Multi-modal Fusion Network for Rumor Detection with Texts and Images

Boqun Li, Zhong Qian, Peifeng Li(✉), and Qiaoming Zhu

School of Computer Science and Technology, Soochow University, Suzhou, China
20205227046@stu.suda.edu.cn, {qianzhong,pfli,qmzhu}@suda.edu.cn

Abstract. Currently, more and more individuals tend to publish texts and images on social media to express their views. Inevitably, social media platform has become a media for a large number of rumors. There are a few studies on multi-modal rumor detection. However, most of them simplified the fusion strategy of texts and images and ignored the rich knowledge behind images. To address the above issues, this paper proposes a Multi-Modal Model with Texts and Images (M^{3}TI) for rumor detection. Specifically, its Granularity-fusion Module (GM) learns the multi-modal representation of the tweet according to the relevance of images and texts instead of the simple concatenation fusion strategy, while its Knowledge-aware Module (KM) retrieves image knowledge through the advanced recognition method to complement the semantic representation of image. Experimental results on two datasets (English PHEME and Chinese WeiBo) show that our model M^{3}TI is more effective than several state-of-the-art baselines.

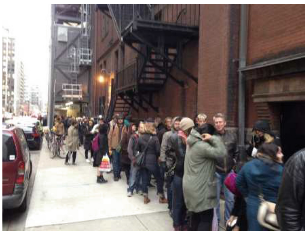
Keywords: Rumor detection · Coarse-grained and fine-grained · Image knowledge

1 Introduction

Rumors, a variety of false information which is widely spread on the social media, can publicize false information, spread panic, mislead the public, and cause terrible effects. For example in Iran, with the rise of COVID-19 cases and deaths, the use of so-called "traditional" and "Islamic" anti-coronavirus drugs has been sought after by some people. However, these drugs without safety certification are very dangerous to the human body. Hence, considering the potential panic and threat caused by rumors, rumor detection on social media efficiently and as early as possible is a way to control rumor propagation.

Twitter and Microblog have become the focus of rumor detection research because of their huge number of users and wide propagation. To reduce manpower and implement automatic rumor detection, a large number of methods based on neural network have been proposed [2,8,11,17] and almost all typically only considered single modality, i.e., text, and largely ignored the visual data and knowledge behind images. Only a few studies focused on both texts

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13141, pp. 15–27, 2022.
https://doi.org/10.1007/978-3-030-98358-1_2



These folks sure hope so. Lineup surprisingly short.

Fig. 1. Example of multi-modal tweet.

and images [5,16,18]. However, most of them simplified the fusion strategy of texts and images and ignored the richer knowledge behind images. For example, Wang et al. [16] simply concatenated textual features and visual features to get multi-modal features. Jin et al. [5] proposed to use attention mechanism to fuse text and image features. Zhang et al. [18] proposed a multi-channel CNN for combining text, text knowledge, image features. However, all of them only used the high-dimensional representation learnt by images, and did not pay attention to the rich semantic knowledge behind images.

Nevertheless, take Fig. 1 as an example, the methods above have following issues: 1) These strategies adopted similar fusion method for different samples which are unable to pinpoint the key region of text ("*Lineup surprisingly short*"), let alone the simple concatenating strategy; 2) All of them used the high-dimensional representation learnt by images which only captured underlying features of the visual layer (such as "*person*"), for better understanding the image, we prefer to acquire semantic knowledge of the concept layer (such as "*crowd*" or "*mass*") to prove "*Lineup surprisingly short*" belongs to rumor.

To address the above issues, we propose a Multi-Modal Model with Texts and Images (M^{3}TI) to detect rumors, which mainly considers three features, i.e., texts, images, and image knowledge. Firstly, our Granularity-fusion Module (GM) using a coarse-grained and fine-grained fusion mechanism is introduced to learn the latent connection from images and source texts. Then, our knowledge-aware Module (KM) integrates the image knowledge[1] and the source texts to capture the full semantic meaning of images. Finally, the multi-modal representation and knowledge representation extracted by GM and KM, respectively, are fed into a classification module with reply representation for rumor detection. Experimental results on two publicly available datasets (English PHEME and Chinese WeiBo) show that our model M^{3}TI is more effective than several state-of-the-art baselines. To sum up, our main contributions can be summarized as the following three aspects:

(1) We applied a novel fusion strategy on multi-modal rumor detection to generalize better under different cases in reality, in which Granularity-fusion

[1] The image knowledge consists of the entities with brief introduction and is extracted by an object recognition tool(https://ai.baidu.com/tech/imagerecognition/general).

Module uses a coarse-grained and fine-grained fusion mechanism to consider the relevance between images and texts.

(2) We exploited image knowledge consisting of entities with introduction to complement semantic meaning of images. As far as we know, it is the first time to use image knowledge for rumor detection and Knowledge-aware Module helps the model intuitively understand image and context together.

(3) We experimentally demonstrated that our model is more robust and effective on two public datasets from the real world.

2 Related Work

The neural network models have been proved effective in rumor detection. Compared with traditional machine learning, neural network can save a lot of manpower and learn the representation of various information more effectively.

Rumor Detection on Texts. Various methods using text modality have been proposed in rumor detection. Among them, Ma et al. [9] proposed the hierarchical attention to notice the specific sentences conducive to rumor detection. Li et al. [7] used user information, attention mechanism and multi-task learning to detect rumors. They added the user credibility information to the rumor detection layer, and applied the attention mechanism in the hidden layer of the rumor detection layer and stance detection layer. Khoo et al. [6] proposed a rumor detection model named PLAN based on attention, in which only texts were considered, and user interaction was achieved by self-attention.

Multi-modal Rumor Detection. There are only a few work on multi-modal rumor detection. EANN proposed by Wang et al. [16] included three components: feature extractor, fake news detector and event discriminator. It only simply concatenated the visual features and text features to obtain multi-modal features. We believe that the concatenation strategy breaks up the connection among words and visual data. Jin et al. [5] proposed a multi-modal detection model att-RNN and emphasized the importance of social context features. The attention mechanism was used to capture the association between the visual features and the text/social joint features. Zhang et al. [18] proposed a model MKEMN to learn the multi-modal representation of the tweet and retrieve the knowledge about the entities from texts. Different from Zhang et al. we further use the image information to obtain the entities in the image, which can make use of images more intuitively and effectively.

Other Multi-modal Tasks. Emotion analysis and sarcasm detection are related to rumor detection, and there are a few relative methods based on multi-modal content in these aspects. Although the inputs are different, their fusion mechanism can inspire our tasks. Poria et al. [12] used multi-core learning to fuse different modalities. Gu et al. [4] aligned text and audio at the word level and applied attention mechanism to fuse them. Cai et al. [1] proposed to use the attribute features in the image, then applied early fusion and late fusion between these attributes and texts to detect whether a sentence is ironic.

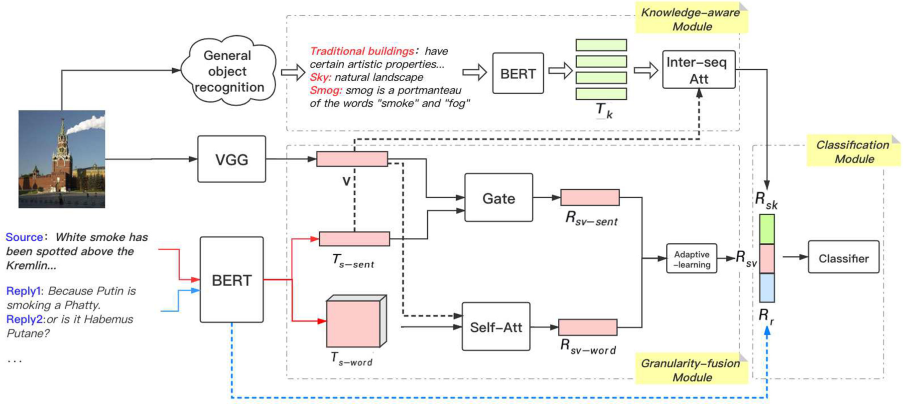

Knowledge-aware Module
General object recognition
Traditional buildings: have certain artistic properties...
Sky: natural landscape
Smog: smog is a portmanteau of the words "smoke" and "fog"
BERT
T_k
Inter-seq Att
VGG
v
Gate
$R_{sv-sent}$
T_{s-sent}
Source: White smoke has been spotted above the Kremlin...
Reply1: Because Putin is smoking a Phatty.
Reply2: or is it Habemus Putane?
...
BERT
T_{s-word}
Self-Att
$R_{sv-word}$
Adaptive -learning
R_{sv}
R_{sk}
R_r
Classification Module
Classifier
Granularity-fusion Module

Fig. 2. Multi-modal model with texts and images (M^{3}TI) for rumor detection

3 Approach

3.1 Overview

We define a thread, i.e., $thread = \{S_0, I_0, R_1, R_2, ..., R_n\}$, where S_0 means source tweet, I_0 is the image attached to S_0 and it may be null which depends on whether the user published the image, R_i is the i-th tweet in chronological order, n means that the thread has n relevant replies. The goal of our model is to classify the thread as *"Rumor"* or *"Non-Rumor"*.

Figure 2 shows the architecture of our proposed Multi-Modal Model with Texts and Images (M^{3}TI) for rumor detection. Our model mainly includes the following three parts: Granularity-fusion Module (corresponding to Sect. 3.2), Knowledge-aware Module (corresponding to Sect. 3.3), and Classification Module (corresponding to Sect. 3.4).

We briefly introduce the process of three modules above. Firstly, source tweet and image are fused through Coarse-grained and Fine-grained Mechanism to obtain two fused vectors. Adaptive Learning Mechanism is used to balance them. Secondly, we combine the source text and image knowledge through Inter-Sequence Attention. At last, we make use of two vectors obtained by two modules mentioned above and information of replies to detect rumors.

3.2 Granularity-Fusion Module

Text Encoder. Text encoder module aims to produce the text representation for a given source text S_0. Firstly, we feed source text into a Bert-Base-Uncased [3] model to capture the contextual information, "[CLS]" is used to represent its sentence-level vector $T_{s-sent} = [s_1, s_2, ..., s_{768}]$ and the embedding corresponding to each token in the source text is used to represent the word-level vector matrix $T_{s-word} = [s_{w1}, s_{w2}, ..., s_{wn}]$, where n is the number after word segmentation of each source text, $s_{wi} \in \mathbb{R}^{768}$ is used as the embedding of each token after word segmentation. We also get sentence-level vector of replies $R_r \in \mathbb{R}^{768}$.

Visual Encoder. Given an image (I_0) , we firstly resize the image to 224 × 224 pixels, then feed it into the vision sub-network to generate $v \in \mathbb{R}^{512}$. The structure of its front layers is the same as VGG-16 [14] network. We modify the last two full connection layers of VGG-16. Only the parameters of the last two full connection layers will be updated when training.

Coarse-grained Fusion. When fusing different modalities, one representation may overwhelm another, resulting in biased representation. Sangwan et al. [13] have proved that the Gating Mechanism is very effective for balancing two different modalities. We concatenate T_{s-sent} and v to obtain the weight $w^{'}$ which represents the weight of the text modality w.r.t. image modality, and then obtain the transformed representation $R_{sv-sent} \in \mathbb{R}^{768}$ as follows.

$$w = T_{s-sent} \oplus v \tag{1}$$

$$w^{'} = softmax(w) \tag{2}$$

$$R_{sv-sent} = w^{'} T_{s-sent} \tag{3}$$

Fine-grained Fusion. In order to accurately pinpoint the specific region of source text highly relevant to the image. We apply an image-guided text attention method. Given an image vector v, we firstly map v into text space through fully connected layer to obtain $v^{'}$, and use Q to imply the relationship between image and words. M represents the new embedding matrix calculated from the weight matrix W and T_{s-word}. Finally we obtain the attention distribution over each region of text, where $R_{sv-word} = [r_1, r_2, ..., r_{768}]$, the r_i denotes the weight of attention from the i-th region of text features towards the image representation. The Self-Att Mechanism equations are as follows.

$$v^{'} = fully\_connected(v) \tag{4}$$

$$Q = v^{'} T_{s-word} \tag{5}$$

$$W = softmax([\sum_{i=1}^{768} Q_{1_i}, ..., \sum_{i=1}^{768} Q_{n_i}]) \tag{6}$$

$$M = W^T T_{s-word} = [m_1, ..., m_n] \tag{7}$$

$$R_{sv-word} = [\sum_{t=1}^{n} m_{t_1}, ..., \sum_{t=1}^{n} m_{t_{768}}] \tag{8}$$

Adaptive Learning. We utilize an adaptive gate $e \in \mathbb{R}^{768}$ to combine relevance from the coarse-grained representation and the fine-grained representation. The final representation $R_{sv} \in \mathbb{R}^{768}$ is computed as follows.

$$e = softmax(W_e(R_{sv-sent} \oplus R_{sv-word}) + b_e) \tag{9}$$

$$R_{sv} = (1 - e) \odot R_{sv-sent} + e \odot R_{sv-word} \tag{10}$$

where W_e, b_e are the trainable parameters of adaptive learning mechanism. $\oplus$ is concatenation and $\odot$ is the element-wise multiplication.

3.3 Knowledge-Aware Module

Knowledge Encoder. In previous work, Cai et al. [1] have proved that using image recognition method to identify image entities is beneficial for model to understand context. The image entities with brief introduction are extracted as image knowledge through object recognition method, where only image entities higher than the recognition threshold (i.e., 0.6) will be recognized. We concatenate the entities with their introduction, and then feed them into the shared Bert to obtain several image knowledge vectors $T_k \in \mathbb{R}^{m*768}$, where m is the number of entities recognized in the image.

Inter-sequence Attention. Then, the Inter-sequence Attention is used to obtain the fusion representation. M_{sk} including $\{T_{s-sent}, T_{k1}, T_{k2}, ..., T_{km}\}$ represents a sequence set. We need to initialize a training parameter $v \in \mathbb{R}^{768}$ to obtain the knowledge fusion representation $R_{sk} \in \mathbb{R}^{768}$ as follows.

$$M_{sk} = [T_{s-sent}, T_k] \tag{11}$$

$$M = tanh(M_{sk}) \tag{12}$$

$$\alpha = softmax(v^T M) \tag{13}$$

$$R_{sk} = tanh(M_{sk} \alpha^T) \tag{14}$$

3.4 Classification Module

We get the fusion vector of image knowledge and source text (R_{sk}), fusion vector of source text and attached image (R_{sv}), and vector of replies (R_r). We use two full connection layers as our classifier, the activation function of the output layer is tanh, and the loss function is the cross entropy function.

$$R_I = R_{sk} \oplus R_{sv} \oplus R_r \tag{15}$$

$$R = tanh[W_b(W_a(R_I) + b_a) + b_b] \tag{16}$$

where W_a, b_a, W_b, b_b are the trainable parameters of classifier.

4 Experimentation

4.1 Experimental Settings

In this paper, two public datasets are used to evaluate our model, i.e., English PHEME and Chinese WeiBo, in which the source tweets containing images and we crawl them. The data distribution of PHEME and WeiBo is shown in Table 1.

Table 1. Distribution of datasets

	PHEME	WeiBo
Events	6425	4664
Users	50,593	2,746,818
Tweets	105,354	3,805,656
Average length of tweets	13.6	99.9
Rumors	2403	2313
Non-rumors	4022	2351
Images in rumors	739	1982
Images in non-rumors	1667	1861
Percentage of rumors with images	30.75%	85.69%
Percentage of non-rumors with images	41.45%	79.16%
Percentage of all the threads with images	37.45%	82.40%

For PHEME and WeiBo, we randomly divided the data, and used the same processing method as Sujana et al. [15], 80% of the data was used as the training set, 10% as the validation set, and the remaining 10% as the test set. Similar to Ma et al. [10], we calculated the accuracy, precision, recall and F1 score to measure the performance of rumor detection.

We firstly preprocess our data such as deleting tags, emotions and removing retweets. In order to better fit the situation that may occur in reality, the source tweets in our datasets contain images randomly, it just depends on whether the user published the image. If the source tweet does not contain image, we uniformly give them a blank image without any information. For the images contained in the source tweets, we adjust the size to 224 × 224 pixels and normalize them. We use Adam optimizer to update the parameters, dropout is 0.2 and the learning rate is 10^{-5}.

4.2 Experimental Results

The corresponding baselines will be briefly described below.

SVM. A SVM classifier using bag-of-words and N-gram (e.g., 1-gram, bigram and trigram) features.

CNN. A convolutional neural network model (CNN) for obtaining the representation of each tweet followed by a softmax layer.

BiLSTM. A bidirectional RNN-based tweet model that considers the context of the target and tweet.

PLAN [6]**.** A model uses the maximum pooling layer at the word-level to obtain the tweet representation and the transformer encoder to encode at tweet-level, and finally obtains a high-level representation.

Table 2. Results of comparison with different baselines

Dataset	Method	Acc	Pre	Rec	F1
PHEME	SVM	0.688	0.518	0.512	0.515
	CNN	0.795	0.731	0.673	0.701
	BiLSTM	0.794	0.727	0.677	0.701
	PLAN	0.871	0.859	0.863	0.861
	att-RNN	0.880	0.868	0.869	0.869
	MKEMN	0.891	0.884	0.879	0.881
	Our Model	**0.901**	**0.893**	**0.890**	**0.892**
WeiBo	SVM	0.818	0.819	0.818	0.818
	CNN	0.897	0.897	0.897	0.897
	BiLSTM	0.929	0.931	0.929	0.929
	PLAN	0.939	0.939	0.938	0.938
	att-RNN	0.932	0.932	0.932	0.931
	MKEMN	0.942	0.942	0.942	0.942
	Our Model	**0.957**	**0.958**	**0.957**	**0.957**

att-RNN [5]**.** A multi-modal detection model based on LSTM, using attention mechanism to capture the association between the visual features and the text/social joint features.

MKEMN [18]**.** A model combining multi-modal content with external knowledge for rumor detection, and using memory network to measure the differences between different events.

Comparing the experimental results in Table 2, it can be seen that the proposed method achieves the best results both on Acc and F1. From Table 2, we can draw the following observations:

(1) The SVM model performs worst among all methods. This is for the reason that it is built with the hand-crafted features which have limitation for the task of rumor detection.

(2) CNN and BiLSTM have comparable performance in both datasets. In addition, BiLSTM outperforms CNN in WeiBo. We can see that the average length of tweets in WeiBo is much longer than PHEME from Table 1. This suggests that BiLSTM can better capture the forward and backward semantic dependence which is more suitable for long sentences.

(3) In PHEME, multi-modal att-RNN performs better than single-modal PLAN, but in WeiBo, att-RNN performs worse than PLAN. The phenomenon we can see from Table 1 that fewer users in PHEME publish more events, which indicates that users in PHEME are more social intensive. In this case, users are more inclined to publish images strongly related to texts in order to prove narration. att-RNN adopts the similar fusion strategy regardless of the relevance between image and text. So it will only perform better when images are highly relevant to texts.

Table 3. Ablation experiment results (**Text**: Text modality; **Image**: Image modality; **Coarse**: Coarse Grained Fusion; **Fine**: Fine Grained Fusion; **Knowledge**:Image Knowledge)

	Text	Image	Coarse	Fine	Knowledge	Acc	Pre	Rec	F1
PHEME	✓					0.866	0.853	0.861	0.856
		✓				0.639	0.320	0.500	0.390
	✓	✓				0.869	0.870	0.843	0.854
	✓	✓	✓			0.854	0.845	0.835	0.840
	✓	✓		✓		0.883	0.874	0.873	0.873
	✓	✓	✓	✓		0.891	0.883	0.880	0.882
	✓	✓			✓	0.888	0.885	0.869	0.876
	✓	✓	✓	✓	✓	**0.901**	**0.893**	**0.890**	**0.892**
WeiBo	✓					0.931	0.932	0.931	0.931
		✓				0.523	0.526	0.525	0.515
	✓	✓				0.933	0.934	0.934	0.934
	✓	✓	✓			0.936	0.936	0.936	0.936
	✓	✓		✓		0.872	0.875	0.872	0.872
	✓	✓	✓	✓		0.947	0.947	0.947	0.947
	✓	✓			✓	0.944	0.945	0.944	0.944
	✓	✓	✓	✓	✓	**0.959**	**0.959**	**0.960**	**0.959**

(4) MKEMN performs well on both datasets due to its effective fusion strategy, which integrates text, image and textual knowledge on different channels for better capturing semantic information.

(5) Compared with all baselines, our model outperforms other methods in most cases. We attribute the superiority to two properties: a)The GM module took the relevance between image and text into account and learned a multi-modal representation according to the correlation size that can accurately express the user's view. b)The KM module extracted the semantic knowledge from visual data to provide evidence for the judgment.

4.3 Ablation

In this section, we compare among the variants of M^{3}TI with respect to the following two aspects to demonstrate the effectiveness of our model: the usage of fusion strategy, and the usage of image knowledge. We conducted the ablation experiments as shown in Table 3 and we can conclude that:

(1) The model based on text modality is superior to model based on image modality, which is obvious because the text features contain more rich semantic information.

(2) When we simply concatenate image features and text features, the improvement of result is not obvious compared with text features alone, which shows that concatenation can not be effectively used as multi-modal representation.

(3) In PHEME, the effect of fine-grained fusion vector is better, but in WeiBo, the effect of coarse-grained fusion vector is better. As described in Sect. 4.2, the sample collection of PHEME and WeiBo is different, in PHEME, the relevance between texts and images is higher than WeiBo dataset. We can prove that fine-grained is fit for the dataset where image has strong relevance with text and coarse-grained is fit for the other case.

(4) However, when GM module uses Adaptive-learning ("✓" on both "Coarse" and "Fine") to balance coarse-grained vector and fine-grained vector, the result is significantly improved, which shows that the GM module can mine the relevance between images and texts and learn an optimal representation according to the relevance.

(5) When we use image knowledge, the results are also improved. This is obvious because it provides more understandable semantic information for the model.

Based on the analysis above, the following conclusions can be drawn, i.e., 1) The coarse-grained and fine-grained fusion mechanism combined by adaptive-learning plays an important role in fusing text and image; 2) Image knowledge is useful for our model to better understand context.

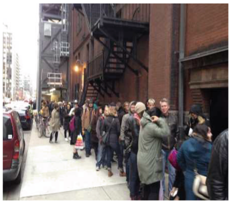
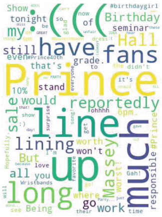
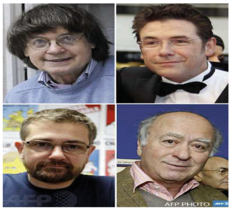
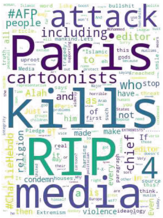
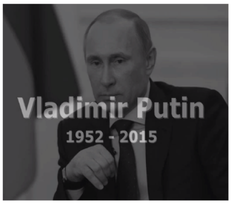
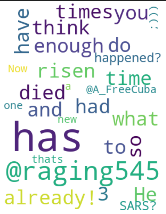
4.4 Case Study

To illustrate the importance of text, image and image knowledge for rumor detection, we present several intuitive cases into Table 4 and compare the predicted results of our model with the single-modal model. These cases explain our fusion strategy and how image knowledge and reply information work.

For Case 1, the rumor case was successfully detected by our proposed model, but ignored by the single-modal model based on text. Because the text content alone is not enough to give sufficient evidence while the image attached to it provides clues for detection. When the image and text are highly correlated, our model may increase the proportion of fine-grained vector to pinpoint important region of text "*Lineup surprisingly short*"; In the image knowledge, from the word "*demonstration*", we can more intuitively find that this thread belongs to rumor; With the help of the wordcloud of replies (we preprocess some unnecessary tags), the frequent words like "*much*" and "*long*" are contrary to the "*surprisingly short*". Therefore, such a thread that seems difficult to judge from the text alone can be accurately recognized by our model.

While in Case 2, the non-rumor case was successfully detected by our proposed model, but ignored by the single-modal model based on image. From the image alone, we can only get the information of four men in the photo, which can't judge the correctness of the statement that "*four cartoonists were killed*", but the view of replies (i.e., "*RIP*","*kill*") is more firm. When the relevance between source text and attached image is not strong, and the image can not provide effective evidence, our model may try to enhance the proportion of

Table 4. Cases

Number	Image	Source/Image knowledge	Reply-wordcloud
1		**Source:** *These folks sure hope so.Lineup surprisingly short.* **Image knowledge:** **street:** *Street is one of administrative division...* **demonstration:** *public activities and events: a way of modern society public opinion*	 Prince line up much long have fans Birthday Hall reportedly lining Massey sale
2		**Source:** *Paris media attack kills 4 cartoonists including chief editor.#AFP* **Image knowledge:** **group photo:** *a close-up, a photograph of several people together* **man:** *male human being*	 Paris kills RIP media attack cartoonists #AFP including editor chief 4 religion #CharlieHebdo
3		**Source:** *The internet has now pronounced @PutinRF dead #putindead* **Image knowledge:** **Putin:** *Public figure and President of Russia, Vladimir Putin has been nominated for the Nobel Peace Prize*	 has @raging545 already! risen died think enough do times you have happened? time had and what to so He SARS? 3 Now one new thats @A_FreeCuba

coarse-grained vector, so as to reduce the noise of irrelevant images. Combined with the views of commentators, we can successfully determine it as a non-rumor.

As shown in Case 3, the image and text are highly related. The image seems to learn "strong" evidence to prove that the text content is correct, and the replies are inadequate, our model may make mistakes. Although we all know that this is a rumor now, under the circumstances at that time, it will be difficult for our model to judge the authenticity of this tweet.

5 Conclusion

In this paper, we propose a Multi-Modal Model with Texts and Images (M^{3}TI), which explores a novel fusion strategy according to the relevance of texts and images, and semantic knowledge behind images. Our Granularity-fusion Module utilizes the adaptive gate to balance coarse-grained features and fine-grained features which can generalize well for different cases. Moreover, Knowledge-aware Module offers more intuitive image knowledge. In the comparative experiments of PHEME and WeiBo datasets, compared with other methods, our method uses three modalities, i.e., text, image, image knowledge, and achieves the best effect. In the future work, we will continue to study the fusion method to integrate images and texts effectively under misleading circumstances.

Acknowledgments. The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 61836007 and 62006167.), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

1. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2506–2515 (2019)
2. Chen, W., Zhang, Y., Yeo, C.K., Lau, C.T., Lee, B.S.: Unsupervised rumor detection based on users' behaviors using neural networks. Pattern Recogn. Lett. **105**, 226–233 (2018)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I.: Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2018, p. 2225. NIH Public Access (2018)
5. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 795–816 (2017)
6. Khoo, L.M.S., Chieu, H.L., Qian, Z., Jiang, J.: Interpretable rumor detection in microblogs by attending to user interactions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8783–8790 (2020)
7. Li, Q., Zhang, Q., Si, L.: Rumor detection by exploiting user credibility information, attention and multi-task learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1173–1179 (2019)
8. Lu, Y.J., Li, C.T.: GCAN: graph-aware co-attention networks for explainable fake news detection on social media. arXiv preprint arXiv:2004.11648 (2020)
9. Ma, J., Gao, W., Joty, S., Wong, K.F.: Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics (2019)
10. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks (2016)
11. Ma, J., Gao, W., Wong, K.F.: Detect rumor and stance jointly by neural multitask learning. In: Companion Proceedings of the Web Conference 2018, pp. 585–593 (2018)
12. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2539–2544 (2015)
13. Sangwan, S., Akhtar, M.S., Behera, P., Ekbal, A.: I didn't mean what i wrote! exploring multimodality for sarcasm detection. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)

14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Sujana, Y., Li, J., Kao, H.Y.: Rumor detection on twitter using multiloss hierarchical bilstm with an attenuation factor. arXiv preprint arXiv:2011.00259 (2020)
16. Wang, Y., et al.: 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, pp. 849–857 (2018)
17. Wu, L., Rao, Y., Zhao, Y., Liang, H., Nazir, A.: DTCA: decision tree-based co-attention networks for explainable claim verification. arXiv preprint arXiv:2004.13455 (2020)
18. Zhang, H., Fang, Q., Qian, S., Xu, C.: Multi-modal knowledge-aware event memory network for social media rumor detection. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1942–1951 (2019)