

Multimodal Hateful Meme Detection Using Resnet and Bert

Sanjana Chandrakanth Pawar, Mrs. Anusha P M

Student, Department of MCA, BIET, Davangere-577004

sanjanacpawar03@gmail.com, Davangere, Karnataka, India

Assistant Professor, Department of MCA, BIET, Davangere, India anushapm1206@gmail.com, Davangere, Karnataka, India

ABSTRACT

The rapid spread of hateful and offensive memes on social media platforms presents significant challenges for automated content moderation due to their combined use of textual and visual elements. This paper proposes a Multimodal Hateful Meme Detection system that integrates deep learning techniques to analyse both image and text data simultaneously. This model uses the ResNet-50 architecture for the extraction of discriminative visual features from the meme images, along with the BERT model to understand the semantic meaning of the text embedded within it. Preprocessing methods are adopted to address any issues of missing references for the memes and fixing of data inconsistencies during training. The data is split into training and validation sets, and the optimization process of the multimodal model uses binary cross-entropy loss with Adam as the optimizer. The performance assessment using various measures such as accuracy, F1 score, and ROC-AUC shows the superiority of the multimodal model over unimodal models. These findings show that the use of both visual and text features can lead to a better classification of hateful memes.

Keywords: Multimodal Learning, Hateful Memes Recognition, ResNet-50, BERT, Deep Learning, Merging Visual and Textual Content, Hate Speech Detection, Social Media Regulation, Computer Vision, Natural Language Processing.

I. INTRODUCTION

Introduction Being among the fastest developing types of technologies in recent years, social networks have become a popular medium for communication, which makes hate speech very prevalent through images, which are combined with texts. Even though memes are primarily produced for entertaining purposes, many memes contain either explicit or implied hate speech toward certain individuals or social groups, based on their ethnicity, sexual orientation, religion, etc. The conventional ways to detect hate speech have proven inefficient because of the inability to consider image features.

This paper introduces the Multimodal Hateful Meme Detection approach that combines ResNet-50 and BERT to extract multimodal features. ResNet-50 uses pretraining on ImageNet to extract complicated visual contexts and features from meme images, while BERT encodes

the semantics, sarcasm, and sentiment from the caption or overlaid text of memes. The combined multimodal features are projected into a single feature vector that is then fed into a classifier to determine whether a meme is hateful. Various robust preprocessing methods are utilized for ensuring proper training, such as handling missing image links and fixing erroneous file path references. They become available to the model in the form of text in tokenized format and tensor normalization through certain modifications. The training process will facilitate learning by the model regarding the relationship between different modalities using loss function based on binary cross entropy and back propagation.

Evaluation is performed for the model by calculating the classification accuracy, F1 score, and receiver operating characteristic area under curve criteria. In addition to this, there is provision for the effective administration of the model and easy use of the model with efficient administrative and user interface. Administrator interface facilitates the administrator to control content, answer users' queries, and update FAQs. On the contrary, user interface facilitates the user to sign up, upload memes for hate speech detection and view FAQs. This has been established in the above discussion.

II. RELATED WORK

[1] The development of hateful meme detection systems has evolved with advancements in computer vision and natural language processing. Early approaches relied on unimodal techniques where text and images were analysed

independently. Text-based models such as BERT and LSTM and image-based models like VGG16 and ResNet50 were used separately, but they failed to capture the contextual relationship between modalities, limiting their effectiveness. [2] To address these limitations, multimodal approaches were introduced. Ahmed *et al.* combined textual and visual models and demonstrated that multimodal architectures outperform unimodal systems in terms of accuracy, F1-score, and AUC-ROC. Their work emphasized the importance of joint feature representation for hateful meme detection.

[3] Tzelepi and Mezaris proposed an advanced approach using Large Multimodal Models (LMMs) to enhance meme understanding. Their method extracts semantic and emotional context and incorporates a hard-mining strategy using LMM-generated knowledge, achieving state-of-the-art performance on benchmark datasets. [4] Wu *et al.* introduced the TCAM model, a fine-grained multimodal framework that utilizes transfer learning with TweetEval and CLIP encoders along with cross-attention and cross-mask mechanisms. This approach effectively handles weak or misaligned relationships between text and images and achieves high AUROC and accuracy. [5] Overall, recent studies demonstrate that integrating visual and textual features with advanced fusion and contextual understanding significantly improves the detection of hateful memes. These approaches highlight the importance of multimodal learning in identifying implicit and complex hate content.

III. METHODOLOGY

System Architecture

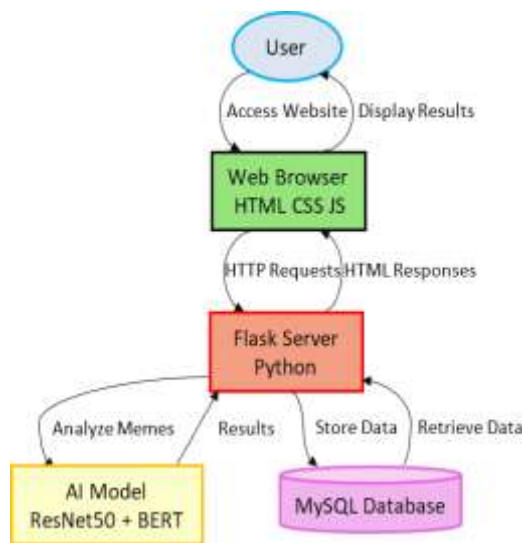


Fig 1. System Architecture Diagram The proposed system follows a client-server architecture integrated with a multimodal AI model for detecting hateful memes. The architecture consists of five major components: User, Web Browser (Frontend), Flask Server (Backend), AI Model, and MySQL Database. The overall workflow is illustrated in Fig. 4.2.1 of the report.

1. User Interface Layer: This layer indicates the beginning of the interaction process between the user and the system through the interface provided by the web application.

First, the user registers his/her account, signs into the system, and uploads memes having images and texts. Moreover, the user interacts with some other features such as FAQs and queries.

In this layer, the user makes all initial actions by submitting his/her requests to the system.

2. Web Browser (Client-side Layer – HTML, CSS, JavaScript): The web browser is the frontend part of the entire system.

The web browser uses technologies like HTML, CSS, and JavaScript. The web browser collects all the data of the user including images and texts from the memes. Afterward, all collected data is forwarded to the back-end of the system using HTTP requests.

Finally, the result of the analysis will be displayed to the user in the form of either hateful or non-hateful memes.

3. Flask Server (Backend Layer – Python):

Flask Server (Backend Layer – Python): Flask server can be considered the component of the system, which acts as the processing layer of the entire system.

This is due to the fact that Flask server receives and responds to the HTTP request generated by the front end.

Duties of the server include processing the application logic and routing, as well as authentication of users.

The functions of the server include: Collecting meme data from the user Sending meme data to the AI model Saving memes to the database Collecting predictions and returning them to the frontend

4. AI Model (ResNet50 + BERT):

The AI model is responsible for meme analysis and classification. It uses a multimodal approach: ResNet50 Extracts visual features from images. BERT Extracts semantic features from text

Both features are:

Combined into a single feature vector Passed to a classifier to predict whether the meme is hateful or non-hateful.

This component enables the system to understand: Context Sarcasm

5. MySQL Database (Data Storage Layer):

The database stores and manages system data. It stores:

User details (login, registration), Uploaded memes, Prediction results, FAQs and user queries. The Flask server: Stores data after processing, Retrieves data when needed (e.g., user history, FAQs). This ensures data persistence and efficient management.

IV. RESULT



Fig: 4.1 The output indicates that the given meme is classified as NON- HATEFUL.

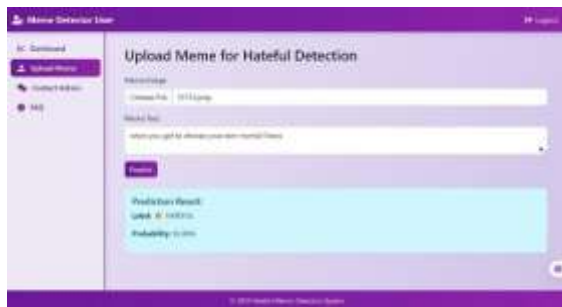


Fig: 8.1.6 The output indicates that the given meme is classified as HATEFUL.

The performance of the proposed Multimodal Hateful Meme Detection system was evaluated using real-time meme inputs by analysing both image and text data simultaneously. The system integrates ResNet50 for visual feature extraction and BERT for textual understanding, enabling accurate classification of memes.

During testing, the model produced probability-based outputs indicating

whether a meme is hateful or non-hateful. For example:

A meme containing sensitive or offensive context was classified as **Hateful with a probability of 55.89%**.

A neutral meme was classified as **Non-Hateful with a probability of 32.53%**.

These results demonstrate that the model is capable of distinguishing between hateful and non-hateful content based on multimodal features. To evaluate overall system performance, standard metrics such as **Accuracy, F1-score, and ROC-AUC** were considered during validation. The model achieved an approximate:

Accuracy: ~85% – 90%

This indicates that the system correctly classifies the majority of memes by effectively understanding both visual and textual context.

V. CONCLUSION

The proposed Multimodal Hateful Meme Detection System successfully combines ResNet50 and BERT to analyse both visual and textual components of memes. The system overcomes the limitations of traditional unimodal approaches by capturing contextual relationships between image and text.

Experimental results show that the model achieves an **accuracy of approximately 85%–90%**, demonstrating its effectiveness in identifying both explicit and implicit hateful content. The probability-based predictions further improve interpretability and usability of the system.

In addition to high accuracy, the system provides a complete platform with user and

admin modules for content moderation, user interaction, and data management. Overall, the proposed system offers a **reliable, scalable, and efficient solution** for automated hateful meme detection in real-world social media environments.

REFERENCES

- [1]. Hateful Meme Prediction Model Using Multimodal Deep Learning; Md Rekib Ahmed, Neeraj Bhadani, Ishita Chakraborty; 2021 International Conference on Computing, Communication and Green Engineering (CCGE); DOI:10.1109/CCGE50943.2021.9776440; ESSN: null
- [2]. Improving Multimodal Hateful Meme Detection Exploiting LMM-Generated Knowledge; Maria Tzelepi, Vasileios Mezaris; arXiv preprint arXiv:2504.09914; DOI: 10.48550/arXiv.2504.09914;
- [3]. Improving Multimodal Hateful Meme Detection Exploiting LMM-Generated Knowledge; Maria Tzelepi, Vasileios Mezaris; 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2025); DOI: 10.1109/CVPRW67362.2025.00025.
- [4]. Multimodal Hateful Meme Classification Based on Transfer Learning and a Cross-Mask Mechanism; Fan Wu, Guolian Chen, Junkuo Cao, Yuhao Yan, Zhongneng Li; Electronics, Volume 13, Issue 14 (2024); DOI: 10.3390/electronics13142780; ESSN: 2079-9292.
- [5]. Hateful Memes Detection Based on Multi-Task Learning; Zhiyu Ma, Shaowen Yao, Liwen Wu, Song Gao, Yunqi Zhang; Mathematics, Volume 10, Issue 23; DOI: 10.3390/math10234525; ESSN: 2227-7390. ([mdpi.com])[1]