# Deciphering Hate: Identifying Hateful Memes and Their Targets

**Eftekhar Hossain♣, Omar Sharif♠, Mohammed Moshiul Hoque¥, Sarah M. Preum♠**

¥Department of Computer Science and Engineering
♠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

Internet memes have become a powerful means for individuals to express emotions, thoughts, and perspectives on social media. While often considered a source of humor and entertainment, memes can also disseminate hateful content targeting individuals or communities. Most existing research focuses on the negative aspects of memes in high-resource languages, overlooking the distinctive challenges associated with low-resource languages like Bengali (also known as Bangla). Furthermore, while previous work on Bengali memes has focused on detecting hateful memes, there has been no work on detecting their targeted entities. To bridge this gap and facilitate research in this arena, we introduce a novel multimodal dataset for Bengali, **BHM** (Bengali Hateful Memes). The dataset consists of 7,148 memes with Bengali as well as code-mixed captions, tailored for two tasks: *(i) detecting hateful memes*, and *(ii) detecting the social entities they target (i.e., Individual, Organization, Community, and Society)*. To solve these tasks, we propose DORA (Dual cO-attention fRAmework), a multimodal deep neural network that systematically extracts the significant modality features from the memes and jointly evaluates them with the modality-specific features to understand the context better. Our experiments show that DORA is generalizable on other low-resource hateful meme datasets and outperforms several state-of-the-art rivaling baselines.

## 1 Introduction

In recent years, social media has brought a distinct form of multimodal entity: *memes*, providing a means to express ideas and emotions. Memes are compositions of images coupled with concise text. While memes are often amusing, they can also spread hate by incorporating socio-political elements. These hateful memes pose a significant threat to social harmony as they have the potential to harm individuals or specific groups based on



Figure 1: Example of hateful memes with associated targets. The first meme directly refers to a telecom organization as a bandit, and the second one deliberately attacks a religious community.

factors like political beliefs, sexual orientation, or religious affiliations. Despite the significant influence of memes, their multifaceted nature and concealed semantics make them very hard to analyze. The prevalence of highly toxic memes in recent times has led to a growing body of research into the negative aspects of memes, such as hate (Kiela et al., 2020), offensiveness (Shang et al., 2021), and harm (Pramanick et al., 2021b). However, most works focused on the memes in high-resource languages, while only a few studied the objectionable (i.e., hate, abuse) memes of low-resource languages (Kumari et al., 2023). This is particularly true for Bengali.

Despite being the seventh most widely spoken language, having 210 million speakers globally, Bengali is considered one of the notable resource-constrained languages (Das and Mukherjee, 2023). Moreover, it is the official language of Bangladesh and holds recognition as one of the official languages in the constitution of India. So, developing resources in Bengali is important to build more inclusive language technologies. Statistics indicate that over 45 million users engage with Bengali on various social media platforms daily (Sharif and Hoque, 2022). Recently, memes have gained significant traction in social media, reaching a broad

8347

audience and influencing public sentiment. Many of these memes contain hateful content targeting various social entities. The limited availability of the benchmark datasets primarily constrains the identification of such hateful memes. Only two prior works (Karim et al., 2022; Hossain et al., 2022) developed hateful meme datasets in Bengali. However, both of them overlook the targets associated with the hateful memes. For instance, the first meme in figure 1 is hateful towards an organization because it depicts a company (i.e., Grameenphone) as a robber. Similarly, the second meme propagates hate towards a specific religious community (i.e., Muslim) by highlighting that they produce many children. Therefore, identifying targets of hateful memes is crucial for 1) understanding targeted groups and developing interventions to counter hate speech and 2) personalizing content filters, ensuring that users are not exposed to hateful content directed at them or their communities.

To bridge this research gap, we develop a novel Bengali meme dataset encompassing the targeted entities of hate. The captions in the dataset contain code-mixed (Bengali + English) and code-switched text (written Bengali dialects in English alphabets). This makes the dataset more distinctive and challenging compared to previous studies in the field. On the technical front, prior research on hateful meme detection (Kiela et al., 2019; Pramanick et al., 2021a) revealed that off-the-shelf multimodal systems, which often perform well on a range of visual-linguistic tasks, struggle when applied to memes. Besides, most of the existing visiolinguistic models (Radford et al., 2021; Li et al., 2019, 2022) are primarily trained on image-text pairs of English languages, thus limiting their capability on low-resource languages. Moreover, the existing state-of-the-art multimodal models (Pramanick et al., 2021c; Lee et al., 2021) for hateful meme detection can not be replicated because several components of their architectures are not available in low-resource languages (i.e., Bengali). To tackle these issues, we developed a multimodal framework, and our major contributions are as follows:

- We develop a benchmark multimodal dataset comprising 7,148 Bengali memes. The dataset includes two sets of labels for (i) detecting hateful memes and (ii) identifying targeted entities (individuals, organizations, communities, and society). We also provide detailed annotation guidelines to facilitate resource creation for other low-resource languages in this domain.

- We propose DORA, a multimodal framework to identify hateful memes and their targets. We also perform extensive experiments on **BHM** dataset and show that DORA outperforms nine state-of-the-art unimodal and multimodal baselines in terms of all the evaluation measures. We further establish the generalizability and transferability of DORA on two existing benchmark hateful meme datasets in Bengali and Hindi.

## 2 Related Work

**Hateful memes dataset:** Over the past few years, several meme datasets have been developed regarding hate speech and its various intensity levels, such as offense, harm, abuse, and troll. Sabat et al. (2019) developed a hateful memes dataset containing 5,020 memes. Facebook AI (Kiela et al., 2020) contributed to developing a hateful memes dataset consisting of around 10K synthetic memes labeled in hateful and not-hateful classes. Similarly, another large-scale dataset comprising 150K memes was introduced by Gomez et al. (2020) for hate speech detection. In another work, Suryawanshi et al. (2020) developed an offensive memes dataset comprising 743 memes collected during the event of the 2016 US presidential election. Pramanick et al. (2021a) built a dataset to detect harmful memes containing around 3.5K related to COVID-19.

Some works have also been conducted concerning low-resource languages. Perifanos and Goutsos (2021) developed a dataset of 4,004 memes for detecting hate speech in Greek. Kumari et al. (2023) developed an offensive memes dataset consisting of 7,417 Hindi memes. Recently, Das and Mukherjee (2023) introduced a dataset comprising 4,043 samples for detecting the Bengali abusive memes. Two prior studies focused on hateful meme detection in the Bengali language. Karim et al. (2022) introduced a synthetic hate speech dataset comprising around 4,500 Bengali memes. Similarly, Hossain et al. (2022) developed another Bengali hateful memes dataset having 4,158 memes labeled hateful or not-hateful. However, none of these datasets annotated the targets of hateful memes.

**Multimodal hateful meme detection:** Over the years, various approaches have been employed for

detecting hate speech using multimodal learning. Earlier researchers used conventional fusion (i.e., early and late) (Suryawanshi et al., 2020; Gomez et al., 2020) produce producing multimodal representation. Later, some researchers employed bilinear pooling (Chandra et al., 2021) while others developed transformer-based multimodal architectures such as MMBT (Kiela et al., 2019), ViLBERT (Lu et al., 2019), and Visual-BERT (Li et al., 2019). Besides, some works attempted to use disentangled learning (Lee et al., 2021), incorporate additional caption (Zhou et al., 2021), and add external knowledge (Pramanick et al., 2021c) to improve the hateful memes detection performance. Recently, Cao et al. (2022) applied prompting techniques for hateful meme detection in English.

**Differences with existing studies:** Though many studies have been conducted on hateful meme detection, only a few studies have focused on Bengali. We point out several drawbacks in the existing research on Bengali. Firstly, the existing hateful memes datasets are small. They framed the task as a binary (hateful or not-hateful) classification problem, overlooking the social entities (i.e., individuals, society) that a hateful meme can target. Our dataset provides both levels of annotation. The richness of our dataset contributes to a comprehensive understanding of the dynamics of Bengali memes. Only one particular work (Das and Mukherjee, 2023) studied the targets of abusive memes in Bengali. Their work is completely different from ours as we attempt to identify targeted social entities of hateful memes, which are more explicit than abuse. Secondly, the current research overlooked the memes containing cross-lingual captions, while many internet memes are written in code-mixed and code-switched manner. Lastly, none of the works provided any model that can be generalized across low-resource datasets.

# 3  BHM: A New Benchmark Dataset

As per our exploration, no dataset in Bengali currently focuses on capturing hateful memes that target specific entities. To fill this gap, we develop a novel multimodal hateful meme dataset. We followed the guidelines outlined by Hossain et al. (2022) and Kiela et al. (2020) to develop the dataset. This section will provide a detailed discussion of the dataset development steps, including the data collection and annotation process and relevant statistics.

## 3.1  Data Collection and Sampling

We collected memes from various online platforms, including Facebook, Instagram, Pinterest, and Blogs. We used keywords such as *"Bengali Troll Memes", "Bengali Faltu Memes", "Bengali Celebrity Memes", "Bengali Memes", "Bengali Funny Memes", "Bengali Political Memes"*, etc to search for these memes. To avoid copyright infringement, we only collect the memes from publicly accessible pages and groups. We accumulated a total of 7,532 memes from March 2022 to April 2023. The distribution of data sources is presented in Appendix B. We collected the memes with Bengali and code-mixed (Bengali + English) embedded texts. During the data collection, we filtered out memes if they (i) contained only visual or textual information, (ii) contained drawings or cartoons, and (iii) had uncleared contents either visually or textually. Appendix Figure B.3 presents some filtered meme samples. We also removed duplicate memes. After filtering and deduplication, we discarded 299 memes and ended up with a curated dataset of 7,233 memes. Following these, we use an OCR library (PyTesseract[1]) to extract captions from the memes. As the captions have code-mixed texts, we manually reviewed and corrected if there were any missing words or spelling errors in the extracted captions. Finally, the memes and their captions are passed for manual annotation.

## 3.2  Dataset Annotation

We develop **BHM** focusing on two tasks: (i) detecting whether a meme is hateful or not and (ii) identifying the targeted entity of a hateful meme. We create guidelines defining the tasks to ensure the annotation quality and mitigate the bias. Appendix Figure B.1 depicts a few annotated meme examples.

### 3.2.1  Definition of Categories

Following the definition of previous studies (Kiela et al., 2020; Hossain et al., 2022), we consider *a meme as hateful if it explicitly intends to denigrate, vilify, harm, mock, abuse any entity based on their gender, race, ideology, belief, social, political, geographical and organizational status*. Moreover, we define four target categories of hateful memes that the annotators can adhere to during annotation. The four target categories are as follows[2]:

---

[1]https://pypi.org/project/pytesseract/
[2]The definitions reference individuals and organizations are based in Bangladesh.

1. **Targeted Individual (TI)**: The hate directed towards a specific person (male or female) based on his fame, gender, race, or status. The person might be an artist, an actor, or a well-known politician such as *Sakib Khan, Mithila, Khaleda Jia, Sajeeb Wazed, Tamim Iqbal* etc.

2. **Targeted Organization (TO)**: When the hate propagates towards any particular organization which is a group of people having certain goals such as a business company (e.g., *Grameenphone, Airtel*), government institution(e.g., *School & Colleges*), political organization (e.g., *BNP, Awami League*), etc.

3. **Targeted Community (TC)**: Hate on any specific group of people who hold common beliefs or ideology towards any religion (e.g., *who follow the ideology of Buddhism*), culture (e.g., *who celebrates the Bengali Pohela Baisakh or valentines day*), person (e.g., *followers of cricketer Tamim Iqbal*) or organization (e.g., *followers of Bangladesh National Party*).

4. **Targeted Society (TS)**: When a meme promotes hate towards a group of people based on their geographical areas such as mocking entire *Indian people* or *British people* it becomes hateful to an entire society.

### 3.2.2 Annotation Process

To carry out the annotation process, we employed six annotators: four were undergraduate students, and two were graduate students, falling within the age range of 23 to 27 years. The group comprised four male and two female annotators, each possessing prior research experience in the field of NLP. Furthermore, to resolve any disagreement among the annotators, we included an expert with 15 years of experience in NLP. We divided the annotators into three groups of two people, each annotating different subsets of memes. Initially, we trained our annotators with the definition of hateful memes, their categories, and associated samples. Our primary goal was to ensure that the annotators comprehended the guidelines and identify hateful memes and their target entities.

Two annotators independently annotated each meme, and the final label was determined through consensus between the annotators. On average, an annotator spent 3 minutes deciding the label of a meme. In disagreement, an expert provided the

final decision by discussing the uncertainties. For a minimal number of memes ($< 2\%$), we observe that memes target multiple entities. Since this number is minimal for annotation simplicity, such samples were annotated with the dominant class label. During the final label assignment, we discarded 85 memes as the annotators, and the expert could not agree on assigning a label. Finally, we get BHM, a multimodal Bengali hateful memes dataset with their targeted entities containing 7,148 memes.

| | Label | $\kappa$-score | Average |
|---|---|---|---|
| Task 1 | Hate | 0.82 | 0.79 |
| | Not Hate | 0.76 | |
| Task 2 | Target Individuals | 0.68 | 0.63 |
| | Target Organizations | 0.66 | |
| | Target Communities | 0.61 | |
| | Target Society | 0.57 | |

Table 1: Cohen's $\kappa$ agreement score during the annotation of each task: Task 1: hateful meme detection (2-class classification) and Task 2: target identification (4-class classification) of hateful memes.

**Inter-annotator Agreement:** We computed the inter-annotator agreement in terms of Cohen's Kappa ($\kappa$) score (Cohen, 1960) to check the validity. Table 1 shows the Kappa scores for each task category. We achieved a high agreement score of 0.79 for the hateful meme detection task. However, for target identification, we attained a moderate score of 0.63. These agreement scores indicate that annotators struggled distinguishing the targeted entities within hateful memes.

### 3.3 Dataset Statistics

We divided the dataset into training (80%), validation (10%), and test (10%) sets for model training and evaluation. Table 2 shows the data distribution across different categories within each split. Task 1 exhibits a slight imbalance, while task 2 presents a significant imbalance, with most data falling under the 'TI' category. The distribution highlights that it will be challenging to accurately identify the targeted entities in hateful memes due to the limited number of samples in the 'TO', 'TC', and 'TS' categories. We analyze the training set memes to acquire more insights into data characteristics. The analysis is presented in Appendix C.

## 4 Methodology

This section describes the proposed multimodal framework for detecting hateful memes and their

| | Class | Train | Valid | Test | Total |
|---|---|---|---|---|---|
| Task 1 | HT | 2117 | 241 | 266 | 2624 |
| | NHT | 3641 | 399 | 445 | 4485 |
| Task 2 | TI | 1623 | 192 | 193 | 2008 |
| | TO | 160 | 17 | 27 | 204 |
| | TC | 249 | 24 | 37 | 310 |
| | TS | 85 | 8 | 9 | 102 |

Table 2: Number of memes in train, test, and validation set for each category.

targeted entities. Figure 2 shows the overall architecture of the proposed system.

## 4.1 Feature Extractor

To encode the visual information of the memes, we leverage the image encoder component of the CLIP (Contrastive Language Image Pretraining) (Radford et al., 2021), a prominent visio-linguistic model. This image encoder incorporates a vision transformer (Dosovitskiy et al., 2020) as its backbone. Meanwhile, we leverage the XGLM (Lin et al., 2022), a multilingual generative language model, to encode meme captions. XGLM has effectively learned from diverse languages within context without parameter updates. Given that our dataset comprises code-mixed captions (Bangla + English), we posited that XGLM could offer a better contextualized representation of these code-mixed captions. We fine-tuned both the image and text encoders, aiming to extract encoded representations. These representations were subsequently fed into a dual co-attention module to produce a multimodal representation.

## 4.2 Dual Co-Attention

We fed the encoded visual and textual features into a dual co-attention block to generate an effective multimodal representation. Specifically, we generate two attentive multimodal feature representations using the Multi-head Self Attention (MSA) mechanism. The MSA takes three matrices: Query (Q), Key (K), and Value (V) as input. In standard NLP applications, all the matrices come from word representations. In contrast, motivated by Lu et al. (2019), we modified the MSA block where queries come from one modality and keys and values from another. This modification will generate an attention-pooled representation for one modality conditioned on another.

Specifically, in our case, the Q will be generated from visual features, whereas the K will be from textual features. Afterward, to determine the similarity between the visual and textual features, we calculated the attention values by performing a dot product between Q and K. We then generated two different Value matrices, one coming from visual features and another from textual features. The goal was to generate two attentive representations where one modality guided another.

To do this, we first weighed the visual features by performing a point-wise multiplication with attention scores and named it vision-guided attentive representation (VGAR). After observing the visual information, we also generate another attentive representation by weighing the textual features. We called this text-guided attentive representation (TGAR). These two attentive representations (VGAR and TGAR) now contain significant cross-
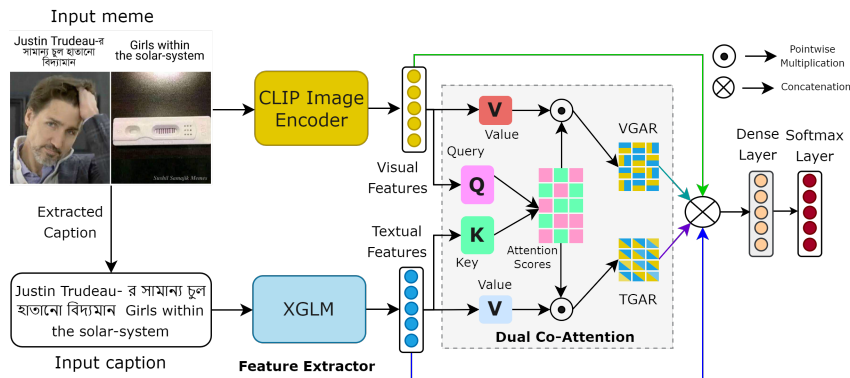


Figure 2: A simplified view of our proposed Dual Co-Attention Framework (DORA). The upper block represents the visual feature extractor, and the lower block is the textual feature extractor. The Dual Co-Attention block takes encoded visual and textual representation and generates two attentive vectors: VGAR (Vision-guided attentive Representation) and TGAR (Text-guided Attentive Representation). Finally, our method generates a richer multimodal representation by concatenating the attentive vectors with the individual modality-specific features.

modal features. This cross-modal representation is further concatenated with the individual modality features (obtained from CLIP and XGLM). This process will boost the gradient flow and help the model learn from the individual and their cross-modal features. Finally, the combined multimodal representation is passed to the dense layer, followed by a softmax operation to predict the meme's category.

# 5 Experiments

This section discusses the baselines and their performance comparison with the proposed method (DORA) and its variants. We developed several state-of-the-art computational models, including only visual, textual, and multimodal models pre-trained on both modalities. We used weighted F1 scores as the primary evaluation metrics. Other metrics, such as precision and recall, are also reported for the comparison. Appendix A provides the details of the experimental settings.

## 5.1 Baselines

We implemented several baseline models that were proven superior in similar multimodal hateful meme detection studies (Pramanick et al., 2021c; Hossain et al., 2022; Sharma et al., 2023).

### 5.1.1 Unimodal Models

For the visual-only models, we employed three well-known architectures: **Vision Transformer (ViT)** (Dosovitskiy et al., 2020), **Swin Transformer (SWT)** (Liu et al., 2021), and **ConvNeXT** (Liu et al., 2022). Three pre-trained textual-only transformer models, namely **Bangla-BERT** (Sarker, 2020), **multilingual BERT** (Devlin et al., 2019), and **XLMR** (Conneau et al., 2020) are used. We fine-tuned the unimodal models on the developed dataset.

### 5.1.2 Multimodal Models

- **MMBT**: Multimodal BiTransformer (MMBT) (Kiela et al., 2019) uses a transformer (Vaswani et al., 2017) architecture for fusing the visual and textual information.

- **CLIP**: It is a multimodal model trained on noisy image-text pair using contrastive learning (Chen et al., 2020) approach. CLIP has been widely used for several multimodal classification tasks (Pramanick et al., 2021b; Kumar and Nanadakumar, 2022). We extract the visual and textual embedding representations by fine-tuning the CLIP on the developed dataset. Afterward, we combined both representations and trained them on top of a softmax layer.

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

## 5.2 Results

Table 3 shows the performance of all the models for hateful meme detection and its target identification.

**Hateful Meme Detection:** Among the visual approach, the ConVNeXT model obtained the highest score (F1: 0.665). In the case of the textual-only models, Bangla BERT outperformed others (mBERT, XLMR) with an F1 score of 0.644. Notably, this score falls approximately 2.1% short of the best visual model performance, indicating that visual information is more distinguishable in identifying hateful memes. In contrast, ALBEF surpassed both unimodal counterparts in the state-of-the-art multimodal model, attaining the highest F1 score of 0.670. MMBT and CLIP failed to deliver satisfactory results. However, despite ALBEF's notable performance, our proposed method (DORA) outperforms the best model with an absolute improvement of 4.8% in F1 score.

**Target Identification:** Table 3 illustrates that task 2's textual model (mBERT) achieved the best F1 score of 0.652 among unimodal models. Nonetheless, consistent with our earlier findings, the joint evaluation of multimodal information led to significant enhancements in target identification. In the case of the multimodal models, the ALBEF surpassed the unimodal counterparts, achieving the highest score of 0.668. However, DORA outperforms the best model by 5.2% in terms of F1 score.

To further illustrate the superiority of the DORA, we compared its class-wise performance with the best baseline model (ALBEF) presented in Table 4. ALBEF performs poorly across all the target categories, especially in the 'TC' (0.04) and 'TS' (0.0) classes. In contrast, DORA demonstrated significant improvement in F1 score across all the classes. Overall, DORA yielded an impressive 13% improvement in the macro average F1 score, elevating it from 0.32 to 0.45. This improvement signifies

| Approach | Models | Hateful Meme Detection (Task 1) | | | Target Identification (Task 2) | | |
|---|---|---|---|---|---|---|---|
| | | **P** | **R** | **F1** | **P** | **R** | **F1** |
| **Visual Only** | ViT | 0.677 | 0.682 | 0.645 | 0.704 | 0.631 | 0.645 |
| | SWT | 0.669 | 0.680 | 0.660 | 0.682 | 0.620 | 0.645 |
| | ConvNeXT | <u>0.692</u> | <u>0.699</u> | 0.665 | 0.604 | 0.650 | 0.622 |
| **Text Only** | Bangla BERT | 0.644 | 0.658 | 0.644 | 0.634 | 0.597 | 0.612 |
| | mBERT | 0.628 | 0.648 | 0.610 | <u>0.706</u> | 0.616 | 0.652 |
| | XLMR | 0.640 | 0.655 | 0.638 | 0.555 | 0.672 | 0.601 |
| **Multimodal** | MMBT | 0.629 | 0.646 | 0.587 | 0.704 | 0.657 | 0.662 |
| | CLIP | 0.596 | 0.607 | 0.600 | 0.550 | 0.714 | 0.617 |
| | ALBEF | 0.671 | 0.682 | <u>0.670</u> | 0.649 | <u>0.740</u> | <u>0.668</u> |
| **Proposed System and Variants** | DORA w/o VF | 0.692 | 0.697 | 0.693 | 0.592 | 0.736 | 0.644 |
| | DORA w/o TF | 0.694 | 0.696 | 0.694 | 0.647 | 0.751 | 0.664 |
| | DORA w/o VF+TF | 0.694 | 0.689 | 0.691 | 0.536 | 0.725 | 0.616 |
| | DORA w/o VGAR | 0.688 | 0.696 | 0.675 | 0.639 | 0.740 | 0.679 |
| | DORA w/o TGAR | 0.672 | 0.682 | 0.654 | 0.659 | 0.729 | 0.677 |
| | DORA w/o VGAR + TGAR | 0.693 | 0.696 | 0.665 | 0.662 | 0.744 | 0.686 |
| | DORA | **0.718** | **0.718** | **0.718** | **0.706** | **0.759** | **0.720** |
| $\Delta_{DORA-baseline\_model}$ | | 2.6 | 1.9 | 4.8 | 0 | 1.9 | 5.2 |

Table 3: Performance comparison of visual only, textual only, and multimodal models on the test set. P, R, and F1 denote precision, recall, and weighted F1-score, respectively. The VF, TF, VGAR, and TGAR denote the visual features, textual features, vision-guided attentive representation, and text-guided attentive representation. The best-performing score in each column is highlighted in **bold**, and the second-best score is <u>underlined</u>. The last row shows the performance improvement of the proposed system (DORA) over the best baseline score.

| Model | Categories | P | R | F1 | Ma.F1 | W.F1 |
|---|---|---|---|---|---|---|
| **ALBEF** | TI | 0.77 | 0.98 | 0.86 | | |
| | TC | 0.11 | 0.03 | 0.04 | 0.32 | 0.67 |
| | TO | 0.78 | 0.26 | 0.39 | | |
| | TS | 0.00 | 0.00 | 0.00 | | |
| **DORA** | TI | 0.82 | 0.94 | 0.87 | | |
| | TC | 0.29 | 0.11 | 0.16 | **0.45** | **0.72** |
| | TO | 0.55 | 0.59 | 0.57 | | |
| | TS | 0.50 | 0.11 | 0.18 | | |

Table 4: Class-wise performance comparison of the best model with the DORA. Here, Ma.F1 and W.F1 indicate macro and weighted F1-score respectively.

that DORA maintains a good balance in the performance of all the evaluation measures (precision, recall, and F1 score).

### 5.3 Ablation Study

We perform an ablation study to analyze the contribution of each component (visual features (VF), textual features (TF), vision-guided attentive representation (VGAR), and text-guided attentive representation (TGAR)) of DORA. The last seven rows of Table 3 show the ablation outcome.

For Task 1, it is noteworthy that even in the absence of VF and TF, the model's performance remains superior (0.691-0.694) compared to the best baseline model, ALBEF (0.670). However, a substantial drop in the F1 score (ranging between 0.654 to 0.675) occurs when the attentive representations are removed, underscoring the significant impact of the dual-co attention mechanism in our proposed approach. Conversely, in the case of target identification, the model exhibits diminished performance when the VF and TF are excluded.

Interestingly, removing VGAR and TGAR shows less effect on the performance as the F1 score rises to its highest at 0.686. This implies that, for target identification, VF and TF bear greater significance than attentive representations. However, integrating all components in DORA results in a notable overall performance boost for both tasks.

### 5.4 Transferability and Generalizability of DORA

Table 5 shows the transferability and generalizability of the DORA on two datasets (*MUTE* (Hossain et al., 2022) and *EmoffMeme* (Kumari et al., 2023)) of different languages (i.e., Bengali and Hindi). Both *MUTE* and *EmoffMeme* have code-mixed captions in the memes, similar to the BHM dataset. To compare the performance in this experiment, we consider the best baseline model ( ALBEF). Results exhibit that when training and testing are done on the same dataset, DORA exceeds ALBEF by $\approx$ 4-5% in terms of F1 score across all the datasets. This outcome illustrates that DORA can also generalize well across languages. Similarly, when trained on one dataset and tested on a different one, DORA yields a better score than ALBEF.

Interestingly, the model trained in the Hindi dataset (*EmoffMeme*) exhibits poor performance when tested on Bengali datasets. Conversely, the model trained on the Bengali datasets and tested on the Hindi dataset exhibits suboptimal performance (0.64-0.65). This outcome emphasizes the need for a more sophisticated method that can be transfer-

Figure 3: Example (a) and (b) shows the memes where DORA yields better predictions, and example (c) illustrates a wrongly classified sample. The symbol (✓) and (✗) indicates the correct and incorrect prediction, respectively.

|  |  | BHM | MUTE | EmoffMeme |
|---|---|---|---|---|
|  |  | F1 | F1 | F1 |
| BHM (BEN) | ALBEF | 0.670 | 0.683 | 0.583 |
|  | DORA | **0.718** | 0.744 | 0.655 |
| MUTE (BEN) | ALBEF | 0.673 | 0.724 | 0.627 |
|  | DORA | 0.701 | **0.762** | 0.647 |
| EmoffMeme (HIN) | ALBEF | 0.513 | 0.499 | 0.785 |
|  | DORA | 0.529 | 0.493 | **0.824** |

Table 5: Transferability of best multimodal baseline and (DORA) on two additional benchmark datasets namely *MUTE* and *EmoffMeme*. Here, BEN and HIN indicate the Bengali and Hindi languages, respectively. The models are trained on the dataset specified in the rows and tested on the dataset specified in the columns. All the reported scores are weighted F1. The best transferable results are indicated in blue, and the scores in bold denote the best performance when models are trained and tested on the same dataset.

able across various languages. Overall, it can be stated that DORA is generalizable and also transferable across the datasets of the same languages.

### 5.5 Error Analysis

The results from table 3 demonstrated that our proposed method DORA is superior in identifying the hateful memes and their targets compared to the unimodal counterparts. To gain insights into the model's mistakes, we conduct a qualitative error analysis by examining some correctly and incorrectly classified samples, as illustrated in Figure 3. For better demonstration, we compare DORA's prediction with the best visual (ViT) and textual model (mBERT) predictions specifically for task 2.

In figure 3(a), the visual model incorrectly identified the meme as belonging to the 'Targeted Organization (TO)' class, likely due to the appearance of some famous political person faces. Simultaneously, the presence of a country name may lead the textual model to consider the meme as from the 'Targeted Society (TS)' class. However, when

both information is attended our proposed method correctly identified the meme as from the 'Targeted Individual (TI)' hate category. Similarly, in figure 3 (b), the visual model labeled the meme as the 'TI' category, and the textual model identified it as the '(TC)' class. The presence of multiple persons in visual information and slang words in textual information might have contributed to the misclassification. However, when visual and textual cues were jointly interpreted, the proposed method DORA correctly predicted the meme as 'TO'. Nevertheless, there were instances where DORA failed to provide the correct outcome. For instance, in figure 3 (c), the meme belongs to the 'TS' class and is misclassified by DORA as 'TI'. This misjudgment may be attributed to inconsistent visual features, specifically the presence of two boys' faces, which could misleadingly suggest classification as 'TI'. Incorporating world-level knowledge can help mitigate such model mistakes, which would be a promising avenue for future exploration.

## 6 Conclusion

This paper introduced a new large-scale multimodal dataset of 7,148 memes for detecting Bengali hateful memes and their targeted social entities. We also proposed DORA, a multimodal deep neural network for solving the tasks. Experiments on our dataset demonstrate the efficacy of DORA, which outperformed nine state-of-art baselines for two tasks. We further demonstrated the generalizability and transferability of DORA across other datasets of different languages. We plan to extend the dataset for more domains and languages in the future.

## Limitations

Though the proposed method (DORA) shows superior performance, there still exist some limitations of our work. First, we did not consider the

background contexts, such as visual entities (i.e., detected objects) and textual entities (i.e., person name, organization name), as external knowledge to the model, which could improve the overall performance. Second, it is likely that in some cases, the DORA may focus on less significant parts of the content while attending to the information. If the dataset contains misleading captions or irrelevant textual information, the attention mechanism might align with those parts of the image that are visually unrelated, producing misleading representations. Incorporating adversarial training could be an interesting future direction to mitigate the generation of such biased multimodal representations. Third, we observed that our method DORA struggled with memes that convey hate implicitly. It appeared to have difficulty correctly interpreting cultural references and context-specific content, leading to additional incorrect predictions. We plan to address this aspect in the future.

## Ethical Considerations

**User Privacy:** All the memes in the dataset were collected and annotated in a manner consistent with the terms and conditions of the respective data source. We do not collect or share any personal information (e.g., age, location, gender identity) that violates the user's privacy.

**Biases:** Any biases found in the dataset and model are unintentional. A diverse group of annotators labeled the data following a comprehensive annotation guideline, and all annotations were reviewed to address any potential annotation biases. We randomly collected data from various public social media pages and blogs to reduce data source biases. Moreover, we used neutral keywords (e.g., Bengali Memes, Bengali Mojar Memes, Funny Memes, Bengali Hashir Memes) not explicitly tied to specific hate themes to mitigate biases toward any specific person, community, or organization. Despite our best efforts, there may be inherent biases in the dataset, a common challenge in the dataset development process.

**Intended Use:** We intend to make our dataset accessible to encourage further research on hateful memes. We believe this dataset will help in understanding and building models of low-resource, especially Asian languages.

**Reproducibility:** We present the details of our experimental setting in Appendix A for the system's reproducibility. We will release the source code and the dataset at `https://github.com/eftekhar-hossain/Bengali-Hateful-Memes` upon accepting the paper.

## References

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.

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

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

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

Mithun Das and Animesh Mukherjee. 2023. Banglaabusememe: A dataset for bengali abusive meme classification. *arXiv preprint arXiv:2310.11748*.

Aaron Defazio and Samy Jelassi. 2022. Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. *The Journal of Machine Learning Research*, 23(1):6429–6462.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

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

Abhishek Jain, Aman Jain, Nihal Chauhan, Vikrant Singh, and Narina Thakur. 2017. Information retrieval using cosine and jaccard similarity measures in vector space model. *Int. J. Comput. Appl*, 164(6):28–30.

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

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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

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.

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

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

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*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

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

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

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

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34.

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

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

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021c. 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.

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.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*.

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

Sefik Ilkin Serengil and Alper Ozpinar. 2021. Hyper-extended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE.

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

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

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? *arXiv preprint arXiv:2301.11219*.

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

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

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

# Appendix

## A  Experimental Settings

For the experiment, we used the Google Colab platform. We downloaded the transformer models from the Huggingface[3] library and implemented it using the PyTorch Framework. The BNLP[4] and scikit-learn[5] library has been used for the preprocessing and evaluation measures. We empirically selected the models' hyperparameter values by examining the validation set's performance. All the models are compiled using *cross_entropy* loss function.

For optimizing the errors, in the case of the visual-only models, we used *MADGRAD* (Defazio and Jelassi, 2022) optimizer with a *weight_decay* of 0.01. For task 1, we chose the learning_rate $2e^{-5}$ while for task 2 it is settled to $7e^{-7}$. Conversely, for both tasks, among the textual-only models, mBERT and XLMR were trained using *Adam* (Kingma and Ba, 2014) optimizer with learning rate $1e^{-5}$ while MADGRAD (learning_rate = $2e^{-5}$ ) was utilized for Bangla-BERT model. Meanwhile, in the case of the multimodal models, ALBEF and CLIP were optimized using Adam (learning_rate = $1e^{-5}$), and MMBT with MADGRAD (learning_rate of $2e^{-4}$) optimizer. These settings of multimodal models were kept identical for both tasks.

On the other hand, in the case of the proposed DORA and its variants, we use the two attention heads in the multi-head co-attention block. During training, the models were optimized using MADGRAD with a $2e^{-5}$ learning rate. We used the *batch size* of 4 and trained the models for 20 *epochs* with a learning rate scheduler. We examined the validation set performance to save the best model during training.

## B  Data Sources and Filtering

Figure B.2 depicts the number of memes collected from each source. Most memes were collected from Facebook (50%) and Instagram (30%), while a few were accumulated from Pinterest, blogs, and other sources.

## C  Additional Data Statistics

**Text Analysis:** Table C.1 presents lexical statistics for the training set meme captions. In Task 1, the NHT class exhibits the highest number of

---

[3]https://huggingface.co/
[4]https://github.com/sagorbrur/bnlp
[5]https://scikit-learn.org/stable/

| (a) Targets Individual | (b) Targets Community | (c) Targets Organization | (d) Targets Society |

**English Translation:** Ahsan Habib Piyar He covers songs of Siya but most importantly He stores collection of Mia

**English Translation:** When you find out that your son is a BTS fan Where is my son's machine?

**English Translation:** Famous bandits then now Famous bandits

**English Translation:** Oh my dear I defecate in the street

Figure B.1: Few examples hateful memes targets from **BHM** dataset. The factors based on which the targets were decided (a) demean a person, (b) attack the sexual orientation of a community (BTS Fanbase), (c) state some organizations as Robbers, and (d) denigrate the people of a particular region.
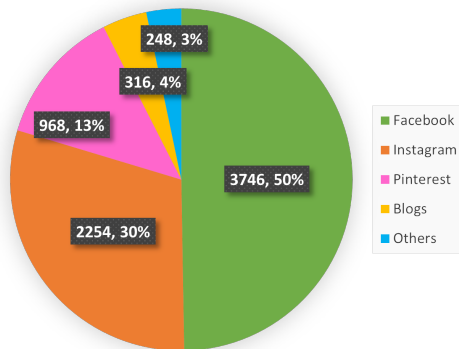


Figure B.2: Distribution of data sources. Each cell represents the number and percentage of samples collected from the corresponding sources.
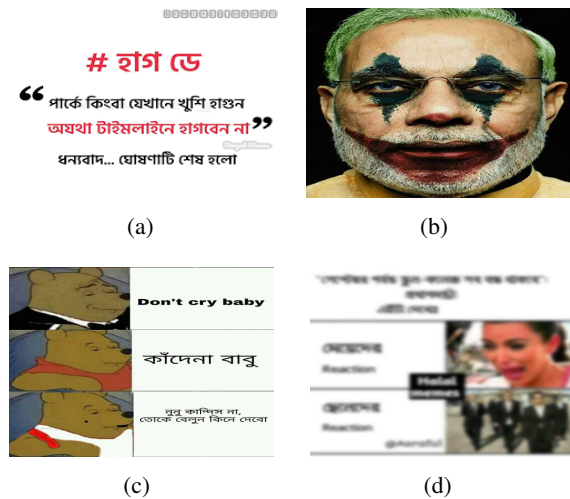


(a)

(b)

(c)

(d)

Figure B.3: Example filtered memes during the data collection process. The reason for the filtering is (a) contains only visual information (b) only textual information (c) contains cartoons (d) the contents are not cleared.

unique words(12,428) compared to the HT class (8,852). This discrepancy is unsurprising as the NHT class has the largest number of instances among all the classes. In the case of Task 2, the TI

class got the highest count of unique words (7,150), while the TS class features the lowest (727). The average caption length remains consistent at 13 words for most classes, with exceptions in TC (14) and TS (12). Figure C.1 displays a histogram illustrating caption length across different classes. The distribution reveals that most captions fall within the 5 to 30-word range. Only the TI and NHT classes contain captions with lengths exceeding 30 words. We further analyzed the captions to quantify word overlap across different classes. Specifically, we computed the Jaccard similarity (JS) (Jain et al., 2017) score between the top 400 most common words of different classes. The Jaccard score between each pair of classes is presented in Table C.2. We observed a substantial JS score of 0.51 between the HT and NHT classes, indicating a significant overlap in the words of these two classes. Regarding the target categories, the TI and TC pair exhibited the highest JS score (0.34), while the scores for the other categories remained below 0.20.

| | Class | #Words | #Unique words | Avg. #words/cap. |
|---|---|---|---|---|
| Task 1 | HT | 28477 | 8852 | 13.45 |
| | NHT | 50344 | 12428 | 13.82 |
| Task 2 | TI | 21583 | 7150 | 13.29 |
| | TO | 2159 | 1362 | 13.49 |
| | TC | 3694 | 2017 | 14.83 |
| | TS | 1041 | 727 | 12.24 |

Table C.1: Lexical analysis of captions in training set in terms of total words, total unique words, and average caption length across categories.

**Image Analysis:** The presence/absence of facial images is an important component in any meme. Therefore, we analyze the faces present in a meme. We employed the *deepFace*(Serengil and Ozpinar, 2021)library to perform this analysis. It allowed us to determine whether a given meme contains any faces. If a face is detected, we also extract in-

|     | TO   | TC   | TS   |
| --- | ---- | ---- | ---- |
| TI  | 0.19 | 0.34 | 0.16 |
| TO  |      | 0.18 | 0.11 |
| TC  |      |      | 0.14 |
|     | **NHT** |   |      |
| HT  | 0.51 |      |      |

Table C.2: Jaccard similarity analysis among the top 500 common words across different class combinations.
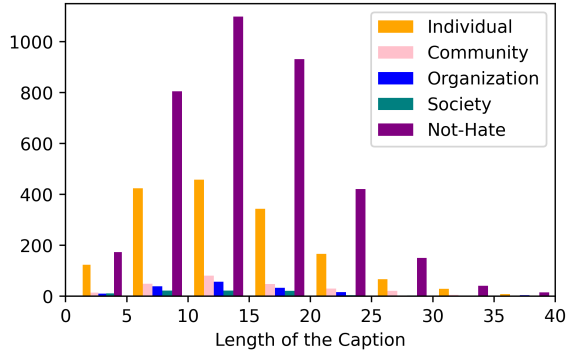


Figure C.1: Caption length distribution of the training set across different classes.

formation regarding the associated gender and age. Our findings reveal that approximately 34.12% of the memes contain no faces. Among this group, 33.28% belong to the 'Hate' class. Of the total memes, 53.29% feature male faces, with 35.3% from the 'Hate' class and 64.64% from the 'Not-Hate' class. Meanwhile, 12.53% feature female faces, with 50.83% within the 'Hate' class and 49.17% within the 'Not-Hate' class. On average, the detected ages of males and females in the memes hover around 31 years. A key observation is the prevalence of male faces in the 'Hate' class, indicating that males are common targets of hateful content in the Bengali community.