Hateful Meme Detection

1st Ankit Kumar*, 2nd Disha Soni†, 3rd Shreya Sinha‡, 4th Avantika Soni§, 5th Taniha Bhutani¶
*†‡§¶ Department of Information Technology,
Indian Institute of Information Technology, Allahabad, India
*iit2022256@iiita.ac.in, †iit2022260@iiita.ac.in, ‡iib2022034@iiita.ac.in,
§iib2022045@iiita.ac.in, ¶iit2022207@iiita.ac.in

***Abstract*—The proliferation of hateful memes on social media platforms has become a pressing issue, as such content often targets individuals or groups based on race, gender, ethnicity, or other personal characteristics. Detecting hateful memes requires the ability to understand both visual and textual cues and their combined context.**

This paper proposes a novel approach to detecting hateful memes by combining advanced machine learning models. We use the Qwen-2-VL-7B-Instruct Vision-Language Model (VLM) for text feature extraction and YOLOv8 for object detection. OpenCV's DeepFace model extracts demographic features such as gender, age, emotion, and ethnicity. Temporal dependencies are handled by an LSTM neural network. The Qwen-2-VL-7B Instruct model classifies memes based on both textual and visual cues. Our approach, integrating these multimodal models, improves classification accuracy and offers a scalable solution for detecting hateful memes.

I. Introduction

A meme contains direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define attack as violent or dehumanizing(comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation.A meme includes mocking hate crime is also considered hateful meme.

Hateful content detection, especially in multimodal formats like memes, has gained significant attention in recent years due to the rising influence of social media platforms where such content proliferates. Memes combine images, text, and sometimes other modalities like video to communicate often nuanced and implicit messages, making the detection of hateful intent complex.

A. Background Information

Hate speech detection, traditionally focused on textual content, has been an active research area for decades. Early methods employed rule-based approaches and keyword detection. However, with the advent of deep learning, machine learning models, especially recurrent neural networks (RNNs) and transformers like BERT (Bidirectional Encoder Representations from Transformers), have been employed for more context-aware text classification tasks (Schmidt & Wiegand, 2017). Despite advances in text-only models, hateful memes require joint interpretation of both visual and textual information, adding a layer of complexity.

B. MultiModal Content and Hateful Memes

As a form of multimodal content, often convey meaning through both visual and textual components. The challenge lies in their contextual nature, where the text and image combination may produce hateful meanings that are not apparent from analyzing one modality in isolation. For example, a benign image might be paired with offensive text or vice versa. Thus, research on hateful meme detection must involve a multimodal approach that combines computer vision and natural language processing (NLP)

II. Dataset

For this project, we utilized the Hateful Memes Dataset, created by Facebook AI. The dataset consists of over 10,000 multimodal examples, combining both images and text, specifically designed to help researchers advance the detection of hateful content in memes. Each example in the dataset includes a meme image paired with associated text, offering a wide variety of content that targets different demographics based on race, gender, ethnicity, and other personal characteristics.

The images in the dataset were licensed from Getty Images, ensuring proper usage rights for research purposes. The multimodal nature of the dataset provides unique challenges, as the hateful intent often lies in the combination of visual and textual cues rather than in either modality alone. This dataset serves as a robust foundation for training and evaluating machine learning models in the task of multimodal hate speech detection.

Additionally, Meta has provided baseline models trained on this dataset, offering researchers a starting point for further experimentation and development. These pre-trained models, alongside the dataset, aim to support the research community in developing more accurate systems for detecting hateful memes.

III. Literature Review

A. Prior Research

Prior research on harmful meme detection has emphasized the importance of integrating both visual and textual data to address offensive content. For instance, Naseem (2024) in *Decoding Memes: A Comprehensive Analysis of Late and Early Fusion Models for Explainable Meme Analysis* explores multimodal meme sentiment detection by combining visual features (from models like ViT and VGG-16) with textual features (from BERT and DistilBERT). This study highlights the

effectiveness of feature fusion but also notes challenges such as class imbalance. Similarly, Ma and Li (2024), in *RoJiNG-CL at EXIST 2024: Leveraging Large Language Models for Multimodal Sexism Detection in Memes*, tackle sexism detection in memes by using GPT-4 for textual descriptions, integrated with vision-language models like CLIP, to detect subtle sexist messaging. Their approach ranked highly, demonstrating the effectiveness of combining large language models with visual data. In *CapAlign: Improving Cross Modal Alignment via Informative Captioning for Harmful Meme Detection*, Ji et al. (2023) propose generating high-quality image captions through dialogues between language and vision models, showing that informative captioning and cross-modal alignment improve meme detection performance, surpassing state-of-the-art methods. Lastly, Huang et al. (2023), in *Evolver: Chain-of-Evolution Prompting to Boost Large Multimodal Models for Hateful Meme Detection*, introduce Evolver, which uses an evolving meme pool to adapt to new and unseen memes, boosting the model's ability to generalize and detect hateful content. These studies collectively emphasize the significance of multimodal data processing and adaptation strategies to enhance harmful meme detection.

B. *Research Gap*

Previous research on hateful meme detection primarily focused on separate text or image analysis, with limited integration of both modalities. Our approach bridges this gap by fine-tuning advanced models like BERT, RoBERTa, and Qwen-2.5-5B Instruct for joint text and image classification, enhancing accuracy. While studies like those by Ma & Li (2024) and Naseem (2024) utilized vision-language models, few leveraged large, fine-tuned models or incorporated demographic features such as gender, age, and emotion from facial analysis, which our approach addresses using DeepFace. Additionally, while previous work often relied on pre-trained models without custom datasets, we develop a tailored dataset for better model fine-tuning, achieving improved performance. Our approach also balances computational efficiency with robust feature extraction, outperforming traditional methods by offering a more integrated and scalable solution for hateful meme detection.

Therefore, compared to previous work, our approach combines advanced multimodal analysis techniques to address the challenges in hateful meme classification. We developed and tested multiple pipelines that integrate both text and image analysis, leveraging state-of-the-art models and methodologies. The detailed methods that we implement will be introduced in detail in next section.

IV. The Proposed Method

In this section we will introduce all proposed methods for the task as we mentioned in the introduction section.

A. *Architecture*

The proposed architecture combines text and image analysis through a well-defined multi-stage architecture, consisting of Pipeline 1 for preprocessing and feature extraction and Pipeline 2, which employs two different methods for model inference and ensemble predictions..Figure 1 illustrates the system.

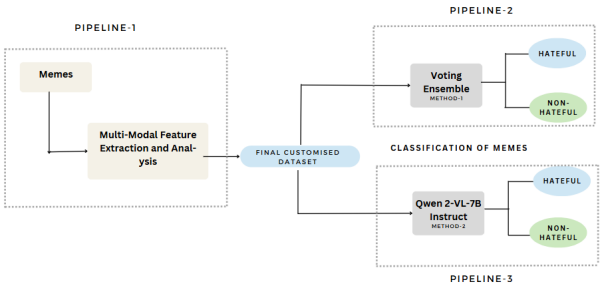


Fig. 1. Architecture of our complete model

B. *Pipeline-1 : Multi-Modal Feature Extraction and Analysis*

This pipeline integrates multi-modal feature extraction, combining textual and visual information for meme analysis. It utilizes advanced models for object detection, text recognition, sentiment analysis, and facial attribute extraction, creating a custom dataset to identify and analyze hateful or offensive content.

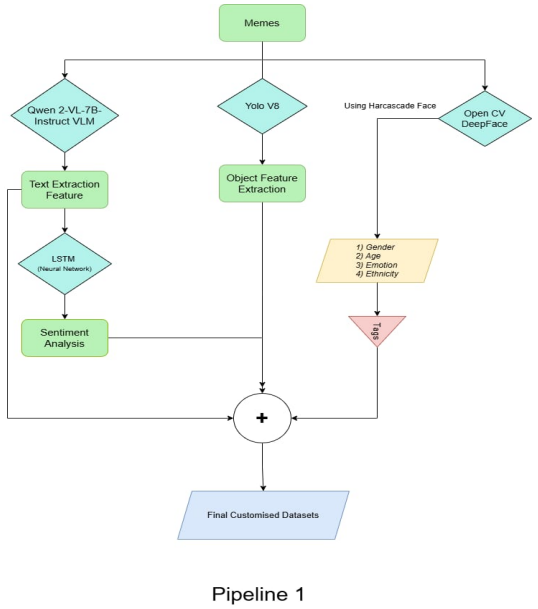


Fig. 2. Multi-Modal Feature Extraction and Analysis

1) *Image, Object, and Text Feature Extraction:*

- **Text Extraction:** The **Qwen2-VL-7B-Instruct** model was utilized for extracting meaningful text features from memes by leveraging its multimodal capabilities. This model processes both the image and a contextual text prompt, generating a structured representation of the embedded text. It effectively captures sentiment, semantics, and contextual relevance, crucial for detecting hate-speech in memes. Pre-trained on diverse multimodal

datasets, the model is adept at understanding intricate text-image relationships, making it ideal for this application. Its robust architecture minimizes the need for extensive pre-processing, ensuring efficiency in handling complex meme content. The extracted text features were integrated into the classification pipeline, enhancing the accuracy of hateful meme detection.

Moreover, the model's capability to adapt to nuanced inputs highlights its versatility for future expansions, such as supporting multilingual text analysis or adding interpretability features. By combining advanced text recognition with image understanding, the model establishes a strong foundation for tackling challenges in multimodal hate-speech detection. Its implementation underscores the potential of leveraging large-scale pre-trained models in solving real-world problems.

- **Object Detection:** We utilized **YOLOv8** for object detection in meme images, leveraging its advanced convolutional neural network (CNN) to identify relevant objects. The model outputs bounding boxes, class labels, and confidence scores for each detected object. To improve clarity, we ensured all detected objects are uniquely labeled, enabling efficient tracking and analysis. This unique labeling facilitates organized handling of detection results in downstream tasks. The pre-trained YOLOv8 model was specifically used to capture contextually significant objects within the memes. Its high accuracy and efficiency make it ideal for analyzing visual elements in hateful content. This approach effectively meets the requirements for precise and systematic object detection in multimodal analysis.

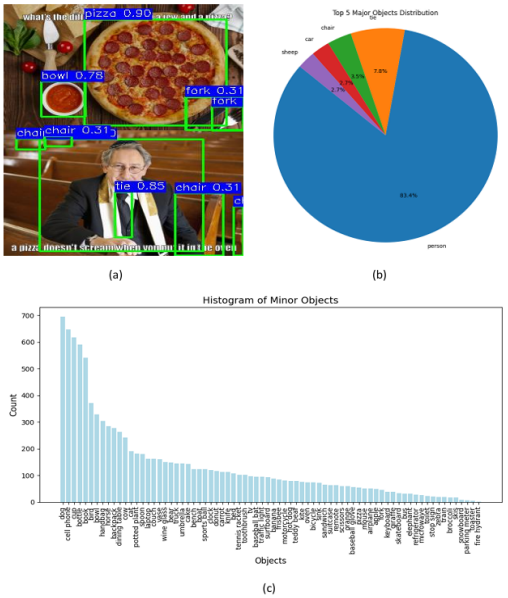


Fig. 3. (a) Object detection highlights items like 'pizza,' 'bowl,' and 'chair' with confidence scores. (b) Pie chart shows the top 5 major objects detected in the dataset. (c) Histogram depicts frequencies of minor objects, with 'dog,' 'cell phone,' and 'cup' being the most common.

- **Face Extraction:** Leveraged **Haar Cascade** classifiers for detecting and isolating faces within images, providing focused face data for additional analysis. Haar Cascade Classifiers, a traditional machine learning-based approach for face detection. It uses a cascade of classifiers trained on positive and negative samples of faces to detect regions of interest in an image. In our model implementation, we identified potential face regions based on features like edges and intensity differences and set *scaleFactor* equals 1.3 and *minNeighbors* equals 5 to ensure robust detection by balancing sensitivity and false positives. The largest face is cropped from the original image and resized to 128x128 pixels for uniformity. It is normalized for downstream tasks like emotion and demographic analysis. In memes, detecting faces is crucial for analyzing emotions, gender, and age, as many memes rely on facial expressions and human subjects to convey messages.

*2) **Tagging of Hate-related Information:**:*

- **Tag Extraction:** Used Face Type Detection and Emotion Detection models to label content with hate-related tags where applicable.
- **Additional Metadata: OpenCV DeepFace** is incorporated for analyzing gender, age, emotion, and ethnicity from facial data, contributing to the context of the image. DeepFace leverages a variety of pre-trained models (e.g., VGG-Face, Google FaceNet, OpenFace) that are trained to recognize facial features and predict attributes such as gender, age, emotion, and race based on facial landmarks. These models are pre-trained on large facial datasets, which allows them to generalize well to unseen faces and classify attributes accurately. DeepFace can classify facial attributes, which help determine the sentiment and potential bias in memes. For example, identifying the gender, age, emotion, and race of individuals in the meme can aid in detecting stereotypes or hateful content.

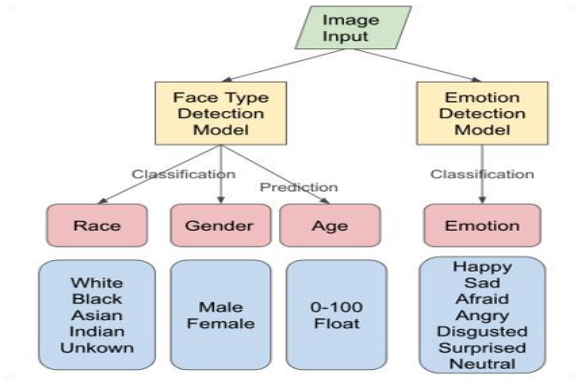


Fig. 4. An illustration of model input and output

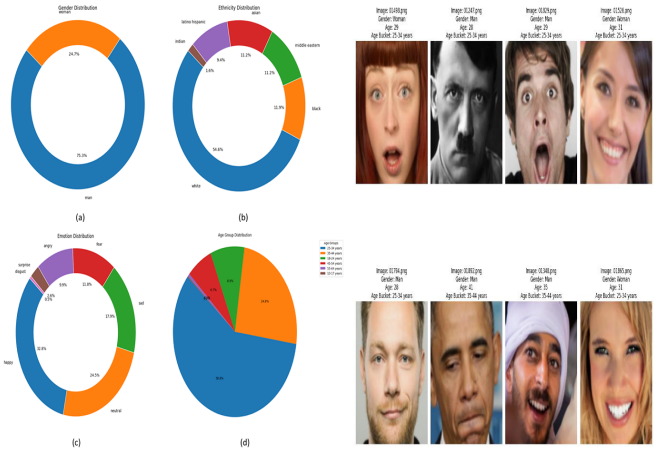


Fig. 5. Illustration of (a) Gender Distribution (b) Ethnicity Distribution (c) Emotion Distribution (d) Age Group Distribution across the dataset.

*3) **Unimodal Classification on Image and Text Features:**:*

- **Image Classification:** Based on the objects and faces detected, models were employed to classify whether the image content might be indicative of hate speech or offensive imagery.
- **Text Classification:** Processed extracted text using NLP techniques to classify content based on hate-related keywords or phrasing.

*4) **Sentiment Analysis of Text:**:*

- LSTM Models are used for sentiment analysis, determining the emotional tone of the text (positive, neutral, negative). The LSTM model helps capture long-term dependencies in text, making it effective for understanding context and sentiment in potentially hateful content.

*5) **Concatenation of Three Channels' Output:**:*

- These extracted features from text, image, and tags are combined into a single input vector. This consolidated input allows for a comprehensive representation of the content across multiple modalities, enabling a more robust analysis. The features and labels generated from Pipeline 1 were used to create a custom dataset for training advanced models.

*6) **Preparing Dataset:**:* Pipeline-1 serves as the input for Pipeline-2 and Pipeline 3. This dataset is meticulously structured, containing features such as:

- Textual content (detected from memes).
- Demographic attributes (e.g., age, gender, race).
- Sentiment and emotion scores.
- Object and visual features extracted from meme images.

C. PipeLine 2: Fine-Tuned Models for Hateful Meme Classification

Pipeline 2 fine-tunes advanced models like BERT, RoBERTa, and Qwen-2.5-5B on a custom dataset derived from Pipeline 1's features. These models integrate text and image inputs for accurate multimodal hateful meme classification.

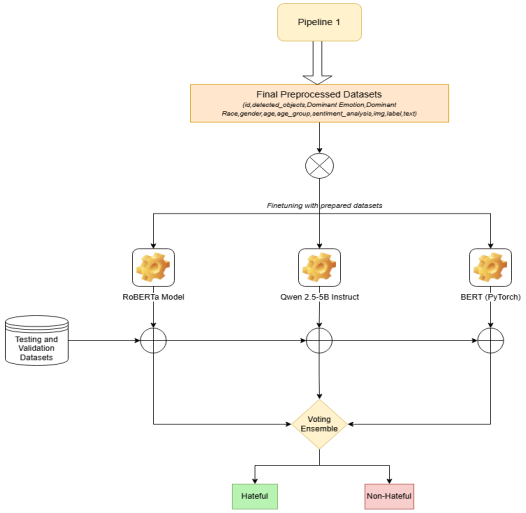


Fig. 6. Pipeline representing voting ensemble method for classification

The models are fine-tuned using the Final Preprocessed Dataset generated from Pipeline 1, which includes features such as dominant emotion, demographic attributes, and sentiment analysis results. Each model provides independent predictions on whether a meme is hateful or non-hateful.

*1) **BERT (PyTorch):*** BERT (PyTorch) was implemented using the Hugging Face Transformers library, which provides a flexible and efficient framework for fine-tuning pre-trained models. This variant was fine-tuned on the custom dataset for meme analysis, focusing on identifying hateful intent within the textual content.

The model consists of 12 transformer layers, each with 768 hidden units and 12 attention heads, enabling it to capture complex contextual relationships in text. It uses a bidirectional attention mechanism, meaning the model considers both left and right contexts simultaneously.

- **Results:**

TABLE I
BERT (PyTorch) Model Evaluation Metrics

Metric	Final Metrics	Evaluation Metrics
Accuracy	67.65% (0.676471)	71.29% (0.712941)
F1 Score	0.5557 (0.555735)	0.5808 (0.580756)
Precision	0.5733 (0.573333)	0.6426 (0.642586)
Recall	0.5392 (0.539185)	0.5298 (0.529781)
Validation Loss	1.6969 (1.696857)	-
Evaluation Loss	-	0.5685 (0.568508)

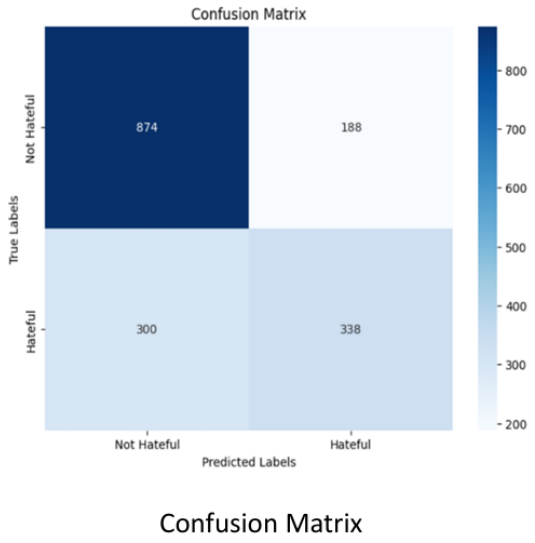
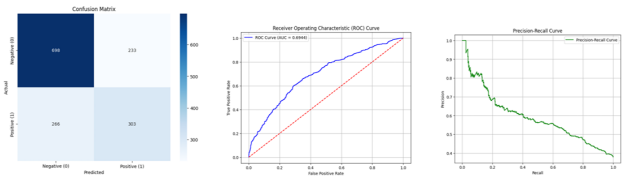


Fig. 7. Visualization of BERT (PyTorch) model performance metrics

*2) **BERT (TensorFlow)**:* BERT (TensorFlow) was implemented using TensorFlow's Keras Functional API, leveraging TensorFlow's robust scalability for handling large datasets. Like its PyTorch counterpart, this variant was fine-tuned for the specific task of hateful meme detection.
The TensorFlow implementation retains the same core architecture as the PyTorch model: 12 layers, 768 hidden units, and 12 attention heads. It supports advanced optimizations such as mixed precision training for faster performance.



(a) Confusion Matrix (b) ROC Curve (c) Precision Recall curve

Fig. 8. Visualization of BERT (TensorFlow) model performance metrics

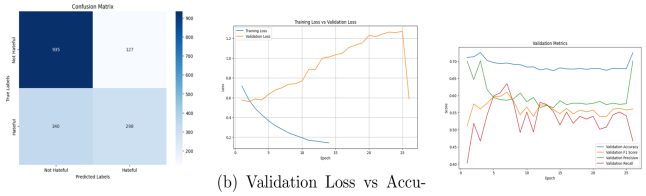
- **Results:**

TABLE II
BERT (TENSORFLOW) MODEL EVALUATION METRICS

Metric	Without Tags	With Tags
Accuracy	65.00%	66.73%
AUC Score	0.6970	0.6944
Precision	0.5364	0.5653
Recall	0.5694	0.5325
F1 Score	0.5524	0.5484

Note: A comparative analysis of BERT implemented in TensorFlow and PyTorch concluded that the performance metrics of BERT in PyTorch are superior. Consequently, we opted to use the PyTorch implementation of BERT.

*3) **Robustly Optimized BERT (RoBERTa)**:* RoBERTa is a state-of-the-art transformer-based model designed to handle complex language tasks through masked language modeling. Its robust architecture and extensive pre-training make it highly suitable for analyzing text in challenging scenarios, such as detecting hateful memes.



(a) Confusion Matrix (b) Validation Loss vs Accuracy Loss (c) Validation Metrics

Fig. 9. Visualization of RoBERTa model performance metrics

- **Results:**

TABLE III
FINAL AND EVALUATION METRICS FOR ROBERTA FINE-TUNING

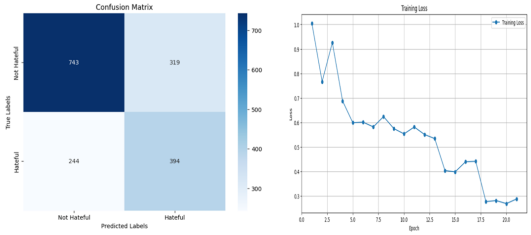
Metric	Final Metrics	Evaluation Metrics
Accuracy	67.82% (0.678235)	72.53% (0.725294)
F1 Score	0.5578 (0.557801)	0.5607 (0.560677)
Precision	0.5759 (0.575960)	0.7012 (0.701176)
Recall	0.5408 (0.540752)	0.4671 (0.467085)
Validation Loss	1.2714 (1.271412)	-
Evaluation Loss	-	0.5858 (0.585753)

*4) **Qwen2.5-5B-Instruct**:* Qwen 2.5-5B Instruct is a transformer-based model fine-tuned on instruction-following tasks, enabling it to better align with user prompts and provide contextual and coherent responses. The model excels in multimodal scenarios, where it combines visual and textual understanding, making it suitable for tasks such as meme analysis.

- **Results:**

TABLE IV
EVALUATION METRICS FOR QWEN2.5 FINE-TUNING

Metric	Evaluation Result
Accuracy	66.88% (0.668824)
F1 Score	0.5833 (0.583272)
Precision	0.5526 (0.552595)
Recall	0.6176 (0.617555)
Validation Loss	0.8865 (0.886505)



(a) Confusion Matrix (b) Training Loss

Fig. 10. Visualization of Qwen2.5-5B-Instruct model performance metrics

*5) **Voting Ensemble** :* Using the outputs of the above three models, we apply a hard weighted voting ensemble to classify memes as either hateful or non-hateful. This approach leverages the predictions from all models, selecting the final label based on the majority vote. Weighted Voting Ensemble: BERT, Qwen, and RoBERTa = [0.90, 0.80, 1.00]

```
Precision (Weighted): 0.7830
Recall (Weighted): 0.7860
F1 Score (Weighted): 0.7840
Precision (Micro Avg): 0.7860
Recall (Micro Avg): 0.7860
F1 Score (Micro Avg): 0.7860
              precision    recall  f1-score   support

           0       0.83      0.86      0.84       658
           1       0.70      0.65      0.68       342

    accuracy                           0.79      1000
   macro avg       0.76      0.75      0.76      1000
weighted avg       0.78      0.79      0.78      1000
```

Fig. 11. Performance metrics obtained from Pipeline 2

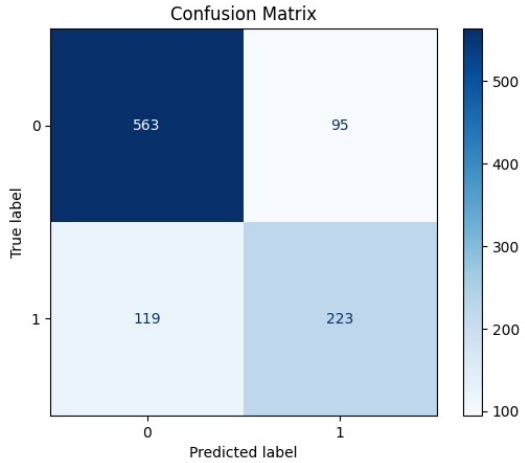


Fig. 12. Confusion Matrix of Voting Ensemble

D. *PipeLine 3: Pre-Trained Model with Multimodal Inputs*

We utilized the Qwen-2-VL-7B Instruct model to process both the extracted text and image tags simultaneously. This pre-trained model directly provided predictions and explanations for the classification task.

*1) **Qwen2-VL-7B Instruct**:* The Qwen2-VL-7B model is a pre-trained large multimodal model designed to handle both images and text. It is used here for the task of hateful meme detection. The model is loaded from Hugging Face's repository using the from pretrained method, where both the model and the processor are initialized to handle the input format and output generation.

Note: Due to computational limitations, the Qwen2-VL-7B model is used without fine tuning in this application. Fine-tuning this model would require a much higher computational power, particularly more GPU resources.

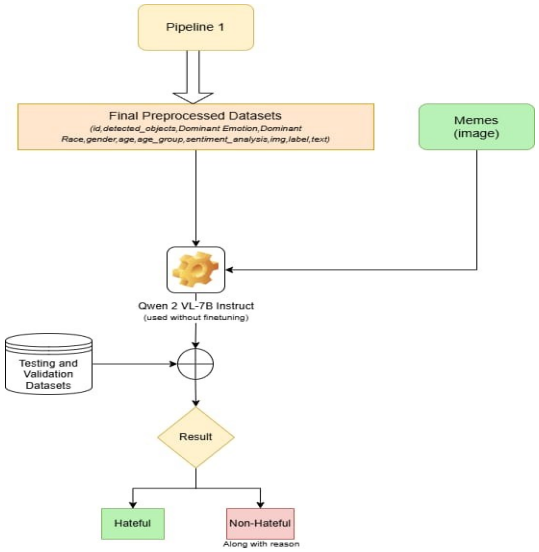
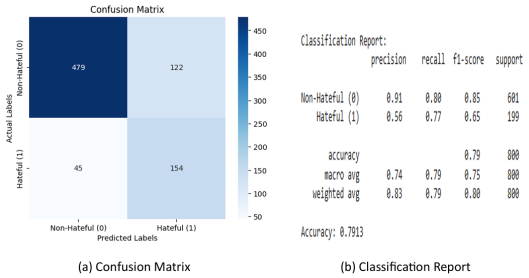


Fig. 13. Pipeline representing Qwen2-VL-7B Instruct model



(a) Confusion Matrix (b) Classification Report

Fig. 14. Visualization of Qwen2-VL-7B model performance metrics

E. *Non-Fine Tuned Model*

*1) **Justification for Not Applying Fine-Tuning**:*

- **High GPU Demands:** Fine-tuning models like Qwen2-VL-7B demands considerable GPU resources, which is challenging given the model's large size and the extensive training data required.
- **Computational Constraints:** The available hardware does not support the efficient fine-tuning of such large models, considering the time and memory needed.

- **Effectiveness of Pre-Trained Model:** Despite the lack of fine-tuning, the Qwen2-VL-7B model is highly capable for meme classification tasks due to its pre-training on a wide range of multimodal datasets, making it well-suited for the task without requiring further fine-tuning.

*2) **Functionality Without Fine-Tuning**:* The model works effectively for meme detection without fine-tuning because it has been pre-trained on diverse multimodal data that includes images and textual content. The pre-trained weights allow the model to generate relevant contextual information based on the input text and image, providing an accurate prediction of whether the meme is hateful or not. The key advantage here is that the model's training on a wide variety of data gives it general knowledge about handling different kinds of inputs, making it robust even without task-specific fine-tuning.

V. Result and Analysis

The evaluation of various models for hateful meme detection was conducted using a diverse set of architectures. Qwen2-VL-7B model outperformed others significantly, achieving an accuracy of 79.13%, an F1 Score of 0.80, and precision of 0.83. As a vision-language model, it leveraged its ability to process both textual and visual inputs simultaneously, leading to higher precision and recall in detecting hateful memes. The results, as summarized in Table, highlight the strengths and limitations of each approach.

TABLE V
Final Model Performance Metrics

Model	Accuracy	F1 Score	Precision
RoBERTa	67.82%	0.5578	0.5758
Qwen 2.5-5B Instruct	66.88%	0.5833	0.5526
BERT (PyTorch)	67.65%	0.5557	0.5733
BERT (TensorFlow) (without tags)	65%	0.5524	0.5364
BERT (TensorFlow) (with tags)	66.73%	0.5484	0.5653
Voting Ensemble	78.60%	0.76	0.76
Qwen 2 VL-7B Instruct	*79.13%	*0.80	*0.83

VI. Conclusion and Future Work

In this study, we developed a multi-modal pipeline to classify hateful memes by integrating text and image processing techniques. Utilizing advanced models like Qwen2-VL-7B Instruct for multimodal analysis, YOLOv8 for object detection, and DeepFace for facial feature extraction, we achieved a comprehensive understanding of meme content. Despite computational constraints preventing fine-tuning of larger models, our pipeline demonstrated the effectiveness of pre-trained models in handling the complex task of hateful meme detection. This work highlights the potential of multi-modal approaches to improve content moderation in online platforms.

Future efforts will focus on addressing the limitations and expanding the capabilities of the proposed pipeline. Fine-tuning large multimodal models like Qwen2-VL-7B, enabled by enhanced computational resources, could significantly improve task-specific performance and accuracy. Additionally, expanding the dataset to include a broader range of memes across diverse languages and cultural contexts will enhance the model's robustness and generalizability. Optimizing the pipeline for real-time hateful meme detection can facilitate seamless integration with social media platforms, enabling proactive content moderation. Furthermore, incorporating explainable AI techniques will provide greater transparency and trust by offering interpretable reasons for classification outcomes. These advancements will ensure that the system evolves into a more effective, scalable, and globally relevant tool for combating hateful content online.

VII. Timelines and Deliverables

- **Phase 1**: Research and Literature Review (July - mid Aug)
  Conduct a comprehensive review of existing literature and current models for hateful meme detection. Analysed about it and understands the working and the shortcomings of all the related works in this field.
- **Phase 2**: Dataset Collection and Preprocessing (mid Aug - mid Sep)
  Gather and preprocess datasets, including labeled examples of hateful memes and related content. Finding datasets was really complex tasks as very limited no of datasets are available for this task.
- **Phase 3**: Model Development and Evaluation (mid Sep - Oct)
  We developed Pipeline 1 for feature extraction, resulting in the final preprocessed dataset. This dataset was then utilized in Pipeline 2, where it was trained using three models: RoBERTa, Qwen, and BERT. Additionally, Pipeline 3 was also designed as a second method to train Qwen2-VL-7B Instruct on both the datasets and images.
- **Phase 4**: Final Model Optimization (end October)
  The outputs from the three models (RoBERTa, Qwen, and BERT) were combined using a Voting Ensemble model to classify memes.
- **Phase 5**: Documentation and Reporting (Starting Nov)
  Compile findings, document methodologies, and prepare a final report and work on the future scope .

VIII. Limitations of the Project

Despite the progress achieved in developing pipelines for hateful meme detection, several limitations were identified that constrained the effectiveness and applicability of the project. These are outlined below:

- **Insufficient Diversity:** The dataset lacked diversity in languages, cultures, and formats, limiting the model's ability to generalize. It only included static image-text memes, excluding mixed media like GIFs, videos, and audio.
- **Handling of Sarcasm and Nuances:** The dataset lacked diversity in languages, cultures, and formats, limiting the model's ability to generalize. It only included static image-text memes, excluding mixed media like GIFs, videos, and audio.

- **Resource-Intensive Processes:** Limited computational resources during the project restricted the ability to train and fine-tune large models, making real-time analysis for high-throughput systems infeasible.
- **Feature Detection and Complex Layouts:** Inaccurate detections by YOLOv8 and Haar Cascade, along with misaligned facial attributes from DeepFace, led to unreliable features. Complex meme layouts with overlapping text and images further hindered feature extraction.
- **Risk of Misclassification:** False positives risk censoring non-hateful content, while false negatives could spread hate, raising ethical and social concerns, especially for sensitive topics.
- **Cultural and Contextual Sensitivity:** The models struggled with cultural and contextual nuances, limiting their effectiveness in global or socio-political contexts.

.

REFERENCES

[1] Abdullakutty, F. and Naseem, U. *Decoding Memes: A Comprehensive Analysis of Late and Early Fusion Models for Explainable Meme Analysis.* School of Computing, Robert Gordon University, Aberdeen, UK and School of Computing, Macquarie University, Sydney, Australia. https://dl.acm.org/doi/pdf/10.1145/3589335.3652504.

[2] Jing Ma, Rong Li, *RoJiNG-CL at EXIST 2024: Leveraging Large Language Models for Multimodal Sexism Detection in Memes.* Notebook for the EXIST Lab at CLEF 2024. University of Zurich, Zurich, Switzerland. https://ceur-ws.org/Vol-3740/paper-100.pdf.

[3] Ji, J., Lin, X., Naseem, U. *CapAlign: Improving Cross Modal Alignment via Informative Captioning for Harmful Meme Detection.* University of Sydney, Sydney, Australia; Shanghai Jiao Tong University, Shanghai, China; Macquarie University, Sydney, Australia. https://dl.acm.org/doi/10.1145/3589334.3648146.

[4] Huang, J., Lyu, H., Pan, J., Wan, Z., Luo, J. (2024) *Evolver: Chain-of-Evolution Prompting to Boost Large Multimodal Models for Hateful Meme Detection.* https://arxiv.org/abs/2407.21004.

[5] Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP features. https://github.com/gokulkarthik/hateclipper

[6] Article for VisualBert. https://medium.com/@raghavr798/visualbert-a-simple-and-performant-baseline-for-vision-and-language-8853de7cb255

[7] Article for Qwen Model. https://medium.com/@farukalamai/qwen2-vl-7b-instruct-a-vision-language-models-vlms-43299b2a196d

[8] Article for YoloV8. https://medium.com/ai-advances/real-time-object-detection-using-yolov8-c8af4f9d206d

[9] Article for Haar Cascade Classifier. https://medium.com/@hansheng0512/haar-cascades-classifier-a-light-weight-face-detection-technique-931b65537a99

[10] Article for DeepFace. https://medium.com/@byte-explorer/deepface-a-library-for-face-recognition-and-facial-analysis-144222eb60bc

[11] https://huggingface.co/google-bert/bert-base-uncased

[12] https://huggingface.co/Qwen/Qwen2.5-Coder-32B-Instruct