EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection

Yaqing Wang[1], Fenglong Ma[1], Zhiwei Jin[2], Ye Yuan[3]
Guangxu Xun[1], Kishlay Jha[1], Lu Su[1], Jing Gao[1]
[1]Department of Computer Science, State University of New York at Buffalo, Buffalo, New York
[2]Institute of Computing Technology, CAS, University of Chinese Academy of Sciences, Beijing, China
[3]College of Information and Communication Engineering, Beijing University of Technology, Beijing, China
[1]{yaqingwa, fenglong, guangxux, kishlayj, lusu, jing}@buffalo.edu
[2]jinzhiwei@ict.ac.cn, [3]yuanye91@emails.bjut.edu.cn

ABSTRACT

As news reading on social media becomes more and more popular, fake news becomes a major issue concerning the public and government. The fake news can take advantage of multimedia content to mislead readers and get dissemination, which can cause negative effects or even manipulate the public events. One of the unique challenges for fake news detection on social media is how to identify fake news on newly emerged events. Unfortunately, most of the existing approaches can hardly handle this challenge, since they tend to learn event-specific features that can not be transferred to unseen events. In order to address this issue, we propose an end-to-end framework named Event Adversarial Neural Network (EANN), which can derive event-invariant features and thus benefit the detection of fake news on newly arrived events. It consists of three main components: the multi-modal feature extractor, the fake news detector, and the event discriminator. The multi-modal feature extractor is responsible for extracting the textual and visual features from posts. It cooperates with the fake news detector to learn the discriminable representation for the detection of fake news. The role of event discriminator is to remove the event-specific features and keep shared features among events. Extensive experiments are conducted on multimedia datasets collected from Weibo and Twitter. The experimental results show our proposed EANN model can outperform the state-of-the-art methods, and learn transferable feature representations.

CCS CONCEPTS

• **Information systems** → **Data mining**; *Web mining*;

KEYWORDS

Fake news detection, adversarial neural networks, deep learning

ACM Reference Format:
Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3219819.3219903

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219903

1 INTRODUCTION

The recent proliferation of social media has significantly changed the way in which people acquire information. Nowadays, there are increasingly more people consuming news through social media, which can provide timely and comprehensive multimedia information on the events taking place all over the world. Compared with traditional text news, the news with images and videos can provide a better storytelling and attract more attention from readers. Unfortunately, this is also taken advantage by fake news which usually contain misrepresented or even forged images, to mislead the readers and get rapid dissemination. The dissemination of fake news may cause large-scale negative effects, and sometimes can affect or even manipulate important public events. For example, within the final three months of the 2016 U.S. presidential election, the fake news generated to favor either of the two nominees was believed by many people and was shared by more than 37 million times on Facebook [1, 7]. Therefore, it is in great need of an automatic detector to mitigate the serious negative effects caused by the fake news.

Thus far, various fake news detection approaches, including both traditional learning [6, 15, 29] and deep learning based models [21, 25], have been exploited to identify fake news. With sufficient verified posts on different events, existing deep learning models have achieved performance improvement over traditional ones due to their superior ability of feature extraction. However, they are still not able to handle the unique challenge of fake news detection, i.e., detecting fake news on newly emerged and time-critical events [27]. Due to lack of the corresponding prior knowledge, the verified posts about such events can be hardly obtained in a timely manner, which leads to the unsatisfactory performance of existing models. Actually, existing models tend to capture lots of event-specific features which are not shared among different events. Such event-specific features, though being able to help classify the posts on verified events, would hurt the detection with regard to newly emerged events. For this reason, instead of capturing event-specific features, we believe that *learning the shared features among all the events* would help us with the detection of fake news from unverified posts. Therefore, the goal of this work is to design an effective model to remove the

nontransferable event-specific features and preserve the shared features among all the events for the task of identifying fake news.

To remove event-specific features, the first step is to identify them. For posts on different events, they have their own unique or specific features that are not sharable. Such features can be detected by measuring the difference among posts corresponding to different events. Here the posts can be represented by the learned features. Thus, identifying event-specific features is equivalent to measuring the difference among learned features on different events. However, it is a technically challenging problem. First, since the learned feature representations of posts are high-dimensional, simple metrics like the squared error may not be able to estimate the differences among such complicated feature representations. Second, the feature representations keep changing during the training stage. This requires the proposed measurement mechanism to capture the changes of feature representations and consistently provide the accurate measurement. Although this is very challenging, the effective estimation of dissimilarities among the learned features on different events is the premise of removing event-specific features. Thus, how to effectively estimate the dissimilarities under this condition is the challenge that we have to address.

In order to address the aforementioned challenges, we propose an end-to-end framework referred to as Event Adversarial Neural Networks (EANN) for fake news detection based on multi-modal features. Inspired by the idea of adversarial networks [10], we incorporate the event discriminator to predict the event auxiliary labels during training stage, and the corresponding loss can be used to estimate the dissimilarities of feature representations among different events. The larger the loss, the lower the dissimilarities. Since the fake news takes advantage of multimedia content to mislead readers and gets spread, our model needs to handle the multi-modal inputs. The proposed model EANN consists of three main components: the multi-modal feature extractor, the fake news detector, and the event discriminator. The multi-modal feature extractor cooperates with the fake news detector to carry out the major task of identifying false news. Simultaneously, the multi-modal feature extractor tries to fool the event discriminator to learn the event invariant representations. For multi-modal feature extractor, We employ Convolutional Neural Networks (CNN) to automatically extract features from both textual and visual content of posts. Experimental results on two large scale real-world social media datasets show that the proposed EANN model outperforms the state-of-the-art approaches.

The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to propose fake news detection for new and time-critical events, which can identify fake news based on multi-modal features and learn transferable features by removing the event-specific features. Towards this end, we propose an end-to-end event adversarial neural networks.
- The proposed EANN model uses event discriminator to measure the dissimilarities among different events, and further learns the event invariant features which can generalize well for the newly emerged events.
- Our proposed EANN model is a general framework for fake news detection. The integrated multi-modal feature extractor can be easily replaced by different models designed for feature extractions.
- We empirically show that the proposed EANN model can effectively identify fake news and outperform the state-of-the-art multi-modal fake news detection models on two large scale real world datasets.

The rest of the paper is organized as follows: related literature survey is summarized in Section 2, the details of the proposed framework are introduced in Section 3, experimental results are presented in Section 4, and the study is concluded in Section 5.

2 RELATED WORK

In this section, we briefly review the work related to the proposed model. We mainly focus on the following two topics: fake news detection and adversarial networks.

2.1 Fake News Detection

There are many tasks related to fake news detection, such as rumor detection [14] and spam detection [26]. Following the previous work [25, 27], we specify the definition of fake news as news which is intentionally fabricated and can be verified as false. In fake news detection task, the main challenge is how to distinguish news according to features. The features can be extracted from posts, social context, and even attached images. Thus, we review existing work from the following two categories: single modality based and multi-modal fake news detection.

Single Modality based Fake News Detection. *Textual features* are statistical or semantic features extracted from text content of posts, which have been explored in many literatures of fake news detection [4, 11, 19, 27]. Unfortunately, linguistic patterns are not yet well understood, since they are are highly dependent on specific events and corresponding domain knowledge [25]. Thus, it is difficult to design hand-crafted textual features for traditional machine learning based fake news detection models. To overcome this limitation, Ma et al. [21] propose a deep learning model to identify fake news. Specifically, it deploys recurrent neural networks to learn the representations of posts in a time series as textual features. Experiments results show the effectiveness of deep learning based models.

Visual features have been shown to be an important indicator for fake news detection [15, 27]. However, very limited studies are conducted on verifying the credibility of multimedia content on social media. The basic features of attached images in the posts are explored in the work [12, 15, 24, 31]. However, these features are still hand-crafted and can hardly represent complex distributions of visual contents.

Social context features represent the user engagements of news on social media [27] such as the number of followers, hash-tag(#) and retweets. In [31], the authors aim to capture propagation patterns such as graph structure of the message propagation. However, social context features are very noisy, unstructured and labor intensive to collect. Especially, it cannot provide sufficient information for newly emerged events.

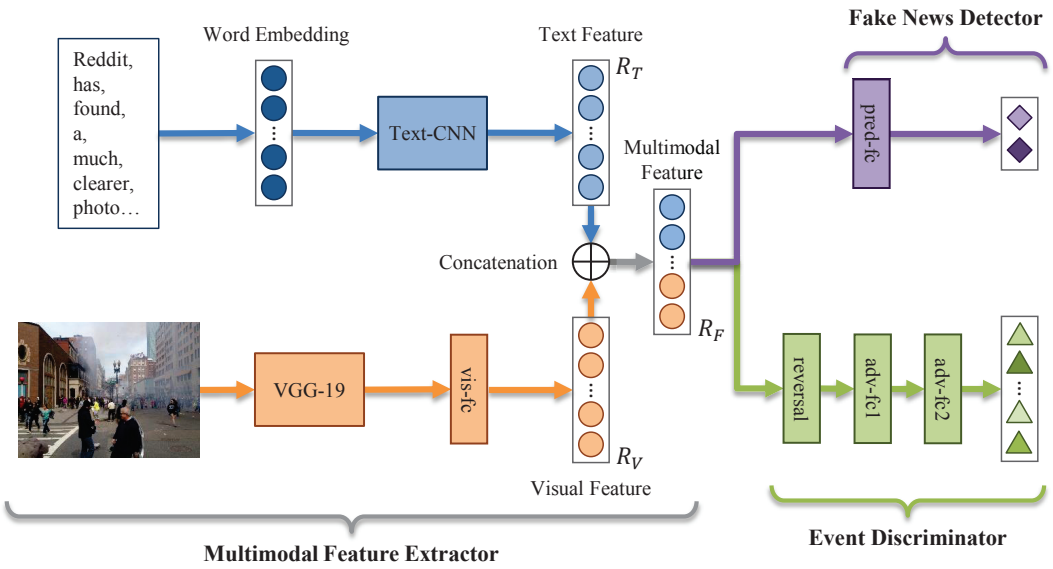

Figure 1: The architecture of Event Adversarial Neural Networks (EANN). The blue colored network is the textual feature extractor, the orange colored network is visual feature extractor, the fake news detector is purple colored, and event discriminator is green colored.

Different from all the aforementioned work, in this paper, we consider multiple types of features simultaneously when identifying fake news on social media.

Multi-modal Fake News Detection. To learn feature representations from multiple aspects, deep neural networks have been successfully applied to various tasks, including but not limited to visual question answering [2], image captioning [17, 30], and fake news detection [13]. In [13], the authors propose a deep learning based fake news detection model, which extracts the multi-modal and social context features and fuses them by attention mechanism. However, the multi-modal feature representations are still highly dependent on specific events in the dataset, and cannot generalize very well to identify fake news on new coming events.

To overcome the limitations of existing work, we propose a novel deep learning model, which significantly improves the performance on fake news detection on different events. The proposed model not only automatically learns multi-modal feature representations, but also generates event invariant feature representations using an adversarial network.

2.2 Adversarial Networks

Our work is also inspired by the idea of adversarial networks [10]. Existing adversarial networks are usually used to generate images which can match the observed samples by a minimax game framework. The adversarial learning framework has been adopted to several tasks, such as learning representations for semi-supervised learning [23], predication of sleep stages [32], discriminative image features [20] and domain adaption [8, 9]. The proposed model also sets up a minimax game between event discriminator and multi-modal feature extractor. In particular, the multi-modal feature extractor is enforced to learn an event invariant representation to fool the discriminator. In this way, it removes tight dependencies on the specific events in the collected dataset and achieves better generalization ability on the unseen events.

3 METHODOLOGY

In this section, we first introduce the three components of the proposed EANN model: the multimodal feature extractor, the fake news detector, and the event discriminator, then describe how to integrate these three components to learn the transferable feature representations. The detailed algorithm flow is also shown in the last subsection.

3.1 Model Overview

The goal of our model is to learn the transferable and discriminable feature representations for fake news detection. As shown in Figure 1, in order to achieve this, the proposed EANN model integrates three major components: the multi-modal feature extractor, the fake news detector, and the event discriminator. First of all, since the posts on social media usually contain information in different modalities (e.g., textual post and attached image), the multi-modal feature extractor includes both textual and visual feature extractors to handle different types of inputs. After the textual and visual latent feature representations are learned, they are concatenated together to form the final multi-modal feature representation. Both of the fake news detector and the event discriminator are built on top of the multi-modal feature extractor. The fake news detector takes the learned feature representation as input to predict whether the posts are fake or real. The event discriminator identifies the event label of each post based on this latent representation.

3.2 Multi-Modal Feature Extractor

3.2.1 Textual Feature Extractor. The sequential list of the words in the posts is the input to the textual feature extractor. In order to

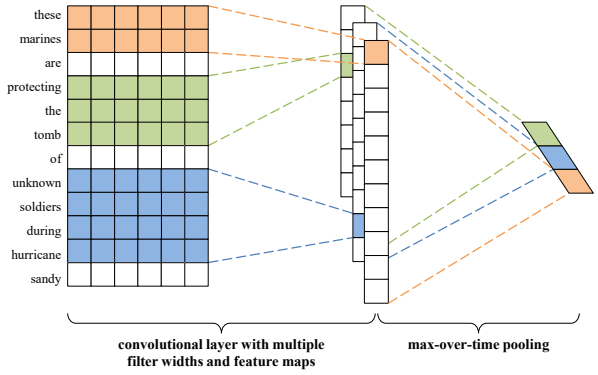

Figure 2: The architecture of Text-CNN.

extract the informative features from textual content, we employ convolutional neural networks (CNN) as the core module of our textual feature extractor. CNN has been proven to be effective in many fields such as computer vision and text classification [5, 16]. As can be seen in Figure 1, we incorporate a modified CNN model, namely Text-CNN [18], in our textual feature extractor. The architecture of Text-CNN is shown in Figure 2. As seen, it takes advantage of multiple filters with various window sizes to capture different granularities of features to identify fake news.

For detailed procedures of the textual feature extractor, each word in the text is represented as a word embedding vector. The embedding vector for each word is initialized with the pre-trained word embedding on the given dataset. For the i-th word in the sentence, the corresponding k dimensional word embedding vector is denoted as $T_i \in \mathbb{R}^k$. Thus, a sentence with n words can be represented as:

$$T_{1:n} = T_1 \oplus T_2 \oplus ... \oplus T_n, \tag{1}$$

where $\oplus$ is the concatenation operator. A convolutional filter with window size h takes the contiguous sequence of h words in the sentence as input and outputs one feature. In order to show the procedure clearly, we take the contiguous sequence of h words starting with the i-th word as example, the filter operation can be represented as:

$$t_i = \sigma(W_c \cdot T_{i:i+h-1}). \tag{2}$$

Here $\sigma(\cdot)$ is the ReLU activation function and W_c represents the weight of the filter. The filter can also be applied to the rest of words and then we get a feature vector for this sentence:

$$t = [t_1, t_2, ..., t_{n-h+1}]. \tag{3}$$

For every feature vector t, we use max-pooling operation to take the maximum value so as to extract the most important information. Now, we get the corresponding feature for one particular filter. The process is repeated until we get the features for all filters. In order to extract textual features with different granularities, various window sizes are applied. For a specific window size, we have n_h different filters. Thus, assuming there are c possible window sizes, we have $c \cdot n_h$ filters in total. The textual features after the max-pooling operation is written as $R_{T_c} \in \mathbb{R}^{c \cdot n_h}$. Following the max-pooling operations, a fully connected layer is used to ensure the final textual feature representation (denoted as $R_T \in \mathbb{R}^p$) has the same dimension (denoted as p) as the visual feature representation through the following operation:

$$R_T = \sigma(W_{tf} \cdot R_{T_c}), \tag{4}$$

where W_{tf} is the weight matrix of the fully connected layer.

3.2.2 Visual Feature Extractor. The attached images of the posts are inputs to the visual feature extractor and are denoted as V. In order to efficiently extract visual features, we employ the pre-trained VGG19 [28]. On top of the last layer of VGG19 network, we add a fully connected layer to adjust the dimension of final visual feature representation to p. During the joint training process with the textual feature extractor, the parameters of pre-trained VGG19 neural network are kept static to avoid overfitting. Denoting p dimensional visual feature representation as $R_V \in \mathbb{R}^p$, the operation of the last layer in the visual feature extractor can be represented as:

$$R_V = \sigma(W_{vf} \cdot R_{V_{vgg}})), \tag{5}$$

where $R_{V_{vgg}}$ is the visual feature representation obtained from pre-trained VGG19, and W_{vf} is the weight matrix of the fully connected layer in the visual feature extractor.

The textual feature representation R_T and visual feature representation R_V will be concatenated to form the multi-modal feature representation denoted as $R_F = R_T \oplus R_V \in \mathbb{R}^{2p}$, which is the output of the multi-modal feature extractor. We denote the multi-modal feature extractor as $G_f(M; \theta_f)$ where M, which is usually a set of textual and visual posts, is the input to the multi-modal feature extractor, and θ_f represents the parameters to be learned.

3.3 Fake News Detector

In this subsection, we introduce the fake news detector. It deploys a fully connected layer with softmax to predict whether the posts are fake or real. The fake news detector is built on top of the multi-modal feature extractor, thus taking the multi-modal feature representation R_F as input. We denote the fake news detector as $G_d(\cdot\,; \theta_d)$, where θ_d represents all the parameters included. The output of the fake news detector for the i-th multimedia post, denoted as m_i, is the probability of this post being a fake one:

$$P_\theta(m_i) = G_d(G_f(m_i; \theta_f); \theta_d). \tag{6}$$

The goal of the fake news detector is to identify whether a specific post is fake news or not. We use Y_d to represent the set of labels and employ cross entropy to calculate the detection loss:

$$L_d(\theta_f, \theta_d) = -\mathbb{E}_{(m,y)\sim(M,Y_d)}\,[y \log(P_\theta(m)) + (1-y)(\log(1-P_\theta(m))]. \tag{7}$$

We minimize the detection loss function $L_d(\theta_f, \theta_d)$ by seeking the optimal parameters $\hat{\theta}_f$ and $\hat{\theta}_d$, and this process can be represented as:

$$(\hat{\theta}_f, \hat{\theta}_d) = arg \min_{\theta_f, \theta_d} L_d(\theta_f, \theta_d). \tag{8}$$

As previously discussed, one of the major challenges for fake news detection stems from the events that are not covered by the training dataset. This requires us to be able to learn the transferable feature representations for newly emerged events. Direct minimization of detection loss only helps detect fake news on the events included in the training dataset, since this captures only

event-specific knowledge (e.g., keywords) or patterns, which cannot generalize well. Thus, we need to enable the model to learn more general feature representations that can capture the common features among all the events. Such representation should be event-invariant and does not include any event-specific features. To achieve this goal, we need to remove the uniqueness of each event. In particular, we measure the dissimilarities of the feature representations among different events and remove them in order to capture the event invariant feature representations.

3.4 Event Discriminator

Event discriminator is a neural network which consists of two fully connected layers with corresponding activation functions. It aims to correctly classify the post into one of K events based on the multi-modal feature representations. We denote the event discriminator as $G_e(R_F;\theta_e)$ where θ_e represents its parameters. We define the loss of event discriminator by cross entropy and use Y_e to represent the set of the event labels:

$$L_e(\theta_f,\theta_e) = -\mathbb{E}_{(m,y)\sim(M,Y_e)}\,[\sum_{k=1}^{K}\mathbf{1}_{[k=y]}\log(G_e(G_f(m;\theta_f));\theta_e)], \tag{9}$$

The parameters of event discriminator minimizing the loss $L_e(\cdot,\cdot)$ are written as:

$$\hat{\theta}_e = arg\min_{\theta_e} L_e(\theta_f,\theta_e). \tag{10}$$

The above loss $L_e(\theta_f,\hat{\theta}_e)$ can be used to estimate the dissimilarities of different events' distributions. The large loss means the distributions of different events' representations are similar and the learned features are event-invariant. Thus, in order to remove the uniqueness of each event, we need to maximize the discrimination loss $L_e(\theta_f,\hat{\theta}_e)$ by seeking the optimal parameters θ_f.

The above idea motivates a minimax game between the multi-modal feature extractor and the event discriminator. On one hand, the multi-modal feature extractor tries to fool the event discriminator to maximize the discrimination loss, and on the other hand, the event discriminator aims to discover the event-specific information included in the feature representations to recognize the event. The integration process of three components and the final objective function will be introduced in the next subsection.

3.5 Model Integration

During the training stage, the multi-modal feature extractor $G_f(\cdot;\theta_f)$ needs to cooperate with fake news detector $G_d(\cdot;\theta_d)$ to minimize the detection loss $L_d(\theta_f,\theta_d)$, so as to improve the performance of fake news detection task. Simultaneously, the multi-modal feature extractor $G_f(\cdot;\theta_f)$ tries to fool the event discriminator $G_e(\cdot;\hat{\theta}_e)$ to achieve event invariant representations by maximizing the event discrimination loss $L_e(\theta_f,\theta_e)$. The event discriminator $G_e(R_F;\theta_e)$ tries to recognize each event based on the multi-modal feature representations by minimizing the event discrimination loss. We can define the final loss of this three-player game as

$$L_{final}(\theta_f,\theta_d,\theta_e) = L_d(\theta_f,\theta_d) - \lambda\, L_e(\theta_f,\theta_e), \tag{11}$$

where λ controls the trade-off between the objective functions of fake news detection and event discrimination. In this paper, we simply set λ as 1 without tuning the trade-off parameter. For the minimax game, the parameter set we seek is the saddle point of the final objective function:

$$(\hat{\theta}_f,\hat{\theta}_d) = arg\min_{\theta_f,\theta_d} L_{final}(\theta_f,\theta_d,\hat{\theta}_e), \tag{12}$$

$$\hat{\theta}_e = arg\max_{\theta_e} L_{final}(\hat{\theta}_f,\theta_e). \tag{13}$$

We use stochastic gradient descent to solve the above problem. The θ_f is updated according to Eq. 14. Here we adopt the gradient reversal layer (GRL) introduced in [8]. The gradient reversal layer acts as an identity function during forward stage, and it multiplies gradient with $-\lambda$ and passes the results to the preceding layer during backprop stage. GRL can be easily added between the multi-modal feature extractor and the event discriminator. We denote it as the reversal layer in the Figure 1.

$$\theta_f \leftarrow \theta_f - \eta\,(\frac{\partial L_d}{\partial\theta_f} - \lambda\frac{\partial L_e}{\partial\theta_f}). \tag{14}$$

In order to stabilize the training process, we follow the approach in [8] to decay the learning rate η:

$$\eta' = \frac{\eta}{(1+\alpha\cdot p)^{\beta}}, \tag{15}$$

where $\alpha = 10$, $\beta = 0.75$, and p is linearly changing from 0 to 1 corresponding to the training progress. The detailed steps of the proposed event adversarial neural networks (EANN) is summarized in algorithm 1.

Algorithm 1 Event Adversarial Neural Networks.

Input: The multi-modal input $\{m_i\}_{i=1}^N$, the auxiliary event label $\{e_i\}_{i=1}^N$, the detection label $\{y_i\}_{i=1}^N$ and the learning rate η

1: **for** number of training iterations **do**
2: Decay learning rate according to Eq. 15
3: Update the parameters of multi-modal feature extractor θ_f according to Eq. 14;
4: Update the parameters of the event discriminator θ_e:
5: $\theta_e \leftarrow \theta_e - \eta\,\frac{\partial L_e}{\partial\theta_e}$
6: Update the parameters of fake news detector θ_d:
7: $\theta_d \leftarrow \theta_d - \eta\,\frac{\partial L_d}{\partial\theta_d}$
8: **end for**

4 EXPERIMENTS

In this section, we first introduce two large social media datasets used in the experiments, then present the state-of-the-art fake news detection approaches, and finally analyze the performance of the proposed model.

4.1 Datasets

To fairly evaluate the performance of the proposed model, we conduct experiments on two real social media datasets, which are collected from Twitter and Weibo. Next, we provide the details of both datasets respectively.

Twitter Dataset

The Twitter dataset is from MediaEval Verifying Multimedia Use benchmark [3], which is used for detecting fake content on Twitter. This dataset has two parts: the development set and test set. We

Table 1: The Statistics of the Real-World Datasets.

Method	Twitter	Weibo
# of fake News	7898	4749
# of real News	6026	4779
# of images	514	9528

use the development as training set and test set as testing set to keep the same data split scheme. The tweets in the Twitter dataset contain text content, attached image/video and additional social context information. In this work, we focus on detecting fake news by incorporating both text and image information. Thus, we remove the tweets without any text or image. For this two sets, there is no overlapping events among them. For model training on Twitter dataset, we adopt early stop strategy.

Weibo Dataset

The Weibo dataset is used in [13] for detecting fake news. In this dataset, the real news are collected from authoritative news sources of China, such as Xinhua News Agency. The fake news are crawled from May, 2012 to January, 2016 and verified by the official rumor debunking system of Weibo. This system encourages common users to report suspicious posts and examines suspicious posts by a committee of trusted users. According to the previous work [21, 31], this system also acts as the authoritative source for collecting rumor news. When preprocessing this dataset, we follow the same steps in the work [13]. We first remove the duplicated and low quality images to ensure the quality of entire dataset. Then we apply a single-pass clustering method [14] to discover newly emerged events from posts. Finally, we split the whole datasets into the training, validation, testing sets in a 7:1:2 ratio, and ensure that they do not contain any common event. The detailed statistics of these two datasets are listed in Table 1.

4.2 Baselines

To validate the effectiveness of the proposed model, we choose baselines from the following three categories: single modality models, multi-modal models, and the variant of the proposed model.

Single Modality Models

In the proposed model, we leverage both text and image information to detect fake news. For each modality, it can also be solely used to discover fake news. Thus, we proposed the following two simple baselines:

• **Text**. We use 32 dimensional pre-trained word-embedding weights of text content from all of posts to initialize the parameters of the embedding layer. Then CNN is used to extract the textual feature R_T for each post. Finally, an additional fully connected layer with softmax function is used to predict whether this post is fake or not. We use 20 filters with window size ranging from 1 to 4, and the hidden size of fully connected layer is 32.

• **Vis**. The input of **Vis** is an image. Pre-trained VGG-19 and a fully connected layer are used to extract the visual feature R_V. Then, R_V is fed into a fully connected layer to make prediction. We set the hidden size of fully connected layer as 32.

Multi-modal Models

All the Multi-modal approaches take the information from multiple modalities into account, including VQA [2], NeuralTalk [30] and att-RNN [13].

• **VQA** [2]. Visual Question Answering (**VQA**) model aims to answer the questions based on the given images. The original VQA model is designed for multi-class classification tasks. In this work, we focus on binary classification. Thus, when implementing VQA model, we replace the final multi-class layer with the binary-class layer. Besides, for fair comparison, we use one-layer LSTM, and the hidden size of LSTM is 32.

• **NeuralTalk** [30]. NeuralTalk is a model to generate captions for the given images. The latent representations are obtained by averaging the outputs of RNN at each timestep, and then these representations are fed into a fully connected layer to make prediction. The hidden size of both LSTM and the fully connected layer is 32.

• **att-RNN** [13]. att-RNN is the state-of-the-art model for multi-modal fake news detection. It uses attention mechanism to fuse the textual, visual and social context features. In our experiments, we remove the part dealing with social context information, but the remaining parts are the same. The parameter settings are the same as [13].

A Variant of the Proposed EANN

The complete EANN model consists of three components: multi-modal feature extractor, fake news detector and event discriminator. Only using multi-modal feature extractor and fake news detector, we still can detect fake news. Thus, we design a variant of the proposed model, named **EANN**−. In EANN−, we do not include the event discriminator.

4.3 Implementation Details

In the textual feature extractor, we set $k = 32$ for dimensions of word-embedding. We set $n_h = 20$, and the window size of filters varies from 1 to 4 in Text-CNN. The hidden size of the fully connected layer in textual and visual extractor is 32. For fake news detector, the hidden size of the fully connected layer is 64. The event discriminator consists of two fully connected layers: the hidden size of first layer is 64, and the hidden size of second layer is 32. For all the baselines and the proposed model, we use the same batch size of 100 instances in the training stages, and the training epoch is 100.

4.4 Performance Comparison

Table 2 shows the experimental results of baselines and the proposed approaches on two datasets. We can observe that the overall performance of the proposed EANN is much better than the baselines in terms of *accuracy*, *precision* and *F1 score*.

On the Twitter dataset, the number of tweets on different events is imbalanced and more than 70% of tweets are related to a single event. This causes the learned text features mainly focus on some specific events. Compared with visual modality, the text modality contains more obvious event specific features which seriously prevents extracting transferable features among different events for the Text model. Thus, the accuracy of Text is the lowest among all the approaches. As for another single modality baseline Vis, its performance is much better than that of Text. The features of image are more transferable, and thus reduce the effect of imbalanced posts. With the help of VGG19, a powerful tool for extracting useful

Table 2: The results of different methods on two datasets.

Dataset	Method	Accuracy	Precision	Recall	F_1
Twitter	Text	0.532	0.598	0.541	0.568
	Vis	0.596	0.695	0.518	0.593
	VQA	0.631	0.765	0.509	0.611
	NeuralTalk	0.610	0.728	0.504	0.595
	att-RNN	0.664	0.749	0.615	0.676
	EANN−	0.648	0.810	0.498	0.617
	EANN	**0.715**	**0.822**	**0.638**	**0.719**
Weibo	Text	0.763	0.827	0.683	0.748
	Vis	0.615	0.615	0.677	0.645
	VQA	0.773	0.780	0.782	0.781
	NeuralTalk	0.717	0.683	**0.843**	0.754
	att-RNN	0.779	0.778	0.799	0.789
	EANN−	0.795	0.806	0.795	0.800
	EANN	**0.827**	**0.847**	0.812	**0.829**

features, we can capture the more sharable patterns contained in images to tell the realness of news compared with textual modality.

Though the visual modality is effective for fake news detection, the performance of Vis is still worse than that of the multi-modal approaches. This confirms that integrating multiple modalities is superior for the task of fake news detection. Among multi-modal models, att-RNN performs better than VQA and NeuralTalk, which shows that applying attention mechanism can help improve the performance of the predictive model.

For the variant of the proposed model EANN−, it does not include the event discriminator, and thus tends to capture the event-specific features. This would lead to the failure of learning enough shared features among events. In contrast, with the help of the event discriminator, the complete EANN significantly improves the performance in terms of all the measures. This demonstrates the effectiveness of the event discriminator for performance improvements. Specifically, the accuracy of EANN improves 10.3% compared with the best baseline att-RNN, and F1 scores increases 16.5%.

On the Weibo dataset, similar results can be observed as those on the Twitter dataset. For single modality approaches, however, contradictory results are observed. From Table 2, we can see that the performance of Text is greatly higher than that of Vis. The reason is that the Weibo dataset does not have the same imbalanced issue as the Twitter dataset, and with sufficient data diversity, useful linguistic patterns can be extracted for fake news detection. This leads to learning a discriminable representation on the Weibo dataset for the textual modality. On the other hand, the images in the Weibo dataset are much more complicated in semantic meaning than those in the Twitter dataset. With such challenging image dataset, the baseline Vis cannot learn meaningful representations, though it uses the effective visual extractor VGG19 to generate feature representations.

As can be seen, the variant of the proposed model EANN- outperforms all the multi-modal approaches on the Weibo dataset. When

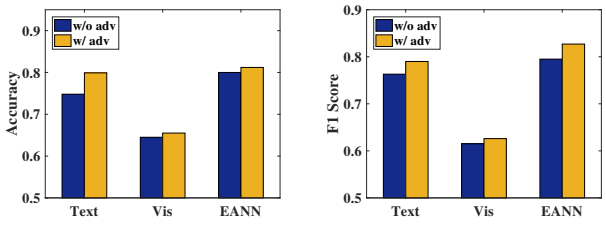

Figure 3: The performance comparison for the models w/ and w/o adversary.

modeling the text information, our model employs convolutional neural networks with multiple filters and different word window sizes. Since the length of each post is relatively short (smaller than 140 characters), CNN may capture more local representative features.

For the proposed EANN, it outperforms all the approaches on accuracy, precision and F1 score. Compared with EANN−, we can conclude that using the event discriminator component indeed improves the performance of fake news detection.

4.5 Event Discriminator Analysis

In this subsection, we aim to analyze the importance of the designed event discriminator component from the quantitative and qualitative perspectives.

Quantitative Analysis

To intuitively illustrate the importance of employing event discriminator in the proposed model, we conduct the following experiments. For each single modality approach, we design its corresponding adversarial model. Then we run the new designed model on the Weibo dataset. Figure 3 shows the results in terms of F1 score and accuracy. In Figure 3, "w/ adv" means that we add event discriminator into the corresponding approaches, and "w/o adv" denotes the original approaches. For the sake of simplicity, let Text+ and Vis+ represent the corresponding approaches, Text and Vis, with event discriminator component being added, respectively.

From Figure 3, we can observe that both accuracy and F1 score of Text+ and Vis+ are greater than those of Text and Vis respectively. Note that for the proposed approach EANN, its reduced model is EANN−. The comparison between EANN and EANN− has been discussed in Section 4.4. Thus, we can draw a conclusion that incorporating event discriminator component is essential and effective for the task of fake news detection.

Qualitative Analysis

To further analyze the effectiveness of event discriminator, we qualitatively visualize the text features R_T learned by EANN− and EANN on the Weibo testing set with t-SNE [22] shown in Figure 4. The label for each post is real or fake.

From Figure 4, we can observe that for the approach EANN−, it can learn discriminable features , but the learned features are still twisted together, especially for the left part of Figure 4a. In contrast, the feature representations learned by the proposed model EANN are more discriminable, and there are bigger segregated areas

(a) EANN− (b) EANN

Figure 4: Visualizations of learned latent text feature representations on the testing data of Weibo.

among samples with different labels shown in Figure 4b. This is because in the training stage, event discriminator tries to remove the dependencies between feature representations and specific events. With the help of the minimax game, the muli-modal feature extractor can learn invariant feature representations for different events and obtain more powerful transfer ability for detection of fake news on new events. The comparison between EANN− and EANN proves that the proposed approach learns better feature representations with the component of event discriminator, and thus achieves better performance.

4.6 Case Studies for Multiple Modalities

In order to illustrate the importance of considering multi-modal features for fake news detection, we compare the results reported by the proposed EANN and single modality feature models (Text and Vis), and report the fake tweets correctly detected by EANN but missed by the single modality feature models.

(a) Five headed snake

(b) Photo: Lenticular clouds over Mount Fuji, Japan. #amazing #earth #clouds #mountains

Figure 5: Some fake news detected by EANN but missed by single text modality model on the Twitter dataset.

We first show two top-confident tweets which are successfully detected by the proposed model but missed by single textual modality model in Figure 5. The text content do not show evidence to identify that the tweets are fake. For both of the examples in Figure 5, they describe the images with common patterns. The textual modality model Text also identifies this news as a real one. Although the experts may be engaged to verify the text content using their domain knowledge, this option may not be available for normal readers. As seen, the two attached images look quite suspicious and are very likely to be forged pictures. By feeding visual content and textual content into the proposed EANN, both tweets are classified as fake with high confidence scores. This shows that the proposed model EANN obtains some clues from the attached images to make correct classification. The additional visual content provides more information for fake news detection beyond single textual modality.

(a) Want to help these unfortunates? New, Iphones, laptops, jewelry and designer clothing could aid them through this!

(b) Meet The Woman Who Has Given Birth To 14 Children From 14 Different Fathers!

Figure 6: Some fake news detected by EANN but missed by single image modality model on the Twitter dataset.

Figure 6 shows another two examples missed by image modality model Vis but successfully spotted by the proposed EANN model. For the first example, the complicated semantic meaning is contained in the attached image, which is challenging to be captured by the visual feature extractor. However, the words with strong emotion and inflammatory intention suggest this is a suspicious post. By combining textual and visual content of tweets, the proposed EANN can easily detect that this is fake news with high confidence. The attached image in the second example looks very normal, but the corresponding textual description seems to misrepresent the image and mislead the readers. Without the textual content, the meaning of the tweets would totally change. Only aligned with the corresponding text description, it can be identified as fake news. The visual modality model Vis does not classify this example as false, but with the help of multi-modal features, the proposed EANN model gives the high confidence in detecting this fake news.

4.7 Convergence Analysis

In order to explore the training process of the proposed EANN model, the development of training, testing and discrimination loss (adversarial losses) has been shown in Figure 7. At the beginning, all of the three losses decrease. Then the discrimination loss increases and stabilizes at a certain level. The decreasing discrimination loss in the beginning represents the event discriminator detecting the event specific information included in the feature representations of multi-modal feature extractor. As the minimax game between the discriminator and the feature extractor is continuing, the feature representations tend to be event invariant. Thus, the event specific information is removed incrementally, and the discrimination loss increases over the time. During the training process, the three losses smoothly converge, which means that a certain level of equilibrium have been achieved. As the training loss decreases steadily, we can observe that the testing loss also decreases steadily, and a very similar pattern of trend is shown. This observation proves that the feature representations learned

by the proposed EANN can capture the general information among all the events, and this representation is also discriminative even on new coming events.

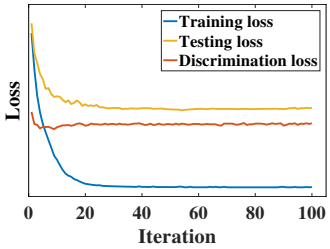

Figure 7: The training, testing and event discrimination loss development.

5 CONCLUSIONS

In this work, we study the problem of multi-modal fake news detection. The major challenge of fake news detection stems from newly emerged events on which existing approaches only showed unsatisfactory performance. In order to address this issue, we propose a novel event adversarial neural network framework which can learn transferable features for unseen events. Specifically, our proposed model consists of three main components, i.e., multi-modal feature extractor, event discriminator, and fake news detector. The multi-modal extractor cooperates with fake news detector to learn the discriminable representations for identifying fake news, and simultaneously learns the event invariant representations by removing the event-specific features. Extensive experiments on two large scale dataset collected from popular social media platforms show that our proposed model is effective and can outperform the state-of-the-art models.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work is supported in part by the US National Science Foundation under grants CNS-1742845. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.
[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.
[3] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. 2015. Verifying Multimedia Use at MediaEval 2015.. In *MediaEval*.
[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.
[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, Aug (2011), 2493–2537.
[6] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.
[7] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. 2017. Fake news mitigation via point process based intervention. *arXiv preprint arXiv:1703.07823* (2017).
[8] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. 1180–1189.
[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.
[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
[11] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.
[12] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 153–164.
[13] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 795–816.
[14] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 230–239.
[15] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia* 19, 3 (2017), 598–608.
[16] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* (2014).
[17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.
[18] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).
[19] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 1103–1108.
[20] Zachary C Lipton and Subarna Tripathi. 2017. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782* (2017).
[21] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks.. In *IJCAI*. 3818–3824.
[22] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.
[23] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).
[24] Dong ping Tian et al. 2013. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering* 8, 4 (2013), 385–396.
[25] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 797–806.
[26] Hua Shen, Fenglong Ma, Xianchao Zhang, Linlin Zong, Xinyue Liu, and Wenxin Liang. 2017. Discovering social spammers from multiple views. *Neurocomputing* 225 (2017), 49–57.
[27] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[28] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[29] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).
[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 3156–3164.
[31] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 651–662.
[32] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*. 4100–4109.