Beyond the Laughter: Detecting Hateful Memes with Deep Learning

Rafeed Mahbub Rafi, Alistair Biswas, Syed Mohtasib Mashruk, Md. Fahim Faisal, Nusrat Islam,
Md. Tanvir Rouf Shawon and Rayhan Tanvir
Department of Computer Science and Engineering,
Ahsanullah University of Science and Technology, Dhaka, 1208, Bangladesh.

***Abstract*—The paper presents an in-depth study on the detection of hateful memes using deep learning models across visual, textual, and multimodal domains. The study assesses the performance of several models, including VGG16, VGG19, ResNet50, BiLSTM + CNN, BanglaBERT, and M-BERT, as well as combinations of these models in a multimodal context. Notably, the BiLSTM + CNN model in the textual modality got the highest overall performance with an accuracy of 66.58% and an F1 score of 66.12%. In the multimodal scenario, the combination of VGG16 and BanglaBERT produced an accuracy of 63.22% and an F1 score of 63.82%. The results highlight the crucial relevance of textual information in hateful meme detection and the potential of leveraging both visual and textual information. The article also analyzes the limits of the existing approach, including class imbalance and fundamental preprocessing approaches, and recommends potential avenues for future research. The insights acquired from this work could pave the way for more effective and robust systems for recognizing and fighting online hate speech.**

***Index Terms*—Hateful meme detection, deep learning, textual modality, visual modality, multimodal analysis.**

I. Introduction

Memes are a popular type of online communication, commonly utilized for humor, amusement, and social commentary. However, some memes are developed for harmful objectives, such as propagating hate, ignorance, and propaganda. These memes are known as nasty memes, and they can have harmful repercussions on individuals and society, such as instigating violence, prejudice, and radicalization.

Hateful memes are difficult to detect mechanically, as they rely on subtle and intricate connections between visuals and words. For example, a meme may utilize a benign image with a caustic or ironic phrase, or a hateful image with a seemingly neutral or positive caption, to express a hidden or ambiguous meaning. Moreover, the meaning of a meme may rely on the cultural, historical, and situational context of the originator and the audience.

The main purpose of this research is to present a hateful meme identification system utilizing deep learning. The objective of this study is to confront the challenge of online hate speech, which poses a severe threat to social harmony, human dignity, and democratic values. By designing a system that can accurately and transparently detect offensive memes, we seek to contribute to the creation of effective and ethical tools for countering online hate speech.

The remaining section of the paper has been arranged as follows. Section II examines the related study on hateful meme detection and explainable AI. Section III describes the MUTE [9] dataset. Section IV discusses the procedure in detail. Section V offers the experimental outcome, Section VI includes a brief performance analysis, Section VII describes the limitations of the study and outcome, and Section VIII wraps up the paper with a brief overview of the research and future prospects.

II. Related Work

Machine learning researchers have been extensively studying hate speech and hateful memes, using different models and datasets.

Kiela et al. [3] provide the Hateful Memes Challenge and dataset in more detail, and examine the performance of multiple baseline models on the dataset. It also explores the ethical and societal consequences of the challenge and the dataset, and makes recommendations for future research.

Das et al. [1] combine visual embedding model [9] with text processing using RNNs to detect hate speech in memes using multimodal information (text and image) using the Facebook Hateful Memes Challenge dataset [3] with ”benign confounders” to challenge unimodal models. The multimodal model outperforms unimodal models, but still struggles with some benign confounders.

Lippe et al. [2] propose a cross-validation ensemble of pre-trained transformer model [10] to develop a multimodal framework for hate speech detection in memes using the same dataset [3] which achieves good accuracy, but interpretability and generalization are lacking.

Pramanick et al. [5] use both unimodal [14], [15], [16], [17], [18] and multimodal models [4], [10], [19] on their presented dataset HarMeme [12] to define the notion of harmful memes. The multimodal systems effectively identify harmful memes and predicted target groups.

Cao et al. [6] present a novel approach to explain hatefulness and humor separately using causal inference framework with BERT and image encoders on the Facebook Hateful Memes Challenge dataset [3].

Deshpande and Mani [7] suggest an interpretable approach to hateful meme detection, employing machine learning and basic heuristics to determine the traits most crucial to identifying a meme as hateful. It compares a gradient-boosted decision tree and an LSTM-based model with human and transformer models on the Hateful Memes dataset.

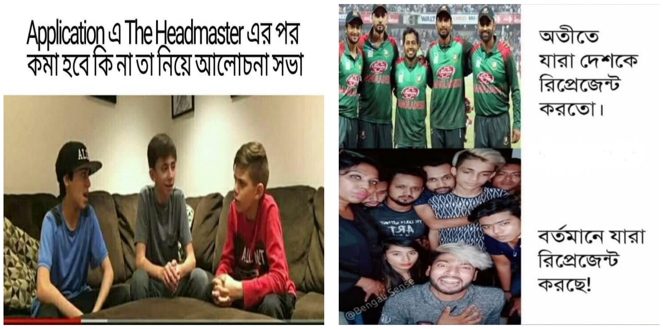
Hee et al. [8] explain multimodal hate speech detection models using SHAP and counterfactual explanations, using VisualBERT [10] model for multimodal analysis on Facebook Hateful Memes Challenge dataset [3]. It provides insights into model reasoning through SHAP and identifies potential biases through counterfactuals.

Miyanishi and Nguyen [9] analyze hate speech in memes considering intersectionality and causal factors using a novel dual gradient descent method. The methodology of the work includes a gradient descent framework with interaction analysis to identify intersectional hate features. It shows the importance of considering intersectionality in hate speech detection and offers causal explanations for hate features.

III. Dataset

A. Data Collection

For our research, we gathered MUTE dataset, a multimodal dataset consisting of 4,369 images annotated with ’hate’ and ’not-hate’. Figure 1 shows an example of the images from the dataset.


Application এ The Headmaster এর পর কমা হবে কি না তা নিয়ে আলোচনা সভা
অতীতে যারা দেশকে রিপ্রেজেন্ট করতো।
বর্তমানে যারা রিপ্রেজেন্ট করছে!
@Bengali Sense

(a) Non-hateful

(b) Hateful

Fig. 1: Example of Bengali Memes (a) A meme that is not hateful. (b) A meme representing hatred based on identity factors.

B. Data Annotation

The MUTE dataset is separated into three sets for training and evaluation: the train set (80%), the test set (10%), and the validation set (10%). Table I presents the distribution of the dataset categorized by class. The dataset exhibits a small imbalance, with the ’Not-Hate’ class containing around 60% of the data. Table II provides the statistics of the training set, which reveal that the ‘Not-Hate’ class has a greater quantity of words and unique words than the ‘Hate’ class. However, the average caption length is practically comparable in both classes. We also did a quantitative research utilizing the Jaccard similarity index to determine the fraction of shared phrases between the classes. We obtained score of 0.498, which demonstrates the presence of shared common words between the classes.

TABLE I: Dataset Summary

Dataset	**Hate**	**Not-Hate**
Train	1275	2090
Validation	152	223
Test	159	257
Total	**1586**	**2570**

TABLE II: Training set statistics for the caption

	Hate	**Not-Hate**
#Words	16,900	30,684
#Unique words	6,036	8,916
Max. caption length	63	106
Avg. #words/caption	13.25	14.68

IV. Methodology

In this section, we describe the data pre-processing, the models, and the evaluation metrics that we used for our project.

A. Data Pre-processing

The section focuses on a systematic and comprehensive pre-processing of the data to prepare it for subsequent analysis. The training, validation, and testing datasets are sourced and the labels in the ’Label’ column of the datasets are transformed into numerical values. Specifically, ’hate’ is encoded as 1, while ’not-hate’ is encoded as 0. All image names are aggregated from a specified directory, and paths for each image in the dataset are generated. The images are loaded and standardized to a uniform size of 128x128 pixels and then converted to an array. The ’Captions’ column in the datasets undergoes cleaning by removing punctuations using the NLTK library in python. Then we had tokenized the captions for the textual and multimodal models.

B. Classification

We applied visual, textual, and multimodal approaches individually. We used VGG-19, VGG-16, and ResNet50 as our primary image-based models, utilizing them to classify the memes using only the image data. Each model began with pre-trained weights to leverage prior visual knowledge. Subsequently, they were fine-tuned using our dataset to adapt to the specific characteristics of the Bengali memes. Then in our textual analysis, we employed BanglaBERT, BiLSTM with CNN and M-BERT. BanglaBERT is a transformer-based model specifically designed for bengali sentences. Then for multimodal approach, we employed a combination of VGG16 and BanglaBERT. The model fuses the outputs of VGG16 and BanglaBERT using a linear layer and then feeds them to a classifier layer that predicts the sentiment polarity of the input. It can leverage both visual and textual cues to capture the nuances of multimodal sentiment expression in Bangla.

C. Evaluation

We used four evaluation measures: Accuracy, Precision, Recall and F1-score. For the measures, higher values are better,

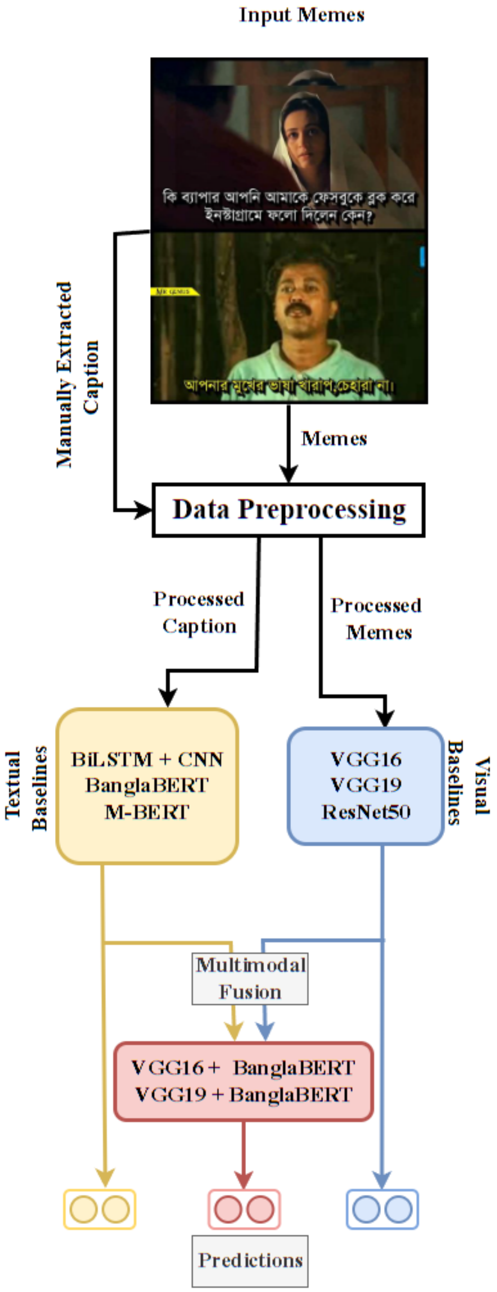


Fig. 2: Experimental Setup.

as they indicate that the classifier is more accurate, precise, sensitive, and balanced in predicting the correct classes.

V. Experiment Results

Table III presents the performance metrics for Hateful Meme Detection across different modalities and models. The evaluation metrics include Accuracy (A), Precision (P), Recall (R), and F1 Score (F1). The results provide insights into the effectiveness of each modality and model combination in accurately identifying hateful memes.

TABLE III: Performance for Hateful Meme Detection

Modality	Model	A	P	R	F1
Visual	VGG16	61.29	61.63	61.29	61.44
	VGG19	59.37	59.52	59.37	59.44
	ResNet50	61.29	61.63	61.29	61.44
Textual	BiLSTM + CNN	66.58	65.93	66.58	66.12
	BanglaBERT	61.77	38.16	61.77	47.18
	M-BERT	61.77	38.16	61.77	47.18
Multimodal	VGG16 + BanglaBERT	63.22	64.44	63.22	63.82
	VGG19 + BanglaBERT	54.80	62.28	54.80	54.52

VI. Performance Analysis

The evaluation of different models for the detection of hostile memes was conducted using four metrics: Accuracy (A), Precision (P), Recall (R), and F1 Score (F1). The models underwent testing in three different modalities: Visual, Textual, and Multimodal.

Three models were evaluated in the visual domain: VGG16, VGG19, and ResNet50. VGG16 and ResNet50 demonstrated similar performance, earning the greatest accuracy and F1 score of 61.29% and 61.44% respectively. In the textual modality, the BiLSTM + CNN model outperformed both BanglaBERT and M-BERT, earning an accuracy of 66.58% and an F1 score of 66.12%. Figure 3 and Figure 4 demostrates the confusion matrix and ROC curve of the BiLSTM + CNN model.

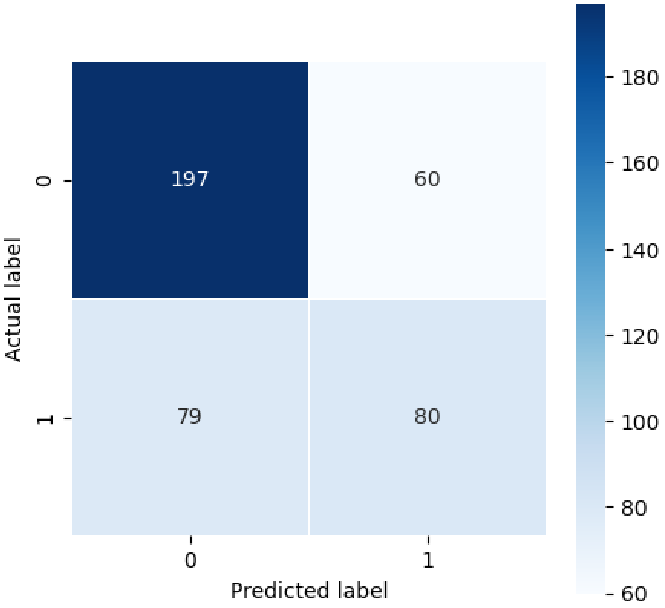


Fig. 3: Confusion matrix of BiLSTM + CNN model.

It is important to mention that although BanglaBERT and M-BERT achieved the same level of accuracy, their precision was much lower, suggesting a larger occurrence of false positives. The multimodal category yielded the greatest accuracy and F1 score of 63.22% and 63.82% respectively when

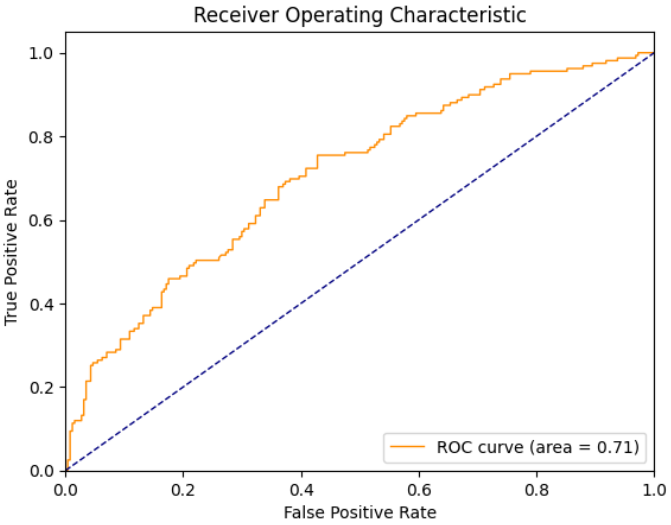


Fig. 4: ROC curve of BiLSTM + CNN model.

VGG16 and BanglaBERT were combined. The amalgamation of VGG19 and BanglaBERT, albeit exhibiting superior precision, yielded inferior overall accuracy and F1 score. Among all the other approaches, the BiLSTM + CNN model performed the best in the textual modality, with an accuracy of 66.58% and an F1 score of 66.12%. These findings indicate that when it comes to identifying hateful memes, the textual content may hold greater importance than the visual elements. Nevertheless, the encouraging outcomes derived from the multimodal models suggest that the integration of visual and textual data can potentially yield favorable results for this objective.

VII. Limitations and Future Work

One of the key reasons behind the unremarkable outcomes of the utilized models may have been the balance of the dataset. The dataset employed in this work is uneven, and hence, this imbalance may bias the models towards predicting the 'Not-Hate' class, thereby impacting the performance of the models. The preprocessing processes comprised simple cleaning and tokenization of the captions. However, no advanced word embedding techniques were applied, which could limit the ability of the models to grasp semantic links between words.

Moreover, all photos in the collection were normalized to a uniform size of 128x128 pixels. This could result in loss of detail, especially for photographs that were originally of a higher resolution.

Future studies could explore with advanced word embedding techniques, such as Word2Vec or GloVe, to better capture the semantic links between words. Efforts might be made to gather a more varied and representative dataset for training the models. This could involve gathering memes from a larger number of sources, or employing approaches such as data augmentation to enhance the size of the dataset. Exploring the impact of image resolution on model performance could involve testing the models on photos of various resolutions, or experimenting with approaches for maintaining detail when shrinking images. Development of more advanced multimodal models that better integrate visual and textual information could involve exploring state-of-the-art techniques in deep learning, such as attention mechanisms or transformer models, which may offer more nuanced interpretations of the complex interplay between image content and captions in memes.

The study of sentimental analysis of memes could also open the door for more ambitious studies, such meme production utilizing powerful generative models like GAN. Finally, the insights gathered from this study, coupled with the prospective pathways for future research, could pave the way for more effective and resilient systems for identifying and countering online hate speech and address reversal problems using generative AI.

VIII. Conclusion

In this study, we addressed the challenging task of hateful meme detection using a range of models across visual, textual, and multimodal domains. Our findings imply that while each modality delivers distinct insights, the mix of modalities can lead to improved performance.

The BiLSTM + CNN model in the textual modality produced the highest overall performance, highlighting the crucial relevance of textual information in hateful meme detection. However, the positive results from the multimodal models underline the potential of harnessing both visual and textual information for this job.

Despite various limitations such as class imbalance and basic preprocessing approaches, this study provides a good platform for future research in the field of hateful meme detection. Future work could explore approaches for addressing class imbalance, experiment with improved word embedding techniques, and test the impact of image resolution on model performance.

In conclusion, the insights gathered from this work, coupled with the prospective possibilities for future research, could pave the way for more effective and robust systems for recognizing and fighting online hate speech. As we continue to advance in this field, we move closer to establishing a safer and more inclusive digital space for all.

References

[1] A. Das, J. S. Wahi, and S. Li, "Detecting Hate Speech in Multi-modal Memes," *arXiv preprint*, 2020.

[2] P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, and H. Yannakoudakis, "A Multimodal Framework for the Detection of Hateful Memes," *arXiv preprint*, 2020.

[3] D. Kiela et al., "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," in *Thirty-Fourth Annual Conference on Neural Information Processing Systems*, 2020.

[4] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, "Supervised Multimodal Bitransformers for Classifying Images and Text," *ArXiv*, 2019.

[5] S. Pramanick et al., "Detecting Harmful Memes and Their Targets," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2783–2796.

[6] R. Cao et al., "Disentangling Hate in Online Memes," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5138–5147.

[7] T. Deshpande and N. Mani, ”An Interpretable Approach to Hateful Meme Detection,” in *ICMI ’21: Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 723–727.

[8] M. S. Hee, R. K.-W. Lee, and W.-H. Chong, ”On Explaining Multimodal Hateful Meme Detection Models,” in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3651–3655.

[9] Y. Miyanishi and M. L. Nguyen, ”Causal Intersectionality and Dual Form of Gradient Descent for Multimodal Analysis: a Case Study on Hateful Memes,” *ArXiv*, 2023.

[10] L. H. Li et al., ”Visualbert: A simple and performant baseline for vision and language,” *arXiv preprint arXiv:1908.03557*, 2019.

[11] Y.-C. Chen et al., ”Uniter: Universal image-text representation learning,” in *European Conference on Computer Vision*, 2020, pp. 104–120.

[12] D. Dimitrov, ”HarMeme,” *GitHub*, http://github.com/di-dimitrov/harmeme, 2021.

[13] H. Eftekhar, O. Sharif, and M. M. Hoque, ”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*, 2022, pp. 32–39.

[14] J. Devlin et al., ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[15] K. Simonyan and A. Zisserman, ”Very deep convolutional networks for large-scale image recognition,” in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[16] G. Huang et al., ”Densely Connected Convolutional Networks,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[17] K. He et al., ”Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] S. Xie et al., ”Aggregated Residual Transformations for Deep Neural Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987–5995.

[19] J. Lu et al., ”In Proceedings of the Conference on Neural Information Processing Systems,” in *NeurIPS ’19*, 2019, pp. 13–23.