Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1542–1554
Marseille, 20-25 June 2022

MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes

Eftekhar Hossain[$], Omar Sharif[Ψ], Mohammed Moshiul Hoque[Ψ]
[$]Department of Electronics and Telecommunication Engineering
[Ψ]Department of Computer Science and Engineering
[$Ψ]Chittagong University of Engineering and Technology, Chattogram-4349, Bangladesh
{eftekhar.hossain, omar.sharif, moshiul_240}@cuet.ac.bd

Abstract

Posting and sharing memes have become a powerful expedient of expressing opinions on social media in recent days. Analysis of sentiment from memes has gained much attention to researchers due to its substantial implications in various domains like finance and politics. Past studies on sentiment analysis of memes have primarily been conducted in English, where low-resource languages gain little or no attention. However, due to the proliferation of social media usage in recent years, sentiment analysis of memes has become a crucial research issue in low resource languages. The scarcity of benchmark dataset is a significant barrier in performing multimodal sentiment analysis research in resource-constrained languages like Bengali. This paper presents a novel multimodal dataset (named **MemoSen**) for Bengali containing 4368 memes with three annotated sentiment labels *positive, negative, and neutral*. A detailed annotation guideline is provided to facilitate further resource development in this domain. Additionally, a set of experiments are carried out on MemoSen by constructing twelve unimodal (i.e., visual, textual) and ten multimodal (image+text) models. The evaluation exhibits that the integration of multimodal information significantly improves (about 1.2%) the meme sentiment classification compared to the unimodal counterparts and thus elucidate the novel aspects of multimodality.

Keywords: Sentiment analysis, Multimodal fusion, Memes, Code-mixing, Low resource languages

1. Introduction

Recently, the usage of social media platforms (i.e. Facebook, Twitter, Instagram) has increased dramatically due to the substantial evolution of the Internet and various web 2.0 applications. These platforms have become a place where people share their opinions concerning business, politics, services, entertainment, and other current affairs. Automatic sentiment analysis of these conversations has grabbed increased attention from the NLP researchers in recent years since it helps to identify a user's viewpoint or expression on a particular event or topic (Hossain et al., 2020b; Bakliwal et al., 2013). To date, most of the researches have tried to classify the textual sentiment into three classes: positive, negative and neutral classes (Mamta et al., 2020; Mamun et al., 2022). However, the mode of information in social media platforms is dramatically transforming day by day. The recent surge in multimodal (i.e., combination of image, text, and videos) content in these platforms has brought a new direction in sentiment analysis research. One such multimodal content is the *meme* which has become a popular form of propagating information. Few pieces of researches (Pranesh and Shekhar, 2020; Walińska and Potoniec, 2020) have been conducted to analyze memes' sentiment in English by combining visual and textual features. Joint evaluation of features of multiple modalities is crucial to accurately infer a meme's sentiment. To the best of our knowledge, no significant attempt has been made to analyze memes sentiment in low-resource languages, especially Bengali. People prefer to use memes in their mother language. In recent years, an increasing trend in using memes is observed in Bengali due to the rapid growth of social media users in Bangladesh. Thus it is crucial to identify the sentiment of the meme to mitigate the spread of negativity and understand the public expression towards an event or topic. Therefore, to initiate research in this arena, this work aims to develop a Bengali benchmark multimodal dataset for analyzing the sentiment of memes.

However, developing an automatic multimodal memes detection model is a complicated task. The most challenging task is to extract the embedded text from the memes, as Bengali has no standard optical character recognizer (OCR). Finding the appropriate sentiment (positive, negative, neutral) of a meme is another challenging task for human annotators. Thus, determining the underlying sentiment of the memes would also be difficult for the machines and humans for several reasons. For example, (i) most of the memes are context-dependent, (ii) the visual and textual information are often disparate (iii) sometimes the embedded text is too short for machines to learn the context. Another obstacle is extracting the code-mixed and code switched text from the memes that the existing OCR's cannot obtain. Finally, when integrating the visual and textual features, it is realized that these complex problems require more sophisticated models for providing accurate inference.

This work developed a novel Bengali multimodal sentiment analysis dataset (named *MemoSen*) containing 4368 memes collected from various social media platforms and then carefully annotated them into three classes to address the above issues. Proper annota-

tion guidelines are presented to mitigate the ambiguity concerning the sentiment labelling. Moreover, several state-of-the-art models are employed for benchmarking and investigating their outcomes. The key findings of the investigations are (i) multimodal features are very effective than unimodal features (i.e., visual or textual) for detecting the meme's sentiment, and (ii) the Word2vec features are more effective compared to the BERT embedding when aggregated with the visual features in multimodal evaluation. The significant contributions of this work can be summarized as follows:

- Created the MemoSen, a multimodal sentiment analysis dataset for Bengali consisting of 4368 memes annotated with Positive, Negative, Neutral labels.
- Performed extensive experiments with state-of-the-art visual and textual models and then synergistically integrated features of both modalities by utilizing different multimodal fusion approaches.

Reproducibility: The entire dataset and the source code are available at https://github.com/eftekhar-hossain/MemoSen. The appendix presents a few samples of the MemoSen, model hyperparameters, and annotation information.

2. Related Work

This section covers past studies on sentiment analysis based on unimodal (i.e., image, text) and multimodal contents.

Image based sentiment analysis: Despite being a notable research topic, sentiment analysis using visual or multimodal information has received very little attention from the researchers compared to the text data. Borth et al. (2013) introduced the visual content based sentiment analysis with the SentiBank system that extracts semantic features from the images and utilizes them to predict the associated emotion of the image. Miller and Sinanan (2017) discussed the importance of image data in determining the sentimental states of users on social media. Both You et al. (2015) and Kumar and Jaiswal (2017) proposed a domain transfer technique using convolutional neural network (CNN) for sentiment analysis of the Flicker image dataset. French (2017) used metadata and images other information for predicting the sentiment from social media memes.

Text based sentiment analysis: Substantial studies have been conducted on sentiment analysis using textual data for the high resource, and resource-constraint languages (Li et al., 2019b). Most early works focused on the traditional feature engineering with various machine learning techniques such as Logistic Regression (LR) (Hamdan et al., 2015), Support Vector Machine (SVM) (Zainuddin and Selamat, 2014), and Naive Bayes (NB) (Hossain et al., 2021). Later, researchers used several deep learning methods such as Bidirectional Long Short Term Memory (BiLSTM) (Hameed and Garcia-Zapirain, 2020), CNN (Liao et al., 2017), and transformers (Islam et al., 2020). For example, both Murthy et al. (2020) and Hossain et al. (2020a) developed textual sentiment analysis framework using BiLSTM network. Li et al. (2019a) proposed an enhanced sentiment feature based deep neural network for sentiment classification. Alam et al. (2017) also proposed CNN based approach to identify the textual sentiment. Similarly, Naseem et al. (2020) employed the transformer-based approach to classify the positive, negative, and neutral sentiment of texts.

Multimodal sentiment analysis: In contrast to the image and text-based analysis, few works have been accomplished on multimodal sentiment analysis, specifically on internet memes. Moreover, the majority of the tasks were conducted in the English language. For instance, Pranesh and Shekhar (2020) proposed a multimodal framework for meme sentiment classification into positive, negative, and neutral classes. Behera et al. (2020) also proposed a multimodal approach for predicting the sentiment of the internet memes. Some researches also focused on multimodal sentiment analysis but not precisely on memes. For example, Poria et al. (2018a) studied the multimodal sentiment classification where the work explored deep learning-based architectures for combining the image and textual features. Similarly, Jiang et al. (2020) proposed a framework that utilizes visual and textual information for predicting the sentiment.

Differences with existing researches: The majority of the past studies readily focused on meme's sentiment analysis considering unimodal information. Though some works accomplished on multimodal sentiment analysis for English language, in other languages most of the sentiment analysis works based on the text modality. In our exploration, none of the research has been found on multimodal sentiment analysis for low-resource languages like Bengali in the context of memes. Though several researchers also explored multimodal sentiment analysis, only a few works are accomplished on internet memes. Moreover, the existing research analyses the memes embedded text written in English. However, memes can contain texts in a code-mixed and cross-lingual manner, which was overlooked in past studies. Considering these shortcomings, the proposed research differs from existing studies in four ways: *(i)* develop a meme sentiment analysis dataset for Bengali (i.e., MemoSen). This work is the first attempt in Bengali as far as we are concerned, *(ii)* provide a detailed annotation guidelines which can be followed in other languages for resource creation *(iii)* consider the memes that contain code-mixed (i.e., English and Bengali) and code-switched (called *Banglish*-a written form where the dialects of the Bengali language correspond in English characters) and *(iv)* evaluated the developed dataset using several state-of-the-art models.

3. MemoSen: a New Benchmark Dataset

As per our exploration, none of the datasets have been constructed to date in Bengali to perform multimodal sentiment analysis. This work developed **MemoSen**: a novel multimodal sentiment analysis dataset in Bengali. *MemoSen* is developed by following guidelines illustrated by Sharma et al. (2020). This section briefly describes the data accumulation process and annotation guidelines with detailed dataset statistics.

3.1. Data Accumulation

To construct the dataset, memes are collected manually from various social media platforms, such as Facebook, Twitter, and Instagram. A set of keywords such as *Bengali Memes*, *Bengali Funny Memes*, *Bengali Troll Memes*, *Bengali Celebrity Memes*, *Bengali Motivational Memes*, *Bengali Offensive Memes*, and *Bengali Political Memes* were used to search the memes. Subsequently, memes were acquired from relevant social media pages and public groups. Appendix B presents a detailed summary of the various sources of collected memes.

A total of 4700 memes were collected from February 2021 to September 2021. This work only considered memes with captions written in Bengali, Bengali and English (code-mixed) or in Banglish (code-switched) manner. The accumulation process contained some inappropriate memes, which are discarded based on the following bases: (i) memes that were unimodal (missing visual or textual information); (ii) memes whose texts are not readable; (iii) memes containing cartoons; (iv) repeated memes. Based on the above criterion **332** memes were removed and thus finished up with a dataset of **4368** memes. Some sample memes are presented in Figure C.1. Afterwards, we manually extracted the caption from the memes since no standard OCR system exists in Bengali. Finally, the memes and their associated captions are passed to the annotators to complete the manual annotation.

3.2. Dataset Annotation

The **MemoSen** is constructed by the manual labelling of the collected memes into three distinct sentiment categories: *Positive*, *Negative*, and *Neutral*. It is crucial to follow a uniform definition to distinguish among these categories, which reduces the annotation bias and helps to ensure the quality of the dataset.

3.2.1. Definition of Categories

After exploring the existing works on multimodal sentiment analysis (Soleymani et al., 2017; Cambria et al., 2018; Poria et al., 2018b), we differentiate the classes like the following: **Positive:** A meme is considered as positive if (i) it expresses affection, support, gratitude, accolade, and motivation; (ii) it has a humorous context that does not convey any covert intention to vilify, contempt or mock an entity[1].

Negative: A meme can be reckoned as a negative class if (i) it intends to denigrate, insult, disregard an entity based on its social, personal and organizational status; (ii) it expressed inappropriate cogitation such as obscene visual or textual content.

Neutral: A meme can be deemed as a neutral class if the expressed intention of the memes can not infer as positive or negative.

3.2.2. Process of Annotation

It is essential to have guidelines for the annotators to ensure the quality of the dataset (Liao et al., 2021). We asked the annotators to follow the class definition during labelling. At first, determine whether a meme expresses positive or negative sentiment. If yes, ascertain the reasons behind choosing the specific sentiment. This reasoning will help the expert when a disagreement arises between the annotators. A meme is considered neutral if it has no potential reasons to classify as positive or negative. To acquire quality labels, we trained the annotators with examples and ensured they could distinguish between the classes with proper reasoning. A manual annotation process was carried out by four annotators (graduate students having a computer engineering background). An expert verified all labels (see Appendix C for more detailed information of the annotators). Annotators were split into two groups (two in each), and each group labelled a different subset of memes. The expert decided the final label in case of disagreement between initial annotators. Final labels are determined by following the steps of Algorithm 1.

Investigation revealed that memes often use sarcastic words, making it difficult for the annotators to infer the sentiment correctly. For each meme m_i, check the two labels y_1, y_2 from the initial annotators. If they agree, it is considered the final label and included in $SL[]$. Otherwise, the expert checks the reasons and set the final labels upon discussion with annotators. Finally, the inter-annotator agreement is computed using Cohen (1960) Kappa coefficient to ensure data and annotation quality. A mean kappa score of 0.674 is obtained, indicating a moderate agreement between the annotators.

3.3. Dataset Statistics

The MemoSen is utilized to build the computational models for multimodal sentiment analysis. Thus, to perform the training and evaluation, the MemoSen is partitioned into three distinct sets: train (70%), test (20%), and validation (10%). Table 1 presents the class-wise distribution of each set. Out of 4368 memes, 1349 and 2728 memes are respectively from positive and negative classes, while 291 are from the neutral class. The distribution indicates that the dataset is imbalanced as the neutral class has only ($\approx$7%) data com-

[1]Here, entity denotes an individual, a group/community, an organization or the society.

Algorithm 1: Sentiment label assigning process

Input: Set of memes with associated captions
Output: Dataset with sentiment annotation

$M \leftarrow \{m_1, m_2, ..., m_n\}$ (set of collected memes);
$MemoSen \leftarrow []$ (Multimodal sentiment dataset);
$SL \leftarrow []$ (final sentiment labels of the memes);
$L[n][2] \leftarrow \{x_1, x_2, .., x_m\}$ (initial labels);

for $m_i \epsilon M$ **do**
    $y_1 = L[i][1]$ (first annotator label);
    $y_2 = L[i][2]$ (second annotator label);
    if $(y_1 == y_2)$ **then**
        $MemoSen.append(m_i)$;
        $SL.append(y_1)$;
    else
        1. expert resolve the issue;
        2. decide final label and add it to 'MemoSen'
    end
    $i = i + 1$;
end

pared to the positive (≈30%) and negative (≈63%) classes.

Class	Train	Test	Valid	Total
Positive	950	285	114	1349
Negative	2001	524	203	2728
Neutral	195	64	32	291

Table 1: Number of samples in train, test and validation set for each class.

Captions of the training memes are further investigated to acquire in-depth insights. Table 2 shows the detailed statistics of the captions, which illustrates that the negative class contributed ≈30K words whereas the positive class contained ≈17K. In contrast, the neutral type has approximately ten times fewer words (3k) than the positive class. Similarly, the negative class has the highest (≈8.8K) while the neutral type has the lowest (2.7K) number of unique words. On average, the positive class contained a maximum of 15 words, while the negative and neutral class consisted of ≈13 average number of words per caption.

	Positive	Negative	Neutral
#Words	16864	29443	3074
#Unique words	5745	8738	1702
Max. caption length	106	63	30
Avg. #words/caption	15.84	13.35	13.54

Table 2: Training set statistics for the captions of the memes

Figure 1 depicts the caption length-frequency distribution for each sentiment class. It is noticed that the majority of the captions' length in positive and negative classes lies between 7 to 25 words. Meanwhile, the neutral class has no captions with more than 30 words, while other classes have a small number of captions (<50) with a length greater than 30 words. We also carried out a quantitative analysis of MemoSen by measuring the Jaccard similarity index. Table 3 presents the similarity values, which are calculated between the most frequent 400 words of each class. The positive-negative pair obtain the highest similarity value of 0.355. On the other hand, the negative-neutral pair has approximately 1% more common words than the positive-neutral pair.

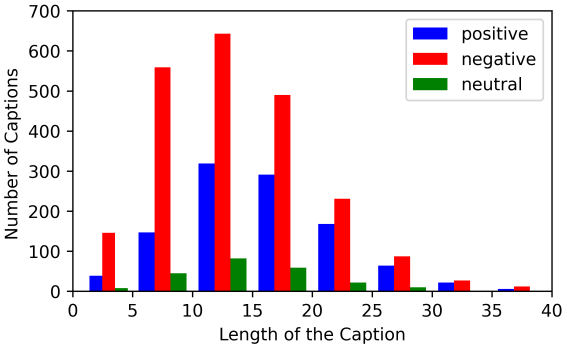


Figure 1: Histogram of the caption length of the memes for each class

	Positive	Negative	Neutral
Positive	-	**0.355**	0.213
Negative	-	-	0.228

Table 3: Jaccard similarity of 400 most frequent words between each pair of classes

4. Methodology

Several computational models are investigated considering unimodal data (i.e., images, texts) and the combination of both modalities (i.e., image+text) to classify the sentiment of the Bengali memes. For visual modality, state of the art pre-trained CNN networks (i.e., VGG19, VGG16, ResNet, DenseNet) are used. On the other hand, several machine learning (ML), deep neural networks (DNN), and transformer-based models are employed for the textual modality. Furthermore, we exploit both visual and textual features to acquire more robust inferences and develop several models using multimodal fusion approaches. Figure 2 shows the abstract view of the overall multimodal sentiment classification system.

4.1. Data Preprocessing

Before feeding them into the network, preprocessing the unimodal data (i.e., image, text) is required. On the

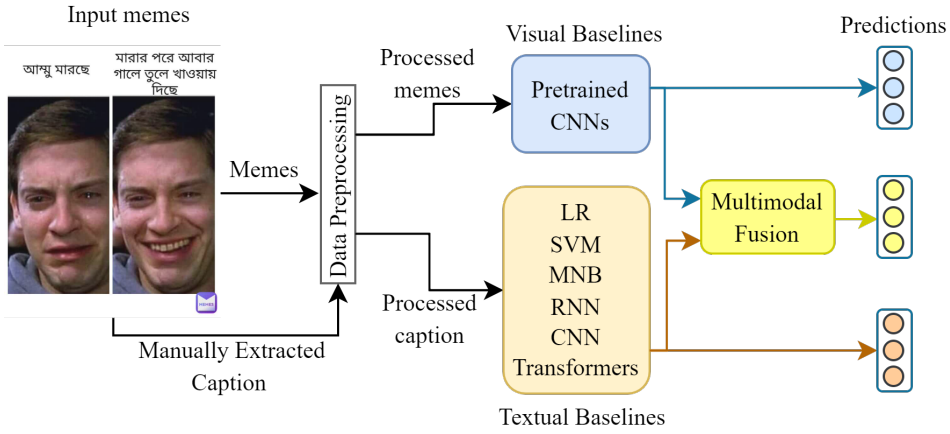


Figure 2: Abstract view of the Bengali meme sentiment classification system

visual modality side, the images are transformed into equal sizes of $150 \times 150 \times 3$. Then we use Keras[2] image preprocessing function to make them suitable before driving into the CNN models. In the case of the textual modality, DNN and transformer architectures require inputs in a specific format. For DNN, we convert the input texts into a vector of unique numbers by using the Keras tokenizer function. Subsequently, the padding method is applied to make the vectors of an equal length of 30. The length is determined by analyzing the caption length frequency distribution (described in Section 3.3). Similarly, we follow the transformer tokenization[3] method and use *encode_plus* function to encode the input texts. This method generates two vectors: unique Ids and attention masks, given as input to the transformers.

4.2. Baseline Models for Visual Modality

This work employed convolutional neural networks to classify the sentiment of the memes based on visual features. Rather than implementing custom neural networks, we adopt the transfer learning (Tan et al., 2018) approach. Several state of the art CNN architectures such as Xception (Chollet, 2017), VGG19, VGG16 (Simonyan and Zisserman, 2015), ResNet50 (He et al., 2016), and DenseNet121 (Iandola et al., 2014) are considered for the transfer learning. For sentiment classification, we keep the upper layers of the models non-trainable and use the weights that are already learned through training on the ImageNet (Deng et al., 2009) dataset for 1000 classes. The top two layers of the models are removed, and a global average max-pooling layer is added, followed by a softmax layer for the classification. Finally, the models are fine-tuned on the dataset.

[2]https://keras.io/

[3]https://huggingface.co/transformers/main classes/tokenizer.html

4.3. Baseline Models for Textual Modality

Various machine learning and deep learning models are investigated to obtain the features from the textual content. The architectures and parameters of the developed models are discussed in the subsequent subsections.

4.3.1. Machine Learning Based Methods

For the initial investigation, several ML-based methods such as LR (Hamdan et al., 2015), MNB (Hossain et al., 2021), and SVM (Zainuddin and Selamat, 2014) have been applied. We calculated the *term frequency-inverse document frequency (Tf-idf)* (Tokunaga and Makoto, 1994) values for the unigram features of the texts. We enabled the inverse document re-weighting technique during the calculation and settled the maximum and minimum document frequency value to 1.0. These Tf-idf values of unigram features have been used to train the ML models. The LR model is constructed with the 'lbfgs' optimizer and 'l2' regularization technique. In MNB, the additive smoothing parameter is set at 0.05 while the prior class probabilities are determined based on the number of instances in a class. Similarly, for SVM, we use the 'linear' kernel along with the 'l2' penalizer. Further, the 'tolerance' of stopping criterion and random state settle to 0.001 and 0, respectively.

4.3.2. Deep Learning Based Methods

Several popular deep learning-based models also investigated for textual sentiment classification task including BiLSTM (Hossain et al., 2020a), CNN (Ouyang et al., 2015), and the combination of BiLSTM and CNN (BiLSTM+CNN) (Sharif et al., 2020). Word embedding (Mikolov et al., 2013) features are used to train these models. To generate the embeddings, keras embedding layer is used which transforms each word into a 64-element vector that holds the semantic meaning of the words. Furthermore, the pre-trained transformer models are also exploited to develop more robust models.

BiLSTM: We construct a BiLSTM network of two layers, each associated with 32 and 16 units, respectively. Initially, the embedding features are propagated to the BiLSTM network. Afterwards, the output of the last BiLSTM layer is directly transferred to the softmax layer for the sentiment classification.

CNN: A two-layer CNN architecture is constructed that takes embedding features as input. The convolutional layers consist of 64 and 32 filters with kernel size (1×2). Max pooling operation is performed on the convolved features with a window of size 1×2 to extract more compact features.

BiLSTM+CNN: This method combined BiLSTM and CNN networks with slight modifications. The embedding features are passed to the BiLSTM layer that generates sentence embeddings of 64-dimension. This vector is propagated to the convolutional layer having 32 filters of kernel size 1×2. Following this, max-pooling is performed for further down-sampling of the features.
For all the models, the softmax layer utilizes the deep layers' features to obtain the models' predictions.

Transformers: Recent studies reveal that transformer (Vaswani et al., 2017) models trained on multilingual settings achieved outstanding result in solving various NLP problems (Naseem et al., 2020; Yang et al., 2020; Cao et al., 2020). As the task deals with a dataset of low-resource language, this work uses the monolingual, multilingual, and cross-lingual transformers for the investigation. This work employed three transformer-based models: (i) Multilingual Representations for Indian Languages (MuRIL) (Khanuja et al., 2021), (ii) Bidirectional Encoder Representations for transformers for Bangla language (Bangla-BERT) (Sarker, 2020), and (iii) Cross-lingual version of Robustly Optimized BERT (XLM-R) (Conneau et al., 2020). The models are taken from the huggingface[4] transformers library and fine-tuned on the developed dataset. We fetched the 'muril-base-cased', 'xlm-roberta-base', and 'bangla-bert-base' models and fine-tuned them using the textual content of the memes. Transformer models take 'Input Ids' and 'attention masks' as the input and generate contextualized sentence embeddings of 768-element vector. This vector is then passed to a fully connected (FC) layer of 28 neurons accompanied by a softmax layer for the classification. Before the softmax layer, we introduced a dropout layer with a 1% dropout rate to reduce the overfitting effect.

4.4. Multimodal Approach

In recent years, learning from multiple modalities (i.e., image, text, and speech) has proven effective in solving several NLP tasks such as visual question answering (Agrawal et al., 2015) and image captioning (Huang et al., 2019). In this work, we employed two main multimodal fusion methods, namely early or feature fusion (Natarajan et al., 2012) and late or decision fusion (Trong et al., 2020) to classify the sentiment of the memes by utilizing the multimodal information. However, considering the computational issues, for the multimodal experiments, we take one visual model (ResNet50) and the five deep learning-based methods (i.e., BiLSTM, CNN, BiLSTM+CNN, MurIL, and Bangla-BERT) described in Sections 4.2 and 4.3.2. The current work did not consider the XLM-R model due to its poor performance during the validation phase. The visual model is selected based on its outcome (i.e., Accuracy, f1-score) on the validation set. Therefore, by utilizing the two fusion methods on these six models, a total $((1 \times 5) \times 2) = 10$ multimodal models are developed where each fusion approach contributed of 5 different models.

Decision Fusion Based Models: For decision fusion, the softmax outputs of the visual and textual models are aggregated to make a joint representation of the multimodal features. These combined features are then passed to an FC layer of 4 neurons followed by a softmax layer for the classification.

Feature Fusion Based Models: In the feature fusion approach, at first, the softmax layer's of the visual and textual models are excluded and add an FC layer of 20 neurons at each modality side. Then concatenate on the FC layer's output of the visual and textual models. We passed these combined outcomes to another FC layer of 8 neurons, generating a learned representation of both features. Eventually, the softmax operation is performed to obtain the sentiment class prediction.

5. MemoSen: Benchmark Evaluation

A train set is used for the training, while the validation set helps to tweak the model parameters. The details of the hyperparameters are presented in Table A.1. Finally, the evaluation is performed with the unseen instances of the test set. As the MemoSen is imbalanced and consists of multiple classes, we chose weighted f_1-score (WF) as the primary metric for the evaluation. However, other metrics such as precision and recall are also considered for the comparison.

5.1. Results

Table 4 presents the outcome of the visual and textual models for sentiment classification of the memes.
In the case of visual models, VGG16, VGG19, and Xception showed varying WF ranging from $0.55 - 0.57$, whereas DenseNet and ResNet50 obtained a WF of approximately 0.60. However, in terms of precision and recall, only ResNet50 performed well and thus, it is considered the best visual model. Meanwhile, the textual model showed slightly improved performance. The MNB obtained the maximum weighted WF of 0.628 among ML-based methods, while LR and SVM got almost identical WF (0.608). Surprisingly,

[4]https://huggingface.co/

Approach	Models	P	R	WF
Visual	Xception	0.587	0.615	0.579
	VGG19	0.588	0.543	0.563
	VGG16	0.582	0.571	0.559
	ResNet50	0.602	0.628	0.600
	DenseNet	0.585	0.609	0.594
Textual	LR	0.617	0.663	0.608
	MNB	0.643	0.663	0.628
	SVM	0.670	0.653	0.608
	BiLSTM (B)	0.587	0.604	0.594
	CNN (C)	0.605	0.600	0.594
	B+C	0.606	0.554	0.576
	MurIL	0.624	0.640	**0.631**
	Bangla-BERT	0.622	0.605	0.605
	XLM-R	0.360	0.600	0.450

Table 4: Performance comparison of visual and textual models on the test set where P, R, WF denotes precision, recall and weighted f_1-score, respectively.

deep learning-based methods (BiLSTM, CNN, BiLSTM+CNN) performance are almost 3% lower than the MNB's outcome. However, the transformer model (MurIL) obtained the highest WF of 0.631 amid all the textual models.

After investigating the outcome of the unimodal models, we fused the information from both visual and textual modalities for getting robust inference. The results of the multimodal models are reported in Table 5.

		Models	P	R	WF
FF	R+	BiLSTM	0.625	0.633	0.626
		CNN	0.575	0.591	0.582
		BiLSTM+CNN	0.615	0.578	0.592
		MurIL	0.525	0.392	0.419
		Bangla-BERT	0.510	0.557	0.508
DF	R+	BiLSTM	0.644	0.631	0.635
		CNN	0.663	0.628	**0.643**
		BiLSTM+CNN	0.566	0.592	0.575
		MurIL	0.552	0.554	0.543
		Bangla-BERT	0.504	0.394	0.329

Table 5: Performance comparison of multimodal models on test set. Here, (+) sign denoted the aggregation of visual and textual models and R indicates the ResNet50 model. FF and DF denotes the feature fusion and decision fusion approaches.

Amid the feature fusion models, ResNet50+BiLSTM achieved a maximum WF of 0.626 while other models (CNN, BiLSTM+CNN) with ResNet50 obtained WF less than 0.60. On the other hand, ResNet50+BiLSTM got 0.635 WF with the decision fusion approach, which is $\approx 1\%$ higher than the best feature fusion based model (ResNet+BiLSTM). Though the model got a comparatively good outcome, it is the ResNet50+CNN model which achieved the highest WF of 0.643 and thus outperformed all the unimodal and multimodal models. One surprising result is noticed in the case of multimodal models developed with transformers (ResNet50+MruIL and ResNet50+Bangla-BERT), where the models did not achieve notable performance. We infer that the choice of the learning rate for the combined models is the possible reason behind this inferior result. As transformer models are trained with some specific 'lr' thus, when we aggregate the models with ResNet50, it does not provide the expected outcome with those 'lr' and thus degraded the overall performance.

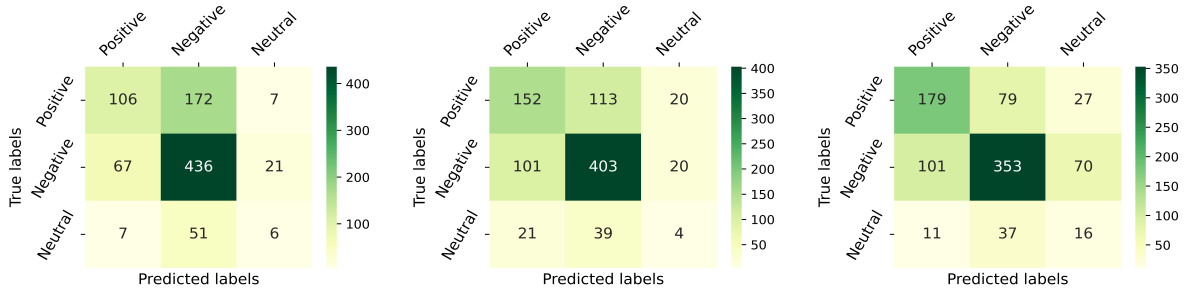
5.2. Error Analysis

The results showed that the multimodal approach is more efficient in classifying the memes' sentiment than the visual and textual approach. A detailed error analysis is carried out in quantitative and qualitative ways to acquire in-depth insights about the model's mistakes. The best visual and textual models are also considered for better demonstration with the multimodal model.

Quantitative Analysis: The quantitative error analysis is performed through the confusion matrices shown in Figure 3. Figures (a), (b), and (c) exhibited that, in case of 'Positive' class, the visual and textual model incorrectly classified 179 (Negative = 172, Neutral= 7) and 133 (Negative = 113, Neutral= 20) instances respectively amid 285 instances. However, when both modalities information's are fused, the number of misclassified instances get reduced to 105 (Negative = 79, Neutral= 27). Similarly, in 'Neutral' class, multimodal model misclassified 48 (Positive = 11, Negative= 37) among 64 samples and significantly improve the performance compared to the visual and textual models where each incorrectly identified 58 (Positive = 7, Negative= 51) and 60 (Positive = 21, Negative = 39) instances respectively. However, the multimodal model did not improve with the 'Negative' class. Although its performance (misclassified 171 instances among 524) degraded with the 'Negative' class, the multimodal model outperformed the unimodal models due to the higher number of accurate predictions in 'Positive' and 'Neutral' classes.

The analysis revealed that the visual information is more appropriate for predicting the negative sentiment of the memes as only 88 instances from 524 are incorrectly classified by the visual model. Meanwhile, visual and textual information is required to get a robust inference for the positive and neutral sentiment.

Qualitative Analysis: Figure 4 provides some example memes that demonstrates how the multimodal approach captures information more effectively and thus lead to better predictions in contrast to the visual and textual models. For instance, in Figure 4 (a), the visual model predicts the meme as 'Negative' whereas the textual model inferred it as a 'Neutral' meme. However, when the visual and textual information is jointly fused, the multimodal model is correctly predicted as a 'Positive' meme. A similar scenario is observed in Figure 4 (b), the visual model reckon it as 'Neutral', and the textual model is considered as 'Positive' meme.



(a) Best Visual (ResNet50) (b) Best Textual (MurIL) (c) Best Multimodal (ResNet50+CNN)

Figure 3: Confusion matrices of unimodal and multimodal models.

(a) **Visual Model:** Negative (✗)
Textual Model: Neutral (✗)
Multimodal Model: Positive (✓)

(b) **Visual Model:** Neutral (✗)
Textual Model: Positive (✗)
Multimodal Model: Negative (✓)

(c) **Visual Model:** Negative (✗)
Textual Model: Negative (✗)
Multimodal Model: Positive (✗)

Figure 4: Example memes where aggregation of the visual and textual modalities yield better predictions. The symbol (✓) and (✗) indicates the correct and incorrect prediction respectively.

Unfortunately, both predictions were incorrect and accurately identified as a 'Negative' meme by the multimodal model. Meanwhile, Figure 4 (c) shows an example where all the models provide incorrect predictions. This outcome illustrates the problematic nature of the memes. It indicates that there is still ample room for improvement and proves the effectiveness of the joint evaluation of multimodal information.

In summary, the error analysis reveals that the model's performance is more biased towards negative class. Perhaps an imbalanced dataset is the main reason for this inclination. Though the overall performance of the multimodal model is improved, the performance across the categories bring some observations that should tackle in future to improve the model's efficacy. One of the profound reasons we found that a large number of words are overlapped between the classes, which is also evident from the Jaccard similarity values (described in Section 3.3). Moreover, the code-mixed and code-switched words might make it more difficult for the model to understand the semantics and overall context. Finally, the consistent visual features (i.e., familiar person faces) across the memes of the different classes also made it arduous for the models to differentiate the appropriate category. Indeed, these significant factors impose a barrier in the Bengali meme sentiment classification problem.

6. Conclusion and Future Work

This paper presented a multimodal classification framework for classifying sentiment from memes in Bengali. For this purpose, this work introduced *MemoSen*, a multimodal benchmark dataset consisting of 4368 memes with three sentiment classes (i.e., positive, negative, neutral). Several computational models have been explored to benchmark the MemoSen, considering only visual, only textual, and both modalities. The key finding is that in classifying memes' sentiments, the incorporation of multimodal information provides more robust inference than the unimodal information. The error analysis revealed that all the models are suffering in identifying the neutral memes, which indicates that more sophisticated models are required to solve this problem with higher accuracy. In future, we aim to alleviate the model biases on specific classes by employing state-of-the-art multimodal models such as ViL-Bert, VisualBert, and CLIP encoder. Since these models are well suited for English, the future attempt will explore the intra-modal and cross-modal attention techniques to improve the sentiment classification of memes.

7. Bibliographical References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. (2015). Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31.

Alam, M. H., Rahoman, M.-M., and Azad, M. A. K. (2017). Sentiment analysis for bangla sentences using convolutional neural network. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., and Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, Georgia, June. Association for Computational Linguistics.

Behera, P., ., M., and Ekbal, A. (2020). Only text? only image? or both? predicting sentiment of internet memes. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 444–452, Indian Institute of Technology Patna, Patna, India, December. NLP Association of India (NLPAI).

Bender, E. M. and Friedman, B. (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.

Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232.

Cambria, E., Hazarika, D., Poria, S., Hussain, A., and Subramanyam, R. B. V. (2018). Benchmarking multimodal sentiment analysis. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 166–179, Cham. Springer International Publishing.

Cao, Q., Trivedi, H., Balasubramanian, A., and Balasubramanian, N. (2020). Deformer: Decomposing pre-trained transformers for faster question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4497.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807.

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

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

French, J. H. (2017). Image-based memes as sentiment predictors. In *2017 International Conference on Information Society (i-Society)*, pages 80–85. IEEE.

Hamdan, H., Bellot, P., and Bechet, F. (2015). Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 753–758.

Hameed, Z. and Garcia-Zapirain, B. (2020). Sentiment classification using a single-layered bilstm model. *Ieee Access*, 8:73992–74001.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hossain, E., Sharif, O., Hoque, M., and Sarker, I. H. (2020a). Sentilstm: A deep learning approach for sentiment analysis of restaurant reviews. In *HIS*.

Hossain, E., Sharif, O., Hoque, M. M., and Sarker, I. H. (2020b). Sentilstm: A deep learning approach for sentiment analysis of restaurant reviews.

Hossain, E., Sharif, O., and Moshiul Hoque, M. (2021). Sentiment polarity detection on bengali book reviews using multinomial naive bayes. In *Progress in Advanced Computing and Intelligent Engineering*, pages 281–292. Springer.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). Attention on attention for image captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4633–4642.

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., and Keutzer, K. (2014). Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*.

Islam, K. I., Islam, M. S., and Amin, M. R. (2020). Sentiment analysis in bengali via transfer learning using multi-lingual bert. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE.

Jiang, T., Wang, J., Liu, Z., and Ling, Y. (2020). Fusion-extraction network for multimodal sentiment analysis. *Advances in Knowledge Discovery and Data Mining*, 12085:785.

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., et al. (2021). Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Kumar, A. and Jaiswal, A. (2017). Image sentiment analysis using convolutional neural network. In *International Conference on Intelligent Systems Design and Applications*, pages 464–473. Springer.

Li, L., Jamieson, K. G., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2016). Efficient hyperparameter optimization and infinitely many armed bandits. *CoRR*, abs/1603.06560.

Li, W., Liu, P., Zhang, Q., and Liu, W. (2019a). An improved approach for text sentiment classification based on a deep neural network via a sentiment attention mechanism. *Future Internet*, 11(4):96.

Li, Z., Fan, Y., Jiang, B., Lei, T., and Liu, W. (2019b). A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications*, 78(6):6939–6967.

Liao, S., Wang, J., Yu, R., Sato, K., and Cheng, Z. (2017). Cnn for situations understanding based on sentiment analysis of twitter data. *Procedia computer science*, 111:376–381.

Liao, Y.-H., Kar, A., and Fidler, S. (2021). Towards good practices for efficiently annotating large-scale image classification datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4350–4359, June.

Mamta, Ekbal, A., Bhattacharyya, P., Srivastava, S., Kumar, A., and Saha, T. (2020). Multi-domain tweet corpora for sentiment analysis: Resource creation and evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5046–5054, Marseille, France, May. European Language Resources Association.

Mamun, M. M. R., Sharif, O., and Hoque, M. M. (2022). Classification of textual sentiment using ensemble technique. *SN Comput. Sci.*, 3:49.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR*.

Miller, D. and Sinanan, J. (2017). *Visualising Facebook: a comparative perspective*. UCL Press.

Murthy, D., Allu, S., Andhavarapu, B., Bagadi, M., and Belusont, M. (2020). Text based sentiment analysis using lstm. *Int. J. Eng. Res. Tech. Res*, 9(05).

Naseem, U., Razzak, I., Musial, K., and Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113:58–69.

Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R., and Natarajan, P. (2012). Multimodal feature fusion for robust event detection in web videos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1305. IEEE.

Ouyang, X., Zhou, P., Li, C. H., and Liu, L. (2015). Sentiment analysis using convolutional neural network. In *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing*, pages 2359–2364. IEEE.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., and Hussain, A. (2018a). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., and Hussain, A. (2018b). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25.

Pranesh, R. R. and Shekhar, A. (2020). Memesem: A multi-modal framework for sentimental analysis of meme via transfer learning.

Röttger, P., Vidgen, B., Hovy, D., and Pierrehumbert, J. B. (2021). Two contrasting data annotation paradigms for subjective nlp tasks.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2021). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.

Sarker, S. (2020). Banglabert: Bengali mask language model for bengali language understading.

Sharif, O., Hossain, E., and Hoque, M. M. (2020). TechTexC: Classification of technical texts using convolution and bidirectional long short term memory network. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): TechDOfication 2020 Shared Task*, pages 35–39, Patna, India, December. NLP Association of India (NLPAI).

Sharma, C., Bhageria, D., Scott, W., PYKL, S., Das, A., Chakraborty, T., Pulabaigari, V., and Gambäck, B. (2020). SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online), December. International Committee for Computational Linguistics.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer.

Tokunaga, T. and Makoto, I. (1994). Text categorization based on weighted inverse document frequency. In *Special Interest Groups and Information Process Society of Japan (SIG-IPSJ*. Citeseer.

Trong, V. H., Gwang-hyun, Y., Vu, D. T., and Jin-young, K. (2020). Late fusion of multimodal deep neural networks for weeds classification. *Computers and Electronics in Agriculture*, 175:105506.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit,

J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *ArXiv*, abs/1706.03762.

Walińska, U. and Potoniec, J. (2020). Urszula walińska at SemEval-2020 task 8: Fusion of text and image features using LSTM and VGG16 for memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1215–1220, Barcelona (online), December. International Committee for Computational Linguistics.

Yang, J., Wang, M., Zhou, H., Zhao, C., Zhang, W., Yu, Y., and Li, L. (2020). Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9378–9385.

You, Q., Luo, J., Jin, H., and Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-ninth AAAI conference on artificial intelligence*.

Zainuddin, N. and Selamat, A. (2014). Sentiment analysis using support vector machine. In *2014 international conference on computer, communications, and control technology (I4CT)*, pages 333–337. IEEE.

Appendix

A. Implementation Settings and Hyperparameters

For the experimentation, we use the Google colaboratory a GPU facilitated platform embedded with Python 3. The machine learning models are implemented using the scikit-learn (0.22.2) packages, while we utilize the TensorFlow (2.3.0) framework for deep learning models. All the models are compiled using the 'sparse categorical cross-entropy loss function' and trained with an 'adam' optimizer for 40 epochs. The visual, some textual (BiLSTM, CNN, BiLSTM+CNN), and multimodal models (ResNet50+BiLSTM, ResNet50+CNN, ResNet50+BiLSTM+CNN) use learning rate of $1e^{-3}$. Likewise, XLM-R and MuRiL models utilize a learning rate $1e^{-5}$ while other transformers (Bangla-BERT) and multimodal models (ResNet50+MurIL, ResNet50+Bangla-BERT) are trained with a learning rate of $3e^{-5}$. All the models utilize a batch size of 32 except the transformer models (batch size = 16). We employed Keras 'callbacks' method to save the best intermediate model during training. The optimum value of the hyperparameters is determined by utilizing the Keras Hyperband (Li et al., 2016) tuner. The summary of the hyperparameters is provided in Table A.1.

Approach		Models	LR	Batch Size
Text		BiLSTM	$1e^{-3}$	32
		CNN	$1e^{-3}$	32
		BiLSTM+CNN	$1e^{-3}$	32
		MurIL	$1e^{-5}$	16
		Bangla-BERT	$3e^{-5}$	16
		XLM-R	$3e^{-5}$	16
Multimodal	R+	BiLSTM	$1e^{-3}$	32
		CNN	$1e^{-3}$	32
		BiLSTM+CNN	$1e^{-3}$	32
		MurIL	$3e^{-5}$	16
		Bangla-BERT	$3e^{-5}$	16

Table A.1: Hyperparameter values for textual and multimodal models.

B. Data Sources

Figure B.1 depicts the distribution of various sources from where memes were collected. Most of the memes were accumulated from Facebook, while only a fraction of the memes were gathered from Instagram and other sources. On the other hand, amid the collected memes, an ample amount (31%) were found through the Bangla Offensive Memes keyword, whereas about 42% were downloaded using the keywords of Bangla Funny and Troll Memes.

C. Annotators Information

How to reduce bias and acquire correct annotations is a critical question to answer while labelling a dataset

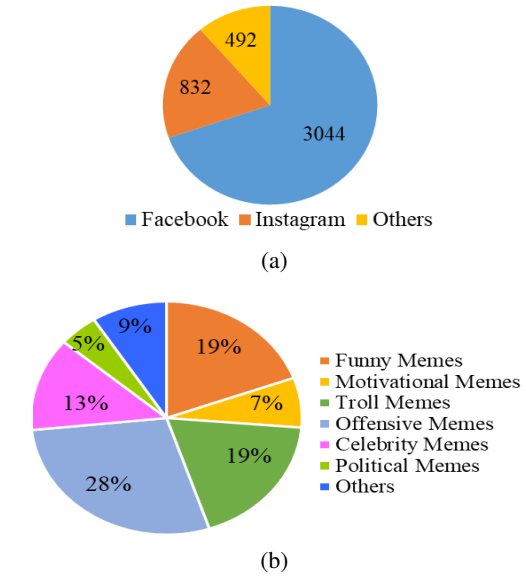
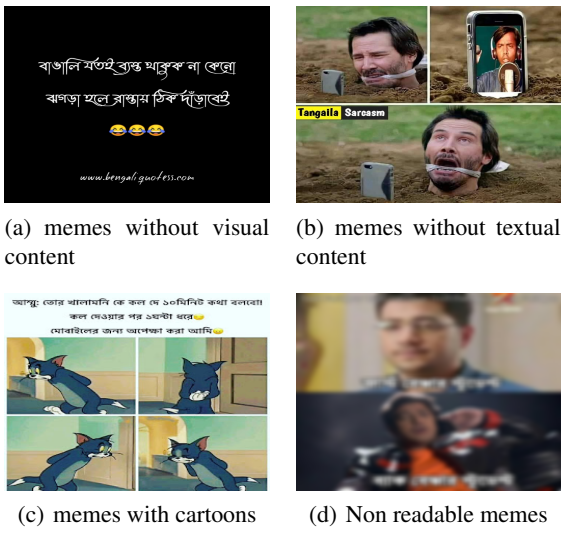


Figure B.1: Statistics of the MemoSen dataset: (a) Distribution of the sources, (b) Percentage of memes culled using the corresponding searched keywords.



(a) memes without visual content

(b) memes without textual content

(c) memes with cartoons

(d) Non readable memes

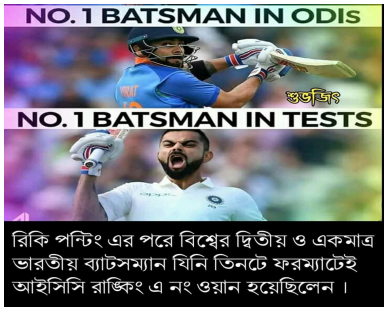
Figure B.2: Few sample of memes that were removed during the process of data collection

(Bender and Friedman, 2018). Several studies (Sap et al., 2021; Röttger et al., 2021) have emphasized knowing the identity of the annotators since their experience and perception influence the annotation. Therefore, relevant information of annotators involved in dataset development are briefly summarized in Table C.1. Four annotators and an expert were involved in the labelling process. Annotators' ages were between 23-27 years, having bachelor degrees in computer engineering and 1-4 years of experience performing research in NLP. Three out of four annotators have prior knowledge of annotating related sentiment samples. All the annotators are native Bengali speakers and have experience using Banglish, Bangla-English code-mixed language.

	Annotator-1	Annotator-2	Annotator-3	Annotator-4	Expert
Research-status	Postgrad	RA	Postgrad	RA	Professor
Research area	NLP	NLP	NLP	NLP	NLP, HCI, Robotics
Experience (in years)	1.5	3	1	4	21
Prior experience of annotation	yes	yes	no	yes	yes
Age	23	26	24	27	46
Religion	Islam	Islam	Hindu	Islam	Islam
Gender	Male	male	Female	Male	Male

Table C.1: Research experience and demographic information summary of the annotators.

(a) meme shows affection



(b) meme shows accolade

(c) meme shows funny humor

(d) insult a person

(e) denigrate a group of celebrities

(f) shows obscene content

(g) memes with inherent sentiment

(h) memes intention is incomprehensible

Figure C.1: Few example memes from **MemoSen**: here (a,b,c) are the positive memes, (d,e,f) are the negative and (g,h) are the neutral memes.