A Multimodal Framework to Detect Target Aware Aggression in Memes

Shawly Ahsan[♣]*, Eftekhar Hossain[♠]*, Omar Sharif[♣], Avishek Das[♣],
Mohammed Moshiul Hoque[♣], M. Ali Akber Dewan[¥]
[♣]Department of Computer Science and Engineering
[♠]Department of Electronics and Telecommunication Engineering
[♣ ♠]Chittagong University of Engineering & Technology, Bangladesh
[¥]School of Computing and Information Systems, Athabasca University, Canada
u1704057@student.cuet.ac.bd, {eftekhar.hossain, moshiul_240}@cuet.ac.bd

Abstract

Internet memes have gained immense traction as a medium for individuals to convey emotions, thoughts, and perspectives on social media. While memes often serve as sources of humor and entertainment, they can also propagate offensive, incendiary, or harmful content, deliberately targeting specific individuals or communities. Identifying such memes is challenging because of their satirical and cryptic characteristics. Most contemporary research on memes' detrimental facets is skewed towards high-resource languages, often sidelining the unique challenges tied to low-resource languages, such as Bengali. To facilitate this research in low-resource languages, this paper presents a novel dataset **MIMOSA** (MultIMOdal aggreSsion dAtaset) in Bengali. MIMOSA encompasses 4,848 annotated memes across five aggression target categories: Political, Gender, Religious, Others, and non-aggressive. We also propose `MAF` (Multimodal Attentive Fusion), a simple yet effective approach that uses multimodal context to detect the aggression targets. `MAF` captures the selective modality-specific features of the input meme and jointly evaluates them with individual modality features. Experiments on **MIMOSA** exhibit that the proposed method outperforms several state-of-the-art rivaling approaches. Our code and data are available at https://github.com/shawlyahsan/Bengali-Aggression-Memes.

1 Introduction

Recently, the rise of social media has given prominence to a distinct multimodal phenomenon known as *meme*, a composition of an image coupled with concise textual content. While memes are often humorous, they can propagate hate, offense, and aggression by incorporating political or cultural elements. Such undesired memes pose a significant threat to social harmony, as they can potentially harm individuals or specific groups based on their political philosophy, sexual orientation, religious beliefs, and more.

(a) (b)

Figure 1: Example of aggressive memes: (a) A meme directly undermining a religion (b) A meme deliberately trying to foster a popular political person as a hypocrite.

As memes have become crucial in influencing social interactions, there has been a notable rise in research focused on meme analysis. This research includes analyzing the emotions (Mishra et al., 2023) conveyed in memes, sarcastic memes detection (Bandyopadhyay et al., 2023), and offensive memes detection (Zhou et al., 2021). The emergence of highly toxic memes has prompted research efforts to explore their negative aspects, such as hate (Kiela et al., 2020), offensiveness (Shang et al., 2021), and harm (Pramanick et al., 2021b). However, most works have focused on the memes of high-resource languages while only a few studied the objectionable (i.e., hate, aggression, offense) memes of low-resource languages (Kumari et al., 2023; Suryawanshi and Chakravarthi, 2021).

Bengali memes have gained significant traction recently, reaching a broad audience and influencing public opinion while promoting negativity and violence. Detecting objectionable Bengali memes is currently in the developing stage due to the limited availability of tools such as OCR. Nonetheless, two works (Karim et al., 2022; Hossain et al., 2022b) accomplished on detecting Bengali hateful memes. Research in this domain (both high-resource and low-resource) has highlighted that the exploration

*Denotes equal contribution

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 2487–2500

of the darker aspects of memes often overlooks the term *'aggression'*, which carries a more explicit and virulent connotation than *'harm'* or *'offense'*. To illustrate, consider the meme depicted in Figure 1 (a);. At the same time, it may be perceived as harmful, a comprehensive analysis of its textual and visual context categorizes it as aggressive due to its explicit undermining of a religious group. Moreover, aggressive meme identification requires separate analysis as it is more target-aware (i.e., religious, political, and gendered) than hate and offense. Considering the pernicious impact of aggression, developing systems to identify aggressive memes and their targets is essential.

With the motivation mentioned above, we develop a novel corpus of Bengali memes encompassing various levels of aggression. On the technical front, prior studies reveal that state-of-the-art multimodal systems, effective in many visual-linguistic tasks, struggle with meme analysis. Memes rely heavily on context and often lack a clear connection between visual and textual elements. Moreover, memes contain much noise, making them distinct from other, more structured multimodal data. To tackle these issues, we develop a multimodal attentive fusion-based model to identify the targets of aggression within these memes. Our significant contributions are as follows.

- We develop a novel multimodal aggression dataset **MIMOSA** consisting of 4,848 Bengali memes labeled with four aggression (Political, Gendered, Religious, and Others) and one non-aggressive class.
- We propose `MAF`, a simple yet effective multimodal fusion approach that utilizes the attentive multimodal representation of the input meme and the individual modality-specific features to learn the subtle aggression elements better.
- Finally, we perform extensive experiments on **MIMOSA** and show that `MAF` outperforms eleven state-of-the-art unimodal and multimodal baselines in terms of all the evaluation measures.

2 Related Work

This section demonstrates the previous studies that have already been conducted on objectionable content (i.e., hate, offense, and aggression) detection based on unimodal and multimodal content.

Unimodal Based Objectionable Content Detection: Most research on objectionable content detection (OCD) focused on analyzing textual data. Over the years, the topic has become a prominent research issue among researchers of different languages (Ross et al., 2017; Lekea and Karampelas, 2018). Several works focused on developing new corpus for various languages (Schneider et al., 2018; Niraula et al., 2021) while others studied to introduce novel methods (Sharif et al., 2021; Sreelakshmi et al., 2020) for OCD. Some works were also performed concerning low-resource languages. Sharif and Hoque (2022) introduced the first dataset for identifying target-aware aggression from Bangla texts. Likewise, two aggression datasets were introduced by Bhattacharya et al. (2020) and Ranasinghe and Zampieri (2021), which cover other low-resource languages like Spanish, Turkish, Greek, and so on.

Various methods were employed over the years for hate, aggression, and offense detection. Earlier studies used machine learning (Sreelakshmi et al., 2020) and recurrent neural network (Sharif and Hoque, 2021; Sadiq et al., 2021) based approaches. Later, transformer-based methods(Kamal et al., 2021; Sharif and Hoque, 2022; Baruah et al., 2020) achieved superior performance for OCD. Apart from the above research, few studies were performed for objectionable content detection from the visual data. For example, identifying the violent objects (Gandhi et al., 2020), nudity (Lin et al., 2021), aggression (Hs et al., 2021), and trolling (Hs et al., 2021) from the images.

Multimodal Based Objectionable Content Detection: In contrast to only text and image-based OCD, several works have been accomplished considering the multimodal information in recent years. Suryawanshi et al. (2020) developed a multimodal dataset for offensive meme detection. Both Kiela et al. (2020) and Gomez et al. (2020) introduced two multimodal datasets for hate speech from online memes. Recently, Pramanick et al. (2021a) introduced a multimodal dataset for harmful memes detection in the context of the COVID-19 pandemic. In recent years, studies have been on multimodal-based OCD for resource-limited languages. Karim et al.

(2022) and Hossain et al. (2022b) developed two multimodal hate speech datasets concerning the Bangla language. Two multimodal datasets are also developed in the Hindi language by Kumari et al. (2023) and Rajput et al. (2022) for identifying offensive and hateful memes. Over the years, several methods have been introduced to detect offense, hate, and harm from the multimodal data. Earlier, researchers used different fusion (Hossain et al., 2021, 2022c; Hasan et al., 2022) strategies, while in recent years, transformer architectures (Kiela et al., 2020) such as MMBT, Visual BERT, ViLBERT, CLIP have been employed. However, these models have broadly applied to the English language, thus limiting their capability to perform highly in resource-constraint languages.

Differences with existing researches: While there has been significant progress in multimodal hate speech and offensive content detection, a notable gap exists in the research landscape regarding multimodal aggression detection, especially in low-resource languages (i.e., Bengali). Our investigation revealed that only two works (Karim et al., 2022; Hossain et al., 2022b) have studied the multimodal data in Bengali. However, they were primarily centered around hate speech detection. It is worth noting that aggression, distinct from hate or offense, has been relatively underexplored in the context of multimodal analysis (Kocoń et al., 2021). Furthermore, most existing datasets in this domain focus on binary classifications (either hateful or not hateful) without delving into the specific targeted entities, such as political, gendered, and religious themes, which can often provide more information about the content. In light of these identified gaps, our work differs from the existing works in three significant ways: (i) develops a multimodal aggression dataset specifically tailored for Bengali, with a focus on internet memes; (ii) instead of treating aggression as a singular construct, we break down the task into distinct dimensions such as political, gendered, religious aggression, others and non-aggression (iii) provides a detailed annotation guideline that can aid in resource creation for other low-resource languages.

3 MIMOSA: A New Benchmark Dataset

Per our exploration, no benchmark dataset is explicitly developed for identifying aggression and its targets from the multimodal data. To fill this void, we developed **MIMOSA:** a novel target-aware multimodal aggressive memes dataset in Bengali. To create *MIMOSA*, we followed the guidelines provided by the Hossain et al. (2022a,b). This section briefly describes the dataset development process, including data accumulation and annotation guidelines.

3.1 Defining Aggressive Meme

Following existing works on aggression detection (Kumari et al., 2021; Sharif and Hoque, 2021), this work defines *aggressive memes* as *multimodal units that include an image with text embedded in it and have the potential to physically threaten, attack, or seek to harm a person, group, or community based on political ideology, religious belief, sexual orientation, gender, race, and nationality, or contain nudity, sexually explicit content, objects used to inspire violence*.

Aggressive memes can be offensive or hateful, but not all offensive or hateful memes represent aggression. Offensive content (Suryawanshi et al., 2020) is defined as any disrespectful, insulting, or inappropriate material and frequently includes abusive or derogatory language. However, unlike aggressive content, offensive content does not always involve direct threats or physical harm. On the contrary, hateful memes (Kiela et al., 2020) contain image and text that promotes discrimination, prejudice, or animosity toward a specific race, ethnicity, religion, gender, or sexual orientation and are fueled by extreme bias against specific groups. As opposed to aggressive memes, hateful content targets entities based on personal attributes.

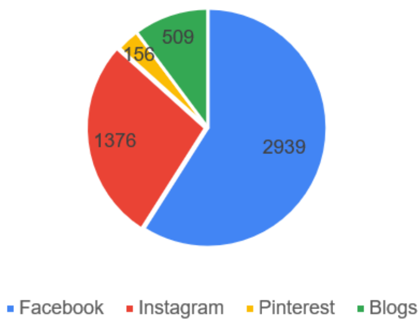


Figure 2: Distribution of data sources. Each cell represents the number of samples collected from the corresponding sources.

3.2 Data Collection

We have collected memes from various social media platforms and online sources to create the dataset. To ensure representativeness and reduce biases to a particular source, we collected data from diverse sources (e.g., Facebook, Instagram, Pinterest, and different Bengali Blogs). Figure 2 depicts the number of memes collected from each source. Most memes were collected from Facebook and Instagram, while a few were accumulated from Pinterest and blogs.

A set of keywords such as *"Bengali Memes," "Bengali Funny Memes," "Bengali Offensive Memes," "Bengali Aggressive Memes," "Bengali Troll Memes," "Bengali Political Memes," "Bengali Political Troll Memes," "Bengali Feminism Troll Memes," "Bengali Islam Troll Memes," "Bengali Hinduism Troll Memes," and "Bengali Celebrity Troll Memes"* were used to search the memes. We used neutral keywords not explicitly tied to specific aggression themes to reduce biases to any specific category. Despite our best efforts, the dataset may have inherent biases, a common challenge in the development process.

We collected the memes only from public domains, social media pages, and groups to avoid copyright infringement. Through this search process, 4,980 memes were collected from March 2022 to February 2023. During the data accumulation period, we have discarded memes that fall under the following categories: (i) memes that have information from only one modality (either visual or textual), (ii) memes that contain cartoons (as AI systems often face difficulty to process them), and (iii) memes that are visibly unclear (blurred). Figure A.2 illustrates some filtered samples. We discarded 132 memes based on the above criterion and finished with a total of **4,848** memes. Afterward, we extract the meme caption using an OCR[1]. However, we manually checked the extracted captions to correct any missing words and spelling as OCR in Bengali is not well-established. Finally, the memes and their associated captions are forwarded to the annotators to start the annotation process.

3.3 Data Annotation

MIMOSA was manually labeled into five categories: four aggression targets categories (political aggression (PAg), religious aggression (RAg), gendered aggression (GAg), others (Oth)) and a non-aggressive (NoAg) category. A detailed definition of each category was supplied to the annotators to ensure consistency and quality in the MIMOSA data annotation process. Figure A.1 shows examples from each category.

3.3.1 Definition of Categories

After reviewing existing works on aggression detection (Kumari et al., 2021; Gasparini et al., 2022; Sharif and Hoque, 2021), this work settled on the following class definitions:

1. **Political Aggression (PAg):** Memes that provoke followers of political parties, condemn political ideology, or excite people in opposition to the state, law, or enforcing agencies are termed political aggression.

2. **Religious Aggression (RAg):** Memes used to incite violence by attacking religion, religious organizations, or the religious beliefs of a person or a community are considered religious aggression.

3. **Gendered Aggression (GAg):** Memes that promote aggression or attack the victim based on gender or contain aggressive reference to one's sexual orientation, body parts, sexuality, or other lewd content, nudity, or sexually explicit content are considered gendered aggression.

4. **Others (Oth):** Memes that express aggression but do not fall under any of the above aggression classes are termed as others. The *Others aggression* class includes the targets based on race, occupation, education, disability, nationality, geography, etc.

5. **Non-aggressive (NoAg):** Memes that do not contain any statement of aggression or express a hidden wish or intent to harm others are included in this category.

3.3.2 Process of Annotation

The annotators were asked to adhere to the class definitions to ensure labeling consistency. Initially, the annotators were asked to determine whether the meme was aggressive or non-aggressive based on the class definition. If an aggressive meme is discovered, they were instructed to further categorize it into one of the specific aggression targets. The annotators were also asked to provide reasoning for annotation decisions, which the expert will

[1] https://pypi.org/project/pytesseract/

Class	Train	Validation	Test
NoAg	846	181	182
PAg	597	128	128
RAg	618	133	132
GAg	672	144	144
Oth	660	141	142
Total	3393	727	728

Table 1: Number of data in train, validation, and test sets

use as a reference in cases of disagreement. Initially, the annotators were trained with a small set of memes before being given a more extensive set to annotate independently. The training assisted in familiarizing the annotators with the task and ensuring consistency in their decisions. Three annotators (computer science undergraduates) each performed manual annotation, and the labels were verified by an expert (a professor with more than 20 years of research experience in NLP). More details of the annotators and the annotation process are provided in the Appendix B. To assess annotation quality, we used inter-annotator agreement metrics like Cohen's kappa coefficient (Cohen, 1960). Our study achieved a Cohen's kappa coefficient of **0.86**, considered almost perfect agreement on the kappa scale.

3.4 Dataset Statistics

For model training and evaluation, the dataset is divided into train (70%), validation (15%), and test (15%) sets. Table 1 depicts the class-wise data distribution of each set. Furthermore, we analyzed the captions of the training set, and Table 2 presents the summary. We noticed that the 'RAg' meme captions have a rich vocabulary and are typically longer than other categories. On the other hand, 'GAg' class captions have the lowest number of unique words (3,163) and the average words per caption (12). In contrast, no significant variation in information is observed in the remaining categories (NoAg, PAg, Oth). We further analyze each category's caption length frequency distribution in the training set shown in Figure 3. We observed that most captions are concise as they are 4 to 25 words long. However, many captions have more than 20 words, implying that some meme captions contain more detailed and elaborate context information.

Class	T_{tw}	T_{uw}	T_{mw}	T_{aw}
NoAg	11257	3813	41	13
PAg	9687	4078	48	16
RAg	11139	4552	61	18
GAg	8307	3163	49	12
Oth	8526	3713	39	13

Table 2: Summary of the training set, where T_{tw}, T_{uw}, T_{mw}, and T_{aw} denotes the number of total words, unique words, maximum words per caption, and average words per caption, respectively)

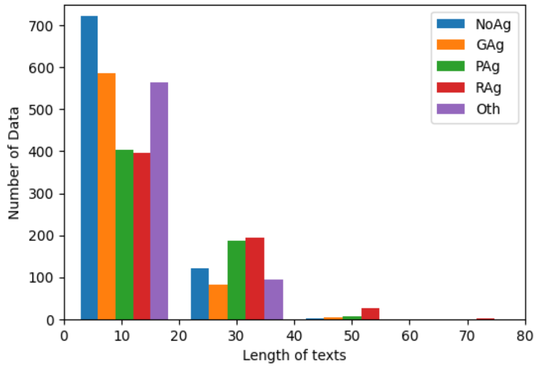


Figure 3: Caption length (in words) distribution for the training set.

	NoAg	GAg	PAg	RAg	Oth
NoAg	-	**0.24**	0.17	0.18	0.22
GAg	-	-	0.16	0.17	0.22
PAg	-	-	-	0.16	0.17
RAg	-	-	-	-	0.18
Oth	-	-	-	-	-

Table 3: Jaccard similarity score between the captions of each class

Apart from the above analysis, we measured quantitatively using the Jaccard similarity index to see how many words overlapped across the categories. Table 3 indicates that the highest similarity (0.24) exists between the 'NoAg' and 'GAg' classes, while other classes did not show any significant variation in similarity score.

4 Methodology

This section describes the proposed multimodal framework for target-aware aggression identification. The system takes memes and their corresponding caption as input. We employed state-of-the-art models to encode the memes' visual and textual information. Afterward, we use an attentive fusion mechanism to create a multimodal representation by selectively focusing on the encoded visual and textual features. Figure 4 shows the overall architecture of the proposed framework.

4.1 Visual and Textual Features Extraction

To encode the visual information of the memes, we use the image encoder of a pre-trained visual-

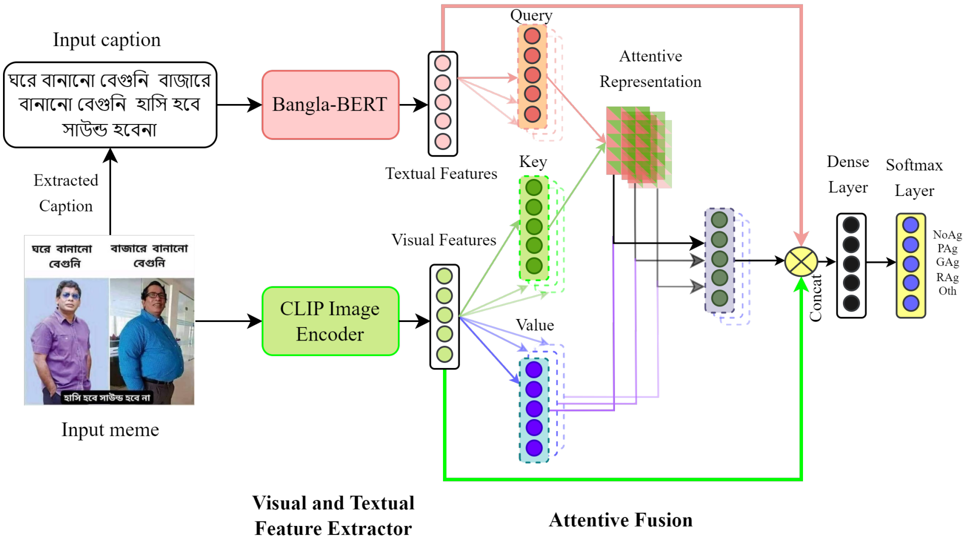


Figure 4: Proposed Multimodal Attentive Fusion (MAF) framework for target aware aggressive meme detection. MAF takes the meme and its corresponding caption as input

linguistic model named CLIP (Contrastive Language–Image Pretraining) (Radford et al., 2021). Though CLIP uses a Vision Transformer (Dosovitskiy et al., 2020) as a backbone in the image encoder, it is compelling compared to other transformer-based vision models (Liu et al., 2021; Bao et al., 2021) as pretraining was performed on millions of noisy image-text pairs from the internet. Similarly, we employed the Bangla-BERT (Sarker, 2020), a language model specifically pre-trained on millions of Bengali texts to extract the textual features. We fine-tuned the image and text encoder for extracting the respective features. Specifically, the CLIP encoder gives an image representation of size 512, and BERT gives a contextualized vector representation of a caption of size 768. These two feature representations are then passed to the multi-head attentive fusion module for generating a multimodal representation.

4.2 Attentive Fusion and Prediction

To make a multimodal representation, the obtained visual and textual vector representations are fused using a multi-head self-attention (MSA) block (Vaswani et al., 2017). The MSA block takes three matrices: query (Q), key (K), and value (V) as input. In standard NLP applications, all the matrices come from the word representations. However, in this research, motivated by Lu et al. (2019), we modified the MSA block where queries come from one modality and keys and values from another. This modification will generate an attention-pooled representation for one modality conditioned on another. Specifically, we generate Q from textual features and K and V from visual features. Afterward, to determine the similarity between the visual and textual features, we calculated the attention values by performing a dot product between Q and K. Then we weighed the visual features using the attention values to get a multimodal representation. This process is intuitive; just like humans, they read the text first and then pay more attention to the image areas similar to the text. Afterward, the attentive multimodal representation is further concatenated with the individual modality features (obtained from CLIP and Bangla-BERT). This process will boost the gradient flow and help the model learn from individual features and their refined, combined representations. Finally, the concatenated multimodal representation is passed to the dense layer, followed by a softmax operation to predict the meme's categories.

5 Experiments

This section discusses the baselines and their performance comparison with the proposed method (MAF). We will also illustrate the proposed approach's superiority by examining the errors. To experiment with **MIMOSA**, we developed several state-of-the-art computational models, including unimodal visual models, unimodal textual models, and multimodal models pre-trained on both modalities. We use two primary metrics for the evaluation: weighted f_1-score (WF1) and macro-averaged mean absolute error (MMAE) (Baccianella et al., 2009). Appendix A presents the details of the experimental settings.

5.1 Baselines

To validate the performance of the proposed multimodal framework, we develop several models considering unimodal information (only visual or textual) and multimodal information (visual and textual).

5.1.1 Unimodal Baselines

For the unimodal visual-only models, we employed three well-known architectures: **ResNet50** (He et al., 2016), **Vision Transformer (ViT)** (Dosovitskiy et al., 2020), and **ConvNeXT** (Liu et al., 2022). Meanwhile, in the case of the unimodal textual-only models, three pre-trained transformer models, namely **Bangla-BERT** (Sarker, 2020), **multilingual BERT** (Devlin et al., 2019), and **XLMR** (Conneau et al., 2020) are used. All the unimodal models are fine-tuned on the developed dataset.

5.1.2 Multimodal Baselines

- **Early Fusion**: We combine the intermediate feature representations of ViT and the Bangla-BERT model for the early fusion approach.

- **Late Fusion**: The softmax prediction scores of the ViT and Bangla-BERT models are utilized to construct the late fusion model.

- **CLIP**: It is a multimodal model trained on noisy image-text pair using contrastive learning (Chen et al., 2020) approach. CLIP has been widely used for several multimodal classification tasks (Pramanick et al., 2021b; Kumar and Nanadakumar, 2022).

- **BLIP**: BLIP (Bootstrapping Language-Image Pre-training) (Li et al., 2022) is a recently developed state-of-the-art multimodal model.

- **ALBEF**: ALBEF (Align Before Fuse) (Li et al., 2021) is another state-of-the-art multimodal model that uses momentum distillation and contrastive learning method for the pre-training on noisy image-text data.

In the case of the CLIP and BLIP models, we extract the visual and textual embedding representations by fine-tuning them on the developed dataset. Afterward, we combined both representations and trained them on top of a softmax layer.

5.2 Results

Table 4 demonstrates the performance of various models (both unimodal and multimodal) for detecting target-aware aggressive memes. Among the visual-only unimodal models, ViT performs best, achieving a weighted f_1-score of 0.582, surpassing ResNet50 and ConvNeXT. However, the textual-only model, Bangla-BERT, outperforms all unimodal models with a weighted F1 score of 0.641. We observed that combining ViT and Bangla-BERT through an early fusion approach improves the model's performance (WF1) by approximately 4% compared to the best unimodal model (Bangla-BERT). Surprisingly, sophisticated multimodal models like CLIP, BLIP, and ALBEF fail to outperform the simple early fusion method. Many of these multimodal models are primarily pre-trained on English image-text pairs, limiting their effectiveness in low-resource languages.

However, the proposed method (`MAF`) stands out, achieving the highest performance (WF1 = 0.742) among all the models. It boasts an absolute improvement of 5.9%, 6.7%, and 14.2% in accuracy, weighted F1 score, and MMAE measurements, respectively, compared to the best baseline model (early fusion).

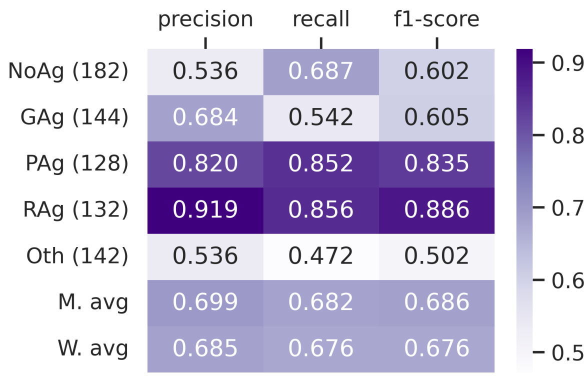
Ablation Study: Apart from this, to justify the effectiveness of the `MAF`, we removed some components from it. We presented their outcomes as the variants of `MAF`. The last four rows in Table

Approach	Model	Acc ↑	WF1 ↑	MMAE ↓
Visual Only	ResNet50	0.551	0.546	1.049
	ViT	0.601	0.582	0.967
	ConvNeXT	0.594	0.572	0.979
Textual Only	m-BERT	0.604	0.608	0.930
	B-BERT	0.646	0.641	0.811
	XLMR	0.585	0.572	0.903
Multimodal	Early Fusion	<u>0.682</u>	<u>0.675</u>	<u>0.787</u>
	Late Fusion	0.645	0.644	0.807
	CLIP	0.621	0.627	0.907
	BLIP	0.632	0.601	0.964
	ALBEF	0.627	0.622	0.906
Proposed System and Variants	`MAF` w/o VF	0.701	0.693	0.743
	`MAF` w/o TF	0.645	0.644	0.807
	`MAF` w/o VF+TF	0.694	0.696	0.735
	`MAF`	**0.741**	**0.742**	**0.645**
Δ_{MAF-BM}		5.9	6.7	14.2

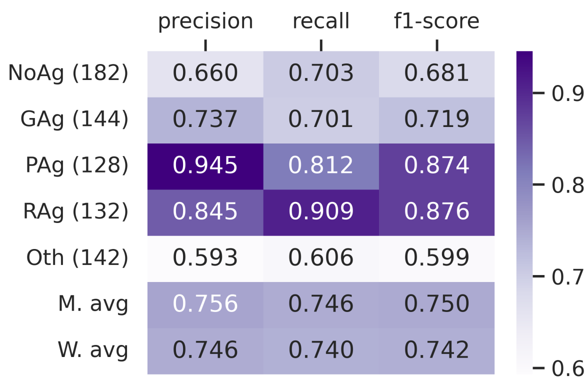
Table 4: Performance comparison of unimodal and multimodal baselines on the test set where Acc, WF1, and MMAE denote accuracy, weighted f_1-score, and macro-averaged mean absolute error. The best baseline score is underlined. The last row shows the performance improvement of the proposed system (`MAF`) over the best baseline model (Early Fusion). Here, VF and TF correspond to visual and textual features, respectively.

4 show the ablation outcomes. We observed that when we don't add the individual modality-specific features (VF or TF or both IF and TF) with the attentive vector, the performance drops up to 10%. This outcome illustrates how each component significantly improves the performance of `MAF`. We also performed an additional ablation study (presented in Appendix C) to illustrate how the number of attention heads impacts the model performance.

Classwise Models Performance: To see the performance across different aggression target classes, we further investigate the classification reports (shown in Figure 5) of the proposed method and compare it with the best baseline model (early fusion). We observed that in terms of f_1-score, the proposed method significantly improves across the 'NoAg'($\approx$ 8%$\uparrow$), 'GAg'($\approx$ 11%$\uparrow$), and 'Oth' ($\approx$ 10%$\uparrow$) classes compared to the baseline model. The proposed method achieved the highest f_1-score



(a) Best baseline model (Early Fusion)



(b) Proposed method MAF

Figure 5: Classwise performance comparison between the best baseline model (early fusion) and the proposed method regarding precision, recall, and weighted f_1-score. M.avg denotes the macro average, whereas W.avg corresponds to the weighted average.

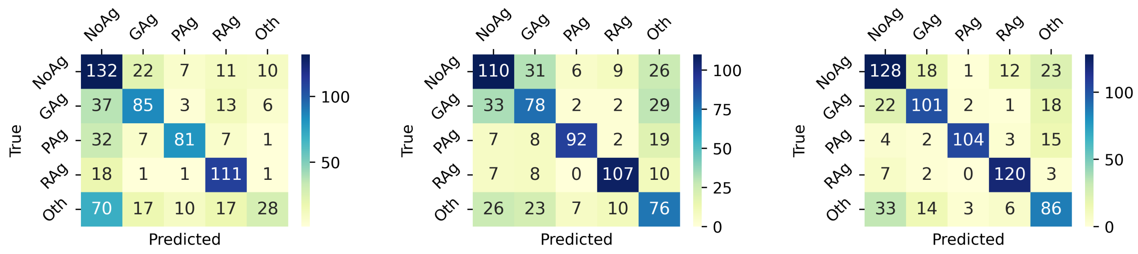
(0.874) and the precision values (0.945) with the 'PAg' class. Overall, with the proposed method, the precision and recall scores in all the classes are significantly higher than in the baseline models. This outcome further demonstrates the efficacy of the proposed method in identifying the targets of aggressive memes.

5.3 Error Analysis

The results showed that the proposed `MAF` is superior in identifying the targets of aggressive memes more accurately compared to the only visual and textual approach. However, to examine the mistakes of the proposed method, we perform a detailed error analysis using quantitative and qualitative ways. We also consider the best visual and textual models for better demonstration.

Quantitative Analysis: To perform quantitative analysis, we use the confusion metrics of the models shown in Figure 6. It is observed that the visual model struggles to correctly classify the 'PAg' and 'Oth' classes compared to the textual model. Moreover, the visual model gets confused with the 'NoAg' class as most of the samples (157) from different classes are misclassified as 'NoAg.' In contrast, the textual model improves the performance by reducing the number of misclassified samples from 114 to 66 in the 'Oth' aggression class. It also yields better performance in identifying the 'PAg' class. However, the proposed `MAF` proved superior by reducing the misclassification rate in almost all classes. Compared to the unimodal approaches, the proposed model `MAF` significantly improves the performance in the 'GAg,' 'PAg,' and 'Oth' classes. One important finding is that most misclassification occurred between the 'NoAg,' 'GAg,' and 'Oth' classes by the `MAF`. This misclassification might be because these classes have overlapping words, as evident from the Jaccard similarity score in Table 3. Besides, we also noticed that the misclassification rate is minimal in the case of the 'GAg,' 'PAg,' and 'RAg' classes, which suggests that our proposed method is good at distinguishing these aggression targets. In summary, visual information is more appropriate for identifying non-aggressive memes, whereas textual data is enough to detect religiously aggressive memes. However, the proposed `MAF` is more effective in obtaining a balanced optimum performance across all the classes.

Qualitative Analysis: We examined some correctly and incorrectly classified memes (shown



(a) Best visual (ViT) (b) Best textual (Bangla-BERT) (c) Proposed method (MAF)

Figure 6: Confusion matrices of the best visual, textual, and proposed multimodal models

in Figure 7) to further investigate the proposed model's mistakes. In the case of Figure 7 (a) textual

(a) **Textual:** NoAg (✗) **Visual:** Oth (✗) **MAF:** GAg (✓)

(b) **Actual:** NoAg **Predicted:** GAg

Figure 7: Example (a) illustrates a meme where the proposed method produces better predictions, and example (b) illustrates a wrongly classified sample. The symbol (✓) and (✗) indicates the correct and incorrect prediction

model incorrectly classified the meme as 'NoAg', whereas the visual model considered it as an aggressive meme but from a different class ('Oth'). However, the proposed model MAF captures the visual and textual relation correctly and identifies it as a Gendered Aggressive ('GAg') meme. However, in some cases, our proposed method can not capture the nuanced context of the memes. For instance, the meme in Figure 7 (b) shows the usual visual content; however, due to some gendered related term in the text part, the proposed method might get confused and yield a false prediction.

6 Conclusion

This paper presented a novel multimodal dataset, **MIMOSA**, consisting of 4,848 memes, for detecting the targets of Bengali aggressive memes into five classes. This research also proposed a multimodal deep neural network MAF for the downstream task. Experiments on **MIMOSA** demonstrated the efficacy of MAF outperformed eleven state-of-art unimodal and multimodal baselines. We plan to extend the dataset for more domains and languages. The future aim is to investigate the proposed model's performance on other datasets to enhance its generalization capabilities.

Limitations

Though the proposed method (MAF) demonstrates superior performance, there still exist some constraints. First, it is likely that in some cases, the MAF may focus on irrelevant parts of the visual and textual features during attentive fusion. For example, suppose the dataset contains misleading captions or irrelevant textual information. In that case, the attention mechanism might align with those parts of the image that are visually unrelated, leading to biased representations and thus providing suboptimal results. Second, upon analyzing the misclassified memes, we observed that the proposed MAF struggled with memes that contained subtle or sarcastic content. Furthermore, it appeared to have difficulty correctly interpreting cultural references and context-specific content, leading to additional incorrect predictions. To address these limitations, expanding the training data set must include a more comprehensive range of threatening objects and more examples of subtle or sarcastic content is critical.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *2009 Ninth international conference on intelligent systems design and applications*, pages 283–287. IEEE.

Dibyanayan Bandyopadhyay, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and Vinutha

BN. 2023. A knowledge infusion based multitasking system for sarcasm detection in meme. In *European Conference on Information Retrieval*, pages 101–117. Springer.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Aggression identification in English, Hindi and Bangla text using BERT, RoBERTa and SVM. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 76–82, Marseille, France. European Language Resources Association (ELRA).

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2020. Scalable detection of offensive and non-compliant content/logo in product images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2247–2256.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in Brief*, 44:108526.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Md Hasan, Nusratul Jannat, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Cuet-nlp@ dravidianlangtech-acl2022: Investigating deep learning techniques to detect multimodal troll memes. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 170–176.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021. Nlp-cuet@ dravidianlangtech-eacl2021: Investigating visual and textual features to identify trolls from multimodal social media memes. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 300–306.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022a. Memosen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022b. 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, Online. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, M Ali Akber Dewan, Nazmul Siddique, and Md Azad Hossain. 2022c. Identification of multilingual offense and troll from social media memes using weighted ensemble of multimodal features. *Journal* *of King Saud University-Computer and Information Sciences*, 34(9):6605–6623.

Chinmaya Hs et al. 2021. Trollmeta@ dravidianlangtech-eacl2021: Meme classification using deep learning. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 277–280.

Ojasv Kamal, Adarsh Kumar, and Tejas Vaidhya. 2021. Hostility detection in hindi leveraging pre-trained language models. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 213–223, Cham. Springer International Publishing.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 293–308. Springer.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Gokul Karthik Kumar and Karthik Nanadakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05916*.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2023. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, pages 1–36.

Kirti Kumari, Jyoti Prakash Singh, Yogesh K Dwivedi, and Nripendra P Rana. 2021. Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. *Future Generation Computer Systems*, 118:187–197.

Ioanna K Lekea and Panagiotis Karampelas. 2018. Detecting hate speech within the terrorist argument: A greek case. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1084–1091. IEEE.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Xinnan Lin, Feiwei Qin, Yong Peng, and Yanli Shao. 2021. Fine-grained pornographic image recognition with multiple feature fusion transfer learning. *International Journal of Machine Learning and Cybernetics*, 12:73–86.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. *arXiv preprint arXiv:2303.09892*.

Nobal B Niraula, Saurab Dulal, and Diwa Koirala. 2021. Offensive language detection in nepali social media. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 67–75.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Kshitij Rajput, Raghav Kapoor, Kaushal Rai, and Preeti Kaur. 2022. Hate me not: detecting hate inducing memes in code switched languages. *arXiv preprint arXiv:2204.11356*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–13.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu Sang Choi, and Byung-Won On. 2021. Aggression detection through deep neural model on twitter. *Future Generation Computer Systems*, 114:120–129.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Julian Moreno Schneider, Roland Roller, Peter Bourgonje, Stefanie Hegele, and Georg Rehm. 2018. Towards the automatic classification of offensive language and related phenomena in german tweets. In *14th Conference on Natural Language Processing KONVENS*, volume 2018, page 95.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58(5):102664.

Omar Sharif and Mohammed Moshiul Hoque. 2021. Identification and classification of textual aggression in social media: resource creation and evaluation. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 9–20. Springer.

Omar Sharif and Mohammed Moshiul Hoque. 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Nlp-cuet@dravidianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 255–261.

K Sreelakshmi, B Premjith, and K.P. Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744. Third International Conference on Computing and Network Communications (CoCoNet'19).

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

Appendix

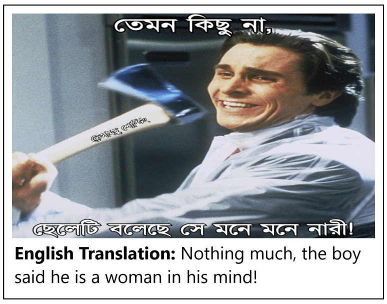
A Experimental Settings

We perform experiments on the Google Colab platform. The transformer architectures are downloaded from the Huggingface[2] library and implemented using the PyTorch Framework. The BNLP[3] and scikit-learn[4] libraries has been used for the preprocessing and evaluation measures. The models' hyperparameter values were selected empirically by examining the performance of the validation set. All the models are compiled using *cross_entropy* loss function. The error is optimized using the *Adam* optimizer with a *weight_decay* of 0.01. For visual and textual models, we use a *learning_rate* of $1e^{-5}$ while for multimodal models it is set to $3e^{-5}$. The proposed MAF and its variants are trained with a *learning_rate* of $5e^{-5}$. We use the *batch size* of 4 and train the models for 20 *epochs* with a learning rate scheduler. We examine the validation set performance to preserve the best model during training.

[2]https://huggingface.co/

[3]https://github.com/sagorbrur/bnlp

[4]https://scikit-learn.org/stable/



(a) GAg

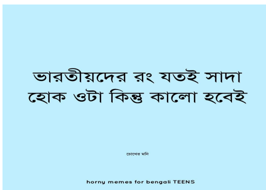
(b) PAg

(c) RAg

(d) Oth

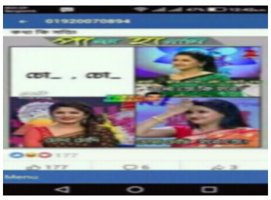
Figure A.1: Example of memes from different aggression classes. The criteria used to decide the classes were: (a) incites violence against people based on sexuality (b) attacks a political leader (c) attacks people based on religion (d) seeks to harm a person.

(a)



(b)

(c)



(d)

Figure A.2: Example memes were filtered out during the data collection process and the reason for the filtering (a) contains only visual information (b) only textual information (c) contains cartoons (d) the contents are not cleared.

B Annotation

Addressing the challenge of mitigating bias and obtaining accurate annotations is a pivotal concern when labeling a dataset (Bender and Friedman, 2018). Many studies (Sap et al., 2021; Röttger et al., 2021) have emphasized knowing the identity of the annotators beforehand because their experience and demographic variety can significantly influence the labeling process. Therefore, in Table B.1, we provide a detailed summary of the annotators' backgrounds in developing the dataset. Three annotators and an expert worked on the data annotation process. The expert was a Professor with 22 years of research experience in AI, while other annotators were computer science undergraduate students with varied research experience in the NLP field. Most annotators had annotation experience, and all were native Bengali speakers.

	Annotator-1	Annotator-2	Annotator-3	Expert
Research status	Undergrad	Undergrad	Undergrad	Professor
Research area	NLP	NLP	NLP	NLP, HCI, Robotics
Research experience (in years)	2	1	3	22
Previous annotation experience	Yes	Yes	No	Yes
Age	23	23	23	47
Religion	Islam	Islam	Hindu	Islam
Gender	Male	Male	Female	Male

Table B.1: A summary of the annotators' research background and demographic details.

We used the majority voting mechanism, where the label with the maximum number of votes was considered the final. In case of conflict, the expert annotator will determine the final label.

C Ablation Study

The proposed MAF has proven effective in aggressive meme classification. One of the core components of the proposed method is how many attention heads we will use to produce a better multimodal representation. In this regard, we performed an ablation study to illustrate the impact of the number of attention heads on the proposed model performance shown in Figure C.1.

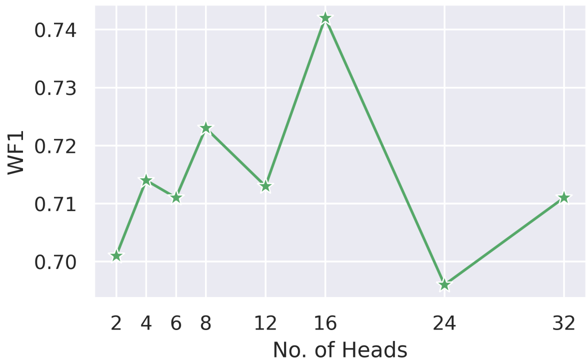


Figure C.1: Impacts of the number of heads on the performance of MAF method. These numbers were chosen because the feature vector dimension (768) is divisible by them.

It observed that the number of heads significantly impacts the model performance (WF1). For instance, it is noticed that between 2-12 heads model yields fluctuating results, however, staying above 70%. The model obtained the highest performance (WF1 $\approx$ 74%) with 16 heads. However, increasing the number of heads to more than 16 does not produce satisfactory results. We hypothesize that adding more heads will not improve the performance as this may make the multimodal representation more complex. However, more investigation is required to unfold the reason behind this performance variation.