Multi-modal affine fusion network for social media rumor detection

Boyang Fu [1] , **Jie Sui** [Corresp. 1]

[1] School of Engineering Science, University of Chinese Academy of Sciences, Beijing, China

Corresponding Author: Jie Sui
Email address: suijie@ucas.edu.cn

With the rapid development of the Internet, people obtain much information from social media such as Twitter and Weibo every day. However, due to the complex structure of social media, many rumors with corresponding images are mixed in factual information to be widely spread, which misleads readers and exerts adverse effects on society. Automatically detecting social media rumors has become a challenge faced by contemporary society. To overcome this challenge, we proposed the multimodal affine fusion network (MAFN) combined with entity recognition, a new end-to-end framework that fuses multimodal features to detect rumors effectively. The MAFN mainly consists of four parts: the entity recognition enhanced textual feature extractor, the visual feature extractor, the multimodal affine fuser, and the rumor detector. The entity recognition enhanced textual feature extractor is responsible for extracting textual features that enhance semantics with entity recognition from posts. The visual feature extractor extracts visual features. The multimodal affine fuser extracts the three types of modal features and fuses them by the affine method. It cooperates with the rumor detector to learn the representations for rumor detection to produce reliable fusion detection. Extensive experiments were conducted on the MAFN based on real Weibo and Twitter multimodal datasets, which verified the effectiveness of the proposed multimodal fusion neural network in rumor detection.

Multi-Modal affine fusion network for social media rumor detection

Boyang Fu[1] **and Jie Sui**[*1]

[1]**School of Engineering Science, University of Chinese Academy of Sciences, Beijing, China**

Corresponding author:
Jie Sui[1]

Email address: suijie@ucas.ac.cn

ABSTRACT

With the rapid development of the Internet, people obtain much information from social media such as Twitter and Weibo every day. However, due to the complex structure of social media, many rumors with corresponding images are mixed in factual information to be widely spread, which misleads readers and exerts adverse effects on society. Automatically detecting social media rumors has become a challenge faced by contemporary society. To overcome this challenge, we proposed the multimodal affine fusion network (MAFN) combined with entity recognition, a new end-to-end framework that fuses multimodal features to detect rumors effectively. The MAFN mainly consists of four parts: the entity recognition enhanced textual feature extractor, the visual feature extractor, the multimodal affine fuser, and the rumor detector. The entity recognition enhanced textual feature extractor is responsible for extracting textual features that enhance semantics with entity recognition from posts. The visual feature extractor extracts visual features. The multimodal affine fuser extracts the three types of modal features and fuses them by the affine method. It cooperates with the rumor detector to learn the representations for rumor detection to produce reliable fusion detection. Extensive experiments were conducted on the MAFN based on real Weibo and Twitter multimodal datasets, which verified the effectiveness of the proposed multimodal fusion neural network in rumor detection.

INTRODUCTION

As Internet technology gradually matures, online social networking (OSN) has become the spiritual ecology. Since OSN information is open and easily accessible, social networking software such as Weibo, Twitter, and Facebook have become the primary sources for millions of global users to receive news and information. They serve as essential approaches for Internet users to express their opinions. However, the authenticity of published information cannot be detected without supervision. Such social networking software has become the source of public opinion in hot events and news media.

For example, during the tenure of Barack Obama as the US President, a tweet from the “so-called” Associated Press said, “Two explosions occurred in the White House, and US President Barack Obama was injured.” Three minutes after the tweet was sent, the US stock index plunged like a “roller coaster,” and the market value of the US stock market evaporated by 200 billion US dollars within a short period, which tremendously affected both the stock and bond futures. Soon after, the Associated Press issued a statement saying that its Twitter account had been hacked, and that tweet proved to be false news. Therefore, it is of great necessity to automatically detect social media rumors in the early stage, and this technology will be extensively applied with the rapid development of social networks.

Nowadays, online rumors are no longer in the single form of texts. Instead, they are often in multiple modalities that combine images and texts. Figure 1 shows the cases of rumors in the Twitter dataset, displaying the texts and images of each tweet. In Figure 1A, the news is fake based on the images and texts; it is hard to identify whether the news in Figure 1B is true or not, but the images are fake; we cannot determine the authenticity of the news in Figure 1C based on the images, but we can confirm that the information is false according to the texts.

Currently, most methods used to detect social media rumors automatically are based on traditional machine learning Tacchini et al. (2017); Dongo et al. (2020); Choi et al. (2020); Chou et al. (2021) and

(A) Text: MH-370 has been found near Bermuda

(B) Text: Sharks in the street...

(C) Text: Woman, 36, gives birth to 14 children from 14 different fathers

Figure 1. Three forms of rumors on Weibo and Twitter datasets

deep learning Song et al. (2021); Jinshuo et al. (2020); Rani et al. (2021); Gokhale et al. (2020). The neural network Rauf et al. (2021), and other learning mechanisms such as federated learning Gao et al. (2021) can learn the constantly changing high-dimensional feature representation of posts in the training process with the superior ability to extract features. The currently available research on rumor detection primarily focus on single modality Jin et al. (2020); Abdulrahman and Baykara (2020); Luo et al. (2021); Balpande et al. (2021), while multi-modal researches are still in infancy, and only a few recent researches have tried to explore the multiple modalities Jin et al. (2017); Wang et al. (2018); Khattar et al. (2019); Jinshuo et al. (2020); Huang et al. (2019).

In current studies, the features of images and texts are mostly fused through feature concentration and averaging results. Nevertheless, this single fusion method fails to represent the posts fully. Firstly, it cannot solve the problem caused by the difference in semantic correlation between texts and images in rumors and non-rumors; secondly, the semantic gap cannot be overcome. Moreover, unlike paragraphs or documents, the texts in posts that are usually short fail to provide enough context information, making our classification fuzzier and more random.

This paper introduces a new end-to-end framework to solve the above problems. This framework is known as the multi-modal affine fusion network (MAFN). In the proposed model, employing affine fusion, we fused the features of images and texts to reduce the semantic gap and better capture the semantic correlation between images and texts. Entity recognition was introduced to improve the semantic understanding of texts and enhance the ability of rumor detection models. MAFN can gain multi-modal knowledge representation by processing posts on social media to detect rumors effectively. This paper makes the following three contributions:

- We proposed the multi-modal affine fusion network (MAFN) combined with entity recognition for the first time better to capture the semantic correlation between images and texts.
- The proposed MAFN model enriched the semantic information of text with entity recognition, and entity recognition was fused with the extracted textual features to improve the semantic comprehension of text.
- Experiments show that the MAFN model proposed in this paper can effectively identify rumors on Weibo and Twitter datasets and is superior to currently available multi-modal rumor detection models.

RELATED WORK

In early research on rumor detection, Castillo et al. (2011); Kwon et al. (2013), the rumor detection model was mainly established based on the differences between the features of rumors and factual information. Castillo et al. Castillo et al. (2011) designed a simple model to evaluate the authenticity of information on Twitter by counting the frequency of words, punctuation marks, expressions, and hyperlinks in texts. On this basis, Kwon et al. Kwon et al. (2013) used the communication structure to build rumors into a communication network and put forward 15 structural features, including the mid-values of network depth and width. Yang et al. Yang et al. (2012) introduced other client-based and location-based functions to identify rumors on Sina Weibo. However, it is time and energy-consuming to design these features manually, and the language patterns are highly dependent on specific time and knowledge in corresponding fields. Therefore, these features cannot be correctly understood.

Rumors on social media have gradually transformed from text-based to multi-modal rumors that combine both texts and images. Data in different modalities can complement each other. An increasing number of researchers have tried to integrate visual information into rumor detection. Isha et al. Singh et al. (2021) manually designed textual, and image features in four dimensions, i.e., content, organization, emotions, and manipulation, and eventually fused multiple features to detect rumors. Jin et al. Jin et al. (2017) detected rumors by fusing the image and textural features of posts using the RNN combined with the attention mechanism. However, multi-modal features still depend highly on specific events in the dataset, which will weaken the model's generalization ability. Therefore, Wang et al. Wang et al. (2018) put forward the EANN model that connected the visual features and textual features of posts in series and applied the event discriminator to remove specific features of events and learn the shared features of rumor events. Experiments show that this method can detect many events that are difficult to distinguish in a single modality.

Ma et al. Ma et al. (2016) introduced recurrent neural networks (RNN) to learn hidden representations from the texts of related posts and used LSTM, GRU, and 2-layer GRU to model text sequences, respectively. It was the first attempt to introduce a deep neural network into post-based rumor detection and achieve considerable performance on real datasets, verifying the effectiveness of deep learning-based rumor detection. Yu et al. Yu et al. (2017) used a convolutional neural network (CNN) to obtain critical features and their advanced interactions from the text content of related posts. Nonetheless, CNN is unable to capture long-distance features. Hence, Chen et al. Chen et al. (2019) applied an attention mechanism to the detection of network rumor and proposed a neural network model with deep attention. This model extracts adequate information and essential features from highly repeated texts, which solves the problems of excessive redundancy of texts in the data to be tested and weak information links between remote sites.

According to Dhruv K et al., Khattar et al. (2019), a single fusion method cannot effectively represent the posts. So, they used the encoder and decoder to extract the features of images and texts and learned across modalities with the help of Gaussian distribution. Liu et al. Jinshuo et al. (2020) put the text vector, the text vector in the image, and the image vector together, and then processed them using Gaussian distribution to get a new fusion vector to discover the association between the two modalities of hidden representation. Besides learning the text representation of posts, Zhang et al. Zhang et al. (2019) retrieved external knowledge to supplement the semantic representation of short posts and used conceptual knowledge as additional evidence to improve the performance of the rumor detection model.

METHODOLOGY

This paper introduced the four modules of the proposed MAFN model in this Section, i.e., the entity recognition enhanced textual feature extractor, the visual feature extractor, the multi-modal affine fuser, and the rumor classifier. Furthermore, we described the integration of the proposed modules to represent and detect rumors.

We instantiated tweets on Weibo and Twitter. The total tweets were expressed as $S = \{t_1, t_2, \ldots, t_n\}$, and each tweet was expressed as $t = \{T, E, V\}$, where T denotes the text content of the tweets, E represents the entity content extracted from the tweets, and V stands for the visual content matched with the tweets. $L = \{L_1, L_2, \ldots, L_m\}$ denotes the corresponding rumor and non-rumor tags of tweets. This paper aims to learn a multi-modal fusion classification model F by using the total tweets S and the corresponding tag sets L. F can predict rumors on unmarked social media. Figure 2 shows the framework of the *proposed* model.

The entity recognition enhanced textual feature extractor and obtained the joint representation R_u of text using Bert pre-training and self-attention mechanism. The visual feature extractor used the pre-trained model VGG19 to capture visual semantic feature R_v. The multi-modal affine fuser fused the joint representation and visual representation to obtain R_s, and the rumor classifier was utilized in the end to detect rumors.

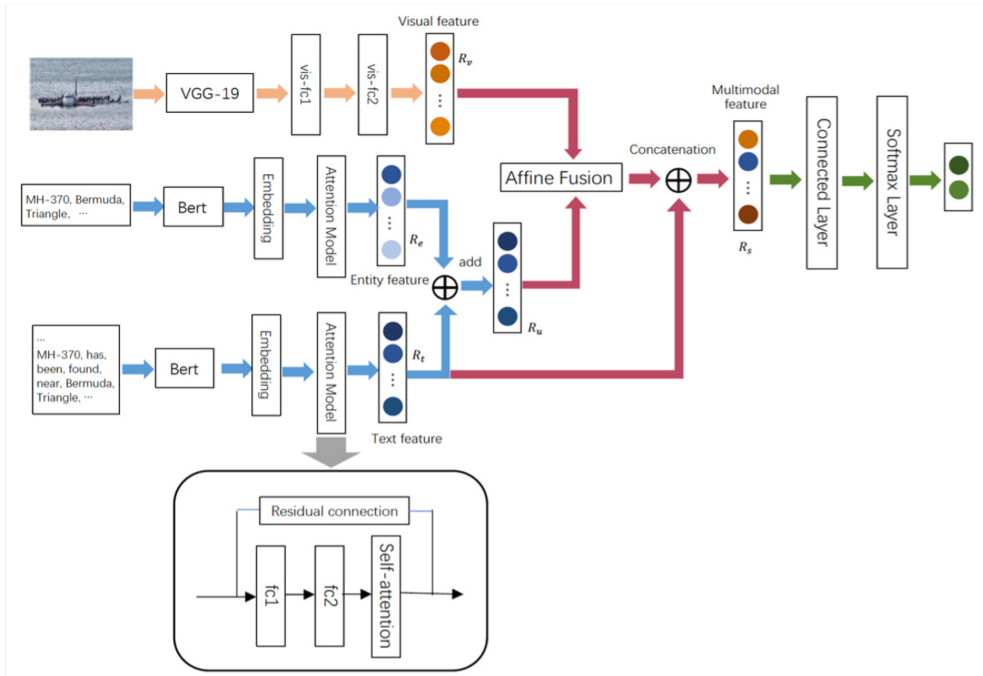


Figure 2. The model diagram of the proposed multimodal network MAFN.The yellow part represents the visual feature extractor, the blue part denotes the entity recognition enhanced textual feature extractor, the pink part stands for the multi-modal affine fuser, and the green part refers to the rumor detector.

Entity Recognition Enhanced Textual Feature Extractor

Extraction of Text Representation

Text representation is a short text representation generated from tweets. Our model extracted the feature vector of tweets through the Bert model to better capture the context's possible meaning and semantic meaning. Bert is a natural language processing model with the transformer bidirectional encoder representation as to the core, which can better extract the text context representation bidirectionally. By inputting the sequential vocabulary of the words in the tweets, the words were first embedded into the vector. The dimension of the ith word in the sentence is denoted by m, which is expressed as $W_i \in R^m$, and by inputting it into the sentence, S, it can be expressed as:

$$S = [W0, W1, W2, \ldots, Wp] \tag{1}$$

Where, $S \in R^{m*p}$, p denotes the total number of words, $W0$ denotes [CLS], and Wp represents [SEP]. By inputting the complete texts of tweets into the Bert model, we obtained the feature vector of the given sentence as

$$Sf = [Wf0, Wf1, Wf2, \ldots, Wfp]$$

Then the sentence feature vectors Sfn were given to the two fully connected layers. The above steps can be defined as follows:

$$Rt' = \sigma(Wft2 \cdot \sigma(Wft1 \cdot Sf + bt1) + bt2) \tag{2}$$

Where $Wft1$ denotes the weight matrix of the first fully connected layer with activation function, $Wft2$ represents the weight matrix of the second fully connected layer with activation function, and $bt1$ and $bt2$ are the bias terms.

The attention-based neural network can better obtain relatively long dependencies in sentences. The self-attention mechanism is a kind of attention mechanism that associates different positions of a single sequence to calculate the representation of the same sequence. To enable the model to learn the correlation between the current word and the other parts of the sentence, we added the self-attention mechanism after the fully connected layer, the process of which was expressed as follows:

$$Attself = softmax[QT \cdot KT^{\top}/\sqrt{m}] \cdot VT \tag{3}$$

Where, $QT = Rt' \times \mathrm{W}QT$, $KT = Rt' \times \mathrm{W}KT$, $VT = Rt' \times \mathrm{WV}T$ .WQT ,WKT ,WVT denote the three matrices learned by Q, K, and V, respectively. To make the model automatically recognize the importance of each word, degrade unimportant features to their original features, and process essential features using the self-attention mechanism, we used the residual connection to extract the features better. Figure 3 shows the architecture of a residual self-attention. A building block was defined as:

$$Rt = Attself + Rt' \tag{4}$$

Where, R_t denotes the eventually extracted text representation, $R_t \in R^k$.

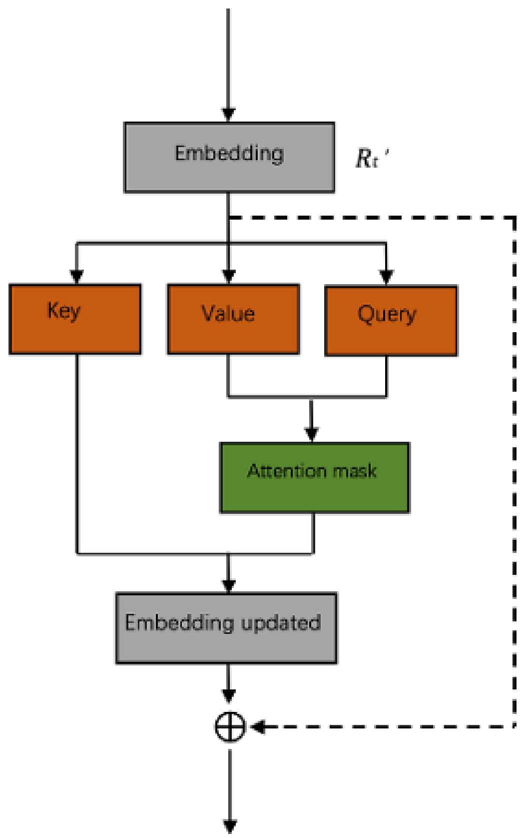


Figure 3. The architecture of a residual self-attention.

Extraction of Entity Representation

Named entity recognition identifies person names, place names, and organization names in a corpus. It was assumed that the combination of entity tagging and text coding in a post could supplement the semantic representation of the short text of the post in a certain way so that the model could identify rumors and non-rumors more accurately. Explosion AI developed spacy, a team of computer scientists and computational linguists in Berlin, and its named entity recognition model was pre-trained on OntoNotes 5, a sizeable authoritative corpus. In this paper, Spacy was applied to train the two datasets and extract the entities of posts. There were 18 kinds of identifiable entities.

First of all, we identified the recognizable word W_i as the entity $e \in E_s$ in every sentence $S = [W0, W1, W2, \ldots, Wp]$ of the tweet, and then obtained the tag $L \in \{L_1, L_2, \ldots, L_n\}$ corresponding to this entity, where L_i is one of the tags $\{PERSON, LANGUAGE, \ldots, LOC\}$. For instance, to instantiate a piece of text, we instantiated the entities in the text, as shown in Figure 4. The extracted entity $L_{European} = \{NORP\}$, $NORP$ means nationalities or religions or political groups; $L_{Google} = \{ORG\}$, OPG represents companies, agencies, institutions etc.

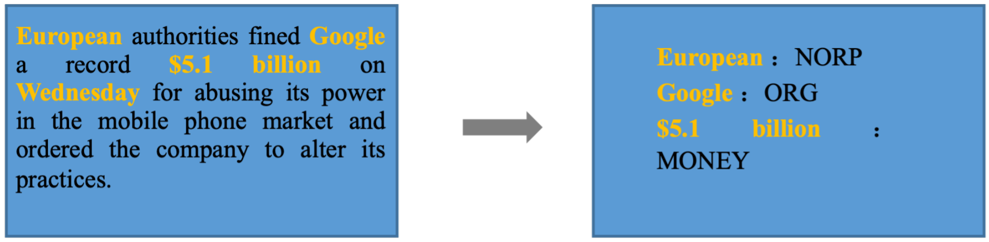


Figure 4. Illustration of entity refining process.

Based on the obtained L_i, the corresponding entity tags were connected in series to capture semantic features by Bert. $E_f \in R^k$, where R^k denotes the embedding dimension of tags. By inputting E_f into the residual attention mechanism, we gained $Re \in R^k$.

In the end, we combined the extracted text representation with the entity representation to obtain the joint representation Ru, $Ru \in R^k$, which was defined as follows:

$$Ru = add\ Re, Rt \tag{5}$$

Visual Feature Extractor

Images in tweets form the input into the visual feature extractor. This proposed framework used the pre-trained model VGG-19 and added two fully connected layers in the last layer to more comprehensively extract the visual features matched with the rumors in the tweet. According to the parameters unchanged after pre-training, VGG-19 adjusted the representation dimension of final visual features to k through two fully connected layers. We added the batch normalization layer and drop-out layer between the two fully connected layers and the activation function to prevent overfitting during the extraction of image representation. The eventually obtained feature of visual representation was expressed as Rv, where $Rv \in R^k$. The equation for extracting image features was defined as follows:

$$Rv' = Wfv2 \cdot \sigma(BN(Wfv1 \cdot Rvgg + bv1)) + bv2 \tag{6}$$

$$Rv = Dropout(\sigma(BN(Rv'))) \tag{7}$$

Where, $Rvgg$ represents the visual features extracted from the network in the pre-trained model VGG19, σ is the activation function, $Wfv1$ denotes the weight matrix of the first fully connected layer with the activation function, and $bv1$ and $bv2$ are the bias terms.

Multi-modal Affine Fuser

Affine transformation transforms into another vector space via linear transformation and translation. Through affine transformation, the multi-modal affine fuser fuses the multi-modal features extracted by the entity recognition enhanced textual feature extractor and the visual feature extractor, the joint representation and visual features of text and entity. It was assumed that the data of the two modalities could be fused more closely and the high-level semantic correlation could be better extracted. The corresponding equation was defined as follows:

$$R_c = \mathscr{F}\ R_v \cdot R_u + \mathscr{H}(R_v) \tag{8}$$

Where, Rc is the feature $R_c \in R^k$ gained after the fusion of all features, and $\mathscr{F}\ \cdot$ and $\mathscr{H}\ \cdot$ were fitted by the neural network. After extracting the fused features, in order to get more robust features, we reconnected the fused features with the textual features to obtain the total feature R_s. The equation was expressed as:

$$R_s = R_c \oplus R_t \tag{9}$$

Where, $\oplus$ denotes concatenation.

Rumor Detector

The rumor detector, based on the multi-modal affine fuser, sent the finally obtained multi-modal feature R_s to the multilayer perceptron for classification to judge whether the message was a rumor or not. The rumor detector consists of multiple completely connected layers with softmax. The rumor detector was expressed as $G(R_s^i, \theta)$, where θ represents all the parameters in the rumor detector, and R_s^i denotes the multi-modal representation of the case of the ith tweet. The rumor detector was defined as follows:

$$p_i = G(R_s^i, \theta) \tag{10}$$

Where p_i denotes the probability that the ith post input by the detector is a rumor, in the process of model training, we selected the cross-entropy function as the loss function, which was expressed as follows:

$$Loss = \sum_{i=1}^{N} -[L_i \times log\,(p_i) + (1 - L_i) \times log(1 - p_i)] \tag{11}$$

Where, L_i denotes the tag of the tweet in the i-th group, and N refers to the total number of training samples.

EXPERIMENTS

This section first described the datasets used in the experiment, namely two social media datasets extracted from the real world. Secondly, we briefly compared the results obtained by the most advanced rumor detection method and those gained by the model proposed in this paper. Through the MAFN ablation experiment, we compared the performances of different models.

Datasets

To fairly evaluate the performance of the proposed model, we used two standard datasets extracted from the real world to assess the rumor detection framework of the MAFN. These two datasets were composed of rumors and non-rumors collected from Twitter and Weibo, which simulated the natural open environment to some extent. They are currently the only datasets with paired image and text information.

Weibo Dataset

The Weibo dataset is a dataset proposed by Jin Jin et al. (2017) for rumor detection. It consists of the data collected by Xinhua News Agency, an authoritative news source in China, and the website of Sina Weibo and the data verified by the official rumor refuting system of Weibo. We preprocessed the dataset using a method similar to that put forward by Jin. First, locality sensitive hashing (LSH) was applied to filter out the same images and then delete irregular images such as very small or very long images to ensure that images in the dataset were of uniform quality. In the last step, the dataset was divided into the training and test sets. The ratio of tweets in training set to those in the test set was 8:2.

Twitter Dataset

The Twitter dataset Boididou et al. (2015) was released to verify the task of social media rumor detection. This dataset contains about 15,000 tweets focusing on 52 different events, and each tweet is composed of texts, images, and videos. The ratio of concentrated development set to test set in the dataset is 15:2, with the ratio of rumors to non-rumors being 3:2. Since this paper mainly studies the fusion of texts and images, we filtered out all tweets with videos. The ratio of development set and test set used to train the proposed model is the same as above.

Experiment Setting

The feature dimension of the images processed by VGG19 was 1000; the image features were extracted and embedded by two linear layers to obtain the feature dimension. After applying Bert and the linear layer were processed, the texts and entities were turned into 32-dimensional vectors. The entire training epochs was 50, and the batch size was 32. Adam served as the model optimizer during the training of the model. The initial learning rate was 0.001, and then lr varied with epoch based on the following equation:

$$\mathrm{p} = float(epoch)/100 \tag{12}$$

$$lr = 0.001/(1. + 10 * p) * *0.75 \tag{13}$$

Baselines

To verify the performance of the proposed multi-modal rumor detection framework based on knowledge attention fusion, we compared it with the single-modal methods, i.e., Textual and Visual, and five new multi-modal models. Textual and Visual were the subnetworks of the MAFN. The following are relatively new rumor detection methods for the comparative analysis:

- Neural Talk generates the words that describe images using the potential representations output by the RNN. Using the same structure, we applied the RNN to output the joint representation of images and texts in each step and then fed the representation into the fully connected layer for rumor detection and classification.
- EANN Wang et al. (2018): extracted textual features using Text-CNN, processes image features with VGG19 and then splices the two types of features together. With the features of specific events removed by the event discriminator, the remaining features were input into the fake news detector for classification.
- MVAE Khattar et al. (2019): used the structure of encoder-decoder to extract the image and textual features and conducted cross-modal learning with Gaussian distribution.
- att-RNN Jin et al. (2017) uses the RNN combined with the attention mechanism to fuse three modalities, i.e., image, textual, and user features. For a fair comparison, we removed the feature fusion in the user feature part of att-RNN, with the parameters of other parts being the same as those of the original model.
- MSRD Jinshuo et al. (2020) obtains a new fusion vector for classification by splicing textual features, textual features in images, and visual features extracted by VGG19 using Gaussian distribution.
- VQA is applied in the field of visual questioning and answering. Initially a multi-classification task, the image question-and-answer task was changed to a binary classification task. We used a single-layer LSTM with 32 hidden units to detect and classify rumors.

Performance Comparison

Table 1 shows the baseline results of single-modal and multi-modal models as well as the performances of the MAFN on two datasets in terms of the accuracy, precision, recall, and F1 of our rumor detection framework. MAFN performed better than the baseline models. The single textual model outperformed the single visual model on the Twitter dataset. Although the image features learned by visual features with the help of VGG-19 had better performance in rumor detection, the extraction of textural features was improved by Bert pre-training and residual attention. However, the single-modal model performed much. Among currently available multi-modal models, att-RNN uses LSTM and attention mechanism to process text representation, but it is not as good as EANN, which shows that EANN's event discriminator can better improve the model when it comes to rumor detection. The variational autoencoder proposed by MVAE can better discover multi-modal correlation, and it outperforms EANN. MAFN outperformed all baselines in terms of accuracy, precision, and F1, with high accuracy increasing from 82.7% to 84.2% and the F1 score going up from 82.9% to 84.0%. This verifies the effectiveness of MAFN in rumor detection.

A similar trend was found on the Weibo dataset. The textual model is superior to the visual model among the single-modal models. The accuracy of single text reaches 77.4%, which verifies the effectiveness of Bert pre-training and residual self-attention mechanism in improving semantic representation. Among the multi-modal methods, att-RNN, EANN, and MSRD proposed for this task outperform NeuralTalk and VQA, proving the necessity of improving modal fusion. The proposed MAFN achieved the best performance among other state-of-the-art models, with accuracy increasing from 74.5 % to 77.1% and the F1 score rising from 75.8% to 78.7%. This implies that the proposed model can better extract the multi-modal joint representation of images and texts.

Table 1. Comparison of performances of MAFN and other methods on Twitter and Weibo datasets.

Dataset	Method	Accuracy	Precision	Recall	F1
Twitter	Textual	0.551	0.680	0.605	0.520
	Visual	0.512	0.655	0.59	0.505
	NeuralTalk	0.610	0.728	0.504	0.595
	VQA	0.631	0.765	0.509	0.611
	att-RNN	0.664	0.749	0.615	0.676
	MSRD	0.685	0.725	0.636	0.678
	EANN	0.715	0.822	0.638	0.719
	MVAE	0.745	0.801	0.719	0.758
	MAFN	0.771	0.790	0.782	0.787
Weibo	Textual	0.774	0.679	0.812	0.739
	Visual	0.633	0.523	0.637	0.575
	NeuralTalk	0.717	0.683	0.843	0.754
	VQA	0.773	0.780	0.782	0.781
	att-RNN	0.779	0.778	0.799	0.789
	MSRD	0.794	0.854	0.716	0.779
	MVAE	0.824	0.854	0.769	0.809
	EANN	0.827	0.847	0.812	0.829
	MAFN	0.842	0.861	0.821	0.840

Component Analysis

To further analyze the performance of each part of the proposed model and to better describe the necessity of adding entity recognition and affine model, we carried out corresponding ablation experiments. We designed several comparison baselines, including simplified single-modal and multi-modal variants that removed some original models' components. The Weibo dataset contains a greater variety of events without strong specificity, better reflecting the rumors in the real world. Therefore, we ran the newly designed simplified variants on the Weibo dataset.

Table 2. Variants of the proposed MAFN's performance on Weibo datasets.

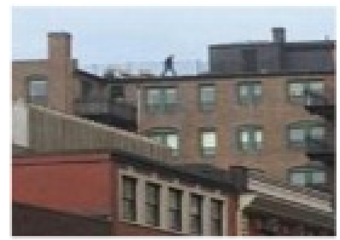
Method	Accuracy	Precision	Recall	F1
MAFN	0.842	0.861	0.821	0.840
w/o entity	0.836	0.826	0.826	0.826
w/o affine fusion	0.829	0.800	0.832	0.816
w/o entity+ affine fusion	0.819	0.750	0.852	0.797
Text-only	0.774	0.679	0.812	0.739
Entity-Link-Only	0.549	0.429	0.529	0.474
w/o image	0.799	0.719	0.834	0.772

As shown in Table 2, "w/o -entity" denotes the proposed MAFN without entity recognition module; "w/o -affine fusion" means removing affine fusion but retaining texts for entity recognition. Images and entity recognition were directly connected in series with the joint representation of texts. "w/o entity+ affine fusion" removed both entity and affine modules. "Text-only" refers to the single-text experiment. After pre-training the text using Bert, we connected the texts to the two fully connected layers and then accessed the residual self-attention to detect rumors directly. We conducted it for comparison. "Entity-Link-Only" results from rumor text detection carried out by only model branch entities. "w/o image" refers to the experiment without images, but only the combination of texts and entities. Furthermore, Table 2 indicates the performance of the simplified variant of MAFN. The experimental results show the necessity for the model to use affine fusion and enhance entity recognition. With entity-link added, the accuracy of single-modal text classification was increased from 77.4% to 79.9%, and F1 increased from 73.9% to 77.2%. 1.9% also improved the accuracy of image text fusion due to the introduction of entity branches. It was found that entity branches could supplement semantic representation, proving our idea effective. According to the experimental results, if we remove affine fusion, the accuracy of MAFN will decrease by 1.3%, and F1 will also decline by 2.4%. If images and texts are only connected without adding fusion and supplement, the accuracy will be lower. This proves the effectiveness of MAFN in rumor detection. MAFN can achieve more reliable multi-modal representation.

Case Study Performance Visualization

A qualitative analysis was performed on MAFN. After analyzing and ranking the examples of rumors successfully classified by MAFN, we selected the best two examples on Twitter and Weibo and showed them in Figure 5 and Figure 6, respectively. Without the support of affine fusion and entity recognition, the examples in Twitter could not be detected. Since the model failed to effectively capture the relationship between texts and images, these examples were misjudged as non-rumors. Insufficient text information and the absence of close connections between information and images are the reasons why the examples in Weibo could not be detected using “w/o entity+ affine fusion.” However, we can identify rumors with affine fusion by judging the image features.

(A) Right now, in Dresden. Over 30,000 at Pegida Anti-Immigrant.



(B) When the bomb exploded, the man on the roof was the one who caused the panic!!

Figure 5. Examples of successfully detecting rumors on Twitter by MAFN

(A) There was a big explosion in Tanggu. Please find out the truth and don’t let firefighters die in vain! Don’t report false death toll!.

(B) Every time there is an accident! All the victims were 35. When some accidents happen, the death toll is doomed.

Figure 6. Examples of successfully detecting rumors on Weibo by MAFN

CONCLUSION

This paper proposed an affine fusion network combined with entity recognition. This network accurately identifies rumors using the affine fusion between the entity recognition joint representation of images and texts. When extracting text representation, we used Bert to generate sentence vector features and learn semantics by extracting knowledge from the outside through entity recognition. Moreover, affine fusion was used for multi-modal fusion to better summarize the invariant features of new events. The Twitter and Weibo datasets experiments show that the proposed model is robust and performs better than the most advanced baselines. In the future, we plan to capture and identify rumor propagation in the field of rumor text and short videos to strengthen the generalization ability of the multi-modal fusion model.

ACKNOWLEDGEMENT

This work is supported by The National Natural Science Foundation of China (Grant No. 61572459).

REFERENCES

Abdulrahman, A. and Baykara, M. (2020). Fake news detection using machine learning and deep learning algorithms. In *2020 International Conference on Advanced Science and Engineering (ICOASE)*, pages 18–23. IEEE.

Balpande, V., Baswe, K., Somaiya, K., Dhande, A., and Mire, P. (2021). Machine learning approach for fake news detection. *International Journal of Research in Engineering, Science and Management*, 4(4):189–190.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., Kompatsiaris, Y., et al. (2015). Verifying multimedia use at mediaeval 2015. *MediaEval*, 3(3):7.

Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Chen, Y., Sui, J., Hu, L., and Gong, W. (2019). Attention-residual network with cnn for rumor detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1121–1130.

Choi, D., Chun, S., Oh, H., Han, J., et al. (2020). Rumor propagation is amplified by echo chambers in social media. *Scientific reports*, 10(1):1–10.

Chou, H.-C., Liu, Y.-W., and Lee, C.-C. (2021). Automatic deception detection using multiple speech and language communicative descriptors in dialogs. *APSIPA Transactions on Signal and Information Processing*, 10.

Dongo, I., Cadinale, Y., Aguilera, A., Martínez, F., Quintero, Y., and Barrios, S. (2020). Web scraping versus twitter api: A comparison for a credibility analysis. In *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*, pages 263–273.

Gao, J., Wang, W., Liu, Z., Billah, M. F. R. M., and Campbell, B. (2021). Decentralized federated learning framework for the neighborhood: A case study on residential building load forecasting. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 453–459.

Gokhale, T., Banerjee, P., Baral, C., and Yang, Y. (2020). Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer.

Huang, F., Zhang, X., Zhao, Z., Xu, J., and Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167:26–37.

Jin, Y., Wu, D., and Guo, W. (2020). Attention-based lstm with filter mechanism for entity relation classification. *Symmetry*, 12(10):1729.

Jin, Z., Cao, J., Guo, H., Zhang, Y., and Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Jinshuo, L., Kuo, F., Pan, J. Z., Juan, D., and Lina, W. (2020). Msrd: Multi-modal web rumor detection method. *Journal of Computer Research and Development*, 57(11):2328.

Khattar, D., Goud, J. S., Gupta, M., and Varma, V. (2019). Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

Kwon, S., Cha, M., Jung, K., Chen, W., and Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE.

Luo, Z., Wang, L., Wang, W., and Ye, A. (2021). Annealing attention networks for user feature-based rumor early detection on weibo. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., and Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks.

Rani, N., Das, P., and Bhardwaj, A. K. (2021). A hybrid deep learning model based on cnn-bilstm for rumor detection. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pages 1423–1427. IEEE.

Rauf, H. T., Bangyal, W. H. K., and Lali, M. I. (2021). An adaptive hybrid differential evolution algorithm for continuous optimization and classification problems. *Neural Computing and Applications*, 33(17):10841–10867.

Singh, V. K., Ghosh, I., and Sonagara, D. (2021). Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, 72(1):3–17.

Song, C., Ning, N., Zhang, Y., and Wu, B. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., and de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (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*, pages 849–857.

Yang, F., Liu, Y., Yu, X., and Yang, M. (2012). Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7.

Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T., et al. (2017). A convolutional approach for misinformation identification. In *IJCAI*, pages 3901–3907.

Zhang, H., Fang, Q., Qian, S., and Xu, C. (2019). Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1942–1951.