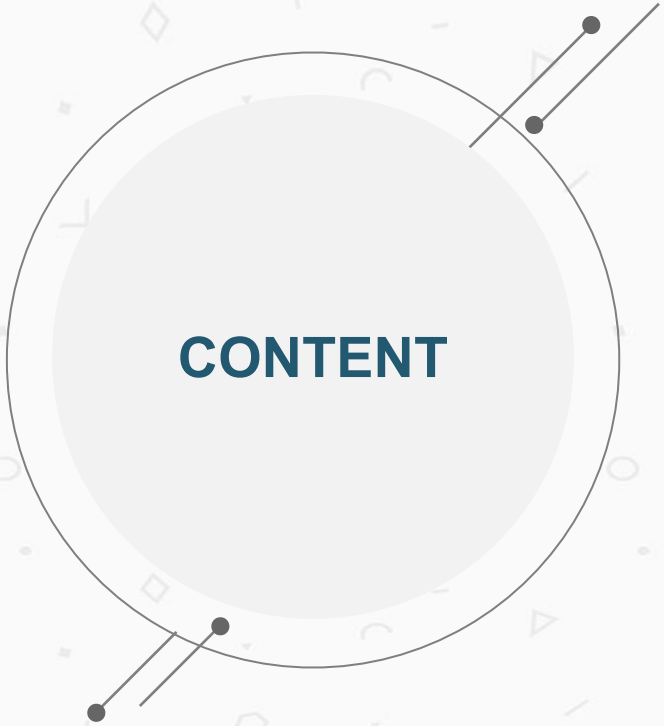
Hateful Meme Detection

Under Supervision of **Prof Vrijendra Singh**

Presented by :
Disha Soni (IIT2022260)
Ankit Kumar (IIT2022256)
Ram Krishan (IIB2022035)
Shreya Sinha (IIB2022034)
Avantika Soni (IIB2022045)



01

Introduction & Application

Introduction

Hateful Meme Detection

Definition of the hateful meme

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.

Introduction

Hateful Meme

Non-Hateful Meme

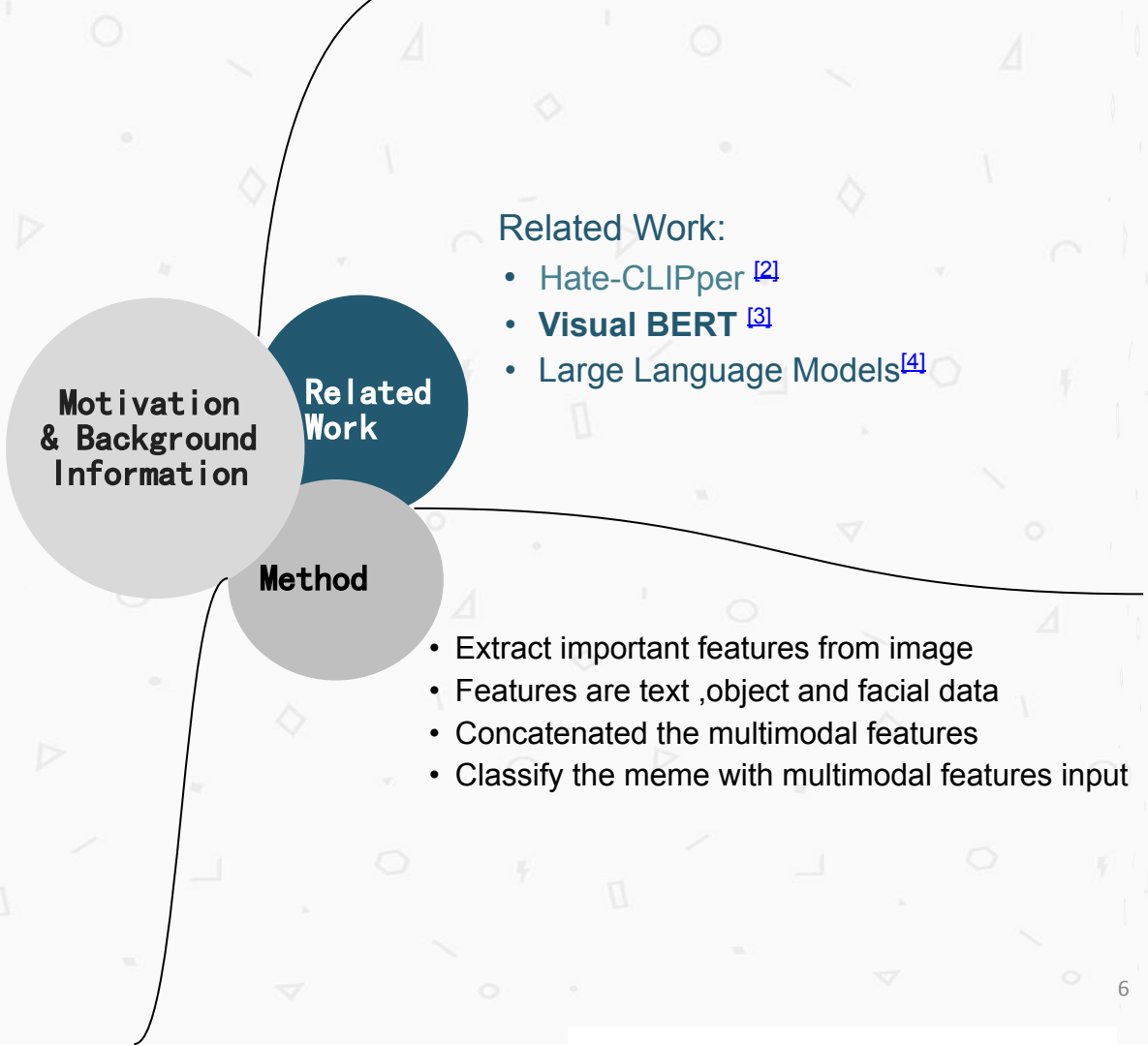
Introduction

Motivation:

- Online hate speech is increasing, especially through memes.
- Memes [1] combine images and text, making detection difficult.
- The project aims to develop an AI system to detect hateful memes.

Background Information:

- Memes are widely used on social media, often blending humor with text and images.
- They are sometimes used to spread hate speech, challenging traditional detection methods.
- Detecting hate in memes requires multi-modal analysis of both visual and textual elements.



Related Work:

- Hate-CLIPper [2]
- **Visual BERT** [3]
- Large Language Models[4]

- Extract important features from image
- Features are text ,object and facial data
- Concatenated the multimodal features
- Classify the meme with multimodal features input

Application

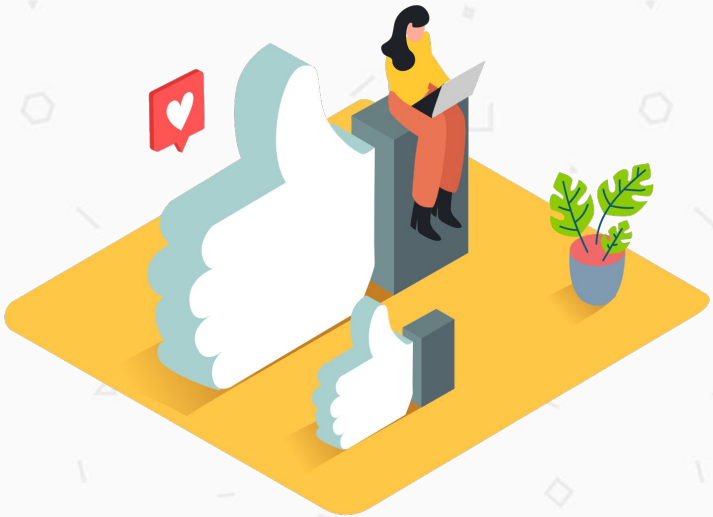
Hateful Meme Detection

Application of the hateful meme detection

Content Moderation: Automatically identify and remove hateful memes on social media platforms.
Law Enforcement: Assist authorities in monitoring and addressing online hate speech.
Anti-bullying Measures: Detect harmful content to prevent online harassment and cyberbullying.
Hate Speech Analysis: Provide insights for researchers studying trends in online hate speech.
User Reporting Systems: Enhance user-flagging mechanisms with automated meme analysis.

02

Related Work



Related Work

- **Paper 1 :** Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text

  Authors: Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Paul Buitelaar Conference: LREC 2020.

- **Paper 2 :** MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes

  Authors: Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque Conference: LREC 2022

- **Paper 3 :** MET-Meme: A Multimodal Meme Dataset Rich in Metaphors

  Authors: Tingting Li, Junzhe Zheng, Bo Xu, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, Feng Xia SIGIR ’22, July 11–15, 2022, Madrid, Spain.

- **Paper 4 :** The Hateful Memes Challenge: Detecting Hate Speech in Multi-Modal Memes

  Authors: Narine Marutyan, Alissa Jouljian,American University of Armenia May 9, 2024

- **Paper 5 :** Mapping Memes to Words for Multimodal Hateful Meme Classification

  Authors: Lorenzo Agnolucci, Alberto Baldrati, Giovanni Burbi, Marco Bertini, Alberto Del Bimbo Conference:ICCVW 2023

Dataset Overview:

Merged Multiple Datasets to create a comprehensive multimodal dataset.

Datasets Used:

- **MET-Meme Dataset:**
  - 10,045 text-image pairs.
  - Includes 6,045 Chinese and 4,000 English images.
  - Manually annotated.
- **CM-Offensive Meme Dataset:**
  - 4,372 memes.
  - Contains Hindi-English offensive content.
- **Facebook Hateful Meme Dataset:**
  - 10,000 examples.
  - Designed specifically for hateful content detection.

Content Diversity:

- Each example includes a meme image + associated text.
- Covers a wide variety of offensive content targeting race, gender, ethnicity, and other personal traits.

Final Merged Dataset:

- Total of **26,432 images** after concatenation and feature integration.

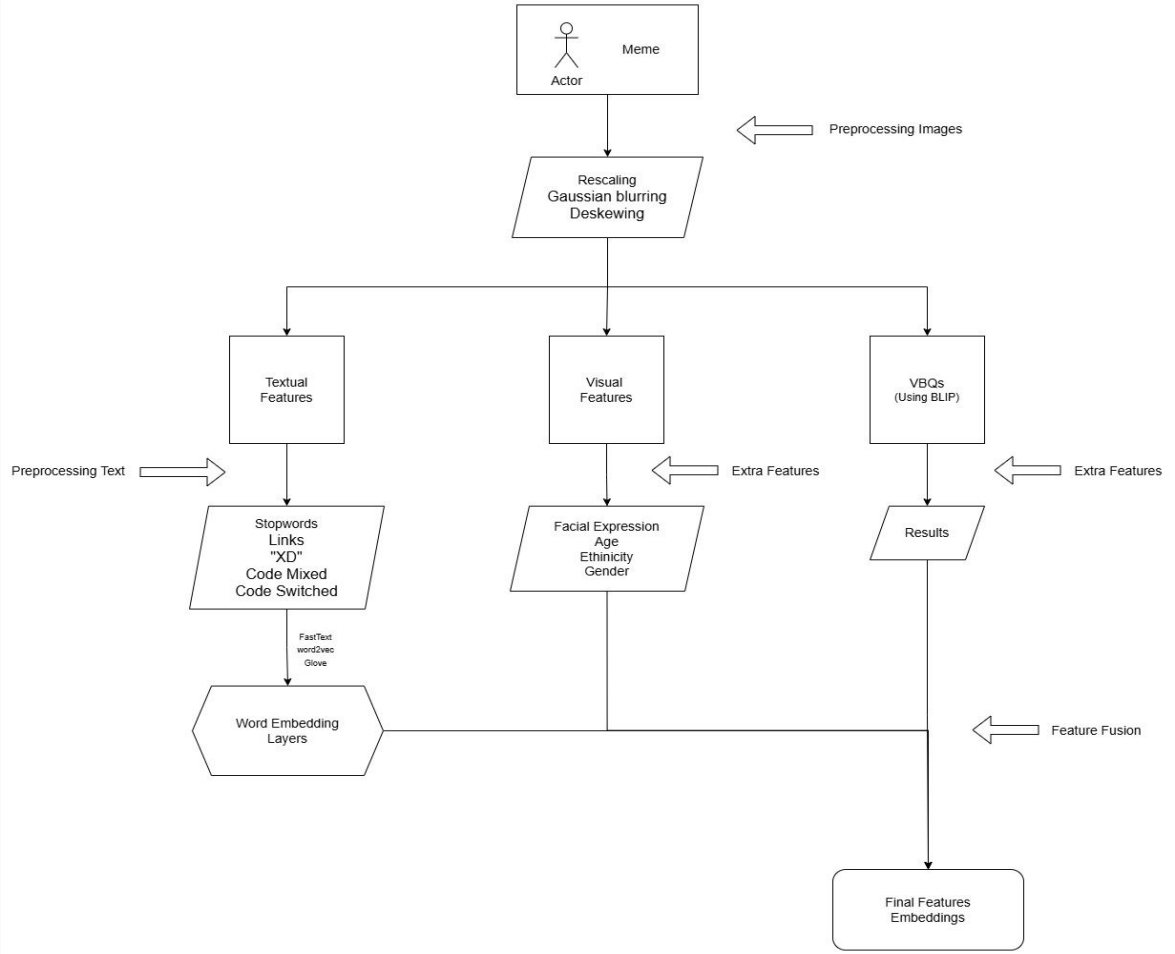
Facial Feature Extraction:

- Extracted from **8,724 images** for enhanced multimodal analysis.

03

Proposed Methodology

Pipeline - 01 : Multi-Modal Feature Extraction and Analysis

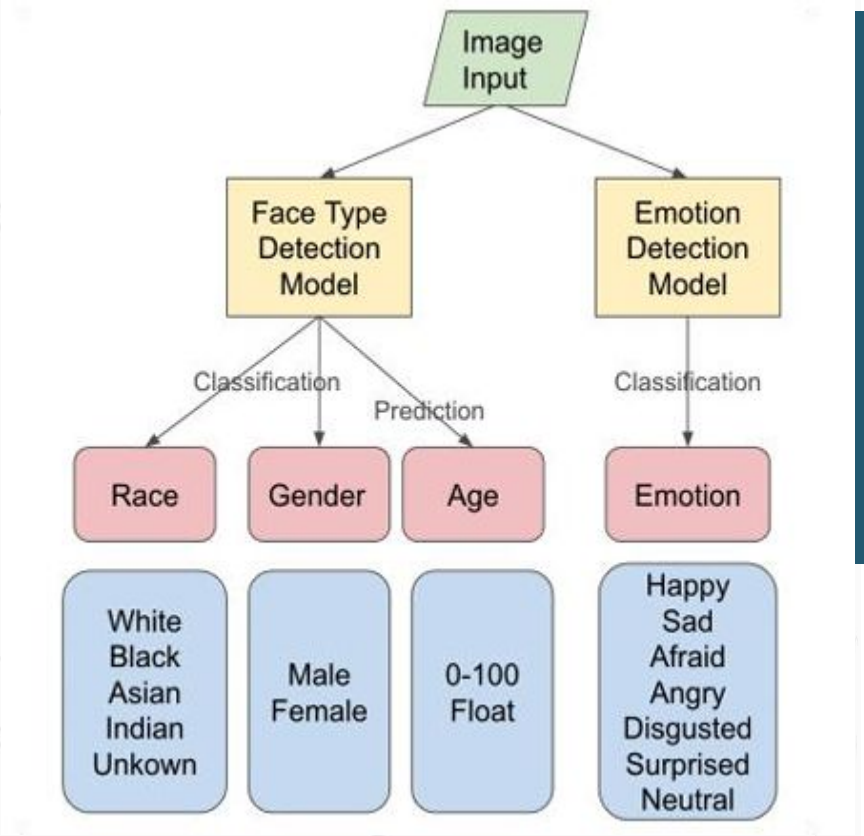


Text Processing and Normalization

1. **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.
2. **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.
3. **Text Preprocessing :** We implement standard NLP preprocessing including removal of stopwords, normalization of links, handling of special characters (”XD”), and processing of code-mixed and code-switched content for improved analysis.

Image Feature Extraction

1. **Image Preprocessing :** Images undergo rescaling, Gaussian blurring, and deskewing for standardization before feature extraction, ensuring consistent input quality for all models.
2. **Facial Analysis :** We use **DeepFace** to extract facial attributes like age, gender, ethnicity, and emotion, aiding hate speech detection. It adds valuable demographic and emotional context to detected faces.
3. **Visual Question Answering (VQA) :** The **BLIP** model is employed for visual question-answering to extract contextual cues from images, helping identify subtle indicators of harmful content.



Tag Extraction — Face Type Detection & Emotion detection Model

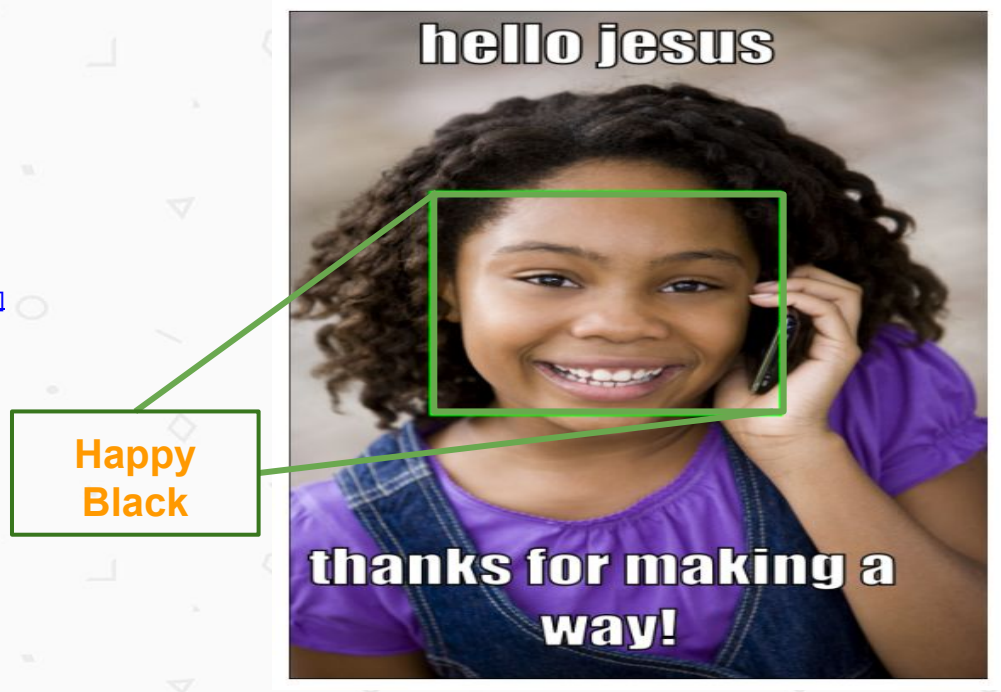
This model aims to tackle the task of predicting the **age** and classifying the **gender, race and emotion** from facial images.

Pretrained model: **DeepFace**[14]

- Backbone: VGG-16[15] and Resnet50[16]

Pretrained dataset:

- Fer2013 dataset
  https://www.kaggle.com/msambare/fer2013
- UTK face dataset
  https://susanqq.github.io/UTKFace/

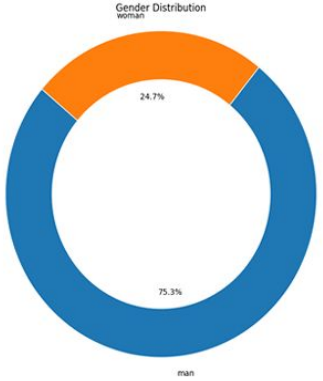


DeepFace Overview and Working

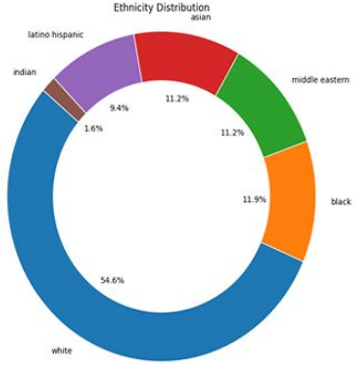
A deep learning facial recognition framework developed by Facebook AI Research (FAIR) that utilizes a convolutional neural network (CNN) to recognize and verify faces in images.

1. **Input Processing:**
   - Input images are pre-processed, including face detection and alignment, to ensure consistent facial orientation and scale.
2. **Feature Extraction:**
   - Uses a CNN architecture (e.g., VGG-Face) to extract facial features from aligned images, transforming them into high-dimensional feature vectors.
3. **Embedding Generation:**
   - Each face image is represented as a fixed-size vector in an embedding space, where similar faces are mapped closer together.
4. **Face Recognition:**
   - Compares the embedding vectors using distance metrics (e.g., Euclidean distance) to determine the similarity between faces, enabling recognition and verification.
5. **Emotion Detection:**
   - Can also predict emotions by using additional classifiers trained on emotion datasets, analyzing facial expressions.

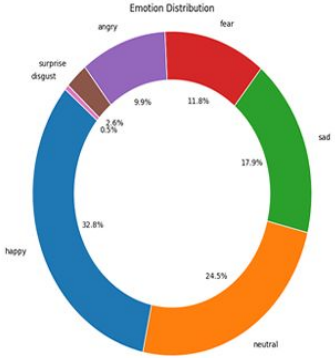
Illustration of (a) Gender Distribution (b) Ethnicity Distribution (c) Emotion Distribution (d) Age Group Distribution across the dataset.



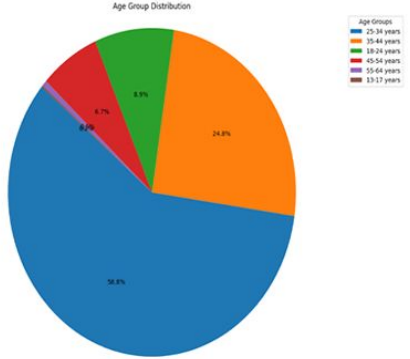
(a)



(b)



(c)

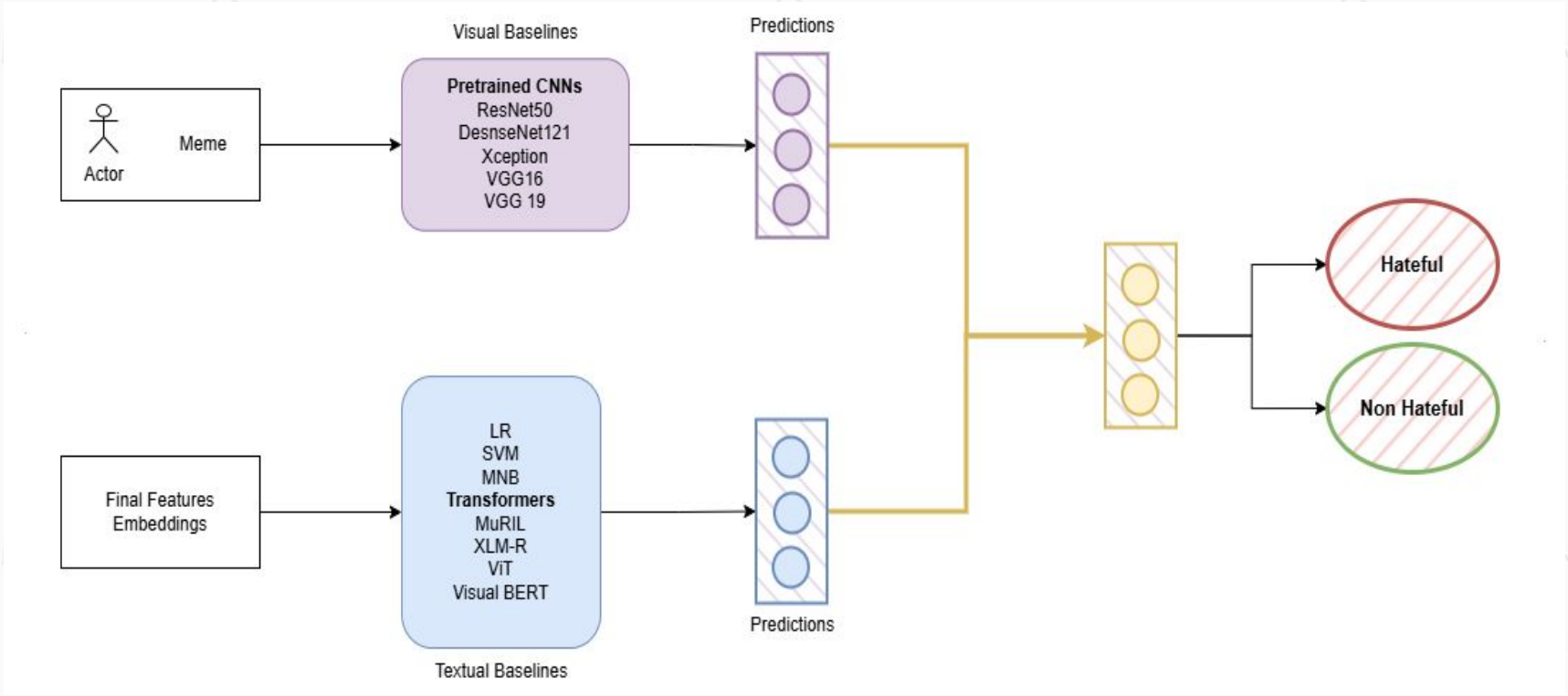


(d)

Multi-modal Integration

1. **Feature Fusion :** We combine word embeddings (**using FastText, Word2Vec, and GloVe**) with visual features from object detection, facial analysis, and visual question answering.
2. **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.
3. **Consolidated Feature Set :** The final feature embeddings incorporate textual, visual, demographic, emotional, and contextual information to provide a comprehensive basis for classification.

Pipeline - 02 : Advanced Model Training and Ensemble Classification



Visual Baselines :

We utilized 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.

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.

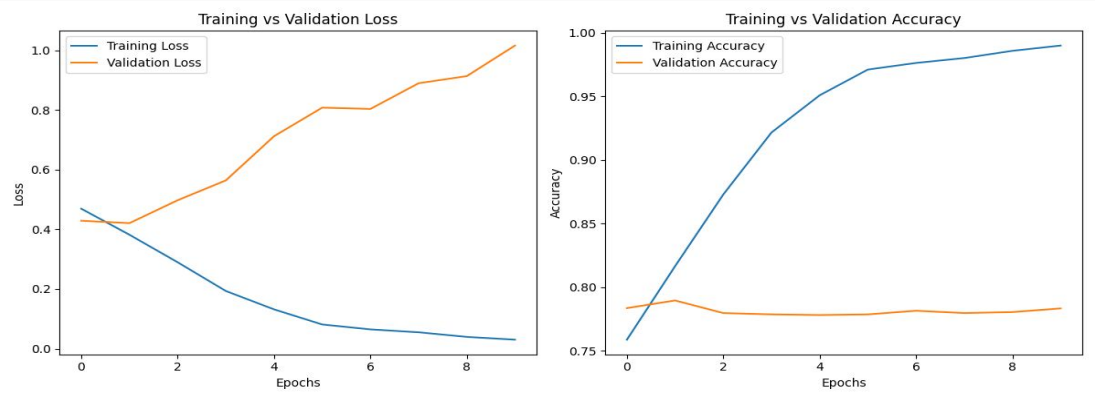
Ensemble Learning & Multimodal Fusion :

- **Model Fine-tuning**: All models are fine-tuned on the custom dataset (from Pipeline 1) for hateful meme detection.
- **Voting Ensemble**: Final classification is based on a voting mechanism combining predictions from multiple models.
- **Decision Fusion**: Enhances accuracy by aggregating outputs from different model types to reduce errors.
- **Multimodal Fusion**:
    - Combines predictions from visual and textual models.
    - Uses weighted confidence scores and model reliability.
    - Final decision via majority voting across ensemble components.
- **Input Features**: Models use features like dominant emotion, demographics, and sentiment from the pre-processed dataset.
- **Independent Predictions**: Each model outputs whether a meme is hateful or non-hateful.

04

Models Used

ResNet50-BERT Multimodal Model

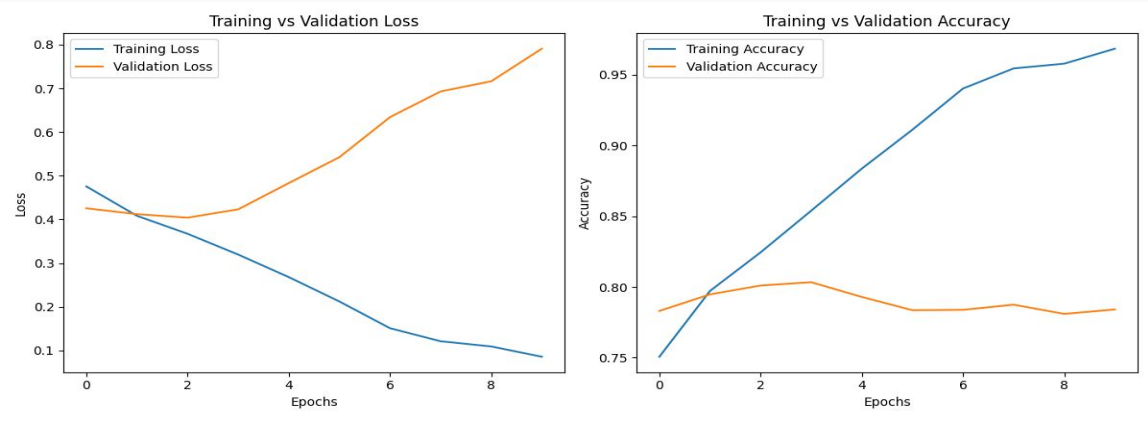


Metric	Value
Accuracy	79.79%
F1 Score (Harmful)	0.84
F1 Score (Non-Harmful)	0.72
Precision (Harmful)	0.90
Precision (Non-Harmful)	0.65
Recall (Harmful)	0.79
Recall (Non-Harmful)	0.81
Macro Avg F1	0.78
Weighted Avg F1	0.80

ResNet50-BERT Multimodal Model Overview

- **Image Features**: ResNet50 (2048-dim, pretrained, final layer removed)
- **Text Features**: Multilingual BERT (CLS token, 768-dim)
- **Fusion**: Concatenated image + text → 2816-dim
- **Classification Head**: 2816 → 512 → 256 → 2 (with ReLU & Dropout 0.3/0.2)
- **Training Details**:
  - Optimizer: AdamW (LR = 2e-5), Loss: Cross-Entropy
  - Batch Size: 16, Scheduler: ReduceLROnPlateau
  - Fine-tuned last 2 layers only; others frozen

DenseNet121 + BERT Classifier

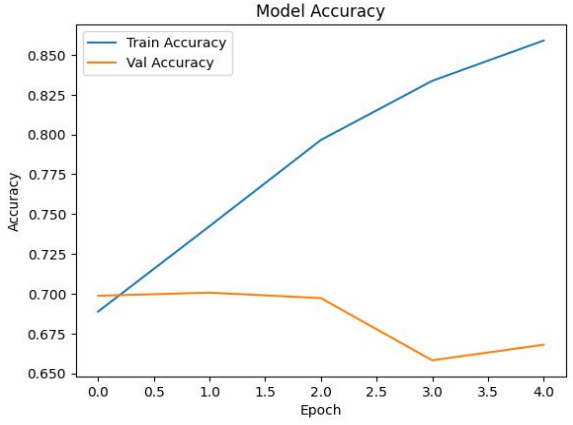


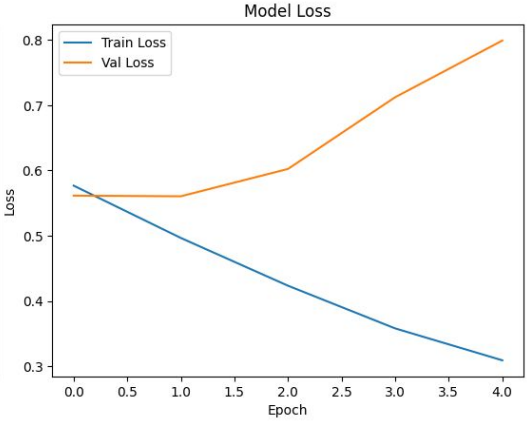
Metric	Value
Test Accuracy	80.86%
Precision (Harmful)	0.86
Recall (Harmful)	0.86
F1 Score (Harmful)	0.86
Precision (Non-harmful)	0.70
Recall (Non-harmful)	0.71
F1 Score (Non-harmful)	0.70

DenseNet121-BERT Multimodal Model Overview

- **Image Features**: DenseNet121 (1024-dim, pretrained, final layer removed)
- **Text Features**: Multilingual BERT (CLS token, 768-dim)
- **Fusion**: Concatenated image + text → 1792-dim
- **Classification Head**: 1792 → 512 → 256 → 2 (with ReLU & Dropout)
- **Training Details**:
  - Transfer Learning: Only final layers trainable
  - Partial Freezing: Earlier layers of DenseNet & BERT frozen

Bidirectional Long Short-Term Memory (BiLSTM)





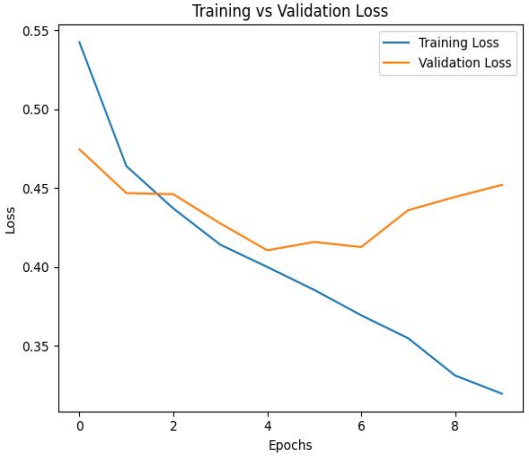
Classification Report:

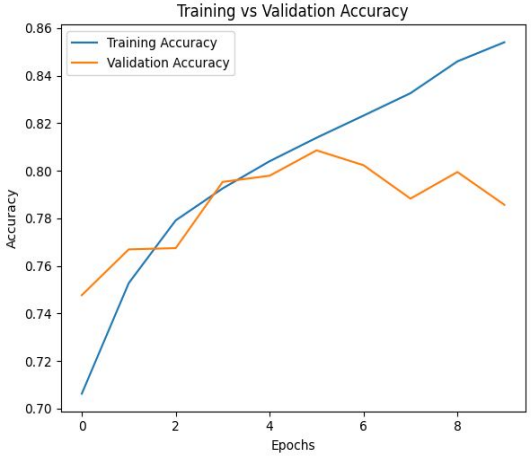
	precision	recall	f1-score	support
Non-Hateful	0.56	0.38	0.45	1644
Hateful	0.74	0.86	0.80	3476
accuracy			0.70	5120
macro avg	0.65	0.62	0.62	5120
weighted avg	0.68	0.70	0.69	5120

BiLSTM Text Classification Model Overview

- **Text Features**: Tokenized multilingual text (vocab size = 10,000, padded sequences)
- **Age Feature**: Standardized using `StandardScaler`
- **Embedding Layer**: 10,000 vocab size, 64-dim embeddings
- **BiLSTM Layers**:First BiLSTM: 64 units (return sequences = True), Dropout: 0.5, Second BiLSTM: 64 units (return sequences = False)
- **Classification Head**: Dense: 128 units with ReLU, Dropout: 0.5, Output: 1 unit with Sigmoid (binary classification)
- **Training Details**: Loss Function: Binary Cross-Entropy, Optimizer: Adam (default learning rate), Batch Size: 32, Validation Split: 10%, Early Stopping: Patience = 3 (monitored `val_loss`)

MuRIL Multimodal Classifier



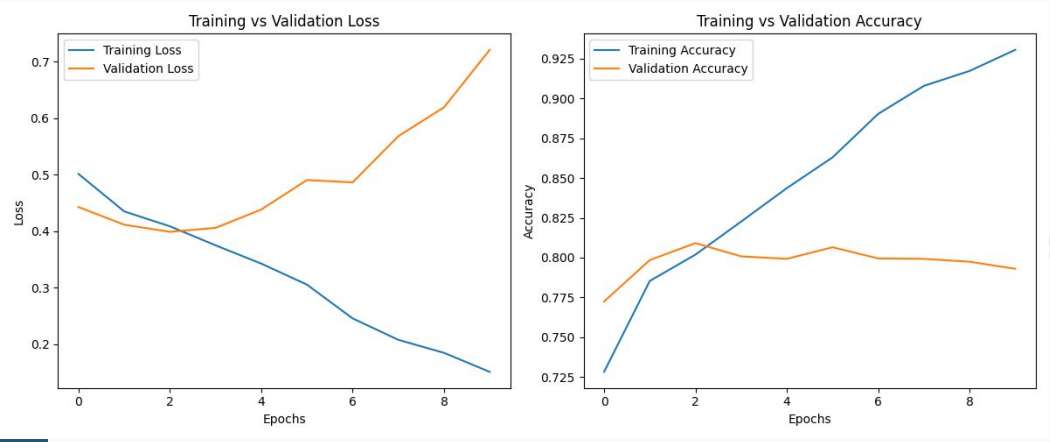


CLASSIFICATION REPORT FOR MURIL MULTIMODAL MODEL

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

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.

XLM-R Text Classifier



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

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.

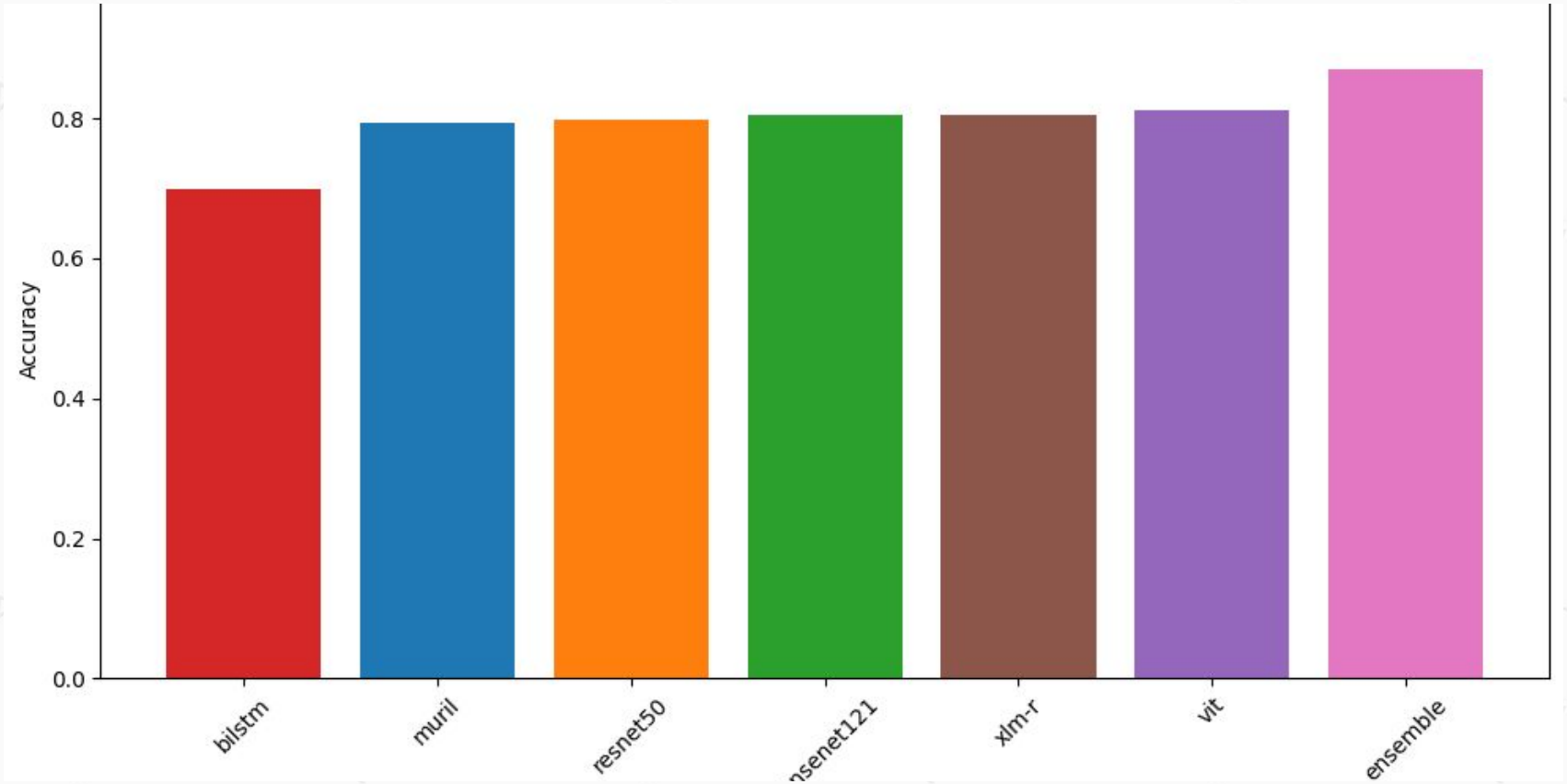
ViT-BERT Multimodal Classifier

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.

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

Model Accuracy Comparison

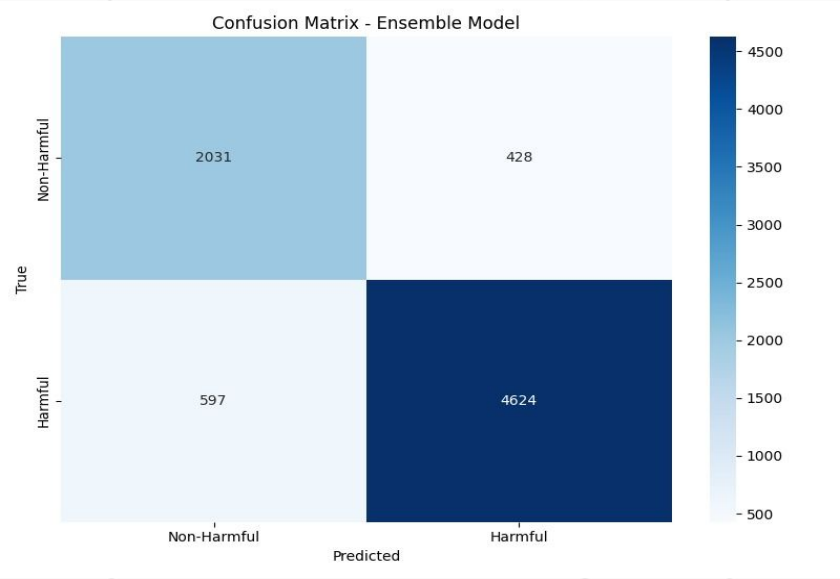


Voting Ensemble

```
===== 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
```



Using the outputs of 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.

Final Results

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

05

Conclusion & Future Works

Conclusion and Future Work

The system integrates multiple architectures — **XLM-R, MuRIL (text)**, **ViT, ResNet50, DenseNet121 (image)**, and **BiLSTM (sequence)** — to handle multilingual, multimodal hateful meme detection.

Majority voting ensemble ensures robust and reliable performance across diverse inputs.

Future Work :

- Replace majority voting with **trainable ensemble methods** (e.g., stacking)
- **Expand dataset** with underrepresented languages and nuanced hate expressions
- Explore **lightweight transformer models** for improved **inference speed and scalability**

Limitations of the Project

- **OCR Inaccuracies**: Errors during Optical Character Recognition (OCR) affect the quality of text fed into downstream models.
- **Translation Limitations**: Translating non-English text to English may lead to **semantic distortion**, weakening multilingual understanding.
- **Computational Overhead**: Running multiple large-scale models (e.g., ViT, BERT, ResNet) increases memory, inference time, and energy use — limiting real-time deployment.
- **Dataset Imbalance**: A skewed class distribution biases the models toward the majority class, leading to **lower sensitivity** to hateful content.
- **Lack of Interpretability**: The ensemble system provides a final label without offering explainability, making it difficult to justify decisions in sensitive applications.
- **Simple Majority Voting**: All models are equally weighted, ignoring **individual model confidence**, reducing ensemble accuracy and adaptability.

06

Timeline

Timeline

Phase 1: Research and Literature Review (Jan - mid Feb)
Conduct a comprehensive review of existing literature and current models for hateful meme detection including memes with different languages. 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 Feb - March)
Gather and preprocess datasets, including labeled examples of hateful memes and related content. Finding datasets was really complex tasks as very limited number of datasets are available for this task. We did `text processing and normalization, feature extraction and multimodal integration to get the final feature embeddings.`

Phase 3: Model Development and Evaluation (March - mid April)
`The final feature embeddings` was utilized in Pipeline 2, where it was trained using six models (**XLM-R, MuRIL, BiLSTM, ViT-BERT, ResNet50, and DenseNet121**).

Phase 4: Final Model Optimization (mid April)
The outputs from the six models (**XLM-R, MuRIL, BiLSTM, ViT-BERT, ResNet50, and DenseNet121**) were combined using a Voting Ensemble model to classify memes.

Phase 5: Documentation and Reporting (end April)
Compile findings, document methodologies, and prepare a final report and work on the future scope if implemented successfully whatever we wanted to achieve.

References

- Abdullakutty, F., & Naseem, U., *Decoding Memes: A Comprehensive Analysis of Late and Early Fusion Models for Explainable Meme Analysis*. Robert Gordon University, UK & Macquarie University, Australia. Link
- Ma, J., & Li, R., *RoJiNG-CL at EXIST 2024: Leveraging Large Language Models for Multimodal Sexism Detection in Memes*. University of Zurich, Switzerland.
- 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, Macquarie University. Link
- Huang, J., Lyu, H., Pan, J., Wan, Z., & Luo, J., *Evolver: Chain-of-Evolution Prompting to Boost Large Multimodal Models for Hateful Meme Detection*. Link
- Karthik, G., *Hate-CLIPper: Multimodal Hateful Meme Classification based on CLIP Features*. GitHub
- Li, L. et al., *VisualBERT: A Simple and Performant Baseline for Vision and Language*.
- Conneau, A. et al., *XLM-R: Unsupervised Cross-lingual Representation Learning at Scale*. ArXiv
- Dosovitskiy, A. et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (ViT)*. ArXiv
- Kakwani, D. et al., *MuRIL: Multilingual Representations for Indian Languages*.ArXiv
- Schuster, M., & Paliwal, K.K., *Bidirectional Recurrent Neural Networks (BiLSTM)*.IEEE
- He, K., Zhang, X., Ren, S., & Sun, J., *Deep Residual Learning for Image Recognition (ResNet50)*. ArXiv
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q., *Densely Connected Convolutional Networks (DenseNet)*.ArXiv

THANK YOU