Detecting and Understanding Hateful Contents in Memes Through Captioning and Visual Question-Answering

Ali Anaissi[1,3], Junaid Akram[1,3,4], Kunal Chaturvedi[2], and Ali Braytee[2]

[1] The University of Sydney, School of Computer Science, Camperdown, NSW 2008, Australia junaid.akram@sydney.edu.au, ali.anaissi@sydney.edu.au
[2] University of Technology Sydney, School of Computer Science, Ultimo, Australia Kunal.Chaturvedi@uts.edu.au, ali.braytee@uts.edu.au
[3] University of Technology Sydney, TD School, Ultimo, Australia ali.anaissi@uts.edu.au, junaid.akram@uts.edu.au
[4] Australian Catholic University, Peter Faber Business School, North Sydney, NSW 2060 Australia junaid.akram@acu.edu.au

Abstract. Memes are widely used for humor and cultural commentary, but they are increasingly exploited to spread hateful content. Due to their multimodal nature, hateful memes often evade traditional text-only or image-only detection systems, particularly when they employ subtle or coded references. To address these challenges, we propose a multimodal hate detection framework that integrates key components: OCR to extract embedded text, captioning to describe visual content neutrally, sub-label classification for granular categorization of hateful content, RAG for contextually relevant retrieval, and VQA for iterative analysis of symbolic and contextual cues. This enables the framework to uncover latent signals that simpler pipelines fail to detect. Experimental results on the Facebook Hateful Memes dataset reveal that the proposed framework exceeds the performance of unimodal and conventional multimodal models in both accuracy and AUC-ROC.

Keywords: Hateful Memes · Multimodal Detection · Optical Character Recognition · Classification.

1 Introduction

Memes have emerged as a widely used medium on social media platforms, combining images and text overlays to convey humor, satire, or cultural commentary. Despite their seemingly innocuous appearance, memes are increasingly exploited to propagate hateful or discriminatory content [3, 27]. Due to their multimodal nature, such content often bypasses conventional text-only or image-only detection algorithms. When image elements and textual components interact in subtle ways, hateful content may remain hidden, allowing offending material to circulate unchecked [6,32]. Empirical evidence from the Hateful Memes Challenge has

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

shown that unimodal approaches typically fail to adequately capture the range of possible hateful expressions embedded within memes [17, 26]. Consequently, there is a critical demand for robust, integrated solutions that can parse both textual and visual cues to identify underlying animosity or prejudice.

Recent studies [11,13,23] have attempted to bridge the gap between language and vision representations, revealing that combined multimodal strategies can achieve promising results for specific domains such as misogynistic memes [34] or harmful COVID-19 memes [29]. While these approaches have shown promise, they exhibit several key limitations that hinder their ability to comprehensively detect nuanced hateful content. First, many existing methods [1, 4, 35] rely on fixed multimodal representations, where text and image features are extracted independently and fused statically. This rigid approach fails to capture the dynamic interplay between textual and visual cues, making it difficult to detect contextually embedded hate signals, such as sarcasm, coded symbols, or ambiguous imagery [18]. Second, these methods typically lack real-time adaptive reasoning, instead relying on predefined classification heuristics [19, 33]. As a result, they struggle with detecting veiled or evolving hate speech that requires contextual reasoning beyond surface-level analysis. Third, existing models often categorize hateful content using coarse-grained labels, such as simply hateful or non-hateful, without distinguishing between different forms of hate speech. This lack of specificity reduces interpretability and makes it harder to apply targeted moderation strategies. Such limitations highlight the need for a more systematic approach that incorporates iterative questioning, refined retrieval, and text-image fusion at a granular level.

To address these challenges, this paper introduces a framework that integrates optical character recognition (OCR), caption generation, retrieval- augmented classification, and a visual question answering (VQA) module. OCR reliably extracts overlaid text, while captioning supplies a neutral description of the visual scene. We enhance classification by leveraging a sub-labeling strategy, segmenting hateful content according to attributes such as race, religion, or others. This fine-grained division increases precision in retrieval-augmented steps, ensuring that exemplars align more closely with the observed meme. Additionally, the VQA system formulates targeted queries about potentially harmful symbols, background contexts, or linguistic cues that might escape notice in single-round analyses. By integrating these components, we aim to offer a system robust enough to detect concealed instances of hate speech.

The paper makes the following contributions:

- A multimodal approach that integrates OCR for textual extraction, neutral captioning for visual context, a sub-label retrieval, and a multi-turn VQA, to detect both explicit and implicit hateful cues in memes is proposed.
- A sub-label classification framework that partitions hateful content into race-based, gender-based, and other sub-dimensions is introduced to improve the accuracy of retrieval-augmented generation (RAG).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

2 Related Works

Over the past few years, research on hateful meme detection has evolved considerably, emphasizing the need for integrated analysis of both textual and visual modalities [15, 30]. Early attempts often separated images from text, applying standard classifiers to each modality in isolation. However, the limitation of such methods became apparent when memes contained subtle or implicit hateful references that only emerged through interaction between visual features and overlaid text. Consequently, various studies started to explore multimodal fusion. Kiela et al. [16] introduced the Hateful Memes Challenge, releasing a dataset that paired each image with short textual content to highlight the complexities of meme-based hate. Badjatiya et al. [5] and Davidson et al. [7] initially concentrated on textual classification, adopting lexicon-based approaches or neural architectures like CNNs and LSTMs, but these did not fully capture the compound nature of memes. Meanwhile, image-based methods such as Gómez et al. [10] and Howard et al. [14] attempted to detect hateful symbols or cues through CNNs and other vision models, yet struggled when the hatred was expressed solely via text.

Subsequent efforts introduced hybrid or multimodal models to process images and text jointly. Transformative architectures such as ViLBERT [24] and Visual BERT [21] harness cross-attention mechanisms to align textual and visual embeddings, thereby improving classification accuracy. In parallel, the Facebook Hateful Memes dataset [16] further prompted researchers to refine their multimodal pipelines, as it contained nuanced and challenging examples of encoded hate speech. Rizzi et al. [34] addressed misogynistic memes by proposing a fine-tuned VisualBERT that excelled at combining textual embeddings from OCR with high-level image features obtained from pretrained CNNs. Pramanick et al. [29] tackled COVID-19-related misinformation with a focus on harmful memes, showing that domain-specific training data could refine detection for medical or pandemic-oriented hate. Although these approaches outperformed unimodal baselines, they occasionally failed on memes whose meaning shifted dramatically depending on cultural or contextual details not captured by purely data-driven models.

Recent works have sought to incorporate advanced language models, retrieval techniques, and Visual Question Answering (VQA) to overcome the remaining challenges. Devlin et al. [8] illustrated the utility of contextual embeddings via BERT for language understanding, while Lewis et al. [20] introduced Retrieval-Augmented Generation (RAG) to infuse external knowledge into classification or generation tasks. Accordingly, sub-label methods emerged to partition hateful content into categories such as race or gender, facilitating more precise retrieval and classification [34]. Additionally, VQA-based systems proved beneficial for generating iterative queries about scene elements, as multi-turn dialogue can reveal latent meaning. By incorporating refined embedding models like CLIP [31] and advanced prompt engineering, researchers succeeded in capturing the interplay between textual overlays and visual symbolism at deeper levels [2, 11]. Collectively, these investigations underscore the vital role of multimodality and contextual verification in tackling hateful memes, thereby guiding the develop-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

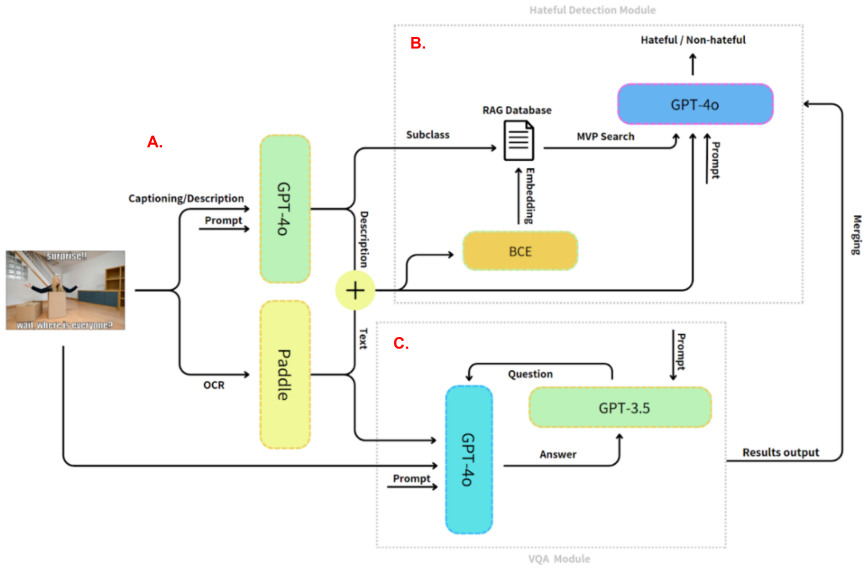

Fig. 1: The overall framework, RAG (sub_label + VQA) for detecting hateful content

ment of more robust pipelines that can identify concealed or culturally coded hatred.

3 Methods

In this section, we describe our proposed multimodal pipeline integrating OCR and captioning, VQA module, and a hateful detection module. The overall framework is shown in Fig. 1.

3.1 Captioning and OCR

A key challenge in detecting hateful content in memes arises from the interplay of textual and visual cues. As shown in Fig. 1A, we first extract all text and generate captions before passing the multimodal information to our detection modules. Specifically, we separate the input processing into two complementary procedures. First, Paddle OCR [9], an optical character recognition system, is used to extract textual messages in memes. To supplement OCR, we generate a caption describing the visual content of the meme using a large language model. The net effect is a more robust representation of the meme, combining the recognized text with a broad contextual description.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

3.2 VQA Module

The Visual Question Answering module integrates OCR, multimodal analysis, and advanced language models to detect and analyze hateful content in memes. As shown in Fig. 1C, the workflow begins with raw meme input, which includes both visual and textual data. Using the Paddle OCR module, textual information embedded within the image is extracted, enabling the identification of captions, phrases, or symbols that may carry hateful messages. This extracted text, along with the raw image, forms the foundation for multimodal analysis in the VQA pipeline. The second stage involves processing the inputs using GPT-based models for dynamic question generation and context-aware answering. Initially, GPT-4.0 generates a broad, context-sensitive question aimed at understanding the overall theme of the meme, with a focus on detecting hate signals such as stereotypes, offensive language, or harmful visual elements. The generated question is then processed by GPT-3.5, which provides detailed answers by integrating visual and textual cues. If hate-related elements are identified, such as racial slurs or stereotypical imagery, the system refines its analysis by generating follow-up questions through GPT-4.0, specifically targeting the hateful components.

To ensure coherence and accuracy, the system employs a multi-turn dialogue mechanism, where all questions and answers are stored in a contextual database. This enables the system to maintain continuity across interactions, eliminating redundancy and ensuring that every aspect of the meme is thoroughly examined. Once the analysis is complete, the structured Q&A pairs are fed into a specialized Hate Detection Module that works in conjunction with the Retrieval-Augmented Generation (RAG) pipeline. The RAG pipeline further contextualizes and validates the findings by cross-referencing against a knowledge base of hate-related symbols, phrases, and behaviors. The final output of the system includes a detailed summary of the detected hateful content, highlighting specific textual elements, visual cues, and their contextual implications. By integrating OCR, multimodal analysis, and dynamic reasoning, the VQA system provides a robust solution for detecting hateful content in memes. Its iterative question-answering logic, combined with adaptive refinement, ensures thorough exploration of nuanced hate signals, making it a powerful tool for content moderation in real-world scenarios.

3.3 Hateful Detection

In the hateful detection module, we aim to classify whether a meme is hateful or non-hateful by leveraging the OCR text, the generated caption, and outputs from the VQA module. Fig. 1B illustrates the schematic of our hateful content detection module, which consists of multiple steps, including RAG, sub-labeling, and content explanation..

RAG Overview: The Retrieval-Augmented Generation (RAG) architecture [20] is utilized as a core component for hate detection by incorporating a vector database, embedding model, and ranking mechanism. The architecture

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

processes each meme's textual input (caption and OCR) by embedding it and querying a repository of labeled data, explanations, and examples. These retrieved chunks provide additional context in a prompt-like fashion to guide the model in producing more accurate and context-aligned inferences about hateful content.

Content Explanation for RAG: Another variant of the RAG architecture augmented the vector database with not only the caption and OCR text but also detailed explanations for why specific memes were labeled hateful or non-hateful. The assumption was that these explanations could help the model identify nuanced hate signals in new memes.

Sub-labeling for RAG: A specific implementation of RAG, known as "sub-label classification", was applied. Instead of treating hatefulness as a single, broad category, the sub-labeling method divides hateful content into finer-grained categories, such as race, religion, and others. By embedding the meme's caption and OCR text, the RAG system retrieves content related to the most relevant sub-label, providing contextual anchors that improve classification. For instance, a meme involving race-based hate speech retrieves examples and contextual references from the race sub-label category, making the detection more precise.

Our final framework, RAG (sub_label + VQA) integrated the outputs of the VQA module with the sub-labeling RAG method. The VQA results, which provide detailed contextual information by combining caption and OCR analysis with visual cues, were incorporated into the RAG pipeline as additional reference material to detect hateful memes.

4 Experiments and Results

4.1 Datasets

We utilized Facebook Hateful Memes (FHM) [26] dataset for our experiments. The dataset contains diverse meme examples with hateful vs. non-hateful labels. The data integrates text captions overlaid on images and is one of the primary resources provided by Facebook for the Hateful Memes Challenge. Next, we apply random transformations such as rotation, scaling, cropping, and mild color jitter. These transformations enrich the model's exposure to diverse visual conditions, thus boosting resilience to typical noise or distortion in user-generated memes. In certain data splits, we note that hateful content is encoded through metaphors, coded language, or domain-specific references. We thus expand the dataset with carefully curated examples reflecting these nuances to reinforce the sub-labeling strategy in the RAG pipeline. This expansion aids in capturing cultural, linguistic, or other contextual factors that might not be evident from standard data subsets. Overall, these strategies enhance the system's capacity to tackle newly emerging hateful memes with novel textual or visual patterns.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

4.2 Evaluation Metrics

For hateful content detection, we employ two primary metrics: Accuracy and AUC-ROC. For VQA, we adopt the VQAScore methodology [22]. We conducted five rounds of scoring to mitigate model variability. In each round, the VQA system generated answers to a collection of queries derived from the meme images. We then calculated VQAScore for each generated answer-image pair, and took the average over these five rounds. This approach ensures a more robust estimation of performance, reducing the influence of any single outlier run.

Table 1: Performance comparison across various models, including unimodal and multimodal methods. Acc. denotes Accuracy; AUROC stands for Area Under the Receiver Operating Characteristic.

Type	**Model**	**Acc. (%)**	**AUROC (%)**
	Human annotators	84.70	82.65
Unimodal	Image-grid [12]	52.00	52.63
	Image-region [16]	52.13	55.92
	Text BERT [8]	59.20	65.08
Multimodal	Late fusion [16]	59.66	64.75
	Concat BERT [16]	59.13	65.79
	MMBT-grid [16]	60.06	67.92
	MMBT-region [16]	60.23	70.73
	ViLBERT [25]	62.30	70.45
	Visual BERT [21]	63.20	71.33
	ViLBERT CC [16]	61.10	70.03
	Visual BERT COCO [16]	64.73	71.41
	GPT-4o mini [28]	69.50	75.02
Our Method	RAG (explanation)	59.20	63.01
	RAG (sub_label)	72.00	76.52
	RAG (sub_label + VQA)	**73.50**	**78.35**

4.3 Results

Quantitative Analysis Table 1 summarizes the performance of various models and methods in detecting hateful memes. The table also includes comparisons with human annotations as an upper bound, as well as benchmark approaches from the challenge. Human annotations remain the most accurate, with 84.70% accuracy and 82.65% AUROC. Among unimodal methods, vision-only models such as Image-grid and Image-region perform poorly around 52% accuracy, highlighting the insufficiency of visual cues alone for detecting nuanced hateful content. Text BERT outperforms these with 59.20% accuracy and 65.08% AUROC, underscoring the greater informativeness of textual features in

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

this domain. Multimodal baselines, which integrate image and text modalities, demonstrate marked improvements. Simple fusion techniques like Late Fusion and Concat BERT offer modest gains (59–60% accuracy). More sophisticated architectures such as MMBT-Region, ViLBERT, and Visual BERT COCO further improve performance, reaching up to 64.73% accuracy and 71.41% AUROC. The strongest baseline, GPT-4o mini, achieves 69.50% accuracy and 75.02% AUROC, setting a high bar for general-purpose large multimodal models.

Our proposed method significantly outperforms all baselines. While the RAG (explanation) variant performs comparably to Text BERT with 59.20% accuracy, incorporating fine-grained sub-labels in RAG leads to a substantial boost with accuracy of 72.00% and AUROC of 76.52%. The highest gains are observed when this sub-label retrieval is further combined with VQA, yielding the best overall results of 73.50% accuracy and 78.35% AUROC. These findings confirm the synergy between sub_label-based retrieval and the contextual enhancements provided by the VQA module. Rather than only relying on raw text or naive retrieval from explanation templates, the sub_label approach retrieves precisely relevant hateful exemplars, while the VQA module helps uncover implicit cues that might not be evident through OCR captioning alone.

Qualitative Observations Fig. 2 offers an illustrative example, showing system outputs for both a positively identified hateful meme and a non-hateful instance. The figure includes how OCR extracts textual content, how the captioning module describes the image, and how the multi-round VQA interacts with the meme to highlight potentially hateful elements.

In the hateful example, OCR precisely captured key terms from the overlaid text, and the captioning module accurately noted contextual objects and background. The VQA dialog then focused on potentially discriminatory language, confirming hateful cues and retrieving relevant sub_label data through the RAG sub_label pipeline. The final classification was correct and accompanied by a short textual explanation consistent with the known ground truth.

In the non-hateful case, the system again accurately recognized textual and visual details but found no hateful signals. The RAG retrieval was less relevant, returning only examples bearing minimal resemblance to hateful content. As a result, the classification was non-hateful, aligning with the ground-truth label. This outcome underlines the pipeline's capacity to remain conservative when the textual and visual signals do not suggest hateful references.

4.4 Discussion

These results indicate that combining large language models with sub_label-based retrieval and VQA modules yields significant gains over unimodal or simpler multimodal baselines. Several observations are worth highlighting:

Performance gaps and remaining challenges. Although RAG (sub_label) + VQA achieves 73.50% accuracy, it still lags behind human annotators. This gap underscores the complexity and ambiguity of hateful memes, which

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

Image		
Text	SCREW FEMINISM MYPLACEISINTHE KITCHEN!	MARRIAGE IS... ALWAYS HAVING SOMEONE TO PITY-LAUGH AT YOUR JOKES RD
Description	This meme features a retro-styled image of a woman dressed in 1950s housewife attire, with a cheerful expression. It includes the text: "SCREW FEMINISM, MY PLACE IS IN THE KITCHEN!" The overall theme suggests a traditional view of gender roles, particularly emphasizing the idea that a woman's place is at home, specifically in the kitchen, which is juxtaposed against the concept of feminism.	The meme features a cheerful couple sitting together, each holding a cup. The text overlays the image and reads, "Marriage is... ALWAYS HAVING SOMEONE TO PITY-LAUGH AT YOUR JOKES." This combines a lighthearted view of marriage with a humorous take on the dynamics of relationship humor, suggesting that partners often support each other despite the quality of the jokes made. The tone is playful and aims to provoke a smile or a laugh.
VQA Dialogue	**Question:** Can you describe the image in detail, focusing on the people, their actions, emotions, notable objects, and the overall setting? **Answer:** A smiling woman in a retro kitchen wears a red dress with a white apron, expressing defiance toward feminism. **Question:** What specific elements in the retro kitchen contribute to the overall nostalgic feel of the setting? **Answer:** The retro kitchen features pastel colors, floral curtains, and vintage utensils, enhancing the nostalgic atmosphere.	**Question:** Can you describe the image in detail, focusing on the people, their actions, emotions, notable objects, and the overall setting? **Answer:** A smiling couple in casual attire share coffee and laughter on a porch, highlighting warmth and companionship. **Question:** What do you think might be the significance of the mugs they are holding in the image? **Answer:** The couple's body language is relaxed and affectionate, with genuine smiles and eye contact, as they enjoy coffee together.
Similar cases (The first one of top5)	that's it, back to the kitchen! where's my sammich? women's day is over! The meme features a woman on the ground in a casual pose, wearing shorts and a sweater, with a man standing behind her. The text overlay includes phrases that suggest traditional gender roles and a dismissal of a specific occasion, Women's Day. The tone appears sarcastic, using humor related to stereotypes about women's roles. The combination of the imagery and text creates a juxtaposition of casualness and a controversial statement regarding gender expectations. (gender)	when you see a black guy in your store but he isn't trying to steal anything impossible This meme features an image of a smiling woman wearing a purple headscarf and floral clothing. The text overlay reads: "i married my brother to commit immigration fraud." The meme appears to imply a controversial claim about the subject, suggesting inappropriate or illegal behavior related to marriage and immigration. The tone is informal and provocative, characteristic of many internet memes that aim to elicit strong reactions. (race)
Ground Truth	Hateful	Non-hateful
Prediction	Hateful	Non-hateful
Explanation	The statement is hateful because it explicitly denigrates feminism and promotes a traditional and restrictive view of women's roles. it perpetuates the stereotype that a woman's place is solely in the kitchen	The statement describes a meme that presents a humorous and affectionate take on marriage. the description emphasizes companionship

Fig. 2: Outputs for a hateful example (left) and a non-hateful example (right). The pipeline includes accurate OCR detection, objective captioning, multi-turn VQA addressing targeted hate cues, and final classification via RAG.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

often rely on cultural references, double meanings, or evolving slang not always captured by static training data. Further refinement of sub_label categories, addition of external knowledge sources, and extended data augmentation strategies may narrow this human-machine divide.

Effectiveness of OCR With the rise of multimodal large language models, one may question the need for a dedicated OCR module. However, we find that incorporating explicit OCR (PaddleOCR) remains valuable, especially when dealing with stylized, distorted, or meme-specific fonts that challenge even state-of-the-art vision-language models. Explicit OCR ensures consistent and controllable extraction of embedded text, which downstream modules such as VQA and RAG rely on for accurate reasoning. Furthermore, separating text extraction from high-level reasoning supports interpretability, and modular debugging.

Utility of VQA dialogue. Notably, RAG showed visible improvements after including VQA-derived context. The VQA system probes the meme with targeted questions, clarifying ambiguous cues and capturing nuanced correlations between text and visuals. This synergy is crucial in uncovering content that is hateful only when certain textual or symbolic aspects align with specific contexts or objects in the image. The average VQAScore of 75.04 also suggests that the system reliably produces answers consistent with the underlying image content, thereby strengthening the subsequent classification.

Explanation-based RAG limitations. The RAG (explanation) configuration performed poorly for reasons related to noise in the textual explanations and potential misalignment with new memes. The assumption that labeled explanations from certain memes would be directly transferable or consistently interpreted appears flawed. In contrast, sub_label retrieval provides a more targeted anchor (e.g., detecting "racial hate" or "religious hate" specifically), improving retrieval precision.

Implications for real-world applications. Content moderation platforms or social media sites that must identify hateful memes in real time can benefit from adopting a pipeline that integrates a carefully designed retrieval mechanism, a multi-turn VQA system, and robust text-image analysis. However, real-time constraints require optimization to reduce computational overhead; our solution underscores the effectiveness of multi-step synergy but also reveals potential latency in large-scale deployments. In particular, the use of large language models for multi-turn VQA and the dependency on retrieval-augmented generation (RAG) with sub-label classification introduce substantial memory and processing demands. These may hinder responsiveness in high-throughput or latency-sensitive settings. Future engineering efforts would thus focus on accelerating sub_label lookups and streamlining the VQA query-response phase.

The quantitative results, shown in Table 1, and the qualitative analysis demonstrate that our integrated system effectively detects hateful memes and outperforms several multimodal baselines. Although there remains a gap relative to human-level comprehension of subtle, context-dependent hate, the results confirm that careful synergy among OCR, captioning, VQA, and specialized retrieval strategies can significantly improve classification performance.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

5 Conclusion

The proposed framework demonstrates a robust approach to addressing hateful content detection within memes by integrating multiple modules for text extraction, captioning, retrieval, and visual question answering. The integrated pipeline achieves significant accuracy and AUC-ROC compared to existing methods on established benchmarks. These results underline the importance of uniting refined language strategies with methods that analyze images more deeply. Future work could extend the framework by incorporating culturally nuanced knowledge graphs, refining VQA prompts to reduce false positives, or integrating dynamic feedback loops for real-time detection.

Acknowledgment

We acknowledge the contributions of Bonnie Zhong, Haodi Yang, Yinuo Wang, Lu Tang, Yiqi Zhao, Yuanhao Huo to this project.

References

1. Aamir, M., Raut, R., Jhaveri, R.H., Akram, A.: Ai-generated content-as-a-service in iomt-based smart homes: Personalizing patient care with human digital twins. IEEE Transactions on Consumer Electronics (June 2024)
2. Agarwal, S., Sharma, S., Nakov, P., Chakraborty, T.: MemeMQA: Multimodal Question Answering for Memes via Rationale-Based Inferencing. arXiv preprint arXiv:2405.11215 (2024)
3. Akram, J., Tahir, A.: Lexicon and heuristics based approach for identification of emotion in text. In: 2018 International Conference on Frontiers of Information Technology (FIT). pp. 293–297. IEEE (December 2018)
4. Anaissi, A., Braytee, A., Akram, J.: Fine-tuning llms for reliable medical question-answering services. In: 2024 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE (December 2024)
5. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep Learning for Hate Speech Detection in Tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion) (2017)
6. Blaier, E., Malkiel, I., Wolf, L.: Caption Enriched Samples for Improving Hateful Memes Detection. arXiv preprint arXiv:2109.10649 (2021)
7. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated Hate Speech Detection and the Problem of Offensive Language. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 11 (2017)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
9. Du, Y., Li, C., Guo, R., Yin, X., Liu, W., Zhou, J., Bai, Y., Yu, Z., Yang, Y., Dang, Q., Wang, H.: Pp-ocr: A practical ultra lightweight ocr system (2020)
10. Gómez, R., Gibert, J., Gómez, L., Karatzas, D.: Exploring Hate Speech Detection in Multimodal Publications. arXiv preprint arXiv:2005.04982 (2020)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

11. Hamza, A., Javed, A.R., Iqbal, F., Yasin, A., Srivastava, G., Połap, D., ..., Jalil, Z.: Multimodal Religiously Hateful Social Media Memes Classification based on Textual and Image Data. ACM Transactions on Asian and Low-Resource Language Information Processing (2023)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
13. Hee, M.S., Lee, R.K.W., Chong, W.H.: On explaining multimodal hateful meme detection models. In: Proceedings of the ACM Web Conference 2022. p. 3651–3655. WWW '22, Association for Computing Machinery, New York, NY, USA (2022)
14. Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.C., Tan, M., ..., Le, Q.: Searching for MobileNetV3. In: IEEE International Conference on Computer Vision (ICCV). pp. 1314–1324 (2019)
15. Khan, M.T.R., Saad, M.M., Tariq, M.A., Kim, D.: Spice-it: Smart covid-19 pandemic controlled eradication over ndn-iot. Information Fusion **74**, 50–64 (October 2021)
16. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. arXiv preprint arXiv:2005.04790 (2021)
17. Kirk, H.R., Jun, Y., Rauba, P., Wachtel, G., Li, R., Bai, X., ..., Asano, Y.M.: Memes in the Wild: Assessing the Generalizability of the Hateful Memes Challenge Dataset. arXiv preprint arXiv:2107.04313 (2021)
18. Kougia, V., Pavlopoulos, J.: Multimodal or Text? Retrieval or BERT? Benchmarking Classifiers for the Shared Task on Hateful Memes. In: Proceedings of the 2021 Workshop on Online Abuse and Harms (WOAH 2021) (2021)
19. Kovács, G., Alonso, P., Saini, R.: Challenges of Hate Speech Detection in Social Media: Data Scarcity, and Leveraging External Resources. SN Computer Science **2**(2), 95 (Apr 2021)
20. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ..., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neural Information Processing Systems (NeurIPS) paper (2020)
21. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language (2019)
22. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., ..., Ramanan, D.: Evaluating Text-to-Visual Generation with Image-to-Text Generation. arXiv preprint arXiv:2404.01291 (2024)
23. Liu, Z., Braytee, A., Anaissi, A., Zhang, G., Qin, L.: Ensemble pretrained models for multimodal sentiment analysis using textual and video data fusion. In: Companion Proceedings of the ACM Web Conference 2024. pp. 1841–1848 (2024)
24. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. Neural Information Processing Systems (NeurIPS) paper (2019)
25. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks (2019)
26. Meta AI Research: Hateful Memes Challenge and dataset. `https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/`, accessed: 2024-11-23
27. Munzni, S., Dixit, S., Bhat, A.: Classification of Hateful Memes by Multimodal Analysis using CLIP. In: Proceedings of ConIT 2024. pp. 1–5 (2024)
28. OpenAI: Gpt-4o: Openai's most advanced multimodal model. `https://openai.com/index/gpt-4o` (2024), accessed: 2024-10-03

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10

29. Pramanick, S., Dimitrov, D., Mukherjee, R., Sharma, S., Akhtar, M.S., Nakov, P., Chakraborty, T.: Detecting harmful memes and their targets. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2783–2796. Association for Computational Linguistics, Online (Aug 2021)
30. Qian, C., Shi, X., Yao, S., Liu, Y., Zhou, F., Zhang, Z.: Optimized biomedical question-answering services with llm and multi-bert integration. In: 2024 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE (December 2024)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ..., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (2021)
32. Rathore, R.S., Jhaveri, R.H., Akram, A.: Galtrust: Generative adversarial learning-based framework for trust management in spatial crowdsourcing drone services. IEEE Transactions on Consumer Electronics (April 2024)
33. Rehman, A.U., Rehman, Z., Ali, W., Shah, M.A., Salman, M.: Statistical topic modeling for urdu text articles. In: 2018 24th International Conference on Automation and Computing (ICAC). pp. 1–6. IEEE (September 2018)
34. Rizzi, G., Gasparini, F., Saibene, A., Rosso, P., Fersini, E.: Recognizing misogynous memes: Biased models and tricky archetypes. Information Processing and Management **60**(5), 103474–103474 (2023)
35. Zhang, Z., Robinson, D., Tepper, J.: Detecting hate speech on twitter using a convolution-gru based deep neural network. In: Gangemi, A., Navigli, R., Vidal, M.E., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M. (eds.) The Semantic Web. pp. 745–760. Springer International Publishing, Cham (2018)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_10