MUTE: A Multimodal Dataset for Detecting Hateful Memes

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

Abstract

The exponential surge of social media has enabled information propagation at an unprecedented rate. However, it also led to the generation of a vast amount of malign content, such as hateful memes. To eradicate the detrimental impact of this content, over the last few years hateful memes detection problem has grabbed the attention of researchers. However, most past studies were conducted primarily for English memes, while memes on resource-constraint languages (i.e., Bengali) remain under-studied. Moreover, current research considers memes with a caption written in monolingual (either English or Bengali) form. However, memes might have code-mixed captions (English+Bangla), and the existing models can not provide accurate inference in such cases. Therefore, to facilitate research in this arena, this paper introduces a multimodal hate speech dataset (named ***MUTE***) consisting of 4158 memes having Bengali and code-mixed captions. A detailed annotation guideline is provided to aid the dataset creation in other resource-constraint languages. Additionally, extensive experiments have been carried out on MUTE, considering the only visual, only textual, and both modalities. The result demonstrates that joint evaluation of visual and textual features significantly improves ($\approx 3\%$) the hateful memes classification compared to the unimodal evaluation.

1 Introduction

With the advent of the Internet, social media platforms (i.e., Facebook, Twitter, Instagram) significantly impact people's day-to-day life. As a result, many users communicate by posting various content in these mediums. This content includes promulgating hate speech, misinformation, aggressive and offensive views. While some contents are beneficial and enrich our knowledge, they can also trigger human emotions that can be considered harmful. Among them, the propagation of hateful content can directly or indirectly attack social harmony based on race, gender, religion, nationality, political support, immigration status, and personal beliefs. In recent years, memes have become a popular form of circulating hate speech (Kiela et al., 2020). These memes on social media have a pernicious impact on societal polarization as they can instigate hateful crimes. Therefore, to restrain the interaction through hateful memes, an automated system is required to quickly flag this content and lessen the inflicted harm to the readers. Several works (Davidson et al., 2017; Waseem and Hovy, 2016) have accomplished hateful memes detection, most of which were for the English language. Unfortunately, no significant studies have been conducted on memes regarding low-resource languages, especially Bengali. In recent years an increasing trend has been observed among the people to use Bengali memes. As a result, it becomes monumental to identify the Bengali hateful memes to mitigate the spread of negativity. However, memes analysis is complicated as it requires a holistic understanding of visual and textual content to infer (Zhou et al., 2021). The visual content of the meme alone may not be harmful (Figure 1 (a)). However,

(a) Attack religious beliefs (b) Insult a person

Figure 1: Examples of hateful memes having (a) only Bengali caption (b) Code-mixed (Bengali + English) caption.

WARNING: This paper contains meme examples and words that are offensive in nature.

Proceedings of the AACL-IJCNLP 2022 Student Research Workshop, pages 32–39
November 20, 2022. ©2022 Association for Computational Linguistics

it becomes hateful with the incorporation of textual content as it directly attacks religious beliefs. A meme's caption can be written in a mixed language (written in both English and Bengali as in Figure 1 (b)), which can evade the surveillance engine in those cases. Developing a hateful meme detection system for such a scenario is complicated as no standard dataset is available. Moreover, developing an intelligent multimodal memes analysis system for Bengali is challenging due to the unavailability of benchmark corpus, lack of reliable NLP tools (such as OCR), and the complex morphological structure of the Bengali language. Therefore, this work aims to develop a multimodal dataset for Bangla hate speech detection and investigate various models for the task. The critical contributions of the work are summarized as follows:

- Created a multimodal hate speech dataset (MUTE) in Bengali consisting of 4158 memes annotated with Hate and Not-Hate labels.
- Performed extensive experiments with state-of-the-art visual and textual models and then integrate the features of both modalities using the early fusion approach.

2 Related Work

This section discusses the past studies on hate speech detection based on unimodal (i.e., image or text) and multimodal data.

Unimodal based hate speech detection: Hate speech detection is a prominent research issue among the researchers of different languages (Ross et al., 2016; Lekea and Karampelas, 2018). Most hate speech detection works were accomplished based on the text data. For example, both Davidson et al. (2017) and Waseem and Hovy (2016) developed hate speech datasets considering the Twitter posts. Similarly, De Gibert et al. (2018) constructs a dataset that considers the hate speech posted in a white supremacy forum. Some works were also accomplished concerning the low resource languages. For instance, Fortuna et al. (2019); Ousidhoum et al. (2019) introduced hate speech datasets for Portuguese and Arabic. A few works have also been done on Bengali hate speech detection (Romim et al., 2021; Mathew et al., 2021; Ishmam and Sharmin, 2019). Several architectures have been employed over the last few years to classify hateful texts. Earlier researchers widely used Recurrent Neural Network (Gröndahl et al., 2018), Long Short Term Memory (LSTM) Network (Badjatiya et al., 2017), and the combination of RNN and convolutional neural network (CNN) (Zhang et al., 2018b) based methods. Recently, Bidirectional Encoder Representations for Transformers or BERT-based models (Pamungkas and Patti, 2019; Fortuna et al., 2021) are applied and achieved superior performance compared to the deep learning-based methods.

Multimodal hate speech detection: In contrast to the text-based analysis, in recent years, few pieces of work considered multimodal information (i.e., image + text) for hate speech detection. For example, Kiela et al. (2020) introduced a multimodal memes dataset for detecting hate speech. Gomez et al. (2020) developed a large scale multimodal dataset (MMHS150k) for detecting hateful memes. In another work, Rana and Jha (2022) introduced a multimodal hate speech dataset concerning three modalities (i.e., image, text, and audio). However, few works have been accomplished on multimodal hate speech detection for resource constraint languages. Perifanos and Goutsos (2021) introduced a multimodal dataset for detecting hate speech in Greek social media. Likewise, Karim et al. (2022) developed a dataset for multimodal hate speech detection from Bengali memes. Several approaches were employed for detecting hate speech using multimodal learning. Some researchers exploited the different fusion (Sai et al., 2022; Perifanos and Goutsos, 2021) techniques (i.e., early and late fusion) to evaluate the image and textual features jointly. Others have employed bi-linear pooling (Chandra et al., 2021; Choi and Lee, 2019) and transformer-based methods (Kiela et al., 2020) such as MMBT, ViLBERT, and Visual-BERT. Despite having the state of the art multimodal transformer architectures, these models have only applied for high resource language (i.e., English).

Differences with existing researches: Though a considerable amount of work has been accomplished on multimodal hate speech detection, only a few works studied low-resource languages (i.e., Bengali). In our exploration, we found a work (Karim et al., 2022) that detects hate speech from multimodal memes for the Bengali language. However, they did not curate the social media memes for analysis; instead artificially created a memes dataset for Bengali by conjoining the hateful texts into various images. Moreover, the current works overlooked the memes containing captions written cross-lingually. Considering these drawbacks, the proposed research differs from the existing studies in three ways: *(i)* develops a multimodal hate speech dataset (i.e., MUTE) for Bengali considering the Internet memes, *(ii)* provides a detailed annotation guideline that can be followed for resource creation in other low resource languages, and *(iii)* consider the memes that contain code-mixed (English + Bangla) and code-switched (written Bengali dialects in English alphabets) caption.

3 MUTE: A New Benchmark Dataset

This work developed MUTE: a novel multimodal dataset for Bengali Hateful memes detection. The MUTE considered the memes with code-mixed and cod-switched captions. For developing the dataset, we follow the guidelines provided by Kiela et al. (2020). This section briefly describes the dataset development process with detailed statistics.

3.1 Data Accumulation

For dataset construction, we have manually collected memes from various social media platforms such as Facebook, Twitter, and Instagram. We search the memes using a set of keywords such as *Bengali Memes, Bangla Troll Memes, Bangla Celebrity Troll Memes, Bangla Funny Memes* etc. Besides, some popular public memes pages are also considered for the data collection, such as *Keu Amare Mairala, Ovodro Memes* etc. We accumulated 4210 memes from January 10, 2022, to April 15, 2022. During the data collection, some inappropriate memes are discarded by following the guidelines provided by Pramanick et al. (2021). The criteria for discarding data are: (i) memes contain only unimodal data, (ii) memes whose textual or visual information is unclear and (iii) memes contain cartoons. In this filtering process, 52 memes were removed and ended up with a dataset of **4158** memes. Afterwards, the caption of the memes is manually extracted as Bengali has no standard OCR. Finally, the memes and their corresponding captions are given to the annotators for annotation.

3.2 Dataset Annotation

The collected memes are manually labelled into two distinct categories: Hate and not-Hate. However, to ensure the dataset's quality, it is essential to follow a standard definition for segregating the two categories. After exploring some existing works on multimodal hate speech detection (Kiela et al., 2020; Gomez et al., 2020; Perifanos and Goutsos, 2021), we define the classes:

Hate: A meme is considered as Hateful if it intends to vilify, denigrate, bullying, insult, and mocking an entity based on the characteristics including gender, race, religion, caste, and organizational status etc.

Not-Hate: A meme is reckoned as not-Hateful if it does not express any inappropriate cogitation and conveys positive emotions (i.e., affection, gratitude, support, and motivation) explicitly or implicitly.

3.2.1 Process of Annotation

We instructed the annotators to follow the class definitions for performing the annotation. It also asked them to mention the reasons for assigning a meme to a particular class. This explanation will aid the expert in selecting the correct label during contradiction. Initially, we trained the annotators with some sample memes. Four annotators (computer science graduate students) performed the manual annotation process, and an expert (a Professor conducting NLP research for more than 20 years) verified the labels. Annotators were equally divided into two groups where each annotated a subset of memes. In case of disagreement, the expert decided on the final label. The expert ruled a total of 113 non-hateful and 217 hateful memes as hostile and non-hateful. An inter-annotator agreement was measured using Cohen (Cohen, 1960) Kappa Coefficient to ensure the data annotation quality. We achieved a mean Kappa score of 0.714, which indicates a moderate agreement between the annotators. Earlier, it is mentioned that this work is the very first attempt at multimodal hate speech detection that considers the social media memes of the Bengali language. Therefore, it requires more extensive scrutiny with more diverse data and a high level of annotator agreement to deploy the model trained on this dataset. The agreement score illustrates the difficulty in identifying the potential hateful memes by humans and brings a question of biases, thus limiting the broader impact of this work.

3.3 Dataset Statistics

For training and evaluation, the MUTE is split into the train (80%), test (10%), and validation (10%) set. Table 1 presents the class-wise distribution of the dataset. It is observed that the dataset is slightly imbalanced as the 'Not-Hate' class contains ≈60% data. Table 2 shows the statistics of the training

Class	Train	Test	Valid	Total
Hate	1275	159	152	1586
Not-Hate	2092	257	223	2572

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

	Hate	Not-Hate
#Code-mixed texts	345	138
#Words	12854	22885
#Unique words	5781	8627
Max. caption length	51	87
Avg. #words/caption	10.08	10.94

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

set, which contains a total of 483 memes with code-mixed captions. Moreover, it is also illustrate that the 'Not-Hate' class has a higher number of words and unique words than the 'Hate' class. However, the average caption length is almost identical in both classes. Apart from this, we carried out a quantitative analysis using the Jaccard similarity index to figure out the fraction of overlapping words among the classes. We obtained a score of 0.391, indicating that some common words exist between the classes.

4 Methodology

Several computational models have been explored to identify hateful memes by considering the single modality (i.e., image, text) and the combination of both modalities (image and text). This section briefly discusses the methods and parameters utilized to construct the models.

4.1 Baselines for Visual Modality

This work employed convolutional neural networks (CNN) to classify hateful memes based on visual information. Initially, the images are resized into $150 \times 150 \times 3$ and then driven into the pre-trained CNN models. Specifically, we curated the VGG19, VGG16 (Simonyan and Zisserman, 2015), and ResNet50 (He et al., 2016) architectures that fine-tuned on MUTE dataset by using the transfer learning (Tan et al., 2018) approach. Before that, the top two layers of the models are replaced with a sigmoid layer for classification.

4.2 Baselines for Textual Modality

For text based hateful memes analysis, various deep learning models are employed including BiLSTM + CNN (Sharif et al., 2020), BiLSTM + Attention (Zhang et al., 2018a), and Transformers (Vaswani et al., 2017).

BiLSTM + CNN: At first, the word embedding (Mikolov et al., 2013) vectors are fed to a BiLSTM layer consisting of 64 hidden units. Following this, a convolution layer with 32 filters with kernel size two is added, followed by a max-pooling layer to extract the significant contextual features. Finally, a sigmoid layer is used for the classification. The final time steps output of the BiLSTM network provides the contextual information of the overall text.

BiLSTM + Attention: We applied the additive attention (Bahdanau et al., 2015) mechanism to the individual word representations of the BiLSTM cell. The CNN is replaced with an attention layer. The attention layer tries to give higher weight to the significant words for inferring a particular class.

Transformers: Pretrained transformer models have recently obtained remarkable performance in almost every NLP task (Naseem et al., 2020; Yang et al., 2020; Cao et al., 2020). As the *MUTE* contains cross-lingual text, this work employed three transformer models, namely Multilingual Bidirectional Encoder Representations for Transformer (M-BERT (Devlin et al., 2019)), Bangla-BERT (Sarker, 2020), and Cross-Lingual Representation Learner (XLM-R (Conneau et al., 2020)). All the models are downloaded from HuggingFace[1] transformer library. We follow their preprocessing [2] and encoding technique for preparing the texts. The transformer models provide a sentence representation vector of size 768. This vector is passed to a dense layer of 32 neurons, and then using the pre-trained weights, models are retrained on the developed dataset with a sigmoid layer.

4.3 Baselines for Multimodal Data

In recent years, joint evaluation of visual and textual data has proven superior in solving many complex NLP problems (Hori et al., 2017; Yang et al., 2019; Alam et al., 2021). This work investigates the joint learning of multimodal data for hateful memes

[1] https://huggingface.co/
[2] https://huggingface.co/docs/tokenizers/index

classification. For multimodal feature representation, we employed the feature fusion (Nojavanasghari et al., 2016) approach. In earlier experiments, all the visual and two textual (i.e., Bangla-BERT and XLM-R) models are used to construct the multimodal models. For the model construction, we added a dense layer of 100 neurons at both modality sides and then concatenated their outputs to make combined visual and textual data representations. Finally, this combined feature is passed to a dense layer of 32 neurons, followed by a sigmoid layer for the classification task.

5 MUTE: Benchmark Evaluation

The training set is used to train the models, whereas the validation set is for tweaking the hyperparameters. We have empirically tried several hyperparameters to obtain a better model's performance and reported the best one. The final evaluation of the models is done on the test set. This work selects the weighted f1-score (WF) as the primary metric for the evaluation due to the class imbalance nature of the dataset. Apart from this, we used the class weighting technique (Sun et al., 2009) to give equal priority to the minority class (hate) during the model training.

5.1 Results

Table 3 illustrates the outcome of the visual, textual, and multimodal models for hateful memes classification. In the case of the visual model, ResNet50 obtained the maximum WF of 0.641. For the text modality, the B-BERT model obtained the highest WF (0.649). The outcomes of the other textual models (i.e., BiLSTM + Attention, BiLSTM + CNN, and XLM-R) are not exhibited significant differences compared to the best model (B-BERT).

Approach	Models	P	R	WF
Visual	VGG19	0.594	0.579	0.584
	VGG16	0.636	0.644	0.638
	ResNet50	0.643	0.639	0.641
Textual	BiLSTM + CNN	0.617	0.663	0.608
	BiLSTM + Attention	0.647	0.653	0.642
	M-BERT	0.627	0.644	0.620
	B-BERT	0.645	0.658	0.649
	XLM-R	0.646	0.656	0.648
Multimodal	VGG19 + B-BERT	0.639	0.649	0.641
	VGG16 + B-BERT	0.676	0.670	**0.672**
	ResNet50 + B-BERT	0.606	0.620	0.609
	VGG16 + XLM-R	0.594	0.581	0.586
	VGG19 + XLM-R	0.515	0.605	0.489
	ResNet50 + XLM-R	0.651	0.600	0.604

Table 3: Performance comparison of the visual, textual, and multimodal models on the test set. Where P, R, WF denotes precision, recall and weighted f_1-score, respectively.

On the other hand, with the multimodal information, the outcomes of the models are not improved. Almost all the models' WF lies around 0.60 except the VGG19 + B-BERT model (0.641). However, the VGG16 + B-BERT model outperformed all the models by achieving the highest weighted WF of 0.672, which is approximately 2% higher than the best unimodal model of B-BERT (0.649).

5.2 Error Analysis

We conducted a quantitative error analysis to investigate the model's mistakes across the two classes. To illustrate the errors, the number of misclassified instances is reported in Figure 2 for the best unimodal (ResNet50 and B-BERT) and multimodal (VGG19 + B-BERT) models. It is observed that the misclassification rate (MR) is increased ≈10% and decreased ≈9% from visual to textual model, respectively, for the 'Hate' and 'Not-Hate' classes. However, the joint evaluation of multimodal features significantly reduced the MR to 38% (from 44% and 54%) in the Hate class and thus improved the model's overall performance. Though the multimodal model showed superior performance compared to the unimodal models, there is still room for improvement. We point out several reasons behind the model's mistakes. Among them, identical words in different written formats (code-mixed, code-switched) made it difficult for the model to identify accurate labels. Moreover, the discrepancy between some memes' visual and textual information creates confusion for the multimodal model. Indeed, these are some significant factors that should be tackled to develop a more sophisticated model for Bengali hateful memes classification.

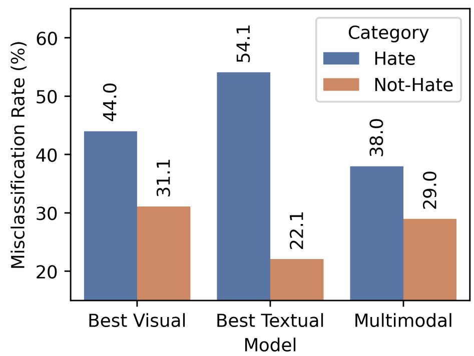

Figure 2: Miss-classification rate across two classes by different models.

6 Conclusion

This paper presented a multimodal framework for hateful memes classification and investigated its performance on a newly developed multimodal dataset (*MUTE*) having Bengali and code-mixed (Bangla + English) captions. For benchmarking the framework, this work exploited several computational models for detecting hateful content. The key finding of the experiment is that the joint evaluation of multimodal features is more effective than the memes' only visual or textual information. Moreover, the cross-lingual embeddings (XLM-R) did not provide the expected performance compared to the monolingual embeddings (Bangla-BERT) when jointly evaluated with the visual features. The error analysis reveals that the model's performance gets biased to a particular class due to the class imbalance. In future, we aim to alleviate this problem by extending the dataset to a large scale and framing it as a multi-class classification problem. Secondly, for robust inference, advanced fusion techniques (i.e., co-attention) and multitask learning approaches will be explored. Finally, future research will explore the impact of dataset sampling and do some ablation study (i.e., experimenting with only English, only Bangla, code-mixed, and code-switched text) to convey valuable insights about the models' performance.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A survey on multimodal disinformation detection. *arXiv preprint arXiv:2103.12541*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Qingqing Cao, Harsh Trivedi, Aruna Balasubramanian, and Niranjan Balasubramanian. 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.

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.

Jun-Ho Choi and Jong-Seok Lee. 2019. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270.

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 *ACL*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

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.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

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.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12.

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

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 555–560. IEEE.

Md Karim, Sumon Kanti Dey, Tanhim Islam, Bharathi Raja Chakravarthi, et al. 2022. Multimodal hate speech detection from bengali memes and texts. *arXiv preprint arXiv:2204.10196*.

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.

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

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. 2020. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113:58–69.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

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. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218*.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, Saiful Islam, et al. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468. Springer.

Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *3rd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media*, pages 6–9. Ruhr-Universitat Bochum.

Siva Sai, Naman Deep Srivastava, and Yashvardhan Sharma. 2022. Explorative application of fusion techniques for multimodal hate speech detection. *SN Computer Science*, 3(2):1–13.

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

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 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. NLP Association of India (NLPAI).

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

Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. 2009. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

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

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 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 Zhang, Jin Wang, and Xuejie Zhang. 2018a. Ynu-hpcc at semeval-2018 task 1: Bilstm with attention based sentiment analysis for affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 273–278.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018b. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

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