BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection

Tong Zhang[1, 2], Di Wang[2, 3], Huanhuan Chen[4, 3], Zhiwei Zeng[2], Wei Guo[5, 6], Chunyan Miao[7, 2, 3], and Lizhen Cui[5, 6]

[1]School of Computer Science and Technology, Shandong University, China
[2]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, Singapore
[3]Joint NTU-WeBank Research Centre on Fintech, Nanyang Technological University, Singapore
[4]School of Computer Science and Technology, University of Science and Technology of China, China
[5]School of Software, Shandong University, China
[6]Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, China
[7]School of Computer Science and Engineering, Nanyang Technological University, Singapore
Email: mr_t@mail.sdu.edu.cn; wangdi@ntu.edu.sg; hchen@ustc.edu.cn; i160001@e.ntu.edu.sg; guowei@sdu.edu.cn; ascymiao@ntu.edu.sg; clz@sdu.edu.cn

***Abstract*—Nowadays, with the rapid growth of microblogging networks for news propagation, there are increasingly more people accessing news through such emerging social media. In the meantime, fake news now spreads at a faster pace and affects a larger population than ever before. Compared with traditional text news, the news posted on microblog often has attached images in the context. So how to correctly and autonomously detect fakes news in a multi-modal manner becomes a prominent challenge to be addressed. In this paper, we propose an end-to-end model, named BERT-based domain adaptation neural network for multi-modal fake news detection (BDANN). BDANN comprises three main modules: a multi-modal feature extractor, a domain classifier and a fake news detector. Specifically, the multi-modal feature extractor employs the pretrained BERT model to extract text features and the pretrained VGG-19 model to extract image features. The extracted features are then concatenated and fed to the detector to distinguish fake news. The role of the domain classifier is mainly to map the multi-modal features of different events to the same feature space. To assess the performance of BDANN, we conduct extensive experiments on two multimedia datasets: Twitter and Weibo. The experimental results show that BDANN outperforms the state-of-the-art models. Moreover, we further discuss the existence of noisy images in the Weibo dataset that may affect the results.**

***Keywords*—Fake news detection, Multimedia, Natural language processing, Data mining, Deep learning**

I. Introduction

Social media platforms such as Twitter and Weibo have become the main stream for people to publish, access and propagate news. According to the statistics shown in [1] and [2], as of the second quarter in 2019, the number of Twitter's and Weibo's monthly active users across the world had already reached over 330 million and 480 million, respectively. Moreover, because anyone may publish posts with minimum authentication, the credibility of news in social platforms becomes crucial. According to [3], people are more likely to believe fake news even if they do not align with the viewer's political inclination. For example, an analysis by BuzzFeed [4] found that the top 20 fake news stories about the 2016 U.S. presidential election received more engagement on Facebook than the top 20 election stories from 19 major media outlets [5], which inevitably affected the fairness of the election process. To improve the credibility of information posted on microblogs and prevent the propagation of fake contents, it is now crucial to correctly and autonomously detect rumors on microblogs.

In addition to texts in tweets and Weibo posts, images have become popular on microblogs. Compared with texts, images present visual contents and thus augment the textual content and attract more attention [6]. Meanwhile, images could also help to distinguish rumors due to their rich visual information. Fig. 1 presents three examples of fake news taken from the Twitter dataset (see Section IV for more details), where each example comprises the text article and the attached image. As shown in Fig. 1(a), we may easily identify that both the image and text article are fabricated. In Fig. 1(b), we may not tell from the text alone that it is fake but the morphed image suggests that it is possibly a piece of fake news. In Fig. 1(c), the image does not present deterministic information but the text suggests that this piece of news may be fake. The examples shown in Fig. 1 motivate us to address the fake news detection challenge from a multi-modal manner, instead of looking at the text or image alone.

This research is supported, in part, by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-003), the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017), and the Joint NTU-WeBank Research Centre on Fintech (NWJ-2019-002), Nanyang Technological University, Singapore. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on October 04,2020 at 22:13:12 UTC from IEEE Xplore. Restrictions apply.

There's a new Olympic game called Turd Dropping in Sochi.

(a)

The incredibly rare Black Lion, only a few of these exist...

(b)

Mysterious 'ghost' boots appear in photo of girl near samurai graveyard.

(c)

Fig. 1: Fake news examples taken from the Twitter dataset.

Recently, the focus of the fake news detection techniques has shifted from text-based to multi-modal. Pioneer studies [7], [8], [9] were developed to validate the authenticity of online text posts based on manually crafted text and social context features. Ma et al. [10] introduced a recurrent neural network (RNN) to extract temporal representations of microblog posts for fake news detection. Following Ma et al.'s work, Chen et al. [11] incorporated the attention mechanism into RNN to selectively extract temporal representations. Recent studies focus on using deep learning methods to autonomously extract both text and image features to detect fake news. Jin et al. [12] proposed a variant of RNN (att-RNN) extract multi-modal features. Wang et al. [13] introduced an end-to-end framework, which extracts event-invariant features that are beneficial to the detection of fake news on unforeseen events. Khattar et al. [14] introduced a bi-modal variational autoencoder coupled with a binary classifier for fake news detection.

In this paper, to further improve the accuracy of fake news detection with multi-modal features, we propose an end-to-end model, named BERT-based domain adaptation neural network (BDANN). BDANN comprises three main modules: a multi-modal feature extractor, a domain classifier and a fake news detector. Specifically, the multi-modal feature extractor employs the pretrained BERT_{base} model to extract text features and the pretrained VGG-19 model to extract image features. The extracted text and image features are then fed to the detector to identify fake news. Other than the extraction of multi-modal features, there is a prominent challenge in fake news detection, i.e., how to identify fake news on unforeseen emerging events. Many existing models merely capture event-specific features that are not shared among different events. Only a few studies attempted to extract generic features across different events (e.g., [13]). In order to better address this challenge, in this paper, we add in the domain classifier to map the multi-modal features of different events to the same feature space. To assess the performance of BDANN, we conduct extensive experiments on two multimedia datasets: Twitter and Weibo. The experimental results show that BDANN outperforms the state-of-the-art baseline models. Moreover, we further discuss the existence of noisy images in the Weibo dataset that may affect the results.

The main contributions of our paper are as follows:

- We propose an end-to-end model, named BERT-based domain adaptation neural network (BDANN) for multi-modal fake news detection. BDANN fuses features from two modalities and removes event-specific dependency.
- We evaluate BDANN on two multimedia datasets: Twitter and Weibo. Experimental results show that BDANN outperforms feature-based methods and state-of-the-art multi-modal models on both datasets.
- We study the phenomenon that images not related to text articles may affect the fake news detection results.

The rest of this paper is organized as follows: In Section II, we review the related work. Section III introduces the technical details of our proposed BDANN model and its different components. In Section IV, we present the datasets used, experiment setup and baselines. In Section V, we report the experimental results with discussion. Finally, Section VI concludes the paper.

II. Related Work

Fake news detection has closely related research topics such as spam detection [15] and rumor identification [16]. Moreover, as individuals may have their own intuitive definition of fake news, prior studies adopted varying definitions, which may conflict or overlap with one another. In this paper, we follow Ruchansky et al.'s definition [17] that fake news are the ones comprising fabricated content.

Most existing studies on fake news detection are feature-based which can be extracted from text articles, social context and images. To extract text features for fake news detection, Castillo et al. [7] use statistics in the text such as the number of URLs, special characters, etc. Gupta et al. [18] use bag-of-words to reveal inter-tweet relations. Feng et al. [19] use context-free grammar phase tree rules to drive text features. However, the linguistic indicators of fake news across topics and media platforms are not yet well understood [17]. As pointed out by Rubin et al. [20], there are many types of fake news, each with different potential textual indicators. Thus, it is difficult to design hand-crafted text features for traditional machine learning based on fake news detection models. To expand beyond hand-crafted features, Ma et al. [10] introduce an RNN-based model to extract temporal representations of microblog events. Following Ma et al.'s work, Chen et al. [11] infuse attention into RNNs to selectively extract temporal representations. Nonetheless, the textual extractor employed by prior studies may not capture the semantics of the text articles posted on microblogs. Thus, in this paper, we employ BERT, a powerful NLP model to extract text features for better performance.

Social context features represent the user engagements of news on social media platforms [21], such as the number of followers, hash-tags (#) and retweets. To better extract the propagation structure of the messages, Wu et al. [22] propose a graph-kernel based hybrid SVM classifier that captures the high-order propagation patterns in addition to semantic features such as topics and sentiments. Jin et al. [12] incorporate

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on October 04,2020 at 22:13:12 UTC from IEEE Xplore. Restrictions apply.

hand-crafted social context features, such as hash-tag topics in tweets and retweets. However, most social context features are noisy, unstructured and labor-consuming. Moreover, they may not provide adequate information for newly emerged events. Therefore, in this paper, we do not incorporate the social context features.

Visual features have been shown as an important indicator in fake news detection [6]. Features extracted from the images attached in the posts have been shown to contribute substantial information [10] [22]. However, the features extracted in the aforementioned prior studies are hand-crafted, which may not represent complex intrinsically embedded contents. Recently, deep neural networks have been shown to be capable of extracting highly complex image and sentence representations [23], [24]. Prior studies [12], [14] leveraging deep learning techniques to extract both text and image features show significantly improved results. In this paper, we adopt VGG-19 [23] to extract image features and concatenate both text and image features for fake news detection.

As the trending events shift fast, features extracted from specific events may lead to poor performance on unforeseen events. To better address this problem, Wang et al. [13] introduce an event classifier to remove event-special features. Thus, inspired by event classifier [13] and domain adaptation [25], in this paper, we incorporate a domain classifier to remove the event-specific dependency from the features extracted by the multi-modal feature extractor.

Fake news detection is related to dynamic system modeling. The classic algorithms include recurrent neural networks, hidden markov model and other algorithms. Recently, Chen et. al [26], [27] proposed a learning in the model space approach for fault diagnosis, and this algorithm transformed the data to the model space and employed the functional analysis to discriminate the models instead of data set. The following work include efficient time series analysis [28] and management of the trade-off between model fitting ability and discrimination ability [29].

III. Dynamics of BDANN

In this section, we present the technical details of each module in BDANN and delineate how the modules are integrated for multi-modal fake news detection.

A. Overall Structure of BDANN

The intuition of BDANN is to learn event-invariant features in a multi-modal way to detect fake news. Its overall architecture is illustrated in Fig 2. BDANN comprises the following three major components:

- multi-modal Feature Extractor: It extracts features from text articles and attached images.
- Domain Classifier: It classifies the posts to different events and removes the event-special features from the extracted features.
- Fake News Detector: It uses latent multi-modal features to determine whether a piece of news is fake or not.

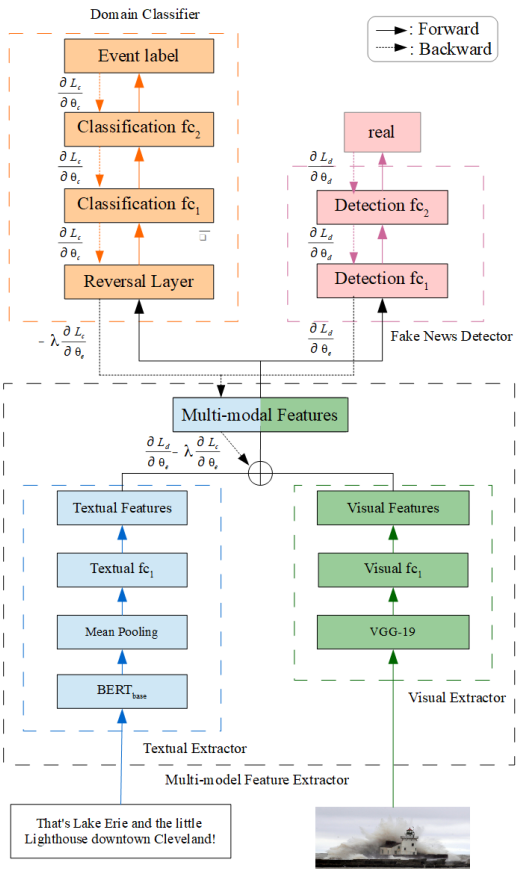

Fig. 2: Network structure of BERT-based Domain Adaptation Neural Network (BDANN).

In the following subsections, we present the technical details of the respective modules and the learning paradigm.

B. Multi-Modal Feature Extractor

Because posts on social media platforms usually comprise information in more than one modal (e.g., text articles and attached images), the multi-modal feature extractor in BDANN comprises textual feature extractor and visual feature extractor.

(i) Textual Feature Extractor: In order to capture the underlying semantic and contextual meanings, we employ Bidirectional Encoder Representations from Transformers (BERT) [24] for textual feature extraction. BERT is essentially a multi-layer bidirectional Transformer encoder based on the original implementation described in [30]. In BDANN, each BERT's layer comprises stacked self-attention and pointwise, fully connected neurons. The outputs of each layer are passed along to the next encoder.

The input of textual feature extractor is a sequential list of words in the text articles, which are first embedded into vectors. We denote the kth dimensional word embedding vector of the ith word in the sentence as $T_i \in R^k$, as such, the input sentence i is represented as follows:

$$T = [T^0, T^1, ..., T^n], \tag{1}$$

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on October 04,2020 at 22:13:12 UTC from IEEE Xplore. Restrictions apply.

where n denotes the number of words in the sentence and T^0 denotes the embedding of [CLS], which is inserted to the top of the sentence as following [24].

In this paper, we use pretrained BERT$_{base}$ that comprises 12 encoder layers and we denote it as R_b. After feeding T into R_b, we get a feature vector of the given sentence as follows:

$$T_f = [T_f^0, T_f^1, ..., T_f^n]. \tag{2}$$

For every feature vector T_f, we use the mean-pooling operation to obtain textual features from all the words based on their importance. The textual features after being applied with the mean-pooling operation are denoted as $R_t \in R^{d_t}$, where d_t denotes the dimensionality of the textual feature obtained from BERT. Finally, we feed R_t to a fully connected layer to ensure that the final output of textual features (denoted as $R_{tf} \in R^p$) has the same dimensionality (denoted as p) as the visual features. As such,

$$R_{tf} = \sigma_t(W_{tf} \cdot R_t), \tag{3}$$

where $W_{tf} \in R^{d_t \cdot p}$ denotes the weight matrix of the fully connected layer in the textual feature extractor and σ_t denotes the Leaky RELU activation function [31] used in the textual feature extractor.

(ii) Visual Feature Extractor: CNN has been successfully applied to various visual understanding problems [32]. In this paper, we employ VGG-19 [23], which has been pretrained on ImageNet, to extract visual features from the images attached in the posts. The dimensionality of imagery features obtained by VGG-19 is denoted as d_v. To map the final output of visual features (denoted as R_{vf}), we append a fully connected layer to VGG-19's last layer. As such,

$$R_{vf} = \sigma_v(W_{vf} \cdot R_{vgg}), \tag{4}$$

where $W_{vf} \in R^{d_v \cdot p}$ denotes the weight matrix of the fully connected layer in the visual feature extractor, R_{vgg} denotes the output of the last layer of VGG-19 and σ_v denotes the Leaky RELU activation function in the visual feature extractor.

The two types of features, i.e., R_{tf} and R_{vf} are then concatenated into a vector of dimensionality $2p$, which is denoted as $R_f \in R^{2p}$. Furthermore, we denote the multi-modal features extractor as $E(P; \theta_e)$, where P denotes the vectorized input post, θ_e denotes the parameter set of the multi-modal extractor, and E denotes the overall mapping function. Thus, we obtain the following equation:

$$R_f = E(P; \theta_e). \tag{5}$$

C. Fake News Detector

In this paper, we use two fully connected layers with the softmax function to devise the fake news detector. We denote the detector as $D(R_f; \theta_d)$, where θ_d denotes the parameter set of the detector and D denotes the mapping function of the detector. The output of the fake news detector $\hat{y}$ for a multi-modal post p^j denotes the probability of the post to be a piece of fake news and thus is defined as follows:

$$\hat{y_j} = D(E(p^j; \theta_e); \theta_d). \tag{6}$$

We use Y to represent the set of labels in which fake news is labeled as 1 (i.e., $y_j = 1$) and real news is labeled as 0 (i.e., $y_j = 0$). Therefore, to compute the classification loss, we employ cross-entropy loss as follows:

$$\mathcal{L}_d(\theta_e, \theta_d) = -\mathbb{E}_{(p,y_j)\in(P,Y)}[y \log(\hat{y_j}) + (1-y)\log(1-\hat{y_j})]. \tag{7}$$

To optimize parameters θ_e and θ_f, we minimize the classification loss, which is defined as follows:

$$(\theta_e^*, \theta_d^*) = \arg\min_{\theta_e, \theta_d} \mathcal{L}_d. \tag{8}$$

D. Domain Classifier

Based on the afore-introduced multi-modal feature extractors, we can obtain the concatenated feature vector for both text and image inputs. However, we still need to devise a robust classifier to better handle the unforeseen circumstance. Inspired by [13] and [25], we devise a domain classifier denoted as $C(R_f; \theta_c)$, where θ_c denotes the parameter set of the domain classifier and C denotes the mapping function of the domain classifier. The domain classifier aims to classify the post into one of the M events based on the input multi-modal features R_f. We use Y_e to denote the event label set and subsequently define the loss of event discriminator by cross-entropy as follows:

$$\mathcal{L}_c(\theta_e, \theta_c) = -\mathbb{E}_{(p,y_e)\in(P,Y_e)}[\sum_{m=1}^{M} y_e \log(C(E(p; \theta_e); \theta_c))]. \tag{9}$$

According to (9), a large loss value means the network learns the event-invariant multi-modal features if the domain classifier is accurate. This approach promotes an adversarial process: The multi-modal extractor tends to extract event-invariant features by maximizing the domain classification loss $\mathcal{L}_c$, while the domain classifier tends to discover the event-special information from multi-modal features by minimizing the domain classification loss $\mathcal{L}_c$. Note that the sequence or timing of the events is not taken into account by BDANN, which may be considered as future work.

E. Model Learning Paradigm

During training, $\mathcal{L}_d(\theta_e, \theta_d)$ needs to be minimized to improve the fake news detection task. In order to obtain invariant event features, the loss of the domain classifier $\mathcal{L}_c(\theta_e, \theta_c)$ needs to be maximized. Simultaneously, the domain classifier tries to discover the event-special information from multi-modal features by minimizing the domain classification loss. Hence, the overall loss is defined as follows:

$$\mathcal{L}(\theta_e, \theta_d, \theta_c) = \mathcal{L}_d - \lambda \mathcal{L}_c, \tag{10}$$

where coefficient $\lambda \in R$ is used to balance the loss function of fake news detection and the domain classification. To achieve the adversarial effect as described in Section III-D, we adopt the gradient reversal layer (GRL) introduced by Ganin et al. [25], which is also applied in [13]. During the forward propagation, GRL acts as an identity transform and during the backpropagation, it multiplies the gradient with λ and

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on October 04,2020 at 22:13:12 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Statistics of two real-world multi-modal datasets

Dataset	Label	Number	all
Twitter	fake	7021	12995
	real	5974	
Weibo	fake	4749	9528
	real	4779	

passes the result to the preceding layer. We place the GRL layer between the multi-modal extractor and domain classifier. Therefore, the optimization process of the model parameters are delineated as follows:

$$\theta_e \longleftarrow \theta_e - \eta(\frac{\partial \mathcal{L}_d}{\partial \theta_e} - \lambda \frac{\partial \mathcal{L}_c}{\partial \theta_e}), \tag{11}$$

$$\theta_d \longleftarrow \theta_d - \eta \frac{\partial \mathcal{L}_d}{\partial \theta_d}, \tag{12}$$

$$\theta_c \longleftarrow \theta_c - \eta \frac{\partial \mathcal{L}_c}{\partial \theta_c}. \tag{13}$$

IV. Experiments Configurations

In this section, we first describe the two social media datasets used in the experiments and we then discuss the state-of-the-art fake news detection baseline models.

A. Dataset Descriptions

(i) Twitter dataset: The Twitter dataset is released for the "verifying multimedia use" task by MediaEval Benchmarking Initiative for Multimedia Evaluation [33]. The dataset comprises a list of tweets that each has textual content with attached images. The dataset is categorized into the development set and the test set. The development set contains about 6,000 rumor and 5,000 non-rumor tweets from 11 rumor-related events. The test set contains about 2,000 tweets of either type. According to [33], there is no overlapping event between the two sets. Because we only consider textual and visual information, we remove the tweets without any text or image in this dataset. Furthermore, to keep the data coherent, we translate non-English content in tweets into English by applying Google Translate. Following the benchmarking models' approach (see Table II), we use the development set for training and the test set for testing.

(ii) Weibo dataset: The Weibo dataset, collected by Jin et al. [12], has been used in many studies on multi-modal fake news detection (e.g., [14], [13]). In Weibo, akin to Tweet in terms of the platform but the content is in Chinese, there exists an official fake news debunking system. Weibo users are encouraged to report suspicious posts and a committee composed of reputable users then would verify them as false or real after examining the reported cases. In addition, according to prior studies [10], [22], this system can serve as an authoritative source to collect fake news. Thus, the fake news in the dataset is the news verified as fake by the debunking system in the period of May 2012 to January 2016. The real news is collected from authoritative news sources in China, such as Xinhua News Agency Weibo. We process this dataset in the same manner as adopted in [13]. Specially, we first remove duplicate images with a near-duplicated image detection algorithm [34] and remove odd-sized images to ensure the dataset's integrity. We then apply a single-pass incremental clustering method [8] to get the news event labels. Finally, we split the dataset into training, validation and testing sets in the ratio of 7:1:2 and ensure the three sets do not overlap with one another. The statistics of the two afore-introduced datasets are listed in Table I.

B. Experiment Setting

For text in Twitter and Weibo datasets, we follow the standard text preprocessing procedure as adopted in [35]. Due to the different languages used in the two datasets, the tokenizer used in preprocessing is the BERT$_{base}$ tokenizer pretrained on the respective languages. In the textural extractor, the dimensionality d_t of textural features obtained from BERT$_{base}$ is 768. For the visual extractor, we first resize images to 224x224x3 and then feed them to VGG-19 pretrained on ImageNet [23]. The dimensionality d_v of imagery features obtained from VGG-19 is 4,096. The hidden size p of the fully connected layer in the textual and visual extractor is set to 32. To avoid overfitting, the parameters of BERT$_{base}$ and VGG-19 are all frozen. The size of the two fully connected layers in the domain classifier is set to 64 and 32, respectively. Every fully connected layer in the model has a Leaky RELU activation function and a dropout probability of 0.5. The model is trained on a batch size of 128 and for 100 epochs with a learning rate of 10^{-3}. To optimize parameters in BDANN, we use the Adam optimizer [36].

C. Baseline Models

To benchmark the performance of BDANN on the fake news detection task, we compare it against the following three types of models: single modality, multi-modal, and the variants of BDANN.

(i) Single Modality Models: Although BDANN detects fake news based on both the visual and textual information, we still compare its performance against text-only and image only models.

- **Text-Only**: The input of this model comprises only text articles, which are fed to a pretrained BERT$_{base}$ followed by a fully connected layer to obtain textual features R_t. Then we feed R_t to a 32-dimension fully connected layer with the softmax function to detect fake news.
- **Image-Only**: The input of this model comprises only images, which are fed to a pretrained VGG-19 to obtain visual features R_v. Then we also feed R_v to a 32-dimension fully connected layer to make predictions.

(ii) Multi-modal Models: Multi-modal models rely on both textual and visual features to detect fake news.

- **VQA** [37]: Visual question answering (VQA) aims to answer questions about the given images. In order to adopt VQA to detect fake news, we modify the element-wise multiplication between text and image to feature concatenation and the multi-class classifier to a binary

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on October 04,2020 at 22:13:12 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Performance of BDANN against baselines on two multi-modal datasets

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	Text-Only	0.706	0.648	0.540	0.589	0.715	0.636	0.673
	Image-Only	0.596	0.695	0.518	0.593	0.524	0.700	0.599
	VQA	0.631	0.765	0.509	0.611	0.550	0.794	0.65
	NeuralTalk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN-	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	att-RNN	0.682	0.780	0.615	0.689	0.603	0.770	0.676
	EANN-	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	EANN	0.719	0.642	0.474	0.545	0.771	0.870	0.817
	MVAE	0.745	0.801	**0.719**	**0.758**	0.689	0.777	0.730
	BDANN-v	0.763	0.747	0.421	0.538	0.765	**0.930**	0.840
	BDANN-d	0.821	0.790	0.610	0.690	**0.830**	0.920	0.870
	BDANN	**0.830**	**0.810**	0.630	0.710	**0.830**	**0.930**	**0.880**
Weibo	Text-Only	0.804	0.800	0.860	0.830	0.840	0.760	0.800
	Image-Only	0.633	0.630	0.500	0.550	0.630	0.750	0.690
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	NeuralTalk	0.726	0.794	0.613	0.692	0.684	0.840	0.754
	att-RNN-	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	att-RNN	0.788	0.862	0.686	0.764	0.738	0.89	0.807
	EANN-	0.794	0.790	0.820	0.800	0.800	0.770	0.780
	EANN	0.816	0.820	0.820	0.820	0.810	0.810	0.810
	MVAE	0.824	0.854	0.769	0.809	0.802	**0.875**	0.837
	BDANN-v	**0.851**	**0.869**	0.836	**0.852**	0.832	0.866	**0.849**
	BDANN-d	0.814	0.800	0.860	0.830	0.840	0.760	0.800
	BDANN	0.842	0.830	**0.870**	0.850	**0.850**	0.820	0.830

classifier. We also change the embedded LSTM to one layer with the size of 32 as adopted in [12], [13], [14].

- **NeuralTalk** [38]: NeuralTalk aims to generate natural sentences to describe an image. Following the main network structure of NeuralTalk, we can get a joint representation of images and text by averaging the output of RNN at each time step. Then we feed the joint representation to a 32-dimension fully connected layer to detect fake news. This approach is consistent with [12], [13], [14].
- **att-RNN** [12]: att-RNN uses LSTM to extract both text and social context features and obtain the joint representation, which is then combined with visual features extracted from a pretrained deep CNN with attention. We denote the original model as att-RNN and its variant without the social context information as att-RNN-.
- **EANN** [13]: EANN comprises three main components: the multi-modal feature extractor, the fake news detector and the event discriminator. In the multi-modal extractor, Text-CNN is employed to extract textual features and pretrained VGG-19 is employed to extract visual features. Then the multi-modal features are fed to the fake news detector to predict whether a post is fake or not. The event discriminator consists of two fully connected layers aiming to remove the event-specific dependency. We denote the original model as EANN and its variant without event discriminator as EANN-.
- **MVAE** [14]: MVAE aims to learn a shared representation between textual and visual modalities to detect fake news. A variational autoencoder is leveraged to get the shared representation by the input data reconstruction and a binary classifier is employed to detect fake news.

(iii) Variants of BDANN: To analyze the impact of different modules in BDANN, we also assess the performance of its variants:

- **BDANN-v**: This variant denotes BDANN without the visual feature extractor. The fake news detector only gets textural features as inputs.
- **BDANN-d**: This variant denotes BDANN without the domain classifier. Thus, this variant does not remove the event-specific dependency.

V. Experimental Results and Analysis

To assess the performance of BDANN, we conduct extensive experiments on two multimedia datasets: Twitter and Weibo. Table II shows the results of the baselines, BDANN and the variants of BDANN on these two datasets. The metrics we report include accuracy, precision, recall and F1 score for both fake news and real news. It is clearly shown in Table II that the overall performance of BDANN (including its variants) is much better than the baselines.

In terms of performance comparison on the Twitter dataset, the image-only model performs worse than the text-only model, which suggests that the textual information is still much more important than the visual information in fake news detection. Nonetheless, though BERT demonstrates competent

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on October 04,2020 at 22:13:12 UTC from IEEE Xplore. Restrictions apply.

TABLE III: Performance comparison before and after Weibo dataset being filtered

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	EANN	0.816	0.820	0.820	0.820	0.810	0.810	0.810
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	BDANN-v	0.851	**0.869**	0.836	0.852	0.832	**0.866**	0.849
	BDANN	0.842	0.830	0.870	0.850	0.850	0.820	0.830
Weibo (Filtered)	EANN	0.822	0.820	0.850	0.840	0.820	0.790	0.800
	MVAE	0.836	0.802	0.902	0.851	0.884	0.762	0.818
	BDANN-v	0.846	0.849	0.862	0.855	0.844	0.829	0.837
	BDANN	**0.865**	0.850	**0.920**	**0.880**	**0.890**	0.810	**0.850**

capability, its performance is still not comparable with the state-of-the-art multi-modal methods, which complement the textual features with visual features. Among the benchmarking multi-modal models, MVAE performs the best, probably due to the employment of multi-modal variational autoencoder. Such that a unified representation of textual and visual modalities is learned to detect fake news. Nonetheless, the performance difference between EANN and EANN- indicates the importance of the incorporated event discriminator, which can better handle emerging unforeseen events. This similar pattern is observed in BDANN as well that the accuracy increases (from 0.821 obtained by BDANN-d to 0.83) if the domain classifier is incorporated. In addition, comparing the accuracy of BDANN with BDANN-v, the performance improves significantly ((0.83-0.763)/0.763=8.8%). This finding suggests that in tweets, the images play an important role in fake news detection. In summary, it is encouraging to see BDANN obtains the best result in most measures (except Recall and F1 score on fake news).

In terms of performance comparison on the Weibo datasets, we can observe similar patterns as in the Twitter dataset that MVAE performs the best among all the baselines, BDANN performs better than MVAE, BDANN performs better with the incorporation of the domain classifier, etc. However, there is an interesting phenomenon that the accuracy of BDANN is lower than that of BDANN-v, which means the complement of the extracted visual features decreases the model performance. Such observation is not found in the Twitter dataset. Thus, it is reasonable for us to suspect that there exist noisy images in the Weibo dataset, which do not contribute relevant information for fake news detection. By noisy, we mean the image is unrelated to the corresponding text article on a certain extent. Therefore, noisy images may lead to inferior results when multi-modal features are used, i.e., they undermine the possibly correct decisions made only based on the textual features. Specifically, in a post, the text article shows that the post is a piece of fake news while the image which has no relationship with the text article may mislead BDANN to classify it as a piece of real news.

To further investigate the existence of suspicious noisy images, we adopt the cross-validation approach to identify them. Specifically, we equally split the original Weibo dataset into 5 folds and perform 5-fold cross-validation using both BDANN and BDANN-v. When getting the testing results in each fold, we archive the indexes of the posts which are detected as fake news by BDANN-v but as real news by BDANN. After cross-validation, we remove all the archived posts[1] from the original dataset and retrain BDANN and BDANN-v. The results are shown in Table III. Compared with models trained on the original Weibo dataset, the accuracy of BDANN-v trained on the filtered Weibo dataset decreases. However, the accuracy of BDANN trained on the filtered Weibo dataset increases ((0.865-0.842)/0.842=2.7%). In addition, the accuracy of BDANN trained on the filtered Weibo datasets is higher than the accuracy of BDANN-v trained on the original Weibo dataset (0.865 vs. 0.851). This finding shows that after filtering the posts with potentially noisy images, the model performs as expected that it achieves better results when using multi-modal features (0.865) than using textual features alone (0.846). This finding may also indicate the existence of such noisy images in the Weibo dataset, which leads to inferior performance. Moreover, to further investigate this phenomenon, we also retrain prior state-of-the-art baselines, EANN [13] and MVAE [14], on the filtered Weibo dataset. As shown in Table III, the accuracy of EANN and MVAE trained on the filtered Weibo dataset both increase ((0.822-0.816)/0.816=0.7%, (0.836-0.824)/0.824=1.5%), which further indicate the existence of noisy images in the Weibo dataset.

Qianxi bus exploded! My friend witnessed this terrible explosion! How many innocent people have died like this!

Capital migration in Xinyang is a foregone conclusion.

Fig. 3: Examples of removed posts from the Weibo dataset.

Fig. 3 shows two pieces of fake news which are detected as fake news by BDANN-v but as real news by BDANN. Looking at the two examples, the text articles may be adequate to identify that the posts are fake and the attached images are irrelevant to the text articles.

VI. Conclusion

In this paper, we proposed BDANN, a BERT-based domain adaptation neural network to detect fake news in a multi-modal manner. BDANN comprises three main modules: a

[1]See the comprehensive list of indexes of the filtered Weibo posts at: https://github.com/xiaolan98/RemovedPostsFromWeibo

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on October 04,2020 at 22:13:12 UTC from IEEE Xplore. Restrictions apply.

multi-modal feature extractor, a domain classifier and a fake news detector. The multi-modal features are first extracted by the feature extractor, concatenated and then fed to the domain classifier to remove the event-specific dependency. After that, the detector is used to distinguish fake news. To assess the performance of BDANN, we conduct extensive experiments on two multimedia datasets: Twitter and Weibo. The experimental results show that BDANN outperforms the state-of-the-art models. Moreover, based on the interesting results obtained in the Weibo dataset, we further discuss the existence of noisy images in that dataset, which leads to inferior results.

In the future, besides applying BDANN to other similar fake news detection datasets or similar detection tasks, we also plan to employ the probabilistic model [39] and propose a deep model to distinguish whether the attached image is relevant to the corresponding text article prior to performing fake news detection.

REFERENCES

[1] J. Clement, "Twitter: Number of monthly active users 2010-2019," https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/, Aug. 2019.

[2] Statista Research Department, "Number of Sina Weibo users in China 2017-2021," https://www.statista.com/statistics/941456/china-number-of-sina-weibo-users/, Aug. 2019.

[3] G. Pennycook, T. D. Cannon, and D. G. Rand, "Prior exposure increases perceived accuracy of fake news," *Journal of Experimental Psychology: General*, vol. 147, no. 12, pp. 1856–1880, 2018.

[4] J. Chang, J. Lefferman, C. Pedersen, and G. Martz, "When fake news stories make real news headlines," https://abcnews.go.com/Technology/fake-news-stories-make-real-news-headlines/story?id=43845383, Nov. 2016.

[5] CBCnews, "Probe reveals stunning stats about fake election headlines on Facebook," https://www.cbsnews.com/news/facebook-fake-election-news-more-popular-than-real-news-buzzfeed-investigation/, Nov. 2016.

[6] 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, 2016.

[7] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proceedings of International Conference on World Wide Web*, 2011, pp. 675–684.

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

[9] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2016, pp. 2972–2978.

[10] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with Recurrent Neural Networks," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2016, pp. 3818–3824.

[11] T. Chen, L. Wu, X. Li, J. Zhang, H. Yin, and Y. Wang, "Call attention to rumors: Deep attention based Recurrent Neural Networks for early rumor detection," *arXiv preprint arXiv:1704.05973*, 2017.

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

[13] 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 *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 849–857.

[14] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal Variational Autoencoder for fake news detection," in *Proceedings of International World Wide Web Conference*, 2019, pp. 2915–2921.

[15] X. Hu, J. Tang, and H. Liu, "Online social spammer detection," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2014, pp. 59–65.

[16] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 3, pp. 99–108, 2015.

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

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

[19] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of Annual Meeting of the Association for Computational Linguistics: Short Papers*, vol. 2, 2012, pp. 171–175.

[20] V. Rubin, Y. Chen, and N. Conroy, "Deception detection for news: Three types of fakes," in *Proceedings of ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, 2015, pp. 1–4.

[21] 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.

[22] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on Sina Weibo by propagation structures," in *Proceedings of IEEE International Conference on Data Engineering*, 2015, pp. 651–662.

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

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[25] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[26] H. Chen, P. Tiňo, A. Rodan, and X. Yao, "Learning in the model space for cognitive fault diagnosis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 124–136, 2013.

[27] H. Chen, P. Tiňo, and X. Yao, "Cognitive fault diagnosis in tennessee eastman process using learning in the model space," *Computers & Chemical Engineering*, vol. 67, pp. 33–42, 2014.

[28] H. Chen, F. Tang, P. Tino, and X. Yao, "Model-based kernel for efficient time series analysis," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 392–400.

[29] H. Chen, F. Tang, P. Tino, A. G. Cohn, and X. Yao, "Model metric co-learning for time series classification," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015, pp. 3387–3394.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[31] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[32] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[33] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris, "Verifying multimedia use at MediaEval 2015," in *Proceedings of MediaEval Benchmark 2015*, vol. 3, no. 3, 2015, p. 7.

[34] M. Slaney and M. Casey, "Locality-sensitive hashing for finding nearest neighbors [Lecture Notes]," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 128–131, 2008.

[35] K. Fortney, "Pre-processing in natural language machine learning," https://towardsdatascience.com/pre-processing-in-natural-language-machine-learning-898a84b8bd47, Nov. 2017.

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

[37] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, V. Tech, and M. Research, "VQA: Visual Question Answering," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[38] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *arXiv preprint arXiv:1411.4555*, 2014.

[39] H. Chen, P. Tino, and X. Yao, "Probabilistic classification vector machines," *IEEE Transactions on Neural Networks*, vol. 20, no. 6, pp. 901–914, 2009.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on October 04,2020 at 22:13:12 UTC from IEEE Xplore. Restrictions apply.