

An Analysis of Multimodal Approaches for Fake News Detection

Xinhai Zhang^{1,a,*}

¹*School of Electronics Engineering and Computer Science, Peking University, Beijing, 100000
China*

a. 2200013138@stu.pku.edu.cn

**corresponding author*

Abstract: Fake news refers to factually incorrect reports that are spreading at an increasing rate, negatively affecting public opinion, media credibility, and social stability. If not screened and prevented promptly, fake news can influence the public's judgment of the facts. More seriously, it may affect the direction of international public opinion and the credibility of the national media. Thus, detecting and managing fake news is the focus of public opinion issues. To solve the problem, the multimodal fake news detection method came into being, which is a detection method that extracts and fuses features from multiple modalities of news (such as text, pictures, and audio), and then evaluates the results. This paper provides an overview of current relevant news detection techniques and analyzes the advantages of multimodal approaches in the current field of fake news detection, as well as the problems that may be encountered. The results indicate that multimodal fake news detection can improve detection accuracy by fusing multiple data modalities, but still faces a series of challenges such as dataset diversity and model interpretability, with great potential for future development.

Keywords: Multimodal Fake News Detection, Feature Fusion, Deep Learning, Neural Network.

1. Introduction

The proliferation of the Internet and the accelerated growth of social media platforms have notably increased the velocity and reach of the dissemination of misinformation, posing a significant challenge to the trust mechanism, political stability, as well as economic order of society. False information not only tends to mislead the public but may cause public opinion crisis and market manipulation. The complexity and multifaceted nature of fake news presents a significant challenge for the traditional unimodal detection methods, which rely on a single source of information, such as text or images, to cope with it. In recent years, multimodal fake news detection has emerged as an important research direction. In comparison to unimodal methods, multimodal detection captures the inherent complexity and intrinsic connections of news content in a more comprehensive manner. Moreover, the integration of data from disparate modalities, including text, image, and video, enhances the accuracy of detection. However, the application of multimodal methods still encounters many technical challenges, such as the timing of modal fusion, the effectiveness of feature extraction and the interpretability of models. This study aims to analyze the recent progress of multimodal fake

news detection, explore the fusion methods of different modalities and their advantages and disadvantages, and propose new ideas and improvement methods to deal with the dissemination of fake news. By combining the existing literature, this paper hopes to provide a reference for subsequent research and promote the further development and application of multimodal fake news detection technology.

2. Overview of Multimodal Fake News Detection Methods

2.1. Fake News Detection Datasets and Detection Criteria

With the continuous development of fake news detection research, various datasets have emerged to support diverse studies in the field. The FakeNewsNet dataset is one of the significant multi-source datasets that Potthast et al. utilized to evaluate text style based on fake news detection methods [1]. Fake news datasets are divided into unimodal and multimodal types. Unimodal fake news datasets primarily rely on a single type of data source, such as single text, images, and videos, and typically contain a lot of news content labeled as true or false for training machine learning models. For instance, the Weibo 21 and LIAR datasets are typical unimodal datasets suitable for simple fake news detection tasks. In contrast, multi-peak fake news datasets combine multiple data sources, including text, images, videos, and social contexts, to provide more comprehensive information and use the complementarity between different modalities to improve the accuracy and robustness of fake news detection. In addition, Liu et al. pointed out that certain multimodal datasets can be regarded as news dissemination trees, such as Twitter15 and Twitter16, which record news content and retain relevant user behavior information, so that each tree represents a news event and covers the user's interactive behavior with news content [2].

In the evaluation of fake news detection, the commonly used metrics include Accuracy, Precision, Recall and F1 Score. Each of these evaluation metrics has its own importance, which can fully reflect the model performance. Accuracy is the basic index that indicates the proportion of correctly predicted samples in the test set to the total samples, reflecting the model's ability to distinguish between true and fake news. Precision measures the proportion of samples predicted by the model to be fake news that are actually fake news, and a higher precision rate implies that the model has a lower false alarm rate. Moreover, Recall indicates the proportion of fake news samples correctly identified by the model as a percentage of all fake news samples, reflecting the comprehensiveness of the model's detection. And F1 Score is the reconciled average of Precision and Recall, which provides a more comprehensive assessment of the model performance, especially if the two are contradictory. In general, a higher F1 Score indicates better model performance.

2.2. Unimodal Fake News Detection Methods

Unimodal fake news detection relies on a single modality in the news content, such as text, image or video, to determine its authenticity. In this approach, a trained model determines the credibility of the news by analyzing the features of a specific modality. In general, this detection methods mainly focus on the textual features of the news. Different text features have significant differences on the detection performance of the model. Baarir and Djeflal proposed a SVM-based (support vector machine-based) classification method for news detection. They concluded that among the five key text features—text content, author, source, date, and sentiment, text content and author information contribute the most to fake news detection. [3]. Khanam et al. summarized existing models and methods and found that most models using plain Bayesian algorithms have 70-76% prediction accuracy. In contrast, methods using Neural Networks and Support Vector Machines can achieve higher accuracy rates, reflecting a positive correlation between model complexity and performance [4].

However, due to advances in media technology, news reports are not limited to textual content, but are often supplemented with images and videos to enhance the credibility of the information. Unimodal technology in news detection has several disadvantages. Firstly, it relies on only a single modality in the news content, such as text, image, or video. This leads to incomplete information acquisition. For instance, analyzing only text may miss potential false clues in images or videos. News authenticity often requires a comprehensive consideration of multiple aspects of information, which unimodal detection cannot achieve. Secondly, its adaptability is limited. With the continuous progress of media technology, news presentations are increasingly diverse. Unimodal methods that solely depend on text cannot meet current needs. Different types of news may exhibit different characteristics in different modalities, and unimodal detection methods struggle to adapt to this diversity. For example, some fake news may seem reasonable in text but contradictions can be found through image or video analysis. Thus, unimodal detection relying only on text cannot meet the current requirements and multimodal methods are needed to detect news [5]. Research by Isabel and Santiago indicates that a multimodal approach based on a Convolutional Neural Network (CNN) that uses both text and picture data is able to detect fake news with 87% accuracy on the Fakeddit dataset. In contrast, a multimodal approach based on Bidirectional Encoder Representation (BERT) achieved 78% accuracy on the same dataset, while an unimodal approach using only text data struggled to reach this level [6]. This suggests that the fusion of images and other data sources not only significantly improves the accuracy of fake news detection, but also enhances the model's ability to adapt to complex scenes. Therefore, a multimodal approach is especially necessary in modern fake news detection.

2.3. The Necessity for Shifting to Multimodal Detection

As the proliferation of fake news intensifies, traditional unimodal fake news detection methods have been inadequate to meet the increasingly complex detection needs. The shift to multimodal detection methods thus has become an inevitable trend. Multimodal detection enhances accuracy and efficiency by integrating features from different modalities, such as text, images, and videos. Data fusion can be categorized into pre-fusion, mid-fusion and post-fusion based on the fusion timing. Pre-fusion merges data from different modalities to form a unified input in the feature extraction stage. For example, when processing a news report that includes both text and images, the two can be merged to extract relevant features. This approach helps to fully utilize multimodal information in the initial stage of the model, but may lead to interference between features. Middle fusion occurs at the intermediate layers of the model, where features from different modalities are integrated, which is usually implemented in deep learning frameworks. Upon reaching a specific middle layer, the model combines text and image features to optimize the subsequent decision-making process. This method allows for dynamic interactions between different modal features, enhancing the flexibility of the model. In addition, post-fusion means that the features of each modality are first extracted independently, fed into the model for processing separately, and finally fused at the output stage, and commonly used fusion methods include weighted averaging or harmonic averaging. This method facilitates independent processing of different modalities and helps maximize the potential of each modality, but the final fusion process may lead to information loss.

In terms of feature extraction, text features are extracted by term frequency or word embedding (such as Word2Vec or BERT), while image features focus on attributes like color, texture, and shape. For video and audio data, corresponding features can also be extracted using deep learning networks, thus enhancing the depth of multimodal processing. More advanced multimodal models such as ViLT (Visual and Linguistic Transformer) and MBT (Multimodal Bottleneck Transformer) can handle different input data and excel in feature fusion [2]. Basic fusion methods include weighted averaging and harmonic averaging, while more complex structures such as attention mechanisms can assign

different weights to the outputs of different modalities, making the model more focused on the most critical information. For example, when determining the authenticity of news, the attention mechanism can dynamically adjust the weights of text and image features according to the relevance of the content, thus improving the accuracy of fake news detection. Through this transformation, fake news detection from unimodal to multimodal not only improves the accuracy of detection, but also makes the model more adaptable and flexible when dealing with complex information.

3. State of the Multimodal Fake News Detection Methods

3.1. Early Multimodal Fake News Detection Methods

Early multimodal fake news detection included EANN and MVAE, which combined the extracted unimodal data features with both text and vision data to improve the accuracy and F1 value of the fake news detection models compared to the state-of-the-art unimodal methods at that time [7, 8]. The SpotFake framework, proposed by Singhal et al, consists of three parts. The first part extracts the textual features using a language model (e.g., BERT), and the second part learns the image features from the pre-trained VGG-19 on the ImageNet dataset. The second part learns image features from pre-trained VGG-19 on the ImageNet dataset. Finally, the two are fused using a simple concatenation technique to obtain a machine model for fake news detection [9]. Later, they proposed SpotFake+, in which they used XLNet instead of BERT, further improving the performance of their model again [10]. In addition, the MAM approach uses a combination of a location-coded self-attention mechanism and a pre-trained convolutional neural network to extract text and image features, respectively. A hybrid attention mechanism module is introduced for text-image feature interaction, which uses a hierarchical feature processing approach to reduce redundancy in multimodal interactions and bi-directional feature fusion to ensure the integrity of the training information [11]. The module uses a hierarchical feature processing method to reduce redundant information during multimodal interaction. A bidirectional feature fusion ensures the integrity of the training information [2].

3.2. Recent Multimodal Fake News Detection Methods

In terms of technological innovation, Ying et al. trained four-dimensional models from image patterns: image semantics, fusion (image semantics and text), and text, which can be visualized to get different evaluation scores under different dimensions, and then finally fused them to get the final test results [9]. In addition, Zhang et al. found that most training models focus on the decomposition of the news content itself, arguing that the news dissemination process, such as the user's comments and replies, will have relevant clues. Besides, some malicious content, such as swiping and diversion, can interfere with this news. Thus, they added the features of the user's information and dissemination perspectives to the model and made a Fake News Detection Model Based on Multi-View Representation, which can extract the representation of news in three views: news content, user information, as well as news dissemination. It uses the attention mechanism and bi-directional graph neural network to mine the features of the news and ultimately achieve the detection of fake news through the interaction and fusion between the high-level features of multiple views of the news [12].

In addition, Yuan et al. argued that the current multimodal detection models, such as EANN and MAVE, are only a simple one-to-one fusion of text and image features, which ignores the correlation between multiple images and text within a news story. Therefore, they proposed a multimodal fake news detection model based on one-to-many relationships, which can make full use of the contextual relationships in the article and the correlation between graphs, to improve the performance of the detection model [13]. Furthermore, the FND-CLIP framework, a multimodal fake news detection network based on Contrastive Language-Image Pre-training (CLIP), extracts the depth representation

in the news using two unimodal encoders and two paired CLIP encoders. The multimodal features generated by CLIP are weighted according to the CLIP similarity of the two modalities, and a modal attention module is also introduced to aggregate features. The framework has been shown to exhibit superior capabilities in the mining of key features for the detection of fake news. [14]. The Game-on model and the GraMuFeN model are based on the loss of information in graph neural networks, which improves the F1 value by about 10% on the same dataset and significantly reduces the trainable parameters [15, 16].

4. Technological Challenges and Bottlenecks

Despite considerable advancements in multimodal fake news detection, existing research highlights numerous technical challenges and bottlenecks that may hinder future development.

4.1. Object-Oriented Challenges and Dataset Issues

In the future, text and picture news may be replaced by video news. With the popularity of short video platforms, such as TikTok (Douyin), much news and information can be released and disseminated through videos. On TikTok, the official account of CCTV News has 160 million followers, and almost every video it releases has tens of thousands of likes and comments, which shows the carrying speed and influence of short video platforms in the dissemination of information. In the future, short video news may become mainstream. At present, video news generally includes text, supplemented by narration. Accurately separating background and text or audio from video and making judgment on relevance will be an important challenge in detecting multimodal fake news. In addition, the existing datasets are still small and are generally collected and organized for news in different fields. In the future, if a more comprehensive fake news dataset can be proposed to generalize more data to multiple domains, the model specificity can be greatly reduced and the generalization ability of future trained models can be improved so that they can be applied to different types of news censorship.

4.2. Interpretability and Scalability Issues

The interpretability issue presents a major challenge in fake news detection. While these models excel at distinguishing between real and fake news, they fail to explain the underlying principles guiding their categorization. Lack of transparency in fake news detection systems can create distrust among the public and the media, thus hindering their development. Therefore, it is essential for multimodal fake news detection techniques to progress toward more explainable AI models. These models should strike a balance between robust fake news detection performance and interpretability, which can increase transparency and public trust in the technology [17]. In addition, it is difficult to achieve scalability without increasing the number of modalities because deep neural networks are complex and expensive to build, and most existing multimodal models use multiple deep neural networks (one for each modality) [18]. If scalable models can be developed in the future, such as interface-based fake news detection techniques like audio modal interfaces, resource utilization and model operation efficiency will be greatly improved by activating audio modal interfaces when they are available.

4.3. Fine-Grained Detection and Timeliness

The majority of extant fake news detection models are predicated on a binary classification problem, namely the distinction between veracity and falsity. However, sometimes we still want to know which parts of the news are false, how much text is false, or the text is true but the accompanying image is wrong. In some cases, the degree of falsehood of the news is a topic of more interest, so there is a need for a finer-grained segmentation of the news, such as funny, satirical, offensive, and other fake

news [2]. At the same time, the speed of information dissemination is increasing. It may take only a few minutes for a piece of fake news to ferment or even trigger a public opinion crisis on the Internet. It is crucial to predict and prevent the spread of fake news as early as possible. Speeding up the detection of new fake news and optimizing the algorithm while ensuring accuracy is a significant challenge for future multimodal fake news detection technology.

5. Conclusion

Multimodal fake news detection methods have shown their excellent effectiveness in today's massive news environment, which become an important tool for detecting fake news. Compared with unimodal fake news detection methods, multimodal techniques are more adaptable to the actual characteristics and diverse forms of today's news dissemination. Through multimodal detection, researchers are able to extract rich feature information from various dimensions of news. This method can not only analyze the semantic, lexical and syntactic features of text, but also extract color, texture and shape features from images, and comprehensively consider the audio information in videos. In addition, combining information such as users' profiles and comments and fusing features from different sources can help determine the authenticity of news more accurately. In the future, as the dominant position of online media in news sharing is further consolidated, multimodal fake news detection methods will face more opportunities for development and innovation. With the continuous evolution of news dissemination methods as well as technological advances, further research should enhance the analysis and detection of video content, construct large-scale multi-domain fake news datasets, promote the interpretability and scalability of fake news detection models, and explore fine-grained fake news identification methods.

References

- [1] Potthast, M., et al. (2017) *A stylometric inquiry into hyperpartisan and fake news*. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 231-240.
- [2] Liu, H., et al. (2023) *Survey of Fake News Detection with Multi-model Learning*. *Journal of Frontiers of Computer Science & Technology*, 17(9): 2015-2029.
- [3] Baarir N F, Djeflal A. *Fake news detection using machine learning*. 2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH). IEEE, 2021: 125-130.
- [4] Khanam Z, Alwasel B N, Sirafi H, et al. *Fake news detection using machine learning approaches*. *IOP conference series: materials science and engineering*. IOP Publishing, 2021, 1099(1): 012040.
- [5] Peng, G., Wu, F., Han, L., et al. (2024) *Cross-modal Interaction and Feature Fusion Network for Fake News Detection*. *Computer Science*, 1-10
- [6] Alonso-Bartolome, S. and Segura-Bedmar, I. (2022) *Multimodal Fake News Detection*. *Inf.*, 13(6): 284.
- [7] Ying, Q., Hu, X., Zhou, Y., et al. (2023) *Bootstrapping multi-view representations for fake news detection*. *Proceedings of the AAAI conference on Artificial Intelligence*, 37(4): 5384-5392.
- [8] Zhang, X., Sun, F., Feng, L., et al. (2024) *Multi-View Representations for Fake News Detection*. *Netinfo Security*, 24(03): 438-448.
- [9] Singhal, S., Shah, R.R., Chakraborty, T., et al. (2019) *Spotfake: a multi-modal framework for fake news detection*. *2019 IEEE fifth international conference on multimedia big data*, 39-47.
- [10] Singhal, S., Kabra, A., Sharma, M., et al. (2020) *Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)*. *Proceedings of the AAAI conference on artificial intelligence*, 34(10): 13915-13916.
- [11] Wang, Y., et al. (2018) *EANN: Event adversarial neural networks for multi-modal fake news detection*. *Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*, 849-857.
- [12] Khattar, D., et al. (2019) *MVAE: Multimodal variational autoencoder for fake news detection*. *Proceedings of the World Wide Web Conference*, 2915-2921.
- [13] Yuan, Y., et al. (2023) *Multi-modal Fake News Detection Based on the One-to-many Relationship*. *Journal of Chinese Information Processing*, 37(09): 131-139.
- [14] Zhou, Y., et al. (2023) *Multimodal fake news detection via clip-guided learning*. *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2825-2830.

- [15] Dhawan, M., Sharma, S., Kadam, A., et al. (2024) *Game-on: graph attention network based multimodal fusion for fake news detection*. *Social Network Analysis and Mining*, 14(1): 114.
- [16] Kananian, M., Badiei, F. and Ghahramani, S.A.G. (2024) *GraMuFeN: graph-based multi-modal fake news detection in social media*. *Social Network Analysis and Mining*, 14(1): 104.
- [17] Tufchi, S., Yadav, A. and Ahmed, T. (2023) *A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities*. *International Journal of Multimedia Information Retrieval*, 12(2): 28.
- [18] Hua, J., Cui, X., et al. (2023) *Multimodal fake news detection through data augmentation-based contrastive learning*. *Applied Soft Computing*, 136: 110125.