joy_2004114 at MAHED Shared Task : Filtering Hate Speech from Memes using A Multimodal Fusion-based Approach

Joy Das, Alamgir Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology
{u2004114, 23mcse701}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Social media platforms have become major spaces for sharing opinions, humor, and information through memes that blend images with text. While many memes are harmless, some promote hate speech against individuals or communities based on cultural, religious, gender, or national identity. Detecting such content in Arabic is particularly challenging due to linguistic complexity, cultural context, and limited annotated data. In this study, we present an effective approach for detecting hateful content in Arabic memes using the **QCRI Prop2Hate-Meme** dataset, which contains image–text pairs labeled for hatefulness. We experimented with several multimodal configurations, and the best performance was achieved using a combination of InceptionNet for visual features and multilingual BERT for text. These representations were fused after applying normalization and augmentation to enhance robustness. Our **InceptionNet** with **mBERT** configuration achieved a macro F1-score of 63 percent and secured the sixth position on the official CodaBench leaderboard. These findings highlight the strength of our multimodal model and support its potential for detecting harmful Arabic content in low-resource settings.

1 Introduction

With the exponential growth of social media platforms, memes have evolved into a dominant means of communication, often blending visual and textual elements to express humor, satire, or commentary. However, this same format has been increasingly exploited to propagate hateful narratives targeting individuals or communities based on attributes such as culture, religion, sex or nationality (Kiela et al., 2021; Pramanick et al., 2021; Sharma et al., 2020a). Unlike conventional text based hate speech, hateful memes present a unique detection challenge since the offensive intent may only emerge when text and image are interpreted together (Das et al., 2020; Zhao et al., 2023), making unimodal approaches insufficient.

In Arabic speaking contexts, the task is further complicated by several factors. First, there is a scarcity of large scale, high quality annotated datasets for multimodal hate speech detection (Zaghouani et al., 2025) . Second, most state-of-the-art detection models have been trained primarily on English datasets, limiting their transferability due to linguistic, cultural, and script specific nuances . Text only models risk overlooking visual sarcasm or symbolism, while image only systems may fail to capture hateful meaning embedded in overlaid text, leading to both false positives and false negatives.

The NeurIPS "Hateful Memes" Challenge (Kiela et al., 2021) highlighted how benign confounders, individually innocuous text and images that form hateful meaning only when combined, require models to perform genuine multimodal reasoning. Large scale vision language transformers such as **UNITER** (Chen et al., 2020), **ViLT** (Kim et al., 2021), and **CLIP** (Arya et al., 2024; Radford et al., 2021) have achieved strong results in high-resource settings but their reliance on vast amounts of paired data and high computational cost renders them impractical for low resource languages like Arabic. While dual-encoder fusion strategies (Ahsan et al., 2024; Hossain et al., 2022; Lippe et al., 2020; Zhou et al., 2021) have shown promising performances in other languages, systematic evaluations for Arabic meme moderation remain scarce.

In order to overcome these challenges, we proposed a lightweight dual-encoder multimodal framework that combined a fine-tuned **Inception-ResNetV2** image encoder (Szegedy et al., 2016) with a **multilingual BERT (mBERT)** text encoder (Pires et al., 2019). Visual features were extracted from resized meme images, while textual features were derived from OCR-extracted Arabic text after normalization (Kaundilya et al., 2019; Hossein-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 683–691

mardi et al., 2015). These representations were concatenated and passed through a compact multilayer perceptron for binary classification, following prior dual-encoder fusion strategies in multimodal hate speech detection (Pramanick et al., 2021; Ahsan et al., 2024).The design maintained a trade-off between accuracy and efficiency, making it practical for use in environments with limited computational resources.

The key contributions of this study are :

- We developed a lightweight multimodal architecture for Arabic hateful meme detection by integrating InceptionResNetV2 with mBERT.
- We conducted comparative experiments across multiple model combinations and found that InceptionNet + mBERT achieved the best macro F1.
- We designed a preprocessing pipeline with OCR-based text extraction, normalization, and image augmentation to improve robustness.

2 Background & Related Work

2.1 Task Definition

We participated in **Subtask 3: Multimodal Hateful Meme Detection**, which is part of **Shared Task 4 (MAHED 2025: Multimodal Detection of Hope and Hate Emotions in Arabic Content)**, organized under **Track 1: Speech and Multimodal Processing** at the **ArabicNLP 2025** workshop. We used the QCRI/Prop2Hate-Meme[1](Alam et al., 2024), which was released for this shared task. This subtask focuses on classifying Arabic memes that contain both images and text as either *hateful* or *non-hateful*. Each sample contains:

1. **Image**: Visual content, symbols, or scenes conveying context or sentiment.
2. **Embedded Arabic text**: Extracted text from the image, providing essential linguistic context.

A meme is classified as hateful if it explicitly or implicitly promotes hostility, discrimination, or stereotypes toward a targeted group. Non-hateful memes lack such harmful content, even when expressing strong opinions or satire.

[1]https://huggingface.co/datasets/QCRI/Prop2Hate-Meme

Figure 1: Examples of hateful and non-hateful memes.

Formally, the problem is modeled as a binary classification task:

$$f(image, text) \rightarrow \{hateful, non-hateful\}$$

The key challenge lies in effectively capturing the interplay between visual and textual information, as hateful intent often arises from their combined interpretation rather than either modality alone.

2.2 Related Work

Detecting hateful memes inherently requires joint vision–language reasoning. The Hateful Memes benchmark introduced by Kiela (Kiela et al., 2021) incorporated benign confounders, text and images that appear harmless in isolation but convey hateful meaning when combined, to prevent models from exploiting unimodal shortcuts. Their multimodal baselines, such as late-fusion BERT(Devlin et al., 2019) + ResNet(He et al., 2015), achieved around 65% accuracy, whereas unimodal baselines rarely exceeded 60%. The follow-up competition report showed that improved cross-modal grounding, use of auxiliary supervision and more effective fusion strategies enabled some teams to surpass 70% accuracy.

Subsequent works explored architectural variations. Zhou (Kiela et al., 2021) integrated auxiliary image and text matching tasks with a BERT + ResNet pipeline, reporting macro F1 gains of approximately 3–4 percentage points over vanilla fusion. Lippe (Lippe et al., 2020) employed contrastive learning to better align modalities, improving robustness to adversarial confounders and achieving competitive leaderboard rankings.

Large-scale vision–language transformers such as ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019), UNITER, ViLT, and CLIP have demonstrated strong modality alignment across a range of multimodal tasks. These models were not originally trained on the Hateful Memes dataset, but some have been later fine-tuned for it and achieved over 70 percent accuracy. However, their reliance on large paired datasets and substantial computational resources limits their suitability for low-resource language settings.

In the related ArAIEval 2024 propaganda meme classification task (Hasanain et al., 2024), the top multimodal system (AlexUNLP-MZ)(Zaytoon et al., 2024) reached a macro F1 of 0.8051. These results demonstrate the benefit of combining visual and textual features for Arabic meme moderation. Despite this progress, lightweight dual-encoder fusion commonly explored for English and multilingual settings which remains under investigated for Arabic multimodal content moderation.

Cross-lingual studies further demonstrate the viability of compact fusion strategies. Datasets such as Memotion(Sharma et al., 2020b) and MUTE (Hossain et al., 2022) have enabled systematic benchmarking, while dual-encoder pipelines (Ahsan et al., 2024) and CLIP-based transfer learning can perform competitively in low-resource contexts, with reported results ranging from 60–68% depending on modality balance and data quality.

Building on these findings, our work adopts a lightweight vision backbone combined with multilingual BERT to balance accuracy, OCR robustness, and deployability in resource constrained settings for Arabic hateful meme detection.

3 Dataset

We use the QCRI/Prop2Hate-Meme corpus released for the Arabic multimodal meme shared task. The dataset pairs each meme image with aligned Arabic text and provides both coarse (hateful vs. non-hateful) and fine-grained hatefulness labels, while preserving related propaganda annotations. Appendix A shows the dataset's statistics, including important measures and how the data is spread out.

4 Methodology

Our approach combines text and image information to detect hateful content in Arabic memes. The workflow begins with preprocessing of both text and image inputs, followed by feature extraction using pre-trained models, and finally a fusion step that integrates the two modalities for classification. An outline of the complete pipeline is presented in Figure 2, which illustrates how data flows through preprocessing, feature extraction, and multimodal fusion before reaching the classifier.

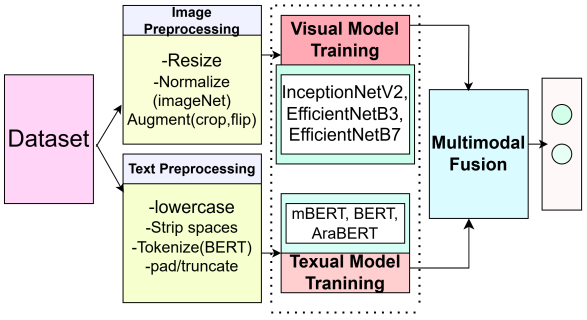


Figure 2: An overview of the methodology for our proposed system

4.1 Data Preprocessing

The experiments employed the publicly available QCRI/Prop2Hate-Meme dataset, hosted on the Hugging Face repository. The dataset contains multimodal entries, each consisting of a meme image and associated text, annotated for binary hate speech classification ($hate\_label \in \{0, 1\}$). The dataset was supplied in `Parquet` format and presplit into training, validation, and test sets.

Text preprocessing: For each meme text, we first measured its length in characters and recorded both the maximum and median values. Based on these lengths, the data was divided into three intervals to analyze distribution patterns. To reduce noise, extremely long non-hateful samples (those exceeding 62 characters) were excluded, while all hateful samples were kept to preserve minority class information. Texts were converted to lowercase, unnecessary spaces were removed, and the content was tokenized using the BERT tokenizer. Finally, each sequence was padded to a maximum length of 128 tokens.

Image preprocessing: All meme images were decoded from their byte format using the PIL library and then resized to a resolution of 128 × 128 pixels to maintain consistency. The resized images were converted into NumPy arrays, and their pixel values were normalized to fall within the range [0, 1]. Finally, the processed images were stored as stacked NumPy arrays, allowing efficient loading. All meme images were first decoded from their

byte format using the PIL library and then resized to a fixed resolution of 128 × 128 pixels to maintain consistency across samples. The resized images were converted into NumPy arrays, and their pixel values were normalized to fall within the range [0, 1] for stable model training. Finally, the processed images were stored as stacked NumPy arrays, allowing efficient loading and batch processing during training.

4.2 Feature Extraction

Text was encoded with the bert-base-multilingual-cased transformer. After tokenization, batches of 32 samples were forwarded through the model, and the `[CLS]` representation from the hidden layer was taken as the sentence-level embedding. The resulting vectors were stored as NumPy arrays of shape $(N, 768)$ for downstream use. Text was encoded with the bert-base-multilingual-cased transformer. After tokenization (max length 128 with padding/truncation), batches of 32 samples were forwarded through the model, and the `[CLS]` representation from the final hidden layer was taken as the sentence-level embedding. The resulting vectors were stored as NumPy arrays of shape $(N, 768)$ for downstream use.

Images were processed with a pre-trained InceptionResNetV2 backbone (ImageNet weights) with the classification head removed (include_top=False). Preprocessed inputs of size 128×128 were passed through the network, and a Global Average Pooling layer yielded a 1536-dimensional descriptor per image. This descriptor was flattened to obtain a fixed-length visual feature vector. The text and image embeddings produced here serve as inputs to the subsequent multimodal fusion module.

4.3 Baselines

Different unimodal models (image only and text only) and multimodal models (combining image and text) were analyzed and fused using an **early fusion** strategy with appropriate hyperparameter tuning.

4.3.1 Unimodal Baselines

To extract textual features, we utilized AraBERT, mBERT, and BERT. For visual features, we experimented with InceptionResNetV2, EfficientNetB3, and EfficientNetB7. The effectiveness of these models was assessed before integration into the multimodal framework. Various deep learning models were employed to establish unimodal baselines. For textual feature extraction, we utilized AraBERT(Antoun et al., 2020), mBERT, and BERT. For visual features, we experimented with InceptionResNetV2, EfficientNetB3, and EfficientNetB7(Tan and Le, 2020). These models were individually trained and evaluated to assess their effectiveness before integration into the multimodal framework. Appendix B.1 outlines the hyperparameters configured for both the textual and visual unimodal models.

4.3.2 Multimodal Baselines

We adopt an early fusion strategy where the 1536-D image vector and the 768-D text embedding are concatenated to form a 2304-D joint representation. The fused vector is then passed through fully connected layers with ReLU activations (1024 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 128), with dropout (0.5) applied after the 512- and 128-unit layers to mitigate overfitting. Finally, a dense layer with two units and a softmax activation outputs the class probabilities for hateful vs. non-hateful.

We evaluated several multimodal combination models, including InceptionNet + mBERT, InceptionNet + BERT, InceptionNet + AraBERT, EfficientNetB3 + mBERT, and EfficientNetB7 + mBERT. The hyperparameters used in this work include learning rate, number of epochs, batch size, dropout rate, optimizer and activation function, as summarized in the appendixB.2.

5 Results Analysis

The proposed multimodal fusion approach was evaluated on the official test split of the QCRI/Prop2Hate-Meme dataset as part of the shared task hosted on CodaBench . Among the tested configurations, the best-performing setup, combining InceptionNet (Szegedy et al., 2016) with mBERT (Devlin et al., 2019), achieved a macro F1-score of 69%, outperforming several strong vision–language baselines commonly used in hate speech detection tasks. (Kiela et al., 2021; Gomez et al., 2020)

Table 1 provides a comparative evaluation of textual and visual models. Within the text-only approaches, AraBERT and mBERT demonstrated the highest performance, both reaching an F1-score of

[1]Source code available at: GitHub Repository

Approaches	Classifiers	F1	P	R	G
Textual only	**AraBERT**	**0.63**	0.61	0.69	0.64
	BERT-base-uncased	0.55	0.57	0.67	0.61
	mBERT	**0.63**	0.61	0.69	0.64
Visual only	EfficientNetB3	0.45	0.54	0.61	0.57
	EfficientNetB7	0.48	0.58	0.62	0.60
	InceptionNetV2	**0.58**	0.57	0.60	0.58

Table 1: Result comparison on validation data of unimodal models, where F1, P, R, and G represent F1-score, precision, recall, and the geometric mean of precision and recall, respectively

0.63. In the visual-only category, InceptionNetV2 achieved the best result with an F1-score of 0.58.

Classifiers	F1	P	R	G	F_F1
InceptionNet + AraBERT	0.59	0.56	0.70	0.62	0.59
InceptionNet + BERT-base-uncased	0.67	0.75	0.61	0.68	0.60
EfficientNetB3 + mBERT	0.56	0.60	0.53	0.55	0.56
EfficientNetB7 + mBERT	0.67	0.70	0.64	0.67	0.60
InceptionNet + mBERT(Proposed)	**0.69**	0.66	0.76	0.71	**0.63**

Table 2: Result comparison on validation data of multimodal models, where F1, P, R, G, and F_F1 denote F1-score, precision, recall, geometric mean of precision and recall, and official F1-score, respectively

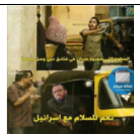
In multimodal comparative evaluations, as shown in Table 2, InceptionNet + AraBERT obtained 59%, InceptionNet + BERT-base-uncased (Devlin et al., 2019) reached 67%, EfficientNetB3 + mBERT scored 56%, and EfficientNetB7 + mBERT achieved 67% macro F-1 score. InceptionNet + mBERT achieved the highest score of 69% among all tested architectures, demonstrating its effectiveness in jointly leveraging visual and textual cues for hateful meme detection in Arabic content. The official shared task result was 63%, securing 6th place on the leaderboard.

5.1 Error Analysis

We present representative examples of both correctly and incorrectly classified hateful and non-hateful memes. Misclassifications often result from subtle visual cues, implicit expressions, or cases where the hateful intent is weak or context-dependent. The selected instances are summarized in Table 3, which lists the input meme, its true label, and the model's prediction.

Detailed error analysis is provided in the appendix C.

Table 3: Examples of predicted outputs

Input Meme	True Label	Predicted Label
	Hateful	Non-Hateful
	Non-Hateful	Non-Hateful
	Non-Hateful	Hateful
	Hateful	Hateful

6 Conclusion and Future Work

This study presented a multimodal fusion approach for detecting hateful content in Arabic memes by combining visual and textual information. Using the QCRI/Prop2Hate-Meme dataset, our best-performing configuration is InceptionNet with mBERT which achieved a **macro F1-score of 63%** in the official Codabench shared task evaluation. Results show that multimodal integration significantly outperforms unimodal models, especially where meaning depends on text-image interaction. Future work includes exploring advanced fusion methods, Vision-Language Models (VLMs), advance data augmentation to mitigate class imbalance, leveraging external context for subtle cues, and extending to multilingual scenarios.

Limitations

This work is limited to deep learning and transformer models, excluding traditional machine learning comparisons. The imbalanced dataset without advanced augmentation may have led to biased predictions. While multimodal fusion improved results, it also increased overfitting risks and computational costs.

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500.

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Za-

ghouani, and Georgios Mikros. 2024. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*, pages 380–390. Springer.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEe Access*, 12:22359–22375.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024. Araieval shared task: Propagandistic techniques detection in unimodal and multimodal arabic content.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: student research workshop*, pages 32–39.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer.

Chandni Kaundilya, Diksha Chawla, and Yatin Chopra. 2019. Automated text extraction from images using ocr system. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 145–150. IEEE.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020a. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020b. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Georgio Mikros, Abul Hasnat, and Firoj Alam. 2025. MAHED shared task: Multimodal detection of hope and hate emotions in arabic content. In *Proceedings of the Third Arabic Natural Language Processing Conference (Arabic-NLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Mohamed Zaytoon, Nagwa El-Makky, and Marwan Torki. 2024. AlexUNLP-MZ at ArAIEval shared task: Contrastive learning, LLM features extraction and multi-objective optimization for Arabic multimodal meme propaganda detection. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 512–517, Bangkok, Thailand. Association for Computational Linguistics.

Bryan Zhao, Andrew Zhang, Blake Watson, Gillian Kearney, and Isaac Dale. 2023. A review of vision-language models and their performance on the hateful memes challenge.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International conference on multimedia & expo workshops (ICMEW)*, pages 1–6. IEEE.

A Dataset Statistics

We follow the official train, development, and test splits without modification. The memes cover common Arabic social topics, including politics, public figures, religion, security, and social issues, making the dataset representative for real-world vision–language tasks. The training set contains 2,143 samples (1,930 non-hateful, 213 hateful), the development set has 312 samples (281 non-hateful, 31 hateful), and the test set includes 603 samples (452 non-hateful, 151 hateful).

Hate label	**Train**	**Dev**	**Test**	W_T	UW_T
Non_hateful	1930	281	452	36123	18059
Hateful	213	31	151	6484	4770
Total	2143	312	603	42607	22829

Table A.1: Class distribution of training, development, and test sets. W_T denotes total words and UW_T denotes unique words for each class.

Table A.1 shows that the dataset suffers from a notable class imbalance, with non-hateful samples dominating across all splits. This imbalance poses challenges for training reliable classifiers and highlights the need for robust evaluation strategies.

B Hyperparameter Setting

B.1 Unimodal Hyperparameters

Table B.1 summarizes the hyperparameters used for the textual models (AraBERT, mBERT, and BERT). The parameters include the learning rate, number of epochs, batch size, dropout rate, optimizer, and activation function. These values were selected after systematic tuning to achieve stable and consistent performance across the text-only experiments.

Hyperparameter	**AraBERT**	**mBERT**	**BERT**
Dropout rate	0.5	0.5	0.5
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001
Epochs	16	16	20
Batch size	32	32	32

Table B.1: Hyperparameters used for training the textual models

Hyperparameter	**InceptionNet**	**EfficientNetB3**	**EfficientNetB7**
Dropout rate	0.5	0.5	0.5
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001
Epochs	16	16	20
Batch size	32	32	32

Table B.2: Hyperparameters used for training the visual models

Table B.2 presents the hyperparameters adopted for the visual feature extraction networks (InceptionResNetV2, EfficientNetB3, and EfficientNetB7). Similar to the textual models, we optimized learning rate, epochs, batch size, dropout rate, optimizer, and activation function. The visual backbones were initialized with ImageNet weights, and the classification head was fine-tuned to adapt the features to our task.

B.2 Multimodal Hyperparameters

The selected hyperparameters are summarized in Table B.3.

Hyperparameter	Im	Ib	Ia	E3m	E7m
Dropout rate	0.5	0.5	0.5	0.5	0.5
Optimizer	Adam	Adam	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001	0.001	0.001
Epochs	16	16	20	16	16
Batch size	32	32	32	32	32

Table B.3: Hyperparameters used for training the multimodal models, where Im, Ib, Ia, E3m, and E7m denote InceptionNet + mBERT, InceptionNet + BERT, InceptionNet + AraBERT, EfficientNetB3 + mBERT, and EfficientNetB7 + mBERT, respectively

C Error Analysis

Both quantitative and qualitative error analyses were carried out to better understand the strengths and weaknesses of the best-performing model.

C.1 Quantitative Analysis

Table C.1 presents the analysis of our model and shows that it achieved an overall accuracy of 85%, with particularly strong results for the non-hateful class (F1-score of 0.91). For the hateful class, which represented the minority category, precision was lower at 0.36, though recall remained relatively high at 0.65. Given the dataset's class imbalance, macro F1-score (0.69) was used as the primary evaluation metric, as it equally weights both classes and provides a balanced performance measure. Table C.2 shows the official leaderboard, the system maintained strong generalization, with a macro F1-score of 63% on the unseen test set.

Class	Precision	Recall	F1-score	Support	G_score
non_hateful	0.96	0.87	0.91	281	0.91
hateful	0.36	0.65	0.46	31	0.48
macro avg	0.66	0.76	**0.69**	312	0.69
weighted avg	0.90	0.85	0.87	312	0.87
accuracy	-	-	0.85	312	-

Table C.1: Class-wise performance report on the validation set of the best-performing model

The confusion matrix in Figure C.1, provides a clear view of how the model distinguishes between hateful and non-hateful memes.The model correctly identified 267 non-hateful memes and 4 hateful memes. However, it misclassified 14 non-hateful memes as hateful and failed to detect 27 hateful memes.From a qualitative perspective, the fusion approach performs well when both text and image contribute useful and complementary information. In such cases, subtle textual hints combined with strong visual signals enable the model to correctly identify hateful content, where unimodal baselines often fail.

Position	Team_Name	Macro F1-Score (%)
1	NYUAD	80
2	yassirea	75
3	mzaytoon	74
4	itbaan	72
5	annasshaikh2003	68
6	**joy_2004114**	**63**

Table C.2: Leaderboard standings for the task

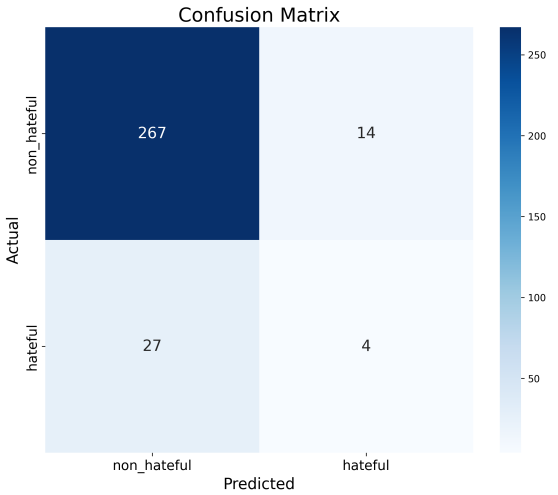


Figure C.1: Confusion matrics of best model

However, the model is not flawless. When one modality introduces misleading or ambiguous information, the fusion method can still occasionally succeed, but it is also vulnerable to misclassification. These errors highlight that while multimodal fusion strengthens overall performance, it remains sensitive to noise or imbalance in either modality.

C.2 Impact of Class Imbalance

The dataset's strong class imbalance (only ∼10% hateful samples) impacted the model's ability to maintain high precision for the hateful class. Although our model improved hateful recall compared to several baselines, targeted class balancing or augmentation strategies could further enhance performance in future work.

C.3 Qualatitative Analysis

Figure presents sample outputs from the model, illustrating both correct and incorrect classifications. In the first image, labeled as hateful, the model predicted non-hateful, likely due to its focus on explicit textual features while overlooking subtle visual cues indicating offensive intent. In the second image, labeled as non-hateful, the model misclassified it as hateful, reflecting sensitivity to certain visual or textual patterns that resemble hateful content. A major factor contributing to these errors is the significant class imbalance in the dataset, which biases the model toward dominant classes and limits its ability to generalize to underrepresented hateful samples. These findings highlight the need for improved context-aware multimodal modeling and strategies to mitigate imbalance effects.