Exploiting Emotions for Fake News Detection on Social Media

Chuan Guo[1,3], **Juan Cao**[1,3*], **Xueyao Zhang**[1,3], **Kai Shu**[4] and **Miao Yu**[2]
[1]Key Laboratory of Intelligent Information Processing Center for Advanced Computing Research, Institute of Computing Technology, Chinese Academy of Sciences
[2]Institute of Information Engineering, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]Department of Computer Science and Engineering, Arizona State University
{guochuan, caojuan}@ict.ac.cn, xueyao_98@foxmail.com, kai.shu@asu.com, yumaio@iie.ac.cn

Abstract

Microblog has become a popular platform for people to post, share, and seek information due to its convenience and low cost. However, it also facilitates the generation and propagation of fake news, which could cause detrimental societal consequences. Detecting fake news on microblogs is important for societal good. Emotion is a significant indicator while verifying information on social media. . Existing fake news detection studies utilize emotion mainly through users stances or simple statistical emotional features; and exploiting the emotion information from both news content and user comments is also limited. In the realistic scenarios, to impress the audience and spread extensively, the *publishers* typically either post a tweet with intense emotion which could easily resonate with the crowd, or post a controversial statement unemotionally but aim to evoke intense emotion among the *users*. Therefore, in this paper, we study the novel problem of exploiting emotion information for fake news detection. We propose a new Emotion-based Fake News Detection framework (EFN), which can i) learn content- and comment-emotion representations for publishers and users respectively; and ii) exploit content and social emotions simultaneously for fake news detection. Experimental results on real-world dataset demonstrate the effectiveness of the proposed framework.

1 Introduction

Social media platforms play a crucial role for people to seek out and spread information, especially in emergencies and breaking news. However, the convenience of publishing and spreading information also foster the wide propagation of *fake news*, commonly referred as intentional false information [Shu *et al.*, 2017]. For instance, an authoritative analysis of BuzzFeed News[1] indicated that, during the 2016 U.S. presidential election campaign, top 20 fake news stories generated more total engagement on Facebook than top 20 major real news, that these fake news earned nearly 9 million shares

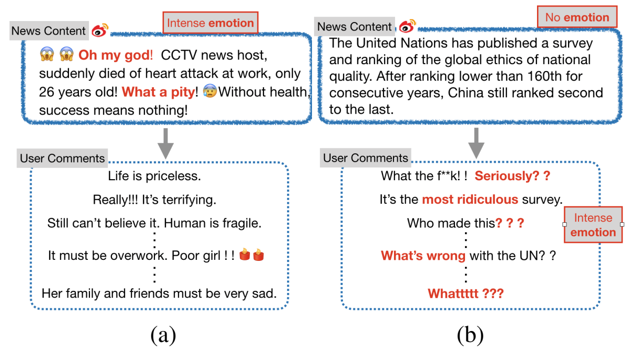


Figure 1: Two fake news posts from Sina Weibo. (a) a post which contains emotions of astonishment and sadness in **news contents** that easily arouse the audience. (b) a post which contains no emotion, but raise emotions like doubt and anger in **user comments** by controversial topics.

on social media. These fakes news seriously do harm to not only the public credibility, but also social stability and economic market. Therefore, it's significantly important to build tools to detect the fake news automatically and effectively.

Existing works on fake news detection mainly focus on *news content* and *social context*. Feature-based classification models extract basic semantic and emotion features from content, and statistical features from users [Castillo *et al.*, 2011]. Propagation-based model construct the relationship network inside the event, and incorporate social conflicting viewpoint towards the event in the network [Jin *et al.*, 2014; 2016]. Recently, deep learning models are proposed to evaluate the credibility of information on social media, which deeply exploits the semantics information from news content[Ma *et al.*, 2016]. Basic social context features are fused into deep learning models in some studies[Guo *et al.*, 2018]. However, emotion information, which is crucial for fake news detection, is underutilized in these studies. Fewer studies leverage emotion in news contents and social context simultaneously for fake news detection.

Fake news publishers often aim to spread information extensively and draw wide public attention. Longstanding social science studies demonstrate that the news which evokes high-arousal, or activating (awe, anger or anxiety) emotions is more viral on social media[Stieglitz and Dang-Xuan, 2013; Ferrara and Yang, 2015]. To achieve this goal, fake news publishers commonly adopt two approaches. First, publishers post news with intense emotions which trigger a high level

*Contact Author

[1]https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

of physiological arousal in the crowd. For example, in Figure 1a, the publisher uses rich emotional expressions (e.g., "Oh my god!") to make this information more impressive and striking. Second, publishers may present the news objectively to make it convincing whose content, however, is controversial which evoke intense emotion in the public, and finally spreads widely. As another example (see Figure 1b), the publisher writes the post in a unemotional way; while, the statement that China ranks second to the last suddenly bring on tension in the crowd, and people express their feeling of anger (e.g., "most ridiculous"), shock and doubt (e.g., "seriously?") in comments. Therefore, learning emotion of the publisher and the users corporately has the potential to improve fake news detection performance.

To exploit emotion information for fake news detection, we first define two types of emotions: (1) ***publisher emotion***: emotion of the publisher while posting information on social media; and (2) ***social emotion***: emotion of users when the information disseminates on social media. In essence, we investigate: (i) how to capture signals of *publisher emotion* and *social emotion* from news content and user comments, respectively; (ii) how to exploit publisher and social emotions simultaneously for fake news detection. Our solutions to these two challenges results in a novel Emotion-based Fake News Detection framework (EFN). Our main contributions are summarized as follows:

- We provide a principled way to capture the *publisher emotion* and *social emotion* signals, and demonstrate the importance of these two emotions from various perspectives on fake news detection.
- We propose a novel framework EFN, which exploits a deep neural network to learn representations from *publisher emotion*, *social emotion* and content simultaneously, for fake news detection.
- We conduct experiments on real world datasets to show the effectiveness of EFN for fake news detection.

2 Related Work

We briefly describe the related work from three-folds: i) Fake News Detection; ii) Emotion Representation; and iii) Multimodal Fusion.

2.1 Fake News Detection

Previous fake news detection studies mostly focus on extracting features and training a classifier to predict the credibility of news. [Castillo *et al.*, 2011] manually extracts a wide range of features including user features, content features, propagation features and topic features. [Yang *et al.*, 2012] demonstrates the effectiveness of location and client; Besides feature-based models, propagation-based approaches aim to mine the relations between various entities in a event. [Gupta *et al.*, 2012] firstly introduces propagation network in credibility evaluation by constructing a relation network. [Jin *et al.*, 2014; 2016] applies similar hierarchical structure on microblogs which consists of news, sub-events and messages. Recently, neural network models are adopted for fake news detection. [Ma *et al.*, 2016] firstly applies RNN for fake news detection on social media, modeling the posts in a event as a sequential time series. [Guo *et al.*, 2018] proposes a social attention network to capture the hierarchical characteristic of events on microblogs.

So far, emotion in these works is used as either emotion feature sets or viewpoints of users, which requires more systematical and comprehensive explorations in future work.

2.2 Emotion Representation

Early studies primarily use hand-crafted features for representing emotion of text, which highly rely on sentiment dictionaries. There are several widely-used emotion dictionaries, including WordNet [Kamps *et al.*, 2004] and MPQA [Wiebe *et al.*, 2005]for English, and HowNet [2] for Chinese. However, this method may encounter problems of emotion migration and low coverage on social media, because of the differences of word usage on social media and in real word.

Learning task-specific emotion embedding with neural network has been prove to be effective. [Tang *et al.*, 2014] leverage a dataset of tweets to obtain emotion embedding. [Agrawal *et al.*, 2018] learns emotion-enriched word-representation on product reviews, with a much smaller corpus. This method could be easily applied on social media by using online emotion corpus, which could largely overcome the limitations of sentiment dictionaries.

2.3 Multimodal Fusion

The simplest approach is concatenation, which fuses different modality vectors by concatenation. [Silberer *et al.*, 2017] employ a stacked autoencoder to learn multimodal representations by embedding linguistic and visual inputs into a common space. However, these methods treat the different modalities equally in fusion. In [Jin *et al.*, 2017], the authors employ the high-level representation of visual modal as the external attention vector to weigh the components of textual modality. Gate mechanism is also widely used in many fusion work. LSTM [Hochreiter and Schmidhuber, 1997] and GRU [Bahdanau *et al.*, 2014] tackle the long-term dependency problems of RNN with different gates that could control how information from last time step and current inputs updates at current unit, which is another form of fusion. [Wang *et al.*, 2018] deploys gates to learn the weights of different modality representations for each word.

Considering the importance of emotion in fake news detection, we proposed a novel fake news detection framework which could exploit emotion information from the publisher and users with the help of emotion embedding and gates.

3 A Preliminary Analysis on Emotion Signals

To explore the relationships between emotions and fake news on social media, we perform a preliminary data analysis on emotions signals for fake and real news in news content and user comments respectively. We collect the datasets from a popular microblog platform Weibo [3] (details of data preparation are introduced in Sec 5.1). We have collected 7,880 fake news pieces and 7,907 real news pieces, and their related user comments on Weibo. The analysis is performed from three perspectives: 1) the emotional category; 2) the emotional intensity; and 3) the emotional expression.

3.1 Emotional Category

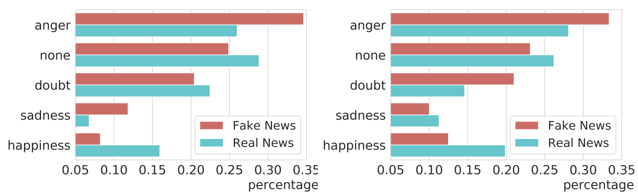
Generally, fake news is sensational and inflammatory. It could arouse specific kinds of emotion among the users, such

[2] http://www.keenage.com/html/e_index.html

[3] www.weibo.com

as suspicion, anxiety or shock. Therefore, we select 5 fine-grained emotion categories to investigate emotions in fake news and real news, including *anger*, *sadness*, *doubt*, *happiness* and *none* (Some contents may not contain emotional information). We adopt the emotion classifier in Sec 4.1 to annotate our experimental data with emotion categories.

Figure 2a and 2b exhibit the distribution of emotion categories in news content and user comments respectively. In fake news' content, the proportion of *anger* is 9% higher than it in real news', while the percentage of *happiness* is 8% lower. Same trend happens in the user comments. In addition, the proportion of sadness in fake news' contents and doubt in fake news' comments is much higher than them in real news. The result demonstrates that, compared to in real news, both the publisher and users tend to express more high-arousal emotions, such as anger, sadness and doubt in fake news.



(a) News Content (b) User Comments

Figure 2: Distributions of emotional category of fake news and real news in: (a) news content and (b) user comments. *Anger* is more likely to appear in fake news, while real news arouse more *happy* emotion in both sources

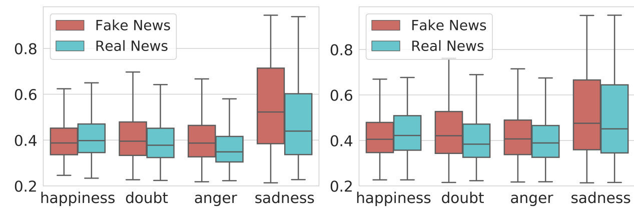
3.2 Emotional Intensity

Each document also owns an emotional intensity level in each emotional category. For example, the intensity of *I'm super happy* is much stronger than intensity of *I'm happy* in happiness emotional category. Fake news is expected to express negative emotion with stronger intensity which could further arouse the intense emotions in the public. In this section, we take the output probability of emotion classifier model's softmax layer as the emotional intensity level for each emotional category, which is a continuous value between 0 and 1.

In Figure 3, we can see that, regardless of sources, the emotions of *anger*, *sadness* and *doubt* in fake news are much more severe than in real news. And this discrepancy is more drastic in news content. In conclusion, both the publisher and users are more possible to express stronger negative emotions in fake news than in real news, while the trend of positive emotion is reverse.

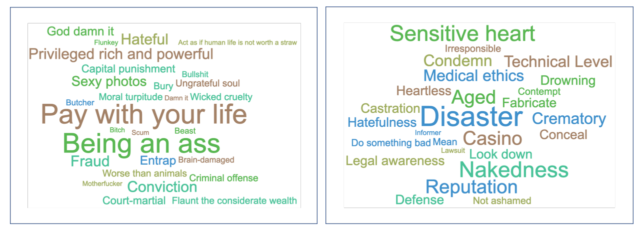
3.3 Emotional Expression

Different people may express their moods with different linguistic usage. For example, some people like using plain words to express their feelings, while others prefer exaggerated words. In fake news, inciting words might be more preferred due to the controversial news content. To analyze the differences of emotional expression, we extract the top-weighted words for expressing *anger* in real news and fake news respectively. We adopt the widely-used method in [Li *et al.*, 2014] to calculate the weight of each word in the dataset for expressing specific kinds of emotion. The top-weighted 30 words in fake news and real news are shown in Figure 4.



(a) News Content (b) User Comments

Figure 3: Distributions of emotional intensities level of fake news and real news in: (a) news content and (b) user comments. The intensities of emotion *anger*, *sadness* and *doubt* in fake news are all stronger than in real news.



(a) Fake News (b) Real News

Figure 4: Emotional expressions for *anger* in fake news and real news. Compared to real news, fake news use more fierce and extreme words to express *anger*.

We can see that fake news conveys *angry* with much more fierce and extreme words like *"damn it"*, *"scum"*. Similar circumstance also exists in other negative emotional categories. Therefore, people use different words for emotional expression in fake and real news.

In summary, we make the following conclusions from these experiments: i) both the publishers and users are more likely to spread more negative emotions in fake news than in real news; ii) participants of fake news tend to express negative emotions with stronger intensities; iii) while expressing a specific kind of emotion, people in fake news prefer exaggerated and inflammatory words.

4 Modeling Emotions for Fake News Detection

In this section, we present the details of the proposed end-to-end emotion based fake news detection framework EFN. It consists of three major components (see Figure 5): i) the content module mines the information from the publisher, including semantic and emotions information in news contents; and ii) the comment module captures semantic and emotion information from users; and ii) the fake news prediction component fuses the high-level features from both news content and user comments and predict fake news.

4.1 Content Module

News contents contain the cues to differentiate fake and real news. We have shown that the distributions of emotion categories are different for fake and real news pieces, which demonstrate the potential to use news content emotions to help detect fake news.

Word Encoder We learn the basic textual feature representations through a recurrent neural network (RNN) based

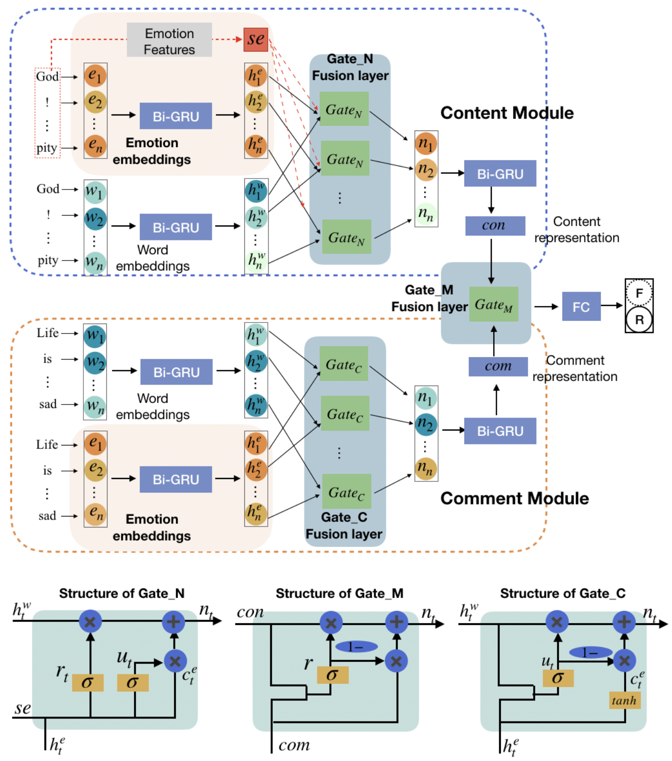


Figure 5: The proposed framework EFN consists of three components: (1) the news content module; (2) the user comments module, and (3) the fake news prediction component. The previous two modules are used to model semantics and emotions from the publisher and users respectively, while prediction part fuse information of these two module and make prediction. Three gates at the bottom are used for multimodal fusion in different layers.

word encoder. Though in theory, RNN is able to capture long-term dependency, in practice, the old memory will fade away as the sequence becomes longer. To make it easier for RNNs to capture long-term dependencies, Gated recurrent units (GRU) [Bahdanau *et al.*, 2014] is designed in a manner to have more persistent memory. To further capture the contextual information of annotations, we use bidirectional GRU to model word sequences from both directions of words. For each word t_i, the word embedding vector t_i^w is initialized with the pre-trained word2vec[Mikolov *et al.*, 2013]. The bidirectional GRU contains the forward GRU $\overrightarrow{f}$ which reads each sentence from word t_0 to t_M and a backward GRU $\overleftarrow{f}$ which reads the sentence from word t_M to t_0:

$$\begin{aligned}\overrightarrow{h_i^w} &= \overrightarrow{GRU}(t_i^w), i \in [0, M],\\ \overleftarrow{h_i^w} &= \overleftarrow{GRU}(t_i^w), i \in [0, M].\end{aligned} \tag{1}$$

for a given word t_i, we could obtain its word encoding vector h_i^w by concatenating the forward hidden state $\overrightarrow{h_i^w}$ and backward hidden state $\overleftarrow{h_i^w}$, i.e., $h_i^w = [\overrightarrow{h_i^w}, \overleftarrow{h_i^w}]$

Emotion Encoder Similar to the word encoder, we adopt bidirectional GRU to model the emotion feature representations for the words. To preserve the emotion signal for each word, we first introduce how to obtain an emotion embedding vector t_i^e for each word t_i.

Inspired by recent advancements on deep learning for emotion modeling [Agrawal *et al.*, 2018], we train a recurrent neural network to learn the emotion embedding vectors. Following traditional settings [Hu *et al.*, 2013], we first obtain a large-scale Weibo datasets that contain emoticons, and use the emoticons as the emotion labels. We initialize words with one-hot vectors instead of word2vecs for not learning too much semantics. After initiation, all word vectors pass an embedding layer which project each word from the original one-hot space into a low dimensional space, and then are sequentially fed into a one-layer GRU model. Then, through back-propagation, the embedding layer get updated during training, producing emotion embedding t_i^e for each word w_i.

After we obtain the emotion embedding vectors, we can learn the emotion encoding h_i^e for word t_i:

$$\begin{aligned}\overrightarrow{h_i^e} &= \overrightarrow{GRU}(t_i^e), i \in [0, M],\\ \overleftarrow{h_i^e} &= \overleftarrow{GRU}(t_i^e), i \in [0, M].\end{aligned} \tag{2}$$

for a given word t_i, we could obtain its word encoding vector h_i^e by concatenating the forward hidden state $\overrightarrow{h_i^e}$ and backward hidden state $\overleftarrow{h_i^e}$, i.e., $h_i^e = [\overrightarrow{h_i^e}, \overleftarrow{h_i^e}]$.

Hand-crafted News Emotion Features The overall emotion information of news content is also important when deciding how much information from emotion embedding should be absorbed for the words. For example, the news which obviously express intense emotions could further strengthen the importance of emotion information on its words in training process. For a given post p_j, we extract the emotion features included in [Castillo *et al.*, 2011] and also add some emoticon features. There are 19 features regarding emotion aspects of news, including *numbers of positive/negative words, sentiment score*, etc. News emotion feature vector of p_j is denoted as se_j.

News Content Representation Gate_N is applied to learn information jointly from word embedding, emotion embedding and sentence emotion features, and yield a new representation for each word(see Figure 5). The units in Gate_N are motivated by the *forget gate* and *input gate* in LSTM. In Gate_N, two emotion inputs corporately decide the value of f_t and i_t with two sigmoid layers, which are used for managing how much information from semantic or emotion modality is added into the new representation. Meanwhile, a tanh layer transfers the emotion inputs to the same dimensional space of word embedding. Mathematically, the relationship in Gate_N is defined as the following formulas:

$$\begin{aligned} r_t &= \sigma(W_r.[se, h_t^e] + b_r)\\ u_t &= \sigma(W_u.[se, h_t^e] + b_u)\\ c_t^e &= tanh(W_c.[se, h_t^e] + b_c)\\ n_t &= r_t * h_t^w + u_t * c_t^e \end{aligned} \tag{3}$$

All the generated vectors of words are fed into a bidirectional GRU layer sequentially, and then the last hidden state of the GRU layer is expected to contain all the information in *Content Module*, which is called *Content Representation*.

4.2 Comment Module

Comment module explores the semantic and emotion information from the users in the event. The architecture of comment module is similar to content module's except: 1) all comments are firstly concatenated before fed into BiGRUs; 2) there is no sentence emotion features; and 3) Gate_C is used for fusion.

Gate_C is introduced for fusion in comment module. Different from Gate_N, Gate_C only two inputs. We adopt the *update gate* in GRU to control the update of information in fusion process (see Figure 5). Two inputs jointly yield a update gate vector $u_t$through a sigmoid layer. A tanh layer creates a vector of new candidate values, h_t^e, which has the same dimension as the w_t. The final output n_t is a linear interpolation between the w_t and h_t^e. Mathematically, following formulas represent the process:

$$\begin{aligned} u_t &= \sigma(W_u.[h_t^w, h_t^e] + b_u) \\ c_t^e &= tanh(W_c.h_t^e + b_c) \\ n_t &= u_t * h_t^w + (1 - u_t) * c_t^e \end{aligned} \tag{4}$$

4.3 The proposed Framework - EFN

Here, Gate_M fuses the high-level representation of content module and comment module, and then yield a representation vector n(see Figure5). Mathematically, following equations demonstrate the internal relationship of Gate_M:

$$\begin{aligned} r &= \sigma(W_u.[con, com] + b_u) \\ n &= r * con + (1 - r) * com \end{aligned} \tag{5}$$

We use a fully connected layer with softmax activation to project the new vector n into the target space of two classes: fake news and real news, and gain the probability distribution:

$$p = softmax(W_c n + b_c) \tag{6}$$

In the proposed model, we employ binary-entropy function to define the loss of the m-th sample S^m as follow:

$$L(S^m) = -[l^m p^m + (1 - l^m) \log(1 - p^m)] \tag{7}$$

where p^m denotes the probability of being fake news of m-th sample, and l^m denotes the ground truth of m-th sample with 1 representing fake news and 0 representing real news.

5 Experiment

In this section, extensive experiments are conducted to evaluate the performance of our framework.

5.1 Datasets

We construct the dataset on Sina Weibo. This dataset includes 7880 pieces of fake news and 7907 pieces of real news, with nearly 160k comments. The fake news is collected from the official rumor debunking system of Weibo[4], and the real news is gathered from *NewsVerify*[5], a real-time news certification system on Weibo which contains a large-scale verified truth posts on Weibo[Zhou *et al.*, 2015]. All comments in 24-hours time interval after publishing time are collected for each news post. The statistics of our datasets is as Table 1. Note that not every post owns comments. In our experiment, we first use K-means algorithm to cluster all news into 200 clusters, and split it into training data and testing data in ratio 4:1 at *cluster level*, to promise that there is no event overlap between the training and testing set, which prevent the model from overfitting on event topics.

[4]http://service.account.weibo.com/

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

	Fake News	Real News	All
# Source Posts	7,880	7,907	15,787
# User Comments	109,154	47,037	1,566,191

Table 1: The Statistics of our Dataset.

Methods	Acc	Prec	Recall	F1
DTC	0.756	0.754	0.758	0.756
ML-GRU	0.799	0.810	0.790	0.800
Basic-GRU	0.835	0.830	0.850	0.840
HSA-BLSTM	0.843	0.860	0.810	0.834
EFN	**0.872**	**0.860**	**0.890**	**0.874**

Table 2: Performance Comparison of Fake News Detection.

5.2 Compared Fake News Detection Methods

We compare our framework with the following state-of-art methods:

- **DTC**[Castillo *et al.*, 2011] uses J48 decision tree to evaluate the credibility of tweets with hand-crafted features. The features also include basic emotion features.
- **ML-GRU** [Ma *et al.*, 2016] models a post event as a variable-length time series and apply a multilayer GRU network to learn these pieces. Since there is no repost in our dataset, we take the comments as replacement.
- **Basic-GRU** contains two generic Bi-GRU network to model semantics of news content and comments with word embedding, with concatenation on top layer.
- **HSA-BLSTM** [Guo *et al.*, 2018] uses a hierarchical attention model to capture the hierarchical structure in a post event. Social context features is incorporated in the model through attention mechanism which also contain some basic emotion features. Similarly, we use comments as replacement of reposts while implementation.

We follow the conventional metrics: Accuracy, Precision, Recall, and F1-Score for a comprehensive evaluation.

5.3 Performance Comparison

Table 2 presents the experiment results of all compared methods and the proposed model. In particular, the EFN model achieves an overall accuracy of 87.2% and 87% of F1-Score on our datasets, which outperforms all the baseline models. The outstanding performance of the proposed model demonstrates that incorporation of emotion through embedding representation and gated fusion could effectively promote the detecting process on fake news.

We can see that all the neural network models earn better performance than the hand-crafted feature based method. This indicates that generic RNN is capable of exploiting deep latent features of text through variable time-series architecture. And the Basic-GRU model outperforms ML-GRU on our dataset mainly because that the comments of each post are somewhat too small to support the complicated structure of ML-GRU which relies on rich repost sources.

Our method shows its strength on fake news detection in these experiments. As is shown in Table 2, EFN rises the accuracy of fake news detection by nearly 12%, from 75.6%

Module	Methods	Acc	F1
Content Module	WE	0.790	0.801
	EE	0.700	0.719
	WSE(c)	0.793	0.790
	WEE(c)	**0.813**	**0.810**
Comment Module	WE	0.667	0.550
	EE	0.619	0.553
	WEE(c)	**0.669**	**0.560**
Content + Comment	WE+WE(c)	0.835	0.840
	(WEE(gn)+WE)(c)	0.860	0.859
	(WEE(gn)+WEE(gc))(c)	**0.866**	**0.868**

Table 3: Evaluation of Emotion Embedding.

Module	Methods	Acc	F1
Content Module	WEE(c)	0.813	0.810
	WEE(att)	0.799	0.793
	WEE(gn)	**0.851**	**0.854**
Comment Module	WEE(c)	0.669	0.560
	WEE(gc)	**0.671**	**0.563**
Content +Comment	(WEE(gn)+WEE(gc))(c)	0.866	0.868
	(WEE(gn)+WEE(gc))(gm)	**0.872**	**0.874**

Table 4: Evaluation of Three Gates.

of decision tree to 87.2%. And its f1-score is also 4% higher than the second one.

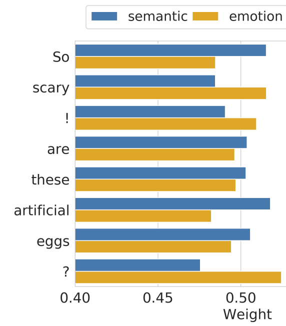
5.4 Evaluation of Emotion Embedding

To analyze the effect of emotion in content, comments and the whole process respectively, we take out the content module alone, comment module alone, and the whole framework for experiments.

WE and **EE** denote that only word embedding or emotion embedding is used. **WEE** means that both of these two embeddings are used. And **WSE** means that both word embedding and emotion representation are used, where the emotion representation for each word is a 21-dimensional one-hot vector from a sentiment dictionary which classifies each sentiment word into 21 emotion categories.

Meanwhile, **c, gn, gc, gm** represent fusion strategies of concatenation, Gate_N, Gate_C and Gate_M respectively. **att** denotes *attention fusion* strategy in [Jin *et al.*, 2017], which takes the high-level representation of one modality as external attention vector to weigh the components of another modality.

From Table 3, we could make the following observations: 1) incorporation of emotion improves the performance of the whole framework, compared to merely using semantic information; 2) in content module, the overall performance rises while using emotion embedding; 3) emotion plays a more important role in content module than in comment module on our dataset. It possibly results from the sparsity of comments data, which limits the effectiveness of emotion in comment module; and 4) emotion embedding method outperforms the sentiment dictionary based emotion representation method.



(a) Gate_C Weights in a sentence.

Fake News Content	Related Comments
This morning, an 11-year-old boy was abducted near Jinxiu Garden Community. 0.362	• Another form of this **rumor** • What? **Rumor**?? It's from my classmates's wechat, **not a rumor** 0.638
Don't give your baby a milk drink with additives. They contain botox!!! 0.364	• A obvious **false news**, you have been **deceived**! 0.636
Typhoon "Crow" is coming again. 0.364	• **Don't spread** it, there's no typhoon called "Crow" 0.636
Although I don't know if this is true, but I hope someone can see the truth. 0.365	• Your words are **self-contradictory**. • News have no truths, and truths will not be told to the public. 0.635
# Tianjin Explosion Accident # Please give them a truth! 0.366	• Waiting the **TRUTH**! 0.634

(b) Fake news whose comment module is top Gate_M weighted.

Figure 6: Analysis of weights in Gate_C and Gate_M: (a) Gate_C Weights of each word in a sentence; and (b) Fake news whose comment modules are top Gate_M weighted.

5.5 Evaluation of Gates

To evaluate the effectiveness of three gates, we take out the content module alone, comment module alone, and the whole framework for experiments, by using different fusion strategies. As is shown in Table 4, various gates further improve the promotion that emotion information brings on fake news detection. In particular, Gate_N in content module evidently increases the F1-Score by 4% compared to simple concatenation, and nearly 5% while contrasting with *attention fusion*. On the other hand, the improvement brought by Gate_C and Gate_M is not as obvious as Gate_N, at less than 1%.

We extract the unit u_t in Gate_C which is a weight vector between semantic vector h_t^w and emotion vector c_t^e for each word. We compute the average of the vector as approximation of the weight at modality. Figure 6a shows an example in fake news. We could observe that: 1) emotion of sentiment words such as *"scary", "!"* and *"?"* gain higher weight than semantic modality. Many of these words even don't appear in sentiment dictionaries; and 2) sentiment words' emotion modality obtain more attention than others words.

Similarly, we also compute the approximate weight between content module and comment module in Gate_M. Figure 6b show 5 samples in fake news whose comment modules are top-weighted. Interestingly, as we can see, most comments in these fake news contain clues for verifying the truth of news content(e.g.,"*deceived*"). This validates the capability of Gate_M on capturing important knowledge while fusing different modalities.

6 Conclusion

In this paper, we propose a end-to-end emotion-based fake news detection framework, EFN, which incorporate the publisher emotion and the social emotion in fake news detection simultaneously. We use content module and comment module to learn semantic and emotion information from the publisher and users. Technically, we adopt embedding for emotion representation for each word and propose three types of gates for fusion at different levels to fully explore emotion information. Extensive experiments on microblogs demonstrate that the proposed EFN model is effective for detecting fake news and outperforms the state-of-art fake news detection methods.

References

[Agrawal *et al.*, 2018] Ameeta Agrawal, Aijun An, and Manos Papagelis. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961, 2018.

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

[Ferrara and Yang, 2015] Emilio Ferrara and Zeyao Yang. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, 1:e26, 2015.

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

[Gupta *et al.*, 2012] Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 153–164. SIAM, 2012.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Hu *et al.*, 2013] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In *Proc. of the 22nd WWW*, pages 607–618. International World Wide Web Conferences Steering Committee, 2013.

[Jin *et al.*, 2014] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 230–239. IEEE, 2014.

[Jin *et al.*, 2016] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI*, pages 2972–2978, 2016.

[Jin *et al.*, 2017] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816. ACM, 2017.

[Kamps *et al.*, 2004] Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. Using wordnet to measure semantic orientations of adjectives. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, 2004.

[Li *et al.*, 2014] Chengxin Li, Huimin Wu, and Qin Jin. Emotion classification of chinese microblog text via fusion of bow and evector feature representations. In *Natural Language Processing and Chinese Computing - Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings*, pages 217–228, 2014.

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

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

[Shu *et al.*, 2017] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[Silberer *et al.*, 2017] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Visually grounded meaning representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2284–2297, 2017.

[Stieglitz and Dang-Xuan, 2013] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013.

[Tang *et al.*, 2014] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1555–1565, 2014.

[Wang *et al.*, 2018] Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Learning multimodal word representation via dynamic fusion methods. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Wiebe *et al.*, 2005] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[Yang *et al.*, 2012] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM, 2012.

[Zhou *et al.*, 2015] Xing Zhou, Juan Cao, Zhiwei Jin, Fei Xie, Yu Su, Dafeng Chu, Xuehui Cao, and Junqiang Zhang. Real-time news cer tification system on sina weibo. In *Proceedings of the 24th International Conference on World Wide Web*, pages 983–988. ACM, 2015.