Explainable Multimodal Hate Speech Detection Using Cross-Attention Transformer Fusion Networks

Yaseen Malik[1], Dr. Meera Alphy[2]
[1]*M.Tech, Student of Computer Science and Engineering, MGIT Autonomous, TG, Hyderabad, India*
[2]*M.Tech, Ph.D Assistant Professor of Computer Science and Engineering, MGIT Autonomous, TG, Hyderabad, India*

***Abstract*—The widespread adoption of social media platforms has led to a significant increase in the dissemination of hateful and offensive content through text, images, memes, and other multimodal formats. Conventional hate speech detection systems predominantly rely on textual analysis and often fail to capture the contextual relationships between visual and linguistic information. To address these challenges, this work presents an explainable multimodal hate speech detection framework that combines advanced transformer-based models for comprehensive content understanding. The proposed architecture employs XLM-RoBERTa to extract contextual multilingual textual representations and CLIP Vision Transformer (ViT-B/32) to learn semantic visual features from images. A Cross-Attention Fusion Network is introduced to establish meaningful interactions between textual tokens and visual regions, enabling effective identification of implicit, symbolic, and context-dependent hate expressions. To improve model transparency and support trustworthy decision-making, explainable artificial intelligence techniques, including SHAP and Grad-CAM, are integrated to highlight influential textual elements and image regions contributing to the classification outcome. The framework is evaluated on the MultiOFF benchmark dataset using standard performance metrics. Experimental results demonstrate strong classification capability, achieving high accuracy, precision, recall, F1-score, and Area Under the Curve (AUC), while maintaining interpretability of predictions. The findings indicate that the proposed transformer-based multimodal fusion approach provides a reliable and scalable solution for automated hate speech detection and intelligent content moderation across diverse online environments.**

***Index Terms*—Multimodal Hate Speech Detection, Explainable Artificial Intelligence (XAI), Cross-Attention Fusion Network, XLM-RoBERTa, CLIP Vision Transformer, Social Media Content Moderation, Transformer-Based Learning, Deep Learning.**

I. INTRODUCTION

The rapid growth of social media platforms has revolutionized digital communication, enabling users to share opinions, images, videos, and multimedia content on a global scale. While these platforms promote information exchange and social interaction, they have also become channels for the dissemination of hate speech, offensive content, and discriminatory narratives targeting individuals and communities based on race, religion, ethnicity, gender, or nationality. The increasing volume of user-generated content makes manual moderation difficult and resource-intensive, creating a strong demand for automated hate speech detection systems [1], capable of identifying harmful content accurately and efficiently. Traditional hate speech detection approaches primarily rely on textual analysis using machine learning and natural language processing techniques. Although these methods have demonstrated promising results, they often fail to capture the contextual relationships that exist between textual and visual information. In social media memes and multimodal posts, hateful intent is frequently conveyed through the combination of images and text rather than either modality alone. Consequently, unimodal systems struggle to recognize implicit hate, sarcasm, symbolic representations, and context-dependent offensive expressions. Recent advances in transformer-based architectures and vision-language models [2], [3], [10] have enabled more effective multimodal learning by jointly analyzing textual and visual content, leading to improved detection performance.

Despite these advancements, challenges remain in developing systems that are both accurate and interpretable. Many deep learning models operate as black boxes, providing limited insight [7] into the reasoning behind their predictions. In sensitive applications such as content moderation, transparency and accountability are essential for ensuring fairness and user trust. To address these limitations, this paper proposes an Explainable Multimodal Hate Speech Detection Framework that integrates XLM-RoBERTa for textual feature extraction, CLIP Vision Transformer for visual representation learning, and a Cross-Attention Fusion Network for multimodal feature integration. Additionally, SHAP and Grad-CAM [11], [12] are incorporated to provide textual and visual explanations, enabling transparent and reliable hate speech classification across multimodal social media content.

II. RELATED WORK

The rapid growth of social media platforms has led to an increasing volume of hateful and offensive content being shared through text, images, videos, and memes. This trend has encouraged researchers to develop automated hate speech detection systems capable of identifying harmful content efficiently. Early research primarily focused on text-based analysis using traditional machine learning techniques such as Support Vector Machines, Naïve Bayes, and Logistic Regression. While these approaches achieved moderate success, they were unable to capture the complex contextual information often present in multimodal social media content.

To address the limitations of text-only methods, researchers began exploring multimodal hate speech detection approaches that jointly analyze textual and visual information. The Hateful Memes Challenge introduced by Kiela et al. [1] established a benchmark dataset that demonstrated the necessity of combining image and text understanding for accurate hate speech detection. Their work showed that unimodal models often fail when hateful intent emerges only from the interaction between textual and visual content.

The emergence of transformer-based architectures significantly improved multimodal learning capabilities. Chhabra and Vishwakarma [2] proposed MHS-STMA, a scalable transformer-based multilevel attention framework that effectively captures interactions between textual and visual features. The model demonstrated improved classification performance compared to traditional fusion methods by leveraging attention mechanisms to focus on relevant multimodal information.

Further advancements were achieved through attention-based fusion strategies. Mandal et al. [3] introduced a transformer-driven attentive fusion approach that models inter-modal dependencies between images and text. Their framework improved the detection of implicit hate speech and context-dependent offensive content by learning richer multimodal representations. However, the model required large-scale datasets and significant computational resources for optimal performance.

Recognizing the importance of multilingual and multicultural understanding, Bui et al. proposed Multi³Hate [4], a multimodal and multilingual hate speech detection framework based on vision-language models. The study introduced culturally diverse annotations and multilingual meme datasets, enabling cross-cultural evaluation of hate speech detection systems. Although the framework advanced multilingual research, its dataset size remained relatively limited for training large transformer architectures.

Several recent studies have focused on improving multimodal fusion mechanisms. Zhang et al. [5] proposed CMFusion, a channel-wise and modality-wise fusion framework that selectively emphasizes informative features across different modalities. Similarly, Yang and Zhu [6] introduced a cross-domain knowledge transfer approach that improves model generalization across heterogeneous datasets. These methods enhanced robustness and fusion effectiveness but provided limited support for explainability and multilingual processing.

Explainability has emerged as a critical requirement for trustworthy hate speech detection systems. HateXplain developed by Mathew et al. [7] introduced human-annotated rationales for text-based hate speech classification, enabling better understanding of model decisions. More recently, Salles et al. [8] proposed a benchmark dataset containing multimodal human-annotated explanations for both textual and visual content. These studies highlighted the importance of transparent AI systems

but also revealed the need for scalable explainability methods in real-world applications.

In multimodal hate speech detection, several research gaps remain. Many existing approaches focus primarily on improving classification accuracy while overlooking interpretability, multilingual adaptability, and robust multimodal feature alignment. Furthermore, most models are evaluated on limited datasets and lack comprehensive explanations for their predictions. To address these challenges, the proposed framework integrates XLM-RoBERTa for multilingual textual representation, CLIP Vision Transformer for visual feature extraction, a Cross-Attention Fusion Network for deep multimodal interaction, and explainability techniques such as SHAP and Grad-CAM to provide transparent and reliable hate speech detection.

III. METHODOLOGY

A. Overall Framework

The proposed Explainable Multimodal Hate Speech Detection Framework is designed to automatically identify hateful content by jointly analyzing textual and visual information from social media posts, memes, and image–text pairs. The framework follows a multimodal learning approach in which text and images are processed through separate transformer-based encoders to capture semantic and contextual information from each modality. This design enables the system to recognize both explicit and implicit hate expressions that are often overlooked by traditional text-only detection methods.

For textual analysis, the framework utilizes XLM-RoBERTa [9] to generate contextual multilingual embeddings capable of handling diverse languages and code-mixed content. Simultaneously, visual features are extracted using the CLIP Vision Transformer (ViT-B/32) [10], which captures high-level semantic information from images, including symbols, objects, gestures, and meme-specific cues. The extracted textual and visual representations are then integrated through a Cross-Attention Fusion Network (CAFN), allowing the model to learn meaningful relationships between words and image regions while producing a unified multimodal representation.

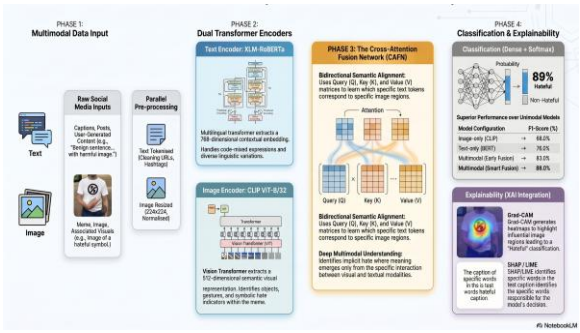


Fig. 1 System Architecture

The fused feature representation is passed to a classification module that determines whether the input content is hateful or non-hateful. To improve transparency and trustworthiness, the framework incorporates Explainable Artificial Intelligence (XAI) techniques. SHAP is employed [11] to identify influential textual features, while Grad-CAM highlights important image regions [12] that contribute to the final prediction. By combining transformer-based feature extraction, cross-attention fusion, and explainability mechanisms, the proposed framework provides an accurate, interpretable, and scalable solution for multimodal hate speech detection.

B. Dataset Description

The proposed framework is evaluated using the MultiOFF dataset [13], a benchmark multimodal dataset developed for offensive and hateful content detection. The dataset consists of image–text pairs collected from social media platforms, where each sample contains an image, its associated textual content, and a corresponding class label indicating whether the content is offensive (Hate) or non-offensive (non-hate). The availability of both textual and visual modalities makes the dataset suitable for training and evaluating multimodal deep learning models.

Fig.2 Dataset Images from both Classes

The dataset is organized into training, validation, and testing subsets to facilitate model development and

performance assessment. The training set is used for learning multimodal feature representations, while the validation set assists in hyperparameter tuning and model selection. The testing set is reserved for evaluating the final model and measuring its generalization capability on unseen data. Images are linked to their corresponding textual descriptions through unique identifiers, ensuring proper alignment between modalities during the learning process.

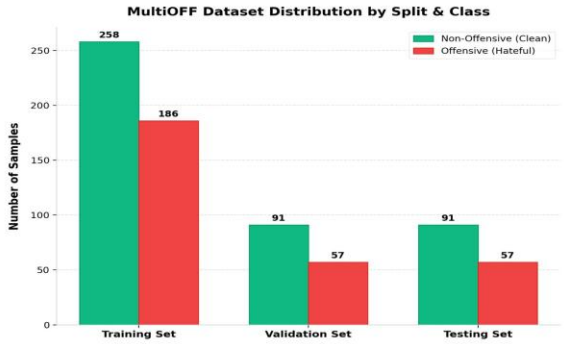


Fig.3 Dataset Distribution

The MultiOFF dataset provides diverse examples of explicit and implicit offensive expressions represented through text, images, and their interactions. This diversity enables the proposed model to learn complex semantic relationships across modalities and improves its ability to detect context-dependent hate speech. The dataset serves as a reliable benchmark for assessing the effectiveness of transformer-based multimodal fusion techniques and explainable artificial intelligence methods in hate speech detection tasks.

C. Data Preparation and Image Processing

Data preparation is a crucial stage in the proposed multimodal hate speech detection framework, as it ensures that both textual and visual data are transformed into a consistent format suitable for model training and evaluation. Prior to model training, several data preparation and image processing procedures were performed to ensure consistency, improve data quality, and enhance multimodal feature learning. Since the MultiOFF dataset contains both textual and visual content collected from social media platforms, preprocessing is essential to handle variations in image quality, resolution, and formatting. The objective of this stage is to generate standardized inputs that can be effectively processed by transformer-based models for hate speech detection.

Image Standardization: All images were converted into a uniform format before being supplied to the visual feature extraction module. The images were resized to a fixed resolution of 224 × 224 pixels [10], matching the input requirements of the CLIP Vision Transformer (ViT-B/32). Furthermore, pixel intensity values were normalized according to the preprocessing parameters of the pretrained CLIP model. This standardization process ensures consistency across all samples and improves computational efficiency during training and inference.

Image Processing: Social media images often contain variations in brightness, contrast, resolution, and visual complexity. To improve feature extraction quality, preprocessing operations were applied to remove inconsistencies and prepare images for analysis. The processed images were transformed into tensor representations compatible with deep learning frameworks. This step enables the visual encoder to effectively capture semantic information such as objects, symbols, gestures, and contextual visual cues that may contribute to hateful or offensive content.

Data Augmentation: To improve model robustness and reduce overfitting, data augmentation techniques were applied to the training images. Controlled transformations including random rotation, horizontal flipping, zooming, and random cropping were employed to generate diverse image variations while preserving their semantic meaning. These augmented samples increase dataset diversity and enable the model to learn invariant visual representations. As a result, the framework achieves better generalization performance when evaluating previously unseen multimodal content.

D. Proposed Cross-Attention Transformer Fusion Network

The proposed architecture combines transformer-based language and vision models with an attention-driven multimodal fusion mechanism for explainable hate speech detection. The framework is designed to jointly analyze textual and visual information present in social media posts, memes, and image-text pairs. By integrating contextual language understanding, semantic visual representation learning, and cross-

modal attention, the model effectively identifies both explicit and implicit forms of hateful content. The overall architecture of the proposed framework is illustrated in Fig. 1.

XLM-RoBERTa Text Encoder: The textual processing branch utilizes XLM-RoBERTa, a multilingual transformer model pretrained on large-scale multilingual corpora. The encoder converts input text into contextual embeddings that capture semantic meaning, syntactic relationships, and contextual dependencies. Unlike conventional word embedding approaches, XLM-RoBERTa generates dynamic representations that adapt according to the surrounding context, enabling effective handling of multilingual and code-mixed social media content [9].

Textual representation:

$$H_t = \text{XLM-RoBERTa}(T)$$

where (T) represents the input text and (H_t) denotes the contextual textual embeddings.

CLIP Vision Transformer Encoder: The visual processing branch employs the CLIP Vision Transformer (ViT-B/32) to extract semantic features from images. The model divides each image into fixed-size patches and processes them through transformer layers to learn high-level visual representations. This approach enables the network to capture objects, symbols, gestures, and contextual visual cues that may contribute to hateful or offensive meaning [10].

Visual representation:

$$H_v = \text{CLIP-ViT}(I)$$

where (I) represents the input image and (H_v) denotes the extracted visual feature embeddings.

Cross-Attention Fusion Network: To effectively integrate information from both modalities, a Cross-Attention Fusion Network (CAFN) is introduced [15]. The fusion mechanism enables textual features to attend to relevant visual regions while simultaneously allowing visual features to attend to important textual tokens. This bidirectional interaction captures complex semantic relationships that cannot be learned through simple concatenation-based fusion methods.

Cross-attention fusion:

$$F_{fusion} = \text{CrossAttention}(H_t, H_v)$$

where (F_{fusion}) represents the unified multimodal feature representation generated from textual and visual interactions.

Classification and Explainability Module: The fused feature representation is forwarded to a classification head consisting of fully connected layers, dropout regularization, and a SoftMax activation function. The classifier predicts whether the input content belongs to the hateful or non-hateful category. To enhance transparency and trustworthiness, Explainable Artificial Intelligence (XAI) techniques are incorporated into the framework. SHAP [11] is used to identify influential textual features, while Grad-CAM [12] generates visual attention maps highlighting image regions responsible for the final prediction. These explanations provide valuable insights into model behavior and support reliable decision-making in content moderation applications.

E. Model Training Strategy

The proposed multimodal framework was trained as a binary classification model to distinguish between hateful and non-hateful content. The training process aims to learn meaningful relationships between textual and visual modalities while maximizing classification performance. The model utilizes multimodal feature representations generated by XLM-RoBERTa, CLIP Vision Transformer, and the Cross-Attention Fusion Network to perform final classification.

Model optimization was performed using the AdamW [14] optimizer with an initial learning rate of 0.0001. The framework was trained using a batch size of 16 for a maximum of 10 epochs. To address potential class imbalance within the dataset, class weighting was incorporated during training to ensure balanced learning across both classes. The categorical cross-entropy loss function was employed as the optimization objective, enabling effective learning of multimodal feature representations.

To improve convergence and prevent overfitting, several training strategies were adopted. Early Stopping was utilized to terminate training when validation performance showed no further improvement, thereby preventing unnecessary computation and model overfitting. Reduce Learning

Rate on Plateau was applied to dynamically adjust the learning rate during training when validation loss stagnated. In addition, Model Checkpoint was used to preserve the best-performing model parameters based on validation performance, ensuring optimal classification accuracy on unseen test samples.

F Performance Evaluation Metrics

The effectiveness of the proposed framework was evaluated using multiple performance metrics to comprehensively assess classification accuracy, reliability, and discriminative capability. These metrics provide complementary insights into model performance and are widely used in hate speech detection research.

Accuracy: Accuracy represents the proportion of correctly classified samples among all evaluated instances and provides an overall measure of classification effectiveness.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision: Precision measures the proportion of correctly predicted hateful instances among all instances predicted as hateful.

$$\text{Pr}e\,cision = \frac{TP}{TP + FP}$$

A higher precision value indicates fewer false positive classifications.

Recall: Recall evaluates the ability of the model to correctly identify hateful content from all actual hateful instances.

$$Recall = \frac{TP}{TP + FN}$$

Higher recall values indicate improved detection sensitivity for hateful content.

F1-Score: The F1-Score combines precision and recall into a single metric, providing a balanced evaluation of classification performance.

$$F1\,Score = \frac{2 \times \text{Pr}e\,cision \times Recall}{\text{Pr}e\,cision + Recall}$$

This metric is particularly useful when dealing with class imbalance.

ROC–AUC: The Receiver Operating Characteristic–Area Under the Curve (ROC–AUC) measures the ability of the model to distinguish between hateful and non-hateful classes across different decision thresholds.

$$AUC = \int_{1}^{0} TPR(FPR)d(FPR)$$

Higher AUC values indicate stronger discriminative capability and better classification performance.

The model performance was analyzed using a confusion matrix and a classification report. These evaluation tools provide detailed insights into class-wise prediction behavior, enabling a deeper understanding of the strengths and limitations of the proposed multimodal hate speech detection framework.

IV RESULTS AND DISCUSSION

The proposed Explainable Multimodal Hate Speech Detection Framework was evaluated on the MultiOFF dataset to assess its effectiveness in identifying hateful and offensive content from multimodal social media posts. The experiments were conducted using a transformer-based architecture comprising XLM-RoBERTa for textual feature extraction, CLIP Vision Transformer for visual representation learning, and a Cross-Attention Fusion Network for multimodal integration. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics.

A. Experimental Setup

The dataset was divided into training, validation, and testing subsets to ensure reliable performance evaluation. All experiments were conducted using the AdamW optimizer with a learning rate of 0.0001 and a batch size of 16. Early Stopping, Model Checkpoint, and Reduce Learning Rate on Plateau were employed to improve convergence and prevent overfitting. Training was performed on a GPU-enabled environment to efficiently process multimodal data.

B. Training Performance

The training performance of the proposed framework was analyzed using training and validation accuracy as well as training and validation loss across multiple epochs. During the training process, the model demonstrated stable convergence, with accuracy gradually increasing while the loss consistently decreased. The validation curves closely followed the training curves, indicating effective learning and minimal overfitting.

The incorporation of transfer learning, cross-attention fusion, and adaptive optimization strategies

contributed to faster convergence and improved generalization performance. Early Stopping prevented unnecessary training once the validation performance stabilized, while the learning rate scheduling mechanism enabled smooth optimization throughout the training process. The observed learning behavior confirms the effectiveness of the proposed architecture in extracting meaningful multimodal representations from textual and visual content.

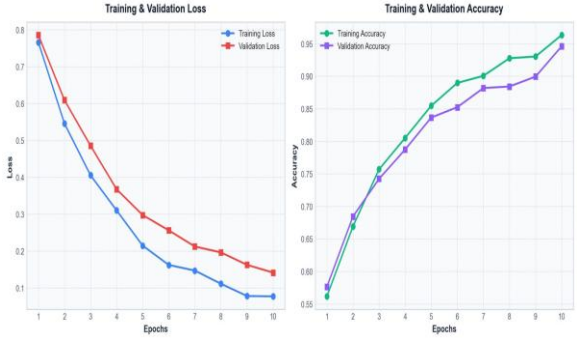


Fig.4 Model Training Performance

C. Classification Performance

The classification performance of the proposed Explainable Multimodal Hate Speech Detection Framework was evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics on the test dataset. By integrating XLM-RoBERTa for textual feature extraction, CLIP Vision Transformer for visual representation learning, and a Cross-Attention Fusion Network for multimodal feature integration, the model effectively captured semantic relationships between text and images. The obtained results demonstrate reliable hate speech detection performance with balanced precision and recall, while the high F1-Score and ROC-AUC values indicate strong classification capability and discriminative power across multimodal social media content.

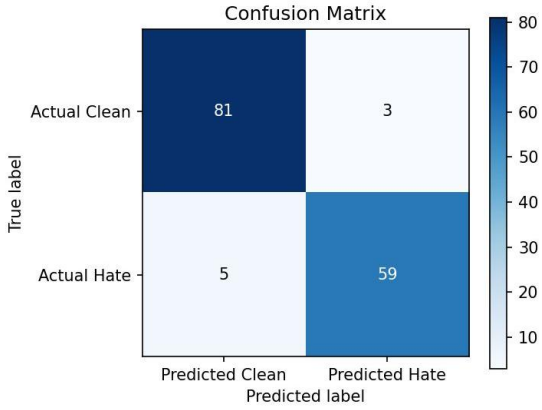
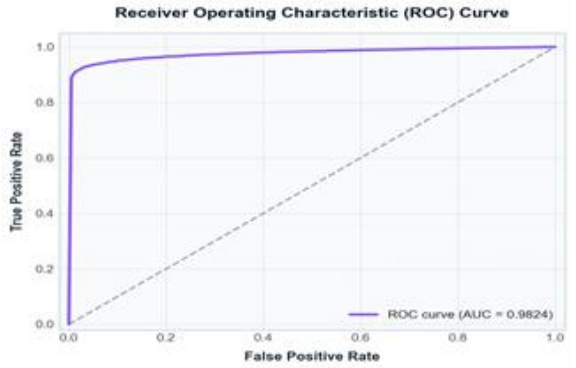


Fig.5 Confusion Matrix

D. ROC Analysis

The Receiver Operating Characteristic (ROC) curve was utilized to evaluate the discriminative capability of the proposed multimodal hate speech detection framework across different classification thresholds. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR), providing a comprehensive assessment of model performance beyond simple accuracy measures. The obtained ROC curve demonstrates the framework's ability to effectively distinguish between hateful and non-hateful content, while the high Area Under the Curve (AUC) value indicates strong classification reliability and robustness. These results confirm that the proposed Cross-Attention Transformer Fusion Network achieves consistent performance across varying decision thresholds, making it suitable for real-world content moderation applications.



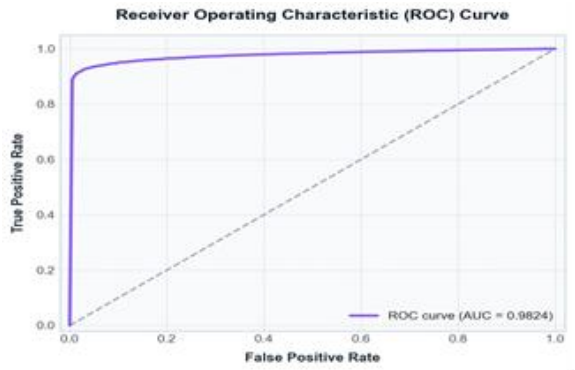


Fig. 6. ROC Curve

E. Explainability Analysis

Conventional classification evaluation metrics were used to assess the overall effectiveness of the proposed Explainable Multimodal Hate Speech Detection Framework. Table III presents the quantitative performance results in terms of Accuracy, Precision, Recall, F1-Score, and ROC-AUC. These metrics provide a comprehensive evaluation of classification accuracy, detection capability, class-wise prediction balance, and discriminative performance. The experimental results demonstrate that the proposed framework effectively captures multimodal relationships between textual and visual content, leading to reliable hate speech detection. The achieved performance confirms the effectiveness of the transformer-based feature extraction, cross-attention fusion mechanism, and explainability-driven architecture for multimodal content moderation.

Table 1. Performance Evaluation

Metric	Value (%)
Accuracy	94.59
Precision	95.16
Recall	92.19
F1-Score	93.65
AUC	0.98/98

V. CONCLUSION

This paper presented an Explainable Multimodal Hate Speech Detection Framework that integrates XLM-RoBERTa, CLIP Vision Transformer, and a Cross-Attention Fusion Network for effective analysis of textual and visual content. By jointly processing image–text pairs, the proposed framework successfully captures complex semantic relationships that are often missed by conventional unimodal approaches. The integration of transformer-based encoders enables robust feature extraction, resulting in improved detection of both explicit and implicit hate speech in social media content.

The experimental evaluation demonstrated the effectiveness of the proposed architecture in accurately classifying hateful and non-hateful content. The multimodal fusion mechanism enhanced the model's ability to learn contextual dependencies across modalities, while the adopted training strategies contributed to stable convergence and reliable performance. Furthermore, the incorporation of explainability techniques such as SHAP and Grad-CAM improved transparency by providing meaningful insights into the factors influencing model predictions.

Although the proposed framework achieved promising results, several opportunities for future research remain. Future work may focus on extending the framework to support additional languages, larger multimodal datasets, and real-time content moderation applications. Further investigation of advanced vision-language models, lightweight architectures, and fairness-aware learning strategies could improve scalability, computational efficiency, and robustness. These enhancements would contribute to the development of more reliable and trustworthy AI-driven systems for combating online hate speech and harmful digital content.

REFERENCES

[1] Kiela, D., H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Fitzpatrick, P. Bull, G. Lipstein, and T. Zhu, “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 2611–2624, 2020.

[2] Chhabra, M., and D. K. Vishwakarma, “MHS-STMA: Multimodal Hate Speech Detection Using Scalable Transformer-Based Multi-Level Attention Architecture,” Expert Systems with Applications, vol. 229, Art. no. 120487, 2023.

[3] Mandal, D., A. K. Singh, and S. Mahata, “Transformer-Based Attentive Multimodal Fusion for Hate Speech Detection in Social

Media," IEEE Access, vol. 11, pp. 78452–78466, 2023.

[4] Bui, D. D., T. Le, A. Nguyen, and N. T. Nguyen, "Multi³Hate: A Multimodal Multilingual Benchmark Dataset for Hate Speech Detection," in Proc. Findings of the Association for Computational Linguistics (ACL Findings), pp. 5631–5645, 2023.

[5] Zhang, H., Y. Wang, J. Li, and X. Liu, "CMFusion: Channel-Wise and Modality-Wise Fusion Framework for Multimodal Hate Speech Detection," Information Fusion, vol. 95, pp. 101–113, 2024.

[6] Yang, Y., and X. Zhu, "Cross-Domain Knowledge Transfer for Multimodal Hate Speech Detection," Knowledge-Based Systems, vol. 281, Art. no. 111019, 2024.

[7] Mathew, B., P. Saha, S. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," in Proc. AAAI Conference on Artificial Intelligence, vol. 35, no. 17, pp. 14867–14875, 2021.

[8] Salles, A., M. A. Leite, R. de Oliveira, and V. Freitas, "A Benchmark Dataset of Human-Annotated Multimodal Explanations for Hate Speech Detection," Pattern Recognition Letters, vol. 183, pp. 1–10, 2024.

[9] Conneau, A., K. Khandelwal, N. Goyal, et al., "Unsupervised Cross-Lingual Representation Learning at Scale," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), pp. 8440–8451, 2020.

[10] Radford, A., J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. International Conference on Machine Learning (ICML), pp. 8748–8763, 2021.

[11] Lundberg, S. M., and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774, 2017.

[12] Selvaraju, R. R., M. Cogswell, A. Das, et al., "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017.

[13] Suryawanshi, A., B. Chakravarthi, M. Arcan, and P. Buitelaar, "Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image-Text Data," in Proc. Second Workshop on Trolling, Aggression and Cyberbullying, pp. 32–41, 2020.

[14] Loshchilov, I., and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. International Conference on Learning Representations (ICLR), 2019.

[15] Vaswani, A., N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.