A Multimodal Fake News Detection Model with Self-supervised Unimodal Label Generation

Yun Liu[1], Zhipeng Wen[1(✉)], Minzhu Jin[1(✉)], Daoxin Fan[1], Sifan Li[1], Bo Liu[2], Jinhe Jiang[3], and Xianda Xiao[3]

[1] School of Cyber Science and Engineering, Liaoning University, Shenyang, Liaoning, China
zhipengwen2023@163.com, mmmzzz0826@foxmail.com
[2] Lianyungang Aids to Navigation Department of Donghai Navigation Safety Admin, Lianyungang, Jiangsu, China
[3] China Tower Corporation Limited Liaoning Branch, Shenyang, Liaoning, China

AQ1 **Abstract.** Fake news detection has become a hot topic. Most multimodal fake news detection models only focus on the semantic correlation between single modalities and often ignore the semantic differences between single modalities, which limited the performance. To deal with the above problem, this paper proposes a multimodal fake news detection model (AFUG), which fully pays attention to the semantic correlation between each modal information by designing a cross-modal fusion module. The self-supervised unimodal label generation model is also added to constrain the overall model optimization. In order to focus on samples with highly differentiated modal information, we design an adaptive weight adjustment strategy to guide the model's learning of unimodal information. Extensive AQ2 experiments on two datasets demonstrate the effectiveness of our AFUG.

Keywords: Fake News Detection · Cross-Modal Fusion · Multimodal Learning · Semantic Correlation

1 Introduction

Nowadays, the development of social media has greatly changed the way of socializing and opinion sharing, which leads to the rise of fake news and seriously endanger the honesty of public discussions and the effectiveness of democratic systems. Therefore, many fake news detection [1] algorithms are proposed. The fake news detection at earlier stage is detected manually [2], which is time consuming and inefficiency. Therefore, the automatic way aroused the public attention, which has become a crucial way in identifying complex and sophisticated fake news narratives.

The automatic fake news detection can be classified into unimodal fake news detection [3] and multimodal fake news detection [4]. The unimodal fake news detection method is proposed based on the analysis of either textual or visual content. However, this approach is becoming progressively insufficient when confronted with the intricate tactics, especially with the multimodal contents. To overcome the limitations of unimodal model, research is shifting towards a multimodal approach for fake news detection, which

D. -S. Huang et al. (Eds.): ICIC 2024, LNCS 14869, pp. 1–12, 2024.
https://doi.org/10.1007/978-981-97-5603-2_11

involves an integrated analysis of various modalities, including text and images. Since each modal has the potential to provide a distinct viewpoint on the news item, and their interaction also can present valuable insights to detect the fake news. The interaction of modalities [5] offers novel perspectives and methodologies, and provided a new way to solve the above challenges.

However, most previous works only focus on the semantic correlation between single modalities and often ignore the semantic differences between single modalities. Besides, there may be deviations between the multimodal true value and the unimodal true value in multimodal fake news detection. For example, the multimodal real label value is true, while the real label value of a single modality or multiple single modalities is false. Therefore, take unimodal fake news detection as a sub-task of multimodal fake news detection tasks will not only help to achieve more accurate determination of the authenticity, but also help to optimize multimodal tasks to achieve a more efficient fake news detection. Motivated by the above motivation, this paper proposes a multimodal fake news detection model (AFUG), which fully pays attention to the semantic correlation between each modal information and a self-supervised unimodal label optimization. This work is of utmost importance, both from a technological standpoint and in terms of its societal implications, as it addresses the problem of misinformation and its harmful effects on public discourse and trust.

The main contributions can be concluded as follows:

1. The combination of self-supervised unimodal label generation and unimodal labels optimization model in the field of multimodal fake news detection is the first time.
2. An adaptive weight adjustment strategy is designed to guide the learning of unimodal information, which can boost the overall performance.
3. The AFUG model we proposed not only focuses on the correlation between cross-modal information through the attention mechanism but also combines the differences between each modal information, which can adaptively reweigh and guide to achieve better multimodal fake news detection.
4. Experimental results conducted on two public datasets prove that AFUG achieves state-of-the-art performance.

2 Related Work

2.1 Unimodal Fake News Detection

The unimodal fake news detection methods are proposed based on the analysis of a single modal, which is rely on either textual or visual content. Gôlo et al. [6] adopt One-Class Learning (OCL) with text-based unimodal representation to the fake news, which is proposed only rely on the linguistic features. Alonso-Bartolome and Segura-Bedmar's research [6] demonstrated the effectiveness of unimodal text analysis approach, which also achieve good result in fake news detection. Besides, Wu et al. [7] considered the automatically excavate semantic features of text and proposed a CED model. Motivated by the above works, Luvembe et al. [8] proposed a fake news detection model based on the attention mechanism and dual emotion features.

However, recently news is usually presented in a multimodal way, including of text and images, the above unimodal methods cannot fully capture the complexity and nuance of multimodal content, which limited their performance.

2.2 Multimodal Fake News Detection

Li et al. [9] proposed a semantic-enhanced multimodal fusion network by considering textual and visual features separately for detecting fake news. Wang et al. [10] proposed a detection model by adopting the graph convolutional network to enhance semantic analysis. However, these early approaches treated modalities as separate entities, and merging their outputs without deeply exploring the intermodal interactions. Considering the relationship between visual modal and textual modal, Yadav et al. [11] built an efficient detection framework by leveraging the strengths of both visual and textual cues. Wu et al. [12] utilized the co-attention mechanisms to facilitate a more cohesive integration of textual and visual features, which achieves good results. Singh et al. [13] further emphasized the importance of combining text and visual analysis in automatic fake news detection area, and highlighted the need for a comprehensive approach to multimodal information processing. Focus on the semantic information between text and images, Zeng et al. [14] demonstrated the context-specific challenges of multimodal fake news detection. Then Xue et al. [15] built a Multimodal Consistency Neural Network (MCNN), which also illustrated the emphasis on ensuring consistency in multimodal data for effective fake news identification. Considering the effectiveness of attention mechanism, Lv et al. [16] proposed a TICCA (Text, Images, Comments Co-Attention) model by fusing images and text features based on a text-image co-attention model. Besides, Zeng et al. [17] built a multimodal inconsistency contrastive learning framework (MMICF) to address global inconsistencies in multimodal semantics.

Overall, effective fake news detection model should not only consider the unique characteristics inherent to each modal but also pay close attention to the semantic difference and interactions between different modalities.

3 Proposed Method

The proposed model is consisted in three parts: feature extraction module, feature fusion module, and the unimodal label generation module, shown in Fig. 1. Our model aims to achieve fine-grained partitioning of tasks through self-supervised unimodal label generation, and back-assisted models learn cross-modality ambiguity information.

3.1 Feature Extraction Module

1) Text Feature Encoding: Text features are the main features of multimodal fake news. We use the pre-trained BERT [18] language model to encode text features. The pre-trained language model can convert text information into a fixed encoding length while retaining rich language information. Specifically, for a text T, the text feature encoding module can be expressed as follows:

$$F_t = \mathrm{BERT}(\{T_1, T_2, \ldots, T_n\}) \tag{1}$$

where n represents the batch size, and F_t represents the text feature of a single batch size sample. $F_t \in R^{n\times 200\times 256}$ means that the sentence length is 200, the word vector length is 256, and the output feature dimension is 3.

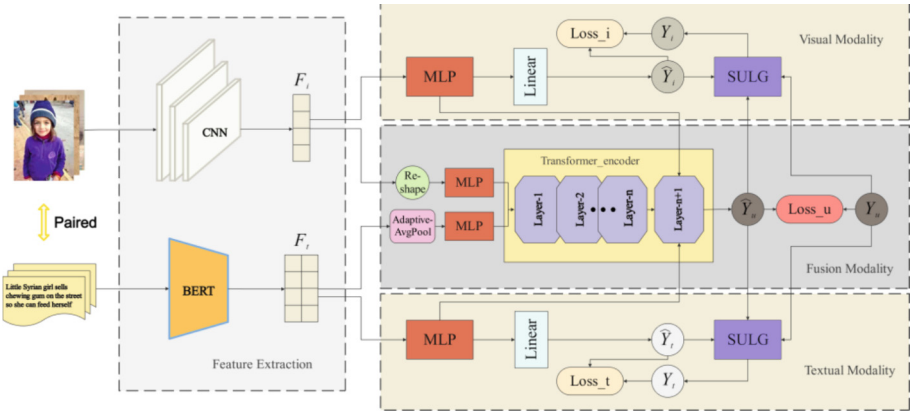

Fig. 1. The flowchart of the proposed multimodal fake news detection model

2) Visual Feature Encoding: In multimodal fake news detection, image also hide a large amount of information. Considering the number of parameters and inference speed of the model, we choose ResNet-34 [19], which has relative advantages in both aspects, as our visual feature encoder. Specifically, for an image I, the visual feature encoding module is calculated as follows:

$$\mathrm{F_i} = \mathrm{ResNet34}(\{\mathrm{I_1}, \mathrm{I_2}, \ldots, \mathrm{I_n}\}) \tag{2}$$

where n represents the batch size, and $\mathrm{F_i}$ represents the visual feature of a single batch size sample. $\mathrm{F_t} \in \mathrm{R}^{\mathrm{n}\times 512}$ means that the length is 512, and the dimension of the output feature is 2.

3.2 Multimodal Feature Fusion

In recent years, the self-attention mechanism has been widely used in various fields due to its ability of capturing long-distance dependencies, including the cross-modal tasks [20]. Considering the possibility of semantic gap between the extracted text features $\mathrm{F_t}$ and image features $\mathrm{F_i}$, we transform them into a shared space to align different modal features. Specifically, we align text features $\mathrm{F_t}$ and image features $\mathrm{F_i}$, and then use the characteristics of the self-attention mechanism to capture the dependency relationship between image features and text features to achieve multimodal feature disambiguation. Text features are represented as follows:

$$\widetilde{F_t} = AdaptiveAvgPool(F_t[0], \frac{F_t[1]}{\alpha_1}, \frac{F_t[2]}{\beta_1}) \tag{3}$$

Where $\widetilde{\mathrm{F_t}} \in \mathrm{R}^{\mathrm{n}\times 16\times 32}$ indicates that the dimension of the output feature is 3. α_1 represents the adaptive parameter in the dimension of the sentence length, and β_1 represents the adaptive parameter in the dimension of the word vector length. $\lfloor * \rfloor$ means rounding down.

The image features are expressed as follows:

$$\widetilde{F}_i = Reshape(F_i[0], \frac{F_i[1]}{\gamma_1}, \frac{F_i[1]}{\gamma_2}) \tag{4}$$

Where $\widetilde{F}_i \in R^{n\times 16\times 32}$ indicates that the dimension of the output feature is 3. γ_1 represents an optional parameter for aligning the dimension of sentence length, and γ_2 represents an optional parameter for aligning the dimension of word vector length. $\gamma_1\gamma_2 = 512$. $\lfloor * \rfloor$ means rounding down. Then, the multimodal fusion features are obtained as follows:

$$F = Concat(\widetilde{F}_t, \widetilde{F}_i) \tag{5}$$

$$Q = W^Q F, K = W^K F, V = W^V F \tag{6}$$

$$F_u = Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

Where W represents the adjustable parameter matrix. The attention function mines the high correlation between visual features and text features, and transforms a set of matrices of the same order with multimodal semantic consistency.

3.3 Unimodal Label Generation

Considering the label const of traditional manual unimodal labeling method, we design a self-supervised parameter-free learning unimodal label generation module (SULG) to achieve dynamic unimodal labels generation, which not only avoids obvious fluctuations in the generated unimodal labels caused by dynamic updates of parameters during training, but also does not increase the number of parameters of the model. Multimodal features F_u, unimodal features F_σ, and multimodal label values y_u are designed as the update basis of unimodal labels. Assuming that the unimodal label is expressed as follows:

$$y_\sigma = SULG(F_u, F_\sigma, y_u) \tag{8}$$

Where $\sigma \in \{t, i\}$. For the two-classification problem, we obtain the positive center (C_ε^p) and negative center (C_ε^n) of different modal representations, as follows:

$$C_\varepsilon^p = \frac{\sum_{j=1}^N G(y_\varepsilon(j) > 0.5) \cdot F_{\varepsilon j}^c}{\sum_{j=1}^N G(y_\varepsilon(j) > 0.5)} \tag{9}$$

$$C_\varepsilon^n = \frac{\sum_{j=1}^N G(y_\varepsilon(j) \le 0.5) \cdot F_{\varepsilon j}^c}{\sum_{j=1}^N G(y_\varepsilon(j) \le 0.5)} \tag{10}$$

Where $\varepsilon \in \{u, t, i\}$. N represents the number of training samples. $G(\cdot)$ represents the indicator function, which determines whether the sample is a positive sample or a negative sample. $F_{\varepsilon j}^c$ is the global representation of the $j-th$ sample. L2 normalization is

then applied to calculate the distance between the modal representation and the modal center, as follows:

$$D_{\varepsilon}^{p} = \frac{\|F_{\varepsilon} - C_{\varepsilon}^{p}\|_2^2}{\sqrt{d_{\varepsilon}}} \tag{11}$$

$$D_{\varepsilon}^{n} = \frac{\|F_{\varepsilon} - C_{\varepsilon}^{n}\|_2^2}{\sqrt{d_{\varepsilon}}} \tag{12}$$

Where $\varepsilon \in \{u, t, i\}$. d_{ε} represents the dimension, which is a scaling factor. In order to avoid excessive difference values and cause abnormal label value ranges, we perform normalization calculation and fix the distance value between [0, 1], as follows:

$$\alpha_{\varepsilon} = D_{\varepsilon}^{p} - D_{\varepsilon}^{n} \tag{13}$$

$$\alpha_{\varepsilon}^{*} = \frac{\alpha_{\varepsilon} - \mathrm{Min}(\alpha_{\varepsilon})}{\mathrm{Max}(\alpha_{\varepsilon}) - \mathrm{Min}(\alpha_{\varepsilon}) + \theta} \tag{14}$$

Where $\varepsilon \in \{u, t, i\}$. θ is a very small number, which is set to $1e{-}8$.

The relationship between unimodal supervision values and multimodal real values is diverse. Here, we only consider the following two ways, as follows:

$$\frac{y_{\sigma}}{y_u} \propto \frac{\hat{y}_{\sigma}}{\hat{y}_u} \propto \frac{\alpha_{\sigma}}{\alpha_u} \rightarrow y_{\sigma} = \frac{\alpha_{\sigma} * y_u}{\alpha_u + \theta} \tag{15}$$

$$y_{\sigma} - y_u \propto \hat{y}_{\sigma} - \hat{y}_u \propto \alpha_{\sigma} - \alpha_u \rightarrow y_{\sigma} = y_u + \alpha_{\sigma} - \alpha_u \tag{16}$$

$$\bar{y}_{\sigma} = \mathrm{Sigmoid}(y_{\sigma}) \tag{17}$$

Where $\sigma \in \{t, i\}$. When y_u is 0, y_{σ} is always equal to 0, and θ is a very small number. We use $\bar{y}_{\sigma}$ as our final unimodal label value.

Since the generation of unimodal label values involves modal features, so the modal label values fluctuate excessively. In order to alleviate this negative impact, we design an iterative update strategy to alleviate this problem, as follows:

$$y_{\sigma}^{(i)} = \begin{cases} \bar{y}_{\sigma} & i = 1 \\ \frac{i-1}{i+1}\bar{y}_{\sigma}^{(i-1)} + \frac{2}{i+1}\bar{y}_{\sigma}^{i} & i > 1 \end{cases} \tag{18}$$

Where $\sigma \in \{t, i\}$. i represents the current epoch number. The unimodal label value will gradually stabilize after several epoch iterations (about 20 in our experiments). This may also be related to the model structure. The self-supervised unimodal label generation algorithm is shown in Algorithm 1.

Algorithm 1.Self-supervised Unimodal Label Generation

Input: unimodal inputs T, I, multimodal label y_u.

Output: unimodal labels $y_t^{(k)}$, $y_i^{(k)}$, where k means the number of training epochs.

1. Initialize model parameters ξ.
2. Initialize unimodal labels $y_t^{(1)} = y_u$, $y_i^{(1)} = y_u$.
3. Initialize global representations $F_t^g = 0$, $F_i^g = 0$, $F_u^g = 0$.
4. For $e \in [1, end]$ do
5. For mini-batch steps, do
6. Compute mini-batch modality representations F_t , F_i , F_u.
7. Compute loss L using Equation (23).
8. Compute parameters gradient $\frac{\partial L}{\partial \xi}$.
9. Update model parameters: $\xi = \xi - \eta \frac{\partial L}{\partial \xi}$.
10. if $e \neq 1$ then
11. Compute relative distance values α_u^* , α_t^* and α_i^* using Equation (9-14).
12. Compute y_t , y_i using Equation (15-17).
13. Update $y_t^{(e)}$, $y_i^{(e)}$ using Equation (18).
14. End if
15. Update global representations F_ε^g using F_ε , where $\varepsilon \in \{u, t, i\}$.
16. End for
17. End for

3.4 Model Optimization and Prediction

The output of the proposed multimodal fake news detection network has two parts, namely unimodal prediction value $\hat{y}_\sigma$ and multimodal prediction value $\hat{y}_m$, as follows:

$$\begin{cases} \hat{y}_\sigma = \mathrm{Sigmoid}(W_\sigma F_\sigma + b_\sigma) \\ \hat{y}_u = \mathrm{Softmax}(W_u F_u + b_u) \end{cases} \tag{19}$$

Where $\sigma \in \{t, i\}$. $W_\sigma \in R^{d\times 1}$ and $b_\sigma \in R^1$ represent the trainable parameters of the unimodal output layer. $W_u \in R^{d\times 2}$ and $b_u \in R^2$ represent the multimodal output layer trainable parameters.

In order to avoid overfitting of unimodal label values and multimodal label values, we use different loss functions to constrain unimodal predicted values and multimodal predicted values respectively. For unimodal label, the optimization method is expressed as follows:

$$\mathrm{Loss}_\sigma = \frac{\sum_{i=1}^{N} \left| \hat{y}_\sigma - y_\sigma \right|}{N} \tag{20}$$

Where $\sigma \in \{t, i\}$. N is the total number of training samples. For the multimodality label, the optimization method is expressed as follows:

$$\mathrm{Loss}_u = -\frac{1}{N} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} y_{n,c} \log(\hat{y}_{n,c}) \tag{21}$$

Where N is the total number of training samples. $y_{n,c}$ is the c − th element of the label vector of sample n, and $\hat{y}_{n,c}$ is the c − th element of the network output result of sample n. C represents the number of categories, which is set to 2 in this paper.

For unimodal tasks, we design the difference between the unimodal label value and the multimodal label value as an adaptive weight, which means that the optimization of the model should pay more attention to samples with obvious difference. The adaptive weight expression is designed as follows:

$$H(y_\sigma, y_u) = \frac{1}{N}\tan|y_\sigma - y_u| \tag{22}$$

The overall optimization method is as follows:

$$Loss = Loss_u + \sum\nolimits_{\sigma}^{\{t,i\}} H(y_\sigma, y_u) \cdot Loss_\sigma \tag{23}$$

The model is considered to have converged when the test metrics have reached a relatively stable state and total losses are observed to be essentially stable.

4 Experimental Analysis

4.1 Experimental Configurations

Dataset: Twitter [21]. It was released for the Media Eval validation multimedia usage task. We maintain the same data segmentation benchmark as in previous works. The training set contains a total of 6840 real news and 5007 fake news. The test set contains a total of 1460 news items.

Weibo [24]. The fake news was collected and disclosed through Weibo's official fake news debunking system from May 2012 to January 2016. In our experiments, we maintain the same data segmentation benchmark as in previous works [5, 20]. The training set contains 3783 real news and 3749 fake news. The test set contains a total of 1996 news items.

Implementation Detail: For the text information of a single sample, we use the pre-trained BERT [18] to extract text features. We follow the same setup as in the work [5], the word vector length is set to 256, and the maximum input length of the text is set to 200 words. For the included images of a single sample, we resize the image to 224 × 224, and use the pre-trained ResNet-34 [19] without the last layer classifier to extract image features. If a single sample contains multiple images, we randomly select one of them as input. In multimodal feature fusion, we align the image feature dimension size to the text feature dimension size. Considering the computational resources, for the input of the Transformer encoding layer [22], the length of the word vector dimension is set to 64 and the heads is 8. In unimodal label generation, we set the generation range of a single label value to [0, 1], where the larger the label value, the higher the credibility of the modality. The batch size is set to 64, and select Adam as the optimizer. The initial learning rate is set to 0.001, and have no weight decay strategy.

4.2 Overall Performance

Table 1 presents the performance comparison of the proposed AFUG model and other fake news detection models on Twitter and Weibo datasets. The overall performance of EANN and RA is relatively poor. The reason is that both of them only focus on unimodal information and ignore the correlation between modalities, which makes the poor overall performance. Because multiple common attention layers are stacked to mine the correlation between modalities, MCAN has relatively excellent performance in precision and F1 score. However, MCAN ignores the characteristics of unimodal and does not use the unimodal label value to reversely constrain the overall prediction value of multimodality during the model optimization process, which limited its performance. Our model not only transforms unimodal features into a shared space with the same semantics to mine the correlation between modalities, but also fully considers the differences between modalities, which can achieve higher performance. Specifically, on the Twitter dataset, compared to the second best value, AFUG achieves a 9.4% improvement in accuracy, and increases the F1 scores for fake news and real news by 8.6% and 8.7%, respectively. In the Weibo dataset, AFUG outperforms the state-of-the-art MCAN model by 1.2% in terms of accuracy, and by 1.3% and 1.1% in F1 scores for fake news and real news, respectively. The model demonstrates a significant enhancement in recall for fake news detection and precision for real news detection. Furthermore, our model exhibits good performance in terms of accuracy for fake news detection and recall for real news, which proves the generalization of our model. Overall, our proposed method can be effectively applied to conduct the fake news detection.

4.3 Ablation Study

In order to prove the effectiveness of each module of our model, we carry out the ablation experiments on Twitter, as shown in Table 2. In each experiment, we only deleted the corresponding module based on the original model for evaluation, and the parameters and content were kept unchanged. 1) w/o TIC: We delete the cross-modal fusion module and only performed simple splicing between unimodal features and output. 2) w/o FLG: We remove the two unimodal branches and only perform cross-modal fusion. 3) w/o ULG: We remove the unimodal label generation algorithm. 4) w/o TIF: We delete the attention-based fusion part of cross-modal fusion and only use simple linear fusion output.

From Table 2, it can be seen that all other models perform worse than the proposed model, which proves the reasonable of combination of each component. In addition, w/o FLG has better performance than w/o TIC, which indicates that cross-modal fusion has an important impact on model performance. Pay attention to the inter-modal correlation may lead to greater performance improvements. The performance of w/o ULG demonstrates that single-modal label generation algorithms can not only generate labels for each modality, but can further exploit the differences between modalities to assist model optimization, which can yield more higher performance. w/o TIF illustrates that the attention mechanism can make the model focus on more important parts, and can further improve the performance of the model.

Table 1. Performance comparison of the proposed AFUG and other fake news detection models. Optimal values are shown in bold.

	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Twitter	EANN [23]	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	RA [24]	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	MKEMN [25]	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE [26]	0.762	0.831	0.724	0.774	0.695	0.811	0.748
	MCNN [15]	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE [5]	0.806	0.807	0.799	0.803	0.805	0.813	0.809
	MCAN [27]	0.809	**0.889**	0.765	0.822	0.732	**0.871**	0.795
	AFUG	**0.903**	0.841	**0.986**	**0.908**	**0.984**	0.823	**0.896**
Weibo	EANN [23]	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	RA [24]	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	MKEMN [25]	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE [26]	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN [15]	0.823	0.858	0.801	0.828	0.787	0.848	0.816
	CAFE [5]	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MCAN [27]	0.899	**0.913**	0.889	0.901	0.884	**0.909**	0.897
	AFUG	**0.911**	0.877	**0.953**	**0.914**	**0.950**	0.870	**0.908**

Table 2. Ablation study results of the proposed AFUG on Twitter. Optimal values are shown in bold

Dataset	Method	Accuracy	F1-score	
			Fake News	Real News
Twitter	w/o TIC	0.866	0.875	0.855
	w/o FLG	0.891	0.899	0.880
	w/o ULG	0.893	0.901	0.883
	w/o TIF	0.894	0.902	0.885
	AFUG	**0.903**	**0.908**	**0.896**

5 Conclusion

In this paper, a multimodal fake news detection method based on self-supervised unimodal label generation is built. Specifically, we use the attention mechanism to design a multimodal information fusion module to mine the correlation between various modalities. In order to pay attention to the differences between modalities, a self-supervised

unimodal label generation module is designed to instantiate the semantic differences presented between each modality and use this difference to reversely constrain the optimization of our model. In order to focus on samples with highly differentiated modal information, we designed an adaptive weight adjustment strategy to guide the learning of unimodal information. Extensive experiments on two public datasets demonstrate that our AFUG can effectively to detect fake news. It needs to mention that our model also has disadvantages. The efficacy of our method is contingent upon the quality of the dataset. An imbalanced distribution of positive and negative samples is detrimental to our self-supervised unimodal label generation method. In the future, we will further study effective model to solve the above problem.

Acknowledgments. This work is supported by Shenyang science and technology plan project under Grant 23-407-3-32, Liaoning Province Natural Science Foundation under Grant 2023-MS-139 and National Natural Science Foundation of China under Grant 61901205.

References

1. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
2. Pathak, A., Srihari, R.K.: BREAKING! Presenting fake news corpus for automated fact checking. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 357–362, July 2019
3. Himdi, H., Weir, G., Assiri, F., Al-Barhamtoshy, H.: Arabic fake news detection based on textual analysis. Arab. J. Sci. Eng. **47**(8), 10453–10469 (2022)
4. Segura-Bedmar, I., Alonso-Bartolome, S.: Multimodal fake news detection. Information **13**(6), 284 (2022)
5. Chen, Y., et al.: Cross-modal ambiguity learning for multimodal fake news detection. In: Proceedings of the ACM Web Conference 2022, pp. 2897–2905, April 2022
6. Gôlo, M., Caravanti, M., Rossi, R., Rezende, S., Nogueira, B., Marcacini, R.: Learning textual representations from multiple modalities to detect fake news through one-class learning. In: Proceedings of the Brazilian Symposium on Multimedia and the Web, pp. 197–204, November 2021
7. Wu, L., Rao, Y., Zhang, C., Zhao, Y., Nazir, A.: Category-controlled encoder-decoder for fake news detection. IEEE Trans. Knowl. Data Eng. (2021)
8. Luvembe, A.M., Li, W., Li, S., Liu, F., Xu, G.: Dual emotion based fake news detection: a deep attention-weight update approach. Inf. Process. Manag. **60**(4), 103354 (2023)
9. Li, D., Guo, H., Wang, Z., Zheng, Z.: Unsupervised fake news detection based on autoencoder. IEEE Access **9**, 29356–29365 (2021)
10. Wang, Y., Qian, S., Hu, J., Fang, Q., Xu, C.: Fake news detection via knowledge-driven multimodal graph convolutional networks. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 540–547, June 2020
11. Yadav, A., Gaba, S., Khan, H., Budhiraja, I., Singh, A., Singh, K.K.: ETMA: efficient transformer-based multilevel attention framework for multimodal fake news detection. IEEE Trans. Comput. Soc. Syst. (2023)
12. 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, August 2021

13. Singh, V.K., Ghosh, I., Sonagara, D.: Detecting fake news stories via multimodal analysis. J. Am. Soc. Inf. Sci. **72**(1), 3–17 (2021)
14. Zeng, J., Zhang, Y., Ma, X.: Fake news detection for epidemic emergencies via deep correlations between text and images. Sustain. Cities Soc. **66**, 102652 (2021)
15. Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., Wei, L.: Detecting fake news by exploring the consistency of multimodal data. Inf. Process. Manag. **58**(5), 102610 (2021)
16. Lv, L., Liu, L.: TICCA-A co-attention network for multimodal fake news detection. In: 2023 4th International Conference on Computer Engineering and Application (ICCEA), pp. 212–215. IEEE, April 2023
17. Zeng, Z., Wu, M., Li, G., Li, X., Huang, Z., Sha, Y.: Correcting the bias: mitigating multimodal inconsistency contrastive learning for multimodal fake news detection. In 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 2861–2866. IEEE, July 2023
18. Devlin, J, Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Associationfor Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics (2019)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, 770–778 (2016)
20. Wang, L., Zhang, C., Xu, H., et al.: Cross-modal contrastive learning for multimodal fake news detection. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5696–5704 (2023)
21. Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., Kompatsiaris, Y.: Detection and visualization of misleading content on Twitter. Int. J. Multimed. Inf. Retr. **7**(1), 71–86 (2018)
22. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
23. 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. ACM (2018)
24. 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. ACM (2017)
25. 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)
26. Zhou, X., Wu, J., Zafarani, R.: SAFE:similarity-aware multi-modal fake news detection. In: Lauw, H., Wong, R.W., Ntoulas, A., Lim, E.P., Ng, S.K., Pan, S. (eds.) PAKDD 2020. LNCS, vol. 12085, pp. 354–367. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-47436-2_27
27. 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)