Theme Article: Big Cross-Modal Social Media Data Analytics with Deep Intelligence

Detecting Disaster-Related Tweets Via Multimodal Adversarial Neural Network

Wang Gao, Xun Zhu, and Yuwei Wang
Jianghan University

Lin Li
Wuhan University of Technology

***Abstract*—Recently, during natural disasters, the use of social media as a source of actionable information has increased significantly. Finding disaster-related tweets and analyzing their textual content and images can help government agencies and rescue organizations make better decisions. The main challenge of disaster-related message detection on social networking sites is how to identify posts related to emerging disaster events. Since most existing methods extract disaster-specific features that cannot be shared between different disaster events, they are difficult to deal with this challenge. In this article, we propose a novel multimodal adversarial neural network (MANN) to handle the above challenge. MANN consists of three modules: a feature extraction module, a tweet detection module, and a disaster discrimination module. The feature extraction module uses BERT and VGG-19 to learn textual and visual feature representations from posts. Based on multimodal features, the tweet detection module is responsible for identifying posts related to disasters. MANN exploits the disaster discrimination module and adversarial training to captures disaster-invariant features for unseen disaster events. Experimental results on real-world datasets show the proposed model outperforms baseline methods in terms of precision, recall, and F1-measure.**

Digital Object Identifier 10.1109/MMUL.2020.3012675
Date of publication 31 July 2020; date of current version 20 November 2020.

1070-986X © 2020 IEEE Published by the IEEE Computer Society

Natural disasters such as wildfires, earthquakes, and floods can cause serious ecological damage, and require extensive efforts by society to deal with them. During natural and man-made disasters, humanitarian organizations need to provide timely assistance to all affected people. However, due to the limited location information of victims, this task is very difficult for government agencies and rescue organizations, and may cause significant economic and life loss. From 2005 to 2014, an average of 380 disasters occurred annually in the world, resulting in economic losses of US $159.7 billion and 199.2 million victims[1]. Nowadays, in an emergency, a large number of users tend to post various information on social media platforms such as Facebook and Twitter. Victims and witnesses often post their own status, report injured persons, inform infrastructure damage, and seek help through these platforms with textual messages and images. These data generated by social media are ubiquitous and up-to-date, which can be used to help government agencies and rescue organizations obtain actionable information to save lives and rebuild infrastructure[2]. During the disaster of Hurricane Harvey, a woman was rescued when she sought help on Twitter because the emergency contact number "911" could not be reached at the time.*

Textual messages and images are the two main forms of information on social media platforms[3]. Most previous systems and research have focused on using textual content to detect disaster information from social media. Nevertheless, in addition to textual messages, recent research has shown that images posted on social networking sites during disasters can help rescue organizations in many ways[2,4,5]. However, few studies have been reported using both textual messages and images to detect disaster-related tweets from large amounts of social media content.

Furthermore, the above methods cannot cope with the unique challenge of disaster-related message detection on social media, i.e., identifying disaster-related messages related to newly emerging disaster events. Because of the lack of prior knowledge, it is difficult to obtain verified tweets about such events in time, resulting in the unsatisfactory performance of existing methods. In fact, existing methods tend to extract disaster-specific features that cannot be shared between different disaster events. Although these disaster-specific features help to categorize posts about disasters that have occurred, they would hurt the detection of tweets related to new disaster events.

In this article, we propose a novel end-to-end model to deal with the above challenges. The main idea comes from the answers of the following two questions: (1) How to create a multimodal system that efficiently leverages the text and images of tweets to identify disaster-related posts. (2) How to capture transferable features to benefit the detection of tweets related to newly emerging disaster events.

Specifically, we propose a multimodal adversarial neural network (MANN) based on multimodal features and adversarial training for disaster-related tweets detection. A generative adversarial network is a generative model with two submodels: a generator and a discriminator, which are adversarially trained together in a zero-sum game[6,7]. Inspired by this, in the training stage, we add a disaster discrimination module to predict disaster labels, and employ the corresponding loss to measure the similarity of features between different disaster events. The smaller the loss, the less the similarities, indicating that the extracted features contain more disaster-specific information. Our model consists of three modules: a feature extraction module, a tweet detection module, and a disaster discrimination module. The feature extraction module and the tweet detection module jointly perform the main task of identifying disaster-related tweets. MANN utilizes the bidirectional encoder representations from transformers (BERT) model[8] to learn textual features and the convolutional neural network (CNN)-based VGG-19 network to capture visual features. We conducted extensive experiments on real-world datasets of various disasters such as hurricanes, wildfires, earthquakes, and floods. Experimental results show that the proposed model significantly outperforms baseline approaches. The main contributions of this article are summarized as follows.

*http://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/

- We propose a novel MANN, which is able to identify disaster-related tweets based on multimodal features and learn disaster-invariant features by adversarial training. To the best of our knowledge, this is the first work to integrate adversarial training into a neural network for disaster-related tweet detection.
- MANN leverages a disaster discrimination module to estimate the similarity between different disasters, and further captures transferable features that generalize well for unseen disaster events.
- The performance of MANN is evaluated on real-world datasets against state-of-the-art methods. Experimental results demonstrate that our model can effectively identify disaster-related tweets, and outperforms other baselines on several evaluation metrics.

RELATED WORK

In this section, we briefly summarize the work related to MANN from the following two perspectives: single modality tweet classification and multimodal tweet classification.

Single Modality Tweet Classification

The semantic features extracted from the textual content of tweets are called textual features, which have been discussed in many text classification studies. Caragea and Tapia proposed a new model that utilizes CNN to extract informative information from social media streams during emergencies[9]. Similarly, based on a deep CNN, Aipe *et al.* proposed a novel crisis-related tweet multilabel classification model[10]. Based on the Twitter dataset during Hurricane Harvey, Sandy, and Irma, Yu *et al.* studied the capability of the CNN model in a cross-event topic classification task[11].

There are still many studies that use visual features as important features for tweet classification. Alam *et al.* proposed an end-to-end model that categorizes imagery content posted on social networking sites to help rescue organizations[12]. Chaudhuri and Bose collected images from the smart city environment in the earthquake-stricken area, and used a CNN network to detect human body parts in the ruins[13]. Their approach can help practitioners (crisis managers and rescue organizations) expand the knowledge base of big data image analysis. Unlike the above models, we incorporate two types of features when identifying tweets related to disasters.

Multimodal Tweet Classification

Recently, many researchers have used deep neural networks to capture feature representations from multimodal or multiviews patterns, and made progress in various tasks[2,14,15]. Jin *et al.* proposed a new recurrent neural network with an attention mechanism (att-RNN), which integrates textual features and visual features to classify posts on social networks[16]. Kumar *et al.* proposed a multimodal method that uses both text and images to detect disaster-related tweets in the Twitter streams[2]. Their model utilizes CNN and long short term memory to extract visual and textual features, respectively. However, the multimodal features extracted by the above method depend on specific disaster events, and cannot be generalized well to identify Twitter data of new disaster events. Unlike the above methods, MANN can not only automatically extract multimodal features of tweets, but also learn disaster-invariant features using adversarial training.

MULTIMODAL ADVERSARIAL NEURAL NETWORK

In this section, we first introduce the three modules of MANN, and then discuss how to combine these three modules to capture transferable features. Figure 1 shows the overall architecture of our model.

Feature Extraction Module

Visual Features Recent research has established deep CNN as a state-of-the-art method in image classification tasks, such as Xception, ResNet50, VGG-16, and VGG-19, all of which use pretrained deep CNN to extract visual features[17]. As a result, we choose the pretrained VGG-19 to extract the visual features of tweet images, as suggested by several recent works[2]. VGG-19 is a CNN-based model that consists of 19 layers (1 softmax layer, 5 maxpool layers, 3 fully

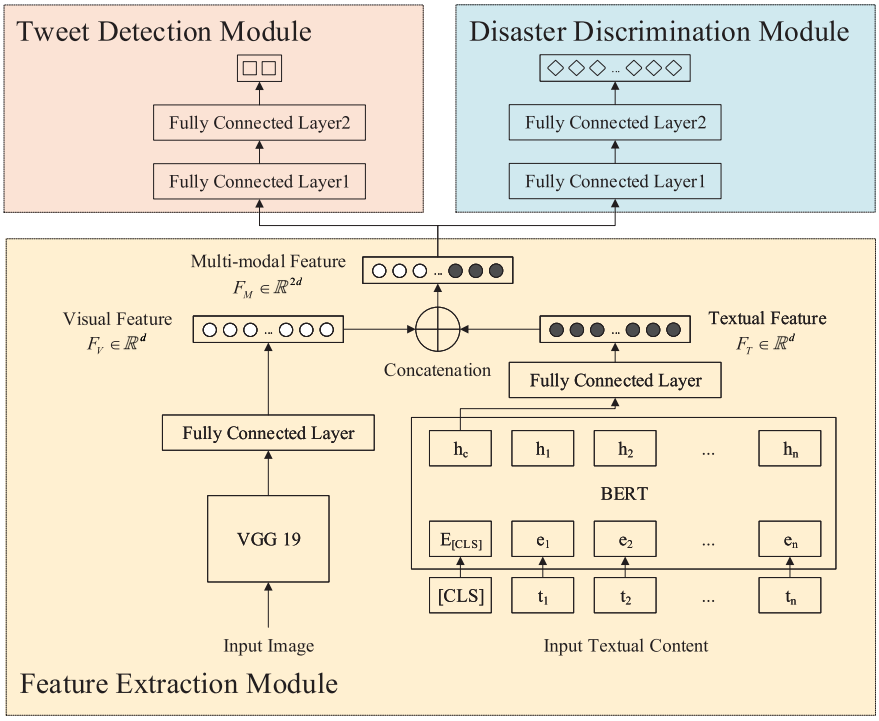

Figure 1. Overall architecture of MANN.

connected layers, and 16 convolution layers). To ensure that the visual feature representation has the same dimension as the textual feature representation, we add a fully connected network above the last layer of the VGG-19 model. To prevent overfitting, during the training process of the feature extraction module, the parameters of the pretrained VGG-19 model remain unchanged. Suppose $F_V \in \mathbb{R}^d$ denotes the representation of visual features, and d is the dimension of the representation. The fully connected layer of the last layer in the visual feature extraction network can be defined as

$$F_V = \sigma(W_V \cdot F_{VGG}) \tag{1}$$

where σ represents the ReLU activation function, W_V denotes the weight matrix of the fully connected network, and F_{VGG} is the output of the pretrained VGG-19 network.

Textual Features The input of the textual feature extractor is a sequential list of words in the tweet. In our model, BERT is employed to learn textual features from posts. In recent years, a novel language representation model BERT has become the most advanced model for various NLP tasks by pretraining on large-scale corpora[8]. Since the maximum length is predefined during pretraining, BERT is suitable for short texts such as tweets. However, there has not been much research on exploring BERT to classify disaster-related tweets and normal tweets. BERT is capable of learning long-range dependencies of word sequences, which is essential in our task. Therefore, the proposed model utilizes BERT to extract textual features from posts.

Let $T = \{t_1, t_2, \ldots, t_j, \ldots, t_n\}$ be the input sentence, where t_j is the jth word of the sentence and n is the length of the sentence. The input representation is a concatenation of token embeddings, segment embeddings, and positional embeddings. A token embedding layer transforms each word into a vector representation, and segment embeddings are used to distinguish different sentences. The BERT model consists of transformer networks, which cannot encode the sequential information of input sentences. As a result, BERT utilizes positional embeddings to capture the sequential nature of input sentences. The three representations are summed element-wise to create a single representation $E = \{e_1, e_2, \ldots, e_j, \ldots, e_n\}$, which is then used as input to the BERT encoding layer. After that, multihead attention and self-attention can

be employed to encode E into a series of hidden states $H = \{h_1, h_2, \ldots, h_j, \ldots, h_n\}$. Self-attention has become an important part of sequence-to-sequence modeling in various NLP tasks such as machine translation and topic modeling[18]. The self-attention mechanism can be seen as mapping a series of key-value pairs and queries to the output[8]. The output of queries can be calculated using the weighted sum of values, and the dot product of keys. Queries can be used to calculate the weight of values. Keys, values and queries are packed together into matrices *K*, *V*, and *Q*, and the matrix of outputs can be computed as follows:

$$\text{Attention}(K, V, Q) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where d_k denotes the dimension of keys and queries.

The multihead attention mechanism obtains different representations of queries, keys, and values, allowing BERT to pay attention to information from different representations. Therefore, the mechanism enables the model to learn multiple relationships at different positions in a sentence. All heads are concatenated, and the multihead attention can be expressed as:

$$\begin{aligned}\text{MultiHead}(K, V, Q) &= \text{concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{head}_j &= \text{Attention}(KW_j^K, VW_j^V, QW_j^Q)\end{aligned} \tag{3}$$

where $\text{concat}(\cdot)$ denotes a concatenation function, and $W_j^O \in \mathbb{R}^{hd_{\text{model}} \times d_v}, W_j^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_j^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_j^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ represent parameter matrices. Following the work by Devlin *et al.*[8], we set $h = 12, d_k = d_v = d_{\text{model}}/h = 64$. A special embedding [CLS] is inserted into the input sequence as the first token, and its hidden state h_c is used as the out of BERT. The last layer of the text feature extractor is a fully connected layer, which can be represented as

$$F_T = \sigma(W_T \cdot h_c) \tag{4}$$

where $F_T \in \mathbb{R}^d$ denotes the representation of textual features, W_T is a weight matrix.

The representation of multimodal features $F_M \in \mathbb{R}^{2d}$ is formed by concatenating visual feature representation F_V and textual feature representation F_T

$$F_M = \text{concat}(F_V, F_T). \tag{5}$$

F_M is the output of the feature extraction module that can be denoted as $f_F(X; \theta_F)$. X represents the input multimedia post, and θ_F is the parameters of the module.

Tweet Detection Module

The tweet detection module is a neural network which consists of two fully connected layers with softmax, and is used to predict whether tweets are related to disasters. Since the tweet detection module is based on the feature extractor, the multimodal feature representation F_M is the input of the module. This module can be denoted as $f_D(F_M; \theta_D)$, where θ_D represents all parameters that need to be learned. Let x_j represent the jth multimedia tweet. The output of the tweet detection module for x_j is the possibility that the tweet is not related to any disaster

$$p(x_j) = f_D(f_F(x_j; \theta_F); \theta_D). \tag{6}$$

The purpose of the tweet detection module is to identify whether a particular tweet is related to a disaster. We utilize cross entropy loss as a loss function, and this process is defined as follows:

$$\begin{aligned}\mathcal{L}_D(\theta_F, \theta_D) = -\,&\mathbb{E}_{(x,y)\sim(X,Y_d)}[(1-y)(\log(1-p(x))) \\ &+ y\log(p(x))]\end{aligned} \tag{7}$$

where Y_d denotes the labels of tweets. The loss function of the tweet detection module $\mathcal{L}_D(\theta_F, \theta_D)$ can be minimized by finding optimal parameters θ_F^* and θ_D^*

$$(\theta_F^*, \theta_D^*) = \arg\min_{\theta_F, \theta_D} \mathcal{L}_D(\theta_F, \theta_D). \tag{8}$$

As discussed in this article, the main challenge of disaster-related tweets detection comes from disaster events not covered by the training corpus. To identify tweets related to emerging disaster events, our model needs to be able to capture disaster-invariant features. The minimization of loss function $\mathcal{L}(\theta_F, \theta_D)$ only helps to identify tweets related to disaster events that are already included in the training corpus. Since this process

learns disaster-specific features, patterns, or knowledge (e.g., the name of the disaster), it is difficult to generalize well. Therefore, our model needs to capture feature representations that can be generalized, and these representations can learn more common information between different disaster events. To achieve this goal, the representation must be disaster-invariant and contain as little disaster-specific information as possible. Specifically, we introduce adversarial training to eliminate the uniqueness of disaster events. MANN leverages a disaster discrimination module to enhance the similarity of feature representations between different disaster events, which in turn learns disaster-invariant features.

Disaster Discrimination Module

The disaster discrimination module deploys two fully connected layer with activation functions. Its purpose is to classify tweets as one of N disaster events based on multimodal features. The module can be denoted as $f_M(F_M;\theta_M)$, where θ_M represents all parameters. The loss of the disaster discrimination module can be defined as

$$\mathcal{L}_M(\theta_F,\theta_M) = -\mathbb{E}_{(x,y)\sim(X,Y_m)}\left[\sum_{j=1}^{N} y_j \log(f_M(f_F(x;\theta_F);\theta_M))\right] \tag{9}$$

where Y_m denotes the labels of disasters. Optimal parameter θ_M^* of the loss is defined as

$$\theta_M^* = \arg\min_{\theta_M} \mathcal{L}_M(\theta_F,\theta_M). \tag{10}$$

$\mathcal{L}_M(\theta_F,\theta_M)$ is employed to measure the similarity of different disasters. The greater the loss, the more similar the representation of different disasters, and the distributions of representations remove the uniqueness of each disaster. Therefore, to learn disaster-invariant features, MANN maximizes the loss $\mathcal{L}_M(\theta_F,\theta_M)$ by looking for optimal parameter θ_F.

In this article, adversarial training is a minimax game between the feature extraction module and the disaster discrimination module. On one hand, the feature extraction module $f_F(\cdot;\theta_F)$ cooperates with the tweet detection module $f_D(\cdot;\theta_D)$ to minimize the loss $\mathcal{L}_D(\theta_F,\theta_D)$, and improve the performance of detecting disaster-related tweets. On the other hand, the feature extraction module $f_F(\cdot;\theta_F)$ tries to fool the disaster discrimination module to maximize the loss $\mathcal{L}_M(\theta_F,\theta_M)$, and learn disaster-invariant feature representations. Furthermore, the disaster discrimination module attempts to identify each disaster based on multimodal features by minimizing the loss $\mathcal{L}_M(\theta_F,\theta_M)$. The final adversarial loss function of three modules can be defined as

$$\mathcal{L}_{adv}(\theta_F,\theta_D,\theta_M) = \mathcal{L}_D(\theta_F,\theta_D) - \upsilon\mathcal{L}_M(\theta_F,\theta_M) \tag{11}$$

where hyperparameters υ controls the tradeoff between the loss of the tweet detection module and the disaster discrimination module. A larger υ indicates disaster labels are relatively more important, whereas a smaller υ implies the tweet detection module plays a more prominent role in the detection task. It is worth noting that by setting υ to 0, MANN is able to suppress adversarial training. Being a preliminary study on exploiting adversarial training for disaster-related tweets detection, we simply set the values of υ as 1. For the minimax game, the optimal parameters $(\theta_F^*,\theta_D^*,\theta_M^*)$ of MANN can be calculated as follows:

$$\begin{aligned}\theta_M^* &= \arg\max_{\theta_M} \mathcal{L}_{adv}(\theta_F^*,\theta_D^*,\theta_M)\\ (\theta_F^*,\theta_D^*) &= \arg\min_{\theta_F,\theta_D} \mathcal{L}_{adv}(\theta_F,\theta_D,\theta_M^*).\end{aligned} \tag{12}$$

We utilize stochastic gradient descent to find the optimal parameters of MANN.

EXPERIMENTS

To validate the effectiveness of the proposed model, we conduct extensive experiments against multiple baselines. In this section, we will introduce our experimental setup, and analyze the performance of MANN.

Dataset

We use a labeled dataset CrisisMMD[†] as positive examples, which contains tweets with both textual messages and associated images. CrisisMMD is composed of different disaster event datasets: Sri Lanka flood, California wildfire,

[†]https://dataverse.mpi-sws.org/dataverse/icwsm18/

Table 1. Number of data instances for each natural disaster.

Disaster name	Text messages	Images
Sri Lanka flood	832	1025
California wildfire	1486	1589
Mexico earthquake	1241	1239
Hurricane Harvey	4000	4562

Mexico earthquake, Hurricane Irma, and so forth. A detailed description of the dataset can be seen by Alam *et al.*[4]. If a post contains more than one image, then we use the same textual message for all corresponding images. Table 1 lists the number of instances of each natural disaster in the dataset used in this article.

As for negative examples, we collect 8415 English tweets containing pictures, which are labeled by human annotators as irrelevant to disaster events. We perform the following standard text preprocessing steps: (1) convert letters to lowercase; (2) remove stop words,[‡] punctuation, URLs, and special symbols such as "@," "#". Baseline models perform the above preprocessing steps, while the proposed model uses raw text. This is because BERT used unpreprocessed corpora with noise during the pretraining process, and it can learn that stopwords are useless. Furthermore, BERT has a built-in tokenizer built with a WordPiece model. The tokenizer generates a fixed-size vocabulary containing individual words, subwords, and characters that are most suitable for pretraining data.

All images are resized to $(224 \times 224 \times 3)$, which is the input size that VGG-19 needs. In the dataset, 20% of samples are used to evaluate the performance of models, and the remaining 80% of them are used for training.

Baseline Methods

We compare the performance of our model with two single modality models, two multimodal models and a variant of MANN, which are introduced as follows.

- *Text-only* leverages BERT to learn textual features from tweets, and add a fully connected layer with softmax to predict whether the tweets are related to disasters.
- *Image-only* uses the VGG-19 network to capture visual features of images in tweets. Similarly, a fully connected layer with softmax is added for prediction.
- *DMN* is a deep multimodal neural network, which exploits text and images to detect informative disaster-related information in the Twitter stream[2].
- *att-RNN* is a recurrent neural network with an attention mechanism, which can integrate multimodal features for detection of rumors on social media[16].
- *MANN-* is a variant of MANN, which only includes the feature extraction module and the tweet detection module.

In order to select appropriate hyperparameters such as the hidden size and dropout probabilities, iterations of random search are conducted for possible combinations of hyperparameters. Based on the previous iteration step, each iteration reduces the number of possible hyperparameter combinations. In the experiments, we utilize talos,[§] which is a powerful tool for hyperparameter optimization, to perform a random search and evaluate the hyperparameter optimization in each iteration.

In the feature extraction module, MANN leverages $\mathrm{BERT}_{\mathrm{BASE}}$[8] to capture textual features. We set the number of heads and layers to 12, and the dimension of h_c is set to 768. Gelu activation function is used in the BERT model. In the textual and visual feature extractor, the hidden size of the fully connected layer is set to 32 $(d = 32)$. For the tweet detection module and the disaster discrimination module, we set the hidden size of the first fully connected layer to 64, and the second layer to 32. For Text-only and Image-only, we set the hidden sizes of fully connected layers to 32. The dropout probability of each fully connected layer is 0.5. For DMN and att-RNN, if not explicitly mentioned, the parameter settings are the same as their original papers.

[‡] NLTK Stopword List: http://www.nltk.org/

[§] https://github.com/autonomio/talos

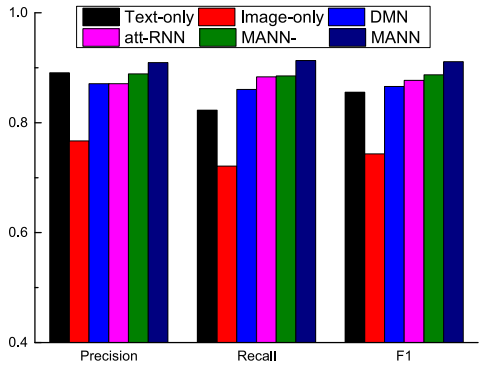

Figure 2. Classification results of six models.

Classification Results

Figure 2 shows the classification results of MANN and baseline models. From the figure, we observe that the performance of the proposed model is better than baseline models in terms of precision, recall, and F1-measure.

For single modality models, the performance of Text-only is much better than Image-only. The reason may be that images contain more disaster-specific features, making it difficult for the Image-only model to extract transferable features between different disasters. Furthermore, in the CrisisMMD dataset, more images are labeled as noninformative, which cannot provide useful information about disasters[4]. Compared with visual features, textual features are more transferable, resulting in better performance of the Text-only model. As a powerful model for extracting textual features, BERT is able to help MANN learn more disaster-invariant patterns contained in the text.

Although textual features are effective for detecting disaster-related tweets, the performance of multimodal models is still better than Text-only. The experimental results verity that textual and visual features are complementary to each other. For multimodal models, the performance of att-RNN is slightly better than the DMN model. The reason may be that att-RNN applies a neuron-level attention mechanism to assign more weight to visual neurons with similar semantics to the corresponding words, which helps to improve classification performance.

MANN-, as a variant of our model, does not contain the disaster discrimination module. Therefore, most of the features extracted by the model are disaster-specific features. This makes it difficult to learn enough sharable patterns between different disasters. On the contrary, MANN utilizes the disaster discrimination module to remove nontransferable features and retain disaster-invariant features. As a result, the complete MANN model achieves the best results on all metrics. The experimental results validate the effectiveness of the disaster discrimination module to improve the performance of disaster-related tweet detection.

Effect of Adversarial Training

In this section, we study the effect of adversarial training in the proposed model [see (12)]. For single modality models Text-only and Image-only, we apply adversarial training to these models. Specifically, by adding a disaster discrimination module, we use adversarial loss to train these models. Next, we employ new single modality models (called Text+ADV and Image+ADV) to classify tweets.

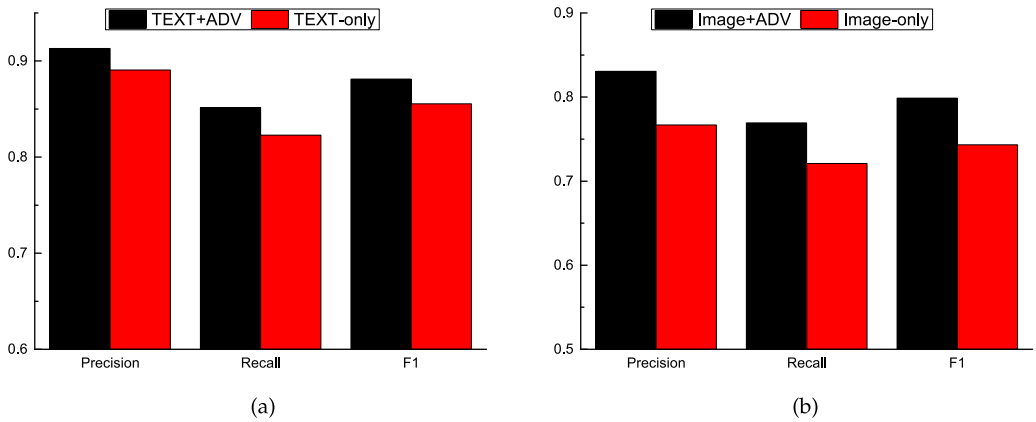

Figure 3. Effect of adversarial training. (a) Classification results of Text+ADV and Text-only. (b) Classification results of Image+ADV and Image-only.

Figure 3 shows the classification results in terms of precision, recall, and F1-measure. As shown in Figure 3(b), we observe that Image+ADV gains significant improvements over Image-only on three metrics. The accuracy of Image+ADV is 8.3% higher than Image-only, and the recall and F1-measure are 6.7% and 7.4% higher, respectively. As shown in Figure 3(a), we observe a similar pattern on the Text+ADV model. The improvement of Image+ADV is slightly greater than Text+ADV, which may be because there more images are noninformative in the CrisisMMD corpus[4]. Experimental results demonstrate that the adversarial training is essential and effective for the extraction of transferable features.

CONCLUSION

In this article, we propose a new MANN that is capable of capturing disaster-invariant features for emerging disaster events. The MANN model consists of three modules: a feature extraction module, a tweet detection module, and a disaster discrimination module. The feature extraction module and the tweet detection module are used to extract multimodal representations and identify disaster-related tweets. The role of the disaster discrimination module is to learn transferable and disaster-invariant features for unseen disaster events. Experiments on a large disaster dataset validate the effectiveness of MANN. In the future, we will explore how to apply the videos of posts to improve the performance of the proposed model.

ACKNOWLEDGMENTS

The authors would like to thank reviewers for their valuable comments. This work was supported in part by Doctoral Start-up Fund of Jianghan University and National Science Foundation of China (NSFC, No. 61772382).

REFERENCES

1. D. Guha-Sapir, P. Hoyois, and R. Below, "Annual disaster statistical review 2015: The numbers and trends," Centre Res. Epidemiology Disasters, Tech. Rep., pp. 1–48, 2016.
2. A. Kumar, J. Singh, Y. Dwivedi, and N. Rana, "A deep multi-modal neural network for informative twitter content classification during emergencies," 2020. [Online]. Available: https://doi.org/10.1007/s10479-020-03514-x
3. Y. Wang, X. Lin, L. Wu, and W. Zhang, "Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1393–1404, Mar. 2017.
4. F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," in *Proc. Int. Conf. Web Soc. Media*, 2018, pp. 465–473.
5. Y. Wang, "Survey on deep multi-modal data analytics: Collaboration, rivalry and fusion," 2020. [Online]. Available: https://arxiv.org/abs/2006.08159
6. I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
7. L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1602–1612, Apr. 2019.
8. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Ch. Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, pp. 4171–4186.
9. A. S. Caragea Caragea and A. H. Tapia, "Identifying informative messages in disaster events using convolutional neural networks," in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage.*, 2016, pp. 137–147.
10. A. Aipe, A. Ekbal, M. NS, and S. Kurohashi, "Linguistic feature assisted deep learning approach towards multi-label classification of crisis related tweets," in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage.*, 2018, pp. 705–717.
11. M. Yu, Q. Huang, H. Qin, C. Scheele, and C. Yang, "Deep learning for real-time social media text classification for situation awareness—Using hurricanes sandy, Harvey, and IRMA as case studies," *Int. J. Dig. Earth*, vol. 12, pp. 1230–1247, 2019.
12. F. Alam, M. Imran, and F. Ofli, "Image4act: Online social media image processing for disaster response," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining*, 2017, pp. 601–604.
13. N. Chaudhuri and I. Bose, "Application of image analytics for disaster response in smart cities," in *Proc. Hawaii Int. Conf. Syst. Sci.*, 2019, pp. 1–10.
14. Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, Oct. 2018.

15. Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, and S. Pan, "Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2153–2159.
16. Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. ACM Multimedia Conf.*, 2017, pp. 795–816.
17. L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1412–1424, Jun. 2019.
18. W. Gao *et al.*, "Generation of topic evolution graphs from short text streams," *Neurocomputing*, vol. 383, pp. 282–294, 2020.

Wang Gao is currently a Lecturer with the School of Mathematics and Computer Science, Jianghan University, Wuhan, China. His main research interests include natural language processing and information retrieval. He is the corresponding author of this article. He received the Ph.D. degree in computer software and theory from Wuhan University, Wuhan, China, in 2019. Contact him at gaow@jhun.edu.cn.

Xun Zhu is currently working toward the Ph.D. degree with Wuhan University, Wuhan, China. She is also an Assistant Professor with the School of Mathematics and Computer Science, Jianghan University, Wuhan, China. Her research interests include natural language processing and data mining. Contact her at zhuxun@jhun.edu.cn.

Yuwei Wang is currently a Lecturer with the School of Mathematics and Computer Science, Jianghan University, Wuhan, China. His current research interests include computer vision and deep learning. He received the M.S. and Ph.D. degrees from the Wuhan University, Wuhan, China, in 2015 and 2019, respectively. Contact him at weberwang@jhun.edu.cn.

Lin Li is currently a Full Time Professor and Ph.D. supervisor with the Wuhan University of Technology, Wuhan, China. Her current research interests include text mining, machine learning, information retrieval, and recommender system. Contact her at cathylilin@whut.edu.cn.