user characteristics along the propagation path.
- **Proposed (w/o CIM)**: This is the proposed model without the contextual inter-modal attention module used for validating the effectiveness of CIM.
- **Proposed (w/o time)**: This model comprises two features for learning; i.e., it uses linguistic and user features for validating the effectiveness of the temporal features.
- **Proposed (freq)**: This model replaces the infectiousness values with the number of posts during each period for validating the effectiveness of the infectiousness values.

Experimental Settings

Our model has been trained to minimize the binary/categorical loss function while predicting the class label of each news item in the training set. During training, all model parameters were updated using gradient-based methods following the AdaDelta update rule. Additionally, Dropout, for which the value was set to 0.5, was applied on hidden layers $\tilde{\mathbf{h}}_t^*, \mathbf{h}_t^*, \mathbf{f}^1$, and $\mathbf{f}^2$ to avoid overfitting. The number of training epochs was set to 500. Early stopping was applied as the validation loss saturated for 10 epochs.

The network structure and hyper-parameters were set based on the validation set and on previous studies (Ma et al. 2016; Liu and Wu 2018). We set 5,000 vocabularies as top-K based on the tf-idf values as input to the linguistic module. These tf-idf values were converted to embedding vectors with a dimension I^l of 100. I^u was set to eight, as described in the user module in the "Fake News Detection Model" section. The sequence lengths of the GRUs for the linguistic and user features, T^l and T^u, were chosen as above 30 in the Weibo dataset and above 40 in the Twitter15 and Twitter16 datasets, based on the results of a previous study (Liu and Wu 2018). Namely, we used the first 30 or 40 posts in a story time sorted in ascending order as the input of T^l and T^u.

In the case study, most time series of the number of fake news posts showed a second upsurge after approximately one day after post publication. Therefore, we set the infectiousness values on the first two days with a length T^s of 47 as the input of the GRUs for the temporal features $\mathbf{s}^h$. These 47 infectiousness values were calculated using all data from the point publication time up to at each hourly point; i.e., $s^h_{t_3}$ is calculated by all posts up to 3 hours elapsed from the post publication.

The output size of each GRU (E^l, E^u, and E^s) is selected from (16, 32, 64, and 128) and the hidden dimension of the output FC layer $\mathbf{f}^2$ is selected from (E^{con}, $\frac{E^{con}}{2}$, $\frac{E^{con}}{4}$, and $\frac{E^{con}}{8}$) in the validation period, where E^{con} is the size of $\mathbf{f}^1$, equal to $(E^l + E^u + 2E^l + E^s)$.

We used the accuracy and F1-measure as metrics to evaluate the model capabilities. Classification tasks, such as fake news detection, are commonly evaluated by the accuracy while F1-measure works complementary to address class imbalance. We used the accuracy over all categories and the F1-measure for each class to evaluate the model performance.

Results and Discussion

The experimental results are presented in Table 3 and indicate that the proposed model outperforms most baseline methods, confirming the benefits of the multi-modal method and temporal features. The baseline **SVM-TS**, based on hand-crafted features, was a better model because it combined various features, including linguistic, user, and temporal features. Contrarily, **CSI** achieved low accuracy. The model calculates the user relation score from the training data and then detects fake news from the test data by using the scores of users who appear in both training and test data. Because few users appeared in both the training and test datasets in our experiments, CSI performed poorly. Most deep learning-based models, such as the **Proposed model**, **GRU-2**, and **PPC**, outperformed feature engineering-based models, such as **SVM-TS**. Deep neural networks helped to learn better hidden representations of people's responses to the news on SNS for fake news detection. The results show that **GRU-2** and **PPC**, which used linguistic and user features, respectively, to capture complex hidden features indicative of the corresponding responses, achieved a high accuracy and high F1-measure.

To validate the effectiveness of each module, we also conducted experiments using models that excluded CIM and the temporal features of the proposed model. Compared to **Proposed (w/o CIM)**, **Proposed model** achieved a higher accuracy and F1-measure on all datasets, except for the unverified label data. This result demonstrates that it was insufficient to learn the hidden representations of the user and linguistic features differently. Moreover, inter-dependencies between the linguistic and user features were useful to detect whether a news item was fake or not because posts consist of both features. Compared to **Proposed (w/o time)**, **Proposed**

Table 3: Fake news detection results on each dataset

Dataset	Weibo			Twitter15					Twitter16				
Method.	Acc.	F_1		Acc.	F_1				Acc.	F_1			
		T	F		T	F	U	D		T	F	U	D
SVM-TS	0.827	0.831	0.837	0.599	0.772	0.598	0.608	0.544	0.574	0.743	0.488	0.551	0.549
CSI	0.780	0.750	0.803	0.556	0.601	0.631	0.550	0.530	0.507	0.552	0.511	0.475	0.443
GRU-2	0.876	0.872	0.879	0.794	0.822	0.815	0.849	0.697	0.750	0.761	0.750	**0.771**	0.723
PPC	0.914	0.912	0.917	0.806	0.748	0.840	0.807	0.730	0.778	0.803	0.760	0.711	0.767
Proposed (w/o CIM)	0.920	0.922	0.917	0.814	0.807	0.813	**0.870**	0.745	0.791	0.850	0.782	0.747	0.791
Proposed (w/o time)	0.912	0.913	0.910	0.814	0.857	0.806	0.868	0.677	0.791	0.864	0.829	0.717	0.776
Proposed (freq)	0.921	0.931	0.908	0.807	0.872	0.815	0.828	0.660	0.805	0.864	0.801	0.740	0.699
Proposed model	**0.937**	**0.937**	**0.936**	**0.831**	**0.880**	**0.850**	0.833	**0.758**	**0.819**	**0.870**	**0.831**	0.739	**0.841**

model achieved higher scores, except for the unverified label data in Twitter15. In a previous study (Kwon, Cha, and Jung 2017), the time series of rumors is useful to detect rumors in long-term observation periods (56 days). However, these results support our claims that temporal features can be useful for short-term fake news detection (2 days). **Proposed (freq)** replaced the infectiousness values with the number of posts in each period for validating the conversion to the infectiousness values. Its accuracy was slightly higher than that of Proposed (w/o time) for the Weibo and Twitter16 datasets when the number of posts was added. Simultaneously, the degree of increased accuracy was not significantly higher than that of the Proposed model. This result shows that conversion to infectiousness values is useful to catch latent information from the temporal features for the fake news detection.

Proposed model overall performed the best for most measures and datasets, demonstrating the effectiveness of our model compared to baseline methods. Specifically, our model achieved the highest accuracy for the Weibo test subset (0.937), the Twitter15 test subset (0.831), and the Twitter16 test subset (0.819). Additionally, our model achieved the highest performance in terms of the F1 score on the True, Fake, and Debunking news data labels. However, similarly to the compared methods, our model did not produce good results for classifying unverified labels. Presumably, effectively classifying ambiguous labels, such as unverified, is challenging even when implementing the temporal features.

Finally, we evaluated the details of the contributions of the temporal features. To examine the contributions, we compared the proposed models with varying time frames to obtain the temporal features from zero (w/o time) over six days (see Figure 3). The accuracy of the proposed model improves gradually as the time frame lengthens. However, the proposed model performance remains more or less unchanged for a time frame over three days. Specifically, the model accounting for five days of the Weibo dataset achieved an accuracy of 0.939. When accounting for 4 days of the Twitter15 and Twitter16 datasets, our model achieved an accuracy of 0.867 and 0.830, respectively. Although we set the time frame to the first 2 days in the experimental settings, the results show that time periods of approximately 4 or 5 days would be more appropriate for obtaining the temporal features for fake news detection.

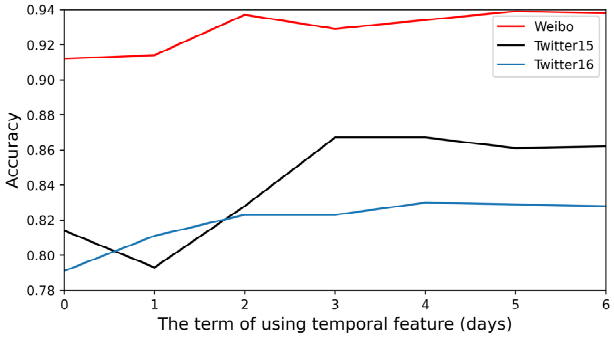

Figure 3: Accuracy of the proposed model with temporal features obtained for different time frames for each dataset: the horizontal axis represents the time frames from 0 (w/o time) to 6 days; the vertical axis represents the accuracy. It is observed that better accuracy is achieved when a longer time frame is used.

Although our model demonstrates that incorporating temporal features, which are difficult to manipulate, in fake news detection models is useful, limitations also exist; it is difficult to detect "early" fake news. The comparative method PPC claims to achieve fake news detection within an hour. However, it is difficult to accurately estimate the infectiousness values s_t^h of the information within an hour, so our model is not suitable for detecting early fake news. Therefore, our results suggest the use of different models depending on the circumstances; models without temporal features are better for early detection, while the proposed model with temporal features are better for robust and high-precision detection.

Conclusion

We conclude this paper by highlighting the key points of our study: (1) We ascertained the differences in time series behaviors between real and fake news from short-term observations. (2) We proposed a novel multi-modal method for fake news detection, combining text and user features and infectiousness values. (3) The experimental results empirically showed the effectiveness of the proposed model for the fake news detection problems. However, it remains unclear whether the temporal features are useful in ambiguously la-

beled data (e.g., debunking label). Future studies must examine how temporal features can be used flexibly effectively classifying ambiguous data labels.

References

Bovet, A., and Makse, H. A. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications* 10(1):7.

Budak, C. 2019. What happened? the spread of fake news publisher content during the 2016 us presidential election. In *The World Wide Web Conference*, 139–150.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proc. of the WWW*, 675–684.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; and Zubiaga, A. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proc. of the SemEval-2017*, 69–76.

Ghosal, D.; Akhtar, M. S.; Chauhan, D.; Poria, S.; Ekbal, A.; and Bhattacharyya, P. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proc. of the EMNLP*, 3454–3466.

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58(1):83–90.

Kai, S.; Deepak, M.; Suhang, W.; and Huan, L. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proc. of the ICWSM*.

Karimi, H., and Tang, J. 2019. Learning hierarchical discourse-level structure for fake news detection. In *Proc. of the NAACL*, 3432–3442.

Kwon, S.; Cha, M.; and Jung, K. 2017. Rumor detection over varying time windows. *PLOS ONE* 12(1):1–19.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Liu, Y., and Wu, Y.-F. B. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *AAAI Conference on Artificial Intelligence*, 354–361.

Lukasik, M.; Srijith, P.; Vu, D.; Bontcheva, K.; Zubiaga, A.; and Cohn, T. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proc. of the ACL*, volume 2, 393–398.

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 *Proc. of the CIKM*, 1751–1754.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, J.; Wong, K.-F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proc. of the IJCAI*, 3818–3824.

Ma, J.; Gao, W.; Joty, S.; and Wong, K.-F. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proc. of the NAACL*, 1391–1400.

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proc. of the ACL*, 708–717.

Ma, J.; Gao, W.; and Wong, K.-F. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proc. of the ACL*, 1980–1989.

Matsubara, Y.; Sakurai, Y.; Prakash, B. A.; Li, L.; and Faloutsos, C. 2012. Rise and fall patterns of information diffusion: model and implications. In *Proc. of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 6–14.

Nguyen, D. M.; Do, T. H.; Calderbank, R.; and Deligiannis, N. 2019. Fake news detection using deep Markov random fields. In *Proc. of the NAACL*, 1391–1400.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. Csi: A hybrid deep model for fake news detection. In *Proc. of the CIKM*, 797–806.

Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2016. Hoaxy: A platform for tracking online misinformation. In *Proc. of the WWW*, 745–750. International World Wide Web Conferences Steering Committee.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature communications* 9(1):4787.

Sharma, K.; Qian, F.; Jiang, H.; Ruchansky, N.; Zhang, M.; and Liu, Y. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM TIST* 10(3):21.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proc. of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 849–857. ACM.

Wu, K.; Yang, S.; and Q. Zhu, K. 2015. False rumors detection on sina weibo by propagation structures. *Proc. of the ICDM* 2015:651–662.

Yang, F.; Liu, Y.; Yu, X.; and Yang, M. 2012. Automatic detection of rumor on sina weibo. In *Proc. of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, 13:1–13:7.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2017. A convolutional approach for misinformation identification. In *Proc. of the IJCAI*, 3901–3907.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522.