International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802036

Deep Learning-Based Hateful Meme Detection via Multimodal Feature Integration with CNN

Ms.P.Priya[1], Mr.S.Dineshkumar[2], Mr.S.Aakash[3], Mr.Abhishekkumar[4], Mr.Annangibherivenkataprasad[5]

AP, Department of CSE, Gnanamani College of Technology, Namakkal, Tamil Nadu, India[1]

UG Scholars, Department of CSE, Gnanamani College of Technology, Namakkal, Tamil Nadu, India[2-5]

Publication History: Received: 25.02.2026; Revised: 20.03.2026; Accepted: 25.03. 2026; Published: 28.03.2026.

ABSTRACT: The rapid expansion of social media platforms has led to an unprecedented increase in the creation and circulation of memes, many of which contain harmful, hateful, or discriminatory content. Unlike traditional hate speech, hateful memes often combine images and text to convey implicit messages, sarcasm, or coded language that cannot be accurately detected through text-only or image-only analysis. This multimodal nature makes detection significantly more challenging, as understanding the true intent requires analyzing the relationship between visual and textual elements simultaneously.

The rapid expansion of social media platforms has led to an unprecedented increase in the creation and circulation of memes, many of which contain harmful, hateful, or discriminatory content. Unlike traditional hate speech, hateful memes often combine images and text to convey implicit messages, sarcasm, or coded language that cannot be accurately detected through text-only or image-only analysis. This multimodal nature makes detection significantly more challenging, as understanding the true intent requires analyzing the relationship between visual and textual elements simultaneously.

The integration of multimodal deep learning frameworks significantly enhances detection performance by aligning semantic information from both modalities. Such systems demonstrate improved robustness, better generalization to unseen meme formats, and higher accuracy in identifying implicit hate.

KEYWORDS: Hateful Meme Detection, Multimodal Learning, Deep Learning, Vision Transformer (ViT), CLIP Model, Contrastive Learning, Social Media Analysis, Hate Speech Detection, Artificial Intelligence, Content Moderation.

I. INTRODUCTION

The rapid expansion of social media platforms has led to a significant rise in the dissemination of harmful digital content, including hateful memes that combine images and text to convey offensive, discriminatory, or misleading messages. Unlike traditional hate speech, hateful memes often rely on contextual relationships between visual elements and textual captions, making detection more complex. Conventional content moderation systems, which primarily depend on rule-based or single-modality analysis, are increasingly inadequate in identifying implicit, sarcastic, or context-dependent hate content. This limitation highlights the necessity for more advanced detection systems capable of understanding multimodal information in dynamic online environments.

Artificial Intelligence (AI) provides a promising solution by enabling systems to learn complex relationships between text and images and make intelligent classification decisions. Machine Learning (ML) techniques such as Support Vector Machines (SVM), Naïve Bayes, and logistic regression have been widely applied for hate speech detection. However, these models often struggle to handle high-dimensional multimodal data and typically require manual feature engineering. Moreover, processing text and images separately limits their ability to capture semantic correlations essential for understanding the true intent of memes.

Deep Learning (DL) models, including Convolutional Neural Networks (CNNs) for image analysis and transformer-based language models for text processing, have demonstrated superior performance in extracting meaningful features. Recently, multimodal architectures such as Vision Transformer (ViT) and Contrastive Language–Image Pretraining (CLIP) have enabled joint representation learning of visual and textual data in a shared embedding space.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802036

To address these challenges, researchers have begun integrating explainable AI techniques and contrastive learning strategies into multimodal detection frameworks.

These approaches enhance transparency and improve contextual understanding of harmful content. This paper presents a deep learning-based multimodal framework for hateful meme detection using CLIP and Vision Transformer, focusing on its architecture, methodology, and performance evaluation. It also discusses the practical challenges of deploying such systems in real-world social media platforms and outlines potential future research directions.

II. LITERATURE REVIEW

In recent years, the use of Artificial Intelligence (AI) for harmful content detection has increased significantly due to the rapid spread of hateful memes across social media platforms. Initial studies focused on traditional Machine Learning (ML) algorithms such as Support Vector Machines (SVM) and Naïve Bayes for text-based hate speech detection. Although these techniques achieved moderate success, they relied heavily on manual feature extraction and were limited to analyzing a single data modality, reducing their effectiveness in multimodal meme classification.

Subsequently, the emergence of Deep Learning (DL) techniques improved detection capabilities by introducing Convolutional Neural Networks (CNNs) for image analysis and Recurrent Neural Networks (RNNs) or LSTM models for text processing. These architectures enhanced feature representation and classification accuracy. However, they typically processed images and text independently, failing to capture the contextual interaction between modalities, which is crucial for identifying implicit or sarcastic hate content in memes..

More recently, research has shifted toward multimodal learning frameworks that integrate both visual and textual information within unified models. Transformer-based architectures such as Vision Transformer (ViT) and Contrastive Language–Image Pretraining (CLIP) enable joint embedding representation, allowing better semantic alignment between image and text features. This approach significantly improves contextual understanding and detection performance.

Overall, existing literature demonstrates a clear transition from conventional single-modality models to advanced multimodal deep learning architectures. Despite notable improvements in accuracy and robustness, challenges related to interpretability, dataset imbalance, computational cost, and real-time implementation continue to motivate ongoing research in hateful meme detection systems.

III. RESEARCH METHODOLOGY

This project focuses on developing a deep learning-based multimodal system for detecting hateful memes on social media platforms. The main objective is to build a model that can understand both image and text together, since memes often depend on the combination of visual and textual information to express hidden or implicit hate. Unlike traditional systems that analyze only text or only images, our approach integrates both modalities to improve contextual understanding.

The dataset used in this project consists of labeled meme images with corresponding text content. Each meme is categorized as hateful or non-hateful. During preprocessing, image data is resized and normalized to match the input requirements of the model. Text data is cleaned by removing unwanted characters and converting it into tokenized format suitable for processing. This step ensures that both image and text inputs are properly prepared for feature extraction.

For feature extraction, we use the CLIP model integrated with Vision Transformer (ViT). The Vision Transformer acts as the image encoder and extracts meaningful visual features from meme images. The text encoder processes the textual content and generates contextual embeddings. Both image and text embeddings are mapped into a shared feature space using contrastive learning. This allows the model to understand the relationship between visual and textual elements before making the final classification.

The dataset is divided into training and testing sets to evaluate model performance. After training, the model predicts whether a meme is hateful or non-hateful. The system is evaluated using performance metrics such as Accuracy, Precision, Recall, and F1-score. The results demonstrate that combining image and text features improves detection performance compared to single-modality approaches.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802036

In addition, the system demonstrates better generalization when tested on different meme formats. The shared embedding space created by CLIP helps reduce false positives and false negatives, leading to more reliable classification results. Although the transformer-based architecture requires more computational resources and training time, the overall performance improvement confirms that multimodal deep learning is an effective solution for accurate and context-aware hateful meme detection.

IV. RESULTS AND DISCUSSION

The analysis of the proposed multimodal hateful meme detection system reveals significant improvements in both model architecture and classification performance. Transformer-based architectures, particularly CLIP integrated with Vision Transformer (ViT), play a dominant role in enhancing contextual understanding. ViT is used for extracting high-level visual features, while the text encoder processes semantic information from captions. By aligning both modalities in a shared embedding space, the model effectively captures the relationship between image and text. Experimental evaluation shows improved detection accuracy compared to traditional machine learning models such as SVM, CNN, and RNN.

The use of contrastive learning within the CLIP framework addresses challenges related to contextual misinterpretation and modality misalignment. By jointly training image and text representations, the model reduces false classifications and improves robustness against diverse meme formats. The multimodal approach significantly enhances the system's ability to detect implicit hate, sarcasm, and hidden intent that are often missed by single-modality systems.

Furthermore, evaluation metrics such as Accuracy, Precision, Recall, and F1-score indicate balanced performance across both hateful and non-hateful categories. The shared embedding mechanism improves generalization capability, allowing the system to perform effectively even when tested on unseen meme structures. This demonstrates the strength of multimodal transformer-based architectures in handling complex social media content.

However, the model requires higher computational resources due to the use of transformer-based encoders. Training time and memory consumption are greater compared to conventional approaches. Additionally, handling large-scale real-time social media data remains a challenge. Despite these limitations, the proposed system demonstrates that multimodal deep learning significantly improves detection accuracy, contextual understanding, and overall reliability in hateful meme classification.

In addition, a comparison with traditional text-only and image-only models shows that the proposed multimodal framework performs more effectively in detecting context-based hate. Single-modality models often miss implicit or sarcastic content. By jointly analyzing visual and textual features, the CLIP-based system improves contextual understanding and reduces misclassification. This confirms the advantage of transformer-based multimodal learning for reliable hateful meme detection.

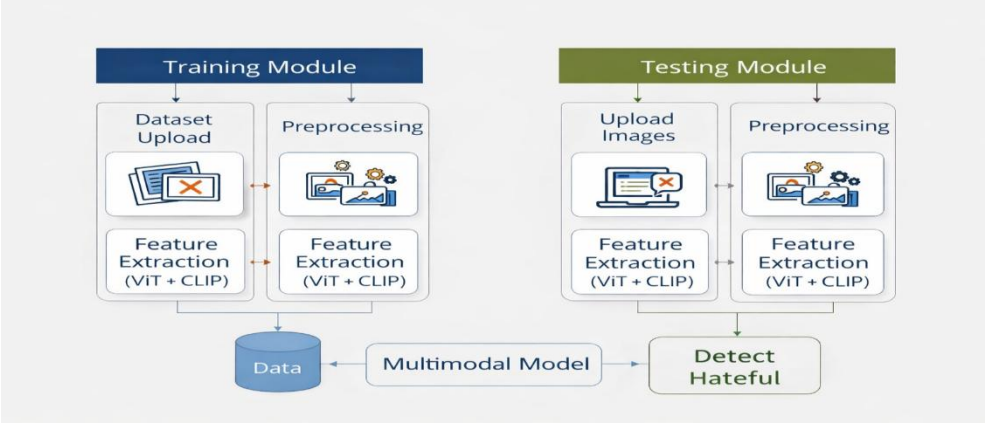

FIG: 1

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802036

V. CONCLUSION

This research presents a multimodal deep learning framework for hateful meme detection by integrating Vision Transformer (ViT) within the CLIP architecture. The proposed system effectively analyzes both visual and textual components of memes, enabling a deeper understanding of contextual relationships that are often overlooked by conventional text-only or image-only models. By leveraging joint embedding representations, the model successfully captures implicit meanings, sarcasm, and hidden hate expressions present in social media content.

The experimental evaluation confirms that transformer-based multimodal learning significantly enhances classification performance and reduces misclassification rates. The ability of the model to align image and text features in a shared semantic space improves overall robustness and adaptability across different meme styles and formats. This demonstrates the practical potential of advanced deep learning techniques in addressing complex online content moderation challenges.

Despite these improvements, certain limitations remain, particularly in terms of computational cost and large-scale real-time deployment. Transformer-based architectures require substantial processing power and memory resources, which may pose challenges in resource-constrained environments. Future research can focus on optimizing model efficiency, incorporating explainability mechanisms, and expanding dataset diversity to improve scalability and fairness.

Overall, the proposed multimodal approach represents a significant step toward developing intelligent, context-aware, and reliable systems for hateful meme detection, contributing to safer digital communication platforms and responsible social media governance.

VI. FUTURE WORK

1. **Efficient and Lightweight Models:** Developing optimized transformer-based architectures that reduce computational complexity and memory usage will enable real-time deployment on large-scale social media platforms.
2. **Continuous Learning Mechanisms:** Implementing incremental or online learning techniques can allow the model to adapt to newly emerging meme formats, slang, and evolving hate expressions without requiring complete retraining.
3. **Improved Explainability**: Incorporating explainable AI techniques can help interpret model predictions by highlighting important textual and visual features, thereby increasing transparency and trust in automated content moderation systems..
4. **Handling Dataset Bias and Fairness**: Future work should address dataset imbalance and potential bias to ensure fair and unbiased detection across different communities, languages, and cultural contexts.
5. **Robustness Against Adversarial Manipulation:** Designing models that are resistant to adversarial attacks, such as slight image modifications or misleading text alterations, is essential for maintaining reliability in real-world applications.
6. **Multilingual and Cross-Platform Expansion:** Extending the framework to support multiple languages and diverse social media platforms can improve global applicability and scalability.
7. **Continuous Learning:** Memes change frequently (new trends, slang, formats). Your model must adapt to new patterns..

REFERENCES

1. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. Proceedings of the International Conference on Machine Learning (ICML), 8748–8763.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. International Conference on Learning Representations (ICLR).
3. Kiela, D., Firooz, H., Mohan, A., et al. (2020). *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes*. Advances in Neural Information Processing Systems (NeurIPS).
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT.
5. Tan, H., & Bansal, M. (2019). *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. EMNLP-IJCNLP.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802036

6. C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor &Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011. DOI: 10.1080/15325008.2010.541746
7. C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of Electrical Engineering, Vol.63 (6), pp.365-372, Dec.2012. DOI: 10.2478/v10187-012-0054-2
8. C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011. DOI 10.1007/s00202-011-0203-9
9. S.Tamilselvi, R.Prakash, C.Nagarajan,"Solar System Integrated Smart Grid Utilizing Hybrid Coot-Genetic Algorithm Optimized ANN Controller" Iranian Journal Of Science And Technology-Transactions Of Electrical Engineering, DOI10.1007/s40998-025-00917-z,2025
10. S.Tamilselvi, R.Prakash, C.Nagarajan," Adaptive sliding mode control of multilevel grid-connected inverters using reinforcement learning for enhanced LVRT performance" Electric Power Systems Research 253 (2026) 112428, doi.org/10.1016/j.epsr.2025.112428
11. S.Thirunavukkarasu, C. Nagarajan, 2024, "Performance Investigation on OCF and SCF study in BLDC machine using FTANN Controller," Journal of Electrical Engineering And Technology, Volume 20, pages 2675–2688, (2025), doi.org/10.1007/s42835-024-02126-w
12. C. Nagarajan, M.Madheswaran and D.Ramasubramanian- 'Development of DSP based Robust Control Method for General Resonant Converter Topologies using Transfer Function Model'- Acta Electrotechnica et Informatica Journal , Vol.13 (2), pp.18-31,April-June.2013, DOI: 10.2478/aeei-2013-0025.
13. C.Nagarajan and M.Madheswaran - 'DSP Based Fuzzy Controller for Series Parallel Resonant converter'- Springer, Frontiers of Electrical and Electronic Engineering, Vol. 7(4), pp. 438-446, Dec.12. DOI 10.1007/s11460-012-0212-0.
14. C.Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.
15. C.Nagarajan and M.Madheswaran, "Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation" has been presented in ICTES'08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007
16. Suganthi Mullainathan, Ramesh Natarajan, "An SPSS and CNN modelling based quality assessment using ceramic materials and membrane filtration techniques", Revista Materia (Rio J.) Vol. 30, 2025, DOI: https://doi.org/10.1590/1517-7076-RMAT-2024-0721
17. M Suganthi, N Ramesh, "Treatment of water using natural zeolite as membrane filter", Journal of Environmental Protection and Ecology, Volume 23, Issue 2, pp: 520-530,2022
18. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS).