Multimodal Fake News Detection Combining Social Network Features with Images and Text

Lawrence Y. H. Low, Yen-Tsang Wu, Yan-Hong Liu, Jenq-Haur Wang
Department of Computer Science and Information Engineering
National Taipei University of Technology
Web Information Retrieval Lab
t113999402@ntut.org.tw t107599005@ntut.edu.tw
t112598044@ntut.org.tw jhwang@ntut.edu.tw

Abstract

The rapid development of social networks, coupled with the prevalence of Generative AI (GAI) in our society today, has led to a sharp increase in fake tweets and fake news on social media platforms. These fake media led to more in-depth research on fake news detection. At present, there are two mainstream methods used in detecting fake news, namely content-based fake news detection and propagation / network-based fake news detection. Early content-based detection method inputs an article's content and uses a similarity algorithm to identify fake news. This method improved by using single-modality features such as images and text as input features. However, existing research shows that single-modality features alone cannot identify fake news efficiently. The most recent method then fuses multimodal features such as images and text, as features to be input into the model for classification purposes. The second propagation / network-based fake news detection method creates graphs or decision trees through social networks, treating them as features to be input into the model for classification purposes. In this study, we propose a multimodal fake news detection framework that combines these two mainstream methods. This framework not only uses images and text as input features but also combines social metadata features such as comments. The framework extracts these comments and builds them into a tree structure to obtain its features. Furthermore, we also propose different feature fusion methods which can achieve better results compared with the existing methods. Finally, we conducted ablation experiments and proved that each module is required to contribute to the framework's overall performance. This clearly demonstrated the effectiveness of our proposed approach.

Keywords: Fake news detection, Multimodal fusion, Multimodal learning, Social media

1 Introduction

With the development of the Internet, social media has replaced most traditional media such as newspapers and magazines. Although these network developments provide convenience for users (Khattar et al., 2019), due to the increasingly rapid speed of information dissemination on social networks and the development of Generative AI (GAI), fake news is generated in increasing quantities and its spread is also transforming significantly. Therefore, how to quickly and effectively classify fake news is an important research topic.

Over the years, online social media content has evolved from plain text to multimodal content that combines pictures and text, and in some cases, videos, sounds, and text (He et al., 2021; Pinnaparaju et al., 2021). Early research on fake news detection mainly used single-modality classification, but studies have found that single-modality can no longer efficiently detect fake news. Therefore, current research topics are moving towards using multimodality features to detect fake news. Multimodality refers to the combination of multiple types of modalities, including but not limited to pictures, text, and sound. Chen et al. (2022), and Singhal et al. (2019) reported that past multimodal fake news detection tends to solve the fake news problem by considering additional subtasks such as event discriminators and finding cross-modal correlations. Fake news detection relies heavily on subtasks. Without subtasks, the detection results

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 266–276
November 20-22, 2025 ©2025 Association for Computational Linguistics

will be significantly reduced. It was found that cross-modal content can now provide additional supplementary features for fake news detection, but these studies mainly focus on the integration of cross-modal content. Previous studies did not take into account the differences between content in different modalities, resulting in poor model performance. The Modular Co-Attention Network (MCAN) method is inspired by the way humans read news with pictures and text (Wu et al., 2021). They proposed a multimodal joint attention network and found that the interdependence between multimodal features can achieve better detection results.

However, different modalities may express the same thing at certain times. In this case, adding multimodal fusion features will create noise and affect the performance of the classification task. On the other hand, when the unimodal detection performance is subpar, multimodal fusion features may be added to increase the feature input for better model performance. Therefore, researchers should be aware of the impact of the modifications between different modalities on the model. Other than the usage of unimodal features, a timely addition of multimodal features can obtain better classification results (Qian et al., 2021; Zhou et al., 2020; Wang et al., 2018), therefore most recent research studies focus on how to fuse different modalities and understand the consistency between different modalities to achieve better accuracy.

Currently, almost all research on multimodal fake news detection uses images and text. However, with the rapid development of GAI, the generation of fake news has become faster and more compelling, therefore using only images and text as input is no longer sufficient to accurately detect fake news. More diverse features must be considered to assist fake news detection, but relatively few studies have incorporated social metadata features. In addition to user information such as “retweets”, “likes”, and “number of friends”, common social metadata features also use shared comments or articles to establish a propagation path structure of a graph or tree structure (Rahimi et al., 2024; Li et al., 2020). However, these existing methods rarely combine the features of the two methods.

In order to solve the above-mentioned problems, this study refers to the architecture MMFN, Multi-grained Multi-modal Fusion Network (Zhou et al., 2023) and proposes a novel framework, Multi-Modal Title Comment (MMTC) combining social metadata features and multimodal fusion of text and images. In addition to the integration of unimodal features and multimodal fusion features of text and images, our proposed method adds social metadata features with comment tree structure as input to achieve more accurate fake news detection. The MMTC framework includes:

1) Multimodal fusion module: The module obtains single-modal feature input through pre-trained BERT (Devlin et al., 2019) and Swin Transformer v2 (Liu et al., 2022). The pre-trained model CLIP (Radford et al., 2021) is used to extract semantic information between different modalities to solve the problem of semantic inconsistency between the different modalities.
2) Title and comment module: The similarity between the text and the image summary is used as a weight, multiplied by the features of the comment tree structure and subsequently concatenated with the image summary features to evaluate whether the social background feature is important based on the relevance between the image and the text.

We summarize our key contributions in this paper as follows:

- We propose a novel framework, MMTC that uses social network comments, pictures, and text as features, taking into account both the details and overall aspects of the news.
- We demonstrate the effectiveness of our framework by comparing with existing baseline methods using well-known datasets. MMTC outperforms the existing fake news detection methods.
- We perform ablation tests to verify the effects of the various modules in our framework are effective.

2 Related Works

The purpose of fake news detection is to distinguish the authenticity of news based on the relevant information of the news released on social media platforms. This information may include text content, image content, comments, communication structure and other user characteristics. Related research can be divided into two categories based on the data. The first category is based on article content. The features of this method usually use the content of the article, such as pictures, text, and in some cases, news videos. The second category is based on social background. This method uses the information about the user in the news as features, such as numerical features such as “retweets”, “likes”, and “friends”. It also uses graphs or tree structures to transform the user’s information into the dissemination structure of the article in order to use it as a social interaction feature.

2.1 Content-based fake news detection

In recent years, fake news has been spreading frequently on social media platforms. According to previous studies (Liu and Wu, 2020; Shu et al., 2017; Rubin et al., 2016), it is crucial to detect fake news. Research on fake news detection can be divided into two broad categories: Based on (i) news content and (ii) social context. The method based on news content can be sub-divided into unimodal fake news detection and multimodal fake news detection.

Unimodal Fake News Detection Previous studies on fake news detection have mostly focused on single modality, with a large portion of them using text content analysis (Nan et al., 2021; Ajao et al., 2019) and image content analysis (Jin et al., 2017). The amount of existing information makes traditional manual detection more difficult, and fake news detection models based on Machine Learning is used to mitigate this limitation. Ma et al. (2016) proposed a method that uses Recurrent Neural Networks (RNN) to learn features. The results show that models based on Deep Learning are more effective. Kaliyar et al. (2021) reported several methods of fake news detection such as Features-based approaches, Knowledge-based approaches, Learning-based approaches, and proposes a BERT-based method that only uses text data as input training. Xue et al. (2020) proposed a Multi-Vision Fusion Neural Network (MVFNN) for the detection of fake news pictures, combining the pixel domain, frequency domain and tampering detection features of the image.

Multimodal Fake News Detection Although fake news can be effectively identified by simply using text or pictures. Online social platforms include rich multimodal information such as picture, text, video (Zhang et al., 2019) and uses existing post datasets to achieve multimodal fake news detection by extracting visual emotion features, text emotion, behavioral responses, and metadata (Leung et al., 2023). Ying et al. (2023) reported that the consistency between cross-modalities and the features of different modalities affecting model decisions are still unresolved. The authors proposed a method of extracting features from different perspectives of text, image patterns, and image semantics, and using the representation of each image to approximately predict the authenticity of the news. This multimodal representation can predict the consistency across the different modalities, thereby obtaining accurate fake news detection. Qi et al. (2023) focused on short videos to detect fake news. Their model added social metadata features such as “comment”, “user information” and used a Cross-Model Transformer to learn the relationship between different modalities. These added social metadata features are in addition to the video content features such as pictures, text, and images. Palani et al. (2022) used images and text, together with CapsNet and BERT as input models to extract features in order to combine the features for fake news classification.

2.2 Fake news detection based on social context

The interaction between social media allows news to have a variety of social activities. For example, after the news is released, users can share, discuss and analyze it with their friends. These constitute the social interaction of news, which includes not only the authenticity of news reports, but also users’ comments on the news and their emotions towards the news. Usually, social metadata features are obtained using structure-based methods. User information can be obtained from social media, and unstructured data can be combined into structured data such as graphs and tree-structures. Graph based methods have achieved remarkable results because they can closely simulate the social interaction and the process of spreading online news.

Related work on graph methods includes Qian et al. (2016), who propose a novel Multi-modal Multi-view Topic Opinion Mining (MMTOM) model for social event analysis in multiple collection sources. MMTOM can effectively combine multimodal and multi-view attributes in a unified and principled manner for social event modeling. It not only discovers multimodal common topics from all collections and summarize the similarities and differences of these collections on each specific topic, but also automatically mine multi-perspective opinions on the learned topic in different collections. Gong et al. (2023) presented a systematic survey of graph-based fake news detection research and Deep Learning-based techniques. We further discuss the challenges and unsolved issues in graph-based fake news detection and identification as well as the future research directions. Zhang et al. (2021) proposed a model based on Graph Attention Networks to extract information from user interactions. In the communication graph, nodes represent user text content and edges represent response interactions. The authors implemented an attention mechanism to decide the edge weights between pairs of nodes.

Regarding tree-structed research methods, Ni et al. (2021) aim to solve the problem of fake news detection in real-world scenarios. The authors developed a new Neural Network-based model to detect fake news and provide explanations on social media. Only the source short tweet, and its retweets are provided as features, however user comments are omitted. A Multi-task Attention Tree Neural Network (MATNN), proposed by Bai et al. (2023) jointly classify stance and detect the authenticity of rumors. The authors designed a structural representation, which converts irregular rumor conversation trees into Regular rumor Conversation Trees (RC-Trees). When extracting features, the authors use the tree's word attention mechanism to extract local structures for stance analysis. Zhang et al. (2023) organized claim posts in a cycle as a temporary event tree, extracts event elements, and converts them into bipartite graphs of temporary event trees in terms of posts and authors, namely, author tree and post tree. We propose a novel rumor detection model with a hierarchical representation on a bipartite temporal event tree.

3 Model architecture and methods

We propose a multimodal framework, Multi-Modal Title Comment (MMTC) that combines social network features to improve multimodal models that only use content or only use social metadata. MMTC consists of two modules: Multi-Modal Block (MMB) and Title-Comment Block (TCB). MMB (section 3.1) contains two types of inputs. Unimodal feature input, which includes the unimodal input of images and texts in the architecture to represent the overall representation of texts and images. Multimodal fusion input, which represents the detailed input of the consistency of images and texts in the architecture. TCB (section 3.2) includes a similarity weight calculation and a social background feature input. The image title similarity weight, which evaluates the importance of the comment features of the article by calculating the similarity between the image summary and the title text, and the comment tree structure feature, which calculates the comment features related to the article by establishing a tree structure. The final portion of MMTC is a fake news classifier (section 3.3). The overall architectural design of MMTC is as shown in Figure 1. MMTC has two major blocks: Multi-Modal and Title-Comment. The feature strings from these two blocks are sent to the classifier for training and classification.

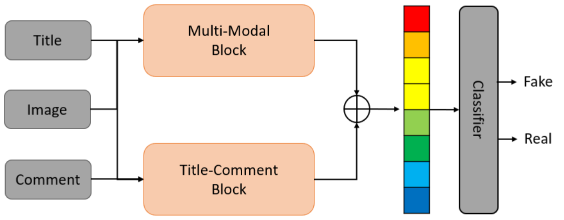

Figure 1. Overall Multi-Modal Title Comment (MMTC) framework.

3.1 Multi-Modal Block (MMB)

The MMB module is divided into two parts: Unimodal feature input and Multimodal feature input. The unimodal features are the title text of the tweet and the image in the tweet. The pre-trained model is used to extract the overall semantic representation of the single modality in the model. In the multimodal feature, the pre-trained multimodal model is used to find the correspondence between words and images, and subsequently, the multimodal detailed semantic representation is obtained through the designed Multiple Feature Fusion network. The Multi-Modal Block (MMB) architecture is as shown in Figure 2.

In MMB, the unimodal inputs of text and images will use a pre-trained BERT and Swin Transformer v2 respectively, whereas the multimodal input will use the pre-trained CLIP model. The unimodal features will be added to the text or image output generated by the multimodal through average pooling, and finally the feature output will be obtained through the projection head. The multimodal part will be input into a multimodal fusion block, Multiple Feature Fusion to obtain feature output.

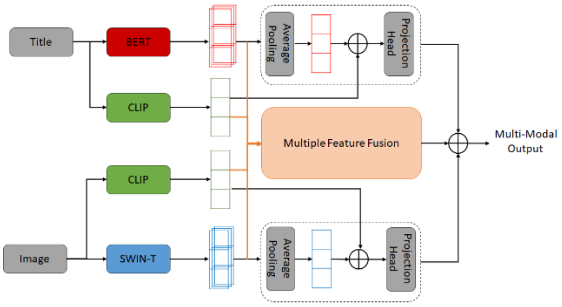

Figure 2. Multi-Modal Block (MMB) architecture.

Unimodal Feature Input The unimodal text features here use the pre-trained BERT to extract text features. BERT is a Transformer based model developed by Google that uses a large amount of unlabeled text in an unsupervised manner. It is one of the most commonly used language models. Therefore, we use BERT here to extract text features in the text content. The text of the title will be input into BERT and output to the last layer of the hidden layer, and average pooling will be used to compress the features to obtain the overall text semantic representation as FT.

The unimodal image features are obtained using a pre-trained Swin Transformer v2, an upgraded version of Swin Transformer (Liu et al., 2021). The upgraded version is a hierarchical Visual Transformer architecture designed for efficient processing of high-resolution images and various downstream tasks such as classification, detection, and segmentation. In this paper, the image is pre-processed and the image size is converted to 224 × 224. The image is input into the last layer of the output hidden layer of Swin Transformer v2. Similar to text processing, average pooling is used to compress the features to obtain the overall image semantic representation as FI.

Multimodal Feature Input The purpose of multimodal feature input is to find the corresponding relationship between different modalities to obtain features. In this paper, the CLIP model is used to obtain multimodal features. CLIP is a novel multimodal model proposed by OpenAI in 2021. Its pre-training task allows the model to predict from scratch on a dataset containing 400 million image-text pairs, from which the model learns the caption to obtain image feature representation.

Our proposed method use CLIP to extract two different embeddings, CLIP text embedding and CLIP image embedding. These two embeddings represent the overall semantic vector of the sentence and the overall semantic vector of the image, respectively. Since single-modal feature inputs FT and FI cannot directly realize information interaction, a multi-modal fusion method, Multiple Feature Fusion Block (MFFB) is designed here to obtain information interaction between single modality and multimodality, as shown in Figure 3.

The MFFB uses the Co-attention Transformer method (Lu et al., 2019) to experiment with cross-modal information complementarity using the features extracted from BERT and Swin Transformer v2. MFFB uses the co-attention for single modal input to obtain the focus of attention of the other modality, and then connects the outputs between the two modalities and multiplies the weighted similarity of the two inputs of CLIP to obtain the multimodal fusion feature.

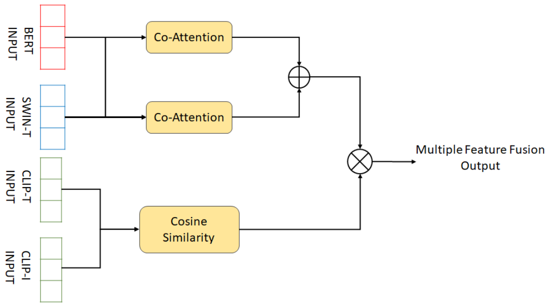

Figure 3. Multiple Feature Fusion Block (MFFB).

The Co-attention Transformer architecture is as shown in Figure 4.

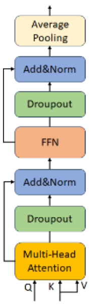

Figure 4. Co-attention Transformer architecture.

We treat the two modalities F1 as Query and F2 as Key and Value and input them into the Co-attention Transformer. Q is the input of one modality, K and V are the input of the other modality. Using this method, the input modality can obtain the attention focus of the other modality.

In this way, we can obtain two different representations. One is the focus of attention in the image guided by the text, read as $FT2I$, and the other is the focus of attention in the text guided by the image, read as $FI2T$. The two features of the previous CLIP are matched through Cosine Similarity to find out whether the two are matched, which is converted into a weight, read as WC. Finally, $FI2T$ and $FT2I$ are concatenated and multiplied by WC to obtain the final output $FMMF$ of MFFB, as shown in Eq. (1):

$$F_{MMF} = (F_{I2T} \oplus F_{T2I})\, W_C \qquad (1)$$

The feature output FMM of Multi-modal Block consists of three features concatenated together as shown in Eq. (2):

$$FMM = F_T \oplus F_I \oplus F_{MMF} \qquad (2)$$

3.2 Title-Comment Block (TCB)

The Title-Comment Block (TCB) uses the similarity between image summary and title to discover the degree of relevance between image and text. It then weights against the comment feature composed of the comment tree, and finally concatenates with the image summary to obtain the feature output. The TCB consists of two blocks. The first block is the similarity calculation between the article title and the text summary. The second block is the comment tree feature established by the article comments. It is weighted by the similarity weight to distinguish which comments are relatively important. Finally, the image summary and comment features are concatenated as the output features of the TCB. The TCB architecture is as shown in Figure 5.

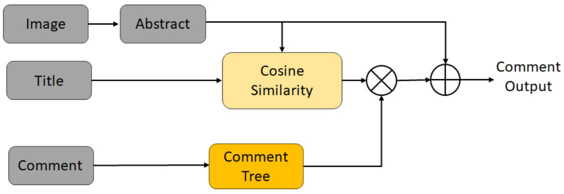

Figure 5. Co-attention Transformer architecture.

Image and Title Similarity Weight In order to find the appropriate social metadata features to reduce noise, similarity calculation is used as weighted weight to calculate the text summary of the article title and the picture, respectively. The text summary of the picture is obtained through the pre-trained multimodal model BLIP (Li et al., 2022). BLIP is a multimodal learning architecture proposed by the Salesforce team. Its purpose is to enhance multimodal performance by aligning images with text. BLIP can perform image question answering, image description generation and multimodal classification tasks. In the paper, we used BLIP to obtain the text summary of the picture and perform cosine similarity calculation with the article title to obtain the image-text semantic similarity weight WIT of the article.

Comment Tree-structured Features In the propagation/network-based methods based on social metadata, comments are often used as one of the features. Some methods treat comments as text and input them into the model together with other text features (Kirchknopf et al., 2021), while others create a tree structure for comments (Ma et al., 2018), which allows people to reply to comments posted by others on social media platforms.

In our proposed method, the article title is taken as the root of the tree structure, and the reply comments are used to establish the leaf nodes of the tree. According to the reply to each comment, a structure tree around the reply-article-title will be established by using a recursive method to calculate from bottom to top. The comment tree-structured features are as shown in Figure 6. In the example, comment 3 replies to comment 1, whereas comment 1 and comment 2 reply to the article title. The features of the parent node are used in a bottom-up manner using a multi-head attention or self-attention, and then the comment feature output of the article is calculated in a cascade manner.

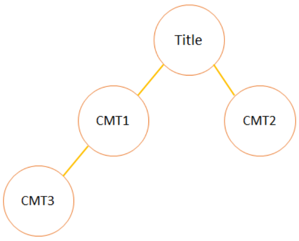

Figure 6. Comment Tree-structured Features.

Since it is impossible for every article on the social platform to have a reply, there are two situations: (i) articles with no comments and (ii) articles with at least one comment. For articles with no comments, our proposed method sets the part without comments to a 256-dimensional tensor vector with all zeros, and uses self-attention to calculate the comment features represented by the tree. Whereas for articles with at least one comment, the parent node of each layer represents each child node and performs multi-head attention calculation with the parent node. Finally, the output of the horizontal child node is concatenated to represent the feature output of the parent node. The comment feature FC represented by the final tree, that is, the article, is calculated through recursion. The feature output of Title-Comment Block is FTC, as shown in Eq. (3):

$$FTC = WIT \times FC + FIA \tag{3}$$

3.3 Fake News Classifier

The fake news classifier in this paper is input from two modules in the architecture, namely the multimodal feature input FMM of the Multiple Feature Fusion Block and the comment social metadata feature FTC of the Title-Comment Block. We concatenate the two feature inputs as shown in Eq. (4):

$$F = FMM \oplus FTC \tag{4}$$

Then we feed the final features into the classifier, as shown in Eq. (5):

$$y = classifier(F) \tag{5}$$

The classifier consists of three fully connected layers and a ReLU activation function. F is the concatenation of all modules. The classification method uses Focal Loss (FL) to classify the prediction result as true or false, as shown in Eq. (6):

$$FL(y) = -\alpha(1 - y)\gamma log(y) \tag{6}$$

α is the balance parameter between positive and negative samples, γ is the penalty for easily classified samples, and y is the probability of predicting the correct category.

4 Experimental Setup and Results

In this section, we present the various parameter settings of our experiments, explain the baseline models and our experimental results.

4.1 Experiment Setup

Datasets Fakeddit is a multimodal fake news dataset released in 2020 (Nakamura et al., 2020). The data is collected from Reddit which contains text, pictures, raw data and comments. The data were collected between March 2008 to October 2019. The data have multiple labels, namely 2-way, 3-way and 6-way. In our experiments, 2-way labels are used for model training, and the features used are "clean title", "title id", "image", "comment", and "comment id" from the dataset.

Preprocessing Because our proposed model simultaneously uses text, images, and comments, we first filtered the dataset to include only articles containing both text and images. We applied a second filter based on the number of comments and replies. However, since not all articles receive replies, we selected those with a reply count between 0 and 5 comments for this study. For the image summaries, we employed the BLIP model to generate captions ranging from 5 to 20 words. After filtering, the dataset size is about 160,000. We randomly selected 50,000 data as our experimental dataset, in which the ratio of true to false news is set to 1:1.

Evaluation Metrics We use accuracy as the evaluation criterion for the binary classification of fake news. Additionally, precision, recall, and F1 score are also used as supplementary evaluation criteria.

Experimental Details In our experiments, we set a dimension of 512. For co-attention, the embedding dimension is 256, and FFN hidden dimension is 512, with 8 multi-heads and dropout is set to 0.1. The text embedding dimension of BERT is set to 256, and uses "bert-base-uncased" model for English data. The maximum input text length is 512. For Swin Transformer v2, we use "microsoft/swinv2-tiny-patch4window8-256" for image features, and the input image size is set to 224*224. The CLIP model uses "openai/clip-vit-base-patch32" with both features set to 256.

The unimodal features of text and images are added together after the output features of CLIP and then reduced to 256 using a linear layer. The features output by multiple feature fusion are also reduced to 256 dimensions. Finally, MMB concatenates text features, image features, and multimodal features and reduces the dimension to

256. In TCB, BERT is also used for text features. The image summary and comment features are concatenated, and then the dimension is reduced to 256. Finally, the features of the two modules are concatenated, and the dimension is reduced to 256.

During the training phase, we scaled the number of classifier layers to 64, 16, and 2, respectively. The batch size is 1. The optimizer uses AdamW, the Learning Rate is set to 1e-6, and the weight drop is set to 1e-1. The model is trained for up to 20 epochs, and early stopping is used to prevent overfitting. For *FL* Loss Function setting, we set alpha to 1 and gamma to 0.25.

4.2 Baseline models

We compare our experimental results with existing method proposed by Uppada et al. (2022). Their method uses Fakeddit dataset with images, text and emotional features to train the model. However, our proposed model does not include emotional features as part of the social metadata; instead, it focuses on images and textual content. We use cross-entropy as our Loss function.

4.3 Experimental results

We present the model comparison between our method and other existing baseline methods. The experimental results showed that our model, MMTC with added social network features achieve better results than models that only uses image and text as input. The results of our method, MMTC and other baseline methods is as shown in Table 1, best results are marked in bold.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F1	Precision	Recall	F1
(BERT+Dense) +Xception	0.826	0.809	0.859	0.833	0.846	0.793	0.819
MMTC	**0.861**	**0.878**	**0.843**	**0.86**	**0.846**	**0.88**	**0.863**

Table 2: Baseline model comparison.

We present the ablation test results that we performed on each input feature of our model, MMTC. From the ablation test results, we observed that, simply by using the multimodal model CLIP, it has the best accuracy among all the modules. However, it is still slightly lower than the accuracy of our proposed framework. An additional finding from the experimental result is the accuracy of using only comments as features is much lower than that of directly obtaining features from the article. Instead, by combining the entire module, the framework can achieve better results. This indicates that the comment features and image summary features can help our model to achieve better classification results. The ablation test results for each modality are as shown in Table 2, best results are marked in bold.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F1	Precision	Recall	F1
Text	0.812	0.869	0.74	0.799	0.769	0.886	0.824
Image	0.777	0.792	0.757	0.774	0.762	0.797	0.779
Multiple	0.856	0.873	0.836	0.854	0.839	0.876	0.857
Image Abstract	0.652	0.651	0.668	0.66	0.652	0.635	0.644
Comment	0.487	0.444	0.059	0.104	0.491	0.925	0.641
MMTC	**0.861**	**0.878**	**0.843**	**0.86**	**0.846**	**0.88**	**0.863**

Table 1: Ablation test results.

5 Conclusion and future work

In this study, we propose a multimodal model that incorporates social network features for fake news detection. Based on previous research, we believe that content-based methods cannot efficiently detect fake news, hence we add a comment tree structure based on social metadata features to assist the detection task. We also designed a different feature fusion method, relying on the similarity of image and text associations to weight single-modal features. Our model also achieves better results than existing methods using the same dataset, and ablation tests on each feature are performed to prove that our inputs are necessary.

For future work, we plan to address the limitations of our proposed model, MMTC through continued research. We aim to use additional social metadata features, such as emotional stance and short videos, to enhance the diversity of the model. To improve the model, we plan to use the full Fakeddit dataset, conduct ablation studies on the effects of comments, and investigate gradient accumulation techniques. For better model evaluation, we intend to compare with recent SOTA models and more up-to-date datasets. With the current advancement of large language models, generative AI, and explainable AI with enhanced reasoning capabilities, we aim to improve our fake news detection model by extending our research in this direction.

Acknowledgments

The authors would like to thank the support from the National Science and Technology Council, Taiwan under the grant numbers: NSTC 114-2221-E-027-068, NSTC 114-2634-F-027-001-MBK.

References

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/ abs/2103.00020

Armin Kirchknopf, Djordje Slijepčević, and Matthias Zeppelzauer. 2021. Multimodal Detection of Information Disorder from Social Media. In *2021 International Conference on Content-Based Multimedia Indexing* (*CBMI*). 1–4. doi:10.1109/ CBMI50038.2021.9461898

Balasubramanian Palani, Sivasankar Elango, and Vignesh Viswanathan K. 2022. CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT. *Multimedia Tools Appl.* 81, 4 (Feb. 2022), 5587–5620. doi:10.1007/s11042-021-11782-3

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference* (San Francisco, CA, USA) (*WWW'19*). Association for Computing Machinery, New York, NY, USA, 2915–2921. doi:10.1145/3308558. 3313552

Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multimodal Knowledge-aware Event Memory Network for Social Media Rumor Detection. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) (*MM'19*). Association for Computing Machinery, New York, NY, USA, 1942–1951. doi:10.1145/3343031.3350850

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Curran Associates Inc., Red Hook, NY, USA.

Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting Microblog Conversation Structures to Detect Rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online),5420–5429. doi:10.18653/v1/2020.coling-main.473

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 1980–1989. doi:10.18653/v1/P18-1184

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) (*IJCAI'16*). AAAI Press, 3818–3824.

Johnathan Leung, Dinusha Vatsalan, and Nalin Arachchilage and. 2023. Feature analysis of fake news: improving fake news detection in social media. *Journal of Cyber Security Technology* 7, 4 (2023),224–241. doi:10.1080/23742917.2023.2237206 arXiv:https://doi.org/10.1080/23742917.2023.2237206

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086 [cs.CV] https://arxiv.org/abs/2201.12086

Junxiao Xue, Yabo Wang, Shuning Xu, Lei Shi, Lin Wei, and Huawei Song. 2020. MVFNN: Multi-Vision Fusion Neural Network for Fake News Picture Detection. In *Computer Animation and Social Agents*, Feng Tian, Xiaosong Yang, Daniel Thalmann, Weiwei Xu, Jian Jun Zhang, Nadia Magnenat Thalmann, and Jian Chang (Eds.). Springer International Publishing, Cham, 112–119.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 6149–6157. https://aclanthology.org/2020.lrec-1.755/

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (Sept. 2017), 22–36. doi:10.1145/3137597.3137600

Marzieh Rahimi and Mehdy Roayaei. 2024. A Multi-View Rumor Detection Framework Using Dynamic Propagation Structure, Interaction Network, and Content. *IEEE Transactions on Signal and Information Processing over Networks* 10 (2024), 48–58. doi:10.1109/TSIPN.2024.3352267

Na Bai, Fanrong Meng, Xiaobin Rui, and Zhixiao Wang. 2023. A multi-task attention tree neural net for stance classification and rumor veracity detection. *Applied Intelligence* 53, 9 (May 2023), 10715–10725. doi:10.1007/s10489-022-038335

Nikhil Pinnaparaju, Manish Gupta, and Vasudeva Varma. 2021. T3N: Harnessing Text and Temporal Tree Network for Rumor Detection on Twitter. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021*, Virtual Event, May 11–14, 2021, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, 686–700. doi:10.1007/978-3-030-75762-5_54

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment Aware Fake News Detection on Online Social Networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2507–2511. doi:10.1109/ICASSP.2019.8683170

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. FakeSV: a multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence* (*AAAI'23/IAAI'23/EAAI'23*). AAAI Press, Article 1620, 9 pages. doi:10.1609/aaai.v37i12.26689

Qi Zhang, Yayi Yang, Chongyang Shi, An Lao, Liang Hu, Shoujin Wang, and Usman Naseem. 2023. Rumor Detection with Hierarchical Representation on Bipartite Adhoc Event Trees. arXiv:2304.13895 [cs.SI] https://arxiv.org/abs/2304.13895

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23).* AAAI Press, Article 601, 9 pages. doi:10.1609/aaai.v37i4.25670

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) (*CIKM'21*). Association for Computing Machinery, New York, NY, USA, 3343–3347. doi:10.1145/3459637.3482139

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools Appl.* 80, 8 (March 2021), 11765–11788. doi:10.1007/s11042-02010183-2

Santosh Kumar Uppada, Parth Patel, and Sivaselvan B. 2022. An image and text-based multimodal model for detecting fake news in OSN's. *J. Intell. Inf. Syst.* 61, 2 (Nov. 2022), 367–393. doi:10.1007/s10844-022-00764-y

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) (*SIGIR'21*). Association for Computing Machinery, New York, NY, USA, 153–162. doi:10.1145/3404835.3462871

Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. 2016. Multi-modal Multi-view Topic-opinion Mining for Social Event Analysis. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands) (*MM'16*). Association for Computing Machinery, New York, NY, USA, 2–11. doi:10.1145/2964284.2964294

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data* (*BigMM*). 39–47. doi:10.1109/BigMM.2019.00-44

Shiwen Ni, Jiawen Li, and Hung-Yu Kao. 2021. MVAN: Multi-View Attention Networks for Fake News Detection on Social Media. *IEEE Access* 9 (2021),106907–106917. doi:10.1109/ACCESS.2021.3100245

Shuzhi Gong, Richard O. Sinnott, Jianzhong Qi, and Cecile Paris. 2023. Fake News Detection Through Graph-based Neural Networks: A Survey. arXiv:2307.12639[cs.SI]https://arxiv.org/abs/2307.12639

Victoria L. Rubin, Yimin Chen, and Nadia K. Conroy. 2016. Deception detection for news: Three types of fakes. In *Proceedings of the Association for Information Science and Technology* 52, 1 (Feb. 2016), 1–4. doi:10.1002/pra2.2015.145052010083

Xinpeng Zhang, Shuzhi Gong, and Richard O. Sinnott. 2021. Social Media Rumour Detection Through Graph Attention Networks. In *2021 IEEE Asia Pacific Conference on Computer Science and Data Engineering*(*CSDE*).1–6. doi:10.1109/CSDE53843.2021.9718466

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multimodal Fake News Detection. In *Advances in Knowledge Discovery and Data Mining*, Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan (Eds.). Springer International Publishing, Cham, 354–367.

Yang Liu and Yi-Fang Brook Wu. 2020. FNED: A Deep Network for Fake News Early Detection on Social Media. *ACM Trans. Inf. Syst.* 38, 3, Article 25 (May 2020), 33 pages. doi:10.1145/3386253

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2560–2569. doi:10.18653/v1/2021.findings-acl.226

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal Fake News Detection on Social Media via Multi-grained Information Fusion. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval* (Thessaloniki, Greece) (*ICMR'23*). Association for Computing Machinery, New York, NY, USA, 343–352. doi:10.1145/3591106.3592271

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. 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* (London, United Kingdom) (*KDD'18*). Association for Computing Machinery, New York, NY, USA, 849–857. doi:10.1145/3219819.3219903

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW'22)*. Association for Computing Machinery, New York, NY, USA, 2897–2905. doi:10.1145/3485447.3511968

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. arXiv:2111.09883 [cs.CV] https://arxiv. org/abs/2111.09883

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs.CV] https://arxiv.org/abs/2103.14030

Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor Detection on Social Media with Event Augmentations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) (*SIGIR'21*). Association for Computing Machinery, New York, NY, USA, 2020–2024. doi:10.1145/3404835.3463001

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia* 19, 3 (2017), 598–608. doi:10.1109/TMM.2016.2617078