KMMN: Knowledge Enhanced Multimodal Multi-grained Network for Fake News Detection

Liyuan Zhang, Zeyun Cheng, Zhongyan Gui, Yan Yang(✉), and Yong Liu(✉)

Heilongjiang University, Harbin, China
{2231976,2231975}@s.hlju.edu.cn,{guizhongyan,yangyan,liuyong123456}@hlju.edu.cn

Abstract. The development of the social media has created an environment for the rapid spread of fake news. The existing automated detection methods may have the following shortcomings: (1) Content-based methods neglect the rich background information related to news; (2) Unable to effectively exploit multimodal information at both fine-grained and coarse-grained levels; (3) Unable to effectively handle ambiguity problem (information from different modalities may contradict each other). To overcome these challenges, we present a Knowledge enhanced Multimodal Multi-grained Network (KMMN) for fake news detection. We obtain background knowledge contained in news based on entities to enhance cross-modal interaction and provide external information. The cross-modal feature fusion process is separated at different granularities (with fine-grained and coarse-grained branches). We design an improved Mixture-of-Experts (iMoE) network for feature fusion and reweight the cross-modal features to alleviate ambiguity problem. Experimental results demonstrate that the proposed framework outperforms state-of-the-art methods on three public datasets.

Keywords: Fake news detection · Multimodal learning · Social media.

1 Introduction

Fake news detection has become a hot topic at present [5–7, 12]. Modern news often have multimodal information that is interrelated. In recent years, many methods that combine multimodal information to detect fake news have been proposed [2, 3, 11]. Although some progress has been made, they still face some shortcomings.

Firstly, most existing methods only model the basic semantics, ignoring the rich background information contained in news. Intuitively, when readers try to understand news, they often place the news in a broader social, historical, and political context, which can help them analyze the news more deeply and uncover potential relationships. As shown in Fig. 1, the text describes news about

This work was supported by the National Natural Science Foundation of China (No. 6247074060), the Natural Science Foundation of Heilongjiang Province in China (No. PL2024F029) and the Basic Research Funds for Provincial Universities in Heilongjiang Province (No. 2023-KYYWF-1486, No. 2024-KYYWF-0115).

celebrity Jaycee Chan, while the image shows a photo of another celebrity, Nicolas Cage. Content-only based model may learn that the visual and textual aspects contain a subject, but cannot correspond to whether the object in different modalities is the same person. That is to say, the lack of background knowledge may lead to one-sided judgments. The background knowledge contained in news is often described around entities. Some works have explored guiding fake news detection from the perspective of entities [4], but these works often based on the differences of entities between different modalities, rather than using the background information that these entities may contain. To compensate for this omission, we explored the use of these background knowledge to enhance the detection process of fake news, which were utilized in two ways: (1) Enhancing fine-grained cross-modal interaction; (2) Capture differences and provide additional knowledge features.

Secondly, although many works have proposed novel methods for aggregating multimodal information, they either only aggregate at the overall level [13], which may result in the loss of detailed information or only match from a fine-grained perspective [4], ignoring global consistency. In our method, the formation process of cross-modal features is clearly separated into two branches, which respectively aggregate coarse-grained and fine-grained information.

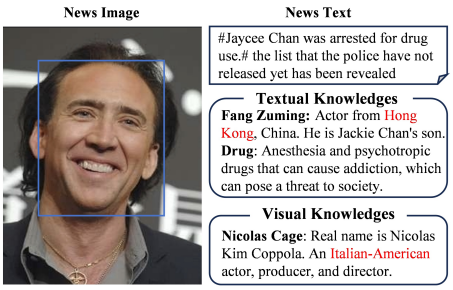

Fig. 1. A typical case of multimodal fake news. Both visual and textual content mention a celebrity, but different modalities of celebrities conflict with each other. The background knowledge of celebrity entities can help us identify this situation.

In addition, the correlation between modalities has a certain degree of ambiguity [11, 12]. The visual and textual information of some news has almost no semantic matching relationship. In this case, cross-modal feature fusion is unnecessary. FND-CLIP [13] uses aligned multimodal features generated by CLIP to evaluate ambiguity. We introduce this idea into the modality fusion process of our different branches to further alleviate ambiguity problem.

In summary, we propose a novel **K**nowledge enhanced **M**ultimodal **M**ulti-grained **N**etwork (KMMN) for detecting fake news, which introduces high-level knowledge semantic information that enhances cross-modal interaction and provides external information. We clearly separate the process of the formation of cross-modal features at coarse-grained and fine-grained, and design an **i**mproved **M**ixture-**o**f-**E**xperts (iMoE) network to aggregate multimodal information at different levels of granularity, and reweights the cross-modal features to alleviate the ambiguity problem.

The contributions of this paper are mainly three-folded, namely:

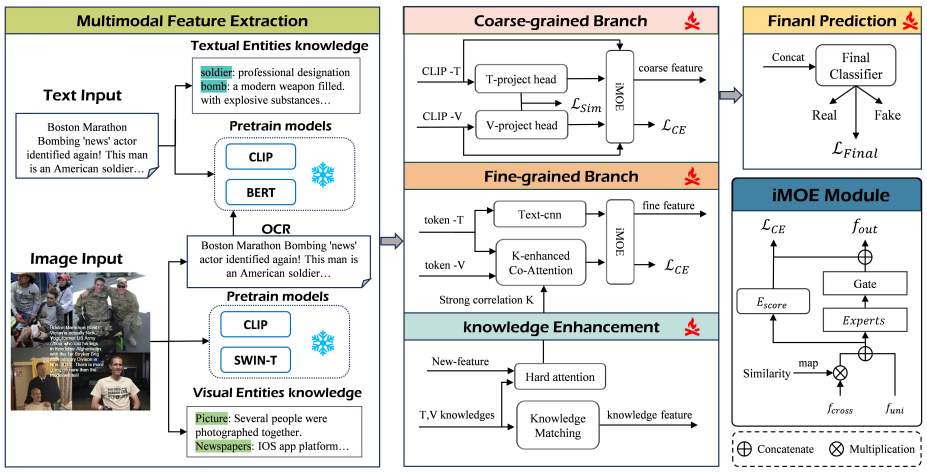

Fig. 2. An overall architecture of the proposed KMMN. It contains four components: feature extraction module, coarse-grained and fine-grained branches, knowledge enhancement module, classification module.

- We proposes a novel multimodal fake news detection framework KMMN that introduces news background knowledge to enhance cross-modal interaction process and provide additional information to the model.
- We specifically design two granularity-level branches and design iMoE to reweight and aggregate multimodal features, forming a comprehensive representation that reflects detailed, global, and background aspects of the news.
- We conducted extensive experiments on three real-world datasets to validate the effectiveness of KMMN. Our codes is publicly available.

2 Method

2.1 Model Overview

The proposed KMMN is shown in Fig. 2. Our model mainly includes the following four parts: (1) Multimodal feature extraction module. We use publicly available API to extract the background knowledge of news. (2) Cross-modal interaction module. We use the extracted knowledge information to enhance the fine-grained interaction process. (3) Knowledge matching module. We capture knowledge differences between modalities through knowledge matching to provide additional information to the model. (4) Fake news classifier.

2.2 Multimodal Feature Extraction

Textual Feature Encoding via BERT. We use a pre-trained BERT to encode the text. The text is the connection between title and embedded text extracted

https://github.com/ZhangLiyuan11/KMMN-main

from images. After applying BERT, a fine-grained text feature matrix $T^b = [t_1^b, t_2^b, ..., t_l^b]$ is obtained, where $t_i^b \in R^{d_b}$ is the *i-th* token, and d_b is the dimension of word embeddings.

Visual Feature Encoding via SWIN-T. For a given image V, we use a pre-trained SWIN-T transforms it into an embedding sequence $V^s = \{v_1^s, v_2^s, ..., v_p^s\}$, where p is the number of patches, $v_i^s \in R^{d_s}$ is the *i-th* patch, and d_s is the dimension of token embeddings.

Multimodal Feature Encoding via CLIP. For a given multimodal news $N = \{T, V\}$, we define the features extracted by CLIP [8] as $N^c = \{t^c, v^c\}$, where $t^c, v^c \in R^{d_c}$ is the bimodal embeddings, and d_c is the vectors dimension.

Background Knowledge Extraction and Encoding. We use publicly available APIs to identify entities with specific meanings in news text and images. In addition, we also extract Baidu Baike descriptions of these entities as their background knowledge and connect them together, such as "Andy Lau" (name entity); "Born in Hong Kong, China, he is a Chinese language actor, pop singer, film producer, and lyricist" (entity description). We also perform the same operation on text embedding of images as a supplement to text entities. For each entity and its background knowledge, we use BERT to extract its embeddings. The entity sets of the extracted text and images are denoted as $TE = \{te_1, ..., te_m\}$, $VE = \{ve_1, ..., ve_n\}$, respectively. Each element in the entity set is a vector of d_b dimension, referring to the encoding of a specific entity and its background knowledge.

2.3 Multi-grained Feature Fusion

We performs cross-modal fusion on two branches with different granularity levels.

Coarse-grained Branch. At coarse-grained level, we construct two independent projection heads, P_T and P_I, consisting of Multi-Layer Perceptrons (MLPs), to process the multimodal features encoded by CLIP. The bimodal feature vectors of the mapping are t_p^c, v_p^c. We achieve cross-modal interaction of coarse-grained branch through similarity loss:

$$\mathcal{L}_{sim} = ylog(1 - M_p) + (1 - y)logM_p \tag{1}$$

where M_p is the similarity between t_p^c and v_p^c, The inherent logic of this loss is that, assuming pure similarity analysis, news articles formed by text and visual information mismatch are more likely to be fake than news articles formed by text statements and image matching. If the sample is real news, t_p^c, v_p^c have higher similarity compared to the original CLIP features. In order to preserve the original features extracted by CLIP, we concatenate t^c, v^c as coarse-grained unimodal feature (without interaction) f_{uni}^c, and use the mapped features t_p^c, v_p^c as coarse-grained cross-modal feature f_{cross}^c.

---

https://cloud.baidu.com/product/nlp_basic/entity_analysis
https://cloud.baidu.com/product/imagerecognition/general

Fine-grained Branch. We designed a **k**nowledge enhanced **C**o-attention **T**ransformer network (KCT) based on Transformer architecture to perform fine-grained interactions. The structure of KCT is shown in Fig. 3, with the goal of highlighting the relevant details between different modalities.

Specifically, we represent the fine-grained inputs of different modalities as I_1 and I_2 , respectively. In our improvement of the self-attention mechanism, we focus on enhancing the model's ability to capture details related to entity background knowledge. Considering that not all entity backgrounds are related to news content, we first use hard attention to filter knowledge information strongly related to news content. We merge the strongly correlated text and visual entity sets into $E_s = \{te_1, ..., te_u, ve_1, ..., ve_w\}$, where u and w represent the number of strongly correlated knowledge in filtered entities set TE_s and VE_s, respectively. In KCT, we firstly calculate the knowledge enhancement matrix M_e based on V^s, T^b, E_s:

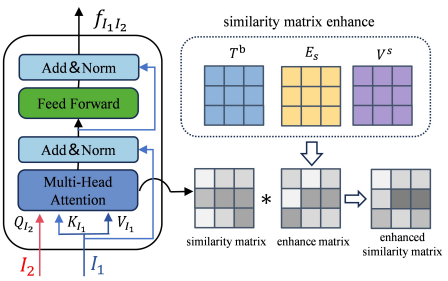

Fig. 3. The architecture of the KCT.

$$M_e = (W^V V^s \cdot W^T E_s) \cdot (W^T T^b \cdot W^T E_s) \tag{2}$$

The values at i and j of the knowledge enhancement matrix M_e represent the affinity between the *i-th* text token and the *j-th* visual token, which is connected by entity knowledge. When two tokens from different modalities have high affinity with a certain entity knowledge at the same time, the value at that position is higher. We define self-attention in KCT as:

$$Attention(Q, K, V) = softmax(\frac{QK^T M_e}{\sqrt{d}})V \tag{3}$$

We input visual features and textual features as I_1 and I_2 into KCT, respectively, and obtain image enhanced text feature f_{vt} and text enhanced image features f_{tv}. We concatenate f_{vt}, f_{tv} as fine-grained multimodal features f^f_{cross}.

However, this approach may to some extent overlook the importance of irrelevant information. To address this problem, we have added TextCnn to fairly capture fine-grained features of text as fine-grained unimodal features f^f_{uni}.

2.4 Improved Mixture-of-Experts (iMoE) Network

At this point, for each granularity-level branch, we have obtained a cross-modal feature f_{cross} and a unimodal feature f_{uni} without interaction. We propose an improved mixture-of-experts network (called iMoE) to fuse features on each branch, as shown in Fig. 2 The objectives of this module include feature fusion,

feature reweighting based on modality correlation (to alleviate ambiguity), and making rough predictions (improve the interpretability of the model). Specifically, we use f_{cross} and f_{uni} as two inputs to the network. Adjust the strength of f_{cross} based on the similarity between modalities. Then, the f_{cross} and f_{uni} are combined as the input of the expert network, that is:

$$f_{fusion} = iMoE(concat(similarity \cdot f_{cross}, f_{uni})) \tag{4}$$

We also added an expert named E_{score} specifically for evaluating the fusion representation of the branch, and add the rating results to the output vector to assist the final decision-making module in evaluating the authenticity of news.

$$f_{out} = concat(score, f_{fusion}) \tag{5}$$

where f_{fusion} is the output of the mixture-of-experts network, *socre* is the output of E_{score}. In this way, for each granularity-level, we obtained representations through feature fusion and reweighting.

2.5 Knowledge Matching

Inspired by [4], for the background knowledge information of news, we also capture knowledge differences between modalities through a comparison based method. Specifically, we first process strongly correlated textual and visual knowledge through soft alignment. Our soft alignment layer defines attention weights as the similarity between knowledge embeddings:

$$\tilde{te}_i = \sum_{j=1}^{w} \frac{exp(s_{ij})}{\sum_{k=1}^{s_{ik}}} ve_j, \forall i \in [1, ..., u] \tag{6}$$

where s_{ij} represents the similarity between te_i and ve_j (calculated by dot multiplication), $\tilde{te}_i$ contains the knowledge semantics related to te_i in visual entities. In order to capture the knowledge differences between modalities, we calculate the difference and integral between $\tilde{te}_i$ and te_i. The difference is used to capture the differences, and the dot product is used to capture the similarities. When concatenated with the original vectors $\tilde{te}_i$ and te_i, we obtain $ec_i \in EC$:

$$ec_i = [te_i, \tilde{te}_i, te_i - \tilde{te}_i, te_i \odot \tilde{te}_i] \tag{7}$$

Finally, perform average pooling and max pooling strategies on EC to generate fixed size representations. And obtain the knowledge vector f^k that is fed into the decision header through the result of the connection pooling operation:

$$f^k = [Avg(EC), Max(EC)] \tag{8}$$

Table 1. Performance comparison between KMMN and other methods on Weibo datasets. The best performance is highlighted in bold.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	MVAE	0.814	0.857	0.772	0.809	0.814	0.878	0.839
	SAFE	0.826	0.829	0.826	0.829	0.816	0.819	0.826
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MCAN	0.896	0.912	0.899	0.914	0.886	0.912	0.898
	FSRU	0.901	<u>0.922</u>	0.892	0.906	0.879	<u>0.913</u>	0.895
	FND-CLIP	0.907	0.915	0.903	0.909	0.916	0.902	<u>0.908</u>
	BMR	<u>0.918</u>	0.882	**0.948**	<u>0.914</u>	**0.942**	0.870	0.904
	KMMN(Ours)	**0.933**	**0.938**	<u>0.924</u>	**0.931**	<u>0.930</u>	**0.941**	**0.934**

2.6 Fake News Classifier

After obtaining the coarse and fine grained multimodal representations f^c, f^f, and background knowledge representation f^k, we connect them together as inputs to the classifier:

$$\hat{y} = FNC([f^c; f^f; f^k]) \tag{9}$$

Fake news detection is a binary classification problem. We apply the cross entropy loss between the ground-truth label y and the final prediction result $\hat{y}$. There is an extra similarity loss $\mathcal{L}_{sim} = \mathcal{L}_{CE}(M_p, y)$. The single branch FND classification loss is the aggregate of $\mathcal{L}_{coarse} = \mathcal{L}_{CE}(score_{coarse}, y)$ and $\mathcal{L}_{fine} = \mathcal{L}_{CE}(score_{fine}, y)$:

$$\mathcal{L}_{signal} = \frac{1}{2}(\mathcal{L}_{coarse} + \mathcal{L}_{fine}) \tag{10}$$

The total loss of KMMN is defined as:

$$\mathcal{L} = \mathcal{L}_{final} + \alpha\mathcal{L}_{signal} + \beta\mathcal{L}_{sim} \tag{11}$$

where α and β are adjustable parameters.

3 Experients

3.1 Experimental Setup

Dataset. We conducted extensive experiments on three real-word datasets collected from social media, Weibo [2] and recently published MR-C and MR-E [1], to evaluate the proposed KMMN. Among them, Weibo and MR-C are chinese datasets, MR-E is an english dataset.

Table 2. Performance comparison between KMMN and other methods on MR^2 datasets.

Method	MR-C				MR-E			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
MVAE	0.821	0.796	0.784	0.789	0.790	0.791	0.781	0.783
FND-CLIP	<u>0.862</u>	0.842	<u>0.885</u>	<u>0.863</u>	0.809	0.828	0.805	0.816
MR^2-RB	0.859	<u>0.852</u>	0.848	0.851	<u>0.856</u>	**0.856**	<u>0.850</u>	<u>0.853</u>
KMMN(Ours)	**0.964**	**0.975**	**0.956**	**0.965**	**0.866**	<u>0.847</u>	**0.885**	**0.861**

3.2 Performance Comparison

Table 1 and Table 2 compares the performance of KMMN with other methods on three datasets. KMMN achieved an accuracy of 93.3% on the Weibo dataset, achieving a significant improvement of 1.5%. On the MR-C and MR-E datasets, the accuracy of KMMN reached 96.4% and 86.6%, which were 10.6% and 1% higher than the best baseline, respectively. In addition, we have the following observations:

Traditional strategies for identifying false information, such as MVAE [3] and MCAN [9], mainly utilize direct connection or attention methods to integrate coarse-grained or fine-grained multimodal features. However, considering the inconsistency between image and visual features, it is difficult to explain how this feature works. In addition, they did not take into account ambiguity issues. When the graphic and textual modalities are inconsistent, cross modal features can easily affect the model's ability to make correct judgments. SAFE [11], CAFE [10] and FND-CLIP [13] both introduce inter modal correlation adjustment to cross modal features. SAFE generates titles instead of the original image, inevitably losing a lot of visual information details.

We inject a large amount of background knowledge into the model from the perspective of entities, enhance the cross-modal interaction process, and provide additional knowledge information, making our model evaluate the authenticity of news closer to people's reading habits. We also evaluate accurate inter-modal correlations and adjust the strength of cross-modal features. Our research findings indicate that these adjustments are crucial for identifying multimodal fake news.

3.3 Ablation Studies

To investigate the effectiveness of each component in KMMN, we explored the impact of key components by evaluating the performance of models with different settings. In each test, we remove different components and train the model from scratch. Specifically, we will simplify the model as follows: **(1) W/o K:** we

Table 3. Ablation study on the architecture design of KMMN on three datasets.

Dataset	Method	Accuracy	Precision	Recall	F1
Weibo	w/o K	0.906	0.886	0.921	0.903
	w/o S	0.925	0.943	0.908	0.925
	w/o C	0.905	0.931	0.883	0.906
	w/o F	0.909	**0.949**	0.877	0.911
	KMMN	**0.933**	0.944	**0.923**	**0.933**
MR-C	w/o K	0.929	0.921	0.939	0.929
	w/o S	0.951	0.951	0.951	0.951
	w/o C	0.939	0.936	0.945	0.939
	w/o F	0.945	0.951	0.941	0.946
	KMMN	**0.964**	**0.975**	**0.956**	**0.965**
MR-C	w/o K	0.836	0.846	0.834	0.835
	w/o S	0.861	**0.904**	0.836	**0.868**
	w/o C	0.827	0.831	0.819	0.825
	w/o F	0.832	0.861	0.809	0.834
	KMMN	**0.866**	0.847	**0.885**	0.861

removed all external background knowledge and only used the basic semantics of the news. **(2) W/o S**: we removed all the correlation measures between modalities and ignored ambiguity problem. **(3) W/o C**: we removed all coarse-grained features and only uses fine-grained features. **(4) W/o F**: we removed all fine-grained features and only uses coarse-grained features.

From the results in Table 3, it can be seen that the performance of all variants is inferior to the original KMMN, which proves the effectiveness of each component. The performance changes of different variants show the same trend on three datasets. In addition, we have the following analysis: (1) The performance of the model that removes all background knowledge has decreased to a certain extent on three datasets, proving that the introduction of high-level knowledge information can provide effective additional references for detecting fake news. (2) Ignoring ambiguity problem on three datasets can lead to performance degradation, indicating that multimodal features are limited by the fuzziness of inter-modal correlations. We can effectively alleviate this by adjusting the strength of cross-modal features. (3) The performance of variants that remove all coarse-grained or fine-grained features has decreased, indicating that features at different granularity levels have made beneficial contributions.

4 Conclusion

In this work, we propose a novel knowledge enhanced multimodal fusion network (KMMN) that enhances cross-modal interaction processes and provides additional knowledge information by introducing background knowledge based on news entities. Extensive experiments conducted on three public benchmarks

for detecting fake news in multiple languages have validated the effectiveness of KMMN. In the future, we plan to further explore how to effectively retrieve and utilize external knowledge in fake news detection.

References

1. Hu, X., Guo, Z., Chen, J., Wen, L., Yu, P.S.: Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval. pp. 2901–2912 (2023)
2. 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)
3. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: Mvae: Multimodal variational autoencoder for fake news detection. In: The world wide web conference. pp. 2915–2921 (2019)
4. Li, P., Sun, X., Yu, H., Tian, Y., Yao, F., Xu, G.: Entity-oriented multi-modal alignment and fusion network for fake news detection. IEEE Transactions on Multimedia **24**, 3455–3468 (2021)
5. Ma, J., Dai, J., Liu, Y., Han, M., Ai, C.: Contrastive learning for rumor detection via fitting beta mixture model. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 4160–4164 (2023)
6. Ma, J., Liu, Y., Han, M., Hu, C., Ju, Z.: Propagation structure fusion for rumor detection based on node-level contrastive learning. IEEE Transactions on Neural Networks and Learning Systems (2023)
7. Ma, J., Liu, Y., Liu, M., Han, M.: Curriculum contrastive learning for fake news detection. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 4309–4313 (2022)
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
9. Wu, Y., Zhan, P., Zhang, Y., Wang, L., Xu, Z.: Multimodal fusion with co-attention networks for fake news detection. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. pp. 2560–2569 (2021)
10. Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., Wei, L.: Detecting fake news by exploring the consistency of multimodal data. Information Processing & Management **58**(5), 102610 (2021)
11. Zhou, X., Wu, J., Zafarani, R.: : Similarity-aware multi-modal fake news detection. In: Pacific-Asia Conference on knowledge discovery and data mining. pp. 354–367. Springer (2020)
12. Zhou, Y., Yang, Y., Ying, Q., Qian, Z., Zhang, X.: Multi-modal fake news detection on social media via multi-grained information fusion. In: Proceedings of the 2023 ACM international conference on multimedia retrieval. pp. 343–352 (2023)
13. Zhou, Y., Yang, Y., Ying, Q., Qian, Z., Zhang, X.: Multimodal fake news detection via clip-guided learning. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 2825–2830. IEEE (2023)