IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Multi-level Multi-modal Cross-attention Network for Fake News Detection

LONG YING[1]*, HUI YU[1]*, JINGUANG WANG[2], YONGZE JI[3], AND SHENGSHENG QIAN[4].

[1]Nanjing University of Information Science and Technology, Nanjing, China (e-mail: lorin_ying@hotmail.com)
[2]Hefei University of Technology, Hefei, China
[3]China University of Petroleum, Beijing, China
[4]National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China (e-mail: shengsheng.qian@nlpr.ia.ac.cn)

Corresponding author: Long Ying (e-mail: lorin_ying@hotmail.com).

*First Author and Second Author contribute equally to this work.
This work is supported in part by the National Natural Science Foundation of China (Grant No. 61902193) and in part by the PAPD fund.

ABSTRACT With the development of the Mobile Internet, more and more users publish multi-modal posts on social media platforms. Fake news detection has become an increasingly challenging task. Although there are many works using deep schemes to extract and combine textual and visual representation in the post, most existing methods do not sufficiently utilize the complementary multi-modal information containing semantic concepts and entities to complement and enhance each modality. Moreover, these methods do not model and incorporate the rich multi-level semantics of text information to improve fake news detection tasks. In this paper, we propose a novel end-to-end *Multi-level Multi-modal Cross-attention Network* (MMCN) which exploits the multi-level semantics of text and jointly integrates the relationships of duplicate and different modalities (textual and visual modality) of social multimedia posts in a unified framework. Firstly, pre-trained BERT and ResNet models are employed to generate high-quality representations for text words and image regions, respectively. A multi-modal cross-attention network is then designed to fuse the feature embeddings of the text words and image regions by simultaneously considering data relationships in duplicate and different modalities. Specially, due to different layers of the transformer architecture have different feature representations, we employ a multi-level encoding network to capture the rich multi-level semantics to enhance the presentations of posts. Extensive experiments on the two public datasets (WEIBO and PHEME) demonstrate that compared with the state-of-the-art models, the proposed MMCN has an advantageous performance.

INDEX TERMS Multi-level Neural Networks; Fake News Detection; Multi-modal Fusion

I. INTRODUCTION

WITH the development of the Mobile Internet and Mobile Communication technologies, social media has become more and more extensive and deeply integrated into our daily life. With easy accessibility, people tend to acquire and share information as well as express and exchange opinions through social multimedia. Unfortunately, due to the openness of social multimedia, a large number of users, and the complexity of sources, various fake news have been fostered on websites. These widespread fake news are utilized by some evil guys to mislead readers, which could do serious harm to society and may cause great economic loss. Ordinary users do not have the time and skill to check the genuineness of every piece of info. Thence, it is necessary and pressing to discover fake news on social multimedia and guarantee users receive authentic info.

Nowadays, there are different kinds of methods [1]–[4] proposed for fake news detection tasks, including traditional machine learning-based and modern deep learning-based methods mainly. Traditional methods [1] such as Support Vector Machine (SVM), Random Forest, and Decision Tree heavily depend on hand-craft features to debunk fake news, which is labor-intensive and time-consuming. For instance, SVM-TS [1] uses a linear classifier based on SVM along with heuristic rules to classify the news as fake or real. With the massive success of the neural networks, existing deep learning-based models have achieved better performance than traditional ones due to their outstanding feature extraction ability. Some early works tried to extract features from plain textual content to identify fake posts. Then it

FIGURE 1. An example of multi-modal post on social media platforms. The upper part of the post is textual content, and the lower part is the attached image.

further exploited Recurrent Neural Networks (RNN) [2] and its variants, such as Long Short-Term Memory (LSTM), to extract temporal language features for fake news detection. On this basis, another study introduces the attention mechanism into RNNs and extracts the sequence language features of specific focal points. Also, some researchers introduce the convolutional neural networks (CNN) [4] to learn the high-level representations extracted from posts on social media platforms for identification.

To date, news content evolves from pure text content to multi-modal content with text, images, and videos. An example of multi-modal news is illustrated in Figure 1. Fake news detection with multi-modality has received more and more attention. Many works [5]–[8] utilize deep schemes to extract and combine textual and visual representation in posts. However, advanced methods which effectively integrate complementary and noisy multi-modal information containing semantic concepts and entities to complement and enhance each modality have not been sufficiently researched. For instance, some models [6], [7] just simply concatenate features extract in different modalities such as text and image together to form the final representation. In [7], the authors propose a multi-modal variational autoencoder (MVAE) to encode and aggregate information of each modality independently and the employ a fully connected network to conduct multi-modal fusion. However, much important information is suppressed due to the modality-independent aggregation operation and the limitation of fully connected networks.

In addition, these methods are still insufficient to utilize the different semantic information of textual content. Most state-of-the-art methods attempt to employ the pre-trained Bidirectional Encoder Representations from Transformers (BERT) [9] model as textual feature extractors due to that rich feature representations can be generated by different layers. However, they usually utilize the representations of the last output layer of the BERT model to conduct fake news detection, which cannot make full use of the intermediate hidden state to capture the abundant textual semantics.

In order to build an effective model to conduct fake news detection, we should address the subsequent challenges:

- *Challenge 1:* How to effectively integrate multi-modality information containing semantic concepts and entities to enhance the performance of fake news detection?
- *Challenge 2:* How to explore and capture the multi-level semantics of textual information to learn the outstanding feature representations of multi-modal posts?

To address the above challenges, we propose a novel end-to-end Multi-level Multi-modal Cross-attention Network (MMCN), which conducts fake news detection via jointly modeling the multi-modal information and the multi-level semantics of textual content into a unified model. To obtain a robust fake news detection model, we employ two available modules consisting of the multi-modal cross-attention module and the multi-level encoding module, which play vital roles in modeling the multi-level semantics and relationships of the textual and visual content of social multimedia posts. (1) For *Challenge 1*, we adopt a novel multi-modal cross-attention network based on extracted fine-grained representations for sentence words and image regions. In particular, BERT is utilized as the language model to encode the sentence words of textual content of news, and a pre-trained ResNet [10] is employed to capture a better image region representation. By jointly considering relationships of data in duplicate and different modalities, the representations of text words and image fragments will complement and enhance each other in semantic space, which can effectively assist in fake news detection. (2) For *Challenge 2*, we design a multi-level encoding network to capture the abundant multi-level semantic features of the post, which integrates the multi-level semantics of the textual information with visual content. The multi-modal features and multi-level semantics are employed simultaneously to calculate the probability that the post is fake or real.

In conclusion, the contributions of our work are as follows:

- We propose a novel Multi-level *Multi-modal Cross-attention Network* (MMCN), which jointly models multi-modal information and multi-level semantics of posts into a unified end-to-end framework for fake news detection.
- A multi-modal cross-attention network is designed to incorporate multi-modal information for each post, which can utilize the relationships between sentence words and image regions to supplement and promote each other for high-quality multi-modal representation. Moreover, the multi-level semantics of the textual information is integrated with visual content to generate multi-level semantic features by the multi-level encoding network, combined to form a comprehensive representation.
- We evaluate the proposed model on the two public real-world datasets (WEIBO and PHEME). Experimental results demonstrate that the proposed MMCN performs better than state-of-the-art (SOTA) baselines.

II. RELATED WORK

With the massive growth of social multimedia content on the Mobile Internet, how to identify fake news becomes increasingly challenging. Many researchers have discussed fake news detection tasks and proposed various approaches [3], [5], [6], [11], [12] that can be roughly divided into two categories: single-modal (e.g., text or images) and multi-modal fake news detection.

For single-modality analysis, most existing methods [3], [11]–[13] extract the features from the textual or visual content of the posts, which have already been explored in the fake news detection literature. For instance, Castillo et al. [12] make use of the decision tree to classify the posts via capturing the topic-based representations of the textual information. Ma et al. [2] extract hidden representations from the textual content of relevant posts via recurrent neural networks (RNN). Yu et al. [4] obtain high-level interactions and critical features of related posts via convolutional neural networks (CNN). In [14], the authors merely utilize the abundant visual information of the posts with different pixel domains and use a novel Multi-domain Visual Neural Network (MVNN) to detect fake news. Dun et al. propose a novel Knowledge-aware Attention Network (KAN) in paper [15] which incorporates external knowledge from knowledge graph for fake news detection via utilizing the knowledge-level relationships among news entities along with the entities and their contexts jointly. Mishra et al. [16] propose a novel Higher-order User to User Mutual-attention Progression (HiMaP) method to capture the cues related to authority or influence of the users by modeling direct and indirect (multi-hop) influence relationships among each pair of users in propagation paths. This method exploits latent relationships among users to model the influence of the users with high prestige on the other users for detecting fake news efficiently. To obtain text representation for fake news detection, Jiang et al. [17] propose a stacking model based on five machine learning models and three deep learning models, such as CNN, LSTM, SVM, etc. Umer et al. [18] design a hybrid Neural Network architecture combining the capabilities of CNN and LSTM, which is used with two different dimensionality reduction approaches containing Principle Component Analysis (PCA) and Chi-Square to reduce the dimensionality of the feature vectors.

In recent years, posts on social multimedia evolve to contain multi-modal content with text, images, and videos. Utilizing the complementary information between different modalities such as textual and visual to extract more comprehensive and accurate features is beneficial to recognize fake news [19]–[22]. Multi-modal fake news detection has received widespread attention.

Primary works [23]–[25] exploit the basic representations of the corresponding images in the given posts, which are mainly hand-crafted, and difficult to learn the complex distributions of image content effectively. Besides, neural networks based on deep learning have already achieved extraordinary performance on nonlinear representation learning [26]–[28]. Most multi-modal representation methods [5]–[8] use deep schemes to extract the representative features and achieve superior performance in fake news detection. In paper [5], Jin et al. design a multi-modal model based on deep learning to identify fake news, which can learn the multi-modal and social context features and make use of attention mechanisms to fuse them. Wang et al. [6] capture event-invariant feature representations between posts to generate the multi-modal features of every post for fake news detection by using a novel adversarial network together with a multi-modal feature extractor. Khattar et al. [7] propose a novel multi-modal variational autoencoder for fake news detection, which obtains the multi-modal representations by feeding the multi-modal features into a bimodal variational autoencoder and jointly learning with a classifier to recognize fake news. Zhou et al. [8] propose the Similarity-Aware Fake news detection method (SAFE) in [8], which adopts neural networks to gain the latent representations of both textual and visual content of the post, and then takes the relationships (similarities) between modalities to form similarity feature. The similarity feature is combined with the concatenation of textual and visual features to conduct fake news detection. In paper [29], Hsu et al. present a multi-modal feature fusion framework to characterize multiple types of features, including image feature, text information, and meta-data, in an effective manner to predict social media popularity.

Except for the above methods, some researchers utilize pre-trained models such as BERT to extract better textual multi-level representations conducting fake news detection tasks. For instance, certain works [30], [31] explore the architecture of the BERT model to extract the multi-level semantic representations of the intermediate hidden layers of BERT, which are favorable to enhance the performance of the model for better conducting the given downstream tasks.

Although these approaches have made a fairly good performance, most of them only use plain text content to learn post representation for fake news detection and ignore the influence of visual information. More importantly, they treat post texts as word sequences to obtain the embedding, which cannot capture the fine-grained and hierarchical textual representation. Moreover, the above multi-modal fusion-based methods are simple and coarse in modeling semantic space across multi-modalities. These methods do not effectively integrate complementary and multi-level multi-modal information to complement and enhance each modality in semantic space for extracting high-quality multi-modal feature representations.

In this paper, we propose Multi-level Multi-modal Cross-attention Networks (MMCN), which can exploit the relationships between sentence words and image regions to supplement and promote each other for high-quality multi-modal representation. Simultaneously, rich multi-level semantics of the textual information is captured and combined to further improve multi-modal fusion for better performance in the fake news detection task.

III. THE PROPOSED ALGORITHM

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114093, IEEE Access

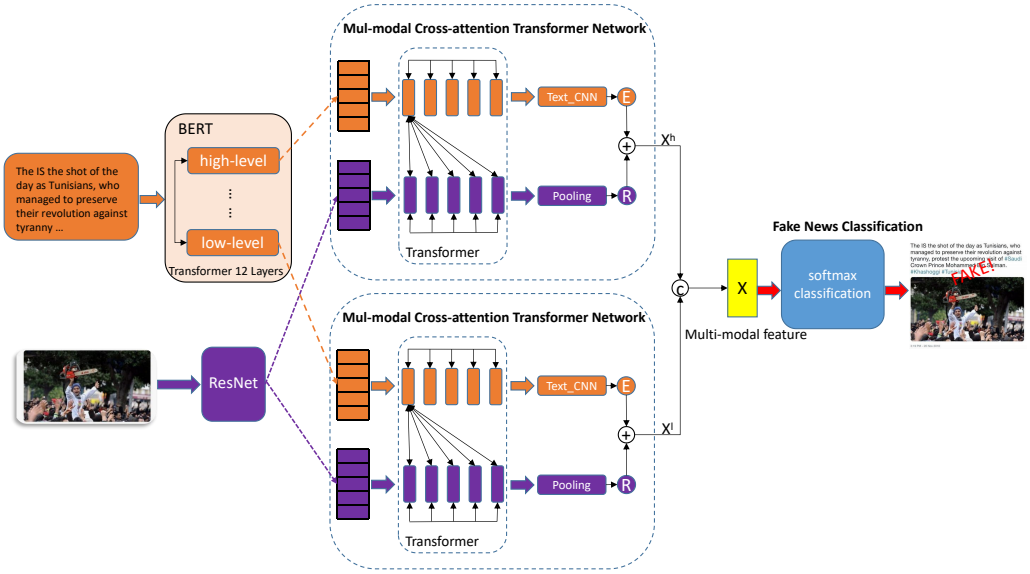


FIGURE 2. The overall framework of the proposed **MMCN**: The inputs consist of the textual content and the attached image of the post. The pre-trained BERT and ResNet50 models are respectively utilized to obtain the embeddings of text words and features of image regions. The dashed box indicates the multi-modal cross-attention network, which can be adopted to jointly model the inter-modality and intra-modality relationships for text words and image regions and aggregate textual and visual fragments for overall features. Multiple multi-modal cross-attention modules constitute the multi-level encoding network, which captures the abundant multi-level semantics of multi-modal content by integrating the multi-level semantics of the textual information with visual content. The final representation denoted as $\boldsymbol{X}$ is formed by the concatenate operation and then fed into a classification network to calculate whether the post is fake or real.

A. PROBLEM DEFINITION

Generally speaking, we define the fake news detection task as a binary classification problem, aiming to classify the post on social multimedia as fake or real news. Specially, given a set of multi-modal posts collected from social multimedia $\mathcal{S} = \{S_1, ..., S_M\}$, where S_i indicates a post consisting of text words and corresponding visual content that mainly comprises images, where M is the number of the posts. Our purpose is to learn a model $f : \mathcal{S} \rightarrow \mathcal{Y}$, to classify each post S_i into the predefined categories $\mathcal{Y} = \{0, 1\}$ where 1 denotes fake news while 0 denotes real news.

B. OVERALL FRAMEWORK

We propose a novel Multi-level Multi-modal Cross-attention Network (MMCN) to enhance the performance of fake news detection. By designing a multi-modal cross-attention network for multi-modal feature fusion and a multi-level encoding network for textual and visual information, our model can capture the intra-modality and inter-modality relationships of multi-modal content in posts, and integrate the multi-level semantics of the textual information with visual content to generate multi-level semantic features. The overall architecture of our model is exhibited in Figure 2. Specifically, our model involves the following components:

- **Text and Image Feature Extractor**: Given a post with multi-modal content which contains text and image, we use word piece tokens of text as the fragments in the textual modality. The pre-trained BERT [9] model is employed to fetch embeddings of word piece tokens. Meanwhile, for each image in the post, we utilize a pre-trained ResNet50 [10] model to extract features corresponding to the partitioned regions. Note that during the training stage, the pre-trained models are fixed.
- **Multi-modal Cross-attention Network**: Based on the extracted fine-grained representations for text words and image regions, we design a multi-modal cross-attention network to jointly model the inter-modality and intra-modality relationships for image regions and text fragments. By taking these relationships into account, features of different modalities can be enhanced and complemented. The text_CNN module and pooling operation are then utilized to aggregate the obtained fragment representations to the multi-modal features, resembling the bag of visual words model.
- **Multi-level Encoding Network**: In order to exploit the multi-level semantics in the given posts, we design a novel multi-level encoding network to model and jointly learn the abundant multi-level semantics of the multi-modal content, which integrates the multi-level semantics of the textual information with visual content. The captured multi-level semantic features are combined to form the comprehensive representation.
- **Fake News Classification Network**: The aim of fake news classification network is to classify each post on social multimedia as fake or real. The classifier takes the learned multi-modal features as input and then and then feeds them into a fully connected network with corresponding activation function to classify posts into predefined categories.

IV. METHODOLOGY

This section presents our multi-level multi-modal cross-attention network for fake news detection.

A. TEXT AND IMAGE FEATURE EXTRACTOR

As mentioned above in subsection IV-A, the input of our model is a multi-modal post $S = \{T, F\}$, where T and F denote the textual content and visual content, respectively.

Text Feature Extractor: To precisely model the semantics of text words, we use BERT model to encode them, which has already been proved to be effective in most domains such as reading comprehension, text classification, and translation [32]–[34].

Given textual content T of the post, we model T as the sequence of textual words $T = \{t_1, t_2, ..., t_n\}$, where n indicates the number of words in the text. Word embeddings of text can be obtained by hidden representation calculated on one output layer of the BERT model, depicted as below:

$$\boldsymbol{E} = \{\boldsymbol{e}_1; \boldsymbol{e}_2 ...; \boldsymbol{e}_n\} = \text{BERT}(T) \tag{1}$$

where $\boldsymbol{E} = \{\boldsymbol{e}_1; \boldsymbol{e}_2; ...; \boldsymbol{e}_n\}$ is the concatenation of word embeddings of text. $\boldsymbol{e}_i \in \mathbb{R}^{d_t}$ is the hidden representation of one output layer of BERT corresponding to text word t_i, d_t indicates the dimension of the word embedding.

Image Feature Extractor: For the processing of visual content in multi-modal data (here, in particular, image), most previous methods use a VGG-based model to cature visual features. However, features extracted by ResNet are more representative and discriminative compared with those extracted by VGG. Thence, in our work, the ResNet model is employed to learn visual features of image information. Some of the current methods simply use the features of the last layer of ResNet, which in some conditions ignore a lot of detailed visual information. In order to better model visual semantic information, we use the penultimate layer of ResNet to extract fine-grained region features of the image.

Given the attached image F, we employ the pre-trained ResNet50 model to extract region features of the image. The feature map obtained on the penultimate pooling layer of the ResNet50 model is considered as the concatenation of region features of the image, depicted as below:

$$\boldsymbol{R} = \{\boldsymbol{r}_1; \boldsymbol{r}_2; ...; \boldsymbol{r}_k\} = \text{ResNet50}(F) \tag{2}$$

where $\boldsymbol{R} = \{\boldsymbol{r}_1; \boldsymbol{r}_2; ...; \boldsymbol{r}_k\}$ denotes the concatenation of features of image regions, $\boldsymbol{r}_i \in \mathbb{R}^{d_f}$ is feature corresponding to specific image region i, d_f is the dimension of region feature, and k is the number of regions.

These pre-trained models are fixed during the training stage, which are chosen in correspondence to earlier work on this problem to compare the efficacy of our architecture.

Since it is usual that $d_t \neq d_f$, we add a 2D-convolutional layer to transform the dimension of image region features to be the same as the dimension of textual word features, denoted as $d_f = d_t = d_s$, which better adapts to the task.

B. MULTI-MODAL CROSS-ATTENTION NETWORK

To effectively fuse the textual and visual features of posts, we synchronously model both the relationships of multi-modal content.

As shown in Figure 2, the multi-modal cross-attention network takes the concatenation of features corresponding to text words and image regions $\boldsymbol{S} = \begin{pmatrix}\boldsymbol{E}\\ \boldsymbol{R}\end{pmatrix} = \{\boldsymbol{e}_1; ...; \boldsymbol{e}_n; \boldsymbol{r}_1; ...; \boldsymbol{r}_k\}$ as the input, where $\boldsymbol{S} \in \mathbb{R}^{(n+k)\times d_s}$, $\boldsymbol{E} = \{\boldsymbol{e}_1; \boldsymbol{e}_2; ...; \boldsymbol{e}_n\}$ and $\boldsymbol{R} = \{\boldsymbol{r}_1; \boldsymbol{r}_2; ...; \boldsymbol{r}_k\}$ are concatenations of features obtained respectively by pre-trained models for text words and image regions, in which image region features have been adapted to the same dimension.

The concatenation of features $\boldsymbol{S}$ is fed into a transformer unit. The query, key, and value for the fine-grained features are formulated as follows:

$$\boldsymbol{K}_S = \boldsymbol{S}\boldsymbol{W}^K = \begin{pmatrix}\boldsymbol{R}\boldsymbol{W}^K\\ \boldsymbol{E}\boldsymbol{W}^K\end{pmatrix} = \begin{pmatrix}\boldsymbol{K}_R\\ \boldsymbol{K}_E\end{pmatrix} \tag{3}$$

$$\boldsymbol{Q}_S = \boldsymbol{S}\boldsymbol{W}^Q = \begin{pmatrix}\boldsymbol{R}\boldsymbol{W}^Q\\ \boldsymbol{E}\boldsymbol{W}^Q\end{pmatrix} = \begin{pmatrix}\boldsymbol{Q}_R\\ \boldsymbol{Q}_E\end{pmatrix} \tag{4}$$

$$\boldsymbol{V}_S = \boldsymbol{S}\boldsymbol{W}^V = \begin{pmatrix}\boldsymbol{R}\boldsymbol{W}^V\\ \boldsymbol{E}\boldsymbol{W}^V\end{pmatrix} = \begin{pmatrix}\boldsymbol{V}_R\\ \boldsymbol{V}_E\end{pmatrix} \tag{5}$$

Then, the Scaled Dot-Product Attention is executed as:

$$Attention(\boldsymbol{Q}_S, \boldsymbol{K}_S, \boldsymbol{V}_S) = softmax\left(\frac{\boldsymbol{Q}_S\boldsymbol{K}_S^T}{\sqrt{d}}\right)\boldsymbol{V}_S \tag{6}$$

To make the derivation understand easily, we remove the softmax function and the scaled function in Eq.(6) without affecting the main idea of our attention mechanism. The equation will be rewritten as:

$$\begin{aligned}\boldsymbol{Q}_S\boldsymbol{K}_S^T\boldsymbol{V}_S &= \begin{pmatrix}\boldsymbol{Q}_E\\ \boldsymbol{Q}_R\end{pmatrix}\begin{pmatrix}\boldsymbol{K}_E^T & \boldsymbol{K}_R^T\end{pmatrix}\begin{pmatrix}\boldsymbol{V}_E\\ \boldsymbol{V}_R\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{Q}_E\boldsymbol{K}_E^T & \boldsymbol{Q}_E\boldsymbol{K}_R^T\\ \boldsymbol{Q}_R\boldsymbol{K}_E^T & \boldsymbol{Q}_R\boldsymbol{K}_R^T\end{pmatrix}\begin{pmatrix}\boldsymbol{V}_E\\ \boldsymbol{V}_R\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{Q}_E\boldsymbol{K}_E^T\boldsymbol{V}_E + \boldsymbol{Q}_E\boldsymbol{K}_R^T\boldsymbol{V}_R\\ \boldsymbol{Q}_R\boldsymbol{K}_R^T\boldsymbol{V}_R + \boldsymbol{Q}_R\boldsymbol{K}_E^T\boldsymbol{V}_E\end{pmatrix}\end{aligned} \tag{7}$$

After the above attention layer $\begin{pmatrix}\boldsymbol{E}_{up}\\ \boldsymbol{R}_{up}\end{pmatrix} = \boldsymbol{Q}_S\boldsymbol{K}_S^T\boldsymbol{V}_S$, the updated features of the textual and visual fragments are calculated as follows:

$$\boldsymbol{E}_{up} = \{\boldsymbol{e}_{up1}; ...; \boldsymbol{e}_{upn}\} = \boldsymbol{Q}_E\boldsymbol{K}_E^T\boldsymbol{V}_E + \boldsymbol{Q}_E\boldsymbol{K}_R^T\boldsymbol{V}_R. \tag{8}$$

$$\boldsymbol{R}_{up} = \{\boldsymbol{r}_{up1}; ...; \boldsymbol{r}_{upk}\} = \boldsymbol{Q}_R\boldsymbol{K}_R^T\boldsymbol{V}_R + \boldsymbol{Q}_R\boldsymbol{K}_E^T\boldsymbol{V}_E, \tag{9}$$

These results reveal that the output of the multi-head sublayer in the transformer unit considers the multi-modal relationships. And then $\begin{pmatrix}\boldsymbol{E}_{up}\\ \boldsymbol{R}_{up}\end{pmatrix}$ is fed into the followed position-wise forward sublayer. Ultimately, we obtain the output of the transformer unit and formulate it as follows: $\boldsymbol{S}_c = \begin{pmatrix}\boldsymbol{E}_c\\ \boldsymbol{R}_c\end{pmatrix}$.

We split $\boldsymbol{S}_c$ into $\boldsymbol{E}_c = \{\boldsymbol{e}_{c1}, ..., \boldsymbol{e}_{cn}\}$ and $\boldsymbol{R}_c = \{\boldsymbol{r}_{c1}, ..., \boldsymbol{r}_{ck}\}$. In general, a fully connected layer or pooling layer is used to obtain the aggregated final feature of text words. However, we find that employing the Text_CNN module, which is designed to capture local context textual patterns resembling n-gram before aggregating, can achieve better performance. Thence, the concatenation of word features $\boldsymbol{E}_c$ is fed into Text_CNN for aggregating, which can be written as:

$$\boldsymbol{E}_o = \text{Text\_CNN}(\boldsymbol{E}_c) \tag{10}$$

An average pooling layer is utilized to aggregate the concatenation of image region features $\boldsymbol{R}_c$, which can be depicted as:

$$\boldsymbol{R}_o = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{r}_{ci} \tag{11}$$

The final multi-modal feature representations of posts are got by sum operation between $\boldsymbol{E}_o$ and $\boldsymbol{R}_o$, which can be denoted as:

$$\boldsymbol{X} = \lambda \boldsymbol{E}_o + (1-\lambda)\boldsymbol{R}_o \tag{12}$$

where λ is the balance factor of the proportion of text and visual info in the multi-modal features.

C. MULTI-LEVEL ENCODING NETWORK

As we know, BERT can provide multi-level semantics for text, which includes the outputs of 11 intermediate layers and one final output layer. According to the attention allocated characteristics of the different layers in Transformer architectures, the features in different layers emphasize different views for data samples. For instance, the features in lower layers tend to encode more local contents with basic syntactic representations, while higher layers capture more complex semantics and usually produce higher-level semantic representations. Intuitively, in order to utilize the rich semantics in the layers of BERT sufficiently, we employ multiple multi-modal cross-attention networks on representations of different output layers of BERT.

Specially, we utilize the representations of the first output layer of BERT model as the low-level basic syntactic features of text words and the representations of the last output layer of BERT as the high-level semantic features of text words. In order to explore the multi-level semantic information of posts, we design two multi-modal cross-attention network units to construct the multi-level encoding network, which employs different text word features such as high-level and low-level features and the corresponding image region features as the input. Then two multi-modal representations corresponding to different semantic levels are calculated, denoted as $\boldsymbol{X}^h$, $\boldsymbol{X}^l$ respectively. According to Eq.(12), the high-level and low-level semantic features can be obtained as follows:

$$\boldsymbol{X}^h = \lambda_1 \boldsymbol{E}_{ho} + (1-\lambda_1)\boldsymbol{R}_{ho} \tag{13}$$

$$\boldsymbol{X}^l = \lambda_2 \boldsymbol{E}_{lo} + (1-\lambda_2)\boldsymbol{R}_{lo} \tag{14}$$

Note that the two multi-modal cross-attention transformer networks have shared weights.

Ultimately, we concatenate the outputs of two units:

$$\boldsymbol{X} = concat(\boldsymbol{X}^h, \boldsymbol{X}^l) \tag{15}$$

where $concat$ denotes concatenate operation, $\boldsymbol{X}$ is the final output multi-modal representation of the given post generated by the proposed multi-level multi-modal cross-attention transformer network.

D. FAKE NEWS CLASSIFICATION NETWORK

In this part, we present the fake news classification network. Based on the multi-modal representation of posts $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\}$, the fake news classification network is utilized to classify posts as fake or real news. It applies a fully connected layer along with the corresponding softmax activation function to predict whether the post is fake or not. It can be formalized as follows:

$$\hat{p}_m = \sigma(\boldsymbol{W}_f \boldsymbol{x}_m + b) \tag{16}$$

where $\sigma(.)$ denotes softmax activation function, $\hat{p}_m$ denotes the classifying probability that post m is fake, and $\boldsymbol{x}_m$ is the feature representation of the post m. We use y_m to represent the ground-truth labels of post m and utilize the cross entropy loss function to calculate the total loss:

$$\mathcal{L}(\theta) = \frac{1}{M}\sum_{m=1}^{M} -[y_m \log(\hat{p}_m) + (1-y_m)\log(1-\hat{p}_m)] \tag{17}$$

where M is the number of posts. We minimize the classification loss by seeking the optimal parameters θ^*, which can be defined as follows:

$$(\theta^*) = \arg\min_{\theta} \mathcal{L}(\theta) \tag{18}$$

V. EXPERIMENTAL RESULTS

In this section, we carry out experiments to evaluate our model (MMCN) against SOTA models on two public datasets. Moreover, we provide an experimental analysis in detail into point out more insights in our model.

TABLE 1. The statistics of two Real-World Datasets.

News	WEIBO	PHEME
# of Fake News	4749	1972
# of Real News	4779	3830
# of Images	9528	3670

A. DATASET

Considering the sparse availability of structured multimedia data, we compare the proposed approach with SOTA baselines on two public real-world datasets (WEIBO [5] and PHEME [35]). Each dataset consists of a large number of texts and the attached images with labels. The statistics of the two datasets are shown in Table 1.

1) WEIBO Dataset

The data of WEIBO dataset [5] is collected from Xinhua News Agency[1] and Weibo[2]. The former is an authoritative news source, and the latter is a Chinese microblog website. The data has been collected from a timespan of May 2012 to January 2016. The dataset consists of 9528 posts, including 4749 fake posts, 4779 real posts, and 9528 images corresponding to posts. Each post in WEIBO dataset contains text and the attached unique image. Posts in the dataset are verified by Xinhua News Agency as fake or real news.

2) PHEME Dataset

The PHEME dataset [35] consists of data based on five breaking news, including charliehebdo, ferguson, germanwingscrash, ottawashooting, and sydneysiege. Each news involves a set of posts, including a sizable amount of texts and images corresponding to the tweets with labels.

B. BASELINES

To validate the performance, We tend to compare our model(MMCN) with two categories of the SOTA models: unimodal and multi-modal models for fake news detection.

- *SVM-TS* [1]: SVM-TS uses a linear classifier based on SVM along with heuristic rules to predict the news fake or real.
- *GRU* [2]: GRU views the content of posts as variable-length time series via employing a multi-layer GRU network to detect fake news.
- *CNN* [4]: CNN uses a convolutional neural network with fixed-length windows on posts to capture the features.
- *TextGCN* [36]: Text Graph Convolution Network (TextGCN) models the whole corpus as a heterogeneous graph and feeds it into the GCN to obtain the textual semantic features.
- *att_RNN* [5]: att-RNN proposes a novel RNN based on attention mechanism to learn better multi-modal features. Specially, it incorporates the feature representations of the text and the relevant social context via utilizing the LSTM module. In order to a fair comparison, we eliminate the part addressing social context info.
- *EANN* [6]: Event Adversarial Neural Network (EANN) learns event-invariant multi-modal features of each post for fake news detection by employing an adversarial network to eliminate event-specific feature representations from the posts features base on the concatenation of extracted textual and visual features.
- *MVAE* [7]: Multimodal Variational AutoEncoder (MVAE) employs a variational autoencoder with encoder and decoder for each modality to obtain a shared multi-modal representation between text and image, which is trained jointly with a classifier for fake news detection.
- *SAFE* [37]: SAFE adopts neural networks to gain the latent representations of both texts and images and then takes the relationship (similarity) between modalities as feature combined with the concatenation of feature and visual feature to conduct fake news detection.

In addition, we also design several variants to demonstrate the effectiveness of every component in our proposed model. Then, we will introduce the details of the variants in the analysis of MMCN components in subsection V-E.

C. EXPERIMENTAL SETTING

Generally speaking, we use Accuracy as the evaluation metric for fake news detection task, which can be viewed as a binary classification task. However, when the dataset suffers from class imbalance, its reliability will be reduced. Thence, apart from Accuracy, we additionally add Precision, Recall and F1 score as complementary evaluation metrics of different categories for fake news detection tasks.

For the given dataset, we split it into the training set and test set according to the ratio of 8:2.

Given multi-modal posts, for textual content, we exploit the pre-trained BERT [9] module for the textual branch, which consists of 12 heads and 12 attention layers, where the dimension of hidden units is 768 for each token. For simplicity, we fix the weights of BERT during the training phase. For the visual branch, feature map calculated on the penultimate layer in the pre-trained ResNet50 model [10] is fetched as the feature tensor concatenated by region features, whose shape is $4 \times 4 \times 2048$. And we add a 2D-convolutional layer to transform the last dimension from 2048 to 768 to adapt our task. We directly use the pre-trained BERT and ResNet50 models provide by the relevant works on the Internet[3]. For the whole model, we utilize the Adam optimizer [38] during training stage. We set the learning rate as 0.001 for 150 epochs, and the batch size is set to 64 to start training on the WEIBO dataset. On the PHEME dataset, we start training along with a learning rate of 0.001 for 150 epochs and the batch size is set to 256.

D. QUANTITATIVE RESULTS

We show detailed fake news detection results across all methods for WEIBO and PHEME in Table 2, from which we can obtain the subsequent observations:

1) In two real-world datasets (WEIBO and PHEME), we can see that SVM-TS has the worst performance regardless of the single-modal methods or the multi-modal methods, which indicates that the hand-crafted feature representations cannot be enough to detect fake news.
2) Among the two datasets, we can see that deep learning models are superior to traditional machine learning models. Compared with SVM-TS, deep learning models such as CNN, GRU, and TextGCN have better performance. On the two experimental datasets, we can observe that CNN performs inferior than most baseline

[1]http://www.xinhuanet.com/
[2]https://weibo.com/
[3]https://huggingface.co/models

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114093, IEEE Access

TABLE 2. The results of comparison among different models on WEIBO and PHEME datasets.

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	SVM-TS	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	GRU	0.702	0.671	0.794	0.727	0.747	0.609	0.671
	CNN	0.740	0.736	0.756	0.744	0.747	0.723	0.735
	TextGCN	0.787	0.975	0.573	0.727	0.712	0.985	0.827
	att_RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	MMCN	**0.879**	0.886	0.871	0.879	0.873	0.888	0.880
PHEME	SVM-TS	0.639	0.546	0.576	0.560	0.729	0.705	0.717
	GRU	0.832	0.782	0.712	0.745	0.855	0.896	0.865
	CNN	0.779	0.732	0.606	0.663	0.799	0.875	0.835
	TextGCN	0.828	0.775	0.735	0.737	0.827	0.828	0.828
	att_RNN	0.850	0.791	0.749	0.770	0.876	0.899	0.888
	EANN	0.681	0.685	0.664	0.694	0.701	0.750	0.747
	MVAE	0.852	0.806	0.719	0.760	0.871	0.917	0.893
	SAFE	0.811	0.827	0.559	0.667	0.806	0.940	0.866
	MMCN	**0.872**	0.837	0.780	0.807	0.888	0.920	0.904

TABLE 3. The results of comparison among different variants of MMCN on WEIBO and PHEME dataset.

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	MMCN$\neg v$	0.819	0.842	0.786	0.813	0.799	0.852	0.825
	MMCN$\neg a$	0.839	0.807	0.892	0.848	0.879	0.786	0.830
	MMCN$\neg l$	0.865	0.878	0.849	0.863	0.853	0.882	0.867
	MMCN$\neg h$	0.776	0.795	0.744	0.769	0.759	0.807	0.782
	MMCN$\neg t$	0.813	0.838	0.777	0.806	0.791	0.849	0.819
	MMCN	**0.879**	0.886	0.871	0.879	0.873	0.888	0.880
PHEME	MMCN$\neg v$	0.870	0.782	0.863	0.821	0.924	0.874	0.898
	MMCN$\neg a$	0.841	0.803	0.713	0.755	0.857	0.908	0.882
	MMCN$\neg l$	0.861	0.784	0.825	0.804	0.905	0.880	0.893
	MMCN$\neg h$	0.856	0.784	0.805	0.794	0.896	0.883	0.890
	MMCN$\neg t$	0.870	0.840	0.773	0.805	0.885	0.923	0.904
	MMCN	**0.872**	0.837	0.780	0.807	0.888	0.920	0.904

methods. The main reason is that CNN cannot capture the long-distance semantic relationships between words, which are beneficial to detect fake news. Moreover, TextGCN performs better than SVM-TS and CNN on experimental datasets, indicating that using graph convolutional network can enhance the performance of the model for fake news detection.

3) Across all datasets, most of the multi-modal methods lead to better accuracy compared with single-modal methods, indicating that the additional visual information can be used as complementary information to facilitate fake news detection. For instance, as multi-modal models, att-RNN has relatively better performance, showing that the attention mechanism can take the regions of images corresponding to text segments into consideration and enhance the performance of the whole model. The performance of SAFE on WEIBO and PHEME datasets demonstrates the effectiveness of integrating similarity features between elements of different modalities.
4) The performance of MVAE is better than other multi-model methods on all of the two datasets, showing that self-supervised loss incorporated in the multi-modal representation generation process may take the role of regular to improve the generalization ability. However, the performance of EANN is relatively worse, leaking that in many situations removing event-specific features depraves the discriminative power of multi-modal representations for posts.
5) The MMCN we proposed consistently performs superior to all the SOTA baselines among the two datasets, which shows that our model has the ability to generate more accurate, complementary, and comprehensive multi-level multi-modal representation by joint modeling the intra-modal and inter-modal relationship with the multi-level semantics of text in a unified end-to-end model to detect fake news.

E. ANALYSIS OF MMCN COMPONENTS

Because the proposed MMCN includes multiple vital components, in this section, we will compare the variants of MMCN about the following respects to illustrate the effectiveness of MMCN:

- MMCN$\neg a$: A variant of MMCN with the multi-modal cross-attention transformer network being removed.
- MMCN$\neg l$: A variant of MMCN with the high-level and low-level representations in Transformer being removed, and only utilizes the output of the last layer of BERT.
- MMCN$\neg h$: A variant of MMCN with the high-level and high-level representations in Transformer being removed, and only utilizes the output of the first layer of BERT.
- MMCN$\neg t$: A variant of MMCN with the Text_CNN module being removed, and only utilizes the pooling layer to gain fine-grained feature representations of text words.
- MMCN$\neg v$: A variant of MMCN with the visual information being removed.

We perform ablation analysis with the variants of MMCN, and the results are displayed in Table 3. And then, we make the following conclusions.

(1) *Effects of multi-modal cross-attention transformer network*: We compare the performance of MMCN with MMCN$\neg a$ on the two datasets (WEIBO and PHEME). It can observe that It can observe that our proposed MMCN has better performance than MMCN$\neg a$. The result confirms the superiority of introducing the multi-modal cross-attention network to our model.

(2) *Effects of multi-level encoding network*: We compare the performance of MMCN with MMCN$\neg l$ and MMCN$\neg h$ on the two datasets (WEIBO and PHEME). It can observe that the proposed MMCN performs better than than MMCN$\neg l$ and MMCN$\neg h$. The experimental result confirms the superiority of introducing the multi-level encoding network to capture high-level and low-level semantic features in our model, which exploits the multi-level semantics of textual information of posts. Besides, it can be seen that compared with low-level features that include more basic syntactic information, high-level features containing more semantic information have a larger impact on the performance of the proposed model.

(3) *Effects of the Text_CNN module*: We compare the performance of MMCN with MMCN$\neg t$ on the two datasets (WEIBO and PHEME). It can observe that the proposed MMCN has superior performance than MMCN$\neg t$, which shows that introducing the Text_CNN module to our model is sensible.

(4) *Effects of the visual information*: We compare the performance of MMCN with MMCN$\neg v$ on the two datasets (WEIBO and PHEME). It can observe that the proposed MMCN has superior performance than MMCN$\neg v$. The result shows that the visual information can consistently provide complementary information to improve our model.

F. IMPACTS OF THE VALUES OF λ_1 AND λ_2

In order to learn the rich multi-level semantic representations of multi-modal information, we utilize two different multi-modal cross-attention networks to capture the high-level and low-level representations. The output of the high-level multi-modal cross-attention network unit is $\boldsymbol{X}^h = \lambda_1 \boldsymbol{E}_{ho} + (1 - \lambda_1)\boldsymbol{R}_{ho}$, where $\lambda_1 \in \{0, 1\}$. Also, the output of the low-level multi-modal cross-attention network unit is $\boldsymbol{X}^l = \lambda_2 \boldsymbol{E}_{lo} + (1 - \lambda_2)\boldsymbol{R}_{lo}$, where $\lambda_2 \in \{0, 1\}$. To find suitable λ_1 and λ_2 values, we respectively vary λ_1 and λ_2 from 0.0 to 1.0, and evaluate the impacts for the accuracies of fake news detection on the two datasets. The results are shown in Figure 3(a) and Figure 3(b).

When λ_1 grows from 0.0 to 1.0 and λ_2 is fixed to 0.7 according to experience, the accuracy of our model keeps changing. In Figure 3(a), when the value of λ_1 is 0.7, the accuracy has the highest results on WEIBO and PHEME datasets simultaneously.

For λ_2, we fix λ_1 to 0.7 and vary the value of λ_2 from 0.1 to 1.0 to represent the impact for the accuracy of fake news detection on the two datasets. It can observe in Figure 3(b) that when λ_2 is 0.1, the accuracy is the highest on WEIBO dataset. On PHEME dataset, when λ_2 is 0.5, the accuracy is the highest. Considering the compositive performance of the proposed model on WEIBO and PHEME datasets simultaneously, the accuracy on the two datasets can achieve satisfactory results when λ_2 is 0.7.

Therefore, we set $\lambda_1 = 0.7$ and $\lambda_2 = 0.7$ on the two datasets so that MMCN can achieve relatively good performance.

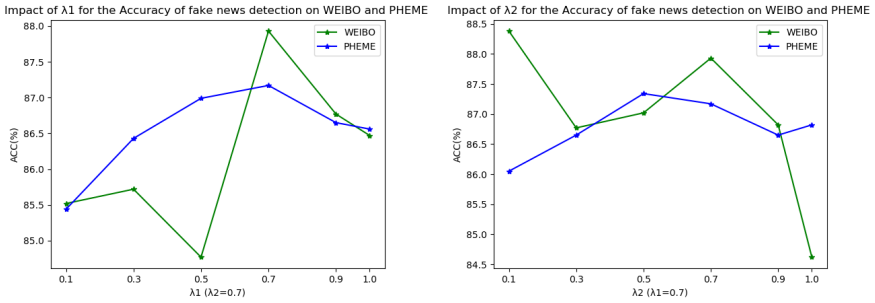
G. FAILURE CASES STUDY

To further illustrate the performance of the proposed method, we collect and analyze some failure cases. Figure 4(a) and Figure 4(b) are two fake news examples that the proposed method failed to detect. In Figure 4(a), the text content of the post is very short, leading to poor performance of the proposed multi-level multi-modal cross-attention network. In Figure 4(b), the scene of the attached image in the post is very complicated, which also contains text information, and the meaning expressed by this image is highly abstract. All of these make the model difficult to extract and represent the contained semantic information and fail to detect it as a piece of fake news. Besides, in order to explore better performance, we will try to fix these weaknesses in our future work.

VI. CONCLUSION

In this paper, we have proposed a novel end-to-end *Multi-level Multi-modal Cross-attention Network* (MMCN), which jointly models the multi-modal information and the multi-level semantics of textual content into a unified model for fake news detection. We argue that existing methods are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114093, IEEE Access



(a) (b)

FIGURE 3. Impacts of different $\lambda 1$ and $\lambda 2$ for the Accuracy of the proposed model on WEIBO and PHEME datasets.

(a) 天哪！ (Oh, My god!) (b) WAKE UP Ferguson http://tco/4N8kRGBXUf

FIGURE 4. Illustration of some fake posts which cannot be correctly classified by the proposed MMCN.

difficult to utilize the multi-modal context information for each post. In addition, existing approaches mainly ignore the rich multi-level semantics of textual content that can help learn a better news feature representation to enhance the detection of fake news. To address these limitations, we design a novel multi-modal cross-attention network based on learned fine-grained representations for sentence words and image regions. In addition, we design a multi-level encoding network to capture the abundant multi-level semantics for fake news detection. The multi-modal representations and multi-level semantics are employed simultaneously to calculate the probability that the input post is fake or real. Existing methods mainly ignore the variability of word relations in the different backgrounds and different types of information appearing in posts. In future work, we intend to seek a more effective way to utilize background knowledge in deep networks, which can supply useful complementary information for fake news detection.

ACKNOWLEDGMENT

This work is partially funded by the National Natural Science Foundation of China (Grant No. 61902193); and in part by the PAPD fund. The first author and the second author make equal contributions to this work.

REFERENCES

[1] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1751–1754.

[2] 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," 2016.

[3] 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.

[4] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan *et al.*, "A convolutional approach for misinformation identification." in *IJCAI*, 2017, pp. 3901–3907.

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

[6] 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 the 24th acm sigkdd international conference on knowledge discovery & data mining*, 2018, pp. 849–857.

[7] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The World Wide Web Conference*, 2019, pp. 2915–2921.

[8] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020, pp. 354–367.

[9] 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.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR 2016*, 2016, pp. 770–778.

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

[12] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th international conference on data mining*. IEEE, 2013, pp. 1103–1108.

[13] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: Real-time credibility assessment of content on twitter," in *International Conference on Social Informatics*. Springer, 2014, pp. 228–243.

[14] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 518–527.

[15] Y. Dun, K. Tu, C. Chen, C. Hou, and X. Yuan, "Kan: Knowledge-aware attention network for fake news detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 81–89.

[16] R. Mishra, "Fake news detection using higher-order user to user mutual-attention progression in propagation paths," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 652–653.

[17] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, "A novel stacking approach for accurate detection of fake news," *IEEE Access*, vol. 9, pp. 22 626–22 639, 2021.

[18] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (cnn-lstm)," *IEEE Access*, vol. 8, pp. 156 695–156 706, 2020.

[19] X. Wu, C.-W. Ngo, and A. G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 188–199, 2008.

[20] I. Kalamaras, A. Drosou, and D. Tzovaras, "Multi-objective optimization for multimodal visualization," *IEEE transactions on multimedia*, vol. 16, no. 5, pp. 1460–1472, 2014.

[21] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 64–78, 2014.

[22] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1062–1075, 2018.

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

[24] 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.

[25] D. ping Tian *et al.*, "A review on image feature extraction and representation techniques," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 8, no. 4, pp. 385–396, 2013.

[26] L. Zhao, Q. Hu, and W. Wang, "Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1936–1948, 2015.

[27] T.-K. Yan, X.-S. Xu, S. Guo, Z. Huang, and X.-L. Wang, "Supervised robust discrete multimodal hashing for cross-media retrieval," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 1271–1280.

[28] Z. Zhao, Q. Yang, H. Lu, T. Weninger, D. Cai, X. He, and Y. Zhuang, "Social-aware movie recommendation via multimodal network learning," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 430–440, 2017.

[29] C.-C. Hsu, L.-W. Kang, C.-Y. Lee, J.-Y. Lee, Z.-X. Zhang, and S.-M. Wu, "Popularity prediction of social media based on multi-modal feature mining," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2687–2691.

[30] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the dark secrets of BERT," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 2019, pp. 4364–4373.

[31] Y. Song, J. Wang, Z. Liang, Z. Liu, and T. Jiang, "Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference," *CoRR*, vol. abs/2002.04815, 2020. [Online]. Available: https://arxiv.org/abs/2002.04815

[32] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[33] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019. [Online]. Available: http://arxiv.org/abs/1910.01108

[34] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, ser. Lecture Notes in Computer Science, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds., vol. 11856. Springer, 2019, pp. 194–206. [Online]. Available: https://doi.org/10.1007/978-3-030-32381-3_16

[35] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in *International Conference on Social Informatics*. Springer, 2017, pp. 109–123.

[36] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7370–7377.

[37] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," *ArXiv*, vol. abs/2003.04981, 2020.

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