International Journal of Informatics and Computation (IJICOM)
Vol. 7, No. 2, 2025
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Multimodal Fake News Detection in Social Media using Transformer Models with Text Image Fusion

Tigus Juni Betri[1], Naufal Hisyam Akmali[2]

Abstract
Fake news on social media increasingly combines misleading textual narratives with manipulated or unrelated images, making detection more challenging for unimodal approaches. This study proposes a transformer-based multimodal fake news detection framework that integrates BERT for textual feature extraction and Vision Transformer (ViT) for visual representation through feature-level fusion. The proposed framework aims to capture complementary semantic and contextual information from both modalities to improve fake news classification performance. We evaluate the model using standard classification metrics and compare it with unimodal baseline models under identical experimental settings. Experimental results show that the proposed multimodal framework achieves 91.3% accuracy, 90.5% precision, 92.4% recall, and 91.4% F1-score, outperforming text-only and image-only models across all evaluation metrics. The findings demonstrate that multimodal fusion effectively captures cross-modal inconsistencies and improves classification robustness in social media environments. Overall, this study confirms that transformer-based multimodal learning provides an effective and reliable solution for fake news detection and contributes to the development of intelligent misinformation monitoring systems.

Keywords:
Detection, Fake News, Multimodal, Social Media, Transformer

1. Introduction

Social media rapidly transforms the way people communicate, distribute information, and consume digital content. Platforms such as Facebook, Instagram, TikTok, and X allow users to share news instantly without editorial verification. This condition increases information accessibility but simultaneously creates serious challenges related to misinformation and fake news dissemination. Social media strongly influences public opinion, political perception, marketing behavior, and social interaction patterns. The massive growth of user-generated content makes manual verification increasingly difficult, allowing fake information to spread faster than factual news. Current social media ecosystems therefore require intelligent automated systems that can identify misleading information efficiently and accurately [1], [2], [5].

Fake news detection becomes increasingly important because misleading information creates negative impacts on society, politics, economics, and public trust. False information often manipulates emotional responses through sensational headlines, misleading narratives, and edited multimedia content. Traditional fact-checking approaches depend heavily on human experts and require substantial time, making them ineffective for large-

Corresponding Author: Tigus Juni Betri (tigusjuni.betri@staff.uinsaid.ac.id)
1 Tigus Juni Betri, State Islamic University of Raden Mas Said Surakarta, tigusjuni.betri@staff.uinsaid.ac.id
2 Naufal Hisyam Akmali, State Islamic University of Raden Mas Said Surakarta, naufalhisyamakmali@gmail.com

scale social media environments. Machine learning approaches begin to address this issue by automating classification tasks using textual features and linguistic patterns. However, many conventional models still struggle to capture contextual semantics, sarcasm, ambiguity, and rapidly evolving misinformation patterns across social media platforms [4], [6], [12], [13].

Recent developments in Natural Language Processing (NLP) significantly improve fake news detection through transformer-based architectures such as BERT. Transformer models effectively learn contextual word representations and semantic relationships from large-scale textual datasets. Studies demonstrate that BERT-based methods outperform traditional machine learning techniques in fake news classification tasks because they better understand sentence context and linguistic dependencies. Enhanced transformer architectures further improve performance by refining attention mechanisms and contextual embeddings. Despite these improvements, text-only approaches still face limitations because fake news on social media increasingly combines textual manipulation with misleading visual content [4], [7], [9].

Visual misinformation emerges as another major challenge in modern fake news dissemination. Social media posts frequently include manipulated images, AI-generated visuals, misleading memes, or unrelated photographs to strengthen false narratives. Advances in generative AI further complicate detection because synthetic images now appear highly realistic and difficult to distinguish from authentic content. Vision Transformer (ViT) models and image-based deep learning approaches show promising performance in identifying manipulated visual patterns. However, image-only analysis cannot fully capture semantic inconsistencies between text and visual information, which limits detection reliability in multimodal misinformation scenarios [8], [11].

The increasing complexity of fake news motivates researchers to explore multimodal approaches that combine textual and visual information. Multimodal fake news detection systems analyze relationships between text and images simultaneously to improve semantic understanding and classification accuracy. Prior studies show that multimodal fusion significantly enhances detection performance compared to unimodal systems. Researchers apply attention mechanisms, semantic alignment, and feature fusion strategies to capture cross-modal inconsistencies between images and textual claims. These approaches demonstrate that integrating complementary modalities enables models to identify deceptive patterns more effectively [16], [17], [18].

Transformer-based multimodal fusion models receive substantial attention because they effectively learn long-range dependencies and semantic interactions across different modalities. Advanced architectures such as tri-transformers, progressive fusion networks, and context-aware fusion frameworks achieve strong performance in multimodal fake news classification. These models utilize self-attention and cross-attention mechanisms to align image and text representations while preserving contextual relationships. Several studies report improved accuracy, robustness, and generalization capability using transformer fusion strategies. Nevertheless, many existing models remain computationally expensive and difficult to optimize for real-world large-scale social media deployment [19], [20], [21], [22].

Researchers also investigate semantic fusion and knowledge-enhanced frameworks to improve multimodal fake news understanding. Co-attention networks, mutual knowledge distillation, entity-enhanced fusion, and cross-modal semantic aggregation techniques successfully strengthen feature interaction between modalities. These methods help models identify subtle inconsistencies between textual statements and accompanying images. However, many previous studies rely on static datasets and controlled benchmarks that may not fully represent the diversity and dynamic nature of real-world social media content. Dataset imbalance, noisy annotations, and rapidly changing

misinformation trends continue to challenge multimodal detection systems [17], [23], [24], [25].

Based on these challenges, this study focuses on developing a multimodal fake news detection framework using transformer models with text-image fusion. This paper aims to improve semantic understanding between textual and visual modalities while enhancing detection accuracy in social media environments. By leveraging transformer-based fusion architectures, this study seeks to capture contextual relationships, cross-modal inconsistencies, and multimodal semantic patterns more effectively than conventional unimodal methods. The proposed approach contributes to the development of more robust and scalable fake news detection systems capable of supporting information credibility and reducing misinformation spread across digital platforms [16], [19], [20].

2. Related Works

Early fake news detection studies primarily relied on traditional machine learning algorithms using textual features extracted from social media posts. Sudhakar and Kaliyamurthie applied Support Vector Machine (SVM) techniques to classify fake news from social media content and demonstrated that machine learning methods could effectively identify deceptive textual patterns [6]. Zarger also investigated machine learning-based fake news detection and showed that feature engineering and supervised learning improved classification performance compared to manual verification approaches [12]. These studies provided important foundations for automated misinformation detection. However, they mainly depended on handcrafted linguistic features and shallow semantic representations. As a result, the models struggled to understand contextual ambiguity, sarcasm, and evolving misinformation styles commonly found on social media platforms [6], [12].

Researchers later adopted deep learning and transformer-based approaches to overcome the limitations of traditional machine learning. Fitri Brianna et al. proposed a fake news detection model using BERT combined with Bi-LSTM and achieved stronger contextual understanding compared to classical models [4]. Oad et al. further enhanced transformer performance through an Enhanced BERT framework that improved feature extraction and semantic representation in fake news classification tasks [7]. These studies showed that transformer architectures effectively captured contextual dependencies and linguistic semantics from textual data. Nevertheless, both approaches focused primarily on textual information and ignored the visual components that frequently accompany fake news content on social media [4], [7].

Several studies also examined fake news from the perspective of user behavior, literacy, and social interaction. Orhan analyzed the relationship between fake news susceptibility, critical thinking disposition, and new media literacy among university students [5]. The findings showed that users with stronger critical thinking skills demonstrated better ability to identify misinformation. Dewanti also highlighted the strong influence of social media on communication and information dissemination within business and marketing environments [1]. These studies successfully explained the human and social dimensions of fake news propagation. However, they did not address the development of automated multimodal detection systems capable of handling large-scale social media data [1], [5].

The advancement of computer vision and Vision Transformer technologies encouraged researchers to investigate visual misinformation detection. Lamichhane developed a Vision Transformer-based framework for detecting AI-generated images and demonstrated that transformer-based visual models effectively identified synthetic image artifacts [8]. Hendryani et al. combined CNN and Vision Transformer architectures with Explainable AI techniques for medical image analysis and showed improved feature extraction and interpretability [11]. These studies proved that transformer-based visual architectures

performed well in extracting complex image representations. However, image-only approaches failed to analyze semantic relationships between images and textual claims, which limited their effectiveness in fake news detection tasks involving multimodal content [8], [11].

Researchers subsequently introduced multimodal fusion frameworks to jointly analyze textual and visual information. Lin et al. proposed a text-image multimodal fusion model using BERT and attention mechanisms for fake news detection [16]. Their approach improved semantic understanding by integrating image features with contextual textual representations. Wang et al. also developed Progressive Fusion Networks that gradually fused multimodal information to improve classification performance [18]. Both studies showed that multimodal learning significantly enhanced fake news detection accuracy compared to unimodal systems. Despite these improvements, the models still experienced challenges in aligning semantic inconsistencies between modalities under highly diverse social media conditions [16], [18].

Several advanced studies focused on improving semantic fusion and cross-modal interaction mechanisms. Zhu et al. proposed intra-modality feature aggregation and inter-modality semantic fusion techniques to strengthen multimodal fake news classification [17]. Hu et al. developed matching-aware co-attention networks combined with mutual knowledge distillation to improve feature interaction between text and images [23]. Qi et al. introduced an entity-enhanced multimodal framework that fused diverse semantic clues to improve contextual understanding [25]. These studies demonstrated strong capability in capturing semantic correlations across modalities. However, the architectures often required high computational resources and complex optimization strategies, making deployment more challenging for real-time social media monitoring systems [17], [23], [25].

Recent studies further refined transformer-based multimodal architectures to improve contextual reasoning and detection robustness. Tufchi et al. proposed AMTCF, which integrated multimodal transformers and ConvNext fusion for contextualized fake news detection [19]. Xu et al. introduced enhanced transformer architectures with multimodal semantic understanding to improve fake news classification accuracy on social media datasets [20]. Chi et al. designed a compact GPT-based multimodal framework with context-aware fusion mechanisms that reduced computational complexity while maintaining strong performance [21]. Choi et al. proposed CroMe, a cross-modal tri-transformer architecture combined with metric learning for improved multimodal alignment [22]. These approaches significantly advanced transformer-based multimodal detection. Nevertheless, many models still faced limitations related to scalability, computational cost, and generalization across rapidly evolving misinformation patterns [19]–[22].

Based on the limitations of previous studies, this research focuses on developing a transformer-based multimodal fake news detection framework using text-image fusion. Prior works successfully demonstrated the effectiveness of transformer architectures, multimodal fusion, and semantic alignment strategies. However, many existing approaches either focused only on textual analysis, image analysis, or computationally intensive fusion mechanisms. This study therefore aims to improve multimodal semantic understanding while maintaining efficient feature interaction between textual and visual modalities. By integrating transformer models with text-image fusion techniques, this paper seeks to develop a more robust and scalable fake news detection system capable of handling diverse misinformation content on modern social media platforms [16], [17], [19], [20].

3. Proposed Method

This study proposes a multimodal fake news detection framework that integrates textual and visual information using transformer-based architectures. We design the framework through five main stages: data pre-processing, feature extraction, multimodal fusion, classification, and evaluation. This paper aims to capture semantic relationships from textual content and contextual visual patterns from social media images to improve fake news detection performance. In the pre-processing stage, we utilize a social media dataset containing paired text and image content. We clean the textual data by removing URLs, special characters, and stop words to reduce noise and improve semantic consistency [9]. After cleaning, we tokenize the text and convert it into contextual embeddings compatible with transformer-based language models. For image pre-processing, we resize all images into fixed dimensions, normalize pixel values, and apply augmentation techniques such as flipping and rotation to improve model robustness and generalization capability during training.

This paper utilizes transformer-based models to extract contextual textual features and deep visual representations from multimodal data. We apply a pretrained BERT model or its variants to process tokenized text sequences and generate contextual embeddings that preserve semantic dependencies and linguistic information [10]. The final hidden representation from the transformer encoder is then used as the textual feature vector T. For the visual modality, we utilize image-based deep learning architectures such as Convolutional Neural Networks (CNN) or Vision Transformers (ViT) to extract high-level image representations [11]. The image encoder generates a visual feature vector V that captures important visual patterns, contextual information, and image semantics relevant to fake news detection. By combining textual and visual representations, this study aims to strengthen multimodal semantic understanding and improve classification accuracy for fake news detection on social media platforms.

The proposed multimodal fake news detection model combines textual and visual feature representations through transformer-based fusion. Let $T \in R^{d_t}$ represent the textual embedding extracted from the transformer model (e.g., BERT), and let $V \in R^{d_v}$ represent the visual embedding extracted from the image encoder (e.g., CNN or ViT). The multimodal fusion process is formulated as:

$$F = [T; V]$$

Where F denotes the concatenated multimodal feature vector, and $[;]$ represents feature concatenation.

The fused representation is then passed into a fully connected classification layer followed by the Softmax activation function:

$$\hat{y} = \mathrm{Softmax}(W_f F + b) \tag{1}$$

where W_f and b denote the trainable weight matrix and bias vector, respectively, while $\hat{y}$ represents the probability distribution of fake and real news classes.

The model optimization uses the Cross-Entropy Loss function:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2}$$

where y_i denotes the ground-truth label, $\hat{y}_i$ represents the predicted probability, and N is the total number of samples.

4. Experimental Setup

To evaluate the effectiveness of the proposed multimodal fake news detection framework, a series of controlled experiments were conducted using a labeled dataset of social media posts. The experimental design aims to ensure a fair and systematic evaluation by maintaining consistency in data distribution, model configuration, and evaluation procedures. This setup allows for a comprehensive comparison between the proposed multimodal approach and baseline unimodal models, while also ensuring that the results are reproducible and scientifically valid.

1. Dataset

The dataset used in this study consists of 6,037 social media posts collected from publicly available sources, where each post contains both textual content and an associated image. The dataset is specifically constructed to support multimodal analysis by ensuring that each instance includes both modalities required by the proposed model. Each data sample is annotated into two classes, namely fake news and real news, with a slightly imbalanced yet realistic distribution of 3,012 fake news samples and 3,025 real news samples. The textual component represents captions or written content accompanying each post, while the visual component consists of images that provide contextual information. The annotation process is conducted based on credibility assessment using verified references and fact-checking guidelines to ensure labeling consistency. Prior to model training, the dataset is divided into training, validation, and testing subsets to support proper model development and evaluation. The distribution of the dataset is presented in Table 1.

Table 1. Dataset Distribution

Dataset Split	Number of Samples	Precentage
Training	4.225	70 %
Validation	906	15 %
Testing	906	15 %
Total	6037	100 %

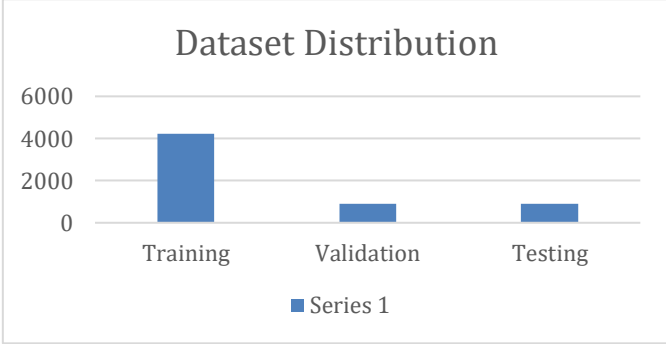

Fig 1. Dataset Distribution

2. Training & Evaluation

The proposed multimodal fake news detection model is implemented using Python with deep learning frameworks, specifically TensorFlow and PyTorch, to ensure flexibility and computational efficiency. The experiments are conducted on a system equipped with GPU acceleration to support the training of transformer-based architectures. For the textual modality, a pre-trained BERT model is utilized to generate contextual embeddings

from input text. The input sentences are tokenized and transformed into input representations consisting of token IDs and attention masks. The final hidden representation of the [CLS] token is extracted as the textual feature vector T, which captures the overall semantic meaning of the input sequence. For the visual modality, a Vision Transformer (ViT) model is employed to extract high-level visual features from images. Each image is resized to a fixed dimension and normalized before being processed into patch embeddings. The resulting feature vector V represents the visual modality and encodes relevant contextual information.

The textual and visual features are integrated using a feature-level fusion approach by concatenating T and V into a unified representation F. This fused feature vector is then passed to a fully connected layer followed by a softmax activation function for classification. The training process is configured using several hyperparameters to ensure optimal performance. A batch size of 16 is used to balance memory efficiency and convergence stability. The learning rate is set to 2×10^{-5}, which is commonly used for fine-tuning transformer-based models. The Adam optimizer is employed due to its adaptive learning capability. The model is trained for a maximum of 10 epochs, and early stopping is applied based on validation loss to prevent overfitting.

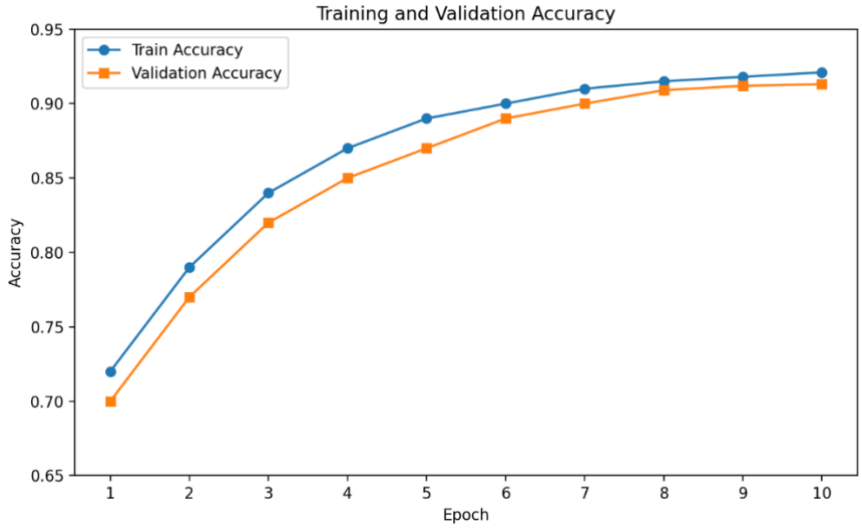

Fig 2. Training and validation accuracy

Figure 2 shows that both training and validation accuracy consistently improved during training and converged after several epochs, indicating stable learning performance. All hyperparameters are selected based on commonly adopted configurations in prior studies and are kept consistent across experiments to ensure fair comparison. The dataset, consisting of 6,037 samples, is split into training, validation, and testing subsets as described previously. All experiments are conducted using the same data split and configuration to ensure reproducibility and consistency in evaluation.

To evaluate the effectiveness of the proposed multimodal framework, this study implements two baseline models for comparative analysis using unimodal approaches. We utilize a text-only model based on the pretrained BERT architecture, where textual inputs are processed through the transformer encoder and the final [CLS] token representation is used for classification through a fully connected layer. We also apply an image-only model using the Vision Transformer (ViT), which extracts visual representations from images. Both baseline models are trained and evaluated using the same dataset partitions, hyperparameter settings, and experimental conditions. To evaluate the performance of the proposed model and baseline approaches, several standard classification metrics are employed, including accuracy, precision, recall, and F1-score.

5. Result and Analysis

The performance of the proposed multimodal fake news detection model and the baseline models was evaluated using the testing dataset consisting of 906 samples. The experiments aimed to compare the effectiveness of unimodal and multimodal approaches in classifying fake and real news posts. The results are summarized in Table 2 using four standard evaluation metrics: accuracy, precision, recall, and F1-score.

Table 2. Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1 Score
Text Only (BERT)	0.847	0.839	0.856	0.847
Image-Only(ViT)	0.781	0.774	0.792	0.783
Multimodal (Proposed)	0.913	0.905	0.924	0.914

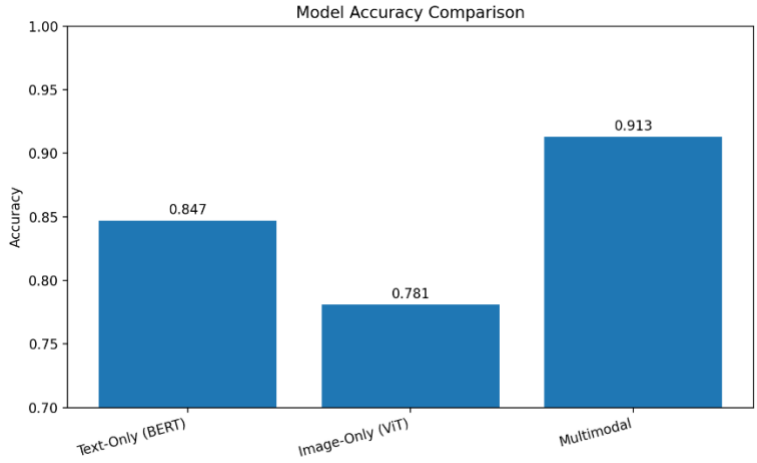

Fig.3 Model Accuracy Comparison

As illustrated in Fig. 3, the proposed multimodal model achieved the highest accuracy compared to both unimodal baseline models. The results show that the proposed multimodal model achieved the highest performance across all evaluation metrics. Specifically, the model obtained an accuracy of 0.913, which is substantially higher than the text-only model (0.847) and the image-only model (0.781). This demonstrates that integrating textual and visual information provides a more comprehensive understanding of social media content compared to unimodal approaches. In terms of precision, the proposed model achieved a score of 0.905, indicating that the majority of posts predicted as fake news were correctly classified, with relatively few false positive errors. The recall value of 0.924 further indicates that the model successfully detected most fake news instances, minimizing the number of false negatives. Additionally, the F1-score of 0.914 confirms that the model maintained a strong balance between precision and recall.

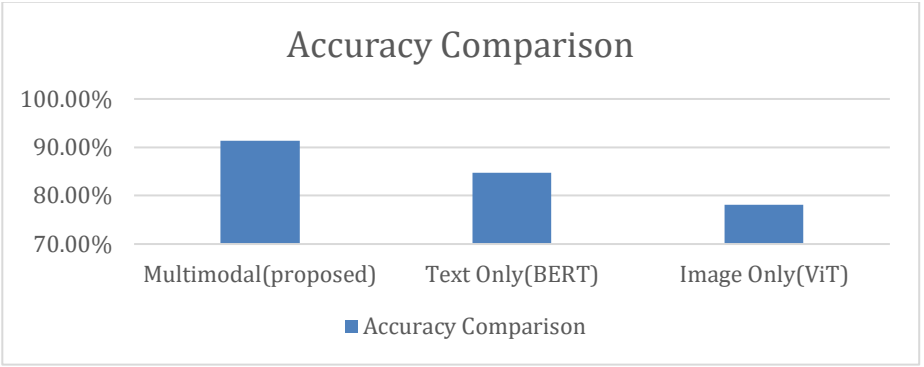

Fig 4. Accuracy Comparison of Models

Fig. 4 clearly illustrates that the proposed multimodal model significantly outperformed both baseline approaches. The performance gain confirms that combining multiple modalities enables the classifier to learn richer and more discriminative patterns.To further analyze the prediction behavior of the best-performing model, a confusion matrix was constructed for the proposed multimodal framework.

Table 3. Confusion Matrix of Proposed Model

Actual/Predicted	Fake News	Real News
Fake News	420	35
Real News	44	407

Table 3 presents the confusion matrix of the proposed multimodal fake news detection model and demonstrates strong classification performance for both fake and real news categories. The model correctly classified 420 fake news instances and 407 real news instances, indicating that the proposed transformer-based text-image fusion framework effectively learned discriminative semantic and visual patterns from multimodal social media data. However, the model also produced 35 false negatives, where fake news content was incorrectly classified as real news, and 44 false positives, where real news was incorrectly identified as fake news.

The relatively balanced distribution between correctly predicted fake and real news samples shows that the model maintained stable classification capability across both classes without significant bias toward a specific category. These results indicate that the integration of textual embeddings and visual representations successfully improved contextual understanding and semantic consistency during classification. Thus, the confusion matrix confirms that the proposed multimodal framework achieved reliable detection performance and demonstrated strong potential for practical fake news identification on social media platforms.

6. Conclusion

The model combines BERT for textual feature extraction and Vision Transformer (ViT) for visual feature extraction through a feature-level fusion strategy. It is to enabling the system to capture complementary patterns from both modalities. This study proposed a transformer-based multimodal fake news detection framework that integrates BERT and Vision Transformer (ViT) through feature-level fusion. The proposed model successfully captured semantic textual information and contextual visual patterns from social media content. Experimental results showed that the multimodal framework outperformed unimodal baseline models and achieved 91.3% accuracy, 90.5% precision, 92.4% recall, and 91.4% F1-score.

These findings confirm that combining text and image representations improves fake news classification performance and enables more balanced detection capability across fake and real news categories. The analysis also demonstrated that multimodal fusion effectively identified cross-modal inconsistencies that are difficult to detect using single-modality approaches. This study therefore confirms that transformer-based multimodal learning provides an effective and reliable solution for fake news detection in modern social media environments. The main contribution of this paper lies in the integration of BERT, ViT, and feature-level fusion to produce stronger semantic understanding and more robust multimodal classification performance compared to conventional CNN- or RNN-based approaches.

References

[1] P. Dewanti, "Analysis of the Effect of Social Media on the Marketing Process in a Store or Business Entity 'Social Media Store,'" *Budapest International Research and Critics Institute Journal*, vol. 2, no. 2, 2024. doi: https://doi.org/10.33258/birci.v4i4.3002.

[2] S. Shakya and E. Ceh-Varela, "Social Media Analytics for Investigations: A Survey of Recent Trends, Challenges and Future Research Direction," *Journal of Social Media Research*, vol. 2, no. 4, pp. 297–318, Dec. 2025. doi: https://doi.org/10.29329/jsomer.57.

[3] Y. R. Sipayung, M. A. Wibowo, and R. Sanjaya, "Multimodal Implicit Sentiment Analysis for Tourism Development: A Systematic Literature Review," *Journal of Information Systems and Informatics*, vol. 8, no. 1, pp. 781–808, Mar. 2026. doi: https://doi.org/10.63158/journalisi.v8i1.1436.

[4] D. Fitri Brianna, M. Apreza Saputra, and M. Al Hapiz, "Fake News Detection Model Using BERT and Bi-LSTM Based on a Discriminative Approach," *JSAI: Journal Scientific and Applied Informatics*, vol. 8, no. 3, 2025. doi: https://doi.org/10.36085.

[5] Orhan, "Fake News Detection on Social Media: The Predictive Role of University Students' Critical Thinking Dispositions and New Media Literacy," *Smart Learning Environments*, vol. 10, no. 1, Dec. 2023. doi: https://doi.org/10.1186/s40561-023-00248-8.

[6] M. Sudhakar and K. P. Kaliyamurthie, "Detection of Fake News from Social Media Using Support Vector Machine Learning Algorithms," *Measurement: Sensors*, vol. 32, p. 101028, Apr. 2024. doi: https://doi.org/10.1016/j.measen.2024.101028.

[7] Oad, M. Hamza Farooq, A. Zafar, B. Ayesha Akram, R. Zhou, and F. Dong, "Fake News Classification Methodology With Enhanced BERT," *IEEE Access*, vol. 12, pp. 164491–164502, 2024. doi: https://doi.org/10.1109/ACCESS.2024.3491376.

[8] D. Lamichhane, "Advanced Detection of AI-Generated Images Through Vision Transformers," *IEEE Access*, vol. 13, pp. 3644–3652, 2025. doi: https://doi.org/10.1109/ACCESS.2024.3522759.

[9] P. W. Cahyo, U. S. Aesyi, W. A. Setianto, and T. Sulaiman, "A Novel Named Entity Recognition Approach of Indonesian Fake News Using Part of Speech and BERT Model on Presidential Election," Elsevier B.V., Dec. 2025. doi: https://doi.org/10.1016/j.jjimei.2025.100354.

[10] E. Murakami, T. Shionoya, S. Komenoi, Y. Suzuki, and F. Sakane, "Cloning and Characterization of Novel Testis-Specific Diacylglycerol Kinase η Splice Variants 3 and 4," *PLoS One*, vol. 11, no. 9, Sep. 2016. doi: https://doi.org/10.1371/journal.pone.

[11] Atika Hendryani, Vita Nurdinawati, and Agus Komarudin, "Explainable Artificial Intelligence Model for Pneumonia Detection: A Hybrid CNN-ViT and Grad-CAM," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 14, no. 4, pp. 235–244, Nov. 2025. doi: https://doi.org/10.22146/jnteti.v14i4.19822.

[12] T. Zarger, "Fake News Detection on Social Media Using Machine Learning," 2025.

[13] M. Fuzail, F. Nazim, and H. Hussain, "Fake News Detection on Social Media Platforms Using Machine Learning and Ensemble Techniques," *Kashf Journal of Multidisciplinary Research*, pp. 2–6, 2025. [Online]. Available: Kashf Journal of Multidisciplinary Research

[14] N. Verma, S. Getenet, C. Dann, and T. Shaik, "Evaluating an Artificial Intelligence (AI) Model Designed for Education to Identify Its Accuracy: Establishing the Need for Continuous AI

Model Updates," *Education Sciences*, vol. 15, no. 4, Apr. 2025. doi: https://doi.org/10.3390/educsci15040403.
[15] S. Luftensteiner and J. Schrammel, "Streamlining AI Model Development and Evaluation in Industrial Setting by Means of an Application," *7th International Conference on Industry of the Future and Smart Manufacturing*, vol. 2, no. 7, 2025.
[16] S.-Y. Lin, Y.-C. Chen, Y.-H. Chang, S.-H. Lo, and K.-M. Chao, "Text–Image Multimodal Fusion Model for Enhanced Fake News Detection," *Science Progress*, vol. 107, no. 4, 2024. doi: https://doi.org/10.1177/00368504241292685.
[17] P. Zhu, J. Hua, K. Tang, J. Tian, J. Xu, and X. Cui, "Multimodal Fake News Detection Through Intra-Modality Feature Aggregation and Inter-Modality Semantic Fusion," *Complex & Intelligent Systems*, vol. 10, pp. 5851–5863, 2024. doi: https://doi.org/10.1007/s40747-024-01473-5.
[18] Y. Wang, W. Ma, Z. Jin, and X. Guo, "Multimodal Fake News Detection via Progressive Fusion Networks," *Information Processing & Management*, vol. 59, no. 6, p. 103120, 2022. doi: https://doi.org/10.1016/j.ipm.2022.103120.
[19] S. Tufchi, A. Yadav, and T. Ahmed, "AMTCF: An Advanced Multimodal Transformer and ConvNext Fusion for Contextualized Fake News Detection in Digital Landscape," *Language Resources and Evaluation*, vol. 59, pp. 2893–2927, 2025. doi: https://doi.org/10.1007/s10579-025-09838-z.
[20] M. Xu, F. Li, Z. Miao, Z. Han, L. Wang, and G. Wang, "Detecting Fake News on Social Media via Multimodal Semantic Understanding and Enhanced Transformer Architectures," *Traitement du Signal*, vol. 42, no. 3, pp. 1553–1564, 2025. doi: https://doi.org/10.18280/ts.420327.
[21] Z. Chi, P. Guo, and F. Liu, "A Compact GPT-Based Multimodal Fake News Detection Model with Context-Aware Fusion," *Electronics*, vol. 14, no. 23, p. 4755, 2025. doi: https://doi.org/10.3390/electronics14234755.
[22] E. Choi, J. Ahn, X. Piao, and J.-K. Kim, "CroMe: Multimodal Fake News Detection Using Cross-Modal Tri-Transformer and Metric Learning," *IEEE Access*, 2025. doi: https://doi.org/10.1109/ACCESS.2025.3633841.
[23] L. Hu, Z. Zhao, W. Qi, X. Song, and L. Nie, "Multimodal Matching-Aware Co-Attention Networks with Mutual Knowledge Distillation for Fake News Detection," *arXiv preprint arXiv:2212.05699*, 2022. Available: arXiv:2212.05699
[24] N. M. D. Tuan and P. Q. N. Minh, "Multimodal Fusion with BERT and Attention Mechanism for Fake News Detection," *arXiv preprint arXiv:2104.11476*, 2021. Available: arXiv:2104.11476
[25] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Lv, C. Guo, and Y. Yu, "Improving Fake News Detection by Using an Entity-Enhanced Framework to Fuse Diverse Multimodal Clues," *arXiv preprint arXiv:2108.10509*, 2021. Available: arXiv:2108.10509