Enhancing Fake News Detection through Consistency Contrastive Learning with MLP-Mixer

Shaodong Cui[1] Wen Ma[1] Hiroyuki Shinnou[1]
[1]Faculty of Engineering, Ibaraki University, Hitachi, Ibaraki, Japan
22nd312g@vc.ibaraki.ac.jp 19ND302H@vc.ibaraki.ac.jp
hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

Abstract

Detecting and filtering false information has become a critical area of academic research with the rapid spread of multimodal fake news on major social media platforms. However, effectively integrating diverse feature types for reliable fake news detection remains challenging. To address this, we propose a novel fake news detection model based on consistency contrastive learning. Our model uses an MLP-mixer to extract features, and consistency contrastive learning to measure the semantic distance between text features and text attribute features. This approach enhances the MLP-mixer's ability to extract consistent high-level features. Experimental results on the LIAR dataset demonstrate that our proposed model outperforms existing methods in detecting fake news.

1 Introduction

The rapid development of the Internet has changed the way we obtain information. However, it has also facilitated the spread of fake news. To detect fake news, many researchers use machine learning to classify fake news. For example, some studies introduce recurrent neural networks into the detection process to better understand the time series characteristics in the text [1]. With the continuous development of fake news detection technology, attention mechanisms focus on fake news features [2]. Ran et al. [3] proposed an end-to-end multi-channel graph attention network, which constructs three sub-graphs in parallel to learn the semantic information of news propagation structure for rumor detection. Although the above studies have achieved good results in fake news detection, there is a lack of research on the consistency between news text and text attributes.

In the multimodal field, contrastive learning has been shown to effectively enhance the multimodal joint feature representation that integrates text and image information. Jia et al. [4] used contrastive learning loss to train a model that merged matching text-image pairs and separated mismatched text-image pairs to align image and text representations. Li et al. [5] proposed a contrastive loss to calculate the similarity of image and text feature representations and dynamically construct negative samples to align multimodal representations. However, current research only considers contrastive learning between different modalities. Inspired by this, we adopt Consistency Contrastive Learning to consider the consistent representation learning between text and text attributes to extract more consistent features. We use MLP-Mixer to replace the convolutional network and attention mechanism to achieve better performance. The main contributions of this work are as follows:
- We use MLP-mixer to extract text features and text attribute features to replace the attention model to capture long-distance dependencies.
- We use consistency contrast loss to shorten the semantic distance between text and text attributes, thereby extracting more consistent high-level features.

2 Proposed Method

Figure 1 shows the proposed model framework. We divide the LIAR dataset into three types of data: news text, text attributes, and numerical data. MLP-mixer is used to extract news text features and text attribute features. LSTM is used to process numerical features. Consistency Contrastive Learning shortens the semantic similarity between news text and text attributes to improve consistency. Cancat concatenates the three features together and feeds them into the classifier.

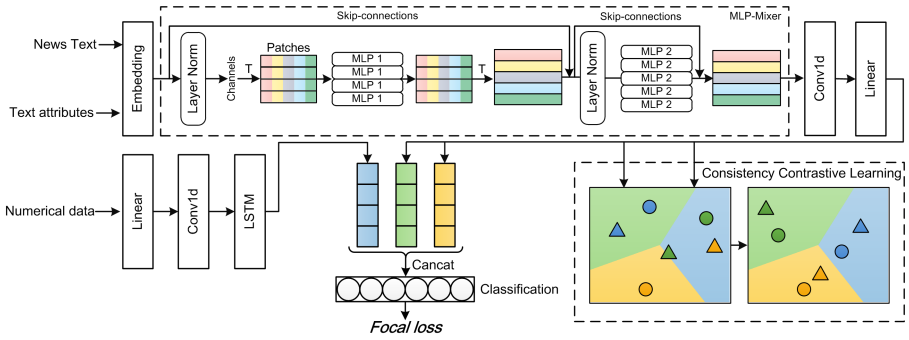

Figure 1 The architecture of our model. The circle represents the news text, and the triangle represents the news attribute. Shapes of different colors represent different samples. Consistency Contrastive Learning aims to shorten the similarity distance between different representations of the same sample.

2.1 MLP-mixer

MLP-Mixer [6] is a pure MLP neural network developed by the Google research team in 2021. It was initially used in image classification tasks in the CV field. MLP-Mixer is mainly composed of token-mixing MLP and channel-mixing MLP. Token-mixing is responsible for the information exchange of spatial positions, and channel-mixing is responsible for the information exchange of feature channels. MLP-mixer takes multiple text tensor patches as input, and each patch is projected onto the hidden dimension C.

Token-mixing is first performed on each patch to fuse adjacent values within each patch and mix spatial information, then channel-mixing is performed. Token MLP acts on the columns of T to fuse spatial information at different positions. Channel MLP acts on the rows of T to fuse positional feature information of different channels. In addition, Mixer also draws on the idea of the residual structure in ResNet, uses skip-connection to add the input and output, and uses Layer Norm before the fully connected layer. A single MLP consists of two fully connected layers sandwiched by a GELU activation function. MLP-mixer can be represented mathematically as:

$$\mathbf{U}_{*,i} = \mathbf{X}_{*,i} + \mathbf{W}_2 \sigma \left(\mathbf{W}_1 \mathrm{LN}(\mathbf{X})_{*,i} \right), \quad \text{for} i = 1 \ldots C \quad (1)$$

$$\mathbf{Y}_{j,*} = \mathbf{U}_{j,*} + \mathbf{W}_4 \sigma \left(\mathbf{W}_3 \mathrm{LN}(\mathbf{U})_{j,*} \right), \quad \text{for} j = 1 \ldots S \quad (2)$$

where X is the input feature of the Mixer layer. LN is the layer normalization operation. σ is the activation function. W_1, W_2 and W_3, W_4 are the weight parameters of the two fully connected layers in the MLP1 and MLP2 modules respectively.

2.2 Consistency Contrastive Learning

We introduce a consistency contrast loss in the network to reduce the semantic distance between similar parts of text and text attributes in the feature space and achieve consistent representation learning. Considering that text and text attributes from the same sample have similar semantic representations, consistent contrast learning is introduced to help MLP-mixer extract more consistent high-level features.

We measure the similarity of text and text attributes by transforming them into cosine space. First, mark all instances: anchor instance $A = c_i^{(t)}$, positive instance $A^+ = c_i^{(v)}\Big|_{t \neq v}$, negative instance $A^- = c_j^{(t)}\Big|_{j \neq i}$. $c_i^{(t)} \in \mathbb{R}^d$ is called the instance of sample i in text t. $c_i^{(v)} \in \mathbb{R}^d$ is called the instance of sample i in text attribute v. The positive instances and negative instances here are relative to the anchor instance. It should be noted that each instance can be selected as an anchor instance, and the anchor instance is combined with the positive instance or the negative instance. We can get $n-1$ positive instance pairs and $n \times l - n$ negative instance pairs.

We aim to minimize the distance between available positive instance pairs and maximize the distance between available negative instance pairs in the feature space. Cosine similarity is used to measure the distance between instance pairs, which can be represented mathematically as:

$$Similarity(c_i^{(t)}, c_j^{(v)}) = \frac{\langle c_i^{(t)} \cdot c_j^{(v)} \rangle}{||c_i^{(t)}|| \cdot ||c_j^{(v)}||} \quad (3)$$

where $\langle \cdot \rangle$ is the dot product operation. Our optimization goal is $\mathcal{S}(A, A^+) \gg \mathcal{S}(A, A^-)$.It maximizes the consistency between text and text attributes, and the cosine simi-

larity between the anchor instance and the positive instance is much greater than the cosine similarity between the anchor instance and the negative instance.

The instance-level contrastive loss for all text and text attributes can be mathematically expressed as:

$$L_{CL} = -\frac{1}{2n}\sum_{v=1}^{l}\sum_{t\neq v}\sum_{i=1}^{n} \log \frac{e^{S\left(z_i^{(v)}, z_i^{(t)}\right)/\tau}}{e^{S\left(z_i^{(v)}, z_i^{(t)}\right)/\tau} + \sum_{r=t,v}\sum_{\substack{j=1,\\ j\neq i}} e^{S\left(z_i^{(v)}, z_j^{(r)}\right)/\tau}} \tag{4}$$

where τ is the temperature parameter, indicating the distribution concentration degree [7]. The temperature parameter τ controls the weights of hard and soft labels in the loss function.

In many practical applications, such as text classification, class imbalance is a common problem. Traditional cross entropy loss may cause the model to over-focus on negative and ignore positive samples. Focal loss [8] adds a coefficient factor based on the standard cross entropy loss, thereby weakening the learning of easy samples and the learning of difficult samples, thereby improving the classification ability of the model. Focal loss can be represented mathematically as:

$$\begin{aligned} L_{FL} &= -(1-p_t)^{\gamma}\log(p_t) \\ &= \begin{cases} -(1-p)^{\gamma}\log(p), y=1 \\ -p^{\gamma}\log(1-p), y=0 \end{cases} \\ &= -y(1-p)^{\gamma}\log(p) - (1-y)p^{\gamma}\log(1-p) \end{aligned} \tag{5}$$

where $y \in \{\pm 1\}$ specifies the ground-truth class. $p \in [0,1]$ is the model' s estimated probability for the class with label y = 1. p_t represents the predicted probability of the true category of the sample, $p_t = \begin{cases} p, y=1 \\ 1-p, y=0 \end{cases}$. γ is an adjustable factor that reduces the loss of simple samples.

The overall training loss of our modal is represented mathematically as:

$$L = \gamma L_{FL} + \beta L_{CL} \tag{6}$$

where γ and β are the punishment parameters of L_{FL} and L_{CL} respectively.

3 Experiment

3.1 Experimental Setup and Dataset

We implemented our experiments on a machine with a single 8 GB NVIDIA GeForce RTX 3080 GPU. The model has a batch size of 64, an epoch of 10. We use the Adam optimizer for gradient descent with a learning rate of 1e-3. For text tokenizer use 'bert-base-uncased'. The temperature parameter in the consistency contrast loss function is set to 0.9. The value ratio of γ is 2.114, 1.995, 1.962, 1.676, 1.654, 0.839.

Table 1 The LIAR dataset statistics.

Statistics	Num
Training set size	10,269
Validation set size	1,284
Testing set size	1,283
Avg. statement length (tokens)	17.9
Pants on fire	1,050
False	2,511
Barely-true	2,108
Half-true	2,638
Mostly-true	2,466
True	2,063

Our study uses LIAR [9], a dataset for detecting fake news. It contains 12.8K short, manually annotated sentences from the politifact.com API, and the politifact.com editors rate the authenticity of each sentence, see Table 1. The data comes from various scenarios, including press releases, TV or radio interviews, campaign speeches, etc. This data set has 6 labels: pants-fire, false, barely-true, half-true, mostly true and true.

3.2 Results

Table 2 shows the accuracy comparison of various existing models and our proposed model on the LIAR dataset. CNN+LSTM+Fuzzy [10] is a hybrid model based on fuzzy logic that considers a combination of news articles and text and digital context information. This model achieves 0.465 on the LIAR dataset, which is 0.007 lower than our model. This shows that our model outperforms this model regarding loss function and network structure selection. Funnel-CNN [11] uses different classifiers and embedding models for fake news detection, and our model also outperforms it. Table 4 shows the impact of different temperature parameters on the model performance when our model uses the consistency contrast loss on the LIAR dataset. The model's performance is relatively low when the temperature is between 0.1-0.2. The model performs fairly well when the temperature increases to 0.3-0.4. When the tem-

Table 2 Performance of existing models and our proposed model on LIAR dataset.

Model	Accuracy
LSTM attention [12]	0.385
Capsule neural networks [13]	0.409
ANSP [14]	0.428
Deep Ensemble Model [15]	0.448
AENeT [16]	0.464
CNN+LSTM+Fuzzy [10]	0.465
Funnel-CNN [11]	0.467
Ours	0.472

Table 3 Ablation experiments of our model on the LIAR dataset.

Method	Accuracy	Precision	Recall	F1
w/o MLP-mixer	0.446	0.550	0.443	0.436
w/o Focal Loss	0.452	0.522	0.461	0.452
w/o CCL	0.448	0.526	0.460	0.439
Ours	0.472	0.553	0.477	0.467

perature increases to 0.5-0.7, the model performance decreases. When the temperature is 0.9, the model's accuracy and F1 score are both the highest, indicating that the model performs best on the LIAR dataset at this temperature.

Figure 2 shows the confusion matrix of our model on the LIAR dataset, showing the classification effect of different categories. The horizontal axis represents the predicted label of the model, the vertical axis represents the true label, and the color depth indicates the number. Label 0 is 'pants-fire', label 1 is 'false', label 2 is 'barely-true', label 3 is 'half-true', label 4 is 'mostly-true', and label 5 is 'true'. It can be seen that the model has the best classification effect on the half-true category. This indicates that the model is not inclined to recognize the complete truth of the news.

Figure 3 shows the multi-classification ROC curve of our model on the LIAR dataset. The ROC curve shows the classification effect of the model for each category, where the horizontal axis is the false positive rate and the vertical axis is the true positive rate. It can be seen that category 0 has the highest AUC value of 0.94. This shows that the model has the best classification effect on category 0. The AUC values of categories 1, 2, 3, and 4 are similar, which shows that the model has the same classification effect on them. The AUC value of category 5 is the lowest, which is 0.61. This shows that the model has the worst classification effect on category 5. The number of samples in category 5 is the lowest, which may be related to this.

Table 4 Performance comparison of different temperature parameters in the consistency contrastive loss on the LIAR dataset.

Temperature	Accuracy	Precision	Recall	F1
0.1	0.444	0.529	0.460	0.441
0.2	0.453	0.526	0.473	0.446
0.3	0.453	0.535	0.468	0.450
0.4	0.463	0.548	0.467	0.455
0.5	0.452	0.536	0.457	0.445
0.6	0.445	0.531	0.459	0.439
0.7	0.443	0.541	0.445	0.438
0.8	0.462	0.551	0.464	0.448
0.9	0.472	0.553	0.477	0.467

3.3 Ablation Experiment

Table 3 shows the ablation study results of our proposed model on the LIAR dataset. After removing the MLP-mixer, the model's accuracy and F1 score are lower, indicating that the MLP-mixer may significantly improve the overall performance. In particular, the recall and F1 values are reduced, indicating that the MLP-mixer plays a key role in balancing precision and recall. After removing the Focal Loss, the model's accuracy did not decrease much, with the accuracy at 0.452, a decrease of 0.02. This suggests that the Focal Loss plays a role in the model, but not as big as that of the MLP-mixer. After removing the CCL module, the accuracy is 0.448, a decrease of 0.024. This shows that CCL can improve model performance by improving text content consistency and attributes.

4 Conclusion

In this paper, we propose a novel fake news detection model that uses consistency contrastive learning to enhance the consistency of text and attribute features. In terms of feature extraction, MLP-mixer is used to extract text and attribute features, thereby replacing the traditional attention mechanism, enabling it to capture long-range dependencies. Experiments on the LIAR dataset show that our model outperforms existing methods, confirming the effectiveness of consistency contrastive learning for fake news detection.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K11212 and the NINJAL Collaborative Research Projects.

References

[1] Pritika Bahad, Preeti Saxena, and Raj Kamal. Fake news detection using bi-directional lstm-recurrent neural network. **Procedia Computer Science**, Vol. 165, pp. 74–82, 2019.

[2] Tina Esther Trueman, Ashok Kumar, P Narayanasamy, and J Vidya. Attention-based c-bilstm for fake news detection. **Applied Soft Computing**, Vol. 110, p. 107600, 2021.

[3] Hongyan Ran, Caiyan Jia, Pengfei Zhang, and Xuanya Li. Mgat-esm: Multi-channel graph attention neural network with event-sharing module for rumor detection. **Inf. Sci.**, Vol. 592, No. C, p. 402–416, May 2022.

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In **International conference on machine learning**, pp. 4904–4916. PMLR, 2021.

[5] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. **Advances in neural information processing systems**, Vol. 34, pp. 9694–9705, 2021.

[6] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. **Advances in neural information processing systems**, Vol. 34, pp. 24261–24272, 2021.

[7] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 3733–3742, 2018.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 42, No. 2, pp. 318–327, 2020.

[9] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In **Annual Meeting of the Association for Computational Linguistics**, 2017.

[10] Cheng Xu and M-Tahar Kechadi. Fuzzy deep hybrid network for fake news detection. SOICT '23, New York, NY, USA, 2023. Association for Computing Machinery.

[11] Mohammadreza Samadi, Maryam Mousavian, and Saeedeh Momtazi. Deep contextualized text representation and learning for fake news detection. **Information processing & management**, Vol. 58, No. 6, p. 102723, 2021.

[12] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. Fake news detection through multi-perspective speaker profiles. In **International Joint Conference on Natural Language Processing**, 2017.

[13] Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. Detecting fake news with capsule neural networks. **Applied Soft Computing**, Vol. 101, p. 106991, 2021.

[14] Lianwei Wu, Yuan Rao, Ambreen Nazir, and Haolin Jin. Discovering differential features: Adversarial learning for information credibility evaluation. **Inf. Sci.**, Vol. 516, No. C, p. 453–473, apr 2020.

[15] Arjun Roy, Kingshuk Basak, Asif Ekbal, and Pushpak Bhattacharyya. A deep ensemble framework for fake news detection and classification. **arXiv preprint arXiv:1811.04670**, 2018.

[16] Vidit Jain, Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Yashvardhan Sharma. Aenet: an attention-enabled neural architecture for fake news detection using contextual features. **Neural Computing and Applications**, Vol. 34, No. 1, pp. 771–782, 2022.

A Appendix

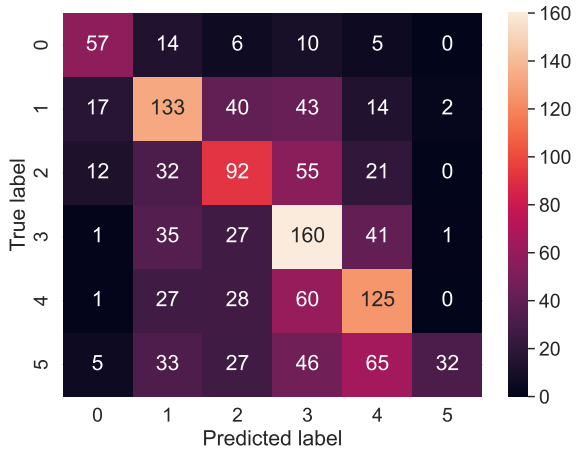

Figure 2 Confusion matrix of our model on LIAR dataset

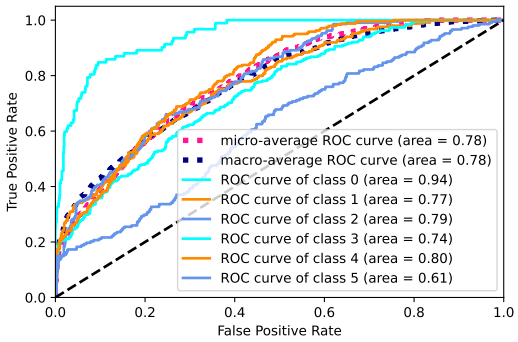

Figure 3 Multi-classification ROC curve of our model on LIAR dataset