




Ashu Sahu

hateful meme

-  Quick Submit
-  Quick Submit
-  Manipal University Jaipur





Document Details

Submission ID**trn:oid:::1:3236209592****Submission Date****May 2, 2025, 2:33 AM GMT+5:30****Download Date****May 2, 2025, 2:38 AM GMT+5:30****File Name****IEEE_mini_report_final.pdf****File Size****1.8 MB****8 Pages****3,611 Words****22,535 Characters**




20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

-  **57 Not Cited or Quoted 19%**
Matches with neither in-text citation nor quotation marks
-  **3 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 15%  Internet sources
- 15%  Publications
- 10%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 57 Not Cited or Quoted 19%**
Matches with neither in-text citation nor quotation marks
- 3 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 15% Internet sources
- 15% Publications
- 10% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	aclanthology.org	2%
2	Internet	link.springer.com	1%
3	Internet	arxiv.org	1%
4	Internet	jestec.taylors.edu.my	1%
5	Internet	researchers.mq.edu.au	<1%
6	Internet	www.coursehero.com	<1%
7	Publication	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P...	<1%
8	Publication	Faseela Abdullakutty, Somaya Al-Maadeed, Usman Naseem. "Chapter 25 Prompt ...	<1%
9	Publication	de Sousa, Lúcia Maria Bessa. "Detecting a Poker Face", Universidade de Aveiro (P...	<1%
10	Student papers	Columbia University	<1%

11	Publication	Souha Nemri, Luc Duong. "Automatic segmentation of echocardiographic images...	<1%
12	Student papers	ICTS	<1%
13	Internet	academic-accelerator.com	<1%
14	Publication	"Advanced Network Technologies and Intelligent Computing", Springer Science a...	<1%
15	Student papers	New College of the Humanities	<1%
16	Student papers	Universitat Politècnica de València	<1%
17	Publication	"Computational Collective Intelligence", Springer Science and Business Media LLC...	<1%
18	Student papers	Vrije Universiteit Amsterdam	<1%
19	Internet	research-management.mq.edu.au	<1%
20	Publication	Hiren Kumar Thakkar, Chintan Bhatt, Victor C.M. Leung, Ilangko Balasingham. "H...	<1%
21	Student papers	Nexford Learning Solutions	<1%
22	Internet	ceur-ws.org	<1%
23	Publication	Xianxun Zhu, Chaopeng Guo, Heyang Feng, Yao Huang, Yichen Feng, Xiangyang ...	<1%
24	Student papers	University of Waikato	<1%

25	Publication	Divya Nimma, Omaia Al-Omari, Rahul Pradhan, Zoirov Ulmas, R.V.V. Krishna, Ts. Y...	<1%
26	Publication	Kumar, Yaman. "Behavior as a Modality.", State University of New York at Buffalo	<1%
27	Publication	Namit Khanduja, Nishant Kumar, Arun Chauhan. "Telugu Language Hate Speech ...	<1%
28	Internet	ebin.pub	<1%
29	Internet	www.mdpi.com	<1%
30	Publication	Luiz Guilherme Kasputis Zanini, Izabel Regina Fischer Rubira-Bullen, Fátima de Lo...	<1%
31	Publication	Sozo Inoue, Guillaume Lopez, Tahera Hossain, Md Atiqur Rahman Ahad. "Activity,...	<1%
32	Internet	web.archive.org	<1%
33	Internet	dokumen.pub	<1%
34	Internet	fatcat.wiki	<1%
35	Internet	www.biorxiv.org	<1%
36	Internet	backend.orbit.dtu.dk	<1%
37	Internet	research-repository.st-andrews.ac.uk	<1%
38	Internet	speakerdeck.com	<1%

39

Internet

www.nature.com

<1%

40

Publication

Tasneem Ahmed, Shrish Bajpai, Mohammad Faisal, Suman Lata Tripathi. "Advanc...

<1%

Multi-Lingual Hateful Meme Detection

1st Ankit Kumar*, 2nd Disha Soni†, 3rd Shreya Sinha‡, 4th Avantika Soni§, 5th Ramkrishan¶

*†‡§¶ 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, ¶iib2022035@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 understanding memes by combining multimodal feature extraction techniques. We preprocess meme images through rescaling, Gaussian blurring, and deskewing to enhance input quality. Textual features are extracted and cleaned by removing stopwords, links, code-mixed and code-switched text, and irrelevant tokens like "XD." Word embeddings are generated using FastText, Word2Vec, and GloVe models. Visual features are obtained through object detection and demographic analysis, where OpenCV's DeepFace model is employed to extract facial attributes such as age, gender, emotion, and ethnicity. These textual and visual features are then fused to create a unified feature embedding. Our approach, leveraging these enriched multimodal embeddings, enhances meme understanding and offers a robust solution for various downstream applications such as meme classification and sentiment analysis.

I. INTRODUCTION

A meme contains direct or indirect attacks on people based on characteristics such as ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define an attack as violent or dehumanizing speech (for example, comparing people to non-human entities like animals), statements of inferiority, and calls for exclusion or segregation. Memes that mock hate crimes are also considered hateful memes.

Hateful content detection, especially in multimodal formats like memes, has gained significant attention in recent years due to the growing 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 particularly complex. Additionally, memes are often created in multiple languages, highlighting the need for a single, unified system capable of identifying hateful content across different linguistic contexts.

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 PREPARATION

For this project, we merged datasets from multiple platforms to create a comprehensive multimodal dataset. The datasets used include the MET-Meme Dataset, which spans across two languages and contains 10,045 text-image pairs with manual annotations, including 6,045 Chinese images and 4,000 English images; the CM-Offensive Meme Dataset, comprising 4,372 Hindi-English offensive memes; and the Facebook Hateful Meme Dataset, which consists of 10,000 multimodal examples specifically designed for hateful content detection. 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. After concatenating all datasets and integrating additional features, the final dataset comprises a total of 26,432 images. Furthermore, facial features were extracted from 8,724 images to enrich the dataset for deeper multimodal analysis.

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) proposes 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*, introduces 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 MuRIL, BERT, XLM-R, and ViT 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.

In addition, we extract facial features using DeepFace and apply Visual Question Answering (VQA) through the BLIP model to capture deeper semantic understanding of the images. We combine the facial analysis results and VQA outputs to create enriched visual representations. These are then fine-tuned using a range of visual models, including ResNet50, DenseNet21, Xception, VGG19, and VGG16. For textual processing, we explore machine learning classifiers such as Logistic Regression (LR), Support Vector Machines (SVM), and Multinomial Naive Bayes (MNB), alongside transformer-based models like MuRIL, BERT, and XLM-R.

Finally, we employ ensemble learning — specifically a voting ensemble — to integrate the outputs from both visual and textual pipelines, leading to stronger and more robust predictions. Unlike previous work, which often relied on pre-trained models without custom datasets, we develop a tailored dataset for better model fine-tuning, achieving improved performance. Our approach balances computational efficiency with robust feature extraction, outperforming traditional methods by offer-

ing a more integrated and scalable solution for hateful meme detection.

IV. THE PROPOSED METHOD

In this section we will introduce our proposed methods for hateful meme detection as 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 multiple models for inference and ensemble predictions. Figure 1 illustrates the system.

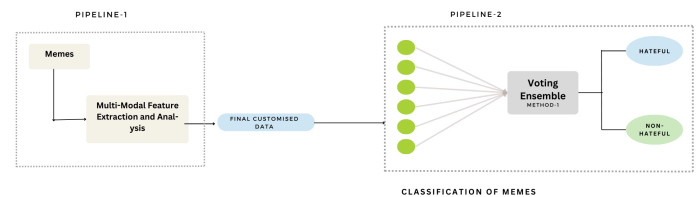


Fig. 1. Architecture of our complete model

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

This pipeline integrates multi-modal feature extraction, combining textual and visual information for comprehensive meme analysis. It utilizes advanced models for text extraction, language normalization, object detection, facial attribute analysis, and visual question answering.

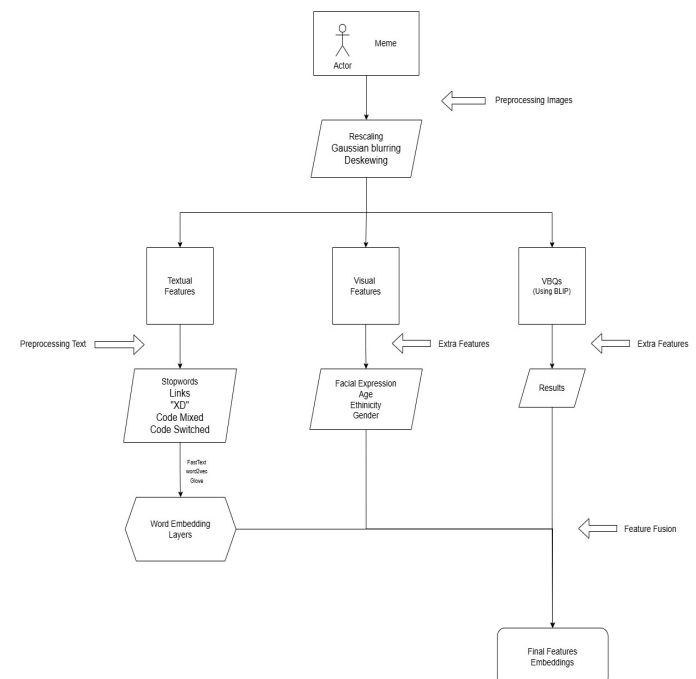


Fig. 2. Multi-Modal Feature Extraction and Analysis

1) Text Processing and Normalization:

- **Text Extraction:** We extract text embedded within memes using optical character recognition techniques. This process captures all textual elements regardless of font, style, or placement within the image.
- **Language Normalization:** All extracted text is converted to standard English to ensure consistent processing. This step handles slang, code-switching, and non-English content, creating uniformity for downstream analysis.
- **Text Preprocessing:** We implement standard NLP pre-processing including removal of stopwords, normalization of links, handling of special characters ("XD"), and processing of code-mixed and code-switched content for improved analysis.

2) Image Feature Extraction:

- **Image Preprocessing:** Images undergo rescaling, Gaussian blurring, and deskewing for standardization before feature extraction, ensuring consistent input quality for all models.
- **Facial Analysis:** We employ **DeepFace** for comprehensive facial attribute extraction from meme images. This tool analyzes detected faces for demographic information (age, gender, ethnicity) and emotional expressions, providing valuable context for hate speech identification. DeepFace leverages pre-trained models (VGG-Face, Google FaceNet, OpenFace) to recognize facial features and predict attributes based on facial landmarks.

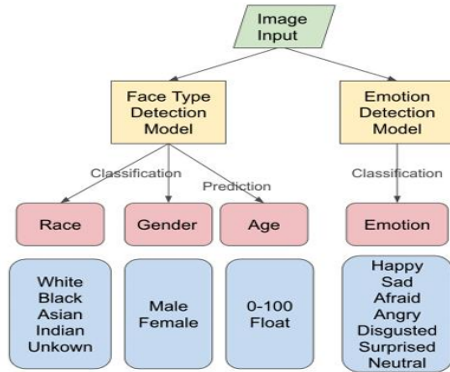


Fig. 3. An illustration of model input and output for facial analysis

- **Visual Question Answering:** The **BLIP** model is employed to generate contextual understanding of visual content through targeted question-answering about the image. By formulating strategic questions about the content and interpreting model responses, we capture nuanced visual information that might indicate harmful content.
- **Multi-modal Integration:**
 - **Feature Fusion:** We combine word embeddings (using FastText, Word2Vec, and GloVe) with visual features from object detection, facial analysis, and visual question answering.

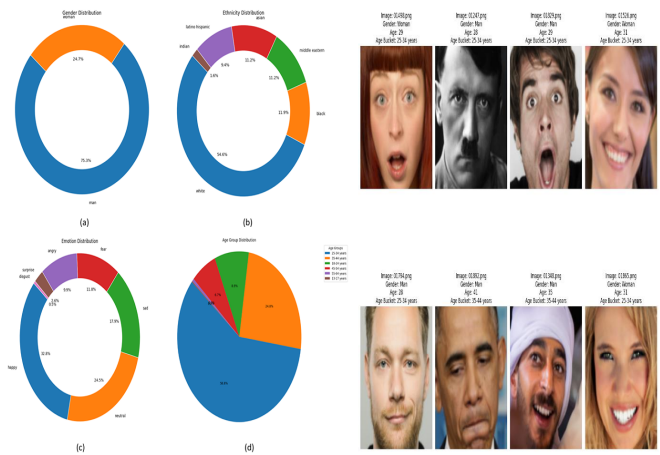


Fig. 4. Illustration of (a) Gender Distribution (b) Ethnicity Distribution (c) Emotion Distribution (d) Age Group Distribution across the dataset.

- **VQA and Facial Feature Integration:** Results from the BLIP visual question answering model are merged with facial attribute data to create a rich representation of visual content, particularly focusing on potentially harmful representations of demographics.
- **Consolidated Feature Set:** The final feature embeddings incorporate textual, visual, demographic, emotional, and contextual information to provide a comprehensive basis for classification.

C. Pipeline-2: Advanced Model Training and Ensemble Classification

Pipeline 2 leverages the rich feature set created in Pipeline 1 to train multiple specialized models that are then combined through ensemble learning techniques.

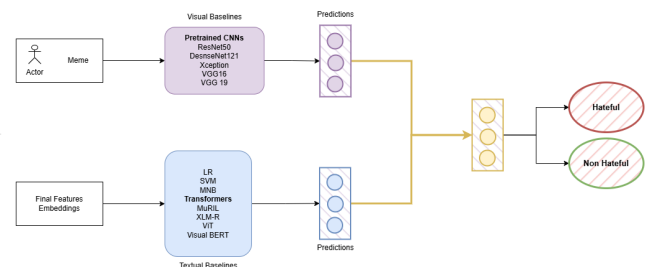


Fig. 5. Pipeline representing voting ensemble method for classification

1) Visual Baselines:

- **Pretrained CNNs:** We utilize ResNet50, DenseNet121, Xception, VGG19, and VGG16 models to extract deep visual features from the meme images. These networks, pretrained on large-scale image datasets, provide robust visual representations that capture different aspects of image content.

2) Textual Baselines:

- **Traditional ML Models:** Logistic Regression (LR), Support Vector Machines (SVM), and Multinomial Naive Bayes (MNB) classifiers process structured text features.

- **Transformer Models:** BERT, XLM-R (cross-lingual version of Robustly Optimized BERT), ViT (Vision Transformer), MuRIL (for multilingual representation of Indian languages), and Visual BERT handle complex textual and multimodal inputs.

3) Ensemble Learning:

- **Model Fine-tuning:** We fine-tune all models on our custom dataset derived from Pipeline 1, optimizing them specifically for hateful meme detection.
- **Voting Ensemble:** The final classification is determined through a voting ensemble mechanism that combines predictions from all models, leveraging the strengths of each architecture.
- **Decision Fusion:** This approach ensures robust classification by aggregating predictions across different model types, reducing error through collective intelligence.

4) Multimodal Fusion:

The system performs fusion of visual and textual predictions through:

- Integration of predictions from visual-based models with those from text-based models
- Weighted combination based on confidence scores and model reliability
- Final decision making through majority voting among ensemble components

The models are fine-tuned using the Final Pre-processed 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.

5) **ResNet50-BERT Multimodal Model:** ResNet50-BERT was implemented using PyTorch and the Hugging Face Transformers library to process both image and text data simultaneously.

a) **Model Architecture:** The model architecture combines ResNet50 for image feature extraction and BERT for text processing:

- **Image Processing:** Pretrained ResNet50 model with the final classification layer removed to extract 2048-dimensional image features.
- **Text Processing:** Multilingual BERT (bert-base-multilingual-cased) for handling text data with demographic features and translated text.
- **Feature Fusion:** Concatenation of image features (2048-dim) and BERT CLS token embeddings (768-dim).
- **Classification Head:** Multi-layer network with dropout for regularization:
 - First layer: 2816 → 512 units with ReLU activation
 - Second layer: 512 → 256 units with ReLU activation
 - Final layer: 256 → 2 units for binary classification

b) **Training Process:** The model was trained with the following hyperparameters:

- **Optimizer:** AdamW with a learning rate of 2×10^{-5}

- **Loss Function:** Cross-Entropy Loss
- **Batch Size:** 16 samples per batch
- **Regularization:** Dropout rates of 0.3 and 0.2 in the classification head
- **Learning Rate Scheduler:** ReduceLROnPlateau with factor=0.5, patience=2
- **Training Strategy:**
 - Froze most ResNet50 and BERT weights for efficiency
 - Only fine-tuned the last two layers of both models
 - Applied image augmentation with resizing, normalization, and random transformations

c) **Results and Evaluation:** Results are shown below:

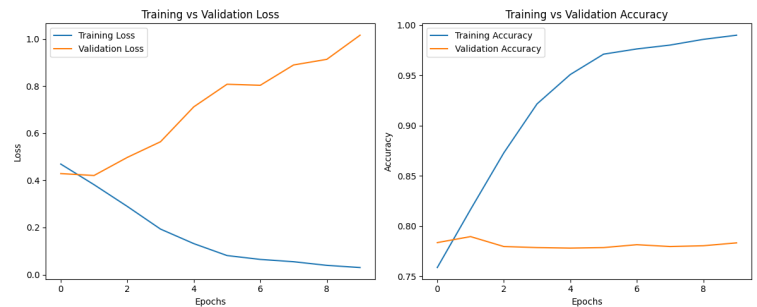


Fig. 6. Training vs. Validation Loss (left) and Accuracy (right) for the ResNet50-BERT model

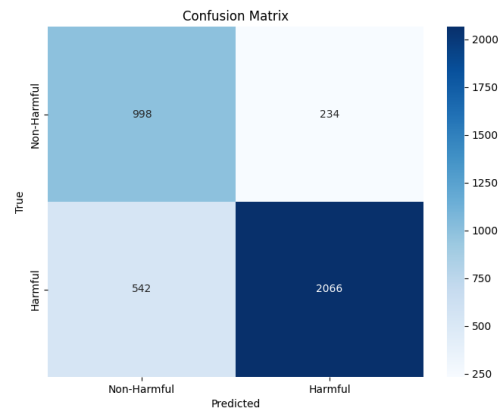
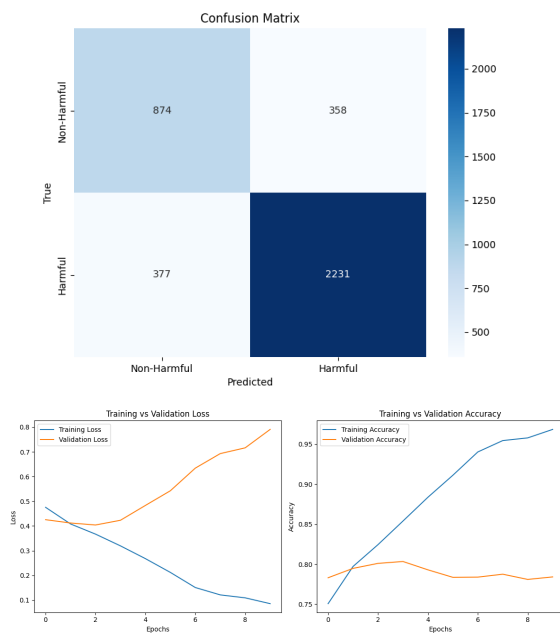


Fig. 7. Confusion Matrix for the ResNet50-BERT model showing 998 true negatives, 2066 true positives, 234 false positives, and 542 false negatives

TABLE I
RESNET50-BERT MODEL EVALUATION METRICS

Metric	Value
Accuracy	79.79%
F1 Score (Harmful)	0.84
F1 Score (Non-Harmful)	0.72
Precision (Harmful)	0.90
Precision (Non-Harmful)	0.65
Recall (Harmful)	0.79
Recall (Non-Harmful)	0.81
Macro Avg F1	0.78
Weighted Avg F1	0.80

6) **DenseNet121 + BERT Classifier**: The model uses a multimodal approach, combining DenseNet121 for image features (1024 dimensions) and multilingual BERT for text features (768 dimensions). Features are fused through direct concatenation and passed through fully-connected layers (512 → 256 → 2). Transfer learning is applied with partial model freezing, keeping only the final layers trainable.



Test Accuracy: 0.8086

Confusion Matrix:

```
[[ 874  358]
 [ 377 2231]]
```

Classification Report:

	precision	recall	f1-score	support
Non-Harmful	0.70	0.71	0.70	1232
Harmful	0.86	0.86	0.86	2608
accuracy			0.81	3840
macro avg	0.78	0.78	0.78	3840
weighted avg	0.81	0.81	0.81	3840

Fig. 8. Visualization of DenseNet121 model performance metrics

Results:

Metric	Value
Test Accuracy	80.86%
Precision (Harmful)	0.86
Recall (Harmful)	0.86
F1 Score (Harmful)	0.86
Precision (Non-harmful)	0.70
Recall (Non-harmful)	0.71
F1 Score (Non-harmful)	0.70

TABLE II

KEY PERFORMANCE METRICS FOR DENSENET121 + BERT

7) Bidirectional Long Short-Term Memory (BiLSTM):

A Bidirectional Long Short-Term Memory (BiLSTM) model was developed for binary classification of hateful and non-hateful content. By processing sequences in both forward and backward directions, BiLSTM effectively captures contextual dependencies in multilingual textual data.

The model was implemented using TensorFlow and Keras frameworks, focusing on clean architecture, regularization, and early stopping to optimize generalization.

a) **Training Process**: The BiLSTM model was trained with the following settings:

- **Tokenizer**: Vocabulary size of 10,000 words with OOV token handling; padding applied to the maximum sequence length.
- **Normalization**: The age feature was standardized using StandardScaler.
- **Training Configuration**:
 - Optimizer: Adam (default learning rate).
 - Batch size: 32.
 - Validation split: 10%.
 - Early stopping with 3-patience epochs and model checkpointing based on validation loss.

b) **Model Architecture**: The architecture consisted of:

- **Embedding Layer**:
 - 10,000 vocabulary size, 64-dimensional embeddings.
- **First BiLSTM Layer**:
 - 64 units with return sequences enabled.
- **Dropout Layer**:
 - 0.5 rate after first BiLSTM.
- **Second BiLSTM Layer**:
 - 64 units without return sequences.
- **Dense Layer**:
 - 128 units, ReLU activation.
- **Dropout Layer**:
 - 0.5 rate for regularization.
- **Output Layer**:
 - 1 unit, sigmoid activation for binary classification.
- **Loss Function**:
 - Binary cross-entropy.

c) **Results**:

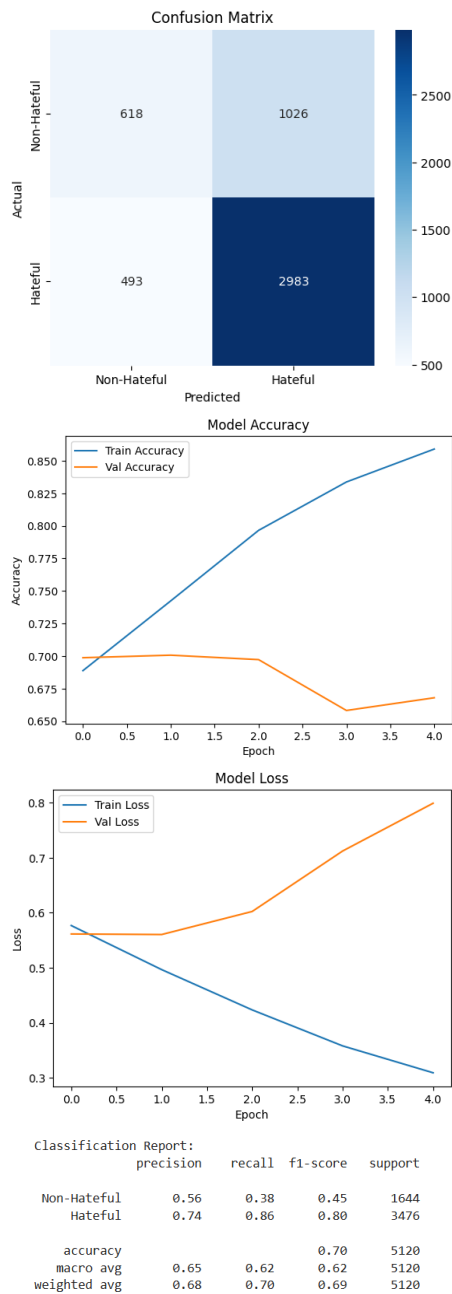


Fig. 9. Visualization of BiLSTM model performance metrics

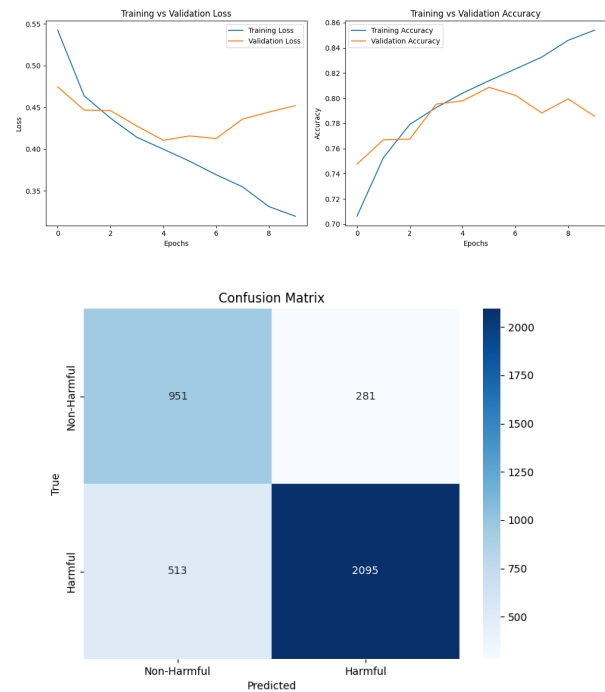


Fig. 10. Visualization of MuRIL multimodal model performance metrics

Results:

TABLE III
CLASSIFICATION REPORT FOR MURIL MULTIMODAL MODEL

Class	Precision	Recall	F1-Score	Support
Non-Harmful	0.65	0.77	0.71	1232
Harmful	0.88	0.80	0.84	2608
Accuracy	0.79			3840
Macro Avg	0.77	0.79	0.77	3840
Weighted Avg	0.81	0.79	0.80	3840

9) **XLM-R Text Classifier:** XLM-RoBERTa (XLM-R) is a multilingual transformer-based model pretrained on 100 languages. This implementation uses XLM-R with an enhanced classification head specifically designed for multilingual harmful content detection across diverse text inputs.

8) **MuRIL Multimodal Classifier:** MuRIL (Multilingual Representations for Indian Languages) is a BERT-based model pre-trained on 17 Indian languages and English. This implementation combines MuRIL with a custom CNN for image processing, creating a multimodal classifier specifically designed for multilingual meme analysis and harmful content detection.



Fig. 11. Visualization of XLM-R text model performance metrics

Results:

TABLE IV
CLASSIFICATION REPORT FOR XLM-R TEXT MODEL

Class	Precision	Recall	F1-Score	Support
Non-Harmful	0.70	0.68	0.69	1232
Harmful	0.85	0.86	0.86	2608
Accuracy	0.8042			3840
Macro Avg	0.78	0.77	0.77	3840
Weighted Avg	0.80	0.80	0.80	3840

10) **ViT-BERT Multimodal Classifier:** This implementation combines a Vision Transformer (ViT) for image processing with a multilingual BERT model for text analysis, creating a robust multimodal classifier for multilingual harmful content detection across both visual and textual elements.

Results:

TABLE V
CLASSIFICATION REPORT FOR ViT-BERT MULTIMODAL MODEL

Class	Precision	Recall	F1-Score	Support
Non-Harmful	0.71	0.71	0.71	1232
Harmful	0.86	0.86	0.86	2608
Accuracy	0.81			3840
Macro Avg	0.78	0.78	0.78	3840
Weighted Avg	0.81	0.81	0.81	3840

11) **Voting Ensemble:** Using the outputs of our six models (XLM-R, MuRIL, BiLSTM, ViT-BERT, ResNet50, and DenseNet121), we apply a majority 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.

```
===== ENSEMBLE MODEL EVALUATION =====
```

	precision	recall	f1-score	support
Non-Harmful	0.77	0.83	0.80	2459
Harmful	0.92	0.89	0.90	5221
accuracy			0.87	7680
macro avg	0.84	0.86	0.85	7680
weighted avg	0.87	0.87	0.87	7680

Fig. 12. Performance metrics of Voting Ensemble

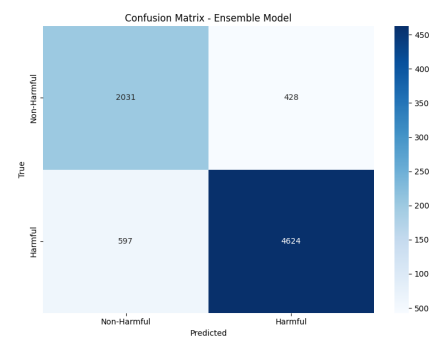


Fig. 13. Confusion Matrix of Voting Ensemble

V. RESULT AND ANALYSIS

The evaluation of various models for hateful meme detection was conducted using a diverse set of architectures. ViT-Bert model outperformed others significantly, achieving an accuracy of 81.20%, an F1 score of 0.81, and precision of 0.81.

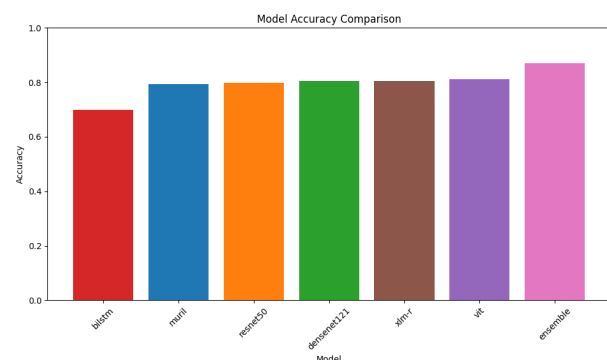


Fig. 14. Comparison of different models

TABLE VI
FINAL MODEL PERFORMANCE METRICS

Model	Accuracy	F1 Score	Precision
BiLSTM	70.01%	0.69	0.68
XLM-R	80.42%	0.80	0.80
ViT-BERT	81.20%	0.81	0.81
MuRIL	79.32%	0.80	0.81
ResNet50	79.79%	0.80	0.82
DenseNet121	80.42%	0.81	0.81
Voting Ensemble	87.00%	0.85	0.86

VI. CONCLUSION AND FUTURE WORK

The proposed system effectively combines diverse architectures—XLM-R and MuRIL for multilingual textual features, ViT for visual understanding, BiLSTM for sequential text patterns, and ResNet50 and DenseNet121 for deep image feature extraction. By aggregating these complementary predictions through majority voting ensemble, the model achieves robust and reliable hateful meme detection across multilingual and multimodal inputs.

Future improvements could focus on replacing majority voting with a trainable ensemble method such as stacking to capture inter-model relationships more effectively. Additionally, expanding the dataset with more underrepresented languages and subtle hate expressions, and exploring lightweight transformer architectures for faster inference, would further enhance the system's scalability and performance.

VII. LIMITATIONS OF THE PROJECT

Although the proposed system combining XLM-R, MuRIL, BiLSTM, ViT, ResNet50, and DenseNet121 with majority voting achieved promising results, several limitations were observed:

- **Text Extraction from Images:** Errors in OCR (Optical Character Recognition) during text extraction can negatively impact the performance of downstream text-based models.
- **Multilingual Text Translation:** Translating non-English text to English may introduce semantic inaccuracies, affecting the quality of text feature representations.
- **Computational Complexity:** Deploying multiple large-scale models increases memory and processing requirements, making real-time applications challenging.
- **Imbalanced Dataset:** Class imbalance in the dataset may bias the models toward the majority class, reducing sensitivity to hateful content.
- **Interpretability Challenges:** The ensemble provides final predictions without offering transparent explanations, limiting user trust in sensitive scenarios.
- **Simple Majority Voting:** Equal weighting across models in majority voting does not account for individual model confidence or reliability, which may reduce ensemble efficiency.

REFERENCES

- [1] Abdullakutty, F. and Naseem, U., *Decoding Memes: A Comprehensive Analysis of Late and Early Fusion Models for Explainable Meme Analysis*. Robert Gordon University, Aberdeen, UK and 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*. 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, Shanghai Jiao Tong University, and Macquarie University. <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] Gokul Karthik, *Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP features*, <https://github.com/gokulkarthik/hateclipper>
- [6] Li, L., et al., *VisualBERT: A Simple and Performant Baseline for Vision and Language*. <https://medium.com/@raghavr798/visualbert-a-simple-and-performant-baseline-for-vision-and-language-8853de7cb255>
- [7] Conneau, A., et al., *Unsupervised Cross-lingual Representation Learning at Scale*. (XLM-R model paper). <https://arxiv.org/abs/1911.02116>
- [8] Dosovitskiy, A., et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. (Vision Transformer - ViT) <https://arxiv.org/abs/2010.11929>
- [9] Kakwani, D., et al., *MuRIL: Multilingual Representations for Indian Languages*. <https://arxiv.org/abs/2103.10730>
- [10] Schuster, M., Paliwal, K.K., *Bidirectional Recurrent Neural Networks*. (BiLSTM introduction) <https://ieeexplore.ieee.org/document/650093>
- [11] He, K., Zhang, X., Ren, S., Sun, J., *Deep Residual Learning for Image Recognition*. (ResNet50 paper) <https://arxiv.org/abs/1512.03385>
- [12] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., *Densely Connected Convolutional Networks*. (DenseNet paper) <https://arxiv.org/abs/1608.06993>
- [13] Faruk Alamai, *Qwen2-VL-7B-Instruct: A Vision Language Model*, <https://medium.com/@farukalamai/qwen2-vl-7b-instruct-a-vision-language-models-vlms-43299b2a196d>
- [14] Real-time Object Detection using YOLOv8, <https://medium.com/ai-advances/real-time-object-detection-using-yolov8-c8af4f9d206d>
- [15] Hansheng, *Haar Cascades Classifier: A Light-weight Face Detection Technique*. <https://medium.com/@hansheng0512/haar-cascades-classifier-a-light-weight-face-detection-technique-931b65537a99>
- [16] Byte Explorer, *DeepFace: A Library for Face Recognition and Facial Analysis*. <https://medium.com/@byte-explorer/deepface-a-library-for-face-recognition-and-facial-analysis-144222eb60bc>