2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

F3ND: Bridging the Semantic Gap with Tri-modal Self-Attention for Enhanced Fake News Detection

Zhonghao Yao(✉) and Huaping Zhang

Beijing Institute of Technology, Beijing, China

Abstract. Current multimodal fake news detection faces the "semantic gap" problem: simple text or image features are unable to fully capture the relationship between the two. To solve this problem, we propose F3ND, a tri-modal integrated self-attention framework for fake news detection. F3ND combines unimodal text and visual features with combined multimodal features. The combined features serve as a link to enhance the correlation between the two unimodal features, which effectively solves the semantic gap problem when understanding multi-modal content. At the same time, we apply self-attention to flexibly determine the importance of each feature, retaining the discriminative information in uni-modal features that helps to determine whether the news is fake. Our experiments on Weibo and Weibo21 datasets show that F3ND can achieve better performance than many previous baseline models, proving the robustness and effectiveness of our method.

Keywords: Fake News Detection, Self-Attention Mechanism, Semantic Gap.

1 Introduction

The advent of internet technology and the rise of social media have simplified the process of content publication, sped up information dissemination, and increased the susceptibility of content to manipulation, thereby facilitating the proliferation of fake news. News articles on the Internet are usually multimodal, which makes their content more credible than similar articles [1], and also brings greater challenges to the detection of fake news. For example, pictures are often an important part of news content. In most cases, these pictures supplement the news text. Evaluating whether an image appears manipulated or misaligned with the news topic and content serves as an effective approach to detecting fake news [2]. Under such circumstances, the unimodal model will inevitably waste part of the news information and cannot achieve good results. From this perspective, incorporating an analysis of multimodal features is crucial for improving the identification of fake news.

Fake news detection (FND) has long been a focal point in data science research [20-22], and deep learning is an effective approach. In the process of understanding deep multimodal content, the semantic gap [3] is a relatively common problem, as establishing the relationship between textual and visual modalities in multimodal content remains a challenging task [2]. The same text may convey completely different meanings

when paired with different pictures. The text and pictures must be combined to get the final result. For example, [27] mentioned sentences like “love the way you smell today” or “look how many people love you”. Individually, these two sentences are innocuous; however, when paired with an equally benign image of a skunk or a tumbleweed, their tone unexpectedly turns harsh. To tackle this issue, we introduce a FND framework named F3ND (Fusion of Three Feature Modalities for Fake News Detection). To improve detection performance, our method focuses on linking heterogeneous modality features, uncovering their correlations, and conducting fake news identification in a unified manner. In particular, we employ RoBERTa [4] to extract textual features and ResNet50 [5] to derive visual features from the target news items. To strengthen the relationship between the two modal features, we utilize ViLT [6] to extract integrated multimodal features from both the textual and visual components of the news content. Since ViLT has captured the interactive information between different modalities, in order to retain the discriminative information that may exist in a single modality, these three features are dynamically weighted via the attention mechanism and finally classified through the classifier.

We conducted experiments on the Weibo and Weibo21 datasets to assess how effectively F3ND identifies multimodal fake news. The experimental results highlight F3ND's strong ability to analyze and process multimodal data, achieving notable performance on these datasets. Through ablation experiments, we further revealed the contribution of the discriminative unimodal information and the capture of correlations between multimodal features to the model effect. Our method’s potential in multimodal FND is highlighted by these findings.

The main contributions of this paper are outlined below:

– We introduce F3ND, a framework for multimodal FND that combines features from three different modalities.

– By adding multimodal fused features extracted by ViLT, we improve the interplay between multimodal features, ensuring that unimodal features retain their independence and discriminative value, and use the attention mechanism to adaptively allocate weights to various features, providing a feasible method to address the semantic gap challenge.

– We perform extensive experiments on the publicly available Weibo and Weibo21 datasets, and the results of the experiments confirm the efficacy of F3ND in detecting multimodal fake news.

2 Related Works

2.1 Unimodal FND

Most existing unimodal FND approaches determine the authenticity of a post by examining either its textual content or its associated image [25,26]. For methods using text, early studies often manually designed data using language features. For example, Wynne [11] investigated the application of character-level and word-level n-grams in their study, and [12] employed TF-IDF and Count Vectorizer for feature extraction.

After the rise of deep learning, many pre-trained language models have become popular. Jwa et al. [13] used BERT [23] for FND, and performed detection by analyzing the relationship between news headlines and text. In the context of visual detection tasks, fake news can usually be detected by analyzing images to determine whether the image is distorted or whether the image description does not match the news topic or content [2]. [14] and [15] perform FND through image analysis.

However, images in news usually supplement text content, and there is often some correlation or consistency information between the two. Detecting only one of the modes will inevitably ignore this information, which may affect the performance of unimodal methods in handling multimodal FND.

2.2 Multimodal FND

As news articles typically contain both textual and visual elements, a lot of research has been done on multimodal FND in recent years. Khattar et al. [16] developed a model called MVAE, which successfully captured the common underlying features between different modalities and improved the accuracy of FND. The SpotFake framework introduced in [17] leverages BERT [23] and VGG-19 [24] to extract textual and visual features, respectively, for the purpose of FND. The CAFE model introduced by [18] can flexibly fuse features from individual modalities along with their cross-modal relationships. [19] Modeled the news posts as graphs, and combined knowledge concepts with text and visual information to capture semantic representations, leveraging supplementary auxiliary information from the posts, such as background knowledge.

However, many current studies do not focus on ensuring the independence of unimodal features. In some cases, unimodal features may contain some important discriminative information, which is sufficient to quickly determine whether the news is fake. Ensuring that this information is not diluted will help improve the model's capability in detecting multimodal fake news, which encourages us to propose F3ND to solve this problem.

3 Method

3.1 Approach Overview

In the context of multimodal FND, we represent each news as $x = (x_t, x_i)$, where x_t represents text input and x_i represents image input. The ground-truth of the news is y. A label of $y = 1$ signifies that the news is genuine, whereas $y = 0$ denotes a fake news instance. We extract independent text features and independent image features from x_t and x_i respectively, and then extract the fused features of the two at the same time. These three features are further processed using the attention mechanism to obtain the features that can ultimately represent the entire news, which are then input into a classifier to calculate the result $\hat{y}$ that is closest to the ground-truth of the news.

$$\hat{y} = C\left(\text{Avg}\left(\text{Att}\left([T(x_t), I(x_i), M(x_t, x_i)]\right)\right)\right) \tag{1}$$

Where T is a text feature extractor, which is used to extract features from text x_t; I is an image feature extractor, which is used to extract features from image x_i; M is a fused feature extractor, which extracts cross-modal interactive features from the combination of text and image; Att() and Avg() denote two fusion strategies for the three modalities: the former employs a multi-head self-attention mechanism, while the latter applies average pooling to the resulting attention outputs. Considering that both text and image modalities might independently carry sufficient discriminative cues for FND, we avoided using a straightforward feature fusion approach. Instead, we introduced an attention mechanism to dynamically assign weights to each feature, so that the information of each modality can complement each other and retain the independence of unimodal features. In order to obtain more comprehensive information about the news, including text and images, ViLT [6] is applied for extracting cross-modal features from textual and visual data, and finally the feature vectors extracted from each modality are fused into a unified representation. After feature extraction and fusion are completed, the final result y is obtained after the calculation of the classifier C.

3.2 Model Architecture

This section will elaborate on the principles of the F3ND architecture we proposed. As shown in Fig. 1, the full architecture consists of four main components: a text feature extractor for unimodal input, an image feature extractor for unimodal input, a module for extracting fused multimodal features, and a final stage for feature integration and classification.

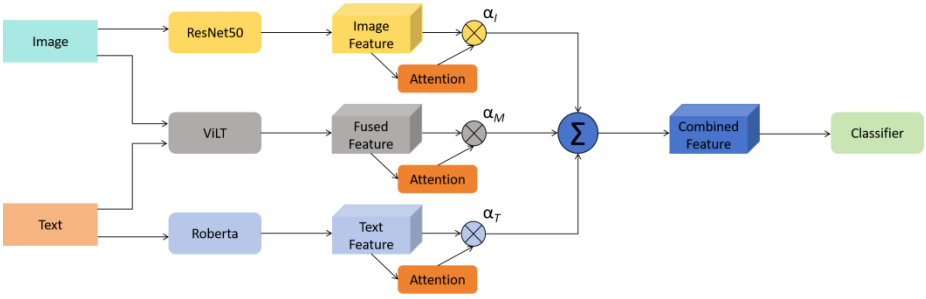

Fig. 1. Roberta and ResNet50 are employed to extract textual and visual features from the news article, while ViLT is used to derive fused features from both the text and images. These three features are then assigned dynamic weights via the self-attention mechanism, and the resulting features are fed into the classifier.

Unimodal text feature extractor. The news text content x_t is input into the pre-trained Chinese Roberta [4] model, which serves as the text feature extractor T. The

extracted unimodal text features are represented as the feature vector $F_t \in \mathbb{R}^{n_{\text{Roberta}}}$, where n_{Roberta} is the output dimension.

$$F_t = T(x_t) \tag{2}$$

Unimodal image feature extractor. For the news input image x_i, we employ the pre-trained ResNet50[5] model and eliminate its final fully connected layer to serve as our image feature extractor I. Unimodal image features are extracted from the news image, which can reflect information such as edges, textures, and overall composition in the image. We use the image feature vector $F_i \in \mathbb{R}^{n_{\text{ResNet50}}}$ to represent it, where n_{ResNet50} is the output dimension of ResNet50.

$$F_i = I(x_i) \tag{3}$$

Multimodal fused feature extractor. Text features or image features on their own may not be sufficient to fully reveal the interaction between the text and images in the news. To accomplish this, we utilize the ViLT [6] model as the multimodal feature extractor M, which captures the cross-modal interactions between text and image to check for consistency between the textual and visual information and to identify possible fake news indicators. It is represented by a fused feature vector $F_m \in \mathbb{R}^{n_{\text{ViLT}}}$, where n_{ViLT} is the output dimension of ViLT.

$$F_m = M(x_t, x_i) \tag{4}$$

Feature fusion and classification. After the above three modules, we obtained the text feature vector F_t, the image feature vector F_i and the fused feature vector F_m. Since the text and image features extracted by Roberta and ResNet50 come from different modal contents of news, they have obvious cross-modal semantic gaps. If we try to directly fuse these two features, capturing the deep semantic correlation between them will be a challenge for the network. At the same time, the text or image of news may contain some discriminative information in a single case. For example, a news article might contain highly exaggerated text alongside ordinary-looking images. In such cases, the news can be classified as fake based solely on the text features. In this case, simply fusing the two features may introduce noise and dilute the discriminative ability of the original unimodal features.

To address the first issue, we incorporated ViLT to extract multimodal fused features, which effectively capture the interaction between text and images, thereby overcoming the limitations of direct fusion methods in terms of semantic alignment. For the discriminative information that may exist in unimodal features, we use a multi-head self-attention mechanism. When the features from a specific modality (like text) play a more crucial role in identifying fake news, the attention mechanism will assign greater weight to these features, preventing their discriminative information from being overshadowed by other modalities. Since different feature vectors have different dimensions, we first project them to the same fusion dimension d, and then stack the projected features on the modality dimension to form a three-dimensional tensor:

$$F_t' = P_T(F_t), \quad F_i' = P_I(F_i), \quad F_m' = P_M(F_m) \tag{5}$$

$$F_{\text{proj}} = \left[F_t', F_i', F_m'\right] \in \mathbb{R}^{3\times d} \tag{6}$$

After that, we use multi-head self-attention to calculate the relationship between each modality to obtain the modal feature F_{att}, and then take the average of the modal dimensions to calculate the final feature F of the news:

$$F_{\text{att}} = \text{Att}(F_{\text{proj}}, F_{\text{proj}}, F_{\text{proj}}) \tag{7}$$

$$F = \text{Avg}(F_{\text{att}}) = \frac{1}{3}\sum_{i=1}^{3} F_{\text{att}}^{(i)} \tag{8}$$

Next, in order to map the feature F to the final classification probability, we designed a two-layer fully connected network. The input is first processed by the fully connected layer and ReLU activation, followed by a Dropout operation to prevent overfitting, which produces the hidden layer representation z:

$$z = \text{Dropout}(\text{ReLU}(W_1 F + b_1)) \tag{9}$$

Finally, the second fully connected layer transforms the hidden layer representation into a single output, which is then passed through the Sigmoid activation function to compute the predicted probability of fake news, $\hat{y}$:

$$\hat{y} = \sigma(W_2 z + b_2) \tag{10}$$

This series of operations constitutes the entire feature fusion and classification process. By applying the attention mechanism, the independent discriminability of each modality is retained and the final classification task is completed. To optimize the model, we utilize the binary cross entropy loss function. The formula for calculating the loss for each training sample is as follows:

$$\text{L}_i = -\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \tag{11}$$

The overall loss is the average of all sample losses, denoted as:

$$\text{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \tag{12}$$

4 Experiments

4.1 Experimental Setup

Datasets. We analyzed the performance of F3ND on two publicly accessible datasets. **Weibo dataset:** The main data comes from Sina Weibo. After preprocessing, there are 7723 samples in the dataset, each of which includes a Chinese text and at least one picture. The dataset is split as follows: 5415 samples (70%) for training, 843 samples (11%) for validation, and 1465 samples (19%) for testing. **Weibo21 dataset:** The data

content of Weibo21 dataset mainly comes from Sina Weibo, and has been strictly screened and manually labeled. It provides news data with nine domain labels (technology, military, education, accidents, politics, health, finance, entertainment, society). After preprocessing, it contains 6156 data. The dataset is split as follows: 4926 samples (80%) for training, 615 samples (10%) for validation, and 615 samples (10%) for testing.

The dataset statistics are shown in Table 1.

Training Details. For the Roberta model, since we are experimenting on the Chinese datasets Weibo and Weibo21, pre-trained model "chinese-roberta-wwm-ext" is used, and the length of the input text should be less than 170 characters. For the ResNet50 model, input images are resized such that their shorter side is scaled to 256 pixels, and then a 224×224 square region is cropped, normalized, and standardized. We use the pre-trained "vilt-b32-mlm" model for ViLT, with input text truncated to fewer than 170 tokens and images resized to 224×224. All parameters of the three pre-models are not frozen. Including the classifier at the end, the total parameter count of the model is approximately 226M.

We use Adam optimizer is used to update the parameters, with the learning rate configured as 1×10^{-5} and the batch size of 48. For the Weibo21 dataset, training 10 epochs takes about 16 minutes on an RTX 4090.

Table 1. Dataset Statistics

Dataset	Total Samples	Real News	Fake News	Language
Weibo	7723	3615	4108	Chinese
Weibo21	6156	3099	3057	Chinese

Baselines. To ensure a fairer and more reproducible comparison, we selected several publicly available multimodal pre-trained models to compare with F3ND:

ALIGN [7]: The pre-trained model used in the experiment is align-base. In this setup, a dual encoder design is implemented with EfficientNet handling visual input and BERT processing the text.

CLIP [8]: The pre-trained model used in the experiment is clip-vit-base-patch32. For visual representation, the ViT-B/32 Transformer is employed, whereas textual information is processed using a Transformer with masked self-attention.

FLAVA [9]: The pre-trained model used in the experiment is flava-full. Both image and text inputs are encoded using the ViT-B/32 Transformer architecture, with a multimodal encoder integrated on top to handle cross-modal tasks.

ViLT [6]: The pre-trained model used in the experiment is vilt-b32-mlm. After splitting the image into uniform patches, each patch is linearly projected to align with the embedding space used for text representations.

We intentionally selected these four vision-language pre-trained models as comparison baselines because they have their own characteristics in cross-modal semantic alignment strategies and can comprehensively cover fusion methods from coarse-grained to fine-grained. By comparing with these four models, we not only examined

the effects of dual encoder alignment methods (ALIGN, CLIP), but also verified the performance of hybrid two-stage fusion (FLAVA) and end-to-end joint encoding (ViLT), thus highlighting the advantage of F3ND in further improving the accuracy of FND through the trimodal self-attention mechanism after comprehensively utilizing the unimodal discriminant information extracted by Roberta and ResNet50 and the cross-modal interaction fused by ViLT.

Experimental Details. Our dataset preprocessing method refers to [10]. A uniform classifier, composed of two fully connected layers and a Dropout layer with a dropout rate of 0.3, is appended to each baseline model. The learning rate and number of training epochs are kept consistent with those used in F3ND.

Evaluation Metrics. Model performance is assessed based on the following set of metrics:

Accuracy: The proportion of samples that the model correctly predicted.

Recall: The ratio of all true positive samples correctly identified by the model.

Precision: The proportion of samples predicted by the model to be positive that are actually positive.

F1-score: The harmonic mean of Recall and Precision, serving to balance the performance of both.

AUROC: An indicator that measures the model's ability to classify positive and negative samples at different decision thresholds. Higher values indicate better performance.

4.2 Performance Analysis

As shown in Table 2 and Fig. 2, our F3ND outperforms several baseline models in all evaluation metrics on Weibo and Weibo21, which reflects its robustness and performance. After analysis, we believe that some discriminative information in the unimodal content contributes to the strong performance of the F3ND model. Compared with several baseline models, our model focuses more on the extraction of unimodal text and image features. In order to retain the discriminative information that may exist in them, we perform separate feature extraction operations on text and images respectively, and introduce a self-attention mechanism to dynamically assign weights to different features, effectively retaining the discriminative information in these unimodal features. Then we fused these unimodal features with multimodal features to complete the classification, and finally achieved excellent performance in all evaluation metrics of the two datasets.

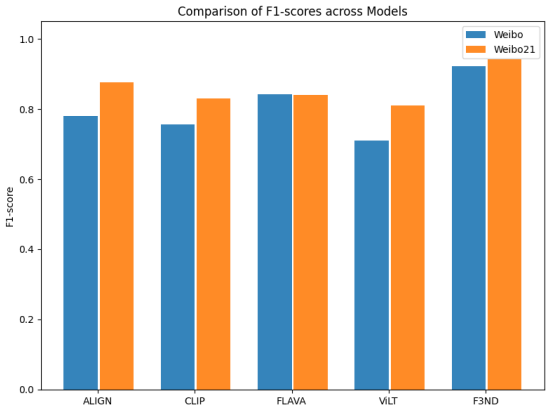

Fig. 2. Comparison of F1-scores across Models

Table 2. Fake News Detection Performance

Model	Weibo					Weibo21				
	Acc.	Rec.	Prec.	F1	AUROC	Acc.	Rec.	Prec.	F1	AUROC
ALIGN	0.765	0.860	0.713	0.780	0.855	0.868	0.923	0.834	0.876	0.938
CLIP	0.769	0.743	0.771	0.757	0.836	0.823	0.861	0.802	0.831	0.894
FLAVA	0.838	0.897	0.794	0.842	0.915	0.844	0.819	0.864	0.841	0.933
ViLT	0.660	0.858	0.605	0.710	0.754	0.808	0.813	0.808	0.810	0.872
F3ND	0.922	0.961	0.888	0.923	0.980	0.942	0.939	0.945	0.942	0.983

4.3 Ablation Study

We performed tests on the Weibo21 dataset to examine the impact of different components in F3ND on its performance. In each test, a different component was removed and the model was retrained to compare its effect with the complete model. Each removed component is as follows:

F3ND w/o ResNet50: We removed ResNet50 and its extracted unimodal image features, and directly used unimodal text features and multimodal features for classification.

F3ND w/o Roberta: We removed Roberta and its extracted unimodal text features, and directly used unimodal image features and multimodal features for classification.

F3ND w/o ViLT: We removed ViLT and its extracted multimodal features, and directly used unimodal text features and image features for classification.

F3ND w/o Attention Mechanism: We removed the attention mechanism and adopted a simpler fusion strategy, which is to directly connect the three features to obtain the final feature.

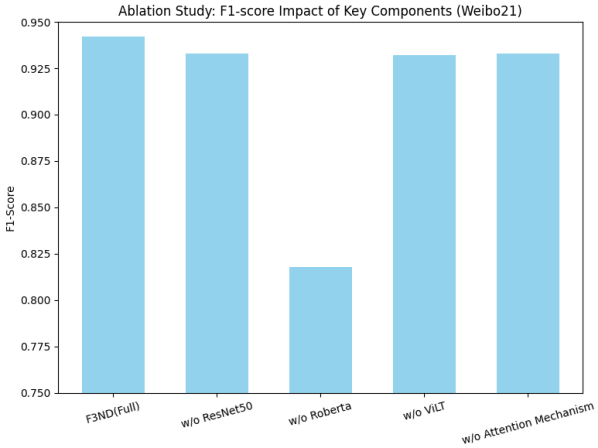

Fig. 3. Ablation Study: F1-score Impact of Key Components (Weibo21)

Table 3. Ablation Study Results

Model Variant	Weibo21				
	Acc.	**Rec.**	**Prec.**	**F1**	**AUROC**
F3ND(Full)	0.942	0.939	0.945	0.942	0.983
w/o ResNet50	0.933	0.919	0.947	0.933	0.980
w/o Roberta	0.811	0.842	0.796	0.818	0.891
w/o ViLT	0.930	0.945	0.919	0.932	0.980
w/o Attention Mechanism	0.932	0.942	0.924	0.933	0.980

Analysis: The results are in Table 3 and Fig. 3. Removing any component leads to performance drops in F1-score and AUROC. The test results of removing ResNet50 and Roberta show that there is indeed discriminative information in the unimodal features that helps to assess if the news is fake. The largest decline occurs without Roberta, indicating that text features carry the most discriminative information in Weibo21. Removing ViLT weakens cross-modal understanding. The decline in model performance after removing the attention mechanism reflects the advantage of the attention mechanism in dynamically assigning weights to each feature. Compared with directly connecting three features, this method can better ensure that some important discriminative information in the features is not diluted by other features, and ultimately obtain better performance.

5 Conclusions

We propose F3ND, a tri-modal self-attention framework that fuses independent text, image and cross-modal features to bridge the semantic gap in FND. On Weibo and

Weibo21, F3ND significantly outperforms strong baselines in both F1-score and AUROC. Ablation studies confirm each component's contribution—particularly the Roberta-based text extractor. Future work will target improved image feature extraction to further boost performance.

References

1. Hangloo, S., Arora, B. Combating multimodal fake news on social media: methods, datasets, and future perspective. Multimedia Systems 28, 2391–2422 (2022). https://doi.org/10.1007/s00530-022-00966-y
2. E. Festus Ayetiran and Ö. Özgöbek, "A Review of Deep Learning Techniques for Multimodal Fake News and Harmful Languages Detection," in IEEE Access, vol. 12, pp. 76133-76153, 2024, doi: 10.1109/ACCESS.2024.3406258.
3. Chen, Wei, et al. "New ideas and trends in deep multimodal content understanding: A review." Neurocomputing 426 (2021): 195-215.
4. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
5. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
6. Kim, Wonjae, Bokyung Son, and Ildoo Kim. "Vilt: Vision-and-language transformer without convolution or region supervision." International conference on machine learning. PMLR, 2021.
7. Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.
8. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.
9. Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
10. Tong, Yu, et al. "MMDFND: Multi-modal Multi-Domain Fake News Detection." Proceedings of the 32nd ACM International Conference on Multimedia. 2024.
11. Wynne, Hnin Ei, and Zar Zar Wint. "Content based fake news detection using n-gram models." Proceedings of the 21st international conference on information integration and web-based applications \& services. 2019.
12. Gravanis, Georgios, et al. "Behind the cues: A benchmarking study for fake news detection." Expert Systems with Applications 128 (2019): 201-213.
13. Jwa, Heejung, et al. "exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert)." Applied Sciences 9.19 (2019): 4062.
14. Masciari, Elio, et al. "Detecting fake news by image analysis." Proceedings of the 24th symposium on international database engineering \& Applications. 2020.
15. Wang, Yuhui, Xin Jin, and Xiaoyang Tan. "Pornographic image recognition by strongly-supervised deep multiple instance learning." 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016.
16. Khattar, Dhruv, et al. "Mvae: Multimodal variational autoencoder for fake news detection." The world wide web conference. 2019.
17. Singhal, Shivangi, et al. "Spotfake: A multi-modal framework for fake news detection." 2019 IEEE fifth international conference on multimedia big data (BigMM). IEEE, 2019.
18. Chen, Yixuan, et al. "Cross-modal ambiguity learning for multimodal fake news detection." Proceedings of the ACM web conference 2022. 2022.

19. Qian, Shengsheng, et al. "Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17.3 (2021): 1-23.
20. Allein, Liesbeth, Marie-Francine Moens, and Domenico Perrotta. "Like article, like audience: Enforcing multimodal correlations for disinformation detection." arXiv preprint arXiv:2108.13892 (2021).
21. Shu, Kai, et al. "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media." Big data 8.3 (2020): 171-188.
22. Xue, Junxiao, et al. "Detecting fake news by exploring the consistency of multimodal data." Information Processing \& Management 58.5 (2021): 102610.
23. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019.
24. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
25. Bian, Tian, et al. "Rumor detection on social media with bi-directional graph convolutional networks." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 01. 2020.
26. Conroy, Nadia K., Victoria L. Rubin, and Yimin Chen. "Automatic deception detection: Methods for finding fake news." Proceedings of the association for information science and technology 52.1 (2015): 1-4.
27. Kiela, Douwe, et al. "The hateful memes challenge: Detecting hate speech in multimodal memes." Advances in neural information processing systems 33 (2020): 2611-2624.