Application Analysis of Multimodal Models in Hateful Meme Detection

Mengqi Yan

Haide College, Ocean University of China, Qingdao, 266100, China

Abstract. Hateful memes are internet memes that spread virally by overlaying short text on images, often containing offensive content targeting groups based on gender, religion, race, or other characteristics. Their rapid dissemination and harmful impact make targeted detection critically important. Multimodal models, capable of simultaneously processing images and text, can accurately identify hateful content in memes. This paper analyzes the image-text fusion methods, optimization strategies, and evaluation metrics of multimodal models in hateful meme detection. Results show that incorporating cross-attention mechanisms during the image-text fusion stage effectively captures complementary information between modalities, thereby enhancing downstream task performance. Furthermore, optimization techniques such as multi-task learning and adversarial training can further improve model robustness and detection accuracy. Model distillation techniques enable faster detection with minimal accuracy loss, facilitating the timely identification of newly released hateful memes. In summary, this paper argues that multimodal models hold significant potential for mitigating the spread of online hate and provides theoretical and practical references for related research through an analysis of image-text fusion methods, optimization strategies, and evaluation metrics.

1 Introduction

Hateful memes, as a unique carrier of online hate speech, have shown a rapid growth trend in recent years. They exacerbate negative emotions such as anxiety and depression among targeted groups and undermine societal inclusivity. As of 2024, major global social platforms handle an average of 47,000 hateful meme reports daily. Unlike traditional single-modality hate content, hateful memes combine images and text metaphorically to construct a more complex semantic expression system. This characteristic renders traditional single-modality detection methods incapable of simultaneously capturing image and text information, often resulting in low detection accuracy.

To address this issue, researchers have focused on constructing and optimizing multimodal machine learning models to capture the complex semantic relationships between images and text, thereby improving detection performance. Current research directions include designing multimodal feature fusion methods to explore effective

yanmengqi@stu.ouc.edu.cn

integration of image and text features for enhanced classification performance, as well as model optimization strategies to improve practical utility. For instance, Kumar et al. proposed a multimodal architecture named Hate-CLIPper, based on the Contrastive Language-Image Pre-training (CLIP) model, which models cross-modal interactions between image and text features through a Feature Interaction Matrix (FIM). This approach achieved the Area Under the Receiver Operating Characteristic Curve (AUROC) of 85.8 on the Hateful Memes Challenge dataset, surpassing human performance (82.65) [1]. Chhabra and Vishwakarma combined a Multi-Scale Kernel Attention Visual Module (MSKAV) with a Knowledge-Distilled Attention Captioning Module (KDAC) to extract diverse visual features and emphasize key regions using attention mechanisms, demonstrating strong generalization through cross-dataset experiments [2].

This paper aims to analyze the technical advancements of multimodal models in hateful meme detection, highlighting challenges in image-text fusion, model optimization, and practical applications. It first reviews the technical evolution of multimodal fusion methods, analyzing the strengths and weaknesses of early fusion, late fusion, and cross-attention mechanisms. It then explores how optimization strategies such as multi-task learning, adversarial training, and model distillation enhance model performance. Finally, through evaluation metrics and qualitative literature analysis, it summarizes the limitations of current research and provides future outlooks.

2 Common Hateful Meme Datasets

The Hateful Memes dataset, developed by Facebook AI (FHM), contains 10,000 image-text combined memes. Each entry undergoes multiple rounds of human annotation and is further reviewed by experts. The dataset deliberately includes cases that appear hateful but are not, ensuring that models must consider both images and text to accurately identify hateful content.

The MMHS150K dataset, compiled by Raul Gomez's team, consists of 150,000 posts from the X platform, each including images and text. Annotations are provided by crowdsourced workers, with the final hate type for each entry determined by majority voting and categorized into six types of hateful content [3].

The MultiOFF dataset, curated by Bharathi Raja Chakravarthi's team, comprises 743 image-text combined memes. Annotations are completed by multiple annotators, with checks for consistency among their labels. All content is related to the 2016 U.S. presidential election and includes material that may be uncomfortable or offensive.

The MAMI dataset, collected by Elisabetta Fersini's team, includes 11,000 image-text combined memes. Annotations are performed by experts and online volunteers, who not only identify whether the memes contain misogyny but also specify the type of discrimination. The content primarily focuses on themes such as insulting women and perpetuating stereotypes [4].

3 Technical Evolution of Multimodal Fusion Methods

3.1 Mainstream Fusion Methods

Based on the stage of image-text fusion, current multimodal feature fusion methods can be broadly categorized into three types: early fusion, late fusion, and intermediate fusion.

Early fusion methods employ a feature concatenation strategy. For example, the framework proposed by Suryawanshi et al. directly combines visual features extracted by a

Convolutional Neural Network (CNN) architecture with 16 layers designed by the Visual Geometry Group (VGG-16) with text embeddings generated by Bidirectional Encoder Representations from Transformers (BERT) to form a joint feature vector. This approach is computationally efficient (with an average inference time of 83ms) but overlooks structural differences between modalities, resulting in an F1-score of only 0.71 on the MultiOFF dataset. Notably, early fusion requires high feature alignment, and its performance significantly declines when the semantic correlation between images and text is weak [5].

Late fusion methods adopt a divide-and-conquer strategy, exemplified by Grasso et al.'s Kontextuell Encoder Representations Made by Insertion Transformations (KERMIT) model. This model first processes images and text separately using Vision Transformer (ViT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), then integrates them at the decision layer through a fully connected network. This method achieved a weighted F1-score of 0.82 on the MAMI dataset but incurs a higher computational cost (47% increase in training time). It is particularly noteworthy that late fusion heavily relies on the quality of single-modality features, and overall performance is noticeably affected when feature extraction in one modality is suboptimal [6].

Intermediate fusion methods, which incorporate cross-attention mechanisms, represent the current technological frontier. Lin et al.'s "multimodal debate framework" dynamically adjusts the interaction strength of image-text features through learnable attention weights. This approach improved the Macro-F1 score by 12% on the Harm-C dataset but requires an additional 15% GPU memory for detection. Technical analysis indicates that the performance advantage of cross-attention primarily stems from its ability to capture fine-grained semantic associations, particularly excelling in handling culturally specific metaphorical content [7].

3.2 Other Fusion Methods

Recent studies have enhanced basic fusion methods by integrating techniques such as bilinear pooling, knowledge augmentation, and contrastive learning, improving the representational capacity of multimodal fusion. These advancements enable models to more precisely integrate image and text information, achieving superior performance in tasks like hateful meme detection.

The application of bilinear interaction techniques has significantly improved the granularity of feature fusion. The Hate-CLIPper model constructs a 256-dimensional interaction matrix to model point-to-point associations between image regions and text tokens. This approach excels in detecting hateful memes with complex compositions, achieving a state-of-the-art AUROC of 0.9098. Notably, the performance gains from bilinear pooling are positively correlated with data scale, with the most pronounced effects observed when training samples exceed 100,000 [1].

Knowledge augmentation methods enhance the detection of hateful memes targeting specific groups. Grasso et al. integrated the ConceptNet knowledge base into the KERMIT model to aid in understanding cultural symbols in images. Empirical studies show that this approach significantly improves the detection of religion-related hateful memes, increasing recall from 54% to 79%. It demonstrates unique advantages, particularly in handling content involving historical events or symbols specific to certain groups [6].

The introduction of contrastive learning techniques increases the sensitivity of fusion representations to subtle image-text differences. Mei et al. designed a retrieval-guided contrastive loss function that actively mines hard negative samples to enhance model discriminative ability. Experiments on the MultiOFF dataset show that this method

improves accuracy by 4.7 percentage points, with particularly notable improvements in recognizing low-sample categories (sample size <100) [8].

4 Model Optimization Strategies

Beyond the feature fusion stage, traditional multimodal detection methods often struggle to balance generalization, robustness, and deployment efficiency when facing challenges such as data scarcity, adversarial perturbations, and resource constraints. By leveraging multi-task learning to exploit semantic complementarity between tasks, adversarial training with data augmentation, and hierarchical knowledge distillation, more stable and efficient detection models can be developed.

Multi-task learning (MTL) was initially proposed to address the overfitting and limited generalization of single-task models under data scarcity or ambiguous task boundaries. Traditional hate detection models typically focus on a single objective, neglecting related emotional information, which leads to significant performance drops when training samples are reduced. To address this, Ma et al. designed a shared-private architecture for multi-task learning. This framework uses a shared visual encoder at the lower level to extract common features and constructs parallel task-specific subnetworks at the top level for the primary task of hate detection and the auxiliary task of sentiment analysis. A gradient conflict suppression algorithm is introduced to prevent negative transfer between tasks. Experiments demonstrate that, even with a 30% reduction in training data, this framework improves the F1-score for hate detection by 8%. The performance gain is particularly pronounced when the auxiliary task is semantically closely related to the primary task, such as sentiment analysis and hate detection [9].

Multimodal detection systems generally exhibit poor robustness against adversarial attacks, with the image modality being particularly sensitive to perturbations. Traditional training methods, optimized only on natural samples, struggle to resist artificially designed adversarial interference, while simple data augmentation techniques like geometric transformations or text synonym replacement offer limited benefits. To address this, Wu et al. incorporated adversarial samples generated by the Fast Gradient Sign Method (FGSM) into the training process. Through adversarial training, the model's Multi-Modal Attention Entropy (MMAE) was significantly reduced from 0.32 to 0.24. They also found that the image modality degrades 40% more than the text modality under the same perturbation strength, highlighting the need to prioritize enhancing the robustness of the visual branch. Additionally, on the small-sample TamilMemes dataset, a combined data augmentation strategy involving geometric transformations and text synonym replacement improved accuracy by $12-15\%$. However, excessive image deformations can disrupt original semantics, necessitating careful adjustment of augmentation intensity based on data characteristics [10].

Although large-scale multimodal models like Vision-and-Language BERT (ViLBERT) excel in accuracy, their large parameter counts and high computational costs limit their deployment on edge devices and real-time inference capabilities. To overcome this bottleneck, Chhabra et al. proposed a hierarchical knowledge distillation framework. This approach extracts soft labels and intermediate features from the teacher model’s visual and text branches in a layered manner to guide the student model in learning critical information. The method compressed ViLBERT from 110 million parameters to 33 million, with only a 2.3% loss in accuracy, while achieving a 3.7-fold increase in inference speed on edge devices. This provides a practical solution for real-time multimodal detection applications [2].

5 Evaluation Metrics

In the field of hateful meme detection, researchers employ various evaluation metrics to comprehensively assess model performance and practicality. These metrics can be divided into three categories: classification performance metrics, modality attention metrics, and computational efficiency metrics.

Classification performance metrics are core tools for evaluating model accuracy, particularly for the binary classification task of hateful memes. AUROC measures a model's ability to distinguish between hateful and non-hateful memes, with values ranging from 0 to 1, where values closer to 1 indicate superior performance. It is calculated by plotting the true positive rate against the false positive rate and computing the area under the curve, reflecting the model's overall performance across different thresholds. The F1-score, the harmonic mean of precision and recall, is particularly suitable for handling imbalanced datasets, where hateful memes may constitute a small proportion of the data.

Modality attention metrics focus on evaluating how models handle multimodal data, such as the joint analysis of images and text, which is critical in hateful meme detection. MMAE, proposed by Pramanick et al., quantifies the extent to which a model neglects key modalities. With an average value of 0.29, it suggests that models may exhibit some degree of modality dependency in multimodal tasks. MMAE measures whether a model balances image and text information by calculating the entropy of the attention distribution. High values may indicate over-reliance on a single modality, such as focusing on text while ignoring image content [11].

Computational efficiency metrics assess a model's feasibility for real-world deployment, directly impacting its utility in real-time applications. Training time refers to the duration required for a model to train from its initial state to convergence, with studies reporting an average of 72 hours. This reflects the substantial computational resources, such as GPU clusters, needed for hateful meme detection models, which may consume significant energy and time during training. Inference latency, typically measured in milliseconds, is the time required for a model to classify a single meme. Studies comparing different architectures show that early fusion, which integrates multimodal data at the input stage, offers higher computational efficiency and is suitable for real-time applications. In contrast, cross-attention mechanisms, which capture modality interactions dynamically, may achieve higher accuracy but incur greater computational costs. These differences are critical in practical deployment scenarios, such as on social media platforms that require rapid processing of large data volumes, where excessive inference latency can degrade user experience. Therefore, balancing model accuracy and computational efficiency is key to developing scalable systems.

6 Qualitative Literature Analysis

When detecting hateful memes using multimodal approaches, selecting appropriate feature fusion methods is crucial, yet there is no standardized criterion for their selection. To understand prominent feature fusion methods, this section conducts a qualitative analysis of recent literature. Table 1 summarizes six representative studies, covering image-text fusion methods, model optimization strategies, experimental datasets, performance outcomes, and primary limitations. Analysis of these studies reveals that current research widely adopts techniques such as cross-attention mechanisms, multi-task learning, and knowledge augmentation during the image-text fusion stage to enhance models' understanding of complex semantics. However, these techniques also expose common issues, including reliance on single datasets, high computational costs, over-dependence on external

knowledge or auxiliary task relevance, and insufficient support for multilingualism and real-time performance. Table 1 below provides detailed information on these studies, aiming to serve as a reference for related researchers.

Table 1 Studies on Complex Multimodal Feature Fusion Methods

Literature authors	Method	Dataset	Performance	Limitations
Kumar & Nandakumar [1]	Hate-CLIPper: Intermediate Fusion Based on CLIP Feature Interaction Matrix	Hateful Memes Challenge	AUROC 85.8%, Surpassing Human Detection Baseline	Validated Only on a Single Dataset; High Computational Load for Feature Matrix
Ma et al. [9]	Multi-Task Learning (Shared-Private Architecture + Gradient Conflict Suppression)	Hateful Memes, etc.	F1-Score for Main Task Improved by 8% with 30% Less Training Data	Strong Dependence on Sentiment Auxiliary Task Relevance
Chhabra & Vishwakarma [2]	MSKAV + KDAC: Multi-Scale Visual Kernels with Knowledge-Distilled Attention	MAMI and Cross-Dataset Experiments	Strong Generalization, Stable Performance Across Datasets, Model Compressed from 110M to 33M Parameters, Inference Speed Increased by 3.7×	High Computational Cost for Visual Branch
Grasso et al. [6]	KERMIT: Late Fusion + ConceptNet Knowledge Augmentation	MAMI	Religion-Related Recall Increased from 54% to 79%; Weighted F1 Reached 0.82	Difficulty in Updating External Knowledge Base; High Dependence on Symbolic Knowledge
Huang et al. [12]	Evolver: large language models (LLM) Dynamic Fusion under Chain-of-Evolution	FHM (Zero-Shot)	Zero-Shot Accuracy of 61.7%	Lower Accuracy than Supervised Models; Multilingual Support and Real-Time Performance

	Prompting			Need Improvement
Wu et al. [10]	Fuser: Enhanced Multimodal Fusion Framework	Hateful Memes, Harm-C, Harm-P	AUROC 87.16%, MMAE 0.1394 (on Harm-C)	Lack of Auxiliary Context for Some Real-World Knowledge

7 Research Limitations and Future Outlook

Despite recent breakthroughs in multimodal hate detection in terms of algorithmic performance and application scenarios, three main limitations persist. First, insufficient data diversity severely restricts model generalization. Approximately 87% of existing studies rely solely on English datasets, with near-zero coverage of low-resource languages like Bengali. Karim et al. found that model performance drops by up to 35% when transferring from English to Bengali. Additionally, the cultural representativeness of datasets is limited, making it difficult to capture the diverse forms of hate expression globally [13]. Second, high computational costs hinder industrial adoption. For instance, cross-modal cross-attention models incur an average training cost of $2,300 on a designation for an Amazon Web Services (AWS) Elastic Compute Cloud (EC2) instance type (AWS p3.2xlarge), 3.2 times higher than early fusion methods, significantly raising the barrier to application and limiting large-scale experiments and commercial deployment [14]. Finally, model biases are unpredictable and urgently require management. Hamza et al. noted that current systems have a 28% misclassification rate for Islam-related content, far higher than other categories, yet existing evaluations often lack systematic assessments of fairness and bias [14].

To address these shortcomings, future research can focus on two directions. First, leveraging LLMs to enhance multimodal fusion capabilities. Huang et al.'s Evolver framework, using Chain-of-Evolution Prompting to simulate the dynamic evolution of meme semantics, achieved a 61.7% accuracy on the FHM dataset in zero-shot scenarios, demonstrating LLMs' significant potential in capturing emerging hate expressions [12]. Overall, LLMs can capture richer linguistic and visual semantic information through large-scale pre-training, maintaining strong adaptability in low-resource languages and few-shot learning scenarios, thus offering greater possibilities for multilingual hate detection. Second, establishing a standardized, comprehensive evaluation system to ensure model robustness and fairness. This system should include multilingual test sets, adversarial robustness benchmarks such as FGSM, Projected Gradient Descent (PGD) attacks, and systematic bias assessments for different types of hateful memes, providing robust support for comprehensive multimodal hate detection research and practical applications.

8 Conclusions

This paper reviews the main advancements of multimodal models in hateful meme detection, with a focus on analyzing image-text fusion methods, model optimization strategies, and evaluation metrics. Through a synthesis of current research, it is found that cross-attention mechanisms excel in handling complex semantic interactions, while multi-task learning and adversarial training enhance model stability in data-scarce or perturbed environments. Knowledge distillation techniques achieve model compression while maintaining performance, improving deployment efficiency. However, current research still

faces challenges such as uneven language coverage in datasets, high computational resource demands, and model biases. Future work can focus on expanding multilingual datasets, leveraging large language models to aid image-text understanding, and developing more systematic evaluation frameworks to improve model generalization and fairness. These efforts hold practical significance for building more reliable mechanisms to identify online hate content.

References

1. Kumar, G. K., & Nandakumar, K. (2022). Hate-CLIPper: Multimodal Hateful Meme Classification Based on Cross-Modal Interaction of CLIP Features. NLP4PI.
2. Chhabra, A., & Vishwakarma, D. (2023). Multimodal Hate Speech Detection via Multi-Scale Visual Kernels and Knowledge Distillation Architecture. Engineering Applications of Artificial Intelligence.
3. Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1470-1478).
4. Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., ... & Sorensen, J. (2022, July). SemEval-2022 task 5: Multimedia automatic misogyny identification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 533-549).
5. Suryawanshi, S., et al. (2020). Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. Workshop on Trolling, Aggression and Cyberbullying.
6. Grasso, B., et al. (2024). KERMIT: Knowledge-EmpoweRed Model In Harmful Meme deTection. Information Fusion.
7. Lin, H., et al. (2024). Towards Explainable Harmful Meme Detection Through Multimodal Debate Between Large Language Models. The Web Conference.
8. Mei, J., et al. (2023). Improving Hateful Meme Detection Through Retrieval-Guided Contrastive Learning. Annual Meeting of the Association for Computational Linguistics.
9. Ma, Y., et al. (2022). Hateful Memes Detection Based on Multi-Task Learning. Mathematics.
10. Wu, F., et al. (2024b). An Enhanced Multimodal Fusion Framework with Congruent Reinforced Perceptron for Hateful Memes Detection. Information Processing & Management.
11. Pramanick, S., et al. (2021). MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. Conference on Empirical Methods in Natural Language Processing.
12. Huang, J., et al. (2024). Evolver: Chain-of-Evolution Prompting to Boost Large Multimodal Models for Hateful Meme Detection. arXiv.
13. Karim, M. R., et al. (2022). Multimodal Hate Speech Detection from Bengali Memes and Texts. arXiv.
14. Hamza, A., et al. (2023). Multimodal Religiously Hateful Social Media Memes Classification Based on Textual and Image Data. ACM Trans. Asian Low Resour. Lang. Inf. Process.