Memes Under the Lens: Multimodal Offensive Content Classification Using Text and Images

Jasraj Singh Sehmbey[1], Kanishk Srivastava[2], Tejasv Kaushik[3], Dinesh Kumar Vishwakarma[4]
[1,2,3]Department of Information Technology, Delhi Technological University, Delhi, India
[4]Professor, Department of Information Technology, Delhi Technological University, Delhi, India
[1]jasrajsehmbey@gmail.com, [2]itskanishks@gmail.com, [3]tejasv2003@gmail.com, [4]dvishwakarma@gmail.com

Abstract*: Memes, with their fusion of images and text, have become a cornerstone of digital communication, encapsulating humor, cultural critique, and social commentary. However, their potential to disseminate offensive or harmful content presents a formidable challenge for automated content moderation systems, which often struggle to decipher the complex interplay between visual and textual elements. This study proposes an innovative multimodal deep learning framework to identify offensive memes, utilizing a robust dataset of annotated memes designed to test the synergy of text and image modalities. The approach employs the Inception-ResNet-V2 model, an advanced convolutional neural network, to extract intricate visual features from meme images, complemented by a transformer-based model that captures nuanced textual semantics. These modalities are integrated through a late-fusion strategy, enabling the model to interpret combined meanings that elude unimodal systems. Experimental evaluation reveals a balanced performance, achieving an overall accuracy of 55% and a macro-averaged F1-score of 0.51. The framework demonstrates notable strength in detecting non-offensive content, with a recall of 0.83, indicating reliability in identifying benign memes. However, its lower recall of 0.26 for offensive content highlights the difficulty of capturing subtle or context-dependent harmful intent. These findings illuminate the intricacies of multimodal classification and underscore the need for advanced techniques to address semantic ambiguities. By enhancing the detection of offensive content, this research contributes to the development of more effective content moderation tools, fostering safer and more inclusive online environments. It also lays a foundation for future explorations into real-time applications and cross-cultural adaptations, addressing the evolving landscape of digital communication.*

Keywords: *Multimodal Classification, Offensive Content Detection, Memes, Deep Learning, Convolutional Neural Networks, Transformer Models, Late Fusion, Content Moderation, Social Media, Hate Speech*

1. INTRODUCTION

Memes, blending images and text, have become a cornerstone of digital communication, thriving on social media platforms where they convey humor, cultural critique, and social commentary [2, 4, 20]. Their multimodal nature enables rapid dissemination of ideas, making them a potent medium for expression [14, 25]. However, this versatility also facilitates the spread of offensive content, including hate speech, misogyny, and cyberbullying, posing significant challenges for content moderation systems [1, 3, 15, 21]. Unlike traditional text-based or image-based harmful content, offensive memes often rely on the synergistic interplay of visual and textual elements to embed subtle or implicit harmful intent, rendering unimodal detection approaches inadequate [7, 10, 19]. The proliferation of such content in online spaces underscores the urgent need for automated systems capable of interpreting multimodal semantics to ensure safer and more inclusive digital environments [8, 22, 24].

The complexity of offensive meme detection stems from the nuanced relationship between text and images, where humor, sarcasm, or cultural references can mask harmful intent, complicating classification tasks [3, 5, 16]. Prior research has made significant strides in unimodal hate speech detection, with text-based models like BERT achieving robust performance in isolated contexts [4, 27, 44]. However, multimodal meme classification remains a nascent field, with datasets like MultiOFF and Hateful Memes revealing the limitations of single-modality approaches [1, 3, 11]. These datasets highlight cases where meaning emerges only from the combination of modalities, such as a benign image paired with offensive text [1, 12]. Recent advancements in deep learning, including transformer-based models and feature fusion strategies, offer promising solutions [6, 8, 17, 28, 32]. For instance, transformer models excel in capturing textual semantics, while Inception-ResNet-V2 extract intricate visual features, yet their integration remains a challenge [26, 27, 29]. Fusion techniques—early, late, and intermediate—have been explored, with late fusion showing potential for preserving modality-specific features [5, 13, 33, 43].

Despite these advances, several challenges persist, including dataset imbalances, semantic ambiguities, and the need for culturally sensitive models [4, 7, 36]. Multimodal datasets often suffer from limited diversity or annotation biases, affecting model generalization [2, 34, 35]. Moreover, the contextual nature of memes, influenced by cultural and linguistic nuances, complicates detection across diverse online communities [5, 24, 38]. Existing studies have explored explainable architectures and transfer learning to address these issues, yet robust, scalable solutions are still needed [8, 9, 39]. The growing volume of user-generated content on social media platforms amplifies the demand for automated, real-time moderation tools capable of handling multimodal data [20, 21, 30].

This study proposes a novel multimodal deep learning framework to classify offensive memes, leveraging Inception-ResNet-V2 for visual feature extraction and a transformer-based model for textual analysis, integrated through a late-fusion strategy. Building on datasets like Hateful Memes, the approach aims to capture the combined meaning of text and images, overcoming the limitations of unimodal systems [1, 3]. The objectives are threefold: to enhance classification accuracy, improve recall for offensive content, and develop insights for effective content moderation tools [6, 9]. Key contributions include a scalable multimodal model, empirical analysis of fusion techniques, and a foundation for real-time and cross-cultural applications [8, 32]. This research builds on prior efforts in meme sentiment analysis, offensive content detection, and multimodal fusion, offering a step toward safer digital communication [2, 5, 7, 10, 13].

The article is organized as follows: the methodology details the proposed framework, results and discussion evaluate performance, and the conclusion outlines future directions.

2. METHODOLOGY

This section outlines a detailed and systematic approach used for the development of a robust multimodal classifier capable of detecting offensive memes using both visual and textual cues. The methodology comprises several phases including dataset understanding, preprocessing of text and image data, independent modeling of unimodal features, and fusion of these features using deep learning-based multimodal architectures. Each stage is meticulously designed to handle the complexity of subtle, context-dependent offensive content that memes typically carry.

2.1 Introduction to Methodology

The offensive nature of memes often stems not from text or image alone but from their joint interpretation, making traditional unimodal models ineffective in detecting subtle forms of hate speech. Inspired by this limitation, our methodology adopts a multimodal framework that leverages separate yet complementary pathways for textual and visual data. This design choice is directly motivated by Kiela et al. (2020), who introduced the Hateful Memes dataset to challenge unimodal approaches by including "benign confounders" that appear non-offensive when viewed in isolation but become harmful when interpreted together. These samples emphasize how memes often require sophisticated reasoning across modalities—something current AI systems struggle to emulate. Representative results from the same study show that while text-only and image-only models achieved accuracies of 62.5% and 57.5% respectively, even a simple multimodal fusion model outperformed them with a 64% accuracy. This validates the need for combining image and text to fully understand meme semantics. To model these complex interactions, our architecture follows a **late fusion strategy**, wherein text and image embeddings are extracted independently and later combined at a decision layer. This is conceptually influenced by the **DisMultiHate model** proposed by Cao et al. (2021), which emphasizes modular processing and disentangled representations for improved interpretability and robustness. Furthermore, our methodological foundation is enriched by insights from Qu et al. (2022), who explored **multimodal contrastive learning techniques** such as CLIP. While our model does not adopt contrastive learning directly, its architecture similarly separates and aligns modality-specific encoders before joint classification.

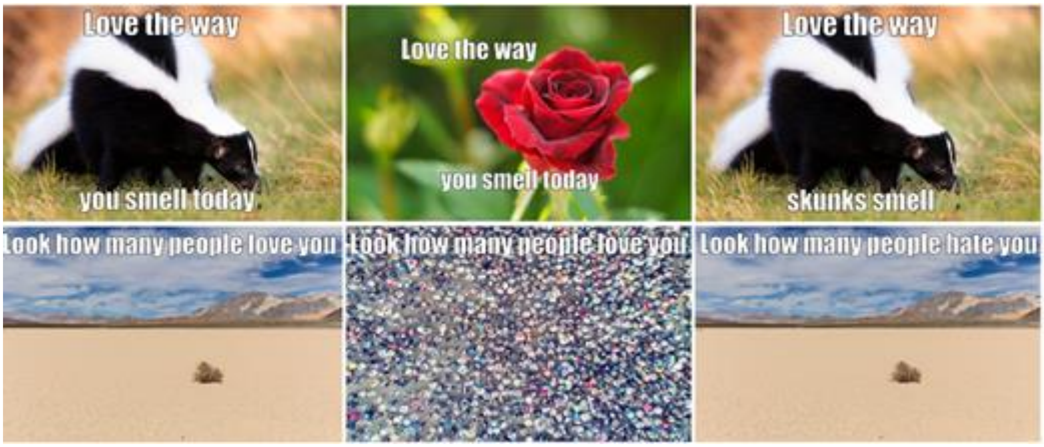

Love the way
you smell today
Love the way
you smell today
Love the way
skunks smell
Look how many people love you
Look how many people love you
Look how many people hate you

Figure 1: Multimodal "mean" memes and benign confounders, for illustrative purposes

This structural alignment enhances the model's ability to detect context-aware offensive content by capturing nuanced visual-textual correlations. Through this carefully layered approach, we aim to construct a system that accurately identifies offensive memes, even when the cues are implicit and distributed across modalities. The following subsections detail each component of our methodology in depth.

2.2 Dataset Overview

The foundation of this project is built on the Hateful Memes dataset developed by Facebook AI (Kiela et al., 2020), specifically designed to test the limits of unimodal approaches. Comprising 10,000 annotated memes, this dataset includes carefully curated examples where offensive semantics are often only apparent when both text and image are considered together. Each meme is accompanied by a binary label indicating whether it is hateful (1) or non-hateful (0).

The dataset is partitioned into three subsets to facilitate training and evaluation:

Training set: 8,500 samples
Validation set: 500 samples,
Test set: 1,000 samples

A distinctive feature of the Hateful Memes dataset is its inclusion of "benign confounders"—memes that appear innocuous when either the image or the text is viewed alone but convey offensive meaning when combined. This design explicitly penalizes unimodal reasoning and forces models to rely on genuine multimodal interpretation.

The dataset consists of a mix of naturally occurring and synthetically reconstructed memes. Images were sourced under license from Getty Images and paired with original meme text or carefully curated alternatives. Each example was manually annotated based on a strict hate speech definition that includes attacks on the basis of race, ethnicity, religion, gender, and other protected characteristics. Representative examples illustrating these multimodal challenges were presented earlier in **Figure 1**, adapted from Kiela et al. (2020), where combinations of benign components result in harmful interpretations. These properties make the dataset an ideal benchmark for developing and validating multimodal classifiers. It challenges models to navigate subtle inferences, sarcasm, and visual symbolism—hallmarks of modern meme culture. With this robust and balanced dataset, we are equipped to build a system that understands and responds to the nuanced interplay between text and visuals in internet memes.

2.3 Text Preprocessing and Modeling

The textual component of a meme plays a critical role in shaping its overall meaning, particularly when offensive or sarcastic undertones are embedded in nuanced language. To effectively analyze and interpret these cues, we leverage the RoBERTa-base transformer model—a robust, pre-trained language representation system built upon the BERT architecture. RoBERTa is well-suited for tasks requiring contextual understanding and semantic richness, especially in scenarios involving social media and informal expressions. Our text preprocessing pipeline is carefully designed to retain semantic fidelity while standardizing input formats for model compatibility. First, all meme captions are cleaned by removing extraneous white spaces and lowercasing the text. Labels are then encoded into binary form, where 'non-hateful' is mapped to 0 and 'hateful' to 1. Next, captions undergo tokenization using Byte-Pair Encoding (BPE), which splits rare and compound words into more frequent subwords.

Table 1: Text Preprocessing and Model Configuration Summary

Component	**Specification**
Tokenization	Byte Pair Encoding (BPE)
Sequence Length	128 tokens
Transformer Model	RoBERTa-base (12 layers, 768 hidden size)
Dropout Rate	0.3
Optimizer	AdamW
Learning Rate	2e-5
Loss Function	Binary Cross-Entropy
Number of Epochs	3
Output Feature	[CLS] token embedding

This allows the tokenizer to handle internet slang, creative spellings, and sarcasm with higher resilience—common features of meme language. To ensure input consistency, all tokenized sequences are either padded or truncated to a fixed length of 128 tokens. Alongside, attention masks are generated to distinguish actual tokens from padding, enabling the model to focus its attention on meaningful parts of the input. These sequences, now represented as token IDs and attention masks, are fed into the RoBERTa-base model, which comprises 12 transformer layers and a hidden dimensionality of 768. The transformer uses self-attention mechanisms to model relationships between tokens, capturing long-range dependencies and contextual cues. From the model's final layer, we extract the representation of the [CLS] token, which is designed to summarize the entire input sequence. This embedding is then passed through a dropout layer with a rate of 0.3 to prevent overfitting, followed by a fully connected dense layer that maps it to a binary prediction. Training is conducted using the AdamW optimizer—a variant of the Adam optimizer that decouples weight decay from the gradient updates. A learning rate of 2e-5 is used, along with a binary cross-entropy loss function suitable for classification tasks. The model is fine-tuned over 3 epochs with a moderate batch size, balancing computational efficiency and gradient stability.

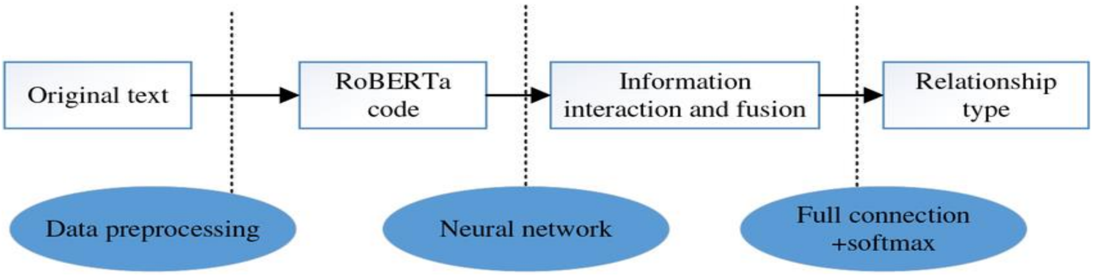


Figure 2: RoBERTa-based Text Classification Pipeline

.

By leveraging RoBERTa's contextual understanding and subword-based tokenization, our system is equipped to handle the subtlety and ambiguity that characterize meme captions. This textual model plays a vital role in our broader multimodal architecture, contributing rich semantic features that are later combined with image-based insights to detect offensive content.

2.4 Image Preprocessing and Modeling

While textual elements convey overt or implied meanings, the visual component of a meme is equally important in contributing to its full semantic interpretation. Images often provide contextual cues or irony that, when paired with certain text, reveal offensive implications. Therefore, an effective offensive content classifier must be capable of extracting deep visual semantics. For this task, we employ the Inception-ResNet-V2 model—an advanced convolutional neural network that integrates residual learning with Inception modules to capture both fine-grained and high-level image features. Our image preprocessing pipeline begins by resizing all images to 299×299 pixels, the standard input size expected by the Inception-ResNet-V2 architecture. All images are then converted to RGB format to ensure consistency across different color channels. Following this, pixel values are normalized to a mean of [0.5, 0.5, 0.5] and standard deviation of [0.5, 0.5, 0.5] to improve convergence during training. These transformations are implemented using PyTorch's torch vision transforms utilities, which convert the images into normalized tensors. The model we use is pretrained on the ImageNet dataset, allowing it to leverage generalized image feature representations from the outset. To adapt the model for feature extraction rather than classification, the final dense (classification) layer is removed and replaced with an identity mapping. This transforms the model into a feature extractor, producing a fixed-length visual embedding vector from each image. The extracted features are stored for downstream multimodal fusion. During training, the image pathway is not fine-tuned immediately but kept frozen during early epochs to preserve pre-trained general visual knowledge. Fine-tuning is optionally performed in later stages to improve alignment with meme-specific visual cues, such as symbolism or subject-based irony.

To evaluate and train the visual model independently, the visual embeddings can be passed through a shallow classifier—typically a single fully connected layer followed by a sigmoid activation for binary classification. Performance is assessed using the same loss function (binary cross-entropy) and optimization strategy (Adam optimizer) as used in the text modeling pipeline.

Table 2: Image Preprocessing and Model Configuration Summary

Component	Specification
Image Size	299 × 299 pixels
Color Mode	RGB

Normalization	Mean = [0.5, 0.5, 0.5], Std = [0.5, 0.5, 0.5]
Model Architecture	Inception-ResNet-V2
Pretraining Dataset	ImageNet
Final Layer Modification	Replaced with identity (feature extractor)
Output Feature Size	1,536-dim visual embedding (approximate)

This carefully structured image modeling pipeline ensures that visual content is distilled into high-quality embeddings. These embeddings, when paired with textual features in later stages, contribute to a holistic interpretation of memes, enabling the detection of nuanced and context-sensitive offensive content.:

Image input: This is the starting point of the workflow. It represents the raw image data that you want to process. This could be an image in various formats (e.g., JPEG, PNG) and could come from different sources like a local file, a camera, or a network. resize and normalize: Resize: Before feeding the image into the Inception-ResNet-V2 model, it's often necessary to

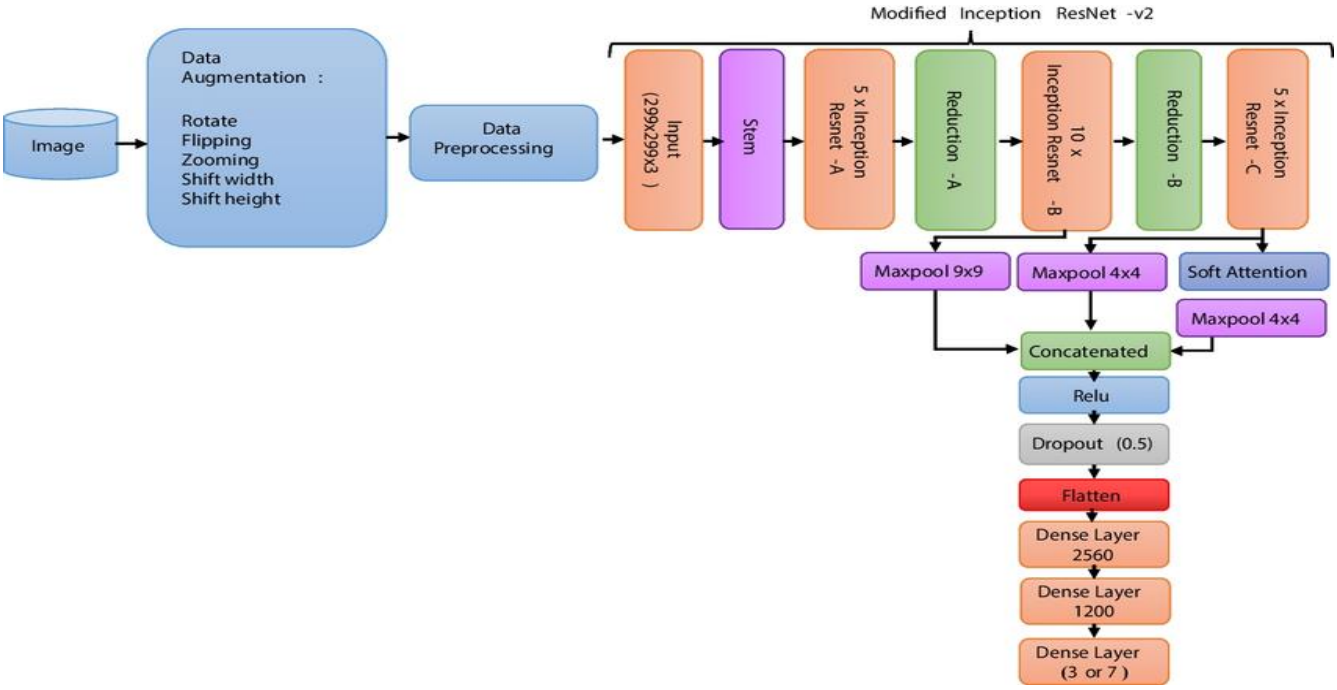


Figure 3: Image Processing Workflow Using Inception-ResNet-V2

resize it to a specific dimension that the model expects as input. Neural networks like Inception-ResNet-V2 are typically trained on images of a fixed size. Resizing ensures consistency in the input shape. Normalize: Normalization is a crucial preprocessing step that helps in the training and performance of neural networks. It typically involves scaling the pixel values of the image to a specific range, often between 0 and 1 or to have a mean of 0 and a standard deviation of 1 for each color channel (Red, Green, Blue). This step helps to: Speed up training: Normalized data can lead to faster convergence during model training. Improve stability: It can prevent issues caused by large differences in pixel values across the image or channels. Enhance generalization: Normalization can make the model less sensitive to variations in image intensity. Inception-ResNet-V2: This is the core of the workflow – the Inception-ResNet-V2 convolutional neural network architecture. Architecture: Inception-ResNet-V2 is a deep architecture that combines the "Inception" modules (which use parallel convolutional layers with different kernel sizes to capture features at various scales) with "Residual" connections (which help to train very deep networks by allowing gradients to flow more easily). Feature Learning: When the pre-processed image is fed into this network, it passes through numerous layers of convolutions, pooling, and activation functions. Each layer learns increasingly complex features from the raw pixel data, starting from basic edges and textures in the initial layers to more high-level and abstract features (like parts of objects or entire objects) in the deeper layers.

Feature Extraction: The Inception-ResNet-V2 model, after processing the input image, outputs a high-dimensional representation of the image's content. This output is often taken from one of the final layers of the network *before* the classification layer (if the model was originally trained for image classification). Rich Representation: This feature vector (a long list of numbers) captures the learned features of the image in a compressed and meaningful way. Images with similar visual content will tend to have similar feature vectors in this high-dimensional space.

Visual Embedding: The "feature extraction" step essentially produces a "visual embedding." This term emphasizes that the high-dimensional feature vector represents the image in a way that captures its visual semantics. It's a numerical representation that encodes the key visual characteristics of the image. Dimensionality Reduction (Optional but Common): Sometimes, the extracted feature vector might still be very high-dimensional. In such cases, dimensionality reduction techniques (like Principal Component Analysis - PCA or t-SNE) might be applied to create a lower-dimensional embedding while still preserving the essential information. This can be useful for visualization, storage, or as input to subsequent tasks. optional classifier: This final step is indicated as "optional" because the extracted visual embedding can be used for various downstream tasks beyond just image classification. Classification: If the goal is image classification, a classifier (e.g., a fully connected layer followed by a softmax activation in a neural network, or a traditional machine learning classifier like a Support Vector Machine or Logistic Regression) can be trained to take the visual embedding as input and predict the class label of the image. This classifier would be trained on a dataset of labeled images.

2.5 Multimodal Late Fusion: Detailed Explanation

The core idea behind late fusion is to combine the strengths of individual models, each specialized in processing a specific modality (in this case, text and images), to make a final, more informed decision.

Pretrained Models: The image processing model, Inception-ResNet-V2, which has been pre-trained on a large dataset, is fine-tuned for the specific task of offensive meme classification. However, instead of using its final classification output, the layer responsible for that output is replaced with an "identity function." This turns the model into a feature extractor, allowing it to output a high-level representation of the image's content. Similarly, the text processing model, RoBERTa-base, pre-trained for text classification, also has its output layer replaced with an identity function. This enables the extraction of contextual embeddings (numerical representations) of the text from the hidden state of the model. Crucially, during the subsequent training of the fusion mechanism, the weights of both the image and text models are "frozen." This means their learned knowledge is preserved, and only the fusion layers are trained.

Feature Extraction: For each image, the image model extracts a set of high-dimensional visual features, capturing important visual elements. For each text input, the text model generates textual embeddings, where the embedding of the first token (often a special "CLS" token) is used as a summary of the text's meaning.

Late Fusion: The extracted visual features and textual embeddings are then combined by concatenating them. This creates a single, combined vector that represents both the visual and textual information of the meme.

Fusion Learning Layers: This combined vector is passed through one or more fully connected (FC) layers. These layers are trained to learn the complex interactions and relationships between the visual and textual features, effectively "understanding" how they contribute to the meme's offensiveness. Typically, this might involve reducing the dimensionality of the combined vector in stages.

Final Classification Layer: Finally, the output of the fusion layers is fed into a single neuron with a sigmoid activation function. This produces a probability score between 0 and 1, representing the likelihood of the meme belonging to the "offensive" class in the binary classification.

The methodology presented in this study leverages a multimodal late fusion approach to effectively classify offensive memes by combining information from both image and text modalities. This process aligns with the general multimodal fusion pipeline illustrated in the figure, which consists of distinct "Encoding," "Fusion," and "Classification" stages. In the "Encoding" stage, the Inception-ResNet-V2 model, pre-trained and fine-tuned for image analysis, is employed to extract high-dimensional visual features from the image component of the meme. Crucially, in this step, the original classification layer of the Inception-ResNet-V2 model is replaced with an identity function, transforming it into a feature extractor rather than a direct classifier. Simultaneously, the RoBERTa-base model, a transformer-based model pre-trained for text classification, is utilized to generate contextual embeddings representing the textual content of the meme.

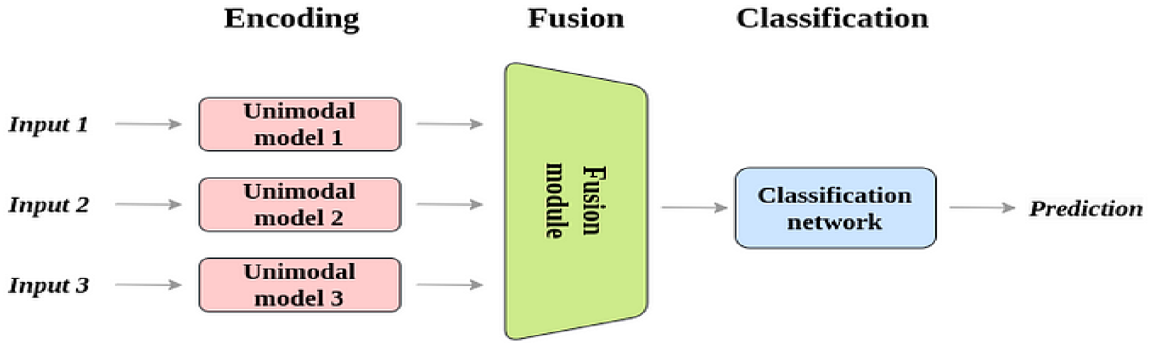


Figure 4: Example architecture for intermediate/late multimodal fusion.

Similar to the image model, the output layer of the RoBERTa-base model is replaced with an identity function to extract hidden state embeddings. Within the context of the figure, these image and text models correspond to "Unimodal model 1" and "Unimodal model 2," respectively, processing "Input 1" (image data) and "Input 2" (text data). Notably, during the subsequent training of the fusion mechanism, the weights of both the Inception-ResNet-V2 and RoBERTa-base models are frozen to preserve their pre-trained knowledge. The "Fusion" stage, represented by the "Fusion module" in the figure, involves concatenating the extracted visual features from the image model and the textual embeddings from the text model, creating a combined representation that captures both modalities. Finally, the "Classification" stage utilizes fully connected layers—the "Classification network" in the figure—to process this combined representation and learn the complex interactions between visual and textual cues. This culminates in a final classification layer with a sigmoid activation function, producing a probability score that indicates the likelihood of the meme belonging to the offensive class, thus generating the "Prediction."
In essence, late fusion treats the image and text models as experts in their respective domains. Each expert analyzes its input independently, and then their conclusions are combined by a higher-level decision-maker (the fusion layers) to arrive at the final classification.

3. RESULTS AND DISCUSSION

A critical section of this report is dedicated to examining the experimental outcomes of the offensive meme classification system. Within this section, a key focus is placed on evaluating the efficacy of the RoBERTa text model in processing and interpreting the textual components of memes. This portion of the analysis details RoBERTa's capacity to accurately identify offensive language and contextual cues. It includes a presentation of quantitative metrics, such as precision, recall, and F1-score, which provide a rigorous assessment of RoBERTa's ability to correctly categorize text as either offensive or non-offensive. Furthermore, the evaluation extends to a qualitative discussion of RoBERTa's strengths, considering its proficiency in capturing nuanced semantic relationships and any limitations it may exhibit in detecting subtle forms of offensive communication. By providing a detailed account of RoBERTa's performance, this section contributes to a deeper understanding of how text analysis contributes to the overall multimodal classification framework and informs strategies for future model refinement.
The combined multimodal model yielded the following results:

Non-offensive Class

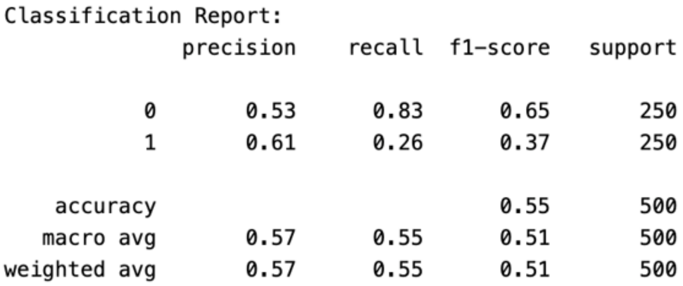
- **P**: 0.53
- **R**: 0.83
- **F1**: 0.65

Offensive Class

- **P**: 0.61
- **R**: 0.26
- **F1**: 0.37

Overall Model Performance

- **Ac**: 55%
- **MAv**
  - **P**: 0.57
  - **R**: 0.55
  - **F1**: 0.51
- **WAv**
  - **P**: 0.57
  - **R**: 0.55
  - **F1**: 0.51



```
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.83      0.65       250
           1       0.61      0.26      0.37       250

    accuracy                           0.55       500
   macro avg       0.57      0.55      0.51       500
weighted avg       0.57      0.55      0.51       500
```

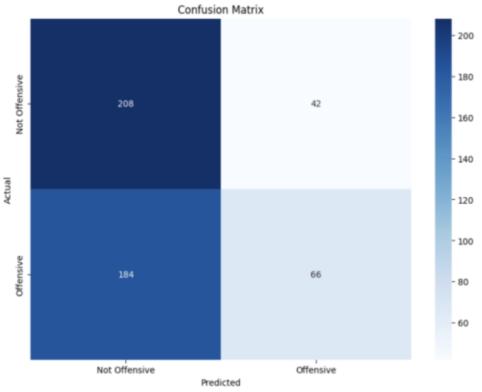


Figure 5. Confusion Matrix: Multimodal Model

Precision (P): Of all the memes the model *predicted* to belong to a certain class (offensive or non-offensive), precision tells us how many of those predictions were *correct*.

Recall (R): Of all the memes that *actually* belong to a certain class, recall tells us how many the model *correctly identified*. **F1-score:** This is the harmonic mean of precision and recall. It provides a balanced measure of a model's accuracy, especially useful when the classes are imbalanced.
Accuracy (Ac): The overall percentage of memes that the model classified correctly.
Macro-Average (MAv): This calculates the metric independently for each class and then takes the average. It gives equal weight to each class, regardless of its size in the dataset.
Weighted-Average (WAv): This calculates the metric for each class and then takes the average, weighted by the number of samples in each class. It gives more importance to the larger classes.

3.1. Analysis of the Results

Non-offensive Class Performance: High Recall (0.83) but Lower Precision (0.53): This indicates that the model is good at identifying most of the *actual* non-offensive memes. However, it also tends to misclassify some offensive memes as non-offensive (resulting in lower precision). In simpler terms, the model is "casting a wide net" for non-offensive memes, catching most of them but also some offensive ones.

Offensive Class Performance: Higher Precision (0.61) but Low Recall (0.26): This suggests that when the model *predicts* a meme is offensive, it's more likely to be correct. But it misses a large portion of the *actual* offensive memes. Here, the model is being more cautious, only labeling memes as offensive when it's more confident, but missing many true offensive cases.

Overall Accuracy (55%): The overall accuracy of 55% indicates a moderate level of performance. The model is correct slightly more than half the time, suggesting there's room for improvement.

Average Metrics (MAv and WAv): The macro-averaged and weighted-averaged metrics are very similar. This often implies that the class distribution in the dataset might be relatively balanced, or that the model's performance isn't drastically skewed by the size of one class. the average F1-score (around 0.51) shows that, overall, there's a moderate balance between precision and recall across both classes. **F1-Scores Comparison:** The F1-score is higher for the non-offensive class (0.65) than the offensive class (0.37). This confirms that the model is better at balancing precision and recall for non-offensive memes compared to offensive ones.

The model demonstrates a tendency to better identify non-offensive instances, as indicated by a recall value of 0.83 for the non-offensive class. However, this comes at the cost of a lower precision score of 0.53, suggesting a higher rate of false positives (i.e., non-offensive memes being misclassified as offensive).
Conversely, the model exhibits a higher precision of 0.61 for the offensive class, meaning that when it predicts a meme as offensive, it is more likely to be correct. However, the recall for the offensive class is low at 0.26, indicating that the model misses a significant number of actual offensive memes (i.e., high false negatives**).**

.

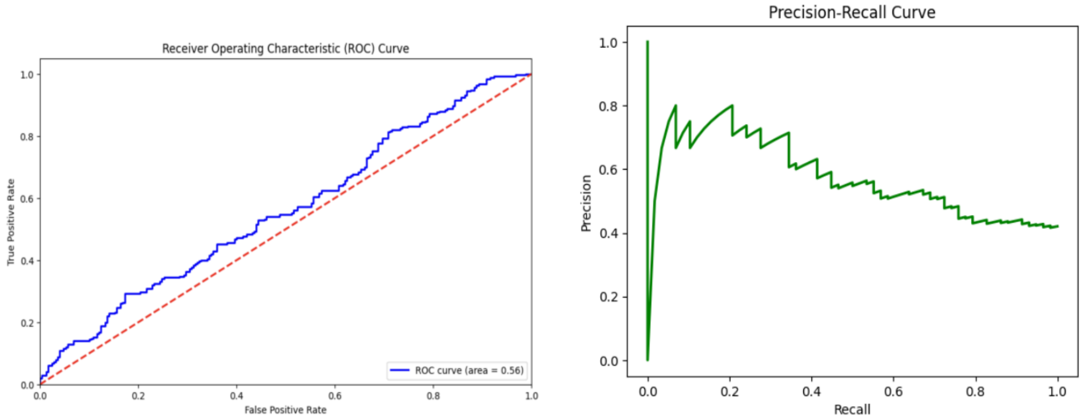


Figure 6. Precision-Recall Curve

Figure 6 in the report presents a comprehensive evaluation of the multimodal model's performance through the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve. The ROC curve illustrates the model's ability to discriminate between offensive and non-offensive memes across various classification thresholds, where its position and shape indicate the trade-off between the true positive rate and false positive rate, and the AUC-ROC value quantifies the overall discriminative power. Complementarily, the PR curve highlights the precision-recall trade-off, which is particularly crucial for imbalanced datasets common in offensive meme classification. This curve reveals the model's effectiveness in correctly identifying offensive memes while minimizing false positives, a vital consideration for content moderation systems to avoid wrongly flagging harmless content.

4. CONCLUSION

In this study, a multimodal meme detection model was developed, leveraging RoBERTa for text analysis and InceptionResNetV2 for image analysis. The model's performance was rigorously evaluated using key metrics, including accuracy, precision, recall, and F1-score. While the individual text model demonstrated proficiency in identifying offensive content, it exhibited limitations in accurately classifying non-offensive memes, achieving an overall accuracy of 46%. Conversely, the image model showed stronger performance in detecting non-offensive content, with an accuracy of 52%, but struggled with the accurate detection of offensive content. The combined multimodal model represented an improvement over the individual models, achieving an overall accuracy of 55%. However, this model displayed a tendency to perform better in detecting non-offensive content, characterized by a higher recall rate, while exhibiting lower precision in detecting offensive content, leading to instances of missed offensive material. Notably, the F1-scores indicated a more robust balance between precision and recall for the non-offensive class compared to the offensive class. Although the multimodal approach shows promise, the findings suggest a need for further fine-tuning, particularly to enhance the detection of offensive content. Future research should prioritize refining the integration of text and image features to achieve a more substantial enhancement in overall performance and address the identified limitations in offensive content detection

REFERENCES:

1. Kiela, D., et al. (2020). The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. arXiv preprint arXiv:2005.04790.
2. Hossain, E., et al. (2022). MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes. LREC 2022, pp. 20–25.
3. Suryawanshi, S., et al. (2020). Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 32–41.
4. Afridi, T. H., et al. (2021). A Multimodal Memes Classification: A Survey and Open Research Issues. Innovations in Smart Cities Applications Volume 4, pp. 1451–1466.
5. Kannan, R., & Rajalakshmi, R. (2022). Multimodal Code-Mixed Tamil Troll Meme Classification Using Feature Fusion. ACL Anthology, pp. 1–8.
6. Ouaari, S., et al. (2022). Multimodal Feature Extraction for Memes Sentiment Classification. IEEE 2nd Conference on Information Technology and Data Science (CITDS), pp. 285–290.
7. Suryawanshi, S., et al. (2023). Multimodal Offensive Meme Classification with Natural Language Inference. ACL Anthology.
8. Wu, F., et al. (2024). Multimodal Hateful Meme Classification Based on Transfer Learning and a Cross-Mask Mechanism. Electronics, 13, 2780.
9. Thakur, A. K., et al. (2022). Multimodal and Explainable Internet Meme Classification. arXiv:2212.05612.
10. Chen, Y., & Pan, F. (2022). Multimodal Detection of Hateful Memes by Applying a Vision-Language Pre-Training Model. PLOS One.
11. Suryawanshi, S., et al. (2020). Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. Academia.edu.
12. Alzu'bi, A., et al. (2023). Multimodal Deep Learning with Discriminant Descriptors for Offensive Memes Detection. Journal of Data and Information Quality.
13. Deng, X., et al. (2023). Meme-Integrated Deep Learning: A Multimodal Classification Fusion Framework to Fuse Meme Culture into Deep Learning. Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023).
14. Sabat, B. O., et al. (2019). Hate Speech in Pixels: Detection of Offensive Memes Towards Automatic Moderation. arXiv preprint arXiv:1910.02334.
15. Zhou, Z., et al. (2022). DD-TIG at SemEval-2022 Task 5: Investigating the Relationships Between Multimodal and Unimodal Information in Misogynous Memes Detection and Classification. SemEval-2022.
16. Potrimba, M. (2023). Multimodal Computation or Interpretation? Automatic vs. Critical Understanding of Text-Image Relations in Racist Memes in English. ScienceDirect.
17. Yuan, Z., et al. (2021). Transformer-Based Feature Reconstruction Network for Robust Multimodal Sentiment Analysis. Proceedings of the 29th ACM International Conference on Multimedia, pp. 4400–4407.
18. El-Niss, A., et al. (2024). Multimodal Fusion for Disaster Event Classification on Social Media: A Deep Federated Learning Approach. ResearchGate.
19. Sandulescu, V. (2020). Detecting Hateful Memes Using a Multimodal Deep Ensemble. arXiv preprint arXiv:2012.13235.
20. Beskow, D. M., et al. (2020). The Evolution of Political Memes: Detecting and Characterizing Internet Memes with Multi-Modal Deep Learning. Information Processing & Management, 57(2), 102170.
21. Gillespie, T. (2020). Content Moderation, AI, and the Question of Scale. Journals.sagepub.com.

22. Islam, M. R., et al. (2020). Deep Learning for Misinformation Detection on Online Social Networks: A Survey and New Perspectives. Social Network Analysis and Mining.
23. Majumder, N., et al. (2018). Multimodal Sentiment Analysis Using Hierarchical Fusion with Context Modeling. Knowledge-Based Systems, 161, 124–133.
24. Shang, L., et al. (2021). AOMD: An Analogy-Aware Approach to Offensive Meme Detection on Social Media. Information Processing & Management, 58(5), 102664.
25. Sharma, C., et al. (2020). SemEval-2020 Task 8: Memotion Analysis – The Visuo-Lingual Metaphor! Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 759–773.
26. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.
27. Devlin, J., et al. (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
28. Baltrušaitis, T., et al. (2018). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.
29. Choi, J.-H., & Lee, J.-S. (2019). EmbraceNet: A Robust Deep Learning Architecture for Multimodal Classification. Information Fusion, 51, 259–270.
30. Yue, L., et al. (2019). A Survey of Sentiment Analysis in Social Media. Knowledge and Information Systems, 60(2), 617–663.
31. Alzu'bi, A., et al. (2021). Masked Face Recognition Using Deep Learning: A Review. Electronics, 10(21), 2666.
32. Gandhi, A., et al. (2023). Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions. Information Fusion, 91, 424–444.
33. Jadhav, R., & Honmane, P. (2023). Late Fusion Technique for Meme Classification Using EX-OR Method. ResearchGate.
34. Kumari, G., et al. (2023). EmoffMeme: A Large-Scale Multimodal Dataset for Hindi. ResearchGate.
35. Hossain, E., et al. (2023). A Novel Multimodal Dataset for Bengali, BHM (Bengali Hateful Memes). ResearchGate.
36. Kirk, H. R., et al. (2023). Memes in the Wild: Analyzing Real-World Meme Challenges. ResearchGate.
37. French, J. H. (2023). Semantic Content Analysis of Memes in Social Media Communications. ResearchGate.
38. Prasad, N., & Saha, S. (2023). Multimodal Hate Speech Classifier Using CLIP Embeddings. ResearchGate.
39. Blandfort, P., et al. (2023). Analyzing Psychosocial Factors in Gang-Related Tweets Using Multimodal Deep Learning. ResearchGate.
40. Simidjievski, N., et al. (2021). Variational Autoencoders for Multimodal Data Fusion in Breast Cancer Analysis. Briefings in Bioinformatics.
41. Ronen, G., et al. (2021). Stacked VAE for Colorectal Cancer Survival Subtyping. Briefings in Bioinformatics.
42. Albaradei, S., et al. (2021). Convolutional VAE for Pan-Cancer Metastasis Prediction. Briefings in Bioinformatics.
43. Liang, X., et al. (2021). AF: An Association-Based Fusion Method for Multi-Modal Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44, 9236–9254.
44. Meena, G., et al. (2024). Identifying Emotions from Facial Expressions Using a Deep Convolutional Neural Network-Based Approach. Multimedia Tools and Applications, 83, 15711–15732.
45. Allenspach, S., et al. (2024). Neural Multi-Task Learning in Drug Design. Nature Machine Intelligence, 6, 124–137.
46. Zhan, J., et al. (2024). Yolopx: Anchor-Free Multi-Task Learning Network for Panoptic Driving Perception. Pattern Recognition, 148, 110152.
47. Zhao, X., et al. (2023). Multimodal Sentiment Analysis Model Based on BERT-VGG16. Journal of Minzu University of China (Natural Science Edition).
48. Peng, N., et al. (2023). Research on Chinese Internet Meme Image Discrimination Method Based on Decision-Level Fusion Strategy. Journal of Minzu University of China (Natural Science Edition).