Construction and Performance Optimization of a Multimodal Transformer-Based Fake News Detection Model

Qifan Huang[1*]

[1] Guangzhou Xinhua University, Guangzhou, 510520, China

* huangqifan@pku.org.cn

https://doi.org/10.70695/IAAI202601A7

Abstract

To address the issues of current false news characterized by multimodal fusion dissemination and the limited interpretability of existing methods, we explore a multimodal transformer detection model (SEL-MSIT) that integrates supervised contrastive learning and multi-stage interaction. Based on feature extraction, the supervised contrastive learning module enhances feature discrimination capability, while the multi-scale cross-modal interaction module is constructed to explore deep correlations. A consistent attention mechanism is employed to achieve efficient feature fusion. Experimental results demonstrate that SEL-MSIT outperforms mainstream baselines in terms of accuracy, precision, recall, and F1-score. Ablation experiments are conducted to verify the effectiveness of each optimization module, and the results can serve as supplementary data for decision-making.

Keywords Transformer; False News; Detection; Cross Modal Interaction; Feature Fusion

1 Introduction

With the development of social media and generative AI technology, false news is becoming more and more common in the world. Unlike traditional false news, which is mostly text-based, modern false news is mostly multimodal, and the falsity of it is enhanced through forging of pictures, tampering of title, grafting of context and so on [1-2]. As for now, the single-mode detection method depends on getting the usual features of false information, it is not easy to do thorough detection for cross modal information [3-4]. Cross-modal analysis, text-semantic abstraction and image-visual presentation cause misalignment and inaccurate detection. Scene variability creates further problems for robust identifcation.

Currently existed multimodal detection technologies can be divided into pre-fusion, post-fusion and deep-fusion [5]. The early fusion method is that we can directly identify the information we want to identify. It does not analyze the internal information. It is like integrating early. Late fusion means we get to know all kinds of information, we can make complete decisions. Features we identify and extract are also diverse. Deep fusion method adopts transformer as the main body of the network and realizes modal interaction by using cross attention mechanism [6].

On the basis of the three categories of multimodal detection methods, the Raemollm framework adds the emotional embedding features to enhance the detection accuracy in cross domain in terms of feature improvement. In terms of data improvement, the rumorllm resolves the issue of category imbalance by fine-tuning the large language model to produce different samples.

It is found through a comprehensive analysis that most of the cross-modal interactions occur at a single stage, there is a lack of full hierarchical correlation, there is no supervision signal guidance for improving feature discrimination, and the features of similar samples are scattered, and the efficient utilization of features in a low-resource scenario is not very complete. Therefore, we must build a detection model with good feature discrimination, good modal adaptability, and good scene robustness to detect false news as soon as possible.

2 Related Theoretical Foundations

2.1 Multimodal Transformer Architecture

Multimodal Transformer is based on the self-attention mechanism and it can process bimodal data using different text encoders and image encoders, and then use cross-modal attention layers for interaction. And the big benefit is being able to learn really long dependecies between those two modalities. Equation (1) gives the calculation way of cross-modal attention:

$$Attention(Q^T,K^I,V^I)=softmax(\frac{Q^T(K^I)^T}{\sqrt{d_k}})V^I \quad (1)$$

In the formula, Q^T is the text query matrix, $K^I$and V^I are the image key matrix and value matrix, respe ctively, and d_k is the feature dimension. This mechanism enables text features to focus on key regions in images and vice versa.

2.2 Supervised Contrastive Learning

Supervised Contrastive Learning (SCL) constructs supervised signals by introducing class labels, narrowing the feature distance of similar samples and pushing away that of dissimilar samples. Its loss function is shown in Equation (2):

$$LSCL=-\frac{1}{N}\sum i=1^{N}\log\frac{\sum_{j\in P(i)}\exp(sim(z_i,z_j)/\tau)}{\sum_{j\in P(i)\cup N(i)}\exp(sim(z_i,z_j)/\tau)} \quad (2)$$

In the formula, N is the number of samples, P(i) and N(i) are the positive and negative sample sets of the i-th sample, respectively, sim(·,·) is the cosine similarity, and τ is the temperature parameter. Applyi ng SCL to multimodal feature learning can simultaneously optimize the feature distribution within text, within images, and across modalities.

2.3 Cross-Modal Consistency Evaluation

Cross-modality consistency evaluation judges the meaning consistency of text and picture, often employing similarity and logic. This paper uses a two-dimensional evaluation of "feature similarity + content consistency", where the former measures feature matching degrees by cosine similarity, and the latter evaluates the logical relevance between text description and image content by large language models (LLMs) to provide weight basis for attention fusion.

3 Model Construction and Optimization

3.1 Overall Architecture Design

The SCL-MSIT model consists of four modules: a bimodal feature encoding module, a supervised contrastive learning module, a multi-stage cross-modal interaction module, and a consistency attention fusion module.

3.2 Core Module Implementation

Bimodal Feature Encoding Module

The text coding adopts the BERT base model, and the input sequence length is set to 512. In order to adapt to low-resource scenarios, LLM is introduced to generate image descriptions to supplement text features, and the Llama2-7b model is used to caption the image to generate visual descriptions.

Supervised Contrastive Learning Module

The module provides a comparison method for bimodal features, which includes the comparison of different news text features for the same news category, the comparison of different news image features for the same news category, the comparison of the same news text and image features, and the cross-modal feature comparison of different news. The classification loss function is optimized by combining

the weighted sum of the loss functions in different ways under the condition of equal weight, in order to carry out comparative analysis of characteristics.

Multi-stage Cross-modal Interaction Module

First, the local interaction phase. The local feature alignment of text and image is realized through the cross modal attention layer;

The cross modal graph structure is constructed. The text semantic unit and image visual unit are used as nodes, and the edge weight is used as feature similarity. The global association features are learned th rough graph convolution network (GCN).:

$$H^{(l+1)}=\sigma(\widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}H^{(l)}W^{(l)}) \quad (3)$$

In the formula, $\widetilde{A}$ is the adjacency matrix, $\widetilde{D}$ is the degree matrix, $H^{(l)}$ is the feature of the l-th layer, $W^{(l)}$ is the trainable parameter, and σ is the activation function.

Consistency Attention Fusion Module

Dynamically allocate feature weights based on the results of cross modal consistency assessment. The steps are as follows:

Step1：Calculate the cosine similarity S_{sim} between text features F_T and image features F_I;

Step2：Invoke GPT-3.5 to judge bimodal content consistency, outputting a confidence score S_{con}(ran ging from 0 to 1);

Step3：Fusion weight calculation: w_T=0.5S_{sim}+0.5S_{con},w_I=1-w_T;

Step4：Generate final features: F_{final}=w_TF_T+w_IF_I.

3.3 Model Optimization Strategies

For the small-sample situation in the experiment, first, the small amount of rumorllm data is augmented, and then the llama2 model is fine-tuned using LoRA to generate false news that resembles the style of real samples. At the same time, the parameters of the model trained on the dataset are transferred to the Chinese low-resource scenario.

4 Experimental Design and Result Analysis

4.1 Experimental Environment and Datasets

Experimental Environment

The hardware environment is NVIDIA A100 GPU (40Gb video memory), Intel Xeon 8375c CPU and 128GB memory; The software environment is vs and transformers 4.34.0.

Experimental Datasets

Three benchmark datasets and one low-resource dataset are selected, with detailed information shown in Table 1:

Table 1. Three benchmark datasets and one low-resource dataset

Dataset	Languag e	Modal Compositio n	Class Distribution	Data Source
Weibo	Chinese	Text - Image	Authentic 32% / Fake 45% / Misleading 23%	Social media crawling
Twitter	English	Text - Image	Authentic 41% / Fake 59%	PHEME project
GossipCop	English	Text - Image	Authentic 63% / Fake 37%	Entertainment news fact-checking platfor m
Tamil-LowRe s	Tamil	Text - Image	Authentic 35% / Fake 40% / Misleading 25%	Language-specific platform dataset

Among them, Tamil-LowRes is a low-resource dataset used to verify the model's adaptability to low-resource language scenarios. All datasets include manually annotated class labels and fact-checking reports.

Evaluation Metrics

Four commonly used metrics are adopted: Accuracy (Acc), Precision (P), Recall (R), and F1-Score (F1), with F1-Score as the core evaluation metric. The formula for F1-Score is:

$$F1=2\times\frac{P\times R}{P+R} \tag{4}$$

To address class imbalance, the weighted F1-score (wF1) is additionally used for supplementary evaluation.

4.2 Comparative Experiment Design

The text of Bert, the image of Vit, MMCNN, CNIFFN and BVT-CNN methods are selected as comparative references for in-depth analysis.

4.3 Experimental Results and Analysis

Multi-dataset Performance Comparison

Table 2. the performance of each model on the three benchmark datasets

Model	Dataset	Acc (%)	P (%)	R (%)	F1 (%)	wF1 (%)
BERT	Weibo	78.2	77.5	76.9	77.2	76.8
ViT	Weibo	75.4	74.8	73.6	74.2	73.9
MMCNN	Weibo	81.3	80.7	80.1	80.4	80.0
MVCNN	Weibo	83.5	82.9	82.4	82.6	82.3
CMIFFN	Weibo	86.9	87.1	85.7	86.4	86.1
BVT-CNN	Weibo	87.2	87.5	86.2	86.8	86.5
SCL-MSIT (Ours)	Weibo	89.1	88.9	88.5	88.7	88.4
BERT	Twitter	76.8	75.9	75.2	75.5	75.1
ViT	Twitter	73.2	72.5	71.8	72.1	71.7
MMCNN	Twitter	79.6	78.9	78.3	78.6	78.2
MVCNN	Twitter	82.1	81.5	80.9	81.2	80.8
CMIFFN	Twitter	85.3	85.6	84.1	84.8	84.5
BVT-CNN	Twitter	85.7	86.0	84.5	85.2	84.9
SCL-MSIT (Ours)	Twitter	87.6	87.8	87.3	87.5	87.2
BERT	GossipCop	82.4	83.1	81.8	82.4	82.0
ViT	GossipCop	79.8	80.5	79.2	79.8	79.4
MMCNN	GossipCop	84.7	85.3	84.1	84.7	84.3
MVCNN	GossipCop	86.2	86.8	85.6	86.2	85.8
CMIFFN	GossipCop	88.5	88.9	87.6	88.2	87.9
BVT-CNN	GossipCop	88.9	89.3	88.0	88.6	88.3
SCL-MSIT (Ours)	GossipCop	90.3	90.6	89.9	90.2	89.9

From Table 2 we can see that the SCL-MSIT model is still better, which proves that the supervised comparative learning and the multi-stage interaction module are effective and can effectively detect the cross-modal deception mode.

Compared with the single mode method and the multi mode method, it can be found that the performance of the multi mode method is better than the single mode method, so it can be seen that the fusion of dual mode information is very important for the detection performance. Deep fusion methodis far better than all others which shows the superiority of transformer architecture on cross-modal modeling.

Low-Resource Scenario Performance Comparison

Table 3. the performance of each model on the Tamil-LowRes low-resource dataset

Model	Acc (%)	P (%)	R (%)	F1 (%)	wF1 (%)
BERT	65.3	64.8	63.7	64.2	63.8
ViT	62.1	61.5	60.4	60.9	60.5
MMCNN	68.7	68.2	67.1	67.6	67.2
MVCNN	71.2	70.7	69.6	70.1	69.7
CMIFFN	75.8	76.2	74.9	75.5	75.1
BVT-CNN	76.3	76.7	75.4	76.0	75.6
SCL-MSIT (Ours)	82.1	82.5	81.7	82.1	81.7

In different scenarios, the advantages of scl-msit model are more significant, which can solve the lack of semantic information caused by data scarcity, and make the model still maintain stable performance in small sample scenarios.

Ablation Experiment Results

To verify the effectiveness of each core module, ablation experiments are designed by sequentially removing the supervised contrastive learning (SCL), multi-stage interaction (MSI), and consistency attention (CA) modules from the SCL-MSIT model. The results are shown in Table 4.

Table 4. Ablation experiments

Model Variant	Dataset	F1 (%)	Performance Drop (ΔF1)
Full Model	Weibo	88.7	-
Without SCL	Weibo	85.2	3.5
Without MSI	Weibo	84.8	3.9
Without CA	Weibo	86.1	2.6
Full Model	Tamil-LowRes	82.1	-
Without SCL	Tamil-LowRes	77.3	4.8
Without MSI	Tamil-LowRes	76.9	5.2
Without CA	Tamil-LowRes	79.2	2.9

Ablation experiments show that all core modules contribute positively to the performance, and removing the multi-stage interaction modules leads to the biggest performance drop. Supervised contrastive learning is more important in the low-resource case, which proves that the optimized feature distribution is effective, and the fusion quality can be further improved by dynamic weight allocation.

In terms of efficiency, after optimizing the SCL-MSIT model, the inference time has been optimized and can meet the requirement of real-time detection. In terms of interpretability, the proposed model is much better than the CMIFFN model.

5 Conclusion

This paper proposes a multimodal transformer false news detection model called SCL-MSIT, which integrates supervised contrastive learning and multi-stage interaction. It solves the problems of cross-modal feature alignment, low-resource adaptation, and lack of interpretability through bimodal feature encoding, supervised contrastive optimization, multi-stage cross-modal interaction, and consistent

attention fusion. From the experiments, we can see that the model has good scene adaptability and its inference efficiency is enhanced.

Acknowledgement

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Ji, L., Wang, Y., Ren, H., et al. (2026). A local dynamic propagation graph-based method for multi-modal fake news detection. Journal of Intelligent Information Systems, Advance online publication.
2. Chen, S. (2026). A multimodal fake news detection method combining adaptive graph convolutional networks. Automatic Control and Computer Sciences, 59(6), 771-782.
3. Habib, A. M., Wadud, H. A. M., Mridha, F. M., et al. (2026). LLM-powered multimodal reasoning for fake news detection. Computers, Materials & Continua, 87(1).
4. Kumar, P. B., Tamilarasi, K., & Thilagavathy, A. (2025). Metaheuristic-assisted deep learning model for fake news detection. Journal of Experimental & Theoretical Artificial Intelligence, 37(8), 1481-1500.
5. Lv, J., Gao, Y., Li, L., et al. (2025). Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. Journal of King Saud University Computer and Information Sciences, 37(9), 306-306.
6. Hashemi, A., Moosavi, R. M., Shi, W., et al. (2026). Enhancing fake news detection through estimating user tendencies to spread fake news. Data and Information Management, 10(2), 100115.

Biographies

1. **Qifan Huang** Associate Professor, Postdoctoral Fellow in Management, University of Cambridge; Senior Research Fellow, Sun Yat-sen University; Member, Dongguan Social Science Federation; Specially Appointed Researcher, Chinese Academy of Social Sciences; B.A., Fudan University.

基於多模態 Transformer 的虛假新聞檢測模型構建與性能優化

黃淇梵[1]

[1]廣州新華學院，廣州，中國，510520

摘要：針對當前虛假新聞多模態融合傳播下解釋性不足的問題，本文提出一種融合監督對比學習與多階段交互機制的多模態 Transformer 檢測模型（SCL-MSIT）。研究在特徵提取的基礎上，通過監督對比學習增強特徵判別能力，分析多尺度跨模態交互性以挖掘模態間深層關聯，採用一致性注意力機制實現高效特徵融合。實驗結果表明，SCL-MSIT 在準確率、精確率、召回率及 F1 值等指標上均優於傳統模型，驗證了模型的有效性，可爲虛假新聞的傳播管理提供科學判斷。

關鍵詞：Transformer架構；虛假新聞；檢測；跨模態交互；特徵融合

1. 黃淇梵，博士，副教授，劍橋大學管理學博士後，中山大學高級研究學者，本科復旦大學，東莞社科聯，社科院特聘研究員。