SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News

Limeng Cui Suhang Wang Dongwon Lee

The Pennsylvania State University, PA, USA
Email: {lzc334, szw494, dongwon}@psu.edu

***Abstract*—How to effectively detect fake news and prevent its diffusion on social media has gained much attention in recent years. However, relatively little focus has been given on exploiting user comments left for posts and latent sentiments therein in detecting fake news. Inspired by the rich information available in user comments on social media, therefore, we investigate whether the latent sentiments hidden in user comments can potentially help distinguish fake news from reliable content. We incorporate users' latent sentiments into an end-to-end deep embedding framework for detecting fake news, named as SAME. First, we use multi-modal networks to deal with heterogeneous data modalities. Second, to learn semantically meaningful spaces per data source, we adopt an adversarial mechanism. Third, we define a novel regularization loss to bring embeddings of relevant pairs closer. Our comprehensive validation using two real-world datasets, PolitiFact and GossipCop, demonstrates the effectiveness of SAME in detecting fake news, significantly outperforming state-of-the-art methods.**

***Index Terms*—Fake news detection, social media, multi-modal.**

I. Introduction

Fake news is a type of false information to deliberately mislead or manipulate public opinion, through traditional mass media and recent online social media. In recent years, due to the explosive growth of online social media, fake news for different political agenda and commercial gains has been coming out in a great amount and widespread online. During the US president election in 2016, for instance, fake news has caused a significant social impact on the election results. For example, "Pizzagate", a scandal that never was true, quickly went viral on multiple social media platforms. A report on the 2016 election indicates that fake news websites rely on online social media for 48% of traffic, which is a much higher share than that of other sources [1]. Therefore, to mitigate the problems of fake news, how to detect it effectively has become an important research problem, which will be the main task of this paper.

ASONAM '19, August 27-30, 2019, Vancouver, Canada

ACM ISBN 978-1-4503-4993-2/17/07?/$15.00
http://dx.doi.org/10.1145/3110025.31100XX

Existing methods for detecting fake news can be generally categorized into two categories based on the heterogeneity of the data, i.e., single-modal based and multi-modal based. In single-modal based methods, single type of, often textual, information such as contents, profiles and descriptions are used. For instance, [2] exploits the linguistic features of misinformation by comparing real news with fake news. Similarly, [3] conducts fake news detection by evaluating the consistency between the body and its claim given a news article. Note that as the content type of news is *not* limited to only text, other data types such as images or videos could also be utilized. In particular, in social media, fake news often comes with multi-modality data including manipulated images, fake videos, or user comments, all of which provide rich information for detecting fake news. As such, multi-modality based fake news detection has gained increased attentions. For example, [4] proposes a Recurrent Neural Network (RNN) with an attention mechanism to fuse multi-modal data from tweets for rumor detection. In addition, [5] proposes the Event Adversarial Neural Networks (EANN), which integrates multi-modal features of images and texts and removes event-specific features via discriminator.

In addition to the issue of modality, another important idea is to exploit the latent sentiment in user comments. Although user's viewpoint has been proved to be useful in fake news detection [6], there are few studies on the impact of user's sentiment. User's comments such as "`I agree..she is a rock star`" or "`No. Its a fake news story specifically targeting 'conservative readers'.`", may potentially add/remove different degrees of credibility to the news in question. Therefore, toward the detection of fake news, we propose to explore to employ both the sentiment analysis in user comments as well as multi-modal fake news data.

In an attempt to solve this problem, these are several challenges. As for incorporating user's sentiment into a detection procedure with multi-modal data, the first step is to represent a user. Each user may comment on or "like" a particular type of news. Such a representation can be measured by the correspondence between user's historical interest and type of news. However, this problem is technically difficult for two reasons. On one hand, the learned features of user's representation are usually high dimensional and sparse, which cannot be processed by traditional methods. On the other hand,

as each modality has an intrinsically different distribution, it is challenging to fuse user's representation with others. For example, a user's sentiment representation is sparse while the image feature is naturally dense, causing a mismatch.

Overcoming these challenges, in this paper, we present a novel method, named as Sentiment-Aware Multi-modal Embedding (SAME), with the emphasis on both sentiment and multi-modality. We propose to use an end-to-end deep architecture to mitigate the heterogeneity introduced by multi-modal data and capture the representation of user's sentiment better. To be specific, first, we use different networks to deal with the triplet relationship among news publishers, users, and news. Second, an adversarial mechanism is introduced to preserve the semantic similarity and enforces the representation consistency between different modalities. To our best knowledge, this is the first attempt to utilize adversarial learning to find semantic correlations between different modalities in news contents. Third, we model a user's sentiment and incorporate it into the proposed framework.

Our main contributions are as follows:

- We propose an end-to-end deep framework to integrate different features of news contents for fake news detection. An adversarial mechanism is added to preserve semantic relevance and the representation consistency across different modalities.
- We validate the effectiveness of user sentiment through statistical analysis and use users' sentimental polarities to facilitate fake news detection.
- We empirically demonstrate that our proposed method, SAME, significantly outperforms five state-of-the-art baselines in detecting fake news on social media using two real-world benchmark datasets.

II. Related Work

In this section, we briefly review two related topics, i.e., fake news detection and sentiment analysis.

A. Fake News Detection

In recent years, researchers have proposed a number of methods for fake news detection. Interested readers are referred to [7], [8] for further information. From the perspective of information used, the fake news detection methods can be roughly divided into two categories: single-modal and multi-modal based methods.

1) Single-modal based Methods: For single-modal based methods, existing works [2], [3], [9], [10] mainly analyze the textual contents of news, including the headline and news content, and extract the characteristics of fake news. Some researchers use methods in linguistics to distinguish the fake news from the real ones. Others check the consistency between the news title and content. For example, Rashkin et al. [2] specifically focus on political coverage verification and fake news detection. They propose to exploit the linguistic features of misinformation by comparing real news with fake news such as satire, hoaxes, and propaganda. Jin et al. [9] assume the images plays a very important role in the news propagation on microblog. The distribution patterns between images of real and fake news are quite different. Thus, they extract the image features from two aspects, including visual content and statistics. In literature [3], Bhatt et al. conduct fake news detection by evaluating the consistency between the body and its claim given a news article. Statistical and external features are used to build a unified classifier for fake news detection. As the content type of news is not limited to text, the above methods do not fully exploit the multi-modal data such as image, video, and network. Thus, they do not yield satisfying results compared with multi-modal based methods.

2) Multi-modal based Methods: In social media, besides the textual features, news often includes various types of data, which provides more comprehensive features for detecting fake news. Thus, investigating multi-modal data for fake news detection is attracting increasing attention [4]–[6], [11]–[14]. A survey on different content types of news and their impacts on readers can be found in [15].

In general, multi-modal based fake news detection focus on extracting features from news content, including news publisher, textual contents and image/video. By using the three types of features mentioned, different kinds of news representations can be built, which capture discriminative aspects of news. In multimedia based methods, researchers usually use deep networks to capture both visual and textual information of news and apply classification models to distinguish fake news from the real ones. In literature [4], the authors propose an attention based Recurrent Neural Network to fuse the multi-modal data from tweets for rumor detection. An attention mechanism is added to find the correlations between images and texts. The architecture of Event Adversarial Neural Networks (EANN) is proposed in literature [5]. Both text and image in an article are taken into consideration. The authors train an event discriminator in order to eliminate the effects of the event-specific features and maintain the common features among all these studied events.

Despite the success of multi-modal based fake news detection approaches, few of them explicitly model user sentiments towards news for fake news detection; while sentiments are very strong signal which have great potential for improving fake news detection performance. Therefore, in this paper, we investigate a novel problem of exploring user sentiments for fake news detection with multi-modal data.

B. Sentiment Analysis

Users opinions or sentiments towards posts or products in social media have been demonstrated to be very effective for many social media mining tasks such as user rating prediction [16], [17], recommender system [18] and stock movement prediction [19]. Detecting user sentiments or stances has become a popular task. In literature [20] authors conduct user's belief classification and in literature [21] authors conduct stance detection. Zhang et al. [10] focus on the news stance detection. The proposed model takes the headline and body of an article, and generates the probabilities of four news stances including "agree", "disagree", "discuss" and "unrelated". The

TABLE I
THE STATISTICS OF THE TWO REAL-WORLD DATASETS.

Dataset	Politifact	GossipCop
# Real News	624	16,817
# Fake News	432	5,323
# User	558,937	1,390,131
# User Replies	552,698	379,996

TABLE II
THE SENTIMENT POLARITY DISTRIBUTION UNDER NEWS.

		Negative	Neutral	Positive
PolitiFact	Fake News	12.6	73.2	14.2
	Real News	9.6	77.9	12.5
GossipCop	Fake News	9.8	69.2	21.0
	Real News	8.9	74.4	16.7

authors use ranking-based to address the problem brought by classification-based algorithms that a clear distinction exists between any two stances. In literature [22], the authors predict the stance of a set of texts representing facts with respect to a given claim by using end-to-end memory networks.

As sentiment features have shown promising results in improving the performance of news stance detection, we introduce sentiment features into the fake news detection task.

III. PRELIMINARY DATA ANALYSIS

Users can express their emotions or opinions, through comments such as sensational or skeptical reactions [23]. These features are useful when detecting fake news. In this section, we conduct preliminary data analysis to demonstrate that users' sentiments towards real news and fake news are statistically different, which lays a foundation for integrating sentiments for fake news detection. Next, we first introduce the datasets followed by preliminary data analysis.

A. Datasets

For preliminary data analysis, we adopt two widely used multi-modal fake news detection datasets, i.e., PolitiFact and GossipCop, which are publicly available from a fake news dataset repository FakeNewsNet[1] [23]. For both datasets, each news entity contains news content, corresponding images, users' retweets/replies and news profile (source, publisher and keywords). Each news has 0 to 1,000 user comments. Some users didn't leave a comment when they retweet, so we excluded such kind of user engagement data.

PolitiFact is a fact-checking website that targets on political news. It rates the authenticity of claims by elected officials and others. The two datasets are crawled from Twitter in order to get users' comments. It contains 432/624 (fake/real) news.

GossipCop is a fact-checking website for celebrity reporting. It investigates the credibility of entertainment stories on magazines, newspapers and social media, to ascertain whether they are real or not. It contains of 5,323/16,817 (fake/real) news.

The statistics of the datasets are summarized in Table I.

B. User Sentiment Analysis Toward Fake and Real News

Intuitively, the comments under fake news can be roughly divided into three classes: (1) Agree (from users who believe in the news); (2) Discuss (from users who doubt the authenticity of the news); and (3) Disagree that the original news is false news (from users who do not believe in the news).

Usually, the first and third types of comments contain polarized emotions ("Negative" and "Positive"), which can be seen from **User1** and **User4** in Figure 1. The second type of comments does not contain such strong emotions. The sentiment is more neutral in skeptical comments or discussions.

Here we perform the sentiment analysis on the users comments with VADER[2] [24], which is a lexicon and rule-based tool to predict the sentiment expressed on social media. For each news piece, we obtain all the replies for this news and apply VADER to predict the sentiment as negative, neutral or positive. As can be seen from Table II, users' comments under fake news often contain more sentiment polarities and are less neutral.

To statistically verify our observation, we conduct hypothesis testing. Positive, neutral and negative sentiment polarity are defined by VADER. For each dataset, two equal-sized collections of tweets are chosen. Each of them contains 50 tweets and each tweet has at least 50 comments. One collection contains the comments randomly selected from fake news, while the other contains comments randomly selected from real news. We use two vectors $\mathbf{s}_f$ and $\mathbf{s}_r$ to denote the sentiment polarities of two groups respectively, where the sentiment polarity is the sum of positive and negative sentiment polarity. A two-sample one-tail t-test is conducted to validate whether there is sufficient statistical correlation to support the hypothesis that the sentiment polarity of the first collection is greater than that of the second.

Let μ_f be the mean of sentiment polarities of the comment in the fake news collection and μ_r the mean of real news. The null hypothesis is H_0, and the alternative hypothesis is H_1. Here the hypothesis of interest can be expressed as:

$$\begin{aligned} H_0 &: \mu_f - \mu_r \leq 0 \\ H_1 &: \mu_f - \mu_r > 0 \end{aligned} \tag{1}$$

The results show that there is statistical evidence on Politifact dataset, with $t = -1.6927$, $df = 98$, $p-value = 0.04684$ to reject the H_0 hypothesis, which is the evidence that the sentiment polarity of comments under fake news is greater than under real news. And we also find statistical evidence on GossipCop dataset, with $t = -2.1012$, $df = 98$, $p-value = 0.01909$.

[1]https://github.com/KaiDMML/FakeNewsNet

[2]https://github.com/cjhutto/vaderSentiment

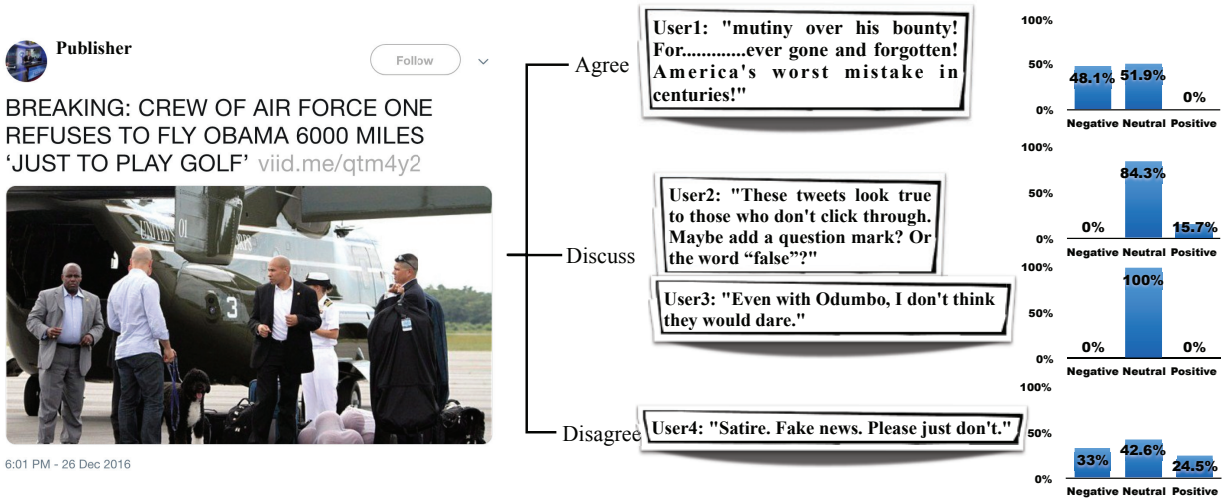

Fig. 1. Sentiment polarity distribution of different stances ("Agree", "Discuss" and "Disagree") in users' comments.

IV. Proposed Method

As we have validated the impact of user's sentiment, in this section, we introduce the proposed multi-modal embedding model by incorporating such information for fake news detection. In multi-modal fake news, we have four objects: news image, content, profile and user comments. The news is multi-modal data which consists of three modalities. Assume that we have N training pairs $\mathbf{D} = \{\mathbf{T}_i, r_i\}_{i=1}^{N}$ in which $\mathbf{T}_i$ denotes news i and $r_i \in \{0, 1\}$ denotes its ground truth label. Further, let $\mathbf{T}_i = (\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$, where $\mathbf{x}_i$ denotes the feature vectors of the image modality, $\mathbf{y}_i$ denotes the feature vectors of the text modality and $\mathbf{z}_i$ can be the one-hot code of news profile related to news i. In addition, we are also given a similarity matrix $\mathbf{S}$, where S_{ij} evaluates the similarity of news i and news j. The similarity is defined by the shared user's sentiment. For example, we can say that news i and news j are similar if they share multiple sentiment words.

We first introduce how to learn the latent news presentation from the multi-modal news data by learning a joint embedding function $f(\mathbf{T}_i)$ map the news to space R^M, where different modalities are distributed consistently. To preserve the similarity matrix $\mathbf{S}$, the distance between embedding vectors $\mathbf{h}_i = f(\mathbf{T}_i)$ and $\mathbf{h}_j = f(\mathbf{T}_j)$ should be small if S_{ij} is relatively large. Thus, a hybrid similarity loss is proposed to embed the user's sentiment into the model. The objective is to maximize the similarity between similar news triplets and minimize it between all dissimilar news triplets. Finally, once each data source is mapped to R^M, we use the embedding vector $\mathbf{h}_i$ to predict the news label r_i.

A. Multi-Modal Feature Extractor

The hybrid deep architecture for learning multi-modal correlation embedding is shown in Fig. 2, which accepts input in a triplet of news image, content and profile, and processes them through deep network: (1) three different networks to deal with the triplet including news image, content and profile; (2) adversarial mechanism to enforce the same distribution between different modalities; and (3) a novel hybrid similarity loss to model the user's sentiment and incorporate it into the proposed framework.

We built the image network based on VGGNet [25], which is pre-trained on the ImageNet database [26]. To fit CNN into our SAME model, we reserve the first seven layers and replace the eighth layer by a layer with R nodes, $\mathbf{fch}^i$. As for the text network, we use GloVe [27] to process text $\mathbf{y}$, in order to capture the complex characteristics of word use (e.g., syntax and semantics). The obtained text representation is used as the input of the text network. We then adopt the Multi-Layer Perceptron (MLP) comprising two fully connected layers. The second layer $\mathbf{fch}^t$ has R hidden units, which transforms the network activation to R-dimensional representation. As for profile information, the features are discrete values such as topic. So we use the one-hot encoding to represent the profile $\mathbf{z}$, and feed it to a two-layer fully-connected MLP, and get the representation $\mathbf{fch}^p$. As for the adversarial networks, we built the discriminator networks by using a three-layer feed-forward neural network.

To integrate the three networks mentioned above, a fully connected layer with M hidden units, which takes the representations of three networks as input, is added on top of the architecture. We denote the multi-modal feature extractor for news i as $f^{(v)}(\mathbf{T}_i^{(v)}; \theta_a) \in \mathsf{R}^M$, which corresponds to the output of the hybrid deep network for multi-modal correlation embedding. Here, θ_a is the network parameter to be learned.

B. Adversarial Learning

With the above network, however, different modalities are usually distributed inconsistently, which is not beneficial if we use the concatenation for fake news detection. In order to bridge this modality gap, we introduce an adversarial learning mechanism. We use two discriminators for image and profile modalities to investigate their distributions. For the image

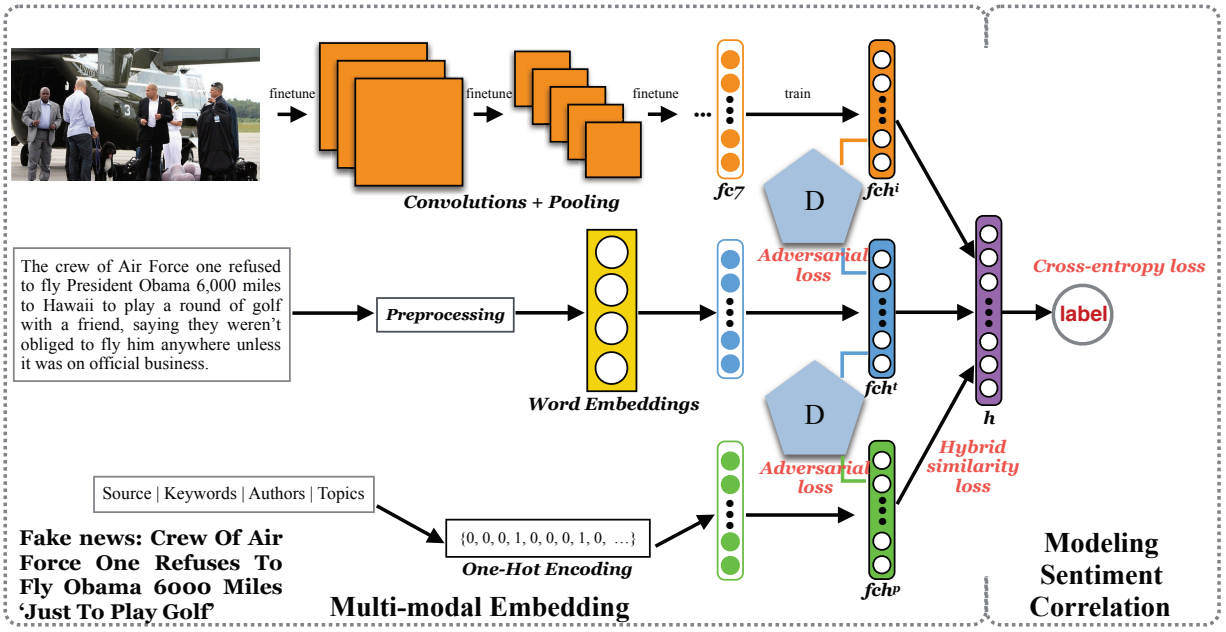

Fig. 2. Multi-modal Embedding (SAME) accepts input in a triplet of news publisher, user and news, and processes them through deep network: (1) three different networks to deal with the triplet including news publishers, users, and news; (2) adversarial mechanism to enforce the same distribution between different modalities; and (3) a novel hybrid similarity loss to model the users representation and incorporate it into the proposed framework.

(profile) discriminator, the inputs are image (profile) features and text features obtained from the feature extractor, and the output is a binary label, either "0" or "1". Specifically, we denote the modality label for the textual feature that has been generated from text network as "1" and define the modality label for image (profile) semantic modality features generated from image network (profile network) as "0". We feed the outputs of image and text network into one discriminator and feed the outputs of profile and text networks into the other discriminator. The loss functions of the two discriminators can be defined as $\mathcal{L}_a^i$ and $\mathcal{L}_a^p$. The two discriminators act as the two adversaries while we are training the SAME.

The loss function $\mathcal{L}_a^i$ can be written as follows:

$$\min_{\theta_c} \mathcal{L}_a^i = \sum_{j=1}^{2N} ||D^{i,t}(\mathbf{fch}_j) - \mathbf{d}_j||_2^2, \tag{2}$$

where $\mathbf{fch}_j$ is semantic features obtained from image network or text network, the modality label is $\mathbf{d}_j$. Specifically we have $\mathbf{d}_j^i = 0$ denoting the modality label for image and $\mathbf{d}_j^t = 1$ denoting the label for text. The result of Eqn. (2) is that the discriminator acts as a binary classifier $D^{i,t}(\mathbf{fch}_j; \theta_c)$, classifying the input features into class "1" and class "0". Similarly, we have $\mathcal{L}_a^p$.

The above idea motivates a MinMax game between the feature extractor and the event discriminator. On one hand, the feature extractor tries to fool the modality that the discriminator tries to maximize the discrimination loss. On the other hand, the modality discriminator tries to discover the modality-specific information included in the feature representations to recognize the modality label. In the experiments (section V-D), we demonstrate the effectiveness of the adversarial learning in detecting fake news.

C. Modeling Sentiment Correlation

In order to make the learned joint embeddings maximally preserve the similarity information, we propose a novel hybrid similarity loss by considering such two issues: (1) entity triplets with lower similarity should be separated and have discriminative embeddings; (2) entity triplets should have similar embeddings only if they are similar in the original feature spaces.

To address the first issue, we propose the *Graph Affinity Metric* between news i and news j. The *Graph Affinity Metric* is defined as follows

Definition 1. Let G_{ij} denotes the similarity of sentiment polarity distribution between the comments of news i and j. We can define the *Graph Affinity Metric* between two news as G_{ij}.

Then, we define the *Local Similarity Metric* to ensure the local similarity in each news to ensure the second issue above.

Definition 2. The *Local Similarity Metric* $L_{ij,m}(m = 1, 2, 3)$ of each modality involves the local similarity information. On modality $\mathbf{x}$, we have

$$L_{ij,1}^{(v)} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathbf{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathbf{N}_k(\mathbf{x}_i) \\ 0. & \text{otherwise} \end{cases}$$

where $\mathbf{N}_k(\cdot)$ denotes the set of k-nearest neighbors. Similarly, we have $L_{ij,2}$ and $L_{ij,3}$ defined on modalities $\mathbf{y}$ and $\mathbf{z}$ respectively.

According to our empirical study, we set the number of nearest neighbors to 5 throughout this paper.

To maintain the similarity between entities and preserve the local structural information in the common embedding space, we propose a hybrid similarity loss loss which ensures the learned embedding space meaningful:

$$\min_{\theta_a} \mathcal{L}_c = \frac{1}{2} \sum_{i,j=1}^{N} S_{ij} \|\mathbf{h}_i - \mathbf{h}_j\|_2^2 \tag{3}$$

where $S_{ij} = G_{ij} + L_{ij,1} + L_{ij,2} + L_{ij,3}$.

D. Fake News Detector

In this section, we introduce how to detect fake news by using the M-dimensional embedding. We use a fully connected layer with softmax, which is shown in Fig. 2. Each network takes embedding vectors $\mathbf{h}_i$of news i as input.

We have a training set $\{r_i\}_{i=1}^{N}$, where $r_i \in \{0, 1\}$ denotes the ground truth label of news i. The goal is to find a set of prediction function g, such that the label for any news i can be predicted. We denote the fake news detector as $g^{(v)}(f^{(v)}(\mathbf{T}_i^{(v)}; \theta_a); \theta_b) \in \mathsf{R}$, where θ_b is the network parameter of the network for fake news detector.

Assume the ranking score is modeled as $\hat{r}_i = [\hat{r}_{i,0}, \hat{r}_{i,1}]$, with $\hat{r}_{i,0}$ and $\hat{r}_{i,1}$ indicate the predicted probability of label being 0 (real news) and 1 (fake news) respectively. r_i denotes the ground truth label of news. Thus, for each news, the goal is to minimize the cross-entropy loss function as follows:

$$\min_{\theta_a, \theta_b} \mathcal{L}_q = -r_i \log(\hat{r}_{i,1}) - (1 - r_i) \log(1 - \hat{r}_{i,0}) \tag{4}$$

E. The Proposed Method: SAME

During the training, the feature extractor and the fake news detector work together to minimize the detection loss $\mathcal{L}_q$. Simultaneously, the feature extractor tries to fool the discriminator to get distribution agreement for different modalities by maximizing the adversarial loss $\mathcal{L}_a^i$ and $\mathcal{L}_a^p$. The

The final objective function of the proposed SAME is:

$$\begin{aligned} \mathcal{J}_g &= \mathcal{L}_c + \gamma \mathcal{L}_q \\ \mathcal{J}_a &= \mathcal{L}_a^i + \mathcal{L}_a^p \end{aligned} \tag{5}$$

where γ is a penalty parameter for trading off the relative importance of multi-modal correlation and news label. We set $\gamma = 1$ based on empirical study.

If we put them together, we can obtain:

$$\begin{aligned} (\theta_a, \theta_b) &= arg \min_{\theta_a, \theta_b} \mathcal{J}_g(\theta_a, \theta_b) - \mathcal{J}_a(\hat{\theta}_c) \\ \theta_c &= arg \max_{\theta_c} \mathcal{J}_g(\hat{\theta}_a, \hat{\theta}_b) - \mathcal{J}_a(\theta_c) \end{aligned} \tag{6}$$

All the parameters in the network are learned through RMSprop, which has been widely used among existing methods. It is an adaptive learning rate method which divides the learning rate by an exponentially decaying average of squared gradients.

V. Experimental Validation

In this section, we conduct experiments to demonstrate the effectiveness of the proposed method SAME. We first describe experimental settings. We then compare SAME against several state-of-the-art baselines for fake news detection followed by ablation study to understand the contribution of each component of SAME. The experiments are conducted on two real-world datasets, PolitiFact and GossipCop, introduced in Section III-A.

A. Compared Methods

We compare SAME with several representative and state-of-the-art fake news detection methods including KNN, SVM, TCNN-URG[3] [12], EANN[4] [5] and CSI[5] [12]. Our implementation of SAME is available here[6].

KNN: This determines the authenticity of news based on the labels of its neighbors, defined in **Definition 2**.

SVM: We concatenate the features including the outputs of VGGNet, GloVe and one-hot encoding, and sentiment polarity distribution vector as the input of Linear SVM. We choose Linear SVM as it is suitable for high-dimensional data.

TCNN-URG: this method exploits the user's historical responses to related articles as soft semantic labels. TCNN generates the representation for each article, which is used for further news classification. URG is trained to learn the users responses to news articles, which can help the classification procedure of TCNN when users response is not available in early detection.

EANN: In this method, both text and image information are taken into consideration. This method uses an event discriminator in order to eliminate the effects of the event-specific features and maintain the common features among all these studied events. We remove the event discriminator of this method as our datasets do not have event labels.

CSI: This method explores all of news content, users responses to the news, and the sources that users promote in detecting fake news. However, as our datasets do not have time interval information in users' comment, we modify the codes accordingly.

For KNN, we set $k = 5$ based on empirical study. We use $C = 1$ in Linear SVM. As for other baseline methods, we use the parameter settings in the paper or in the released source code. For our method, we implemented it using Keras. The news image is re-sized to 128×128 pixels. The image network is pre-trained on the ImageNet classification task [26]. We fine-tune CONV1-FC7 initialized from the pre-trained model, and train layer FCH via back-propagation. Each news content is processed through GloVe. For the text network, we use a two-layer neural network, in which the first layer has 4,096 ReLU

[3]We implemented the code by ourselves.
[4]https://github.com/yaqingwang/EANN-KDD18
[5]https://github.com/sungyongs/CSI-Code
[6]https://github.com/cuilimeng/SAME

TABLE III
PERFORMANCE COMPARISON ON THE TWO DATASETS. THE BEST RESULTS ARE LISTED IN BOLD.

		Training Ratio							
		20%		40%		60%		80%	
Datasets	Measure	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
PolitiFact	KNN	45.30	35.25	56.83	53.87	60.24	55.91	56.53	53.84
	SVM	53.50	51.42	60.74	56.83	64.37	59.39	65.57	60.56
	TCNN-URG	64.53	60.53	68.35	61.50	70.24	67.41	72.35	70.64
	EANN	63.53	59.42	67.93	63.88	70.22	65.65	71.31	69.38
	CSI	65.42	63.42	67.35	65.29	69.64	67.12	74.24	73.24
	SAME	**69.12**	**68.23**	**69.24**	**65.34**	**73.24**	**75.42**	**77.24**	**76.31**
GossipCop	KNN	59.24	56.24	55.46	53.54	54.31	59.32	57.20	53.37
	SVM	56.42	56.58	54.24	57.34	55.24	57.24	61.24	62.34
	TCNN-URG	66.22	62.42	65.33	62.24	67.42	63.42	73.24	68.43
	EANN	65.91	63.62	67.24	65.13	70.23	69.23	71.21	72.24
	CSI	72.35	71.53	74.24	72.24	76.42	74.82	77.24	76.87
	SAME	**76.24**	**76.42**	**78.24**	**75.61**	**77.24**	**78.31**	**80.42**	**81.58**

TABLE IV
COMPARISON OF VARIANTS OF SAME ON TWO DATASETS. THE BEST RESULTS ARE LISTED IN BOLD.

		Training Ratio							
		20%		40%		60%		80%	
Datasets	Measure	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
PolitiFact	SAME w/o I	63.24	56.24	66.52	65.76	69.38	64.73	73.24	71.68
	SAME w/o D	65.74	63.24	65.13	64.31	68.13	67.91	74.86	72.61
	SAME w/o S	60.37	61.42	63.29	63.88	63.24	62.78	70.85	69.54
	SAME	**69.12**	**68.23**	**69.24**	**65.34**	**73.24**	**75.42**	**77.24**	**76.31**
GossipCop	SAME w/o I	71.53	69.53	73.24	72.24	74.24	72.72	75.24	73.23
	SAME w/o D	70.93	71.84	73.15	72.04	75.14	73.93	77.79	75.37
	SAME w/o S	65.67	64.82	67.71	67.93	73.39	71.01	75.91	73.37
	SAME	**76.24**	**76.42**	**78.24**	**75.61**	**77.24**	**78.31**	**80.42**	**81.58**

units with dropout rate 0.5. The news profile is represented by one-hot encoding, which is fed into a two-layer neural network as well. We fix mini-batch size as 128, and set learning rate as 0.001.

B. Evaluation Metrics

As the data is imbalanced, following the common way, we use Macro F1 and Micro F1 as evaluation metrics. Macro Precision is the average precision of all classes, similarly, Macro Recall is the average recall of all classes. Macro F1 is the harmonic mean of Macro Precision and Macro Recall. Macro F1 calculates metrics for each label, and uses their unweighted mean. It does not take label imbalance into account. However, Micro F1 does not calculate on each class, it calculates metrics by counting the total true positives, false negatives and false positives globally.

C. Performance Comparison

We predict the score of the authenticity of news on two datasets respectively. We randomly select x% of data for training and the remaining (100-x)% for testing. To fully understand how SAME performs under different data size, we vary x as $\{20, 40, 60, 80\}$. The process is performed for 5 times and the average performance is reported in Table III. From the experimental results, we make the following observations:

For SVM method, through it concatenates all the features together. However, the results are far from satisfactory. We assume that the features used are highly nonlinear, simple concatenation may cause dense features to dominate the feature space and override the effects of the sparse ones.

For other baseline methods, the information used is not comprehensive (including visual, textual, profile and sentimental features), so the effects are not as good as SAME.

Compared against the best baseline, SAME achieves an absolute increase of 2.8%/3.0% on average in terms of Macro F1 and 4.0%/4.1% on average in terms of Micro F1 on two datasets. This clearly demonstrates that SAME is able to leverage heterogeneous data signals while integrating sentiments for effective fake news detection.

D. Ablation Study

In this section, we conduct an ablation study to fully understand the contribution of each component in SAME. We remove several critical modules in SAME that process images, news profile, and social sentiment (and their corresponding discriminator and loss function) as follows:

SAME without image data (SAME w/o I): this method removes the images network.

SAME without discriminators (SAME w/o D): this methods removes the two discriminators.
SAME without users' sentiment information (SAME w/o S): social sentiment is removed from the proposed model.
SAME: this method is the proposed method, which incorporates not only the three multi-modal networks, but also the sentiment information from users' comments.

We report the Macro F1 and Micro F1 on both datasets in Table IV. We can observe that all the components: visual and textual features, social context features and adversarial mechanism are indispensable for achieving the best performance of SAME. Different components can provide complementary information, which also verifies the effectiveness of our proposed framework.

VI. Conclusion and Future Work

In this paper, we investigate a novel problem of exploring sentiment for fake news detection with multi-modal data. We first use statistical analysis to test the hypothesis in order to validate the effectiveness of users sentiment. Then, we propose a new deep multi-modal embedding architecture for fake news detection, which unifies multi-modal data with adversarial learning and incorporates users sentiment. The experimental results demonstrate the effectiveness of our method as well as the roles of user's sentiment in fake news detection. In addition, we also examine the necessity of each module in the proposed method and thus test the fusion network proposed. The outcome of this work not only has significant contribution in building a machine-based solution for detecting fake news, but also has a far-reaching impact on society by helping improve the quality of information.

There are several interesting directions that need further investigation. First, to mitigate the problem of fake news better, extending SAME to be able to do the early detection is important yet challenging (due to the lack of important signals). Second, most of current fake news detection methods solely focus on the detection. However, in addition to the detecting fake news, being able to "explain" why one is fake news is equally important.

Acknowledgments

This work was in part supported by NSF awards #1742702, #1820609, and #1915801, and ORAU-directed R&D program in 2018.

References

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[2] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *EMNLP*, 2017, pp. 2931–2937.

[3] G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal, "Combining neural, statistical and external features for fake news stance identification," in *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, pp. 1353–1357.

[4] 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 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 795–816.

[5] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *KDD*. ACM, 2018, pp. 849–857.

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

[7] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. American Society for Information Science, 2015, p. 82.

[8] 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.

[9] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia*, vol. 19, no. 3, pp. 598–608, 2017.

[10] Q. Zhang, E. Yilmaz, and S. Liang, "Ranking-based method for news stance detection," in *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, pp. 41–42.

[11] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *CIKM*. ACM, 2017, pp. 797–806.

[12] F. Qian, C. Gong, K. Sharma, and Y. Liu, "Neural user response generator: Fake news detection with collective user intelligence." in *IJCAI*, 2018, pp. 3834–3840.

[13] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause, "Fake news detection in social networks via crowd signals," in *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, pp. 517–524.

[14] L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *WSDM*. ACM, 2018, pp. 637–645.

[15] S. B. Parikh and P. K. Atrey, "Media-rich fake news detection: A survey," in *MIPR*. IEEE, 2018, pp. 436–441.

[16] X. Lei, X. Qian, and G. Zhao, "Rating prediction based on social sentiment from textual reviews," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1910–1921, 2016.

[17] D. Tang, B. Qin, T. Liu, and Y. Yang, "User modeling with neural network for review rating prediction." in *IJCAI*, 2015, pp. 1340–1346.

[18] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks." in *AAAI*, 2015, pp. 1721–1727.

[19] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015.

[20] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *EMNLP*. Association for Computational Linguistics, 2011, pp. 1589–1599.

[21] P. Sobhani, S. Mohammad, and S. Kiritchenko, "Detecting stance in tweets and analyzing its interaction with sentiment," in *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 2016, pp. 159–169.

[22] M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Màrquez, and A. Moschitti, "Automatic stance detection using end-to-end memory networks," in *NAACL-HLT*, vol. 1, 2018, pp. 767–776.

[23] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media," *arXiv preprint arXiv:1809.01286*, 2018.

[24] C. H. E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *ICWSM*, 2014.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[27] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.