Exploiting Emotions for Fake News Detection on Social Media

Anonymous

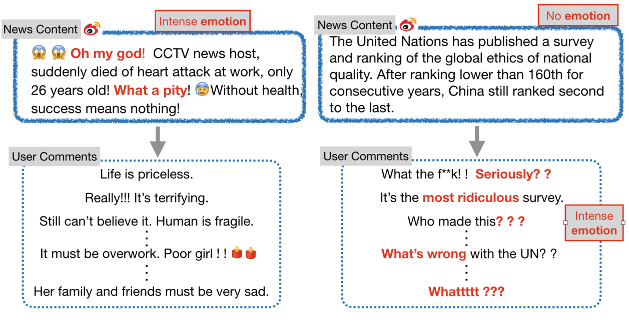
***Abstract*—Microblog has become a popular platform for people to post, share, and seek information due to its convenience and low cost. However, it also facilitates the generation and propagation of fake news, which could cause detrimental societal consequences. Detecting fake news on microblogs is important for the 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 reality, the *publishers* typically post either a tweet with intense emotion which could easily resonate with the crowd, or a controversial statement unemotionally aiming 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 (EFND), 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 datasets demonstrate the effectiveness of the proposed framework.**

***Index Terms*—Fake News Detection, Gate Mechanism, Emotion Embedding**

I. 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 fosters the wide propagation of *fake news*, commonly referred as intentional false information [1]. 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 engagement on Facebook than top 20 major real news, and that these fake news earned nearly 9 million shares on social media. These fake news do serious harm to not only the public credibility, but also social stability and economic market. Therefore, it's important to intensify research efforts of building tools that detect the fake news automatically and effectively.

Existing work on fake news detection mainly focuses on *news content* and *social context*. Feature-based classification models extract basic semantic and emotion features from content, and statistical features from users [2]. Propagation-based models construct the relationship network inside the event, and incorporate social conflicting viewpoints towards event in the network [3], [4]. Recently, deep learning models are proposed to evaluate the credibility of information on social media, which deeply exploits the semantics information from news content [5]. Basic social context features are fused into deep learning models in some studies [6]. However, emotion information, which is crucial for fake news detection, is underutilized in these studies. Few studies leverage emotion in news contents and social context simultaneously for fake news detection.



😱 😱 Oh my god! CCTV news host, suddenly died of heart attack at work, only 26 years old! What a pity! 😰Without health, success means nothing!

Life is priceless.
Really!!! It's terrifying.
Still can't believe it. Human is fragile.
It must be overwork. Poor girl ! ! 🕯️🕯️
Her family and friends must be very sad.

The United Nations has published a survey and ranking of the global ethics of national quality. After ranking lower than 160th for consecutive years, China still ranked second to the last.

What the f**k! ! Seriously? ?
It's the most ridiculous survey.
Who made this? ? ?
What's wrong with the UN? ?
Whattttt ???

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

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 not only in real world but also on social media [7]–[11]. To achieve this goal, fake news publishers commonly adopt two approaches. First, publishers post news with intense emotions which can trigger a high level 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 but with controversial content which evokes intense emotion in the public, encouraging it spreading widely. As another example (see Figure 1b) shows, the publisher writes the post in an unemotional way; while, the statement that China ranks second to the last suddenly brings on tension in the crowd, and

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

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 users corporately has the potential to improve fake news detection performance.

There exists a realistic dilemma of exploiting emotion. Since a large part of fake news doesn’t contain emotional signals in contents, difference of emotion between fake news and real news content may not be obvious. However, fake news content tends to be more controversial and instigating, though unemotional. Hence it is necessary to mine the emotion information from the audience, such as reposts or comments. The majority of existing work utilizing emotion information for detecting fake news either extract emotion feature from news content [2], [6], or model the viewpoints of users in propagation [4], [12].

To exploit emotion information for fake news detection, we define two types of emotion: (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 this paper, news content and user comments are used to capture *publisher emotion* and *social emotion*, respectively. In essence, we investigate: (i) how to capture signals of *publisher emotion* and *social emotion* from news content and user comments, respectively; and (ii) how to exploit publisher and social emotions simultaneously for fake news detection. Our solution to these two challenges is a novel Emotion-based Fake News Detection framework (EFND). Technically, we leverage emotion embedding for better emotional representation of each word. And three gate units are designed to exploit information from different modules simultaneously. 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 EFND, 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 EFND for fake news detection.

The paper is organized as follows. In the next section, we give an overview of related. Next, we demonstrate the difference of emotion between fake news and real news with analysis on a dataset. Section IV presents the details of the framework EFND, including emotion embedding and various gates. Datasets, experiment settings and results are presented in Section V. We conclude our work in Section VI.

II. Related Work

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

A. Fake News Detection

Previous fake news detection studies mostly focus on extracting features and training a classifier to predict the credibility of news. Early work manually extracts a wide range of features including user features, content features, propagation features and topic features [2]. Location and client are proved to be effective in detecting fake news as well [13]; Besides feature-based models, propagation-based approaches aim to mine the relations between various entities in an event. Propagation network is firstly introduced on evaluation in work [14]. Tweets, users, events and their inter-relations construct a hierarchical network, which then evaluates the credibility of each tweet and event by iterative calculation. Then, similar structure is applied on microblogs which consists of news, subevents and messages [3]. And conflicting viewpoints among users are further investigated in the propagation network [4]. Recently, neural network models are adopted for fake news detection. Ma et al. [5] firstly applies RNN for detection on social media by modeling the posts in an event as a sequential time series. Guo et al. [6] proposes a social attention network to capture the hierarchical characteristic of events on microblogs.

So far, emotion in these works is used by tools of either emotion features or viewpoints of users, which requires systematical and comprehensive explorations in future work.

B. Emotion Representation

Early studies primarily use sentiment dictionaries for representing emotion of words. There are several widely-used sentiment dictionaries, including WordNet [15], LIWC[2], MPQA [16]for English, and HowNet [3] for Chinese. Beyond this, many works leverage these dictionaries and man-made rules to exploit emotion information. For example, Vader [17] constructs a list of lexical features and rules to assess the sentiment of microblog-like contents. However, this method could only align each sentiment word with an emotional class, or intensity value which is somewhat not informative for representing words’ emotion. Besides, low coverage and differences of word usage on social media and in real world also limit the effectiveness of sentiment dictionaries in many circumstances.

Learning task-specific emotion embedding with neural network could encodes sentiment information in the continuous representation of words, which has been proved to be effective recently. Tang et al. [18] utilizes a massive distant-supervised tweet dataset to obtain emotion embedding on social media. Agrawal et al. [19] learns emotion-enriched word-representation on product reviews with a much smaller corpus. This sort of methods enable each word to obtain a

[2] www.liwc.net

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

distributed representation vector that contains rich and high-level emotion information.

C. Multimodal Fusion

The simplest approach is concatenation. Silberer et al. [20] employs 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 work [21], the authors employ the high-level representation of visual modal as an external attention vector to weigh the components of textual modality. Gate mechanism is also widely used in many fusion works. LSTM [22] and GRU [23] 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. [24] uses gate units to learn the weights of different modal representations for each word.

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

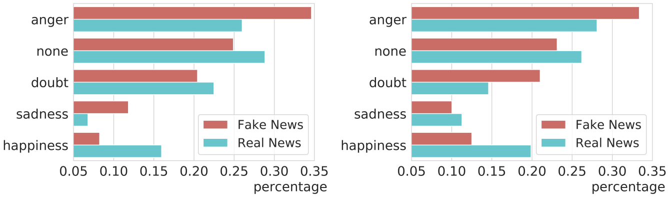
III. A Preliminary Analysis on Emotion Signals

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 dataset from a popular microblog platform Weibo [4] (details of data preparation are introduced in Sec V-A). We have collected 7,880 fake news pieces and 7,907 real news pieces, and their related user comments. The analysis is performed from three perspectives: 1) the emotional category; 2) the emotional intensity; and 3) the emotional expression.

A. Emotional Category

Generally, fake news is sensational and inflammatory. It could arouse specific kinds of emotions among the users such as suspicion, anxiety or shock. Therefore, we select 5 fine-grained emotional categories to investigate emotions in fake and real news, including *anger*, *sadness*, *doubt*, *happiness* and *none* (Some contents may not contain emotional information). We adopt the classifier in Sec IV-A0b to annotate our experimental data with emotion categories.

Figure 2a and 2b exhibit the distribution of emotional 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. This result demonstrates that, compared to in real news, both the publisher and users tend to express more high-arousal emotions in fake news.



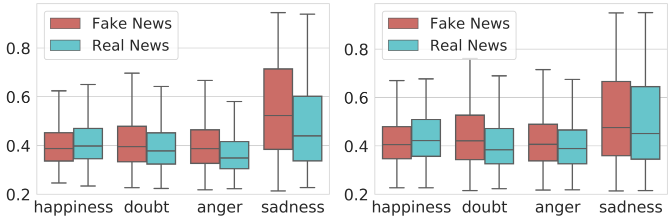
(a) News Content (b) User Comments

Fig. 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

B. 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 emotional category *happiness*. Fake news is expected to express negative emotion with stronger intensity which could further arouse intense emotions in the public. In this section, we take the output probability of emotional category classifier s(in Sec IV-A0b) 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 severer 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.



(a) News Content (b) User Comments

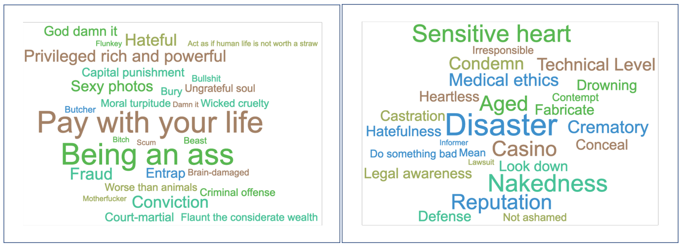
Fig. 3: Distributions of emotional intensities level of fake news and real news in: (a) news content and (b) user comments. The intensities of emotion *anger*, *sadness* and *doubt* in fake news are all stronger than in real news.

C. Emotional Expression

Different people may express their feelings with different linguistic usage. For example, some people like using plain words to express their feelings, while others prefer exaggerated

[4] www.weibo.com

words. In fake news, inciting words might be more preferred. 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 [25] to calculate the weight of each word in the dataset. The top-weighted 30 words in fake news and real news are shown in Figure 4.



(a) Fake News

(b) Real News

Fig. 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”, ”ass”*. 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; and iii) while expressing a specific kind of emotion, people in fake news prefer exaggerated and inflammatory words.

IV. 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 EFND. It consists of three major components (see Figure 5): i) the content module exploits the information from the publisher, including semantic and emotion information in news contents; ii) the comment module captures semantic and emotion information from users; and iii) the fake news prediction component fuses the high-level features from news content and user comments, and then classify it as fake or not.

A. Content Module

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

a) Word Encoder: We learn the basic textual feature representations through a recurrent neural network(RNN) based 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) [23] are 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 w_i is initialized with the pre-trained word2vec [26]. The bidirectional GRU contains the forward GRU $\overrightarrow{f}$ which reads each sentence from word t_0 to t_n and a backward GRU $\overleftarrow{f}$ which reads the sentence from word t_n to t_0:

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

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., $h_i^w = [\overrightarrow{h_i^w}, \overleftarrow{h_i^w}]$

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

Inspired by recent advancements on deep learning for emotion modeling [19], we train a recurrent neural network to learn the emotion embedding vectors. Following traditional settings [27], we first obtain a large-scale Weibo dataset in which each Weibo contains emoticons. Then we categorize the top 200 emoticons into 5 emotional classes(*anger*, *doubt*, *happiness*, *sadness* and *none*), and use the emoticons to label the corpus. Next, we initialize each word with the one-hot vector. Here, we don't use pre-trained word embeddings for not learning too much semantic information in emotion embeddings. After initiation, all words pass an embedding layer which projects each word from the original one-hot space into a low dimensional space, and then are sequentially fed into a one-layer GRU model. Finally, through back-propagation, the embedding layer gets updated during training, producing emotion embedding e^i for each word t_i. Besides, this classifier is also used for analysis in Sec III. For experiment on Twitter, we perform the same procedure on the labeled sentiment tweet corpus which is published in SemEval 2018 [28] to get emotion embeddings of English words.

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

$$\begin{aligned}\overrightarrow{h_i^e} &= \overrightarrow{GRU}(e_i), i \in [1,n],\\ \overleftarrow{h_i^e} &= \overleftarrow{GRU}(e_i), i \in [1,n].\end{aligned} \tag{2}$$

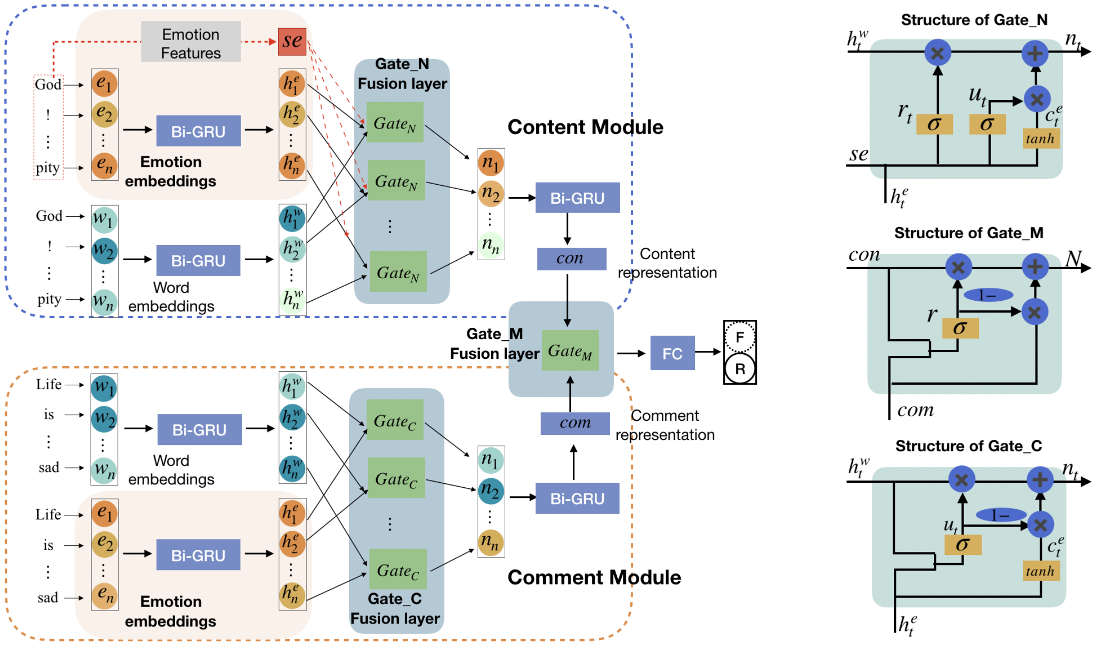


Fig. 5: The proposed framework EFND consists of three components: (1) the news content module, (2) the user comments module, and (3) the fake news prediction component. The previous two modules are used to model semantic and emotions from the publisher and users respectively, while prediction part fuses information from these two modules and makes prediction. Three gates at the right side are used for multimodal fusion at different layers in this framework.

for a given word t_i, we could obtain its emotion 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., $h_i^e = [\overrightarrow{h_i^e}, \overleftarrow{h_i^e}]$.

c) Hand-crafted News Emotion Features: The overall emotion information of news content is also important while deciding how much information from emotion part should be absorbed for each word. For example, the news content which obviously expresses intense emotions could further strengthen the importance of emotion part in each word of the content. For a given post p_j, we extract the emotion features included in work [2] and also add some extra emoticon features. There are 19 features regarding emotion aspects of news, including *numbers of positive/negative words, sentiment score*, etc. News emotion features of p_j is denoted as se_j.

d) News Content Representation: Gate_N is applied to learn information jointly from word embedding, emotion embedding and sentence emotion features, and then yield a new representation for each word(see Figure 5). The units in Gate_N is motivated by the *forget gate* and *input gate* in LSTM. In Gate_N, two emotion inputs corporately decide the value of r_t and u_t with two sigmoid layers, which are used for managing how much information from semantic and emotion is added into the new representation. Meanwhile, a tanh layer transfer the emotion inputs to the same dimensional space with word embeddings'. Mathematically, the relationships between inputs and output of Gate_N are 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*, called *Content Representation*.

B. Comment Module

Comment module explores the semantic and emotion information from 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 BiGRU; 2) there is no sentence emotion features; and 3) Gate_C is used for fusion.

We choose to concatenate all the comments for inputs because over 70% news pieces own less than 5 comments, which reflect the situation in real world as well. As a consequence of concatenation, the input doesn't own the intact information as a *sentence*, so there is no *sentence emotion features*.

Gate_C is introduced for fusion in comment module. Different from Gate_N, there are only two input modalities. 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 create a vector of new candidate values, c_t^e, which has the same dimension as h_t^w. The final output n_t is a linear interpolation between c_t^e and h_t^w. Mathematically, following formulas represent the process:

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

C. The proposed Framework - EFND

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

V. Experiment

In this section, experiments on two real-world datasets are conducted to evaluate the effectiveness of EFND. Specifically, this section aims to answer the following questions:

- **Q1:** Is EFND capable of detecting fake news on social media? And compared to state-of-art methods, how much performance could it improve?
- **Q2:** How effective are emotion and gate mechanism in improving the performance of EFND?
- **Q3:** What is exactly learned behind various gate units?

We firstly introduce the datasets that are used in experiments, and the experiment settings including implementation details and representative methods. Then we compare EFND with these methods to answer **Q1**, and ablation studies are performed to answer **Q2**. Finally, we extract the weight vectors in various gates to investigate the process behind gate units, which answers **Q3**.

A. Datasets

a) Weibo Dataset: We construct a dataset on Sina Weibo. This dataset includes 7880 fake news pieces and 7907 real news pieces, with nearly 160k comments. The fake news pieces are directly collected from the official rumor debunking system of Weibo[5]. This system actually serves as an authoritative source to collect fake news in many literatures [5], [21]. And the real news pieces are gathered from *NewsVerify*[6], a real -time news certification system on Weibo which contains a large-scale verified truth posts on Weibo [29]. All user comments in 24-hours time interval after publishing time are collected. Note that not every post owns comments.

b) Twitter Dataset: For Twitter[7] dataset, we use the dataset built in work [5]. All fake news are crawled for Twitter by searching keywords extracted from fake news on Snopes. Part of non-rumor events are also from Snopes, and others are from two public datasets [2], [30]. This dataset contains 498 fake news pieces and 494 real news pieces. All the reposts of these news pieces are collected as well, which are used to mine the information from users in our experiments.

The statistics of experiment datasets are as Table I. In our experiment, we first use K-means algorithm to cluster all news pieces into 200 clusters, and split them into training data and testing data in ratio 4:1 at *cluster level*. Trough this way could we promise that there is no event topic overlap between the training and testing sets, which could prevent model from overfitting on event topics.

	Weibo	Twitter
# Source Posts of Fake News	7,880	498
# Source Posts of Real News	7,907	493
# All Posts of Dataset	15,787	992
# User Comments/Reposts of Fake News	109,154	134,590
# User Comments/Reposts of Real News	47,037	373,500
# All Comments/Reposts of Dataset	156,191	508,090

TABLE I: The Statistics of Experiment Dataset.

B. Compared Fake News Detection Methods

For word embedding, we align each word with a 32-dimensional vector which is from a pre-trained Word2Vector model on each dataset, and with a 16-dimensional emotion embedding vector from the pre-trained embedding layers(See SecIV-A0b). We set the dimension of the hidden states of Bi-GRU as 32 and use Adam [31] for stochastic optimization. Batch size of the training process is 128. On Weibo datatset, we use up to 5 comments which are closest to the publish time to model comment modules, since over 70% news pieces own less than 5 comments on Weibo dataset. For news pieces which own less than 5 comments, we pad the vacancies with zero vectors. And for a fair comparison, the earliest 10 reposts are used to model comment module on Twitter dataset.

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

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

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

[7] http://www.twitter.com

- **DTC** Catillo et al. [2] uses J48 decision tree to evaluate the credibility of of tweets with hand-crafted features. The features also include basic emotion features.
- **ML-GRU** Ma et al. [5] models a post event as a variable-length time series and apply a multilayer GRU network to learn these pieces. Since there is no reposts in Weibo dataset, we take the comments as a replacement.
- **Basic-GRU** contains two generic Bi-GRU network to model semantic information of news content and comments with word embedding, with simple concatenation on top layer.
- **HSA-BLSTM** Guo et al. [6] use 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 a replacement of reposts while implementation on Weibo dataset.
- **CSI** Ruchansky et al. [32] proposes a framework which is composed of three modules that integrate the response, text and users to classify an article as fake or not. This model uses LSTM to capture the temporal representation of articles, and a fully connected layer to model characteristics of users and then concatenate these two vectors for classification.

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

C. Performance Comparison

Table II presents the experimental results of all compared methods and the proposed model. In particular, the EFND model achieves an overall accuracy of 87.2% on weibo dataset and 75.1% on Twitter dataset, which outperform all the baseline models on both datasets. 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 methods. This may indicate 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 both dataset mainly because that the amount of responses are somewhat too small to support the complicated structure of ML-GRU which rely on rich repost sources.

Our method shows its strength on fake news detection in these experiments. As is shown in Table II, on Weibo dataset EFND rises the accuracy of fake news detection by nearly 12%, from 75.6% of decision tree to 87.2%. And its f1-score is also 4% higher than the second one. On Twitter dataset, the improvement is more obvious by boosting the accuracy from 61.3% of feature-based models to 75.1%. Meanwhile, EFND outperforms the second-best model by over 7% in f1-score. These observations demonstrate the importance of incorporating emotion information into models.

Dataset	Methods	Accuracy	Precision	Recall	F1-Score
Weibo	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
	CSI	0.835	0.735	**0.996**	0.858
	HSA-BLSTM	0.843	0.860	0.810	0.834
	EFND	**0.872**	**0.860**	0.890	**0.874**
Twitter	DTC	0.613	0.608	0.570	0.588
	ML-GRU	0.684	0.663	0.740	0.692
	Basic-GRU	0.695	0.674	0.721	0.697
	CSI	0.696	0.706	0.649	0.671
	HSA-BLSTM	0.718	**0.731**	0.663	0.695
	EFND	**0.751**	0.698	**0.860**	**0.771**

TABLE II: Performance Comparison of Fake News Detection.

D. Component Analysis

To analyze the effectiveness of emotion and gate mechanism in content, comments and the whole framework respectively, we take out the content module alone, comment module alone, and the whole framework for ablation study experiments.

We use **WE** and **EE** to denote that only word embeddings or emotion embeddings is used. **WEE** means that both of these two embeddings are used. Meanwhile, symbols **c, gn, gc, gm** are representing fusion strategies **concatenation**, **Gate_N**, **Gate_C** and **Gate_M**, respectively. Symbol **att** denotes **attention fusion** strategy introduced in [21], which takes the high-level representation of one modality as external attention vector to weigh the components of another modality.

Table III reports the experimental results of different modules on two datasets.

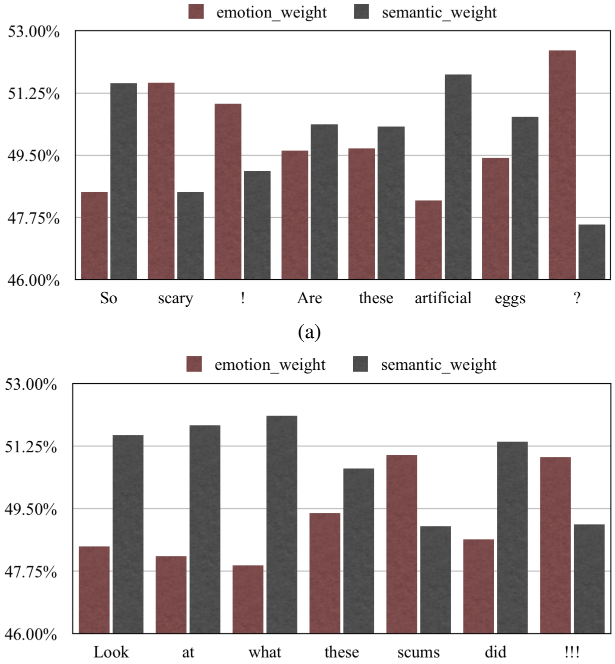
a) Emotion Signals: From Table III. We could make the following observations: 1) in content module, the overall performance rises while using emotion embeddings; Especially on Twitter dataset, adding emotion information increases the f1-score by over 7%; 2) emotions play a more important role in content module than comment module on both datasets. It possibly results from the sparsity of response data, which limits the effectiveness of emotion in comment module. 3) compared to merely using semantic information, incorporation of emotion from one side or two sides all improve the performance of the whole framework, which demonstrates the importance of emotion on fake news detection;

b) Gate Mechanism: We take out the content module alone, comment module alone, and the whole framework for experiments by using different fusion strategies. Here, we compare our gate fusion strategy with concatenation and attention fusion. As is shown in Table III, various gate units

Module	Methods	Weibo Dataset				Twitter Dataset			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Content Module	WE	0.790	0.758	0.849	0.801	0.678	**0.716**	0.558	0.627
	EE	0.700	0.670	0.760	0.719	0.639	0.646	0.609	0.615
	WEE(c)	0.813	0.793	0.826	0.810	0.690	0.650	0.779	0.709
	WEE(att)	0.799	0.788	0.798	0.793	0.701	0.714	0.640	0.675
	WEE(gn)	**0.851**	**0.835**	**0.873**	**0.854**	**0.725**	0.687	**0.791**	**0.735**
Comment Module	WE	0.667	**0.846**	0.407	0.550	0.667	**0.680**	0.593	0.634
	EE	0.619	0.667	**0.472**	0.553	0.655	0.629	0.709	0.667
	WEE(c)	0.669	0.831	0.423	0.560	0.689	0.667	**0.721**	0.693
	WEE(gc)	**0.671**	0.836	0.424	**0.563**	**0.713**	**0.701**	0.709	**0.705**
Content + Comment	(WE+WE)(c)	0.835	0.830	0.850	0.840	0.695	0.674	0.721	0.697
	(WEE(gn)+WE)(c)	0.863	0.860	0.870	0.860	0.736	0.686	0.837	0.754
	(WEE(gn)+WEE(gc))(c)	0.866	0.830	**0.920**	0.870	0.746	0.678	**0.907**	**0.776**
	(WEE(gn)+WEE(gc))(gm)	**0.872**	**0.860**	0.890	**0.874**	**0.751**	**0.698**	0.860	0.771

TABLE III: Component Analysis of Emotion Embedding.

further improve the promotion that emotion information brings on classification. In particular, Gate_N in content module evidently increases the f1-score by around 4% compared to simple concatenation, and nearly 5% while contrasting with *attention fusion* on Weibo dataset. On the other hand, the improvement brought by Gate_C and Gate_M is not as obvious as Gate_N, at less than 1% on both datasets.



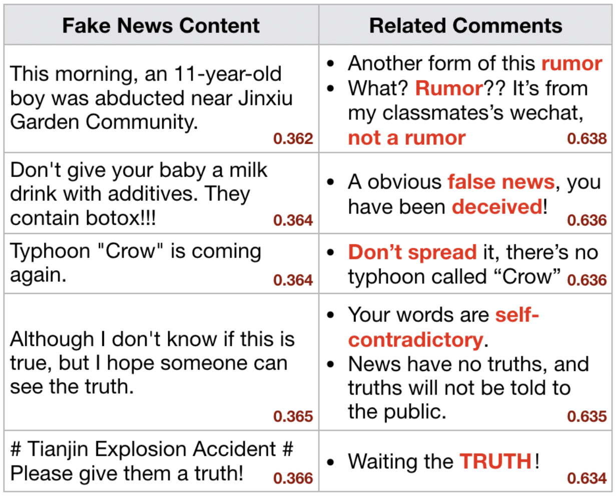
(a)

(b)

Fig. 6: The distribution of weights calculated by Gate_C between semantic and emotion of each word in two sentences.

E. Case Study

To further investigate what is actually learned behind gate units, we extract the vector 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 an approximation of the weight between two modalities. Figure 6 shows two examples in fake news. We could observe that: 1) emotion part in sentiment words such as *"scary", "!" , "?"* and *"scums"* gains higher weight than semantic part. Many of these words even don't appear in sentiment dictionaries; and 2) sentiment words' emotion modalities obtain more attention than unemotional words.



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

Fig. 7: Fake news whose comment modules are top weighed by Gate_M.

Similarly, we also compute the average of vector r as an approximation of weights between content module and comment module in Gate_M. Figure 7 show 5 fake news pieces whose comment modules are top-weighted. Interestingly, most of these news pieces' comments contains cues for verifying the truth of news content(e.g.,"deceived", "rumor"). This validates the capability of Gate_M on capture important knowledge while fusing different modalities.

VI. Conclusion

In this paper, we propose an end-to-end emotion-based fake news detection framework, EFND, which incorporates publisher emotion and social emotion in fake news detection simultaneously. We apply content module and comment module to exploit semantic and emotion information from the publisher and users, respectively. Technically, we adopt embedding to capture emotion information for each word. To fully explore emotion in news event, three types of gates are proposed for fusion at different levels in EFND. Extensive experiments on Weibo and Twitter datasets demonstrate that the proposed EFND model is effective for detecting fake news and outperforms several state-of-art fake news detection methods.

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[2] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 675–684.

[3] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, "News credibility evaluation on microblog with a hierarchical propagation model," in *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 230–239.

[4] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs." in *AAAI*, 2016, pp. 2972–2978.

[5] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks." in *IJCAI*, 2016, pp. 3818–3824.

[6] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp. 943–951.

[7] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social mediasentiment of microblogs and sharing behavior," *Journal of management information systems*, vol. 29, no. 4, pp. 217–248, 2013.

[8] E. Ferrara and Z. Yang, "Quantifying the effect of sentiment on information diffusion in social media," *PeerJ Computer Science*, vol. 1, p. e26, 2015.

[9] R. L. Rosnow, "Inside rumor: A personal journey." *American Psychologist*, vol. 46, no. 5, p. 484, 1991.

[10] N. H. Frijda, "Impulsive action and motivation," *Biological psychology*, vol. 84, no. 3, pp. 570–579, 2010.

[11] R. S. Lazarus and S. Folkman, *Stress, appraisal, and coping*. Springer publishing company, 1984.

[12] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*, 2017.

[13] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 2012, p. 13.

[14] M. Gupta, P. Zhao, and J. Han, "Evaluating event credibility on twitter," in *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 2012, pp. 153–164.

[15] J. Kamps, M. Marx, R. J. Mokken, and M. 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.

[16] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[17] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[18] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. 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*, 2014, pp. 1555–1565.

[19] A. Agrawal, A. An, and M. Papagelis, "Learning emotion-enriched word representations," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 950–961.

[20] C. Silberer, V. Ferrari, and M. Lapata, "Visually grounded meaning representations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2284–2297, 2017.

[21] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 795–816.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[24] S. Wang, J. Zhang, and C. Zong, "Learning multimodal word representation via dynamic fusion methods," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[25] C. Li, H. Wu, and Q. 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*, 2014, pp. 217–228.

[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[27] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in *Proc. of the 22nd WWW*. International World Wide Web Conferences Steering Committee, 2013, pp. 607–618.

[28] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 Task 1: Affect in tweets," in *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[29] X. Zhou, J. Cao, Z. Jin, F. Xie, Y. Su, D. Chu, X. Cao, and J. Zhang, "Real-time news cer tification system on sina weibo," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 983–988.

[30] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 1103–1108.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 797–806.