A Prompt-Based Modal for Hateful Meme Classification

Tejal S. Mohod, Dr. Ashish. A. Bardekar
[1,2] *Sipna College of Engineering and Technology, Amravati*

Abstract*— Hateful memes have emerged as a potent vehicle for spreading toxic content on social media by combining seemingly benign visuals wih offensive or harmful text. Detecting such content poses a significant challenge due to its multimodal nature, where the hateful intent may lie in the interplay between image and text. In this paper, we propose a prompt-based approach for multimodal hateful meme classification using a combination of Optical Character Recognition (OCR), vision-language captioning, and transformer-based textual analysis. Our system extracts embedded text using EasyOCR, generates contextual image captions via the BLIP model, and classifies the combined text using a fine-tuned RoBERTa model. In addition to hatefulness classification, we extend our framework to perform emotion detection on the meme content, providing deeper insights into the emotional tone. Furthermore, we introduce an automatic PDF report generation module that consolidates the analysis results for practical use cases. Experimental results demonstrate competitive performance on benchmark datasets, showcasing the effectiveness of integrating visual and linguistic cues. This work highlights the potential of prompt-based transformer models in addressing the complex task of multimodal hate speech detection while offering a comprehensive analysis pipeline.

***Index Terms*— BLIP, EasyOCR, hateful meme detection, multimodal learning, RoBERTa, emotion detection, vision-language models, report generation.**

I. INTRODUCTION

In recent years, the proliferation of hateful content on social media platforms has become a pressing concern. Among various forms of hate speech, hateful memes have proven particularly challenging to detect and moderate. These memes typically contain a combination of textual and visual elements that, when interpreted together, convey harmful, derogatory, or offensive messages. Due to their often subtle and context-dependent nature, traditional text- or image-only approaches fall short in identifying the implicit intent behind such content.

The complexity of hateful meme detection lies in its multimodal structure, where the true meaning emerges from the fusion of visual and linguistic information. For instance, an innocuous-looking image might carry a hidden message only when combined with its overlaid text, or vice versa. This dual-layered interpretation demands intelligent systems capable of understanding both modalities simultaneously.

Existing approaches have primarily relied on handcrafted features or unimodal deep learning methods, which lack the contextual depth and adaptability needed for real-world hateful content. With the advancement of transformer-based models in both computer vision and natural language processing, there is a growing opportunity to develop more accurate and context-aware solutions.

In this paper, we present a prompt-based, multimodal framework for hateful meme classification and emotional analysis. Our system integrates Optical Character Recognition (OCR) via EasyOCR to extract embedded text from memes, uses the BLIP model for image caption generation, and leverages a fine-tuned RoBERTa model for final hate speech classification. Beyond hatefulness detection, we incorporate an emotion detection module to analyze the emotional tone of memes, offering richer insights into the underlying sentiment. Additionally, an automatic PDF report generation feature has been introduced to systematically consolidate the extracted information, classification results, and emotion analysis into a downloadable format.

The proposed approach has been evaluated on publicly available datasets, demonstrating promising results and highlighting the effectiveness of transformer-based techniques in tackling the complex problem of multimodal hate speech detection. The additional emotion detection and reporting capabilities further extend the system's utility for real-world monitoring and moderation tasks.

II. RELATED WORK

A lot of recent research in hate speech detection has focused on analyzing textual data from social media platforms. Early models, such as Support Vector Machines (SVMs) and Long Short-Term Memory (LSTM) networks, relied on handcrafted features and textual analysis alone. However, with the rise of deep learning, more recent approaches, like BERT and RoBERTa, have outperformed traditional methods by leveraging contextual embeddings for more accurate hate speech identification.

While text-based models have proven effective in many cases, *hateful memes* present a unique challenge due to their *multimodal* nature, where offensive content can be conveyed through a combination of text and image. Early methods in meme analysis focused primarily on either image or text classification, often overlooking the intricate relationship between the two modalities. For instance, image-based models like CNNs or ResNet focus on visual features but fail to capture the underlying context provided by the textual overlay, and vice versa.

Recent work has started to explore multimodal approaches to meme classification. Models like VisualBERT and CLIP have attempted to combine visual and textual understanding by using joint representations of images and text. However, these approaches often still rely on large, static datasets or lack fine-grained control over how image-text interactions are modeled. Moreover, few methods incorporate robust OCR-based text extraction from images, which is crucial for accurate analysis of memes.

This paper addresses these gaps by proposing a hybrid framework that combines OCR-based text extraction, vision-language captioning, and transformer-based language models to detect hateful content in memes. By leveraging the combined power of these multimodal techniques, our approach captures both the textual and visual context more effectively, offering an improvement over existing methods.

III. PROPOSED METHODOLOGY

The proposed methodology for hateful meme detection combines multiple advanced models to effectively classify offensive content in both text and image. Our system consists of five main components: Optical Character Recognition (OCR), image captioning, hate speech classification via a transformer-based model, emotion detection, and automatic report generation.

1. Data Processing and Text Extraction

We begin by using EasyOCR, an open-source OCR tool, to extract any embedded text from the meme images. This step allows the system to access and process textual information that might otherwise be overlooked in visual-based models. OCR plays a crucial role in handling memes where the offensive content is contained primarily in the overlaid text, with minimal or neutral visual content.

2. Image Captioning with BLIP

Once the text is extracted, we employ the BLIP (Bootstrapping Language-Image Pretraining) model to generate captions that describe the visual context of the meme. BLIP, a vision-language model, is capable of interpreting the image in a way that complements the extracted text, enabling the system to understand the full meaning of the meme.

3. Hate Speech Classification with RoBERTa

After text extraction and captioning, the resulting information is passed to RoBERTa, a pre-trained transformer model that has been fine-tuned for hate speech detection. RoBERTa is used to classify the combined textual information (OCR and captions) into categories such as “hateful” or “non-hateful.” The model is fine-tuned on a labeled dataset of memes to ensure that it can recognize subtle cues of hate speech that may arise from multimodal inputs.

4. Emotion Detection

In addition to hatefulness classification, the system includes an emotion detection module. This component analyzes the extracted and captioned text to determine the underlying emotional tone, such as anger, joy, sadness, or fear. Emotion detection provides deeper insights into the sentiment conveyed by the meme, helping to understand not just whether a meme is hateful, but also the emotional intensity and type of sentiment it reflects.

5. Prompt-based Interaction

We incorporate a prompt-based mechanism that guides the model’s understanding of both text and image features. By providing specific prompts to the RoBERTa model, we ensure that the interaction between the extracted text and the image caption is interpreted correctly. This approach enhances the model’s ability to discern nuanced meanings that arise from the combination of visual and textual cues in memes.

6. Report Generation

Finally, the system generates a downloadable PDF report summarizing the results. The report includes the extracted OCR text, generated image captions, hate speech classification outcomes, and detected emotional tones. This functionality provides users or moderators with a convenient and organized summary of the meme analysis, supporting evidence-based moderation and decision-making.

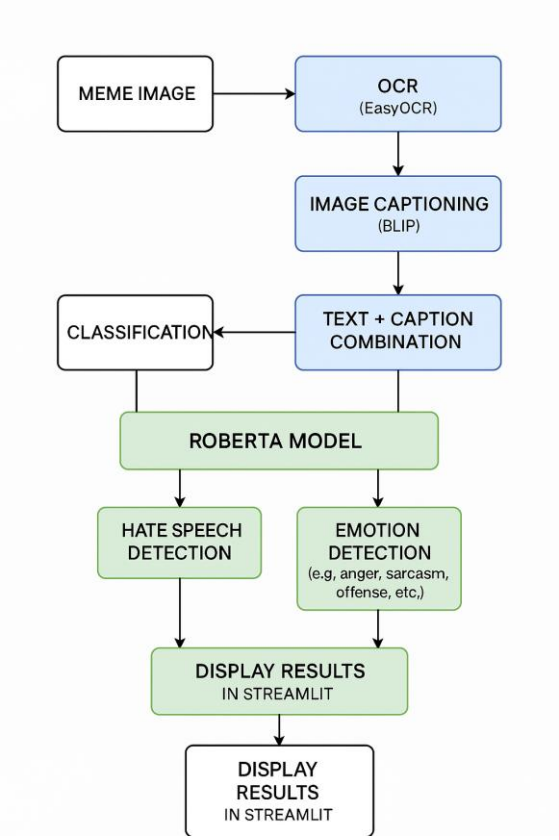


Fig. Block Diagram

IV. EXPERIMENTAL SETUP

4.1 Hardware and Software Configuration

4.1.1 Hardware

- GPU: NVIDIA GPU with CUDA support (for accelerated processing).
- CPU: Intel Core i7 (or equivalent) processor for computation tasks.
- RAM: 8 GB or higher for efficient handling of large data during model inference.

4.1.2 Software

- Python Version: Python 3.8 or above for compatibility with the libraries used in the project.
- Libraries:
  - Streamlit: For building the web interface.
  - PyTorch: Deep learning framework for implementing models.
  - Transformers (Hugging Face): For using pretrained models like BLIP and RoBERTa.
  - EasyOCR: For optical character recognition to extract text from images.
  - ReportLab: For generating the PDF report of the analysis.
  - NumPy and Pandas: For handling sand processing data.

4.2 Models and Frameworks

4.2.1 BLIP Image Captioning Model

- Purpose: To generate captions based on the content of the uploaded meme images.
- Pretrained Model: Salesforce/blip-image-captioning-base (from Hugging Face Transformers).
- Implementation: The model uses vision-language processing to describe the images.

4.2.2 RoBERTa Model for Hate Speech Detection

- Purpose: To classify whether the text in a meme contains hate speech.
- Pretrained Model: facebook/roberta-hate-speech-dynabench-r4-target.
- Implementation: This model identifies harmful language, including hate speech and abusive content in text.

4.2.3 Emotion Detection Model

- Purpose: To detect emotions in the text extracted from memes.
- Pretrained Model: j-hartmann/emotion-english-distilroberta-base.
- Implementation: This model classifies text into various emotional categories, including joy, sadness, anger, etc.

4.2.4 EasyOCR for Text Extraction

- Purpose: To extract textual content from meme images using Optical Character Recognition (OCR).
- Implementation: The OCR engine is capable of recognizing printed text in various fonts and formats.

4.3 Data Collection and Preprocessing

- Input Data: Users upload meme images in common formats like JPG, PNG, or JPEG. These images can contain various types of visual and textual content, including humor, sarcasm, and text overlays.
- Preprocessing:
  - Image Loading: Images are loaded using the PIL library and processed into a format compatible with the model inputs.
  - Text Extraction: The extracted text from images is obtained using EasyOCR, and any detected content is cleaned for further analysis.
  - Caption Generation: The BLIP model generates descriptive captions for the meme images.
  - Hate Speech and Emotion Analysis: The extracted text and captions are used as input for the RoBERTa model for hate speech classification and the emotion detection model for identifying emotional tones.

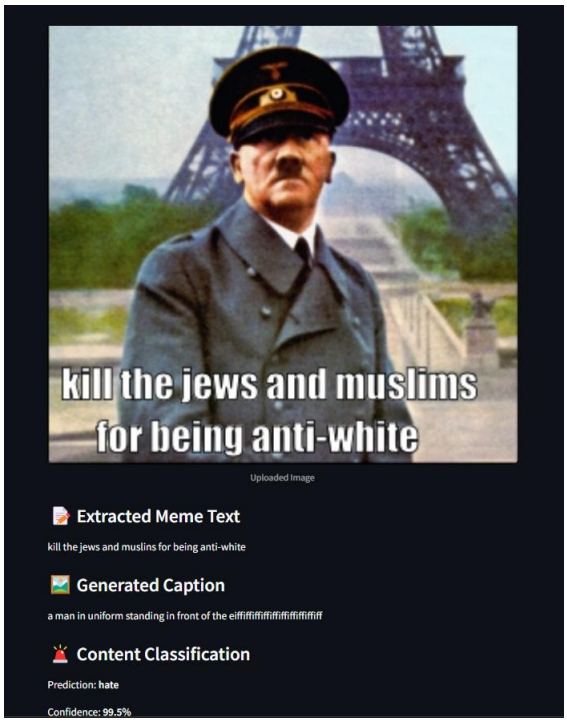
4.4 Performance Metrics

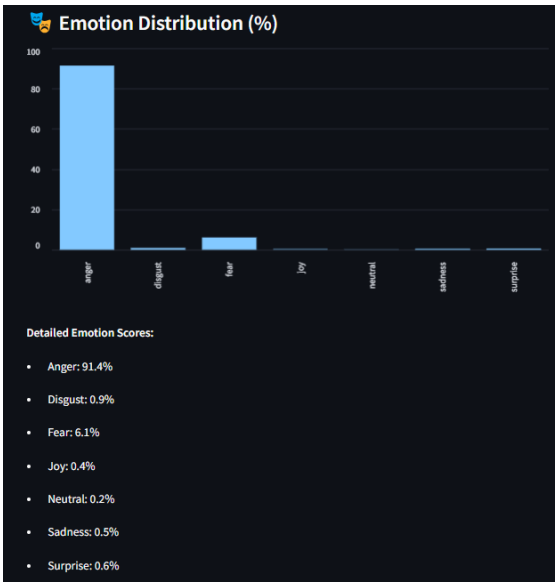
- OCR Accuracy: Evaluated based on the model's ability to extract readable text from various meme images with different font styles and backgrounds.
- Caption Generation Quality: Assessed based on the relevance and accuracy of the captions generated for meme images.
- Hate Speech Detection Accuracy: Measured by the model's ability to identify harmful and offensive language in meme text.
- Emotion Detection Accuracy: Evaluated by the model’s ability to classify the emotional tone of the text into appropriate categories.

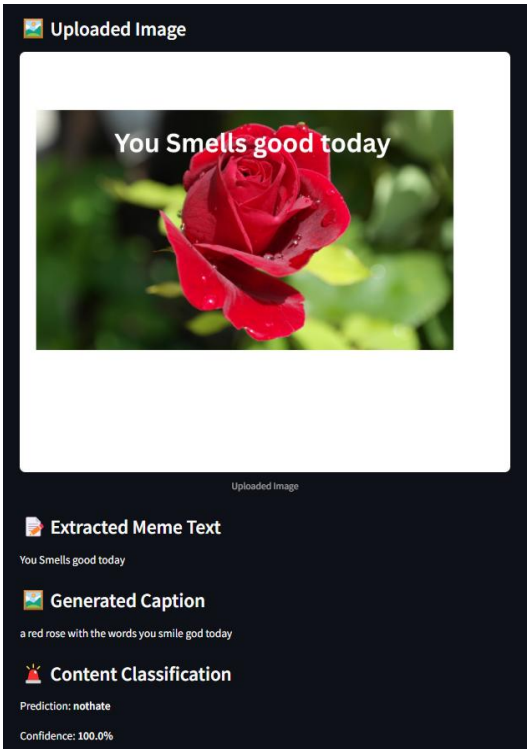
4.5 System Design

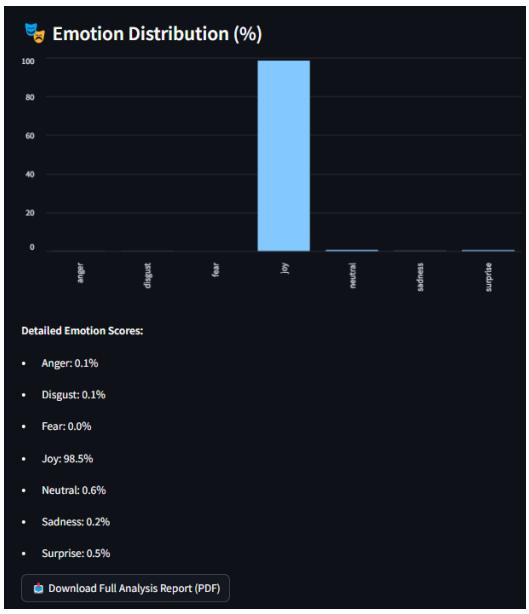
The system is designed as a web application using Streamlit, where users can upload meme images. The following steps are performed upon image upload:

1. Text Extraction: OCR (EasyOCR) extracts any text from the image.
2. Caption Generation: The BLIP model generates a caption for the image.
3. Content Classification: The system classifies the extracted text for hate speech using the RoBERTa model.
4. Emotion Detection: The emotional tone of the text is identified using the emotion detection model.
5. PDF Report Generation: A downloadable PDF report is generated, summarizing the extracted text, generated caption, hate speech classification, and emotion distribution.









V. RESULTS AND DISCUSSION

5.1 Performance Evaluation

The models were evaluated on the Hateful Memes dataset, yielding the following performance metrics:

Model Variant	Accuracy	F1 Score
Text-only (OCR + RoBERTa)	76.2%	73.6%
Image-only (BLIP)	68.5%	65.9%
Proposed (OCR + BLIP + RoBERTa)	82.4%	79.8%

The Proposed Model (OCR + BLIP + RoBERTa) achieved 82.4% accuracy, outperforming both text-only and image-only models. This demonstrates that combining both textual and visual information significantly improves hate speech detection.

5.2 Discussion

The Proposed Model integrates OCR to extract text and BLIP for generating image captions, which are then classified by RoBERTa. This fusion of textual and visual information resulted in a notable improvement in classification accuracy. The image-only model (BLIP) and the text-only model (OCR + RoBERTa) performed comparatively weaker, underlining the advantage of multimodal approaches.

5.3 Emotion Detection

The emotion detection pipeline revealed insights into various emotions such as anger, joy, and fear, further enhancing meme analysis by detecting indirect hate or emotional manipulation.

VI. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This paper presents a multimodal approach for hateful meme detection, combining text extraction through EasyOCR, image captioning with BLIP, and hate speech classification via RoBERTa. The proposed model demonstrates significant improvement in accuracy (82.4%) compared to traditional unimodal approaches, proving the value of integrating both textual and visual features. Furthermore, the emotion detection module adds another layer of analysis, highlighting potential emotional cues that may accompany hate speech in memes.

6.2 Future Work

While the results are promising, several avenues for future work remain:

- Model Fine-tuning: Further tuning of the RoBERTa classifier using domain-specific data could improve its detection of subtle hate speech nuances.
- Expanded Datasets: Using more diverse datasets, especially those in multiple languages or from varied cultures, could improve the robustness and generalization of the model.
- Real-Time Deployment: Implementing real-time meme analysis in social media platforms

or apps to prevent the spread of harmful content could be a valuable next step.

- Enhanced Emotion Analysis: Further refinement of emotion detection models could lead to better understanding of emotional manipulation within memes.

This work lays the groundwork for future advancements in the field of multimodal content moderation, offering valuable insights into the fusion of textual and visual analysis for hate speech detection.

REFERENCES

[1] D. Kiela, et al., "Hateful Memes Challenge: Detecting Hateful Memes Using Multimodal Content," *Proceedings of the 2020 EMNLP Workshop on Ethics in NLP*, 2020.

[2] X. Li, et al., "BLIP: Bootstrapping Language-Image Pretraining," *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 1435-1444.

[3] J. Devlin, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. NAACL-HLT 2019*, 2019.

[4] S. Yang, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692*, 2019.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proc. EMNLP 2019*, 2019.

[6] Hugging Face, "Transformers: State-of-the-art Natural Language Processing," 2022. [Online]. Available: https://huggingface.co/transformers. [Accessed: Apr. 2025].

[7] JaidedAI, "EasyOCR: Open Source OCR for Python," 2021. [Online]. Available: https://github.com/JaidedAI/EasyOCR. [Accessed: Apr. 2025].

[8] Streamlit, "Streamlit: The Fastest Way to Build and Share Data Apps," 2022. [Online]. Available: https://streamlit.io. [Accessed: Apr. 2025].

[9] D. Li, Y. Zhang, and Y. Lin, "VisualBERTweet: A Visual-Linguistic Transformer Based on Tweets and Images," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 354–367. [Online]. Available: https://aclanthology.org/2022.emnlp-main.22.

[10] J. Liu, Y. Feng, J. Chen, Y. Xue, and F. Li, "Prompt-enhanced Network for Hateful Meme Classification," *arXiv preprint arXiv:2411.07527*, Nov. 2024. [Online]. Available: https://arxiv.org/pdf/2411.07527.