ExMute: A Context-Enriched Multimodal Dataset for Hateful Memes

Riddhiman Swanan Debnath[1], Nahian Beente Firuj[1], Abdul Wadud Shakib[1], Sadia Sultana[1], Md Saiful Islam[1, 2]
[1]Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh
[2]Computing Science, University of Alberta, Edmonton, Alberta, Canada

Abstract

In this paper, we introduce ExMute, an extended dataset for classifying hateful memes that incorporates critical contextual information, addressing a significant gap in existing resources. Building on a previous dataset of 4,158 memes without contextual annotations, ExMute expands the collection by adding 2,041 new memes and providing comprehensive annotations for all 6,199 memes. Each meme is systematically labeled across six defined contexts—religion, politics, celebrity, male, female, and others—with language markers indicating code-mixing, code-switching, and Bengali captions, enhancing its value for linguistic and cultural research while facilitating a nuanced understanding of meme content and intent. To evaluate ExMute, we apply state-of-the-art textual, visual, and multimodal approaches, leveraging models including BanglaBERT, Visual Geometry Group (VGG), Inception, ResNet, and Vision Transformer (ViT). Our experiments show that our custom LSTM attention-based textual model achieves an accuracy of 0.60, while VGG-based visual models reach up to 0.63. Multimodal models, which combine visual and textual features, consistently achieve accuracy scores of around 0.64, demonstrating the dataset's robustness for advancing multimodal classification tasks. ExMute establishes a valuable benchmark for future NLP research, particularly in low-resource language settings, highlighting the importance of context-aware labeling in improving classification accuracy and reducing bias.

1 Introduction

The exponential growth of social media platforms such as Facebook, TikTok, Reddit, and Instagram has paralleled the expansion of the internet, transforming them into powerful tools for expressing opinions on politics, business, entertainment, and current events (Oldenbourg, 2024). However, this increased connectivity has also boosted the spread of offensive content targeting individuals or groups based on race, religion, and sexual orientation. The rise of this toxic content poses significant challenges, particularly in the form of hateful memes—visual and textual media repurposed to convey cultural, social, or political views with a mask of humor (Mukhtar et al., 2024). While memes often serve as light-hearted content, they can also propagate harmful and prejudiced messages, exacerbating issues such as cyberbullying, harassment, and societal discord (Sambasivan et al., 2019; Romim et al., 2021b).

Figure 1: **Category -** Hateful, **Context:** Religion

In recent years, the popularity of multimodal memes has surged as an effective means of communication in this era of digital interconnectivity (Abdullakutty and Naseem, 2024). However, identifying and mitigating the spread of such harmful content remains a significant challenge due to the sheer scale of online platforms and the complexity of multimodal content. Significant progress has been achieved in detecting hateful memes in English, with several studies and resources available (Waseem and Hovy, 2016; Davidson et al., 2017). In Bangla, however, existing work focuses primarily on text-based hate speech detection (Al Maruf et al., 2024; Romim et al., 2022; Das et al., 2021; Romim et al., 2021a), leaving hateful meme detection largely unexplored. This gap underscores the

Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP2025), pages 83–89
January 20, 2025.

need for comprehensive multimodal approaches in Bangla. In addition, these advancements have yet to be equally replicated in low-resource languages, particularly Bangla, code-switch (Bangla dialects in English script), and code-mix (Bangla and English) languages. This is noteworthy given that Bangla is the fifth most spoken language worldwide, with over 230 million speakers, including approximately 100 million in Bangladesh and 85 million in India. (Karim et al., 2022).

Despite the rising use of memes in Bengali due to the increasing number of social media users in Bangladesh, there has been limited research focused on the identification and contextual analysis of hate speech in this language (Hossain et al., 2022a,b). Furthermore, existing studies often lack detailed categorization based on different contexts or target audiences (Figure 1). We introduce ExMute, an extended dataset for classifying Bangla hateful memes across various social media platforms to address this gap. Our work also includes categorizing the data into six distinct contexts: religion, politics, celebrity, male, female, and others, providing an enriched framework for nuanced hateful meme analysis. The overall contribution of our paper:

- Curated a human-annotated multimodal hateful memes dataset enriched with six contexts: religious, celebrity, political, male, female, and others.
- Annotated 6,199 memes as hateful or non-hateful, with context labels, using a detailed guideline for Bangla, code-mixed, and code-switched captions.
- Established baselines by extensively testing various textual and visual models, including a custom LSTM with attention, Vision Transformer, and Bangla BERT.
- Released code and data publicly to support further research in this area.

2 ExMute: An Extended Dataset

We extended the **Mute** Hossain et al. (2022b) which consisted of 4,158 labeled memes, and added an additional amount of 2,041 memes along with code mix, code switch, and Banglish captions. For data collection, we followed the approaches shown in these two studies Hossain et al. (2022b) and Kiela et al. (2020).

class	train	test	valid	total
hateful	540	684	925	2149
non-hateful	1943	1182	924	4050

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

2.1 Data Collection

We collected memes and texts containing common slurs and terms from Facebook, Reddit, and Instagram. We searched for these using keywords like "Bangla Memes," "Bangla Troll Memes," etc., on platforms like Wittigenz and Halal Meme posting. During data collection, we exclude some irrelevant memes by considering the rules stated by Pramanick et al. (2021). The criteria for discarding data are **(i)** memes containing only unimodal data (only text or image) and **(ii)** memes whose textual or visual information is unclear. We collected 2,098 memes, and through this filtering process, 57 memes were removed from newly collected data. Afterward, we manually extracted captions from the memes, as Bengali lacks a standardized OCR system, and provided them to annotators for labeling with corresponding memes.

During data collection, memes were sourced from 15 different contexts, such as racial, misogynist, geopolitical, sports, and so on. Emphasis was placed on the frequency of instances across these contexts, with male, female, political, religious, celebrity, and other categories emerging as the most prominent.

2.2 Dataset Annotation

To establish clear annotation guidelines, we followed the approach of prior studies Kiela et al. (2020), Islam et al. (2022), and Perifanos and Goutsos (2021) and defined hateful and non-hateful in the following ways:

- **Hateful:** Targets an entity based on its gender, race, religion, caste, or organizational status and intends to vilify, denigrate, and mock.
- **Non-hateful:** Expresses positive feelings such as affection, gratitude, support, and motivation, whether openly or implicitly.

We also determined the contexts of the memes, hateful or not, by observing the captions and visual characteristics in the following way:

- **Male**: Clearly indicates a male context.

Name	context-wise	Percentage
Political	149	49.83
Religious	293	32.89
Female	677	61.15
Male	772	36.01
Celebrity	870	40.23
Others	3438	26.53

Table 2: Context-wise distribution of hate-non-hate memes and percentage of hateful memes

Characteristics	Hateful	Non-hateful
#Code-Mix Cap	588	1088
#Code-Switch Cap	58	119
#Bangla Cap	223	396
#Words	29245	50215
#Unique Words	9251	13223
Max Caption length	186	241
Avg #Words/Cap	13.6	12.3

Table 3: Distribution of data across various characteristics related to meme captions. Here, cap represents caption

- **Female**: Clearly indicates a female context.
- **Religious**: Refers to an individual or group based on religious beliefs.
- **Political**: Refers to an individual or group based on political beliefs.
- **Celebrity**: Refers to a celebrity.
- **Others**: Does not fit into any of the above categories.

Initially, we hired undergraduate students from different faculties, aged 24-26, with 50% female, and provided training using sample memes. We use five annotators for each instance, which are annotated independently. The final label was assigned based on consensus, with a linguistic expert verifying the labels. For instances with unresolved disagreements, we sought expert adjudication. Annotators were instructed to follow label definitions and guidelines closely and to document their reasoning for each annotation. This documentation helped the expert make informed decisions in cases of conflict. Compensation for annotators was provided according to the university research ethics board's standard local rate, and they were encouraged to pace their work, taking regular breaks to avoid prolonged sessions and negative mental health impacts of annotators Ybarra et al. (2006), Levin (2017).

3 Dataset Statistics

Our final dataset comprises a total of 6,199 instances. The dataset displays an imbalanced distribution, with the 'Non-Hate' class representing about 65% of the dataset, as shown in Table 1. Additionally, Table 2 provides a breakdown of instances by context. Notably, the "Others" category has a disproportionately high number of instances compared to other contexts, as annotators often placed memes here when they didn't clearly fit any other context.

From Table 2, it is evident that memes targeting females are overrepresented in gender-based contexts. Common Bangla words, such as "আমি, না, আমার, কি," appear frequently across all contexts. Words like "girl" are common in female-targeted memes, while terms like "ramadan" and "নামাজ" are primarily associated with religious memes. Figure 2 and Figure 3 further illustrate caption characteristics. Figure 2 shows the number of captions across different contexts based on caption length, providing insights into how caption length varies contextually. Figure 3 displays the distribution of caption lengths between hate and non-hate categories, highlighting any notable differences in caption length within each category. For training and evaluation, we divided the dataset into three parts: 80% for training, 10% for testing, and 10% for validation. The class distribution across these subsets is presented in Table 1.

4 Methodology

In this section, we outline the methods used to develop benchmark models for detecting hateful memes through unimodal and multimodal approaches, utilizing both visual and textual features.

4.1 Data Cleaning and Preprocessing

Initially, HTML tags and URLs were removed from the text captions, followed by the elimination of newline characters to normalize the text layout into a cohesive string. Punctuation marks and special characters were subsequently filtered out to simplify the textual data further.

For compatibility with deep learning architectures, particularly DNN and transformer-based models, the cleaned text was tokenized at the word level using the Keras tokenizer. This step involved

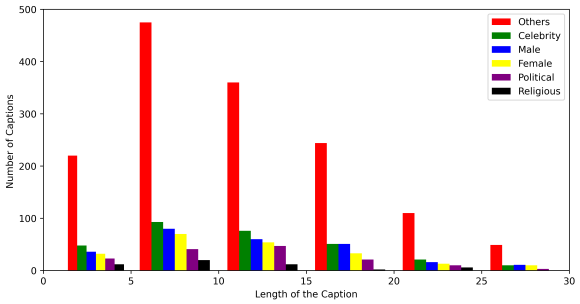

Figure 2: Number of captions according to the length of the captions in different contexts

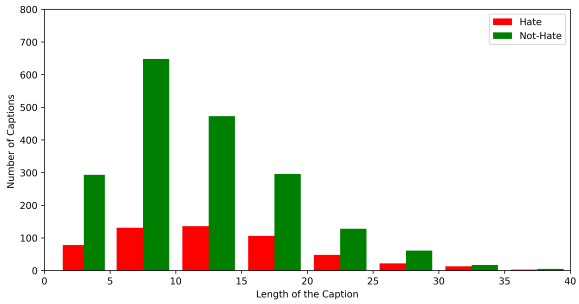

Figure 3: Number of captions according to the length of the captions of hate-nonhate

mapping each unique word to a corresponding integer index, effectively converting the text into a numerical vector representation. To ensure consistent input dimensions across samples, sequences were padded to a maximum length of 50, a necessary step for deep learning models requiring fixed-length input.

For the visual modality, the images were resized to a uniform dimension of 150×150×3, preserving their three-channel (RGB) format. Keras image preprocessing functions were employed to standardize the image data and enhance its compatibility with convolutional neural networks (CNNs). This resizing and adjustment ensured uniformity in input data and facilitated effective model training.

4.2 Textual Model

For text-based hateful memes analysis, various deep learning models are employed, including BiLSTM + CNN (Sharif et al., 2020), BiLSTM + Attention (Zhang et al., 2018), and Transformers (Vaswani, 2017). Additionally, we developed a custom LSTM model with an attention mechanism to enhance performance.

Initially, the word embedding vectors (Mikolov, 2013) are fed into a BiLSTM layer of 64 hidden units. Then a convolution layer with 32 filters with a kernel size of two is added, followed by a max-pooling layer to extract the significant contextual features. Then, a sigmoid layer is used for classification. Finally, the output of the BiLSTM network provides contextual information for the overall text.

Also, we used the additive attention mechanism introduced by Bahdanau (2014) to analyze the representations of individual words in the BiLSTM cell. The CNN is replaced with an attention layer. The attention layer prioritizes significant words to infer a specific class.

Our custom LSTM model integrates an attention mechanism to enhance performance and interpretability. The attention layer computes scores by combining features and hidden states, normalizing them using softmax. A context vector is then derived as a weighted sum of the features. The input sequence is embedded and processed through a bidirectional LSTM, capturing both forward and backward contextual information by concatenating hidden and cell states. The attention layer applies to the LSTM output, producing context vectors and attention weights. The final output layer uses a sigmoid activation function, suitable for binary classification tasks.

4.3 Visual Model

For the visual models, we used advanced architectures, including VGG19, VGG16,(Simonyan and Zisserman, 2014), ResNet50(He et al., 2016), and Vision Transformer(ViT).(Dosovitskiy, 2020). Specifically, VGG19, VGG16, and ResNet50 were fine-tuned on the MUTE dataset through transfer learning. For hate-non-hate classification, the upper layers of these models were frozen, utilizing weights pre-trained on ImageNet (Deng et al., 2009) for 1000 classes, and the top layers were replaced with a sigmoid layer to enable binary classification.

4.4 Multimodal Model

Recent studies, including Hori et al. (2017), Yang et al. (2019), and Alam et al. (2021), indicate that combining visual and textual data improves performance in complex NLP tasks. For multimodal feature representation, we applied a feature fusion approach Nojavanasghari et al. (2016), integrating both visual and textual models such as BanglaBERT(Sarker, 2020; Bhattacharjee et al., 2022). We added a dense layer with 100 neurons to each modality, then concatenated their outputs to cre-

ate a unified feature representation, followed by a dense layer with 32 neurons and a sigmoid layer for classification. We used Bangla-BERT (Sarker, 2020) for text encoding, generating input IDs and attention masks for captions with a maximum sequence length of 50. For the Vision Transformer, we employed ViT_b16 (Ghiasi et al., 2022) with pre-trained weights and resized images to 224 × 224 pixels. The ViT model processes images, and Bangla-BERT processes text, with their outputs fused into a joint feature space. A sigmoid activation at the final output provides binary classification.

5 Benchmark Evaluation and Discussion

Table 4 summarizes the performance of textual, visual, and multimodal models in terms of accuracy, precision, and F1 score. For textual models, BiLSTM + Attention performs poorly (F1 = 0.19), while LSTM + Attention achieves the best results (F1 = 0.60). BiLSTM + CNN (F1 = 0.58) improves performance by leveraging convolutional layers, and BanglaBERT performs similarly (F1 = 0.56), benefiting from pre-trained embeddings.

For the visual-only models, InceptionResNetV2, ResNet-50, and NASNet achieve moderate performance (F1 ranges from 0.34 to 0.50), suggesting room for improvement in extracting meaningful visual features. InceptionV3 and VGG16 both perform slightly better, with VGG16 showing more consistency across metrics. Similarly, among the models with PA (Positional Attention), ResNet-50 achieves slightly higher and more consistent performance compared to VGG16. ViT and InceptionResNetV2 + PA both achieve the highest accuracy of 0.63.

Interestingly, combining modalities did not improve results significantly; most multimodal models achieve similar F1 scores, showing limited gain from integrating visual and textual features. VGG19 + SBB shows the best balance across metrics, with an accuracy of 0.64 and an F1-Score of 0.49, highlighting its potential for multimodal tasks. VGG16 (Att) + SBB achieves comparable performance to other multimodal configurations (F1 = 0.49), though attention did not significantly improve results. These findings suggest that further refinement in model architecture or additional data may be necessary to leverage multimodal features effectively for hate detection in Bangla memes.

App.	Model	A	P	F1
Tex.	Bi-LSTM + Attention	0.36	0.13	0.19
	Bi-LSTM + CNN	0.57	0.59	0.58
	Bangla BERT	0.58	0.56	0.56
	LSTM + Attention	0.60	0.59	0.60
Vis.	InceptionResNetV2	0.41	0.57	0.34
	ResNet-50	0.49	0.56	0.49
	NASNet	0.49	0.56	0.50
	InceptionV3 + PA	0.49	0.54	0.49
	VGG16	0.52	0.54	0.53
	InceptionV3	0.54	0.53	0.54
	ResNet-50 + PA	0.59	0.57	0.58
	VGG16 + PA	0.59	0.55	0.55
	NASNet + PA	0.63	0.40	0.49
	InceptionResnet50V2 + PA	0.63	0.40	0.49
	VIT	0.63	0.40	0.49
MultiM.	VIT + SBB	0.63	0.40	0.49
	VGG19 + SBB	0.64	0.49	0.49
	VGG16 + SBB	0.63	0.40	0.49
	VGG16 + BBB	0.63	0.40	0.49
	VGG19 + BBB	0.63	0.40	0.49
	VGG16(Att) + SBB	0.64	0.40	0.49

Table 4: Performance of the models on the Ex-Mute dataset. Here, A, P, and F1 represent accuracy, precision, and weighted F1 scores. SBB: SagorSarker Bangla BERT, BBB: BUET Bangla BERT, Tex: Textual, Vis: Visual, MultiM: MultiModal

6 Conclusion

In this paper, we introduced ExMute, a multimodal dataset enriched with contextual information to support the detection of hateful memes in Bangla, code-switched, and code-mixed captions. Our findings indicate that textual models outperform visual-only models; however, combining visual and textual features yields the most accurate results, demonstrating the strength of multimodal analysis for identifying hateful content. We observed that model performance can be affected by class imbalance, leading to a bias toward certain classes. To address this, future work will focus on expanding the dataset and exploring advanced computational methods to reduce bias. Additionally, we plan to improve accuracy and incorporate context prediction through Generative AI, CLIP architecture, and comprehensive ablation studies, enhancing the model's interpretability and effectiveness in real-world applications.

References

Faseela Abdullakutty and Usman Naseem. 2024. Decoding memes: A comprehensive analysis of late and early fusion models for explainable meme analysis. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1681–1689, New York, NY, USA. Association for Computing Machinery.

Abdullah Al Maruf, Ahmad Jainul Abidin, Md Mahmudul Haque, Zakaria Masud Jiyad, Aditi Golder, Raaid Alubady, and Zeyar Aung. 2024. Hate speech detection in the bengali language: a comprehensive survey. *Journal of Big Data*, 11(1):97.

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*.

Dzmitry Bahdanau. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md Nur Hossain. 2021. Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591.

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.

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

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. 2022. What do vision transformers learn? a visual exploration. *Preprint*, arXiv:2212.06727.

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

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.

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

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022b. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Khondoker Ittehadul Islam, Tanvir Yuvraz, Md Saiful Islam, and Enamul Hassan. 2022. EmoNoBa: A dataset for analyzing fine-grained emotions on noisy Bangla texts. 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 (Volume 2: Short Papers)*, pages 128–134, Online only. Association for Computational Linguistics.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md. Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from bengali memes and texts. *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.

Sam Levin. 2017. Moderators who had to view child abuse content sue microsoft, claiming ptsd. *The Guardian*, 12.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Shahira Mukhtar, Qurat Ul Ain Ayyaz, Sadaf Khan, Atiya Muhammad Nawaz Bhopali, Muhammad Khalid Mehmood Sajid, Allah Wasaya Babbar, et al. 2024. Memes in the digital age: A sociolinguistic examination of cultural expressions and communicative practices across border. *Educational Administration: Theory and Practice*, 30(6):1443–1455.

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.

Andreas Oldenbourg. 2024. Digital freedom and corporate power in social media. *Critical Review of International Social and Political Philosophy*, 27(3):383–404.

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. *arXiv preprint arXiv:2110.00413*.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2021a. Hs-ban: A benchmark dataset of social media comments for hate speech detection in bangla. *arXiv preprint arXiv:2112.01902*.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. Bd-shs: A benchmark dataset for learning to detect online bangla hate speech in different social contexts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5153–5162.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021b. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.

Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. " they don't leave us alone anywhere we go" gender and digital abuse in south asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

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

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2020. Techtexc: Classification of technical texts using convolution and bidirectional long short term memory network. *arXiv preprint arXiv:2012.11420*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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.

Michele L Ybarra, Kimberly J Mitchell, Janis Wolak, and David Finkelhor. 2006. Examining characteristics and associated distress related to internet harassment: findings from the second youth internet safety survey. *Pediatrics*, 118(4):e1169–e1177.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.