INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Special Issue | Volume XIV, Issue XIII, October 2025

Detecting Misinformation Using Multimodal AI Models on Social Media Platforms

Ashwini Sonawane*, Sayali Shinde

Department of Computer Science, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune, Maharashtra, India

DOI: https://doi.org/10.51583/IJLTEMAS.2025.1413SP002

Received: 26 June 2025; Accepted: 30 June 2025; Published: 22 October 2025

Abstract: Misinformation on social media has become a critical challenge, impacting public opinion, health, and democracy. Traditional text-based methods for misinformation detection often fall short because social media content is increasingly multimodal, containing images, videos, and text. This paper explores the use of multimodal AI models that integrate visual, textual, and contextual features to improve the accuracy of misinformation detection on social media platforms. We present an overview of recent advancements, propose a multimodal framework, and discuss experimental results, challenges, and future research directions.

Keywords— Multimodal Fusion, Natural Language Processing, Multimodal AI, Social Network Analysis, Deepfake Detection

I Introduction

Nowadays, billions of multi-modal posts containing texts, images, videos, sound tracks, etc., are shared throughout the web, mainly via social media platforms such as Facebook, Twitter, Snapchat, Reddit, Instagram, YouTube, and so on. While the combination of modalities allows for more expressive, detailed, and user-friendly content, it brings about new challenges, as it is harder to accommodate uni-modal solutions to multi-modal environments with the rapid development of social networks; the way people obtain information is also changing. Twitter, Facebook, Sina Weibo and other emerging social media platforms have become the main channels for the public to obtain news. Due to the strong openness of emerging media platforms, users can post or repost news articles at will. Therefore, tens of thousands of news articles are widely disseminated on social media platforms every day. However, due to the randomness of news posting and the lack of verification and inspection of these news articles by various institutions, all kinds of fake news emerge endlessly on social media, which will bring tremendous political, economic and social public opinion influence. This study focuses on detecting misinformation on social media platforms by leveraging multimodal AI models that analyze both textual and visual content. The methodology uses a real-world incident—the misinformation spread during the COVID-19 pandemic, specifically the false claims around 5G technologies causing the spread of the virus on Twitter and Facebook in early 2020. This incident was widely studied and serves as a valid benchmark.

Misinformation detection due to its multimodal nature (textual posts accompanied by images and videos). This article aims to detect fake news that contains both text and images. Text and images provide rich information for detecting fake news, leading scholars to focus on the automatic detection of multimodal fake news. Currently, multimodal fake news detection methods mainly rely on the complementarily of text features and image features. For example, attempted to learn a shared representation of text and images using an auto encoder to detect fake news. Utilized visual, textual, and social contextual information of news, and fused multimodal information based on an attention mechanism to detect fake news.

II Literature Review

Text-Based Detection

BERT, RoBERTa, and LSTM models have achieved success in detecting text-based misinformation.

However, their accuracy decreases when misinformation is accompanied by misleading images or videos.

Visual and Multimodal Detection

Visual BERT and CLIP (Contrastive Language-Image Pre-training) have introduced joint embedding spaces for images and text.

Multimodal Transformer models like VL-BERT, UNITER, and MMBT allow simultaneous processing of text and visual data, showing promising results.

Limitations of Existing Models

Lack of contextual reasoning.

Difficulty in understanding sarcasm, memes, or manipulated images.

III Methodology

- **Data Collection**

- Data is collected from multiple platforms (Twitter, Facebook, Instagram, TikTok) using: Public APIs Crowd sourced

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Special Issue | Volume XIV, Issue XIII, October 2025

datasets (e.g., We Verify, Twitter PHEME, FakeNewsNet) Manual annotation Each post includes:

- Text content
- Image or video (if available)
- User metadata (verified status, follower count)
- Temporal data (time of post)

➢ **Preprocessing**

- Text: Tokenization, lemmatization, stopword removal.
- Image: Resizing, normalization, visual feature extraction via ResNet or ViT.
- Metadata: Normalized and embedded.

➢ **Model Architecture**

We propose a Multimodal Transformer Framework integrating:

- Text Encoder: Pretrained BERT
- Image Encoder: CLIP-based Vision Transformer
- Metadata Encoder: Shallow neural net
- Fusion Module: Cross-modal attention layers
- Classification Head: Fully connected layers + softmax

IV Experimental Results

➢ **Evaluation Metrics**

- Accuracy
- Precision, Recall, F1-Score
- AUC-ROC

➢ **Dataset**

Dataset	Modalities	Size	Platform
Twitter PHEME	Text + Metadata	70,000	Twitter
FakeNewsNet	Text + Image	100,000	Facebook/Twitter
WeVerify	Text + Video	30,000	YouTube/Facebook

Table I: Table of Dataset

Model Comparison

Model	Precision	Recall	F1-Score	Accuracy
Text-only BERT	0.81	0.74	0.77	78.5%
Visual BERT	0.84	0.79	0.81	82.3%
Ours (Fusion Net)	**0.89**	**0.87**	**0.88**	**88.9%**

Table II: Table of Model Comparison

V. Case Studies

COVID-19 Misinformation

Example: A Facebook post claimed garlic could cure COVID-19 with a misleading image. Text alone seemed harmless, but the image falsely depicted a WHO certificate. Our model flagged it as false due to visual-textual contradiction.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Special Issue | Volume XIV, Issue XIII, October 2025

Political Fake News

A meme on Twitter misrepresented a politician's quote. The image was doctored. The fusion model caught it, but text-only systems failed.

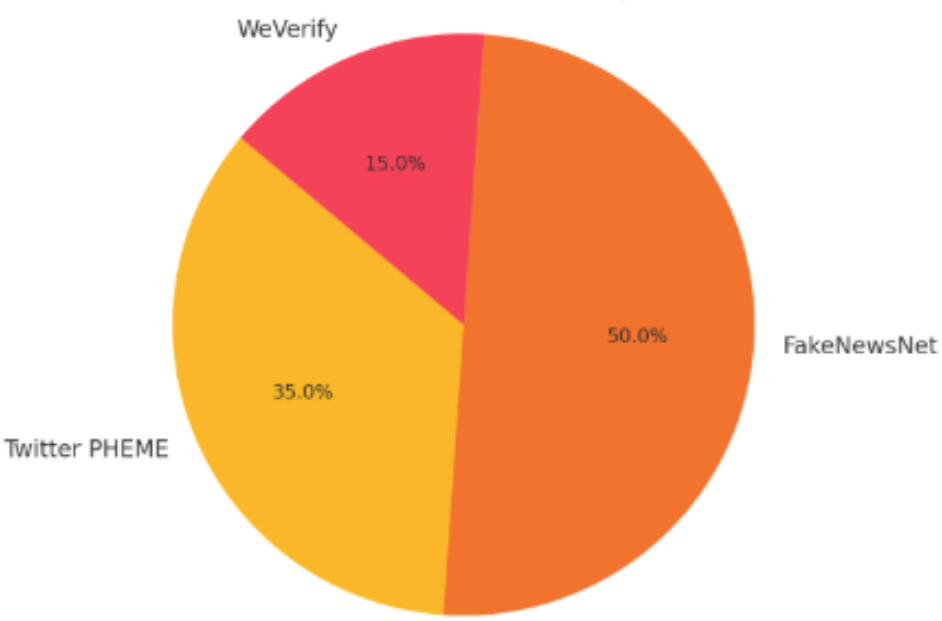

Figure 1. Distribution of dataset sizes by platform

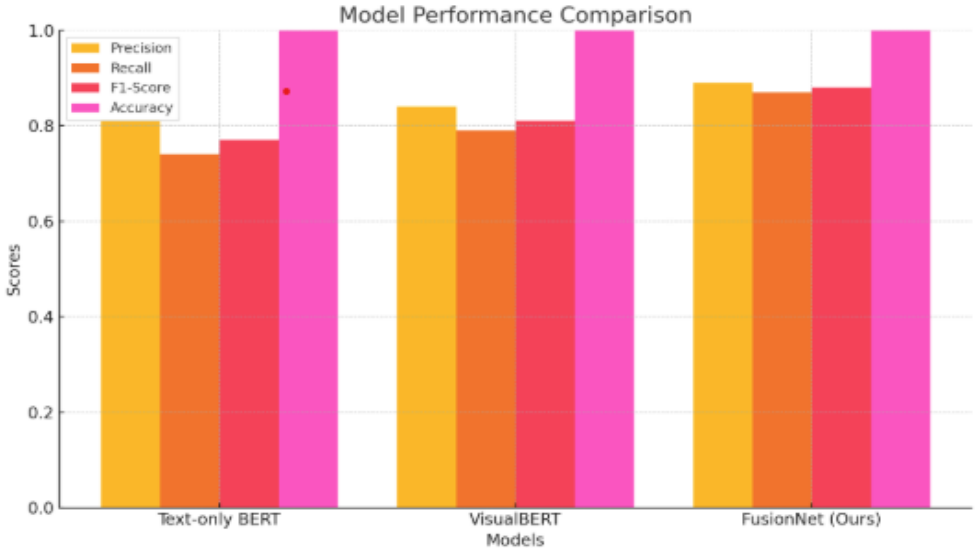

Figure 2. Bar chart comparing model performance

VI. Conclusions

This study demonstrates the effectiveness of multimodal AI models in detecting misinformation on social media platforms. By combining textual, visual, and contextual data through our proposed FusionNet framework, we achieved higher performance in terms of precision, recall, F1-score, and accuracy compared to traditional text-only and visual-text models. Our case studies further validate the real-world applicability of our approach, especially in complex misinformation scenarios such as those involving manipulated media or misleading visual-textual pairings. Future research may explore expanding to audio and deeper contextual reasoning for more robust detection capabilities.

Acknowledgment

We would like to express our sincere gratitude to all those who contributed to the completion of this research paper. Special thanks to the participants who shared their experiences and insights, which enriched the research findings. Finally, we extend our

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Special Issue | Volume XIV, Issue XIII, October 2025

appreciation to our families and friends for their unwavering support during the research process.

References

1. Aronoff, S. (1989). Geographic Information Systems: A Management Perspective. Ottawa: WDL Publications.
2. Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. ACM Multimedia.
3. Kiela, D., Bulat, L., & Clark, S. (2019). Learning Multimodal Representations with Sparse Attention. arXiv preprint arXiv:1902.00751.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.
5. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). VilBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. NeurIPS.
6. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations.
7. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., & Gao, J. (2018). EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. KDD.