Align before Attend: Aligning Visual and Textual Features for Multimodal Hateful Content Detection

Eftekhar Hossain[$*], Omar Sharif[Ψ¥*], Mohammed Moshiul Hoque[¥], Sarah M. Preum[Ψ]
[¥]Department of Computer Science and Engineering
[Ψ]Department of Computer Science, Dartmouth College, USA
[$]Department of Electronics and Telecommunication Engineering
[$¥]Chittagong University of Engineering & Technology, Bangladesh
{eftekhar.hossain, moshiul_240}@cuet.ac.bd, {omar.sharif.gr, sarah.masud.preum}@dartmouth.edu

Abstract

Multimodal hateful content detection is a challenging task that requires complex reasoning across visual and textual modalities. Therefore, creating a meaningful multimodal representation that effectively captures the interplay between visual and textual features through intermediate fusion is critical. Conventional fusion techniques are unable to attend to the modality-specific features effectively. Moreover, most studies exclusively concentrated on English and overlooked other low-resource languages. This paper proposes a context-aware attention framework for multimodal hateful content detection and assesses it for both English and non-English languages. The proposed approach incorporates an attention layer to meaningfully align the visual and textual features. This alignment enables selective focus on modality-specific features before fusing them. We evaluate the proposed approach on two benchmark hateful meme datasets, *viz.* MUTE (Bengali code-mixed) and MultiOFF (English). Evaluation results demonstrate our proposed approach's effectiveness with F1-scores of 69.7% and 70.3% for the MUTE and MultiOFF datasets. The scores show approximately 2.5% and 3.2% performance improvement over the state-of-the-art systems on these datasets. Our implementation is available at *https://github.com/eftekhar-hossain/Bengali-Hateful-Memes*.

Disclaimer: This paper contains hateful images that may be disturbing to some readers.

1 Introduction

Recently, online platforms are witnessing an emerging trend of propagating hateful and offensive content. While most research in this area has focused on detecting hate speech from text-based content (Waseem and Hovy, 2016; Schmidt and Wiegand, 2017), offensive multimodal content is also propagated, such as memes. Memes are images or screenshots with short texts embedded in them. Their sarcastic nature made them an increasingly popular tool for spreading hate and targeting individuals or communities based on various factors such as gender, race, ethnicity, religion, physical appearance, and sexual orientation (Williams et al., 2016; Chhabra and Vishwakarma, 2023). The proliferation of such content poses a significant threat to communal harmony and social stability and has therefore become an area of active research interest (Cao et al., 2022; Pramanick et al., 2021).

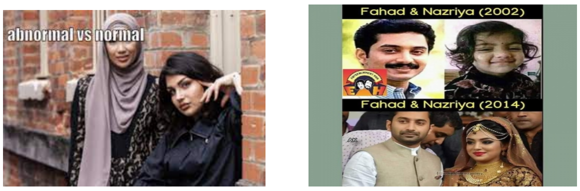


Figure 1: Example of hateful memes. In isolation, neither the image nor the caption may appear hateful, but when combined, they can convey a hateful message.

Multimodal hateful content detection requires a holistic understanding of visual and textual information. When considered separately, the image and caption components in Figure 1(a) may seem innocuous. The image portrays two women—one wearing a hijab and the other without and the caption states, "abnormal and normal". However, as a meme, this composition can be seen as derogatory towards the woman wearing the hijab by labeling her as abnormal. Similarly, the meme in Figure 1(b) insults the marriage of two South Indian celebrities by indicating their age gap in the text. Thus, focusing only on the image or the text is inadequate for complete understanding. Sometimes without the background information of the people and events used in a meme, it is difficult to interpret the meaning because the captions are short, fragmented, and

[*] Denotes equal contribution

sarcastic. Studies have demonstrated that off-the-shelf multimodal systems, which are typically effective in performing various visual-linguistic tasks, encounter difficulties when it comes to detecting hateful memes (Kiela et al., 2020; Cao et al., 2022). Furthermore, the current state-of-the-art systems (Lee et al., 2021; Pramanick et al., 2021) for detecting hateful memes face limitations when applied to resource-constrained languages. This is primarily because several key components within their architectures are not accessible or well-supported in other languages. These challenges underscore the need for language-specific adaptations to address hateful meme detection in a broader linguistic context effectively.

To address this knowledge gap, we present a solution for detecting multimodal hateful memes. The approach leverages an attention-based context-aware fusion framework to create coherent multimodal representations. We hypothesize that by aligning visual and textual features before fusion, the network can better capture essential cues for accurate classification. The key challenge lies in effectively incorporating modality information to enable the network to focus on crucial features. Previous methods (Pramanick et al., 2021; Lee et al., 2021) used background context and additional captions while performing the fusion. In contrast, our approach introduces an attention layer to align modalities which simultaneously facilitates the extraction of contextual representations from both visual and textual modalities. Moreover, without adding external knowledge, the model's learning capability is augmented when the aligned representations are combined with modality-specific (i.e., visual, textual) features. To evaluate our approach, we conducted experiments on two benchmark datasets in different languages: MUTE (Hossain et al., 2022c) and MultiOFF (Suryawanshi et al., 2020). The evaluation results and ablation study demonstrate the effectiveness of our solution over baseline and state-of-the-art methods.

The major contributions of this paper are three-fold: *(i)* develop an attention framework that effectively attends the contributing features of visual and textual modalities to detect multimodal hateful memes (Section 3.1); *(ii)* conduct an extensive evaluation on two different benchmark datasets on real-world memes to demonstrate the effectiveness of the proposed solution (Section 4.3, 4.5); and *(iii)* perform ablation studies in different settings to examine the impact of BERT-base embeddings in detecting hateful memes while also investigate the model's quantitative and qualitative errors to understand its limitations (Section 4.4, 4.4).

2 Related Work

Hateful Content Detection: Over the past few years, offensive/hate speech detection has received a significant amount of attention from researchers. Some works focused on developing new corpus for different languages (Lekea and Karampelas, 2018; Roy et al., 2022) while others studied to develop novel methods (Li and Ning, 2022; Mozafari et al., 2020a). However, most of the studies focused on hateful content detection from textual data and overlooked the multimodal aspects of the user-generated data. One such multimodal data is a meme, which combines both images and text. With the flourishing of internet memes and because of their detrimental impact on society, online hateful meme classification got a considerable amount of traction from the research community (Das et al., 2020; Cao et al., 2022) lately. Suryawanshi et al. (2020) and Kiela et al. (2020) introduced hateful memes dataset in English. Besides developing datasets in English, few works attempted to introduce hateful memes datasets for low-resource languages such as Bengali (Hossain et al., 2022c).

Multimodal Fusion: Over the years, various techniques have been applied to detect multimodal hateful memes. Conventional fusion (Vijayaraghavan et al., 2021; Gomez et al., 2020) by concatenating the modality-specific information is the most commonly used method for learning multimodal representation. Some works employed bilinear pooling (Chandra et al., 2021) while others fine-tuned transformers (Kiela et al., 2020) based architectures such as ViLBERT, MMBT, and Visual-BERT. Besides, some works attempted to use disentangled learning (Lee et al., 2021) and incorporate image captioning (Zhou et al., 2021) to improve the hateful memes detection performance. Recently, Cao et al. (2022) applied prompting techniques for hateful meme detection in English. To the best of our knowledge, no one has attempted to align the visual and textual features for hateful meme detection. Nonetheless, feature alignment is key in creating a successful multimodal representation (Zeng et al., 2022; Liu et al., 2019). This work aims to address this research gap by introducing an alignment technique for hateful meme detection.

Overall our work differs from the existing studies in several significant ways: *(i)* rather than using additional context with conventional (i.e., early, late) fusion for multimodal representation, we align the visual and textual features using attention before fusing them, *(ii)* Existing models are primarily designed for English and challenging to adapt for languages like Bengali. This work presents a model that uses alignment and can be adapted for any language by swapping out language-specific components, and *(iii)* evaluation is performed on real-world meme dataset *(MUTE, MultiOFF)* rather than the synthetic memes as in Kiela et al. (2020).

3 Method

Memes comprise two modalities (i.e., visual and textual); logically, one modality's content can outweigh another's content during prediction. Besides, not all the information from both modalities has an equal effect on determining whether a meme is hateful. We propose a context-aware fusion framework that selectively focuses on modality-specific information to model this complex relationship. The proposed network takes multimodal input and feeds the visual information to a CNN and textual information to an RNN for feature extraction. Then we calculate alignment weights over the visual and textual features through the attention layer. The objective is to capture the contributing features with higher weights by emphasizing both modalities. Subsequently, these alignment weights are utilized to create multimodal contextual representation. Finally, the resulting contextual and modality-specific representations are combined and passed to the softmax layer for classification. We denote our proposed architecture as ***Multimodal Context Aware - Skip Connected Fusion (MCA-SCF)*** framework. An overall architecture of the framework is presented in Figure 2.

To ensure the robustness of the architecture, we experiment with three other variants of the proposed MCA-SCF framework: a) *Vision Guided Contextual Fusion (VGCF)* framework; b) *Text Guided Contextual Fusion (TGCF)* framework; and c) *Multimodal Contextual Fusion (MCF)* framework. The architecture of these variants differs in context vector computation and information fusion. In *VGCF*, we compute contextual information concerning the visual information and fuse it with the textual features. On the other hand, in *TGCF*, the contextual information is computed with respect to textual features and integrated with the visual features. In contrast, we compute the context for both modalities and then combine them in *MCF*. The rest of the components for all the architectures remain the same. The details of the variants *VGCF, TGCF, MCF* can be found in Appendix A.

3.1 Proposed (*MCA-SCF*) Architecture

The *MCA-SCF* framework consists of several components described in the following subsections.

3.1.1 Preprocessing

Before feeding the data into the framework, we preprocess the visual (v) and textual (t) modality. For v, we resize the images to $150 \times 150 \times 3$ and transform the pixel values between 0 to 1 to reduce the computational complexity. On the other hand, we remove unwanted characters (i.e., symbols, URLs, numbers, etc.) from textual data. Then we encode each word with a unique number and make all the text lengths equal to size l, where l=60.

3.1.2 Visual and Textual Feature Extractor

We employ a pre-trained CNN (ResNet50) to obtain the visual features from the memes. We use ResNet50 because of its capability to address the vanishing gradient problem and effectiveness in several multimodal classification tasks (Hossain et al., 2022a,b). To adjust ResNet50, we exclude the top two layers from the main architectures and utilize the weights of the higher-level features previously trained on the ImageNet (Deng et al., 2009) dataset. We add a global average pooling layer followed by a dense layer and retrain the architecture with new weights. The following equation computes the visual features.

$$V_f = Relu\left(\sum_{k}^{d} W_{jk} * G_k + b_j\right) \tag{1}$$

Here, $V_f \in \mathbb{R}^{1\times d}$ represents the visual semantic features extracted by the ResNet50 for the m^{th} memes visual modality (v). Here, d represents the number of neurons (100) in the dense layer. And, G represents the feature map generated by the global average pooling layer while W and b represent the weight matrix and bias respectively.

We employ Recurrent Neural Network to extract both word-level and sentence-level textual features. Specifically, we use Bidirectional Long Short Term Memory (BiLSTM) network to capture the contextual dependency of the words. Initially, we generate

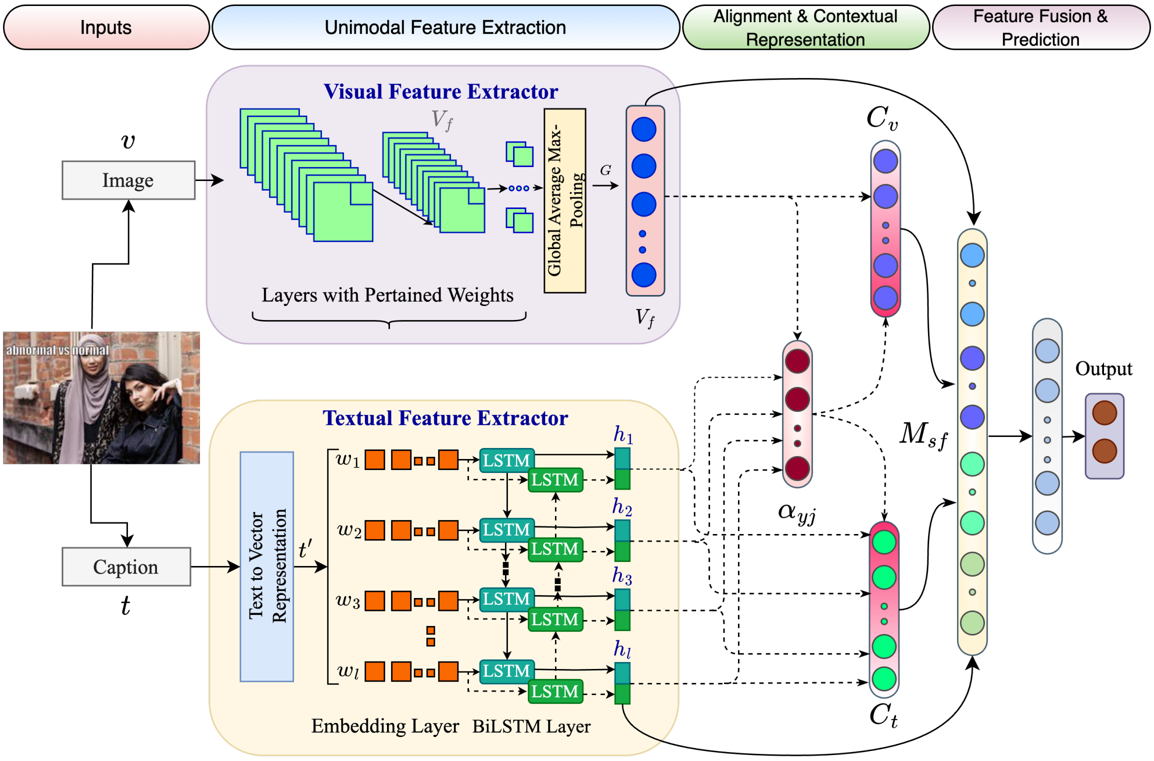

Inputs
Unimodal Feature Extraction
Alignment & Contextual Representation
Feature Fusion & Prediction
Visual Feature Extractor
V_f
Global Average Max-Pooling
G
Layers with Pertained Weights
v
Image
abnormal vs normal
Caption
t
Textual Feature Extractor
Text to Vector Representation
t'
w_1
w_2
w_3
w_l
LSTM
h_1
h_2
h_3
h_l
Embedding Layer
BiLSTM Layer
C_v
α_{yj}
C_t
M_{sf}
Output

Figure 2: Our proposed context-aware multimodal architecture: v and t are the processed image and its corresponding caption. The upper block represents the visual feature extractor, and the lower block is the textual feature extractor. Alignment scores (α_{yj}) are calculated by applying attention on visual (V_f) and textual ($h_1 ... h_l$) features. Subsequently, visual (C_v) and textual (C_t) context vectors are created by aligning (V_f) and ($h_1 ... h_l$) through alignment vector (α_{yj}). Finally, by concatenating these context vectors (C_v, C_t) with modality-specific features (V_f, h_l) our method creates the multimodal context-aware representation M_{sf}.

the embedding vectors that give a semantic meaning to each word. The embedding dimension size is set to (64). The embedding vectors are passed to a BiLSTM which can keep the contextual dependency of the word vectors of t. The output of the BiLSTM network is generated by concatenating the forward and backward LSTM cell's output. It gives a word-level feature vector for every k^{th} time step. The final time step (l^{th}) output is the sentence-level feature vector that we will use during the fusion operation. The features are computed using the following equation.

$$h_j^{[k]} = \vec{h_j} \oplus \overleftarrow{h_j} \tag{2}$$

Here, $h_j^{[k]} \in \mathbb{R}^{1\times 2N}$ and $h^{[l]} \in \mathbb{R}^{1\times 2N}$ respectively denote the BiLSTM word-level and sentence level feature generated for j^{th} word in the k^{th} layer or time step. l is padding length and N is the number of hidden units (50) in the LSTM cell. The $\oplus$ represents the concatenation. All the hyper-parameter values are selected via trial and error fashion by monitoring the validation accuracy.

3.1.3 Alignment and Fusion

Unlike existing approaches that employ early or late fusion techniques for multimodal representation, we align the visual and textual features through attention before joining them. Inspired from (Xu et al., 2015) we apply the additive attention (Bahdanau et al., 2014) mechanism to develop the alignment model. The model assigns a score $\alpha_{y,j}$ to the world-level feature of the j^{th} time step and the visual feature, V_f. The set of weights $\alpha_{y,j}$ determines how much image and text level features are aligned to predict a particular class label (y). The alignment score, α is parameterized by a feed-forward network where each feature vector (i.e., visual and textual) is trained with separate weights. The score function is therefore in the following form, given that $tanh$ is used as the non-linear activation function:

$$\alpha(V_f, h_j) = v_a^T tanh\,(W_1 * V_f + W_2 * h_j) \tag{3}$$

$$\alpha_{y,j} = \frac{exp\,(\alpha(V_f, h_j))}{\sum_{j=1}^{l} exp\,(\alpha(V_f, h_j))} \tag{4}$$

After performing the softmax operation (4), we obtain the normalized alignment scores, where higher weights are assigned to the feature combinations that are important for the prediction (y). Here, v_a, W_1, and W_2 are the weight matrices to be learned by the alignment model.

Afterward, we use these alignment scores to generate context vectors for each modality. The intuition behind this is that not all the features of individual modality are equally important for classification. Thus, focusing only on the significant feature is the key to better prediction. The following equation is computed for the context vectors.

$$C_v = \sum_j \alpha_{y,j} * V_f \tag{5}$$

$$C_t = \sum_j \alpha_{y,j} * h_j \tag{6}$$

Here, $C_v \in \mathbb{R}^{1\times d}$ and $C_t \in \mathbb{R}^{1\times d}$ are referred to as the vision-guided and text-guided context vectors, respectively. These vectors keep the contextual and significant modality-specific information concerning both visual and textual modalities.

The context vectors are concatenated to generate a context-aware multimodal representation. Furthermore, inspired by the residual learning (He et al., 2016) we concatenated each modality feature along with this contextual representation. The idea is to boost the gradient flow to the lower layer and enhance the multimodal representation. The following equation can express the combined feature representation.

$$M_{sf} = C_v \oplus C_t \oplus V_f \oplus h^{[l]} \tag{7}$$

Here, $M_{sf} \in \mathbb{R}^{1\times 4d}$ represents the contextual multimodal representation. This combined feature vector is then passed for the classification.

4 Experiments and Results

In this section, we first describe the datasets and the evaluation settings. We discuss the baselines and their results in comparison with the proposed method. Moreover, we conduct an ablation study to show how replacing components of the ***MCA-SCF*** framework affects the performance. Subsequently, an error analysis will be provided to understand the model's error. Furthermore, we perform a cross-domain analysis to see how the proposed framework performs irrespective of language variation in a zero-shot setting (Appendix C).

4.1 Datasets

We train and evaluate our proposed approach on two benchmark multimodal datasets: the Multimodal Bengali Hateful Memes (MUTE) and a popular English Memes (MultiOFF) dataset. Due to the unavailability of datasets, we limited our performance assessment on these datasets. For this work we only consider real-world memes and avoid synthetic datasets (Kiela et al., 2020). Table 1 presents the distribution of the datasets.

Dataset	Class	Train	Validation	Test
MUTE	Hate	1275	152	159
	Not-Hate	2092	223	257
MultiOFF	Offense	187	59	59
	Not-Offense	258	90	90

Table 1: Distribution of MUTE and MultiOFF datasets.

MUTE (Hossain et al., 2022c): A hateful memes dataset for the Bangla language. It consists of 4158 memes where the captions are code-mixed (Bangla + English) in nature. Among 4158 memes, 1586 are hateful and the rest of them are not hateful. We use the exact train-test split adopted by the authors to compare with our proposed approach.

MultiOFF (Suryawanshi et al., 2020): The MultiOFF consists of a total of 743 memes collected based on the US presidential election. The authors labeled the memes into the *offensive* category. However, these memes can be considered hateful since they substantially overlap with the hatred category and contain derogatory/abusive content targeted toward a group of people. The training, validation, and test set contain 445, 149, and 149 memes.

We adopt the evaluation metrics from the previous works in hateful meme classification (Lee et al., 2021). The superiority of a model is determined based on the weighted F1-score. Besides, weighted precision, recall, and Area Under the Receiver Operating Characteristics (AUC) scores have been reported for comparison. The details of the experimental settings are discussed in Appendix B.

4.2 Baselines

We develop several baselines considering the unimodal (i.e., image or text) and multimodal information. The baseline models are chosen based on the best-performing models on these datasets (*MUTE, MultiOFF*) and popular techniques from the existing literature. The model's hyperparameters are chosen via a trial-and-error approach by monitoring the validation accuracy. The baseline architectures are described in the following subsections.

4.2.1 Unimodal Models

Initially, we implemented models considering only the visual modality. We use the **ResNet50** network where we fine-tuned and retrained it with new weights. The architecture configuration kept the same as described in Section 3.1.2. Besides, we also fine-tuned the **Vision Transformer (ViT)** (Dosovitskiy et al., 2020) architecture on both datasets. On the other hand, for textual modality, we employed several architectures including **BiLSTM** (Baruah et al., 2019), **BiLSTM + Attention** (Altın et al., 2019), **BERT** (Mozafari et al., 2020b), and **XLM-R** (Ranasinghe and Zampieri, 2020). In one architecture we use an LSTM cell with 32 hidden units. Subsequently, the attention mechanism is added with the LSTM in another architecture. We use the language-specific variation of the BERT (i.e., **Bangla BERT** (Sarker, 2020) and **English-BERT** (Devlin et al., 2018)), the multilingual BERT (**m-BERT**), and cross-lingual BERT (**XLM-R**) for our task. We freeze the weights of these architectures and retrain them with new weights by adding a dense layer of 100 neurons. The dense layer takes the sentence embeddings as input and makes a higher-level representation of the text. Finally, this representation is passed to the classification layer for prediction.

4.2.2 Multimodal Models

To develop the models using multimodal information, we use the most popular fusion techniques including **Early Fusion** (Pranesh and Shekhar, 2020), **Late Fusion** (Hossain et al., 2022b), and **Attentive Fusion** (Sharma et al., 2022). We select the best-performing unimodal models (ResNet50 and LSTM) for visual and textual feature extraction.

- For early fusion, a dense layer of 100 neurons is added at both ends of individual modalities to make a joint representation by concatenating them.

- In late fusion, the classification layer's output from each modality is combined and then passed for the classification.

- With attentive fusion, the last dense layer's output is passed to an attention layer, and then the resulting attentive vector is used for classification.

Finally, we employed several state-of-the-art multimodal architectures including **VisualBERT-COCO** (Li et al., 2019), **CLIP** (Radford et al., 2021), and **ALBEF** (Li et al., 2021) and fine-tuned them on our datasets.

4.3 Results

Table 2 presented the outcome of the baselines and proposed method over the test set of *MUTE* and *MultiOFF* datasets. In *MUTE*, the visual models (ResNet50 and ViT) failed to obtain a satis-

Approach	Models	MUTE				MultiOFF			
		P	R	WF	AUC	P	R	WF	AUC
Unimodal	ResNet50 (FT)	0.634	0.646	0.631$_{\pm 0.00}$	0.598$_{\pm 0.01}$	0.624	0.637	0.623$_{\pm 0.02}$	0.593$_{\pm 0.01}$
	ResNet50 (RT)	0.617	0.634	0.614$_{\pm 0.02}$	0.580$_{\pm 0.03}$	0.580	0.557	0.562$_{\pm 0.08}$	0.559$_{\pm 0.01}$
	ViT	0.622	0.639	0.584$_{\pm 0.03}$	0.557$_{\pm 0.02}$	0.603	0.624	0.559$_{\pm 0.06}$	0.542$_{\pm 0.02}$
	BiLSTM	0.660	0.670	0.658$_{\pm 0.02}$	0.626$_{\pm 0.02}$	0.611	0.604	0.606$_{\pm 0.02}$	0.591$_{\pm 0.01}$
	BiLSTM + Attention	0.659	0.622	0.627$_{\pm 0.02}$	0.636$_{\pm 0.01}$	0.577	0.597	0.578$_{\pm 0.02}$	0.548$_{\pm 0.01}$
	BERT	0.645	0.658	0.642$_{\pm 0.08}$	0.609$_{\pm 0.06}$	0.621	0.617	0.610$_{\pm 0.01}$	0.611$_{\pm 0.09}$
	m-BERT	0.627	0.644	0.620$_{\pm 0.02}$	0.586$_{\pm 0.01}$	0.584	0.611	0.574$_{\pm 0.02}$	0.547$_{\pm 0.07}$
	XLM-R	0.646	0.656	0.648$_{\pm 0.04}$	0.618$_{\pm 0.01}$	0.612	0.630	0.580$_{\pm 0.01}$	0.557$_{\pm 0.08}$
Multimodal	Early Fusion	0.634	0.649	0.607$_{\pm 0.02}$	0.575$_{\pm 0.01}$	0.646	0.657	0.645$_{\pm 0.02}$	0.616$_{\pm 0.06}$
	Late Fusion	0.619	0.634	0.619$_{\pm 0.02}$	0.586$_{\pm 0.00}$	0.738	0.657	0.568$_{\pm 0.01}$	0.563$_{\pm 0.07}$
	Attentive Fusion	0.660	0.637	0.642$_{\pm 0.00}$	0.641$_{\pm 0.02}$	0.610	0.624	0.538$_{\pm 0.03}$	0.532$_{\pm 0.06}$
	VisualBERT COCO	0.494	0.572	0.530$_{\pm 0.04}$	0.521$_{\pm 0.01}$	0.396	0.689	0.503$_{\pm 0.07}$	0.502$_{\pm 0.05}$
	CLIP	0.643	0.641	0.560$_{\pm 0.06}$	0.545$_{\pm 0.07}$	0.646	0.651	0.601$_{\pm 0.05}$	0.576$_{\pm 0.03}$
	ALBEF	0.679	0.667	0.668$_{\pm 0.06}$	0.677$_{\pm 0.02}$	0.612	0.617	0.613$_{\pm 0.04}$	0.610$_{\pm 0.04}$
Proposed System and Variants	VGCF	0.671	0.677	0.671$_{\pm 0.02}$	0.644$_{\pm 0.02}$	0.651	0.624	0.628$_{\pm 0.03}$	0.632$_{\pm 0.04}$
	TGCF	0.662	0.665	0.663$_{\pm 0.01}$	0.641$_{\pm 0.01}$	0.667	0.651	0.655$_{\pm 0.01}$	0.651$_{\pm 0.01}$
	MCF	0.692	0.699	0.689$_{\pm 0.02}$	0.659$_{\pm 0.01}$	0.654	0.657	0.655$_{\pm 0.05}$	0.635$_{\pm 0.04}$
	MCA-SCF (Proposed)	0.696	0.696	$\mathbf{0.697_{\pm 0.00}}$	0.674$_{\pm 0.01}$	0.702	0.704	$\mathbf{0.703_{\pm 0.02}}$	0.686$_{\pm 0.03}$

Table 2: Performance comparison of unimodal and multimodal models on test set where P, R, WF, and AUC denote precision, recall, weighted F1-score, and area under the receiver operating characteristics curve respectively. VGCF, TGCF, and MCF are the variants of the proposed MCA-SCF approach. The FT and RT represent the fine-tunned and retrained version of ResNet50, respectively. The standard deviation (±) with five different random seeds is also reported. For space constraints, the score is not shown for precision and recall.

factory outcome, while among the textual models, BiLSTM achieved the highest F1-score of 0.658. Surprisingly, the performance of the pre-trained transformers is lower than BiLSTM. We perform a detailed ablation study to get more insights on this. Meanwhile, when multimodal information is integrated, the attentive fusion approach achieved the highest F1 (0.642) and AUC (0.641) scores compared to its counterparts (i.e., early and late fusion). Among the other multimodal architectures (i.e., VisualBERT, CLIP, and ALBEF), ALBEF showed outstanding performance with an F1 score of 0.668. However, we observed that the variants (VGCF, TGCF, and MCF) of the alignment approach obtained superior performance over the unimodal and other multimodal models except ALBEF. Even though they achieved better outcomes, the proposed MCA-SCF framework outperformed all the models by getting the highest F1 score of 0.697.

In *MultiOFF* dataset, BERT achieved the highest F1-score of 0.610 amid the unimodal models. On the other hand, early fusion showed significantly higher performance (0.645) compared to late fusion (0.568), attentive fusion (0.538), and other multimodal architectures such as VisualBERT (0.503), CLIP (0.601), and ALBEF (0.613). We noticed that the performance is substantially improved with the variants. Nonetheless, MCA-SCF outperforms all the models, obtaining the highest F1 score of 0.703 and AUC score of 0.686.

In summary, the proposed *MCA-SCF* framework and its variants outperformed the baselines in both datasets. Aligning the visual and textual information before fusing them played a crucial role in boosting the model's predictive performance.

4.4 Ablation Study

In addition to the experiments emphasizing the importance of context-aware multimodal representation for hateful meme classification in Table 2, we also examine the effect of contextualized embeddings in MCA-SCF instead of simple word embeddings. We consider three transformer models i.e., language-specific BERT, multilingual BERT, and XLM-R. We employed the architecture with similar parameters described in Section 4.2.1. Two individual models were developed for each transformer architecture. Firstly, BERT word level and sentence level embeddings were used to develop MCA-SCF whereas in the second case, contextualized embeddings were passed to an LSTM layer and utilized the LSTM word level features with the contextualized sentence embeddings to construct MCA-SCF. The training parameters of the models were kept the same as discussed in Appendix B. Table 3 reported the outcomes when contextualized embeddings are used. We observed that, in the case of *MUTE*, MCA-SCF with m-BERT obtained the highest F1 score (0.665), whereas MCA-SCF with m-BERT + BiLSTM achieved the maximum F1 score (0.613) in *MultiOFF* dataset.

Models	MUTE		MultiOFF	
	WF	AUC	WF	AUC
MCA-SCF w/ BERT + BiLSTM	0.657	0.634	0.571	0.542
MCA-SCF w/ only BERT	0.649	0.637	0.612	0.586
MCA-SCF w/ m-BERT + BiLSTM	0.645	0.622	0.613	0.589
MCA-SCF w/ only m-BERT	0.665	0.676	0.575	0.551
MCA-SCF w/ XLM-R + BiLSTM	0.615	0.582	0.525	0.501
MCA-SCF w/ only XLM-R	0.661	0.627	0.540	0.513

Table 3: Effect on the proposed method performance when replacing the text model with various transformer architectures.

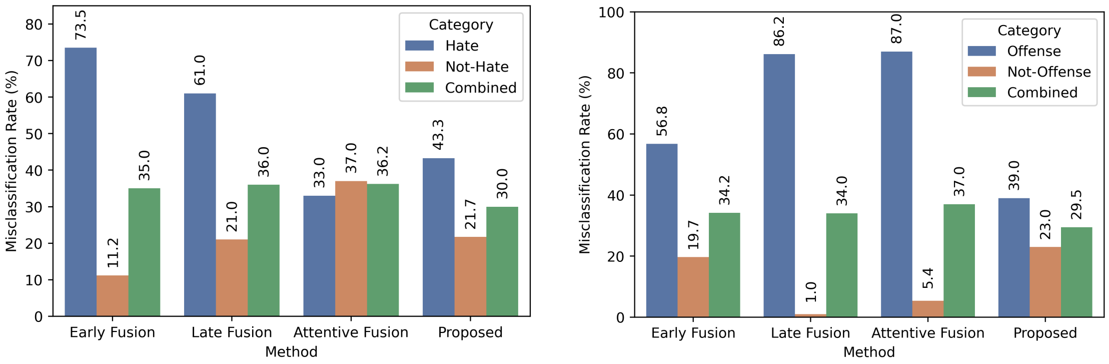
The findings reveal that there is no significant effect of using the BERT-based models for hateful meme detection. Even the BERT-based model outcomes are lower than the variants of the proposed method. Therefore, it can be stated that contextualized embeddings are not suitable for hateful meme detection. The reason behind this lower performance could be the fact that the memes' captions are very different from regular texts. BERT-based models are typically trained on longer and more complete textual inputs, whereas the language used in meme captions is often short, fragmented, and sarcastic. This discrepancy in language style can cause this suboptimal performance.

4.5 Comparison with Existing Studies

Table 4 presents the performance comparison of the proposed method with the existing state-of-the-art systems on the datasets. In *MUTE*, our proposed multimodal framework achieves the best F1 score

Dataset	Approaches	WF (%)
MUTE	Hossain et al. (2022c)	67.2
	Proposed	**69.7**
MultOFF	Suryawanshi et al. (2020)	54.0
	Lee et al. (2021)	64.6
	Hossain et al. (2022d)	66.7
	Zhong et al. (2022)	67.1
	Proposed	**70.3**

Table 4: Comparative analysis of the proposed method with the existing state-of-the-art systems.



(a) MUTE

(b) MultiOFF

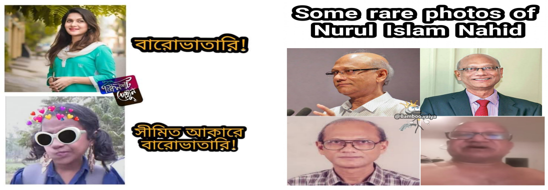
Figure 3: Misclassification rate comparison between various fusion approaches (i.e., early, late, attentive) and proposed (MCA-SCF) method on both datasets.

of 69.7% (↑ 2.5%) as compared to the existing highest score of 67.2%. Likewise, for *MultiOFF* dataset, we obtain the highest F1 score of 70.3% (↑ 3.2%) beating the current state-of-the-art system (67.1%). The performance improvement in both datasets' indicates our proposed method's novelty.

4.6 Error Analysis

We investigate the errors of the proposed MCA-SCF approach both quantitatively and qualitatively.

Quantitative Analysis: Early, late, and attentive fusion techniques have been considered to compare the errors with the proposed approach. We measured the Misclassification Rate (MR) for all the models reported in Figure 3. For *MUTE* dataset, we observed that the MR is reduced at 43.3% (proposed method) from 73.5% (early fusion) in *Hate* class while it is increased ≈10% in *Not-Hate* class. However, the error rate in *Not-Hate* class is minimal with the early fusion approach, whereas for *Hate* class, the attentive fusion approach reduces the error most. To conclude, we computed the combined class error rate and found that the overall system's error is the lowest (30%) with the proposed MCA-SCF method. Likewise, in *MultiOFF*, the proposed method achieves the lowest combined error rate of 29.5%. It is worth noting that the proposed model significantly reduces the error rate in negative classes, enabling effective detection of hateful memes. One interesting aspect observed is that the misclassification rate is higher in the Negative (*Hate or Offense*) class compared to the Positive (*Not-Hate or Not-Offense*) class across all approaches. This discrepancy could be attributed to the uneven distribution of data, with fewer training samples in the negative classes. As a result, the model may have struggled to effectively learn visual and textual patterns, leading to incorrect predictions.



(a) **EF:** Not-Hate (✗) **AF:** Not-Hate (✗) **Proposed:** Hateful (✓)

(b) **Actual:** Hateful **Predicted:** Not-Hate

Figure 4: Example (a) shows a meme where the proposed method yields better predictions, and example (b) illustrates a wrongly classified sample. The symbol (✓) and (✗) indicates the correct and incorrect prediction. EF and AF represent the early fusion and attentive fusion approaches, respectively.

Qualitative Analysis: We also perform qualitative analysis by investigating model predictions on a few samples. For example, the meme in Figure 4 (a) was misclassified as *Not-Hate* by the early and attentive fusion approaches. However, the proposed method captures the image and textual features that represent the context of the meme and therefore can correctly predict them as *Hateful*. We also analyze where the proposed method failed to give accurate inferences. For instance, the model misclassified the meme shown in Figure 4 (b) as *Not-Hate* when the actual label is *Hate*. The reason for this misclassification could be the presence of consistent visual features "Bald Man" and the absence of any trigger word in the text. Moreover, the model needs world-level knowledge to understand that this meme is demeaning the identity of a reputed person in Bangladesh. The above analysis shows that we need to explore more advanced reasoning modules to classify such memes accurately.

5 Conclusion

This paper presents *MCA-SCF*, a multimodal framework that aligns visual and textual features using attention to create a coherent contextual representation. The model aims to improve hateful content detection performance by leveraging contextual and modality-specific representations. We evaluate the model on two publicly available datasets i.e., *MUTE* and *MultiOFF*. Our extensive experiments demonstrate that *MCA-SCF* outperforms the state-of-the-art systems on these datasets. Furthermore, we conducted experiments with different variants of the model and performed an ablation study to ensure the system's robustness. The ablation study reveals that general word embeddings are more suitable than contextualized embedding for multimodal hateful meme detection. Finally, the cross-domain analysis illustrates the model's generalization capability in zero-shot settings.

Limitations

We identify several findings in this work. Firstly, we found that advanced multimodal models (e.g., CLIP, and VisualBERT) can not show satisfactory performance on both datasets. One compelling reason can be attributed that these models are not pretrained on enough Bengali image-text pairs and thus perform poorly when fine-tuning on the MUTE dataset. On the other hand, the lags in the performance in MultiOFF due to having fewer samples. As a result, the model does not get enough examples to learn complex relationships in the task and provides inferior performance. Besides that other advanced multimodal models (i.e., ALIGN, FLAVA, ViLBERT, BLIP) are rarely pretrained for Bengali image text pairs, limiting their applications in such low-resource languages. Therefore, we focus on enhancing the performance of off-the-shelf models with minimal computation by improving intermediate fusion through alignment. Our error analysis indicates that there is still significant room for improvement to effectively align visual and textual features for multimodal hateful content detection. Secondly, while the proposed model can infer the implicit meaning of memes in certain cases, it still falls short in complex reasoning to comprehend the contextual nuances of memes with concise captions. Finally, due to the unavailability of real-world meme datasets, we limited our performance assessment to two benchmark datasets. In the future, we plan to apply the model to detect memes in similar domains like harm and aggression, demonstrating its robustness across diverse and challenging categories.

References

Lütfiye Seda Mut Altın, Àlex Bravo Serrano, and Horacio Saggion. 2019. Lastus/taln at semeval-2019 task 6: Identification and categorization of offensive language in social media with attention-based bilstm model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 672–677.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Arup Baruah, Ferdous Barbhuiya, and Kuntal Dey. 2019. Abaruah at semeval-2019 task 5: Bi-directional lstm for hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 371–376.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. In *13th ACM Web Science Conference 2021*, pages 148–157.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*.

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.

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, Mohammed Moshiul Hoque, Enamul Hoque, and Md Saiful Islam. 2022a. A deep attentive multimodal learning approach for disaster identification from social media posts. *IEEE Access*, 10:46538–46551.

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

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022c. 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. 2022d. 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.

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.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5138–5147.

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.

Jiaxuan Li and Yue Ning. 2022. Anti-asian hate speech detection via data augmented semantic relation inference. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 607–617.

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.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. 2019. Aligning visual regions and textual concepts for semantic-grounded image representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020a. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020b. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

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

Raj Ratn Pranesh and Ambesh Shekhar. 2020. Memesem: a multi-modal framework for sentimental analysis of meme via transfer learning. In *4th Lifelong Machine Learning Workshop at ICML 2020*.

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.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. *arXiv preprint arXiv:2010.05324*.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

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

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Mayukh Sharma, Ilanthenral Kandasamy, and WB Vasantha. 2022. R2d2 at semeval-2022 task 5: Attention is only as good as its values! a multimodal system for identifying misogynist memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 761–770.

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, Marseille, France. European Language Resources Association (ELRA).

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2021. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of internet memes. *Computers in Human Behavior*, 63:424–432.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Qi Zhong, Qian Wang, and Ji Liu. 2022. Combining knowledge and multi-modal fusion for meme classification. In *International Conference on Multimedia Modeling*, pages 599–611. Springer.

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 Variants of *MCA-SCF* Framework

We develop three other variants of the *MCA-SCF* network namely *VGCF*, *TGCF*, and *MCF*. Figure A.1 shows the computation of the variants. The *VGCF* framework does not account for the context vector generated from the text modality. After aligning the visual and textual modalities, we used the obtained alignment score (α_{yj}) to highlight only the significant visual information and combined them with the sentence-level ($h^{[l]}$) textual feature. The VGC vector $V_{gf} \in \mathbb{R}^{1\times 2d}$ can be expressed by the following equation.

$$V_{gf} = C_v \oplus h^{[l]} \tag{8}$$

On the other hand, with *TGCF* framework, we utilize the alignment score to generate a contextual representation (C_t) only for the text modality. This representation is then combined with the visual features (V_f) to compute the TGC vector $T_{gf} \in \mathbb{R}^{1\times 2d}$ by the equation (9).

$$T_{gf} = C_t \oplus V_f \tag{9}$$

In the *MCF* framework, we combined the two context vectors (i.e., C_v and C_t) to make a contextual multimodal representation. The vector $M_{cf} \in \mathbb{R}^{1\times 2d}$ can be expressed by the equation.

$$M_{cf} = C_v \oplus C_t \tag{10}$$

B Experimental Settings

We perform experiments on the Google Colab platform. The transformer architectures were downloaded from the huggingface library and implemented using the TensorFlow framework. All the models are compiled using *binary cross-entropy*

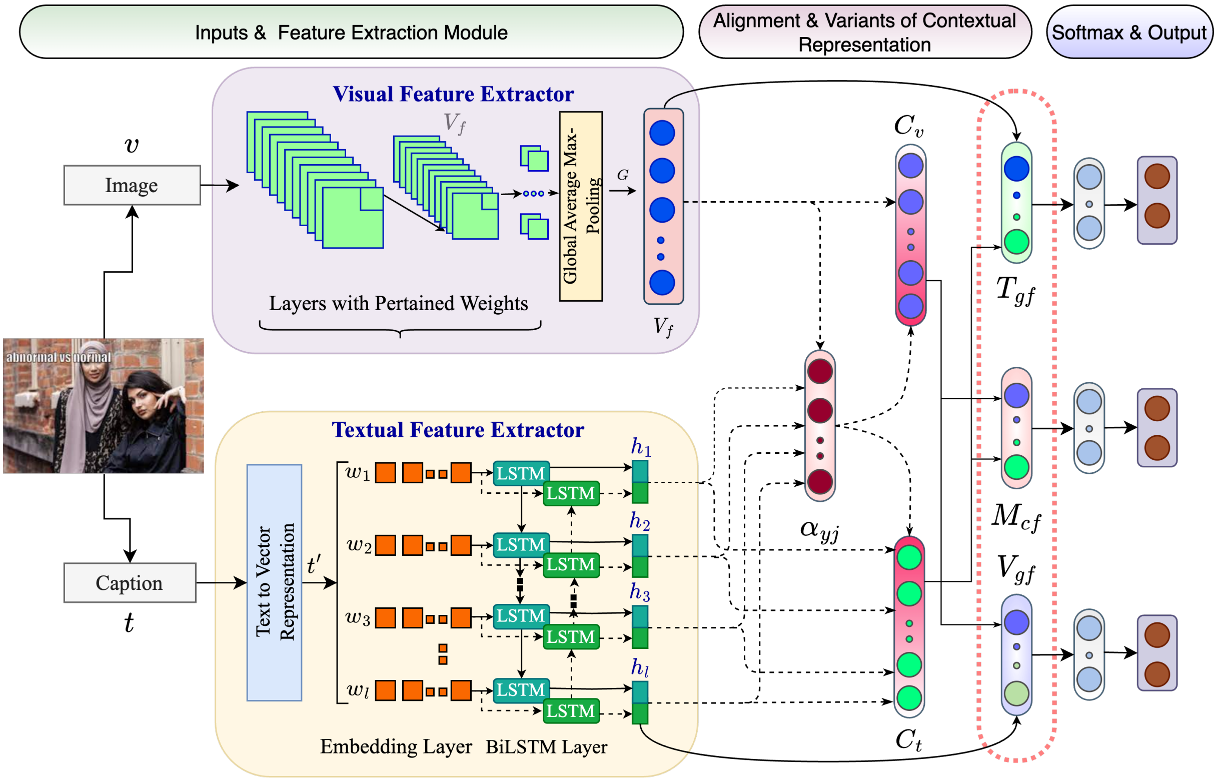


Figure A.1: Variants of the proposed MCA-SCF framework. The majority of the components remain the same as illustrated in figure 2. The three variants (V_{gf}, M_{cf}, T_{gf}) have differences in the way they integrate information to emphasize the context of a particular modality.

loss function. For all the models the error optimization is performed by the *Adam* optimizer with a learning rate of $1e^{-3}$ except for the transformer-based models which are $3e{-}5$. We used the *batch size* of 16 and trained the models for 20 *epochs*. To save the best intermediate models during training Keras checkpoint method has been utilized.

C Zero-shot Cross-Domain Transfer

We examine the cross-domain transfer ability of the proposed method by training it on a source dataset and evaluating it on a target dataset. Besides, we also investigate the proposed method's performance when the training is done on combined datasets but tested only on a particular dataset. We focus on examining the impact of captured phenomena between datasets. The cross-domain performance has been measured by the relative zero-shot transfer ability (Turc et al., 2021). We denoted it as the recovery ratio because it indicates the ratio of how much performance is recovered by changing the source domain and it is given as follows.

$$R(S,T) = \frac{F(S,T)}{F(T,T)} \tag{11}$$

Here, $F(S,T)$ is a model performance (i.e., f_1-score) for the source domain S on the target domain T. If the source and target domains are the same, the R would be 1.0. The recovery scores of both zero-shot and combined dataset settings are given in Table C.1.

			Target	
			MUTE	Multi-OFF
Source	**Zero-shot**	MUTE	0.697	0.585 (84%)
		MultiOFF	0.527 (75%)	0.703
	Cross- domain	MT+MO	0.604 (86%)	0.627 (90%)

Table C.1: Effect of the zero-shot and cross-domain transfer on both datasets. MT+MO indicates the combination of the MUTE and MultiOFF datasets. The major diagonal represents the actual performance, while the minor diagonal indicates how much performance is recovered when we change the source dataset.

In both settings, the recovery rate is comparatively higher when we evaluate on *MultiOFF* dataset and train using the *MUTE* dataset. For instance, in the zero-shot setting, the *MUTE* dataset 75% performance of 0.697 is recovered when the source domain was the *MultiOFF* dataset. Similarly, we observed that 84% is the recovery rate on *MultiOFF* when *MUTE* is the source domain. On the other hand, with a combined setting, 86% and 90% performance is recovered of the *MUTE* and *MultiOFF* datasets. Overall, in zero-shot setting *MUTE* as a source dataset can mostly recover the performance from *MultiOFF*. This could happen because *MUTE* consists of code-mixed captions and has more training samples. This may allow for a greater transfer and sharing of multimodal features between datasets, ultimately contributing to the model's strong performance on the *MultiOFF* dataset. Meanwhile, the proposed method can not generalize well on MUTE when trained with *MultiOFF* dataset. This is because the less number of training samples in *MultiOFF* and the model do not get any information about the Bengali language from the dataset. In contrast to its moderate generalization performance in the zero-shot setting, our proposed method demonstrates strong performance in the test set of each dataset when trained on the combined training set.