2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00215

An Explainable Multi-view Semantic Fusion Model for Multimodal Fake News Detection

Zhi Zeng[1,2,3,4], Mingmin Wu[1,2,3,4], Guodong Li[5], Xiang Li[1,2,3,4], Zhongqiang Huang[1,2,3,4], Ying Sha[1,2,3,4,*]
[1] *College of Informatics, Huazhong Agricultural University, Wuhan, China*
[2] *Key Laboratory of Smart Farming for Agricultural Animals, Wuhan, China*
[3] *Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan, China*
[4] *Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Urumqi, China*
[5] *Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Wuhan, China*
zengmouren@webmail.hzau.edu.cn, wmm_nlp@webmail.hzau.edu.cn, liguodong22@mails.ucas.ac.cn, teoli112@163.com,
zhoqia.h@gmail.com, shaying@mail.hzau.edu.cn

***Abstract*—The existing models have been achieved great success in capturing and fusing miltimodal semantics of news. However, they paid more attention to the global information, ignoring the interactions of global and local semantics and the inconsistency between different modalities. Therefore, we propose an explainable multi-view semantic fusion model (EMSFM), where we aggregate the important inconsistent semantics from local and global views to compensate the global information. Inspired by various forms of artificial fake news and real news, we summarize four views of multimodal correlation: consistency and inconsistency in the local and global views. Integrating these four views, our EMSFM can interpretatively establish global and local fusion between consistent and inconsistent semantics in multimodal relations for fake news detection. The extensive experimental results show that the EMSFM can improve the performance of multimodal fake news detection and provide a novel paradigm for explainable multi-view semantic fusion.**

***Index Terms*—Multimodal fake news detection, Explainable, Multi-View**

I. Introduction

The development of social media has provided a fertile ground for news creation and dissemination. However, without professional control, the creation and dissemination of fake news have caused severe consequences. For instance, the shares and comments of Top 20 fake election stories have influenced the political orders during the 2016 presidential election in the United States [1]; and the misinformation about the COVID-19 pandemic [2] increased public panic and undermined the credibility of government. Furthermore, fake news creators have developed to add visual information to a news article to create more attractive news to deceive readers. Therefore, it is of importance to detect fake news while encountering multimodal news.

Nowadays, there have been many works focusing on multimodal fake news detection. Various deep-learning models have been proposed to explore the joint representation of modalities for multimodal fake news detection [3]–[5]. In addition, researchers have also focused on fusing different modalities based on interactions of modalities by using the attention mechanism [6]–[8]. Futher, many works [9], [10] require extra information to optimize the fake news detection task by jointly training them. Meanwhile, many works considered consistency [12] in multimodal relations. Some researchers [13] captured article-level and entity-level [14] inconsistency in multimodal news for classification. However, they have neglected the interactions of global and local semantics and inconsistency between different modalities.

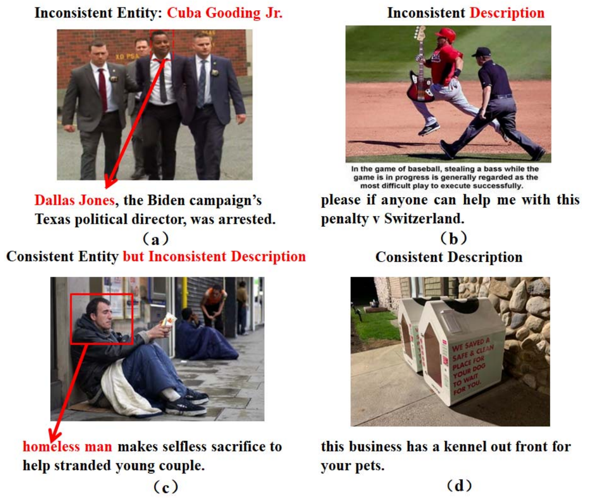

Inconsistent Entity: Cuba Gooding Jr.
Dallas Jones, the Biden campaign's Texas political director, was arrested.
(a)
Inconsistent Description
In the game of baseball, stealing a bass while the game is in progress is generally regarded as the most difficult play to execute successfully.
please if anyone can help me with this penalty v Switzerland.
(b)
Consistent Entity but Inconsistent Description
homeless man makes selfless sacrifice to help stranded young couple.
(c)
Consistent Description
WE SAVED A SAFE & CLEAN PLACE FOR YOUR DOG TO WAIT FOR YOU.
this business has a kennel out front for your pets.
(d)

Fig. 1. Examples of four text-image views

To achieve this, we explore four valuable text-image views in multimodal news, which provide additional dimensions to establish multimodal semantic fusion. 1) **Local Inconsistency**: Visual and textual entities are inconsistent, such as Fig.1(a). [14] The local inconsistent semantics may cheat a model to learn entity-biased information. 2) **Global Inconsistency**: an image and a news article describe a different event, such as Fig.1(b). [20] The global inconsistent semantics may lead optimization of model parameters in opposite directions, resulting in insufficiently training of model. 3) **Local Consistency**: Visual and textual entities are inconsistent, such as Fig.1(c). [20] The local consistent semantics may guide model to learn

[*] is corresponding auther.

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00215

key clues of entites for multimodal news detection. 4) **Global Consistency**: an image and a news article describe a same event, such as Fig.1(d). [20] The global consistent semantics may guide the model learn the useful information.

Therefore, we propose an explainable multi-view semantic fusion model (EMSFM), which interpretatively establishes global and local fusion between consistency and inconsistency of semantics in multimodal relations for fake news detection. Specifically, we use a modal alignment module to align different modalities of news, which can enable the explainable multimodal semantic fusion. Then, we adopt a shared semantic space to learn local lexical distributions of the global semantics from different modalities. Cooperating with the shared semantic space, we conduct a multi-view semantic fusion module to quantify the differences of the consistencies and inconsistencies in global and local semantics for further multimodal fusion. The experimental results on the English and Chinese datasets show that the EMSFM can improve the performance of multimodal fake news detection and provide a novel paradigm for an explainable multi-view semantic fusion. The contributions of this paper can be summarized as follows:

- A new paradigm of explainable multi-view semantic fusion for multimodal news is proposed, which interpretatively establishes global and local fusion between consistency and inconsistency in multimodal relations for fake news detection.
- Our proposed modal alignment module align different modalities of news, which can ground for explainable multimodal semantic fusion.
- We evaluate EMSFM on two large-scale real-world datasets. The results demonstrate that our model outperforms the state-of-the-art models.

II. Related Work

Many deep-learning models have been applied to detect multimodal fake news. The CNN architecture model, like TI-CNN [3], extracted features of texts and images and projected their features into a unified space for classification. With the development of the pre-trained language model, Singhal et al. [5] adopted a pre-trained XLNet and a VGG19 model to extract the visual and textual features to concatenate them for classification. However, simple concatenation or summation operations did not enable the dynamic interaction of multimodal features. Gao et al. [6] used a novel Recurrent Neural Network with an attention mechanism (att-RNN) to fuse multimodal features. The attention mechanism is used to fuse the joint features of text, social context, and images for classification. Kumari et al. [7] proposed ABM-CNN-RNN to use an attention module to explore the interactions between visual and textual features, which were extracted by CNN and RNN models, respectively. Multimodal Co-Attention Networks (MCAN) [8] stacked multiple co-attention layers to better fuse textual and visual features for fake news detection.

Furthermore, Wang et al. [9] introduced event information and derived event-invariant features to optimize the training process of multimodal fake news detection. Khattar et al. [11] proposed a Multimodal Variational Autoencoder (MVAE) to build noise-free multimodal features for classification. Moreover, Xue et al. [12] considers the consistency of multimodal news and captures the general features of social media. Meanwhile, Choi et al. [13] proposed a pre-trained CLIP representation to capture the semantic inconsistency in the multimodal relations, which accurately detected news articles with inconsistent images. Qi et al. [14] proposed a novel entity-enhanced multimodal fusion framework, which exploits three cross-modal correlations to detect fake news.

III. Model

The Model of our proposed EMSFM is shown in Fig. 1. In this section, we will introduce each module in detail.

A. Modal Alignment Module

Considering the huge semantic gap between multimodal news, we employ a modal alignment module [15] to align multimodal news into a unified semantic space by explaining an image with natural language. This module employ an encoder-decoder framework, where the encoder extracts features at different image portions using a convolution neural network (CNN) [16], and the decoder produces different words by focusing on them.

Given a news article and an image pair $\{x, p\}$, we employ an encoder based on a CNN to extract features $R_p = \{r_i\}_{i=1}^{K}$ of p, where r_i denotes the i-th portion feature and K is the number of selected portions.

To understand different portions of an image, we use a decoder, a long short-term memory (LSTM) [17] network, to describe them by different words. The process of implementation of LSTM can be expressed as follow:

$$h_t = \mathrm{LSTM}(y_{t-1} + h_{t-1} + \hat{z}_t) \tag{1}$$

Here, y_{t-1} denotes the $t{-}1$-th word of news article and h_{t-1} denotes $t-1$-th decode state. $\hat{z}_t$ is a context representation, which is a dynamic presentation of the relevant portions of the image at the time t. The calculation process of z_t can be expressed as:

$$\hat{z}_t = \varphi(r_i, \mu_t(i)) \tag{2}$$

$$\mu_t(i) = \frac{\exp(e_t(i))}{\sum\limits_{k=1}^{K} \exp(e_t(k))} \tag{3}$$

where $\varphi(\cdot)$ is a fully-connected layer that connects image portions and similarity weights between image portions and decodes states. $e_t(i)$ denotes the likelihood of the i-th portion whether it correctly produces the next word and $e_t(i)$ is calculated as:

$$e_t(i) = P_{att}(r_i, h_{t-1}) \tag{4}$$

Here, P_{att} is a Portion attention module [15] to calculate the correlation between the hidden states and decode states of images. Based on $\hat{z}_t$, y_{t-1}, and h_t, we employ a probability

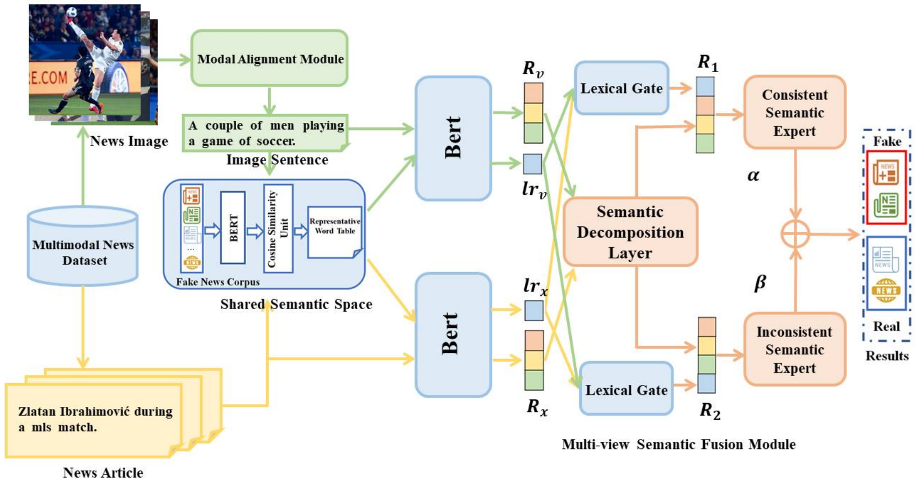


Fig. 2. The framework of EMSFM.

layer to calculate the possible word y_t in a word dictionary [15]:

$$p(y_t|R_p, y_{t-1}) \propto \exp(L_0(Ey_{t-1} + L_h h_t + L_z \hat{z}_t)) \tag{5}$$

where L_0, L_h, L_z, and E are the learnable parameters. The image sentence is $v = \{y_i\}_{i=1}^{M}$, and M is the size of words.

B. Shared Semantic Space

Representative words of fake news are repetitive and high-frequency in fake news, like "Donald Trump" [18]. The Representative words hidden in multimodal sentences can provide a local view to guide our EMSFM to learn fine-grained local information for classification. Considering the multimodal sentences of fake news, we denote them as a fake news corpus $N^f = \{(x^f, v^f)_{i=1}^{Z}\}$ of size Z. Formally, we use a BERT [19] converts N^f into a global presentation set $R_f = \{r_i\}_{i=1}^{P}$ of size P. Meanwhile, we feed all the words of N^f into BERT [19], and obtain the local representation set of them $LR = \{lr_j\}_{j=1}^{J}$, whose element lr_j denotes that j-th local representation, and J is the number of words. In the process of exploring the representative words, a cosine similarity unit can be applied to capture the semantic correlation between fake news and its words. The cosine similarity between r_i and lr_j is calculated as:

$$\cos(r_i, lr_j) = \frac{r_i \cdot lr_j}{|r_i| \times |lr_j|} \tag{6}$$

where the higher cosine similarity value means that lr_j has a larger likelihood to share similar semantics with R_i. Then, we sort cosine similarity in ascending order and find out lr_j of representative words whose cosine similarity ranks top N.

C. Multi-view Semantic Fusion Module

On social media, a news article usually shares an inconsistent image, which can mislead readers. Intuitively, we can employ consistent and inconsistent semantic experts to quantify the different semantics for fake news detection.

Local Semantics Fusion To produce high-quality fused local representations that contain important local lexical distributions appropriately. Note that different representative words specialize in different meanings, and they express distinguishable dimensions of local information. For EMSFM, we would like to select representative words adaptively.

For this, we propose a lexical gate to construct the correlation in local representations adaptively. The output of the lexical gate is a weighting vector γ_i indicating the importance ratio of each representative word. We denote the lexical gate as $LG(;\varphi)$, and φ is the included parameter of γ_i:

$$\gamma_i = LG(\gamma_{i-1}; \varphi)(i = 1, ..., \xi) \tag{7}$$

where γ_0 is the initial lexical weighting vector, W is the representive word table and ξ is the training epoch of $LG(;\varphi)$. The fused representation of representative words of n can be expressed as:

$$F_n = \sum_{i=1}^{T} \gamma_i \cdot W \tag{8}$$

Global Semantics Fusion Different semantic experts have their professional knowledge for detecting fake news in their domain. The multi-view semantic fusion module fuses conclusions of different semantic experts for fake news detection.

Given a multimodal sentence pair $n = \{x, v\}$, we encode them as as R_x, R_v by a BERT model. To capture the consistency and inconsistency of multimodal news, we use a semantic decomposition layer to decompose R_v into the consistent and inconsistent representations as R_1, R_2, respectively. We use a projection vector $Proj(R_v, R_x)$ to denote the relations between R_v and R_x:

$$Proj(h_{v_i}, h_{x_i}) = \frac{h_{v_i} \cdot h_{x_i}}{h_{x_i}} \cdot \frac{h_{x_i}}{|h_{x_i}|} \tag{9}$$

Based on $Proj(R_v, R_x)$, R_1, R_2 are calculated as:

$$R_1 = Proj(R_v, R_x) \tag{10}$$

$$R_2 = R_v - R_1 \tag{11}$$

A “semantic expert” can be expressed as $\delta_i(;\theta_i)(i = 1, 2)$, where δ_1 and δ_2 denote the consistent and inconsistent semantic experts, respectively. θ_1 and θ_2 represent the included parameters of δ_1 and δ_2. For each δ_i, their conclusions can be evaluated as follows:

$$c_i(n) = \delta_i(F_n \oplus R_i; \theta_i) \tag{12}$$

where δ_i is a fully-connected network [20], and $\oplus$ is the concatenation operation. $c_i \in [0, 1]$, a larger c_i value denotes that news n has a larger likelihood to be real in a semantic expert's conclusion; otherwise, n is likely to be fake.

Multi-View Semantics Evaluation Multi-view semantic fusion module fuses the different conclusions from different semantic experts who maximize the detection performance for detecting their specific semantics. Different from existing models, we define joint credibility by considering conclusions c_1 and c_2 of δ_1 and δ_2. Based on the c_1 and c_2, the joint credibility jc_n is expressed as:

$$jc_n = \alpha c_1(n) + \beta c_2(n) \tag{13}$$

Here, $\alpha + \beta = 1 (0 \leq \alpha, \beta \leq 1)$, and a large jc_n value denotes that the n has a high likelihood to be real. The parameters α, β are introduced to balance the conclusions of different experts. The detection loss is calculated as:

$$L_f(\theta) = \sum_{n=1}^{N} y_n \log(jc_n) + (1 - \hat{y}_n)(1 - \log(jc_n)) \tag{14}$$

Here, y_n is the original label and $\hat{y}_n$ is the predicted label. This training procedure enable our EMSFM to learn multi-view semantics for classification.

IV. Experiment

A. Dataset

TABLE I
Statistics of datasets

Datasets	Fakeddit	Weibo
# of fake news	5247	3615
# of real news	3656	4105
# of images	8903	7720

- **Fakeddit [20].** The Fakeddit dataset is an English fake news dataset. It is collected from Reddit, a famous news website containing 22 categories of news. Based on [20], this dataset is verified its effectiveness and diversity, which can strengthen fake news detection.
- **Weibo [9].** This dataset collects real news from authoritative news sources in China, such as Xinhua News Agency. The news was collected between 2012 and 2016 and is verified by Weibo's official evaluation system. The details of the datasets are shown in Table 1.

B. Implementation Details

We split each dataset into the training set and testing set in an 8:2 ratio. For the Fakeddit dataset, we use 2-way classifications for fake news detection. We trained the modal alignment module as same as the model in [15] pre-trained by a MS-COCO dataset. In the EMSFM, we use set as the maximum length of the input news and the Bert-based uncased [19] and Chinese pre-trained BERT with Whole Word Masking [21] for the datasets of Fakeddit and Weibo, respectively. In the shared semantic space, we select $N = 500$ representative words. In the multi-view semantic fusion module, we set Adam as the optimizer, learning rate as 0.001, batch_size as 256, size of the fully-connected layer as 2, and training epochs γ as 30 of δ_i.

C. Performance Comparison

In this research, we conduct comparison experiments to verify the performance of our EMSFM. The comparison performances on the datasets of Fakeddit and Weibo are shown in Table. 2. Their largest values are emphasized in bold.

Among the eleven detection models, the EMSFM achieves the best detection performance. The Accuracy [23], F_1, Precision [22], Recall, and AUC values obtained by the EMSFM are always highest on the datasets of Fakeddit and Weibo. Different detection models are usually designed for different circumstances. In this circumstance of our problem, we exploit the multi-view semantics fusion, which offers us the possibility to make the most of the multimodal semantic fusion. The exploration of multi-view semantics fusion provides an explainable perspective to detect fake news.

D. Ablation Study

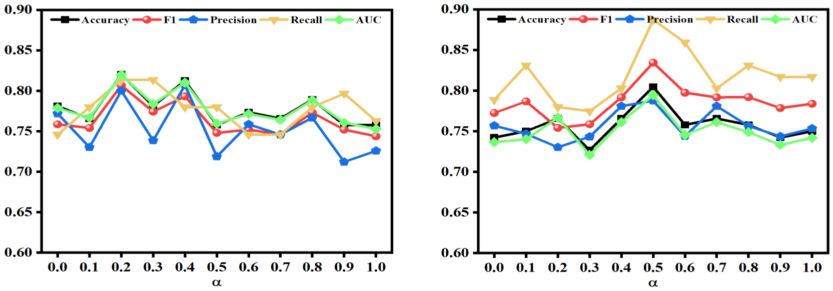
We design five ablation experiments to evaluate the effectiveness of components in EMSFM. The performance comparisons are shown in Table 2.

EMSFM-Text: The EMSFM-Text only uses the news articles as input.
EMSFM-Vis: The EMSFM-Vis only uses the images as input.
EMSFM-VGG: The EMSFM-VGG replaces the modal alignment module with a VGG-19 model [16] in EMSFM.
EMSFM^{s-}: The EMSFM^{s-} removes the shared semantic space in EMSFM.
EMSFM^{m-}: The EMSFM^{m-} removes the multi-view semantic fusion module in EMSFM.

As is seen in Table. 2, our proposed EMSFM achieves best detection performances than those of the five internal models. The experimental results show the multimodal news can capture more effective information than single models. Meanwhile, this is a strong suggestion on the promotional effect of the modal alignment module, shared semantic space, and multi-view semantic fusion module in the multimodal news fusion. Cooperating with the modal alignment module,

TABLE II
EXPERIMENTAL RESULTS OF BASELINES AND THE PROPOSED EMSFM.

	Fakeddit					Weibo				
Method	**Accuracy**	**F$_1$**	**Precision**	**Recall**	**AUC**	**Accuracy**	**F$_1$**	**Precision**	**Recall**	**AUC**
BERT [19]	0.6875	0.6491	0.6727	0.6831	0.7545	0.6719	0.7162	0.6883	0.7465	0.6627
EANN [9]	0.7734	0.7642	0.7344	0.7966	0.7751	0.7656	0.7541	0.7302	0.7797	0.7666
MAVE [11]	0.7818	0.7742	0.7385	0.8113	0.7836	0.7891	0.7731	0.7767	0.7797	0.7884
Spotfake+ [5]	0.7891	0.7731	0.7667	0.7797	0.7884	0.7734	0.7521	0.7586	0.7458	0.7714
MCAN [8]	0.7812	0.7586	0.7719	0.7458	0.7787	0.7656	0.7917	0.7808	0.8028	0.7611
EMSFM-Text	0.7773	0.7349	0.7822	0.6930	0.7690	0.7500	0.7538	0.7778	0.7313	0.7509
EMSFM-Vis	0.6211	0.5446	0.5859	0.5088	0.6100	0.4609	0.1039	0.4000	0.0597	0.4807
EMSFM-VGG	0.7578	0.7350	0.7414	0.7288	0.7577	0.7422	0.7130	0.7321	0.6949	0.7388
EMSFM^{s-}	0.7656	0.7500	0.7377	0.7627	0.7654	0.7656	0.7887	0.7467	0.8358	0.7622
EMSFM^{m-}	0.7578	0.7438	0.7258	0.7627	0.7528	0.7500	0.7838	0.7532	0.8169	0.7418
EMSFM	**0.8203**	**0.8067**	**0.8000**	**0.8136**	**0.8198**	**0.8047**	**0.8344**	**0.7875**	**0.8873**	**0.7945**



(a) Fakeddit

(b) Weibo

Fig. 3. Performance of different α in EMSFM on Fakeddit and Weibo datasets.

the shared lexical space exploits the high-frequency and important lexical information, which is beneficial for fusing fine-grained local semantics. Meanwhile, the multi-view semantic fusion module quantities the differences of the consistencies and inconsistencies of global and local semantics for further multimodal fusion and improve the detection performances.

E. Insight Analysis of Multi-View Semantic Fusion Module

Different α, β describe the difference in evaluating consistent and inconsistent semantic expert's conclusions. The changes of five metrics are shown in Fig. 4.

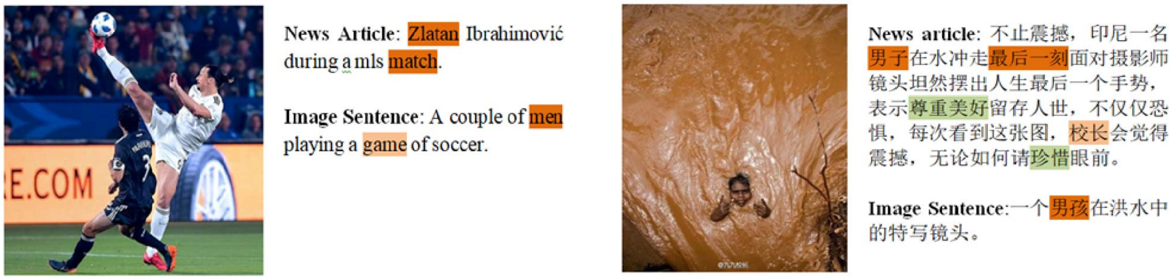
A larger α value indicates that consistent semantic expert plays a more important role in the detection process. Introducing appropriate α is helpful for multimodal fake news detection. Fig. 4 shows that the EMSFM attains optimal performance when $\alpha = 0.2$ and $\alpha = 0.5$ on the Fakeddit and Weibo datasets. This phenomenon can be explained that an appropriate α is helpful for modeling the fine-grained fusion between consistent and inconsistent semantic expert's conclusions, which improves the detection performance.

F. Case Study

To intuitively describe the adequacy of multimodal semantic fusion of EMSFM and the capture of consistent and inconsistent representations between different modalities, we visualize the image sentences, consistent and inconsistent semantics.

1) Visualization of Image Sentences by Modal Alignment Module: To vividly evaluate the superiority of our model in multimodal semantic fusion, we visualize the image sentences by the modal alignment module. Taking two specific samples in Fakeddit and Weibo as two examples, we observe that: the modal alignment module focus on describing the image in an explainable manner (like "A couple of men playing a game of soccer." in Fig. 4(a) and "一个男孩在洪水中的特写镜头。" (meaning: A close-up of a boy in a flood) in Fig. 4(b). Thus, the modal alignment module can provide additional information to enhance the training process. These fully reflect the interpretability of our modal alignment module in the aspect of multimodal semantic fusion.

2) Visualization of Consistency and Inconsistent Seman-tics by Multi-View Semantic Fusion Module: To visually verify the ability of our model to capture consistent and inconsistent semantics, we compare similarity between consistent and inconsistent representation learned by a multi-view semantic fusion module and each word of multimodal sentences. As shown in Fig. 4(a), we found that EMSFM focuses more on consistent entities, like "match", and "men" in news. Meanwhile, our EMSFM not only concentrates on consistent semantics between different modalities "尊重", "珍惜" and "美好"(meanings: "respect", "cherish", "beauty"), but also emphasizes the inconsistent semantics "最后一刻"(meaning: "last time") in Fig. 4(b), which vividly depicts the ability of



News Article: Zlatan Ibrahimović during a mls match.

Image Sentence: A couple of men playing a game of soccer.

(a) Fakeddit

News article: 不止震撼，印尼一名男子在水冲走最后一刻面对摄影师镜头坦然摆出人生最后一个手势，表示尊重美好留存人世，不仅仅恐惧，每次看到这张图，校长会觉得震撼，无论如何请珍惜眼前。

Image Sentence:一个男孩在洪水中的特写镜头。

(b) Weibo

Fig. 4. Visualization of our EMSFM on Fakeddit and Weibo datasets.

EMSFM to explore consistent and inconsistent semantics.

V. Conclusion

In this work, we propose the EMSFM can interpretatively establish global and local fusion between consistent and inconsistent semantics in multimodal relations for fake news detection. Extensive experiments prove our EMSFM can improve the performance of fake news detection and provide a new paradigm of explainable multi-view semantic fusion.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities 2662021JC008, the National Natural Science Foundation of China (No. 62272188), and the National Social Science Foundation of China under Grant 19BSH022.

References

[1] Conroy, Nadia K and Rubin, Victoria L and Chen, Yimin, "Automatic deception detection: Methods for finding fake news," Proceedings of the association for information science and technology, 2015, vol. 52, pp. 1–4.

[2] Zhou, Xinyi.: Fake news early detection: An interdisciplinary study. In: arXiv preprint arXiv:1904.11679, 2019.

[3] Yang, Y., Zheng, L., Zhang, J.: TI-CNN: Convolutional neural networks for fake news detection. In: arXiv preprint arXiv:1806.00749, 2018.

[4] Singhal, Shivangi and Shah, Rajiv Ratn and Chakraborty, Tanmoy and Kumaraguru, Ponnurangam and Satoh, Shin'ichi, Spotfake: A multimodal framework for fake news detection, 2019 IEEE fifth international conference on multimedia big data (BigMM), 2019: 39–47.

[5] Singhal, S., Kabra, A., Sharma, M.: Spotfake+: A multimodal framework for fake news detection via transfer learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13915–13916, 2020.

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

[7] R, Kumari., A, Ekbal.: AMFB: Attention based multimodal Factorized Bilinear Pooling for multimodal Fake News Detection. In: Expert Systems With Applications, vol. 184, pp. 1–12, 2021.

[8] Yang, Wu., Pengwei, Zhan., Yunjian, Zhang., Liming, Wang., Zhen, Xu.: Multimodal semantic fusion with co-attention networks for fake news detection. In Findings of the Association for Computational Linguistics, pp. 2560–2569, 2021.

[9] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Gao, J.: Eann: Event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th acm sigkdd in-ternational conference on knowledge discovery & data mining, pp. 849–857, 2018.

[10] Shah, P., Kobti, Z.: Multimodal fake news detection using a Cultural Algorithm with situational and normative knowledge. In: 2020 IEEE Congress on Evolu-tionary Computation (CEC), pp. 1–7, 2020.

[11] Dhruv, Khattar., Jaipal, Singh.-Goud., Manish, Gupta., Vasudeva, Varma.: Multimodal Variational Autoencoder for Fake News Detection. The World Wide Web Conference, pp. 2915–2921, 2019.

[12] Junxiao, Xue., Yabo, Wang., Yichen, Tian., Yafei, Li., Lei, Shi., Lin, Wei.: Detecting fake news by exploring the consistency of multimodal data. In: Information Processing & Management, vol. 58, no. 5, pp. 102610, 2021.

[13] Choi, H., Yoon, Y., Yoon, S.: How does fake news use a thumbnail? CLIP-based Multimodal Detection on the Unrepresentative News Image. In: arXiv preprint arXiv:2204.05533, 2022.

[14] Qi, Peng and Cao, Juan and Li, Xirong and Liu, Huan and Sheng, Qiang and Mi, Xiaoyue and He, Qin and Lv, Yongbiao and Guo, Chenyang and Yu, Yingchao, Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues,

[15] Xu, K., Ba, J., Kiros, R.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp. 2048–2057, 2015.

[16] Simonyan, Karen and Zisserman, Andrew, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.

[17] Hochreiter, Sepp and Schmidhuber, Jurgen, Long short-term memory, Neural computation, 1997, Vol. 9, pp. 1735–1780.

[18] Zhu, Yongchun and Sheng, Qiang and Cao, Juan and Li, Shuokai and Wang, Danding and Zhuang, Fuzhen, Generalizing to the future: Mitigating entity bias in fake news detection, arXiv preprint arXiv:2204.09484, 2022.

[19] Devlin, J., Chang, M.-W., Lee, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: arXiv preprint arXiv:1810.04805, 2018.

[20] Nakamura, K., Levy, S., Wang, W.-Y.: r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: arXiv preprint arXiv:1911.03854, 2019.

[21] Cui, Y., Che, W., Liu, T.: Pre-training with whole word masking for chinese bert. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3504–3514, 2021.

[22] Zeng Z, Ye L, Liu R, et al. Fake News Detection by Using Common Latent Semantics Matching Method, 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). 2021: 1059-1066.

[23] Zeng Z, Li X, Sha Y. Heterogeneous Propagation Graph Representation Learning for Fake News Detection, 2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). 2022: 456-463.