IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 16, 2021, accepted August 24, 2021, date of publication September 20, 2021, date of current version October 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3113981

Fake News Detection via Multi-Modal Topic Memory Network

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

[1]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China
[3]School of Information Science and Engineering, China University of Petroleum, Beijing 102249, China
[4]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Corresponding authors: Shengsheng Qian (shengsheng.qian@nlpr.ia.ac.cn) and Jinguang Wang (wangjinguang502@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61902193, and in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

ABSTRACT With the development of the Mobile Internet, more and more people create and release multi-modal posts on social media platforms. Fake news detection has become an increasingly challenging task. Although many current works focus on constructing models extracting abstract features from the content of each post, they neglect the intrinsic semantic architecture such as latent topics, etc. These models only learn patterns in content coupled with certain specific latent topics on the training set to distinguish real and fake posts, which will suffer generalization and discriminating ability decline, especially when posts are associated with rare or new topics. Moreover, most existing works using deep schemes to extract and integrate textual and visual representation in post have not effectively modeled and sufficiently utilized the complementary and noisy multi-modal information containing semantic concepts and entities to complement and enhance each modal. In this paper, to deal with the above problems, we propose a novel end-to-end Multi-modal Topic Memory Network (MTMN), which obtains and combines post representations shared across latent topics together with global features of latent topics while modeling intra-modality and inter-modality information in a unified framework. (1) To tackle real scenarios where newly arriving posts with different topic distribution from the training data, our method incorporates a topic memory module to explicitly characterize final representation as post feature shared across topics and global features of latent topics. These two kinds of features are jointly learned and then combined to generate robust representation. (2) To effectively integrate multi-modality information in posts, we propose a novel blended attention module for multi-modal fusion, which can simultaneously exploit the intra-modality relation within each modal and the inter-modality relation between text words and image regions to complement and enhance each other for high-quality representation. Extensive experiments on two public real-world datasets demonstrate the superior performance of MTMN compared with other state-of-the-art algorithms.

INDEX TERMS Fake news detection, multi-modal fusion, topic memory network, blended attention module.

I. INTRODUCTION

Social media has become more and more extensive and has been deeply integrated into our daily life, for the rapid development of Mobile Internet and Communication technologies. With easy accessibility and manipulation, people tend to acquire and share information as well as express and exchange opinions through mobile social media platforms. Unfortunately, due to the openness of social media, a large number of users, and the complexity of sources, various fake news have been fostered. These widespread fake news are utilized by some evil guys to mislead the public, which could do serious harm to society and may cause great economic loss. Ordinary users do not have the time and capability to verify the authenticity of each piece of information. Therefore, it is necessary and urgent to detect fake news on social media [1]–[3] and ensure users receive truthful information.

Nowadays, various fake news detection approaches [1]–[8] have been proposed, including traditional learning and deep learning-based methods. Traditional approaches

The associate editor coordinating the review of this manuscript and approving it for publication was Fahmi Khalifa.

such as Support Vector Machine (SVM) [4], Decision Tree [5], and Random Forest heavily depend on hand-craft features to identify fake news, which is time-consuming and labor-intensive. With the great success of the neural network, existing deep learning models have achieved performance improvement over traditional ones due to their superior ability of feature extraction. Some early studies tried to extract features from the plain text content of news to detect fake news. Then it further explored the use of recurrent neural network (RNN) and its variants [6], such as Gated Recurrent Unit (GRU), to extract sequence language features for fake news detection. Some researchers also introduce the convolutional neural networks (CNN) [7] to learn the high-level representations extracted from posts on social media to identify fake news. Moreover, graph convolutional network (GCN) is employed to learn words and document embeddings [8] which models the whole corpus as a heterogeneous graph.

FIGURE 1. An example of multi-modal post in social media. The upper part of the post is its text content, and the lower part is its attached images.

To date, social media posts content evolves from pure text content to multi-modal content with text, images, and videos. An example of multi-modal post is illustrated in Figure 1. Fake news detection with multi-modality [9]–[11] has received more and more concerns. Many recent works [12]–[15] utilize deep schemes to extract and combine textual and visual representation in posts.

As shown in most scenarios, posts on social media platforms are the elemental objects for fake news detection. Most current works focus on constructing models extracting abstract features from the content of every post. However, they neglect the intrinsic semantic architecture, e.g., latent topics, which can be simply considered as semantic categories. Posts associated with different latent topics (semantic categories) generally contain specific patterns to discriminate between real and fake content. For example, fake political news usually includes extreme compliments or vicious criticisms in narrow perspectives and includes images with crowds gathering or mass violence, while rumors about daily life always contain surprising and exaggerated descriptions for scenes or objects.

Therefore, post presentation model without considering latent topics only learns patterns in content coupled with certain specific latent topics in training set to distinguish real and fake posts. It will suffer discriminating and generalization ability decline, especially when posts are associated with rare topics on the training set and new topics. Therefore, we need to address **Challenge 1**: How to construct post representation model which explicitly decouples and models the effect of latent topic features to generate robust representations for testing data usual with different topic distributions.

Moreover, most recently proposed multi-modal fusion based approaches are coarse and simple when modeling semantic space. Some models [13], [14] just concatenate features extract from different modalities such as text and image or adopt additional fully connected layer to form final representation. Others [12], [15] employ the inter-modal attention mechanism to capture relations between semantic entities in different modals, mapping entities into the same modal space to complement and enhance semantic information.

However, advanced approaches should more elaborately represent and model the entities, concepts, and relations across different modalities. Methods that effectively integrate complementary and noisy multi-modal information containing semantic concepts and entities to complement and enhance each modal have not been sufficiently researched. Therefore, we have to face **Challenge 2**: How to fully utilize the multi-modal information and construct semantic model across different modalities, extracting complementary and comprehensive information to improve the performance of fake news detector?

In order to solve the above challenges, we propose a novel end-to-end *Multi-modal Topic Memory Network* (MTMN), which obtains and combines post representation shared across topics together with global features of latent topics while modeling intra-modality and inter-modality information in a unified framework. (1) For **Challenge 1**, our method incorporates a topic memory network to explicitly characterize final representation for fake news detection as post feature shared across topics and global features of latent topics jointly learned on training data. The stored global topic features are selectively read out by the memory controller to form relevant global topic feature, combined with the corresponding shared post feature to generate final representation. (2) To tackle **Challenge 2**, we propose a novel blended attention module based on extracted fine-grained representations for image regions and sentence words. By considering both intra-modal and inter-modal relations jointly, the features of image and sentence fragments can complement and enhance each other in semantic space.

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

- We propose a novel end-to-end *Multi-modality Topic Memory Network* (MTMN) incorporating topic memory module to explicitly characterize final representation as post feature shared across topics and global features of latent topics. These two kinds of features are jointly learned and then integrated to generate robust

representation for newly arriving posts usual with different topic distribution.

- A novel blended attention module is designed for multi-modal fusion, which is able to exploit the intra-modal relation within sentence words or image regions as well as the inter-modal relation between sentence words and image regions jointly to complement and enhance each other for high-quality multi-modal representation.
- We evaluate our method on the two public real-world datasets (WEIBO and PHEME), and experimental results demonstrate the proposed MTMN approach outperforms the state-of-the-art (SOTA) baselines.

II. RELATED WORK

A. FAKE NEWS DETECTION

With the massive growth of social media content on the Internet, how to recognize and detect fake news becomes more and more challenging. Researchers have been working on fake news detection and propose many different methods [1]–[3], which can be roughly reviewed from two aspects: single-modal (*e.g.*, text or images) fake news detection and multi-modal fake news detection.

In the single-modality analysis, existing methods [4]–[8], [16], [17] mainly extract the textual features or visual features from the text content or image information of posts, which have been explored in various fake news detection works. For example, SVM-TS [4] utilizes heuristic rules and a linear SVM [18] to classify rumors on Twitter while employing a time-series structure to model the social feature variations. Kwon *et al.* [5] adopt the decision-tree to classify the post by learning the topic-based features based on the text content. Ma *et al.* [6] learn hidden representations from the text content of relevant posts by recurrent neural networks. Yu *et al.* [7] obtain high-level interactions and key features of relevant posts by convolutional neural networks. Yao *et al.* [8] employ graph convolutional network (GCN) to learn words and document embeddings which models the whole corpus as a heterogeneous graph. In the paper [16], the authors only exploit the rich visual information with different pixel domains and adopt a novel multi-domain visual neural network (MVNN) to identify fake news. However, social media platforms have rich multi-modal information, such as texts, images, and videos, which can enhance and complement each other and are helpful for social media analysis [19]–[22].

Recently, fake news detection with multi-modality content has received considerable attention. Some works are founded on the conventional features of the attached images in posts [10], [23]. However, these features are hand-crafted, which are difficult to effectively capture the complex distributions of textual and image content.

As deep neural networks (DNN) have achieved extraordinary performance on nonlinear representation learning [24]–[26], many multi-modal representation methods [9], [10], [12]–[15] utilize deep schemes to learn the representative features and obtain superior performance in fake news detection. Jin *et al.* [12] propose a novel Recurrent Neural Network with an attention mechanism (Att-RNN) to fuse multi-modal features for effective rumor detection. The joint features of text and social context obtained with a Long-Short Term Memory (LSTM) network are integrated with image features by neural attention producing reliable representations. Wang *et al.* [13] design an event adversarial neural network (EANN) learning event-invariant features to obtain the multi-modal features of each post for fake news detection. It eliminates event-specific components in the post representations formed by the concatenation of extracted deep textual and visual features. Khattar *et al.* [14] propose a novel multi-modal variational autoencoder (MVAE) for fake news detection, which obtains the shared multi-modal representations by the designed variational autoencoder with encoding and decoding modules for textual and visual modalities. The representation model is learned jointly with the classifier for the fake category. Zhou *et al.* [15] develop a similarity-aware fake news detection method (SAFE) which adopts neural networks to obtain the latent representation of both textual and visual content and then employs the relationship (similarity) among different modalities as the similarity feature. The similarity feature is combined with the concatenation of textual and visual features to recognize fake posts.

Although these approaches have made some breakthroughs, they are mostly simple and coarse in modeling semantic space across multi-modalities. Advanced methods which effectively integrate complementary and noisy multi-modal information to complement and enhance each modality in semantic space have not been sufficiently researched.

B. MEMORY NETWORK

Memory network [27]–[30] is a universal and powerful sequence model which incorporates an external memory bank to capture complex relation and interaction patterns among sequence elements distributed in long-range. Generally, the memory network consists of two components: an external memory bank with several slots to store object representations and a memory controller which performs operations on the memory, including reading, writing, and erasing.

Early works employ simplified versions of memory networks only with attention-based reading mechanism [31], [32] or updating mechanism similar to Long-Short Term Memory (LSTM) [33] to address question-answering in textual discourse. In recent years, memory networks are combined with more sophisticated feature extraction models for explicitly rational or temporal reasoning in many natural language processing (NLP) tasks such as sentiment analysis [34], summarization [35], and task-oriented dialog [36]. They are also utilized to integrate formalized knowledge in external knowledge bases with network structure [37]–[39]. Moreover, many researchers extend memory networks to model reasoning or incorporate external knowledge in computer vision [40], [41], cross-modal tasks such as visual question answering [42], and recommendation [43].

Note that these approaches usually focus on entity-level or sentence-level memories, while our work addresses the global semantic topic-level memory, which shares across the training data.

The event adversarial neural network (EANN) [13] approach aims to generate robust representations of posts in newly emerging events, which is similar to our purpose. Event-invariant features for posts are learned by using an adversarial network along with a multi-modal feature extraction module. However, this approach utilizes additional event information, and in many situations removing event-specific features depraves the discriminating ability of multi-modal representations of posts. We will manifest this in the experiments. (Section V-D).

III. THE PROPOSED ALGORITHM

A. PROBLEM DEFINITION

Fake news detection task can be defined as a binary classification problem, which aims to classify posts in social media into fake news or real news. In real scenarios, newly emerging posts may attach to rare or new topics. Given a set of multi-modal posts $\mathcal{O} = \{o_1, \cdots, o_{k_x}\}$, where o_n is a post consisting of textual words and corresponding visual content which mainly comprises images, k_x represents the number of the posts. Our purpose is to learn a model $f : \mathcal{O} \rightarrow \mathcal{C}$, to classify each post o_n into the predefined categories $\mathcal{C} = \{0, 1\}$ where 1 denotes fake news and 0 denotes real news.

B. OVERALL FRAMEWORK

We introduce a novel multi-modal topic memory network (MTMN) to improve the performance of fake news detection. By employing a multi-modal blended attention network for multi-modal fusion, our model can capture the intra-modal and inter-modal relation of textual and visual content. Topic memory network is incorporated to jointly learn post representations shared across topics and global features of latent topics, which are combined to generate robust representations for newly arriving posts. The overall architecture is illustrated in Figure 2. The proposed model consists of the following components:

- *Text and Image Encoding Network:* Given a multi-modal post containing text and images, we use word piece tokens of text as the fragments in the textual modality. The pre-trained BERT model [44] is employed to fetch embeddings of word piece tokens. Meanwhile, we utilize a pre-trained ResNet50 model [45] for each image in the post to extract region features. Noted that the pre-trained model is fixed during the training stage.
- *Multi-Modal Feature Representation (Section IV-A):* Based on the extracted fine-grained representations for text fragments and image regions, we adopt the Blended Attention Module to jointly model the inter-modal and intra-modal relations for image regions and text fragments. By synthetically considering relations on elements in different modalities, the features of text and image fragments can be aligned and enhanced. Then the self-attention summary layers are used to aggregate these fragment representations.
- *Topic Memory Network (Section IV-B):* Based on the representations of multi-modal posts extracted by the Multi-modal Feature Representation module, a Topic Memory Network is adopted to learn and store specific global features of latent topics, which is conducted jointly with post feature shared across topics by updating mechanism of the memory controller to write in memory slots according to batch sequence iteratively. The stored global topic features are selectively read out by the memory controller to form relevant global topic feature, combined with the corresponding shared post feature to generate final representation.
- *Fake News Detection Network (Section IV-C):* Fake news detector aims to classify each post as fake or true, which takes the learned multi-modal features as inputs and then feeds them into a fully connected network with corresponding activation function to classify whether the posts are fake or real.

IV. METHODOLOGY

This section presents the proposed multi-modal topic memory network for fake news detection.

A. MULTI-MODAL FEATURE REPRESENTATION

To effectively fuse the textual and visual features of posts, we synchronously model both the relation in the same modality and among different modalities of multi-modal content. As shown in Figure 2, the blended attention module takes the stacked features of text words and image regions as the input $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{R} \end{pmatrix} = \{z_1; \cdots; z_{k_z}; \boldsymbol{r}_1; \cdots; \boldsymbol{r}_{k_r}\}$, where $\boldsymbol{X} \in \mathbb{R}^{(k_z+k_r)\times d_x}$, $\boldsymbol{Z} = \{z_1; z_2; \cdots; z_{k_z}\}$ and $\boldsymbol{R} = \{\boldsymbol{r}_1; \boldsymbol{r}_2; \cdots; \boldsymbol{r}_{k_r}\}$ are feature tensors obtained respectively by pre-trained BERT model [44] and ResNet50 model [45] for text words and image regions, in which image region features have been adapted to the same dimension by 1-d convolutional layer.

The concatenation of post features $\boldsymbol{X}$ is fed into a Transformer unit. The query, key, and value for the fine-grained features are formed with the following equations, in which $\boldsymbol{W}^K$, $\boldsymbol{W}^Q$, and $\boldsymbol{W}^V$ stand for the parameter matrices:

$$\boldsymbol{K}_X = \boldsymbol{X}\boldsymbol{W}^K = \begin{pmatrix} \boldsymbol{Z}\boldsymbol{W}^K \\ \boldsymbol{R}\boldsymbol{W}^K \end{pmatrix} = \begin{pmatrix} \boldsymbol{K}_Z \\ \boldsymbol{K}_R \end{pmatrix} \tag{1}$$

$$\boldsymbol{Q}_X = \boldsymbol{X}\boldsymbol{W}^Q = \begin{pmatrix} \boldsymbol{Z}\boldsymbol{W}^Q \\ \boldsymbol{R}\boldsymbol{W}^Q \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q}_Z \\ \boldsymbol{Q}_R \end{pmatrix} \tag{2}$$

$$\boldsymbol{V}_X = \boldsymbol{X}\boldsymbol{W}^V = \begin{pmatrix} \boldsymbol{Z}\boldsymbol{W}^V \\ \boldsymbol{R}\boldsymbol{W}^V \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}_Z \\ \boldsymbol{V}_R \end{pmatrix} \tag{3}$$

Then, the Scaled Dot-Product Attention is carried out as defined:

$$\text{Attention}(\boldsymbol{Q}_X, \boldsymbol{K}_X, \boldsymbol{V}_X) = \text{softmax}\left(\frac{\boldsymbol{Q}_X \boldsymbol{K}_X^\top}{\sqrt{d_k}}\right)\boldsymbol{V}_X \tag{4}$$

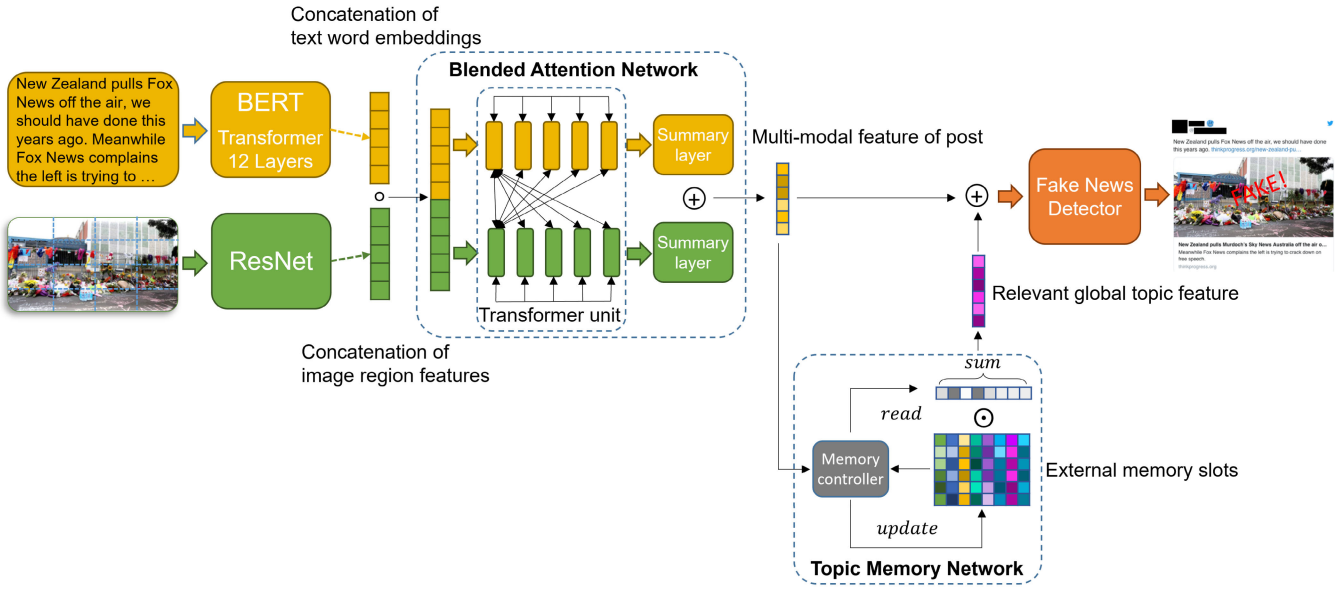


FIGURE 2. The overall framework of MTMN. The inputs consist of the textual content and the attached image of posts. Text words and image regions are then respectively encoded by the pre-trained BERT model and ResNet50. The Blended Attention Network is employed to model the inter-modal and intra-modal relations, aggregate textual and visual fragments, and incorporate different modals to finally get the multi-modal representations. Based on these representations, the Topic Memory Network jointly learns global features of latent topics with post features shared across topics and stores learned global topic features. Post features are combined with corresponding relevant global topic features generated by memory reading to form final representations.

where d_k is value of the last dimension of query and key, and $\frac{1}{\sqrt{d_k}}$ is the scaling factor.

To make the derivation understand easily, the softmax and scaled functions in the above equation are removed, which does not affect the core idea of our attention mechanism. It can be expanded as follows:

$$\begin{aligned}\boldsymbol{Q}_X\boldsymbol{K}_X^\top\boldsymbol{V}_X &= \begin{pmatrix}\boldsymbol{Q}_Z\\ \boldsymbol{Q}_R\end{pmatrix}\begin{pmatrix}\boldsymbol{K}_Z^\top & \boldsymbol{K}_R^\top\end{pmatrix}\begin{pmatrix}\boldsymbol{V}_Z\\ \boldsymbol{V}_R\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{Q}_Z\boldsymbol{K}_Z^\top & \boldsymbol{Q}_Z\boldsymbol{K}_R^\top\\ \boldsymbol{Q}_R\boldsymbol{K}_Z^\top & \boldsymbol{Q}_R\boldsymbol{K}_R^\top\end{pmatrix}\begin{pmatrix}\boldsymbol{V}_Z\\ \boldsymbol{V}_R\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{Q}_Z\boldsymbol{K}_Z^\top\boldsymbol{V}_Z + \boldsymbol{Q}_Z\boldsymbol{K}_R^\top\boldsymbol{V}_R\\ \boldsymbol{Q}_R\boldsymbol{K}_R^\top\boldsymbol{V}_R + \boldsymbol{Q}_R\boldsymbol{K}_Z^\top\boldsymbol{V}_Z\end{pmatrix}\end{aligned} \tag{5}$$

Processed by the above attention layer, the calculated features for the textual and visual fragments are depicted as follows:

$$\boldsymbol{Z}^l = \{\boldsymbol{e}_1^l;\cdots;\boldsymbol{e}_{k_e}^l\} = \boldsymbol{Q}_Z\boldsymbol{K}_Z^\top\boldsymbol{V}_Z + \boldsymbol{Q}_Z\boldsymbol{K}_R^\top\boldsymbol{V}_R. \tag{6}$$

$$\boldsymbol{R}^l = \{\boldsymbol{r}_1^l;\cdots;\boldsymbol{r}_{k_r}^l\} = \boldsymbol{Q}_R\boldsymbol{K}_R^\top\boldsymbol{V}_R + \boldsymbol{Q}_R\boldsymbol{K}_Z^\top\boldsymbol{V}_Z, \tag{7}$$

These results show that the output of the multi-head attention layer in the Transformer unit synchronously takes the inter-modal and intra-modal relation into consideration.

Then, $\begin{pmatrix}\boldsymbol{Z}^l\\ \boldsymbol{R}^l\end{pmatrix}$ is sent into the followed position-wise feed-forward sub-layer. Finally, the output of the Transformer unit is obtained and written as: $\boldsymbol{X}_a = \begin{pmatrix}\boldsymbol{Z}_a\\ \boldsymbol{R}_a\end{pmatrix}$.

$\boldsymbol{X}_a$ can be explicitly split into feature tensor of textual tokens $\boldsymbol{Z}_a = \{\boldsymbol{z}_{a1};\cdots;\boldsymbol{z}_{ak_z}\}$ and feature tensor of image regions $\boldsymbol{R}_a = \{\boldsymbol{r}_{a1};\cdots;\boldsymbol{r}_{ak_r}\}$, and fed in summary layer respectively, getting the aggregated representation $\boldsymbol{H}_z = \{\boldsymbol{h}_{z1};\cdots;\boldsymbol{h}_{zk_z}\} \in \mathbb{R}^{k_z\times d_z}$ and $\boldsymbol{H}_r = \{\boldsymbol{h}_{r1};\cdots;\boldsymbol{h}_{rk_r}\} \in \mathbb{R}^{k_r\times d_r}$ for text and image:

$$\begin{aligned}\boldsymbol{H}_z &= \text{summary}(\boldsymbol{Q}_z, \boldsymbol{V}_z) = \text{softmax}\left(\text{MLP}(\boldsymbol{Z}_a)\right)\boldsymbol{Z}_a\\ \boldsymbol{H}_r &= \text{summary}(\boldsymbol{Q}_r, \boldsymbol{V}_r) = \text{softmax}\left(\text{MLP}(\boldsymbol{R}_a)\right)\boldsymbol{R}_a\end{aligned} \tag{8}$$

where MLP$(\cdot)$ stands for multi-layer perceptron.

The final multi-modal feature representations of posts are produced by sum operation between $\boldsymbol{h}_{zi} \in \boldsymbol{H}_z$ and $\boldsymbol{h}_{ri} \in \boldsymbol{H}_r$, which can be denoted as:

$$\boldsymbol{h}_i = \lambda\boldsymbol{h}_{zi} + (1-\lambda)\boldsymbol{h}_{ri} \tag{9}$$

where λ is the tradeoff factor of the proportion of textual and visual information in the multi-modal features.

B. TOPIC MEMORY NETWORK

After getting the multi-modal representation of posts $\boldsymbol{H} := \{\boldsymbol{h}_1;\cdots;\boldsymbol{h}_{k_x}\}$, most existing fake news detection methods using decision model such as multi-layer perceptron (MLP) with softmax function, to generate final results. However, they neglect the latent semantic structure of multi-modal posts, only learning patterns in content coupled with certain specific latent topics according to training set, which cannot be transferred to rarely or newly emerging posts.

For this reason, our method incorporates topic memory module to capture global feature of each topic while jointly learn post representation shared by topics, which are integrated to generate robust representation.

The architecture of TMN consists of two parts. 1) The memory controller conducts content based addressing approach, which is similar to the multi-head attention module to assign the reading vector and writing vector. 2) The external memory bank contains several slots shared during the whole training process to capture the global latent topic information of posts.

To calculate the relevant global topic features corresponding to posts, given the query tensor $\boldsymbol{H}_B = \{\boldsymbol{h}_1; \cdots; \boldsymbol{h}_{k_b}\} \in \mathbb{R}^{k_b \times d_h}$ corresponding to one batch which is a portion of post representation of training set $\boldsymbol{H}$, the reading process is constructed as follows:

$$\boldsymbol{\Gamma} = \text{softmax}\left(\frac{\boldsymbol{Q}_m \mathbf{M}^\top}{\sqrt{d_h}}\right) = \text{softmax}\left(\frac{\boldsymbol{H}_B \mathbf{W}_Q \mathbf{M}^\top}{\sqrt{d_h}}\right) \quad (10)$$

$$\mathbf{C} = \boldsymbol{\Gamma}\mathbf{M}, \quad \mathbf{c}_i = \sum_{j=1}^{k_m} \boldsymbol{\Gamma}_{ij} \mathbf{m}_j \quad (11)$$

where k_b is batch size, d_h is the dimension of post representation, k_m is number of memory slots, $\mathbf{W}_Q \in \mathbb{R}^{d_h \times d_h}$ is projection matrix, $\mathbf{M} = \{\mathbf{m}_1; \cdots; \mathbf{m}_{k_m}\} \in \mathbb{R}^{k_m \times d_h}$ is the external memory bank containing k_m slots, $\boldsymbol{\Gamma} \in \mathbb{R}^{k_b \times k_m}$ is the attention matrix in which softmax$(\cdot)$ function is calculated along each row, $\mathbf{C} = \{\mathbf{c}_1; \cdots; \mathbf{c}_{k_b}\} \in \mathbb{R}^{k_b \times d_h}$ is relevant global topic feature tensor corresponding to posts in one batch, and $\mathbf{c}_i$ is relevant global topic feature corresponding to representation $\boldsymbol{h}_i$ of post i.

Post feature shared across topics is obtained by putting post representation through residual MLP block:

$$\mathbf{c}_i^P = (\mathbf{h}_i + \text{MLP}(\mathbf{h}_i)) \quad (12)$$

The final representation combined by post feature shared across topics and relevant global topic feature can be depicted as follows:

$$\mathbf{u}_i = (1 - \beta)\mathbf{c}_i^P + \beta \mathbf{c}_i \quad (13)$$

where β is tradeoff factor of the proportion of relevant global topic feature and post feature.

The global features of latent topics are captured and stored in memory slots by updating mechanism iteratively according to data batch sequence, jointly with learning post feature shared across topics.

For memory updating, the writing vector is also produced base on content based addressing, so the writing vector is the same as the reading vector. The updating module is constructed as follows:

$$\mathbf{M}^t = \tanh\left(\mathbf{g}^t \odot \mathbf{M}^{t-1}\mathbf{W}_m + \frac{1}{N_B}\left(\boldsymbol{\Gamma}^\top \boldsymbol{H}_B^t\right)\mathbf{W}_h\right)$$

$$\mathbf{m}_j^t = \tanh\left(\mathbf{g}^t \odot \mathbf{m}_j^t \mathbf{W}_m + \frac{1}{N_B}\sum_{i=1}^{N_B} \boldsymbol{\Gamma}_{ij} \boldsymbol{h}_i^t \mathbf{W}_h\right) \quad (14)$$

where $\mathbf{W}_m, \mathbf{W}_h \in \mathbb{R}^{d_h \times d_h}$ are linear transformation parameters, $\odot$ denotes element-wise multiplication, $\mathbf{g}_t$ is the reset vector for memory slots which is the output of the reset gate, calculated as below:

$$\mathbf{g}^t = \sigma\left(\sum_{j=1}^{k_m}\left(\left[\mathbf{a}^{t-1}, \boldsymbol{\Pi}^t\right] \odot \mathbf{W}_g\right)_{ij}\right) \quad (15)$$

where $\mathbf{W}_g \in \mathbb{R}^{(k_m+1)\times k_m}$ is linear transformation parameters of the reset gate, $\sigma(\cdot)$ stands for sigmoid function, $\mathbf{a}^{t-1}$ is the normalized attention value of posts for memory slots in previous steps, $\boldsymbol{\Pi}^t \in \mathbb{R}^{k_m \times k_m}$ denotes the similarity matrix of normalized update vectors $\frac{1}{N_B}\left(\boldsymbol{\Gamma}^\top \boldsymbol{H}_B^t\right)$ to memory slots vectors.

C. FAKE NEWS DETECTION NETWORK

Based on the combined multi-modal representations of posts $\boldsymbol{U}_x = \{\boldsymbol{u}_1; \cdots; \boldsymbol{u}_{k_x}\}$, a fake news detector is utilized to classify posts as fake news or real news. It deploys a fully connected layer with a corresponding activation function to calculate the probability that post is fake and the probability that post is real, formalized as below:

$$\hat{p}_n = \text{softmax}\left(\text{MLP}(\boldsymbol{u}_n)\right) \quad (16)$$

where MLP$(\cdot)$ stands for multi-layer perceptron, $\hat{p}_n$ denotes the classifying probability that post n is fake, and $\boldsymbol{u}_n$ is the final representation of post n. We use y_n to indicate the ground-truth labels of post n and employ the cross entropy loss function to calculate the total loss:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{k_x} -\left[y_n \log(\hat{p}_n) + (1 - y_n)\log(1 - \hat{p}_n)\right] \quad (17)$$

where k_x is the number of posts in the training set. We minimize the classification loss by seeking the optimal parameters θ^*, which can be depicted as below:

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

V. EXPERIMENTAL RESULTS

In this section, we conduct experiments to evaluate the proposed model against SOTA models on the public real-world datasets. Furthermore, we give a detailed experimental analysis to show more insights into our model.

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 [12] and PHEME [46]). 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.

TABLE 1. The statistics of two real-world datasets.

News	WEIBO	PHEME
# of Fake News	4748	1972
# of Real News	4779	3830
# of Images	38853	2672

1) WEIBO DATASET

The data of WEIBO dataset [12] are 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 have 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 unique images corresponding to each post. Each post in WEIBO dataset contains text and the corresponding image. Posts in the dataset are verified by Xinhua News Agency as real or fake news.

2) PHEME DATASET

The PHEME dataset [46] consists of data based on five breaking news, including charliehebdo, ferguson, germanwings-crash, 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 compare our model with two categories of SOTA approaches: single-modal approaches and multi-modal approaches.

1) SINGLE-MODAL APPROACHES

As against such multi-modal approach, we compare with four single-modal models described below.

- *SVM-TS [4]:* SVM-TS utilizes heuristic rules and a linear SVM classifier to detect fake news.
- *CNN [7]:* CNN employs a convolutional neural network to learn the feature representation by framing relative posts into fixed-length sequences.
- *GRU [6]:* GRU uses the multilayer GRU network to consider the post as a variable-length time series.
- *TextGCN [8]:* TextGCN is an algorithm that uses the graph convolutional network to learn words and document embeddings. The whole corpus is modeled as a heterogeneous graph.

2) MULTI-MODAL APPROACHES

Multi-modal models utilize information from both textual and visual data for the fake news detection task. We also compare with six multi-modal approaches described below.

- *Att-RNN [12]:* Att-RNN uses a RNN based attention mechanism to combine textual, visual, and social content information. Image features are incorporated into the joint features of text and social context, which are obtained with an LSTM network. To make a fair comparison, we remove the component processing social context information in our experiments.
- *EANN [13]:* EANN learns event-invariant multi-modal features of each post for fake news detection by employing an adversarial network base on the concatenation of extracted deep textual and visual features.
- *MVAE [14]:* MVAE utilizes a variational autoencoder with encoding and decoding modules for each modality to obtain a shared latent multi-modal representation between text and image, which is trained jointly with the subsequent classifier to identify fake posts.
- *SAFE [15]:* SAEF is a similarity-aware multi-modal method for fake news detection, which extracts both textual and visual features from news content, and investigates their relationships to obtain the final representation.

In addition, we also design several variants to demonstrate the effectiveness of each component in our model. Details of the variants will be introduced in the analysis of MTNN components in Section V-E.

C. EXPERIMENTAL SETTING

For comparison with the other state-of-the-art approaches, metrics used in most fake news detection works are employed, including Accuracy calculated on all categories, along with F1 score, Precision, and Recall calculated respectively for fake news and real news. The Accuracy can be seen as the overall metric, for the WEIBO and FHEME datasets are not severe label imbalance. Moreover, we add the receiver operating characteristic curve (ROC) analyses for fake category in evaluating the proposed MTMN and the different variants of MTMN on experimental datasets.

The given datasets are partitioned into training sets, validation sets, and test sets according to the ratio of 7:1:2.

Given multi-modal posts, for textual content, we employ the pre-trained BERT module [44] 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 before the fully connected layer in the pre-trained ResNet50 model [45] 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] The transform unit of the blended attention module uses 4 attention heads.

For the whole model, we utilize the Adam optimizer during [47] the training stage. The early stopping strategy based on the validation set is employed, which is not sensitive to the number of epochs. The only requirement is that the number of epochs should be large enough for the early stopping strategy to obtain the best model. Depending on the default values of previous works and some additional experiments, we set the learning rate as 0.001 for 200 epochs, and the batch size is set to 256 to start training on the WEIBO and PHEME datasets.

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

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
	MTMN	**0.884**	0.920	0.841	**0.879**	0.853	0.927	0.888
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
	MTMN	**0.885**	0.819	0.827	**0.823**	0.916	0.912	0.914

D. QUANTITATIVE RESULTS

The detailed experimental results of fake news detection across all methods for WEIBO and PHEME are shown in Table 2. We can obtain the subsequent observations:

1) Across all real-world experimental datasets (WEIBO and PHEME), SVM-TS gets the worst performance in all baseline methods, indicating that the hand-crafted features cannot well characterize multi-modal posts of fake news. Deep learning models are superior to traditional machine learning models.
2) Across all datasets, most multi-modal approaches outperform unimodal approaches, manifesting that the additional visual information can enhance and complement the representation of posts to improve fake news detection.
3) In unimodal approaches, TextGCN performs better than SVM-TS and CNN, showing that model's performance can be enhanced by using a convolutional graph network to capture the relation of post elements and conduct inferring. Even some multi-model approaches such as EANN and SAFE present worse performance than TextGCN. In PHEME datasets, it is observed that CNN performs worse than the other baselines, for CNN cannot capture the long-distance semantic relationships between word and word, which is beneficial to detect fake news.
4) In multi-modal approaches, the performance of MVAE is better than SAFE, att-RNN, and EANN on all of the two datasets, showing that self-supervised loss function, which is incorporated in the multi-modal representation generating process, may take the role of a regular term 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. Method att-RNN has better performance only inferior to MVAE, showing that the attention mechanism can take the parts of the text corresponding to the image into consideration and enhance the performance of the whole model.
5) The proposed MMTN approach consistently performs superior to all the SOTA baselines across the two datasets, which shows that our model has the ability to generate more accurate, complementary, and comprehensive multi-modal representations and jointly learn post representation shared across topics and global features of latent topics which are integrated to obtain robust representation for newly arriving post with different topic distribution.

E. ANALYSIS OF MTNN COMPONENTS

Because the proposed MTMN contains multiple vital components, in this section, we compare variants of MTMN concerning the following aspects to demonstrate the effectiveness of MTMN: (1) the effect of the topic memory network component, (2) the effect of multi-modal blended attention, (3) the effect of the visual information. The following MTMN variants are designed for comparison.

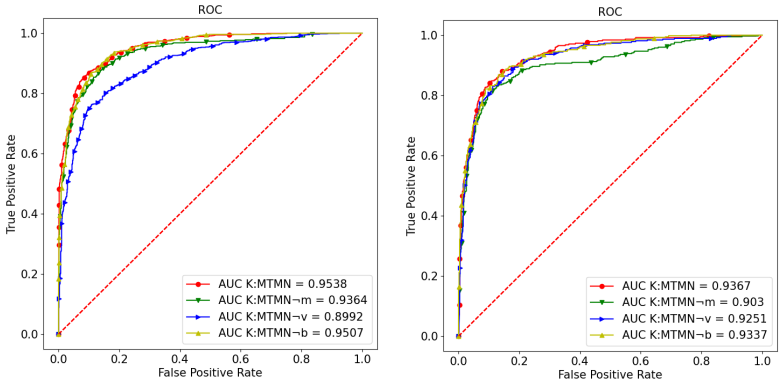
- MTMN$\neg m$: A variant of MTMN with the topic memory network component being removed.
- MTMN$\neg v$: A variant of MTMN with the visual information being removed.
- MTMN$\neg b$: A variant of MMCN with the multi-modal blended attention network being removed.

The metrics of ablation experiments are shown in Table 3. In addition, the receiver operating characteristic (ROC) curve and area under the curve (AUC) of fake category classified by each MTNN variant are exhibited in Figure 3.

(1) *Effects of the Topic Memory Network:* We compare the performance of MTMN with MTMN$\neg m$ on two

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

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	MTMN$\neg v$	0.824	0.887	0.742	0.808	0.778	0.906	0.837
	MTMN$\neg b$	0.878	0.894	0.859	0.876	0.864	0.897	0.880
	MTMN$\neg m$	0.867	0.870	0.864	0.867	0.864	0.871	0.867
	MTMN	**0.884**	0.920	0.841	**0.879**	0.853	0.927	0.888
PHEME	MTMN$\neg v$	0.872	0.843	0.745	0.791	0.884	0.933	0.908
	MTMN$\neg b$	0.876	0.828	0.781	0.804	0.898	0.922	0.910
	MTMN$\neg m$	0.871	0.798	0.803	0.801	0.906	0.903	0.904
	MTMN	**0.885**	0.819	0.827	**0.823**	0.916	0.912	0.914



(a) The ROC curves on WEIBO dataset (b) The ROC curves on PHEME dataset

FIGURE 3. The ROC curves of fake category classified by MTMN variants on WEIBO and PHEME datasets.

real-world datasets (WEIBO and PHEME). It can observe that the proposed MTMN performs better than MTMN$\neg m$ across all these datasets, according to the accuracy, f1 scores of fake and real category, and ROC curve of fake category. The result confirms the superiority of introducing the topic memory network and combining post feature shared across topics and global topic features in our model.

(2) *Effects of the Visual Information:* The performance of MTMN and MTMN$\neg v$ are compared on two real-world datasets (WEIBO and PHEME). It can see that the proposed MTMN performs better than MTMN$\neg v$ across all these datasets, according to the accuracy, f1 scores of fake and real category, and ROC curve of fake category. The result denotes that the visual information can consistently provide complementary information to benefit our model.

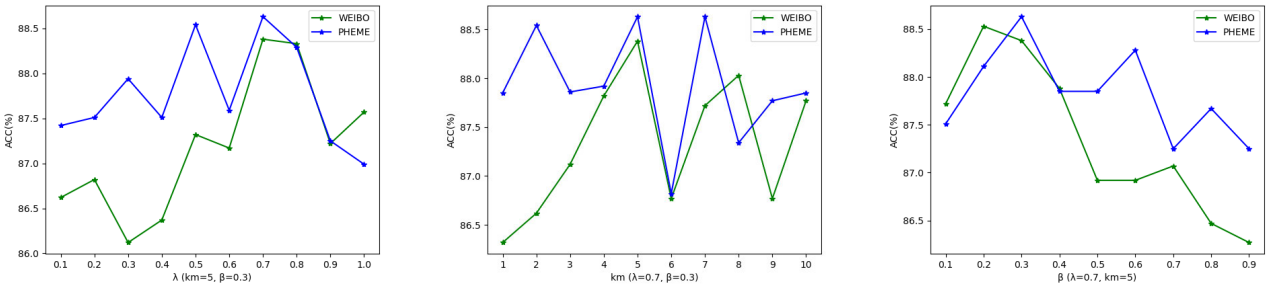
(3) *Effects of the Multi-Modal Blended Attention:* We compare the performance of MTMN with MTMN$\neg b$ on two real-world datasets (WEIBO and PHEME). It can observe that the proposed MTMN performs better than MTMN$\neg b$ on all these datasets, according to the accuracy, f1 scores of fake and real category, and ROC curve of fake category. The result shows that the multi-modal blended attention, which simultaneously exploits the intra-modal relation within each modality and the inter-modal relation between text words and image regions, effectively integrates multi-modal information.

In addition, the impact of the multi-modal blended attention is lower than those of the topic memory network and the visual information, according to the metrics and ROC curve of fake category.

F. IMPACT OF THE VALUE OF λ, k_m, AND β

There are some important hyper-parameters in the proposed MTMN whose impacts on the performances of fake news detection need to be detailed analyzed. When one of these hyper-parameters is varied to obtain the corresponding results, other hyper-parameters are fixed according to experience, leading to relatively better performance.

The output of the multi-modal blended attention network module is generated as Eq.(9): $\boldsymbol{h}_i = \lambda \boldsymbol{h}_{ei} + (1-\lambda)\boldsymbol{h}_{ri}$, where $\lambda \in [0, 1]$ is the tradeoff factor of the proportion of textual and visual information in the multi-modal features. The λ value is varied from 0.1 to 1.0 to represent the impacts for the accuracy of fake news detection on the two datasets. The results are shown in Figure 4(a). It can observe that the accuracies of the proposed model change with different λ values, where peaks exist in the interval [0.7, 0.8] on WEIBO dataset and [0.5, 0.8] on PHEME dataset. When λ is 0.7, the accuracies reach the highest values on WEIBO and PHEME datasets. So in the experiments, λ is set to 0.7.



(a) Impact of λ for the Accuracy (b) Impact of k_m for the Accuracy (c) Impact of β for the Accuracy

FIGURE 4. **Impact of different λ, k_m, and β for the performance of the proposed MTMN on WEIBO and PHEME datasets.**

The number of memory slots d_m in the topic memory network specifies the number of latent topics associated with multi-modal posts. The range of d_m is set as [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] on two datasets to analyze the impacts for the accuracy of fake news detection. Figure 4(b) shows the performance of MTMN with different values of d_m, from which we can observe that the model performance on dataset PHEME increases faster than that on dataset WEIBO when the value of k_m is small. The reason may be that the diversity of semantic space with latent topics is different in each dataset. The PHEME dataset is built by gathering thousands of posts linked with the five major news stories. However, the WEIBO dataset contains a wide range of information, which means the topics shared among posts is sparse. So in the experiments, k_m is set to 5.

The final representation combined by post feature shared across topics and relevant global topic feature formed by memory reading can be depicted as Eq.(13): $\mathbf{u}_i = (1-\beta)\mathbf{c}_i^P + \beta\mathbf{c}_i$ where β is the tradeoff factor of the proportion of relevant global topic feature and post feature. The β value is varied from 0.1 to 0.9 to represent the impacts for the accuracy of fake news detection on the two datasets. The results are shown in Figure 4(c). It can observe that the accuracies of proposed model change with different values of β, where peaks exist in interval $[0.2, 0.3]$ on WEIBO dataset and $[0.2, 0.3]$ on PHEME dataset. Considering the compositive performance of evaluations on WEIBO and PHEME datasets simultaneously, the best performance is achieved when β is 0.3. So in the experiments, β is set to 0.3.

VI. CONCLUSION

In this paper, we propose a novel end-to-end *Multi-modal Topic Memory Network* (MTMN), which obtains and combines post features shared across topics together with global features of latent topics while modeling intra-modality and inter-modality information in a unified framework.

We argue that most existing methods only focus on constructing models extracting abstract features from the content of each post, They neglect the intrinsic semantic architecture such as latent topics, etc. These models only learn patterns in content coupled with certain specific latent topics on the training set for distinguishing real and fake posts, which will suffer generalization and discriminating ability decline, especially when posts are associated with rare or new topics. In addition, advanced multi-modal fusion 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.

To address these limitations, MTMN is proposed to use two technical innovations:

(1) incorporates a topic memory network to explicitly characterize final representation as post feature shared across topics and global features of latent topics, which are jointly learned on the training set and then combined to generate robust representation for fake news detection.

(2) employs a blended attention module for multi-modal fusion, which is able to simultaneously exploit the relation among segments in each modal and the inter-modal relation between text words and image regions to complement and enhance each other for high-quality multi-modal representation.

Our method is evaluated on two real-world datasets (WEIBO and PHEME), and the experimental results demonstrate the proposed MTMN approach outperforms the SOTA baselines.

In future work, we plan to explore a more effective way to exploit background knowledge in deep neural networks, which can provide useful complementary information for fake news detection.

REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[2] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating fake news: A survey on identification and mitigation techniques," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 3, pp. 1–42, May 2019.

[3] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Syst. Appl.*, vol. 153, Sep. 2020, Art. no. 112986.

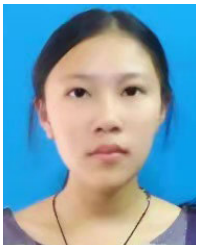
[4] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 1751–1754.

[5] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 1103–1108.

[6] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 2016, 2016, pp. 3818–3824.
[7] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation identification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3901–3907.
[8] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7370–7377.
[9] H. Zhou, H. Yin, H. Zheng, and Y. Li, "A survey on multi-modal social event detection," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105695.
[10] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017.
[11] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5674–5681.
[12] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 795–816.
[13] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 849–857.
[14] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, May 2019, pp. 2915–2921.
[15] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Proc. 24th Pacific-Asia Conf. Knowl. Discovery Data Mining (PAKDD)*. Springer, 2020, pp. 354–367.
[16] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 518–527.
[17] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: Real-time credibility assessment of content on Twitter," in *Proc. Int. Conf. Social Inform.* Cham, Switzerland: Springer, 2014, pp. 228–243.
[18] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2001.
[19] I. Kalamaras, A. Drosou, and D. Tzovaras, "Multi-objective optimization for multimodal visualization," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1460–1472, Aug. 2014.
[20] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.
[21] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via Bi-layer multimodal hypergraph learning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, Apr. 2019.
[22] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.
[23] M. Gupta, P. Zhao, and J. Han, "Evaluating event credibility on Twitter," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2012, pp. 153–164.
[24] L. Zhao, Q. Hu, and W. Wang, "Heterogeneous feature selection with multi-modal deep neural networks and sparse group LASSO," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1936–1948, Nov. 2015.
[25] T.-K. Yan, X.-S. Xu, S. Guo, Z. Huang, and X.-L. Wang, "Supervised robust discrete multimodal hashing for cross-media retrieval," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 1271–1280.
[26] Z. Zhao, Q. Yang, H. Lu, T. Weninger, D. Cai, X. He, and Y. Zhuang, "Social-aware movie recommendation via multimodal network learning," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 430–440, Feb. 2018.
[27] A. Graves, G. Wayne, and I. Danihelka, "Neural Turing machines," 2014, *arXiv:1410.5401*. [Online]. Available: http://arxiv.org/abs/1410.5401
[28] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwinska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, and A. P. Badia, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.
[29] C. Gulcehre, S. Chandar, K. Cho, and Y. Bengio, "Dynamic neural Turing machine with continuous and discrete addressing schemes," *Neural Comput.*, vol. 30, no. 4, pp. 857–884, Apr. 2018.
[30] M. Collier and J. Beel, "Implementing neural Turing machines," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2018, pp. 94–104.
[31] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," 2015, *arXiv:1503.08895*. [Online]. Available: http://arxiv.org/abs/1503.08895
[32] F. Liu and J. Perez, "Gated end-to-end memory networks," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 1–10.
[33] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1378–1387.
[34] N. Majumder, S. Poria, A. Gelbukh, M. S. Akhtar, E. Cambria, and A. Ekbal, "IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3402–3411.
[35] B. Kim, H. Kim, and G. Kim, "Abstractive summarization of Reddit posts with multi-level memory networks," 2018, *arXiv:1811.00783*. [Online]. Available: http://arxiv.org/abs/1811.00783
[36] C.-S. Wu, R. Socher, and C. Xiong, "Global-to-local memory pointer networks for task-oriented dialogue," 2019, *arXiv:1901.04713*. [Online]. Available: http://arxiv.org/abs/1901.04713
[37] R. Das, M. Zaheer, S. Reddy, and A. McCallum, "Question answering on knowledge bases and text using universal schema and memory networks," 2017, *arXiv:1704.08384*. [Online]. Available: http://arxiv.org/abs/1704.08384
[38] K. Xu, Y. Lai, Y. Feng, and Z. Wang, "Enhancing key-value memory neural networks for knowledge based question answering," in *Proc. NAACL-HLT*, 2019, pp. 2937–2947.
[39] Y. Chen, L. Wu, and M. J. Zaki, "Bidirectional attentive memory networks for question answering over knowledge bases," 2019, *arXiv:1903.02188*. [Online]. Available: http://arxiv.org/abs/1903.02188
[40] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9225–9234.
[41] X. Lu, W. Wang, M. Danelljan, T. Zhou, J. Shen, and L. Gool, "Video object segmentation with episodic graph memory networks," in *Proc. ECCV*, Aug. 2020, pp. 661–679.
[42] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1999–2007.
[43] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, and H. Zha, "Sequential recommendation with user memory networks," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 108–116.
[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[46] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in *Proc. Int. Conf. Social Inform.* Cham, Switzerland: Springer, 2017, pp. 109–123.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

LONG YING received the B.E. and M.S. degrees in biomedical engineering from Beijing Institute of Technology, Beijing, China, in 2009, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2015.

He is currently an Assistant Professor with the School of Computer and Software, Nanjing University of Information Science and Technology, China. His current research interests include multimedia content analysis, machine learning, and graph neural networks.



HUI YU received the B.E. degree in software engineering from Hebei GEO University, Shijiazhuang, China, in 2019. She is currently pursuing the MA.Sc. degree with Nanjing University of Information Science and Technology, Nanjing, China. Her main research interest includes multimedia content analysis.

JINGUANG WANG is currently pursuing the master's degree with the School of Computer Science and Information Engineering, Hefei University of Technology. His main research interests include multimedia computing and graph-based small sample learning.

YONGZE JI is currently pursuing the bachelor's degree with the School of Computer Science and Technology, China University of Petroleum. His main research interests include data mining and text analysis.

SHENGSHENG QIAN received the B.E. degree from Jilin University, Changchun, China, in 2012, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include social media data mining and social event content analysis.