Hierarchical Multi-modal Contextual Attention Network for Fake News Detection

Shengsheng Qian
National Lab of Pattern Recognition,
Institute of Automation, CAS
University of Chinese Academy of Sciences
shengsheng.qian@nlpr.ia.ac.cn

Jinguang Wang
HeFei University of Technology
wangjinguang502@gmail.com

Jun Hu
National Lab of Pattern Recognition,
Institute of Automation, CAS
hujunxianligong@gmail.com

Quan Fang
National Lab of Pattern Recognition,
Institute of Automation, CAS
University of Chinese Academy of Sciences
qfang@nlpr.ia.ac.cn

Changsheng Xu
National Lab of Pattern Recognition,
Institute of Automation, CAS
University of Chinese Academy of Sciences
Peng Cheng Laboratory
csxu@nlpr.ia.ac.cn

ABSTRACT

Nowadays, detecting fake news on social media platforms has become a top priority since the widespread dissemination of fake news may mislead readers and have negative effects. To date, many algorithms have been proposed to facilitate the detection of fake news from the hand-crafted feature extraction methods to deep learning approaches. However, these methods may suffer from the following limitations: (1) fail to utilize the multi-modal context information and extract high-order complementary information for each news to enhance the detection of fake news; (2) largely ignore the full hierarchical semantics of textual content to assist in learning a better news representation. To overcome these limitations, this paper proposes a novel hierarchical multi-modal contextual attention network (HMCAN) for fake news detection by jointly modeling the multi-modal context information and the hierarchical semantics of text in a unified deep model. Specifically, we employ BERT and ResNet to learn better representations for text and images, respectively. Then, we feed the obtained representations of images and text into a multi-modal contextual attention network to fuse both inter-modality and intra-modality relationships. Finally, we design a hierarchical encoding network to capture the rich hierarchical semantics for fake news detection. Extensive experiments on three public real datasets demonstrate that our proposed HMCAN achieves state-of-the-art performance.

CCS CONCEPTS

• **Information systems → Multimedia information systems**; *Social networks.*

KEYWORDS

Social media; Fake news detection; Multi-modal learning

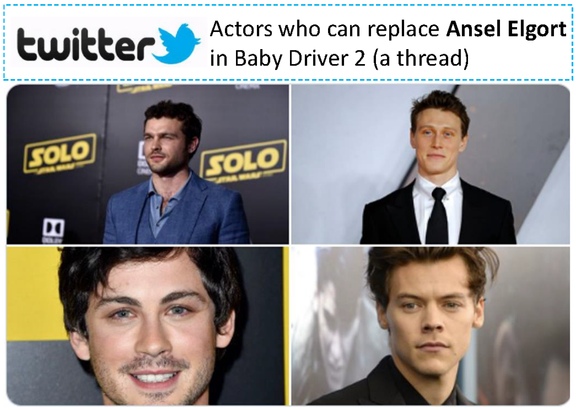
ACM Reference Format:
Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462871

1 INTRODUCTION

Social media websites are convenient platforms for people to share information, express and exchange opinions in their daily life. The ever-increasing number of users has resulted in a variety of information data on social media websites. However, the authenticity of these information is difficult to guarantee since users do not check the reliability of the shared information, which has led to the widespread dissemination of considerable fake news. In addition, without proper supervision, these fake news can easily mislead readers and even cause serious social consequences. Therefore, detecting fake news on social media websites to ensure that users obtain true information has become a top priority.

In order to facilitate the detection of fake news, many approaches have been proposed. The early attempts (e.g., snopes.com) mainly reported the fake news through users, and then invited experts or institutions in related fields to confirm, which is time-consuming and labor-intensive. Therefore, automatically detecting fake news has been a key research direction and drawn much attention in recent years. Basically, existing studies on automatic fake news detection can be summarized into two categories: (1) The first one is traditional learning methods [2, 16, 18, 20], which design plenty of hand-crafted features from the media content of posts and the

SIGIR ’21, July 11–15, 2021, Virtual Event, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462871



Actors who can replace **Ansel Elgort** in Baby Driver 2 (a thread)

Figure 1: An example of multi-modal social media post from Twitter.

social context of users. With these sophisticated features, SVM classifiers [2, 20] and decision tree classifiers [16, 18] have been trained to debunk fake news. However, the content of fake news is highly complicated and hard to be fully captured by hand-crafted features. (2) The second one is deep learning approaches [19, 21, 40], which capture deep features based on neural networks. For example, Ma et al. [19] employ Recurrent Neural Networks (RNNs) to learn the hidden features from posts. Yu et al. [40] use Convolutional Neural Networks (CNNs) to obtain key features and their high-level interactions from fake news. However, most of the above methods focus only on textual content and ignore posts with multi-modal information (such as text, images, etc.), which is a key component of social media platforms.

Recently, as deep neural networks (DNN) have achieved extraordinary performance on nonlinear representation learning, many multi-modal representation methods utilize deep schemes to learn the representative features, and obtain superior performance for fake news detection. Khattar et al. [13] propose the multimodal Variational Autoencoder (MVAE) to obtain the latent multimodal representations of multimedia posts and classify them via a binary classifier. Cui et al. [5] propose an end-to-end deep embedding framework (SAME) to detect fake news, where users' latent sentiments can be utilized to help identify fake news. In [29], Shivangi et al. utilize the pre-trained BERT to learn text features, and apply VGG-19 pre-trained on ImageNet dataset to learn image features. Although these approaches [13, 29] show promising performance on fake news detection tasks, they are still insufficient to take advantage of the multi-modal context information and the hierarchical semantics of text content.

(1) In open systems such as Twitter[1] and Weibo[2], as shown in Figure 1, news posts usually consist of multi-modal content data such as text and images and the well-grounded understanding of the data from each modality relies on the multi-modal context, where data from different modalities can complement each other. For example, due to the openness of these systems, the visual content of news posts usually contain many uncertain elements that are difficult to understand without the help of the text information, and the text content of news posts may refer to some things whose details are shown in the visual content. Although some existing fake news detection approaches already consider the multi-modal context information, the components they employ to capture multi-modal context are too simple, which are insufficient to extract high-order complementary information from the multi-modal context. For example, MVAE [13] encodes and pools information from different modalities independently, and passes the learned representations into a simple fully-connected network to model the multi-modal context. Obviously, a lot of important information is suppressed due to the modality-independent pooling operation, and the high-level complementary information can hardly be learned due to the limitation of fully-connected networks. Therefore, we have to face the *Challenge 1*: How to fully utilize the multi-modal context information and extract high-order complementary information from it to enhance the performance of fake news detection?

(2) Most state-of-the-art fake news detection models use modern text-pretrain deep models such as BERT [7] as text encoders, which can provide hierarchical semantics of text. However, most of them only utilize the output of the last layers of these hierarchical models, while ignoring the intermediate hidden states, which also capture rich linguistic information [11]. There are some works [15, 30] that explore the potential of exploiting the semantic knowledge in the intermediate layers of BERT models, showing that many downstream tasks can benefit from the full hierarchical semantics. Therefore, we need to address the *Challenge 2*: How to explore and capture the hierarchical semantics of text information to learn a better representation of multi-modal news?

To address the above limitations, we propose a novel hierarchical multi-modal contextual attention network (HMCAN) for fake news detection by jointly modeling the multi-modal context information and the hierarchical semantics of text in a unified deep model. To achieve a robust fake news detection, we design two practical modules including multi-modal contextual attention module and hierarchical encoding module, which play important roles in modeling the hierarchical semantics and relationships of textual and visual content of fake news. (1) For *Challenge 1*, we propose a multi-modal contextual attention network to model the multi-modal context for each news posts, where data from different modalities can complement each other to provide a better understanding of the multi-modal data. Specifically, a pre-trained ResNet [9] is utilized to learn better representations of images and the BERT [7] is used as the language model to embed the textual content of news. Then the obtained representations of images and textual content are fed into a multi-modal contextual attention network to fuse both inter-modality and intra-modality relationships to complement each other and assist in fake news detection. (2) For *Challenge 2*, we design a hierarchical encoding network to capture the rich hierarchical semantics for fake news detection.

[1] https://twitter.com/
[2] https://weibo.com/

The multi-modal contextual features and hierarchical semantics are employed simultaneously to predict the verification score of the input fake news. Extensive experiments on three public real datasets demonstrate that the proposed fake news detection model outperforms the state-of-the-art methods.

In summary, our contributions can be summarized as follows:

- We propose a novel hierarchical multi-modal contextual attention network (HMCAN) by jointly modeling of the multi-modal context information and the hierarchical semantics of text in a unified deep model for fake news detection.
- A multi-modal contextual attention network is proposed to model the multi-modal context for each news posts, where data from different modalities can complement each other to provide a better understanding of the multi-modal data. Moreover, the hierarchical semantics of the text model can be utilized by the hierarchical encoding model to capture rich linguistic information.
- We experimentally demonstrate that the proposed HMCAN is more robust and effective than state-of-the-art baselines based on three public benchmark datasets for fake news detection tasks.

2 RELATED WORK

2.1 Fake News Detection

In most of the existing studies on fake news detection, researchers regard the fake news detection as a kind of binary classification task. Early methods [2, 20] design plenty of hand-crafted features to debunk fake news. Here, most of these methods use text content features to train a fake news classifier. Although these manually selected features improve the performance of fake news detection, these approaches typically require extensive preprocessing and feature engineering. With the rapid growth of social media content on the Internet, recognizing and detecting fake news has become increasingly challenging. Researchers have proposed many different and effective methods [2, 12, 16, 27, 34], which can be summarily reviewed from two perspectives: single-modal (e.g., text or images) fake news detection and multi-modal fake news detection.

In single-modal analysis, existing methods [2, 8, 16, 27] mainly focus on extracting textual features or visual features from the text content or image information of the posts. For example, Yu et al. [40] obtain high-level interactions and key features of related posts by convolutional neural networks. Ma et al. [19] learn latent features from the relevant textual posts by recurrent neural networks. In [23], the authors only exploit the rich visual information with different pixel domains, and utilize a multi-domain visual neural network to detect fake news. However, the social media platforms contain rich multimodal information (e.g., images, texts, and videos) [24, 25], which can complement each other and contribute to social media analysis [17, 36].

With deep neural networks having yielded immense success in learning image and textual representations, researchers realize that multi-modal fusion features play a very important role in detecting fake news. Recently, fake news detection with multi-modality has received considerable attentions. Several approaches [12, 13, 28, 29, 35, 41] conduct fake news detection based on the multimedia content and obtain superior performance. Jin et al. [12] propose a multi-modality based fake news detection model, which extracts the multi-modality information including visual, textual and social context features, and then fuses them by attention mechanism. Khattar et al. [?] propose a multimodal variational autoencoder that learns a shared representation of both the modalities,text as well as images. Shivangi et al. [29] make use of the pre-trained BERT to learn text features, and apply VGG-19 pre-trained on ImageNet dataset to learn image features. Wang et al. [35] propose a novel knowledge-driven multimodal graph convolutional network to model the textual information, knowledge concepts and visual information jointly into a unified framework for fake news detection.

Although most existing approaches show promising performance on fake news detection tasks, they are still insufficient to take advantage of the multi-modal context information and the hierarchical semantics of text content. In this paper, we propose a novel hierarchical multi-modal contextual attention network by jointly modeling of the multi-modal context information and the hierarchical semantics of text in a unified deep model for fake news detection.

2.2 Attention Mechanism

Attention mechanisms have been shown to be effective in various tasks such as image captioning [37] , machine translation [1] and recommendation system [3, 33]. Bahdanau et al. [1] firstly introduce attention in the machine translation task to allow the model to automatically search for parts of a source sentence that are relevant to predicting a target word. Soon after, a Transformer [32] model is proposed to solve the problem of sequence to sequence, replacing LSTM with an attention structure, that achieves a state-of-the-art quality score on the neural machine translation task. Recently, attention mechanisms have been incorporated into fake news detection. For example, Chen et al. [4] propose a deep attention model on the basis of recurrent neural networks (RNN) to learn selectively temporal hidden representations of sequential posts for identifying fake news. Motivated by the successful applications of attention mechanism, we introduce a hierarchical multi-modal contextual attention network to compute the intra-modality relationship and inter-modality relationship of image regions and text words.

3 PROBLEM STATEMENT

Fake news detection task can be defined as a binary classification problem, which focuses on whether posts on social media are fake news or not. Given a multi-modal post P from social media consisting of text messages and corresponding images, the model will output $Y = \{0, 1\}$ to indicate to the label of the post, where $Y = 0$ and $Y = 1$ denote that the post is real news and fake news, respectively.

4 METHOD

4.1 Overall Framework

We introduce a novel hierarchical multi-modal contextual attention network (HMCAN) to improve the performance of fake news detection task. By exploiting a multi-modal contextual attention network for multi-modal feature fusion and a hierarchical encoding network for text, our model can capture the intra-modality and inter-modality relationship and the hierarchical semantics of

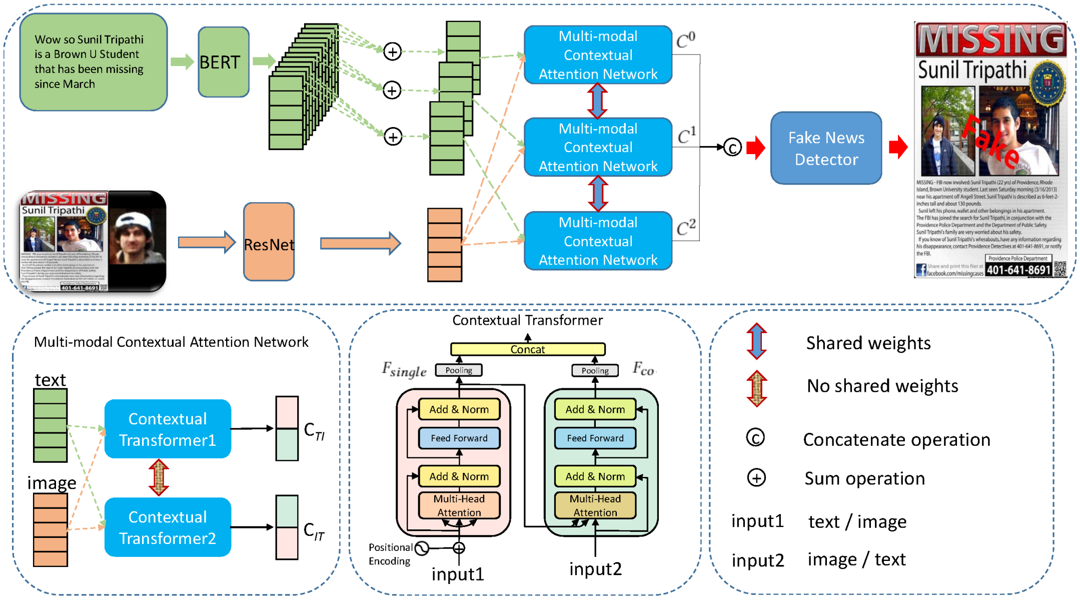


Figure 2: Illustration of the Hierarchical Multi-modal Contextual Attention Network (HMCAN) architecture. The upper part is the main framework of HMCAN. The bottom part is the detailed structure of the Multi-modal Contextual Attention Network and the Contextual Transformer module. We utilize the pre-trained model BERT to generate the embedding vector of words and the pre-trained model ResNet50 to extract region features of image. Then a multi-modal contextual attention network is used to explore the multi-modal context information. Different multi-modal contextual attention networks constitute a hierarchical encoding network to explore and capture the rich hierarchical semantics of multi-modal data.

textual and visual content of fake news. The overall architecture is illustrated in Figure 2. Specifically, our model consists of the following components:

- **Text and Image Encoding Network**: In order to precisely model both the semantic of the word and the linguistic contexts, we employ BERT [7] to generate the embedding vector of words. For each image, we utilize the pre-trained model ResNet50 [9] to extract region features. Here, the pre-trained model is fixed during training.
- **Multi-modal Contextual Attention Network**: As different modalities can complement each other, we propose a multi-modal contextual attention network to explore the multi-modal context information, which can effectively extract high-order complementary information from the multi-modal context.
- **Hierarchical Encoding Network**: In order to make better use of the hierarchical semantics of text information, we propose a novel hierarchical encoding network to explore and capture the rich hierarchical semantics of multi-modal data.
- **Fake News Detector**: Fake news detector aims to classify each post as fake or not. The detector exploits a full-connected layer with the corresponding activation function to generate predicted probability to determine whether the post is a fake news or not.

4.2 Text and Image Encoding Network

As mentioned in the problem statement, the input of our model is a multi-modal post $P = \{W, R\}$, where W and R denote the text content and visual content, respectively.

Text Encoding Network: To precisely model both the semantic of the word and the linguistic contexts, we employ Bidirectional Encoder Representations from Transformers [7] (BERT) as the core module of our textual language model. BERT has been proven to be effective in many fields, such as question answering, translation, reading comprehension and text classification [7, 26, 31].

Given a text content W, we model W as a sequence of words $W = \{w_1, w_2, \cdots, w_m\}$ (m denote the number of words in the text), we denote the transformed feature as $S = \{s_1, \cdots, s_m\}$, with s_i corresponding to the transformed feature of w_i. The word representation s_i is calculated by pre-trained BERT [7]:

$$\mathrm{S} = \{s_1, \cdots, s_m\} = \mathrm{BERT}(\mathbf{W}) \tag{1}$$

where $\mathbf{s}_i \in \mathbb{R}^{d_w}$ is the output layer hidden-state of corresponding token in BERT, and d_w is the dimension of the word embedding.

Image Encoding Network: Given a visual content R, we use the bottom-up attention pre-trained model ResNet50 [9] to extract region features. The output is a set of region features $O = \{o_1, \cdots, o_n\}$ (n denotes the number of regions in the image), where each o_i is defined as the mean-pooled convolutional feature for the i-th region. The pretrained model is fixed during training. In other words, given the attached visual content R, the operation of the penultimate pooling layer in the visual feature extractor can be represented as:

$$\mathbf{O} = \{o_1, \cdots, o_n\} = \text{ResNet50}(R) \tag{2}$$

where $\mathbf{o}_i \in \mathbb{R}^{d_r}$ and d_r is the dimension of the image embedding.

4.3 Multi-modal Contextual Attention Network

To effectively fuse the textual and visual features of posts, we design a multi-modal contextual attention network to build the multi-modal context information and extract high-order complementary information from it. As shown in the Figure 2, the multi-modal contextual attention network consists of two contextual transformer units (Contextual Transformer1 and Contextual Transformer2), where each context transformer unit focuses on different context information for multi-modal representation learning.

The detail of the contextual transformer is illustrated in the middle bottom part of Figure 2. Each contextual transformer consists of two transformer units that take data from different modalities as input (*input1* and *input2*). Taking Contextual Transformer1 as an example, the *input1* and *input2* are text and image, respectively. First, a self-attention network F_{single} (the left part) is utilized to learn the representation of text (*input1*). The self-attention network computes a intra-modality affinity matrix A_s for text as follows:

$$A_s = softmax\left(\frac{FC_s^Q(input1) \cdot FC_s^K(input1)^\top}{\sqrt{d}}\right) \tag{3}$$

where $softmax$ is the row-wise softmax operation, and FC_s^Q and FC_s^K are different full-connected layers. Each entry $A_s[i,j]$ denotes the importance of the j_{th} word for i_{th} word in the text content. Based on the intra-modality affinity matrix, the representation of text H_s can be learned as follows:

$$H_s^{'} = layer\_norm(input1 + A_s \cdot FC_s^V(input1)) \tag{4}$$

$$H_s = layer\_norm(H_s^{'} + FC_s^{ff}(H_s^{'})) \tag{5}$$

where FC_s is a full-connected layer and $layer\_norm$ is the layer normalization. FC_s^{ff} is a two-layer full-connected network that introduces non-linear transformation into the model.

The representation of text H_s is learned independently without considering the multi-modal context. Therefore, we introduce a inter-modality attention network F_{co} (the right part) to further update H_s with the visual information (*input2*) as context. The core idea is to extract information that is relevant to the image from the learned text representation, which can complement the visual information. Thus, different from F_{single}, F_{co} computes a inter-modality affinity matrix A_{co} instead of a intra-modality affinity matrix:

$$A_{co} = softmax\left(\frac{FC_{co}^Q(input2) \cdot FC_{co}^K(H_s)^\top}{\sqrt{d}}\right) \tag{6}$$

where the entry $A_{co}[i,j]$ reflects the importance of the j_{th} word in the text content for the i_{th} block in the image. Then, F_{co} learns the multi-modal context-aware text representation with the inter-modality affinity matrix A_{co} as follows:

$$H_{co}^{'} = layer\_norm(input2 + A_{co} \cdot FC_{co}^V(H_s)) \tag{7}$$

$$H_{co} = layer\_norm(H_{co}^{'} + FC_{co}^{ff}(H_{co}^{'})) \tag{8}$$

Finally, H_s and H_{co} are pooled into two feature vectors, which are then concatenated into a feature vector (C_{TI}/C_{IT}) as the multi-modal contextual representation of the text. Similar to Contextual Transformer1, Contextual Transformer 2 takes the image and text as input1 and input2 respectively, and learns the multi-modal contextual representation of the image.

Note that, the Contextual Transformer1 and Contextual Transformer2 do not share weights. For multi-modal contextual attention network, we let the output of Contextual Transformer1 be C_{TI} and the output of Contextual Transformer2 be C_{IT}. Then, we make the output of the multi-modal contextual attention network $C = \alpha C_{TI} + \beta C_{IT}$, where $\alpha + \beta = 1$.

4.4 Hierarchical Encoding Network

BERT can provide hierarchical semantics for text [15, 30], which consists of the outputs of 11 intermediate layers and 1 output layers. Intuitively, to take advantage of the rich semantics in the intermediate layers, we can apply the contextual transformer on each of the 12 layer outputs. However, this will increase the computational complexity of the model. To address this problem, we first reduce the 12 layer outputs into g group outputs by integrating every $12/g$ adjacent layers of BERT [7]. Specifically, we let $g = 3$, and we add up the output of every 4 adjacent layers of BERT.

$$\mathbf{s_i^0} = \sum_{j=1}^{4} f_B(W)_{j,i},\ \mathbf{s_i^1} = \sum_{j=5}^{8} f_B(W)_{j,i},\ \mathbf{s_i^2} = \sum_{j=9}^{12} f_B(W)_{j,i} \tag{9}$$

where $f_B(W)_{j,i} \in \mathbb{R}^{d_w}$ denotes representation of the j-th layer BERT for the i-th word in text W, and $s_i^k \in \mathbb{R}^{d_w}$ denotes the initial representation of the k-th group of the i-th word. The d_w is the dimension of the word embedding.

So that, we design a hierarchical encoding network to explore the hierarchical semantic information. Through different multi-modal contextual attention network units, we will get different C values, we denote them as C^0, C^1, C^2 respectively. Here, the number of group $g = 3$.

Finally, we concatenate the output of the three units:

$$C = concat(C^0, C^1, C^2) \tag{10}$$

where $concat$ denotes the concatenate operation, C is the final output feature vector of the multi-modal post $P = \{W, R\}$ by the proposed hierarchical multi-modal contextual attention network.

4.5 Fake News Detector

In this subsection, we introduce the fake news detector. Fake news detector takes the multi-modal representation C as input and aims at classifying the post as a fake news or not. It deploys a fully connected layer with the corresponding activation function to predict whether the post is fake or real.

$$\hat{P_n} = \sigma(W_f C_n + b) \tag{11}$$

where $\sigma(.)$ is softmax activation function, $\hat{P_n}$ denotes the predicted probabilities of the n-th post and C_n is the feature representation of the n-th post. We use Y_n to represent the ground-truth labels of

n-th post and employ cross entropy to calculate the detection loss:

$$\mathcal{L}(\Theta) = \sum_{n=1}^{N} -[\mathrm{Y}_n \log(\hat{P}_n) + (1 - Y_n) \log(1 - \hat{P}_n)] \qquad (12)$$

where N is the number of posts.

5 EXPERIMENTS AND RESULTS

5.1 Experimental Setup

5.1.1 Datasets. We compare the proposed HMCAN with state-of-the-art baselines on three public datasets: WEIBO [12], TWITTER [12, 13], and PHEME [42]. The WEIBO dataset is collected from XinHua News Agency[3] and Weibo[4], where each post contains three elements (i.e., tweet id, text and image). The TWITTER dataset [13] consists of tweets that contain textual information, visual information and social context information associated with it. The PHEME dataset [42] is collected based on 5 breaking news, each containing a set of posts. Each dataset includes a large number of texts and images with labels. Following the [13], the WEIBO and PHEME dataset are divided into training and testing set with a ratio of 8:2, and the TWITTER dataset is utilized with the development set for training and the test set for testing. Table 1 shows the statistics of the three datasets.

Table 1: The statistics of three real-world datasets.

News	WEIBO	TWITTER	PHEME
# of Fake News	4749	7898	1972
# of Real News	4779	6026	3830
# of Images	9528	514	3670

5.1.2 Evaluation Metrics. We commonly use the Accuracy as the evaluation metric for binary classification tasks such as fake news detection. However, when a dataset suffers from class imbalance, its reliability is greatly compromised. Therefore, in our experiment, in addition to the Accuracy metric, Precision, Recall and F_1 score are added as complementary evaluation metrics for the task.

5.1.3 Implementation Details. For multi-modal embeddings of posts, we use the pre-trained BERT [7] to extract textual features and the pre-trained ResNet50 [9] to extract visual features. Here, the post embedding size d is 768, the dimension of text embedding is 768, and the dimension of visual embedding is 2048. We add a 2D-convolutional layer to transform 2048 (the dimension of the region features) to 768 to fit our task. Our algorithms are implemented on Pytorch deep learning framework [10, 22] and are trained with Adaptive Moment Estimation (Adam) [14] optimizer. The model is trained for 150 epochs with a learning rate of 0.001 and the mini-batch size is set to 256. Note that, for posts without images attached, we generate dummy images for data alignment.

5.2 Baselines

We compare our model with two types of baselines: single-modal and multi-modal models.

5.2.1 Single-modal Models. For the proposed multi-modal approach, we first compare it with four single-modal models described below.

- *SVM-TS* [20]: SVM-TS is a method for detecting fake news, which utilizes heuristic rules and a linear SVM classifier.
- *CNN* [40]: CNN employs a convolutional neural network to learn the feature representations for misinformation identification and early detection tasks by framing relative posts into the fixed-length sequence.
- *GRU* [19]: GRU is based on recurrent neural networks (RNN) for learning the hidden representations that can use the multilayer GRU network to consider the post as a variable-length time series.
- *TextGCN* [39]: TextGCN employs the graph convolutional network to better learn words and document embeddings, in which the whole corpus is modeled as a heterogeneous graph.

5.2.2 Multi-modal Models. Multi-modal models normally utilize information from both textual and visual data for fake news detection. Here, we also compare our method with six multi-modal approaches described below.

- *EANN* [34]: EANN can derive event-invariant features and thus assist in the detection of fake news on newly arrived events, which consists of the multimodal feature extractor, the fake news detection, and the post discriminator. For fairness of the comparison, we conduct experiments with a simplified version of EANN that excludes the post discriminator.
- *att_RNN* [12]: att_RNN is a novel Recurrent Neural Network with an attention mechanism to fuse multimodal features for effective rumor detection. To make a fair comparison, in the experiments, we remove the component that processes social context information.
- *MVAE* [13]: MVAE uses a bimodal variational autoencoder coupled with a binary classifier for fake news detection.
- *SpotFake* [29]: SpotFake utilizes the pre-trained language models (like BERT) to learn the textual information, and employs VGG-19 (pre-trained on ImageNet [6] dataset) to obtain image features.
- *SpotFake+* [28]: SpotFake+ is an advanced version of SpotFake that extracts the textual feature using a pre-trained XLNet [38] model.
- *SAFE* [41]: SAFE extracts multi-modal (textual and visual) features of news content as well as their relationships through a similarity-aware multi-modal method for fake news detection.

5.3 Results and Analysis

Table 2 displays the experimental results of our proposed HMCAN and all baseline approaches. From Table 2, we can have the following observations:

(1) In the three datasets, SVM-TS performes the worst among all models, indicating that the hand-crafted features are weak and insufficient to identify fake news.

(2) The deep learning models (CNN, GRU) have better performance than SVM-TS, indicating their superior advantages

[3] http://www.xinhuanet.com/

[4] https://weibo.com/

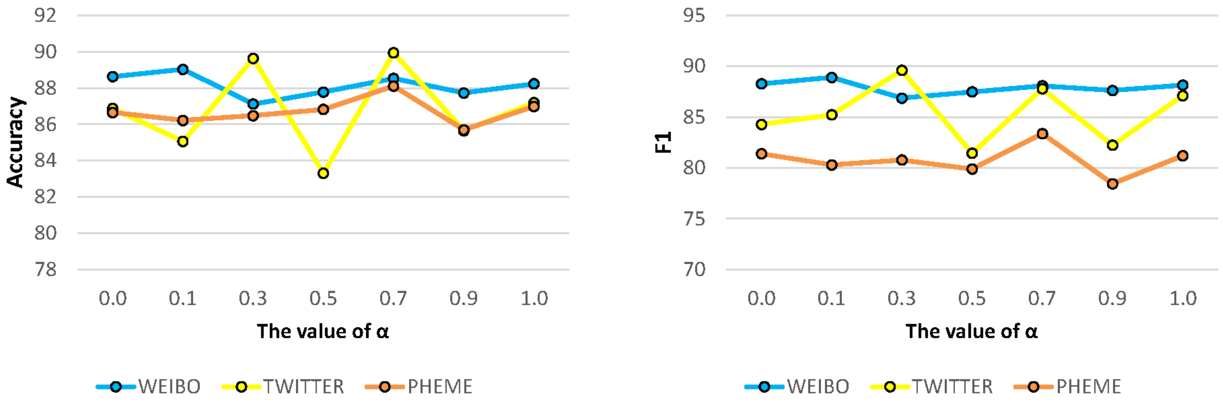
Table 2: Results of comparison among different models on WEIBO, TWITTER and PHEME Datasets. (* means the results are from the baseline paper.)

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
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	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
	TextGCN	0.787	0.975	0.573	0.727	0.712	0.985	0.827
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.869	0.877	0.859	0.868	0.861	0.879	0.870
	SpotFake*	**0.892**	0.902	0.964	**0.932**	0.847	0.656	0.739
	SpotFake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	**0.890**
TWITTER	SVM-TS	0.529	0.488	0.497	0.496	0.565	0.556	0.561
	GRU	0.634	0.581	0.812	0.677	0.758	0.502	0.604
	CNN	0.549	0.508	0.597	0.549	0.598	0.509	0.550
	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	att_RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	TextGCN	0.703	0.808	0.365	0.503	0.680	0.939	0.779
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.771	0.784	0.744	0.764	0.769	0.807	0.787
	SpotFake*	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	SpotFake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
	HMCAN	**0.897**	0.971	0.801	**0.878**	0.853	0.979	**0.912**
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
	SAFE	0.811	0.827	0.559	0.667	0.806	0.940	0.866
	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
	TextGCN	0.828	0.775	0.735	0.737	0.827	0.828	0.828
	MVAE	0.852	0.806	0.719	0.760	0.871	0.917	0.893
	SpotFake	0.823	0.743	0.745	0.744	0.864	0.863	0.863
	SpotFake+	0.800	0.730	0.668	0.697	0.832	0.869	0.850
	HMCAN	**0.881**	0.830	0.838	**0.834**	0.910	0.905	**0.907**

Table 3: Results of comparison among different variants in HMCAN on WEIBO, TWITTER and PHEME Dataset.

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	HMCAN¬V	0.809	0.832	0.774	0.802	0.788	0.843	0.815
	HMCAN¬C	0.872	0.902	0.836	0.868	0.847	0.909	0.877
	HMCAN¬H	0.877	0.871	0.885	0.878	0.883	0.869	0.876
	HMCAN	**0.885**	0.920	0.845	**0.881**	0.856	0.926	**0.890**
TWITTER	HMCAN¬V	0.755	0.828	0.590	0.689	0.719	0.896	0.798
	HMCAN¬C	0.790	0.886	0.622	0.731	0.743	0.932	0.827
	HMCAN¬H	0.879	0.884	0.849	0.866	0.875	0.906	0.890
	HMCAN	**0.897**	0.971	0.801	**0.878**	0.853	0.979	**0.912**
PHEME	HMCAN¬V	0.854	0.814	0.763	0.788	0.873	0.904	0.888
	HMCAN¬C	0.858	0.788	0.821	0.804	0.899	0.878	0.888
	HMCAN¬H	0.871	0.808	0.828	0.818	0.906	0.894	0.900
	HMCAN	**0.881**	0.830	0.838	**0.834**	0.910	0.905	**0.907**

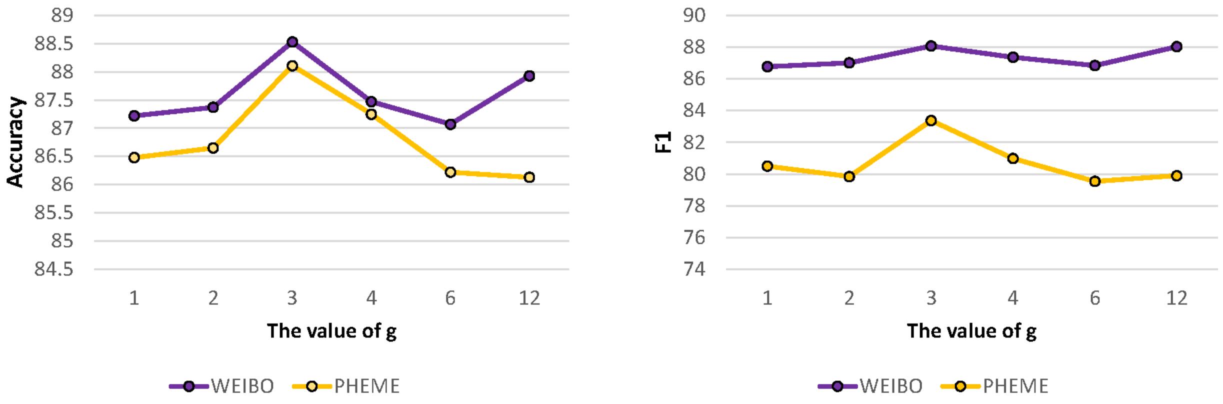
over traditional methods. However, in TWITTER dataset, CNN only outperforms SVM-TS and this is probably because it fails to capture long-range semantic relationships



(a) Accuracy

(b) F1 score of fake news

Figure 3: Impact of the value of α for the accuracy and F1 score of fake news on three datasets.



(a) Accuracy

(b) F1 score of fake news

Figure 4: Impact of the number of group g for the accuracy and F1 score of fake news on three datasets.

between words, which is vitally important for detecting fake news. In addition, the performance of TextGCN is better than that of CNN and SVM-TS on the three datasets. The results show that the graph structure can effectively capture word co-occurrence and document-word relationships through flexible graph convolutional networks.

(3) As multi-modal models, att_RNN has superior performance than GRU, showing the effectiveness of the attention mechanism, which takes into account the text related parts of the image, thus improving the performance of the model. In addition, the MVAE model has better performance than single-modal models, which indicates that additional visual information can be used as complementary information to help detect fake news.

(4) SAFE outperforms CNN on the three datasets, because SAFE jointly uses multi-modal (text and visual) and relational information to learn the representation of posts. In addition, SpotFake and SpotFake+ achieve better results on all baselines on TWITTER and WEIBO datasets, indicating that the pre-trained BERT and XLNet can obtain better textual information to improve model performance.

(5) The proposed HMCAN outperforms all the baselines on TWITTER and PHEME datasets. We can observe that on WEIBO dataset, in the case of fake news, the F1 and accuracy of HMCAN are lower than SpotFake*, while in the case of real news, the F1 of HMCAN is higher. Here, the results of SpotFake* are from the baseline paper and the results of SpotFake are the results of our reproduction of the authors’

model. The results demonstrate that the proposed model can jointly model multi-modal context information and hierarchical semantics of text in a unified deep model, which can better capture the underlying representation of posts, so as to improve the performance of fake news detection.

5.4 Analysis of HMCAN Components

Because the proposed HMCAN contains multiple key components, we additionally compare variants of HMCAN with respect to the following perspectives to demonstrate the effectiveness of HMCAN — (1) the effect of the visual information, (2) the impact of the multi-modal contextual attention network, and (3) the effect of the hierarchical semantics of text information (V, C, H) for the hierarchical multi-modal contextual attention network module. The following HMCAN variants are designed for comparison.

- HMCAN¬V: A variant of HMCAN with the visual information being removed and only use text information.
- HMCAN¬C: A variant of HMCAN with the multi-modal contextual attention network being removed.
- HMCAN¬H: A variant of HMCAN with the hierarchical information of words being removed, and only using the last state of the output of the BERT model.

The ablation study results are shown in Table 3.

5.4.1 Effects of the visual information. We compare the performance of HMCAN with HMCAN¬V on the three datasets (WEIBO, TWITTER and PHEME) to investigate the effectiveness of the visual information. From the result, we can observe that the proposed HMCAN outperforms HMCAN¬V, which shows that the visual information can consistently provide supplementary information to benefit our model.

5.4.2 Effects of the multi-modal contextual attention network. We compare the performance of HMCAN to HMCAN¬C on the three datasets (WEIBO, TWITTER and PHEME) to investigate the effectiveness of the multi-modal contextual attention network component. According to the results, we observe that the proposed HMCAN performs better than HMCAN¬C, which can confirm the superiority of introducing the multi-modal contextual attention network to our model.

5.4.3 Effects of the hierarchical semantics of text informationn. We compare the performance of HMCAN with HMCAN¬H on the three datasets (WEIBO, TWITTER and PHEME) to show the effectiveness of the hierarchical semantics of text information. From the results, we can observe that the proposed HMCAN outperforms HMCAN¬H, which shows that the effectiveness of introducing the hierarchical semantics of text information to our model.

5.5 Impact of the value of α

The output of each multi-modal contextual attention network unit is $C^g = \alpha C_{TI} + \beta C_{IT}$, where $\alpha + \beta = 1$. To find a suitable α value, we vary α form 0.0 to 1.0, and test the impact for the accuracy and F1 score of fake news detection on the three datasets , respectively. The results are shown in Figure 3(a) and Figure 3(b). When α grows from 0.0 to 1.0, the accuracy and the F1 score of our model will keep changing. In Figure 3(a), when the value of α is 0.7, the accuracy has the highest results on TWITTER and PHEME datasets. While the accuracy in case of α is 0.7 is only slightly lower than that in case of α is 0.1 on WEIBO dataset. In Figure 3(b), on WEIBO dataset when α is 0.1, the F1 score is the highest, on TWITTER dataset when α is 0.3 the F1 score is the highest, and on PHEME dataset when α is 0.7 the F1 score is the highest. On the whole, when α is 0.7, the F1 score on the three datasets can achieve satisfactory results. Therefore, we set $\alpha = 0.7$ on the three datasets and HMCAN can achieve relatively good performance.

5.6 Impact of the number of group g

We vary g form 1 to 12, and report the results on WEIBO and PHEME datasets in Figure 4. We can observe that the accuracy and F1 score of fake news show a clear increase from 1 to 3, while it decreases slightly from 3 to 6. When g is set to 12, the F1 score of fake news on both two datasets and the accuracy on WEIBO dataset will decrease slightly, and are also lower than $g = 3$. Besides, $g = 12$ means that we divide the output layer of BERT into 12 groups, which will lead to a large amount of computational cost. Therefore, we make $g = 3$ (ie. we let the output of every 4 adjacent layers of BERT as a group) on the three datasets in our model.

6 CONCLUSIONS

In this paper, we propose a novel hierarchical multi-modal contextual attention network (HMCAN) for fake news detection task. We argue that most existing methods are difficult to utilize the multi-modal context information and extract high-order complementary information for each news. Additionally, existing approaches mainly ignore the full hierarchical semantics of textual content which can help learn a better news representation to enhance the detection of fake news. To tackle the above challenges, HMCAN is proposed to jointly model the multi-modal context information and the hierarchical semantics of text in a unified deep model. Our approach consists of three technical innovations: (1) We employ ResNet to learn better representations of images and utilize BERT to embed the textual content of news. (2) A multi-modal contextual attention network is proposed to fuse both inter-modality and intra-modality relationships. (3) We design a hierarchical encoding network to capture the rich hierarchical semantics for fake news detection. Experiments and comparisons demonstrate that the proposed HMCAN performs more effective and robust than state-of-the-art baselines on three public datasets for fake news detection. In the future, we will explore a more effective way to extract visual information or utilize additional knowledge information, which can provide useful complementary information for fake news detection.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China (No. 2017YFB1002804), National Natural Science Foundation of China (No. 62036012, 61721004, 61720106006, 61802405, 62072456, 61832002, 61936005, U1836220 and U1705262), the Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJSSWJSC039, and the K.C.Wong Education Foundation.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 1–15.

[2] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. 675–684.

[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 335–344.

[4] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection. In *Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2018 Workshops, 2018*. 40–52.

[5] Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. Same: sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*. 41–48.

[6] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, and Fei Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*. 4171–4186.

[8] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016*. 770–778.

[10] Jun Hu, Shengsheng Qian, Quan Fang, Youze Wang, Quan Zhao, Huaiwen Zhang, and Changsheng Xu. 2021. Efficient Graph Deep Learning in TensorFlow with tf_geometric. *CoRR* abs/2101.11552 (2021).

[11] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language?. In *ACL 2019*. 3651–3657.

[12] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 795–816.

[13] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*. 2915–2921.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. 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*. 4364–4373.

[16] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*. IEEE, 1103–1108.

[17] Song Liu, Shengsheng Qian, Yang Guan, Jiawei Zhan, and Long Ying. 2020. Joint-modal Distribution-based Similarity Hashing for Large-scale Unsupervised Deep Cross-modal Retrieval. In *SIGIR 2020, Virtual Event, China, July 25-30*. 1379–1388.

[18] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 1867–1870.

[19] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.. In *Ijcai*. 3818–3824.

[20] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. 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*. ACM, 1751–1754.

[21] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*. 3049–3055.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[23] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. *arXiv preprint arXiv:1908.04472* (2019).

[24] Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. 2016. Multi-modal Multi-view Topic-opinion Mining for Social Event Analysis. In *ACM MM 2016*. 2–11.

[25] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and Jie Shao. 2016. Multi-modal event topic model for social event analysis. *IEEE transactions on multimedia* 18, 2 (2016), 233–246.

[26] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019).

[27] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[28] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2020. SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning (Student Abstract). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.

[29] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'Ichi Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*.

[30] Youwei Song, Jiahai Wang, Zhiwei Liang, Zhiyue Liu, and Tao Jiang. 2020. Utilizing BERT Intermediate Layers for Aspect Based Sentiment Analysis and Natural Language Inference. *CoRR* abs/2002.04815 (2020).

[31] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. 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*, Vol. 11856. Springer, 194–206.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 5998–6008.

[33] Shoujin Wang, Liang Hu, Longbing Cao, Xiaoshui Huang, Defu Lian, and Wei Liu. 2018. Attention-Based Transactional Context Embedding for Next-Item Recommendation. In *AAAI 2018*. 2532–2539.

[34] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. ACM, 849–857.

[35] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake News Detection via Knowledge-driven Multimodal Graph Convolutional Networks. In *Proceedings of the 2020 on International Conference on Multimedia Retrieval, 2020*. 540–547.

[36] Xiao Wu, Chong-Wah Ngo, and Alexander G. Hauptmann. 2008. Multimodal News Story Clustering With Pairwise Visual Near-Duplicate Constraint. *IEEE Transactions on Multimedia* 10, 2 (2008), 188–199.

[37] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 2048–2057.

[38] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS 2019*. 5754–5764.

[39] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7370–7377.

[40] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. 2017. A Convolutional Approach for Misinformation Identification.. In *IJCAI*. 3901–3907.

[41] Xinyi Zhou, Jindi Wu, and R. Zafarani. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. *ArXiv* abs/2003.04981 (2020).

[42] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *International Conference on Social Informatics*. Springer, 109–123.