# Offensive Meme Classification Using Multimodal Sentiment Analysis Applying NLP Techniques

Poroma Biswas
*Department of Computer Science and Engineering*
*BRAC University*
Dhaka, Bangladesh
poroma.biswas@g.bracu.ac.bd

Ashabul Yamin Raad
*Department of Computer Science and Engineering*
*BRAC University*
Dhaka, Bangladesh
ashabul.yamin.raad@g.bracu.ac.bd

Hasnat Md. Imtiaz
*Department of Computer Science and Engineering*
*BRAC University*
Dhaka, Bangladesh
hasnat.md.imtiaz@g.bracu.ac.bd

Fardin Zaman
*Department of Computer Science and Engineering*
*BRAC University*
Dhaka, Bangladesh
fardin.zaman@g.bracu.ac.bd

***Abstract*—Our research delves into the multifaceted impact of internet memes, traditionally seen as humorous expressions but increasingly used for abusive and cyberbullying purposes. Unlike previous research focusing on specific attributes, we take a broader approach, addressing the challenge of identifying these cryptic and humorous digital artifacts. Utilizing a dataset centered around religiously offensive memes, our research employs multimodal simple neural network models demonstrating high accuracy. Our aim is to contribute insights to the ongoing discourse on mitigating the spread of hatred through memes. By unraveling their complexities, we aim to provide a deeper understanding for individuals, researchers, and online platforms grappling with the challenges posed by these dynamic digital expressions.**

## I. Introduction

In the dynamic landscape of the internet, memes emerge as powerful conduits for communicating a diverse array of socio-cultural, psychological, and political ideas. While conventionally regarded as lighthearted and humorous expressions, a noticeable surge in the prevalence of detrimental memes has become apparent, particularly concerning their deployment for abusive, trolling, and cyberbullying purposes. The inherent challenge lies in the cryptic and humorous nature of these memes, rendering their identification a formidable task.

Existing research has primarily fixated on dissecting specific meme attributes, such as hate speech and propaganda. However, there has been a conspicuous oversight in addressing the broader spectrum of harm associated with memes. This study seeks to rectify this gap by conducting a thorough examination of the extensive range of adverse effects linked to memes.

Our research extends beyond the confines of existing studies by building upon a recent dataset that predominantly centers around memes featuring religiously offensive content. This dataset serves as a crucial foundation for our investigation. We applied this dataset to three distinct types of models, each demonstrating commendable accuracy in capturing and understanding the nuances of these complex digital expressions.

Ultimately, our overarching objective is to contribute valuable insights to the ongoing discourse on mitigating the dissemination of hatred on the internet facilitated through memes. By unraveling the multifaceted dimensions of memes and their potential negative implications, we aim to equip individuals, researchers, and online platforms with a more comprehensive understanding of the challenges posed by these digital artifacts.

## II. Literature Review

This paper [1] undertakes a critical examination of the challenging task of identifying religiously hateful memes on social media within the vision-language domain. To address this issue, the authors construct a dedicated dataset of religiously hateful memes and propose a methodology that involves the refinement and comparison of the VisualBERT model for the classification task. Additionally, the dataset is expanded through integration with the Facebook hateful memes dataset.

The study acknowledges and discusses the inherent limitations of deep learning-based models, particularly in managing inductive bias, offering proposed solutions to mitigate these challenges. Employing a multimodal approach, the authors amalgamate visual features extracted through ResNeXT-152 Aggregated Residual Transformations based Masked Regions with Convolutional Neural Networks (R-CNN) and textual encoding via Bidirectional Encoder Representations from Transformers (BERT). This integrated approach yields a notable AUCROC score of 78 %, underscoring the model's efficacy in classifying religiously hateful memes. Preprocessing steps are outlined, encompassing the removal of images lacking textual content and the elimination of unwanted symbols from textual captions. Emphasis is placed on the pivotal role of pre-training and fine-tuning datasets in achieving optimal results.

The utilization of the BERTBASE model for encoding textual inputs is justified, given its provision of context-dependent embeddings, thereby enhancing sentence comprehension. In summary, this paper offers a thorough and systematic approach to the classification of religiously hateful memes, addressing the limitations of deep learning models and showcasing promising outcomes.

This paper [2] introduces the MOMENTA framework, a cutting-edge multimodal deep neural network designed for the detection of harmful memes and the identification of their targets. MOMENTA systematically examines memes from both local and global perspectives in both visual and textual modalities, establishing connections with the background context. Outperforming several strong rivaling approaches, the framework demonstrates superior accuracy in the detection of harmful memes and the identification of their targets.

Leveraging techniques such as multimodal deep neural networks, unimodal and multimodal models, and multimodal pretraining, MOMENTA's effectiveness and generalizability are substantiated through extensive experiments on two substantial datasets, Harm-C which denotes Covid 19 memes and Harm-P which denotes US Presidential Election memes. The paper emphasizes the practical implications of the research, including the identification and detection of harmful memes, the discernment of their targets, and the availability of large-scale datasets for future investigations.

The MOMENTA framework's transferability and interpretability render it applicable to diverse domains and languages, further establishing its utility and significance. Overall, this paper presents a thorough review of the MOMENTA framework, underscoring its comprehensiveness, effectiveness, and practical implications in the realm of harmful meme detection and target identification.

## III. Methodology

### A. Dataset Collection

The dataset consists of two main components: a 'labels.txt' file and an 'images' folder. The 'labels.txt' file likely contains information regarding the classification or labeling of memes within the dataset, specifying whether each meme is categorized as hateful or not. The 'images' folder, on the other hand, contains the actual meme images for analysis. This dataset is designed for the task of hateful meme classification, where the objective is to develop a model capable of accurately identifying and categorizing memes as either hateful or non-hateful based on the provided labels. Such datasets play a crucial role in training machine learning models to understand and differentiate between various types of content on the internet, contributing to efforts in content moderation and online safety.

### B. Data Pre-processing

*1) Text Pre-processing :* The text preprocessing is done on the labels for the hateful meme classification dataset. It begins by reading a 'labels.txt' file containing information about meme images, including their IDs, paths, labels, and associated texts. The script utilizes regular expressions to extract these details and then constructs a Pandas data frame to organize the data. The text preprocessing function is defined to convert text to lowercase, remove punctuation, tokenize, and eliminate common English stop words. This function is applied to the 'text' column of the DataFrame, enhancing the quality of textual data for machine learning tasks. The final data frame reflects the preprocessed data, ready for subsequent steps in the classification pipeline. The code employs popular natural language processing libraries such as NLTK and Pandas to facilitate efficient data handling and text manipulation.

TABLE I
Data Frame

| id | image-path | label |
|---|---|---|
| 12301 | image/12301.png | 1 |
| 12302 | image/12302.png | 1 |
| 12303 | image/12303.png | 1 |
| 12304 | image/12304.png | 1 |
| 12305 | image/12305.png | 1 |

*2) Image Pre-processing :* For preprocessing of the images, the Keras library is utilized to load and preprocess image data for a hateful meme classification dataset. A function takes a list of image paths as input and processes each image by loading it, resizing it to (224, 224) pixels, and normalizing pixel values to a range of [0, 1]. The resulting image arrays are appended to a list, and a NumPy array is constructed by vertically stacking these arrays. The function returns the processed image data.

The function is applied to the column of the previously created DataFrame ‘df. The base path for the images is specified as a default argument. During processing, the code checks if each image file exists and issues warnings if any problems, such as missing files or incorrect image shapes, are encountered.

This image preprocessing step is crucial for preparing the image data for input into a deep learning model, ensuring uniformity in size and pixel values. The normalization of pixel values enhances the convergence and performance of the model during training.

### C. Combining Text and Image Features

The integration of both text and image features for a multimodal hateful meme classification is done. It employs the TF-IDF vectorizer from scikit-learn for text feature extraction and combines it with preprocessed image data.

The TF-IDF vectorizer is configured to consider a maximum of 5000 features. Text features are extracted and converted into a dense array. The image data is reshaped to have two dimensions. To ensure the same number of samples for both text and image features, the code matches the number of samples by slicing the features accordingly.

The final step involves combining text and image features into a single array. The data is then split into training and testing sets, with 80 percent of the data used for training and 20 percent for testing. The resulting feature arrays and corresponding labels are printed, and the total number of features in the combined dataset is displayed.

The dataset is prepared for training a multimodal classification model that considers both textual and visual information for predicting hateful meme labels. The integration of text and image features is essential for leveraging the strengths of both modalities in the classification task.

### D. Multimodal Model

The model is demonstrated through the construction, compilation, training, and evaluation of a simple neural network using the Keras library for the multimodal classification task. The model architecture consists of three layers: two densely connected layers with rectified linear unit (ReLU) activation functions, and a final output layer with a sigmoid activation function, suitable for binary classification.

The model is compiled using binary cross-entropy as the loss function and the Adam optimizer, while accuracy is chosen as the metric for evaluation. The training process involves fitting the model to the training data for 100 epochs, with a batch size of 16 and a validation split of 20

Finally, the trained model is evaluated on the test set, and the test accuracy is printed.

## IV. Result and Discussion

The experimental evaluation of the proposed multimodal classification model was conducted through a systematic process involving construction, compilation, training, and evaluation. The model was compiled with binary cross-entropy as the loss function and the Adam optimizer. Accuracy was selected as the evaluation metric to gauge the model's performance.The trained model underwent evaluation on the test set, and the test accuracy was printed. The achieved accuracy on the test set was found to be 76.8 percent. This indicates a notable level of success in the model's ability to correctly classify instances in the binary classification task. The selected architecture, loss function, and optimization strategy contributed to the model's efficacy in learning and generalizing patterns from the training data to the unseen test data. It's noteworthy that the chosen metrics and hyperparameters played a crucial role in determining the model's performance. The utilization of binary cross-entropy loss ensured that the model effectively learned the binary classification task. The Adam optimizer, known for its efficiency in handling large datasets and non-stationary objectives, facilitated a more robust convergence during training. While the achieved accuracy of 76.8 percent is commendable, further investigation is warranted to explore avenues for improvement. Future work could involve fine-tuning hyperparameters, exploring more complex neural network architectures, and increasing the diversity and size of the dataset. Additionally, a more in-depth analysis of misclassifications and model interpretability could provide valuable insights into potential areas for refinement.

TABLE II
Model Result

| Model | Accuracy |
|---|---|
| Multimodal Neural Network | 0.768 |

## V. Conclusion

This research signifies a critical step towards unraveling the intricate landscape of harmful memes on the internet. We have shed light on the multifaceted dimensions of memes, particularly those with negative implications by addressing the gap in existing literature and taking a comprehensive approach.

As we move forward, the insights gained from this study can serve as a foundation for developing more robust models and strategies to mitigate the dissemination of hatred through memes. While acknowledging the limitations, we remain optimistic about the potential impact of our research on fostering a safer and more informed digital environment. Ultimately, this work contributes to the ongoing discourse surrounding harmful memes, providing valuable perspectives for individuals, researchers, and online platforms grappling with the challenges posed by these complex digital artifacts.

## References

[1] Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Amanullah Yasin, Gautam Srivastava, Dawid Połap, Thippa Reddy Gadekallu, and Zunera Jalil. 2023. Multimodal Religiously Hateful Social Media Memes Classification based on Textual and Image Data. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (September 2023). https://doi.org/10.1145/3623396

[2] Pramanick, Shraman, et al. ”MOMENTA: A multimodal framework for detecting harmful memes and their targets.” arXiv preprint arXiv:2109.05184 (2021).