2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)
Military Institute of Science and Technology (MIST), Dhaka-1216, Bangladesh

BengaliHateCB: A Hybrid Deep Learning model to Identify Bengali Hate Speech Detection from Online Platform

Sagor Kumar Saha[1], Afrina Akter Mim [1], Sanzida Akter [1], Md. Mehraz Hosen [1], Arman Habib Shihab [1], and Md Humaion Kabir Mehedi [2]

[1]Department of Computer Science and Engineering,Bangladesh University of Business and Technology, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, Brac University, Dhaka, Bangladesh
[1]{*sagorsaha746@gmail.com, afrina141mim@gmail.com, sanzidaaktersuraiya@gmail.com, bhuiyanmehraz@gmail.com, swshihab01@gmail.com*}
[2]*humaion.kabir.mehedi@g.bracu.ac.bd*

***Abstract*—Online issues including hate speech, abusive communications, and harassment have been exacerbated by the rising number of Internet users. People in Bangladesh often face online harassment and threats expressed in Bengali on various social media platforms. Also, there has not been nearly enough investigation into the possibility of Offensive language in Bengali literature. Although finding realistic ways to reduce hate speech in Bengali texts is urgently needed, there is a notable lack of study in the area of Bengali abusive speech detection, despite the widespread detrimental impacts of abusive text on people's well-being. The results of this research provide a method for spotting bad hateful comments in Bengali online profiles. This research provides a methodology to identify potentially manipulative hate speech in Bengali social media postings. The BERT architecture is used to gather characteristics of Bengali texts. The next step in hate speech classification is to use a Convolutional Neural Network (CNN) model including a softMax activation function. We propose a new model, BERT-CNN, that combines both models. On the Bengali Hate Speech from Social Platforms (BD-SHS) dataset, the BERT-CNN model outperformed most baseline architectures, with accuracy, precision, recall, and F1-scores of 95.67%, 93.55%,92.67%, and 94.44%, respectively. According to our research, the method we suggested for spotting hate speech in Bengali writings posted on social networking sites works well, which can lessen online hate comments and foster a more civilized online community.**

***Index Terms*—Natural language processing, bidirectional encoder representations from transformers, convolutional neural network, social platform, hate speech**

I. INTRODUCTION

Hateful language is a serious problem on social media, and it's especially prevalent in languages like Bangla which have a big online user population. Hateful language can seriously harm the people who are targeted. Intense feelings of vulnerability, helplessness, anxiety, depression, embarrassment, revenge, retaliation, and, in certain situations, suicidal thoughts are experienced by the victims [1]. The difficulty of identifying hate speech in Bangla stems from the language's intricacy and the paucity of resources.

Because Bangla is a morphologically rich language, words are created by joining together morphemes, or meaning units. Because of this, creating hate speech detection models that can generalize to fresh, untested data is challenging. Furthermore, extensively annotated datasets of hate speech in Bangla are lacking. These obstacles notwithstanding, recent developments in natural language processing (NLP) have allowed for the creation of efficient hate speech detection models for Bangla. In this paper, we propose a novel model for hate speech detection that combines the advantages of two potent natural language processing techniques: CNN and BERT. The key contributions of our paper are:

- First, extracted the data of Bangla Hate Speech and preprocessed it.
- Second, determine what is going to be extracted from the data, such as speech type, category, etc.
- Then, use BERT to extract contextual features from Bangla text, and it then uses a CNN to extract local features from the BERT representations.
- The output of the CNN is fed into a fully connected layer to predict whether the text is hate speech or non-hate speech.
- Finally, find out the most suitable model with the help of performance metrics.

The remaining research is structured as follows: The relevant work from our study is shown in Section II, and the study's framework and complete methodologies for the suggested BERT-CNN—which is utilized to identify hate speech—are provided in Section III. The empirical findings and analysis are presented in Section IV, and the study is brought to a close in Section V.

979-8-3503-8577-9/24/$31.00 ©2024 IEEE

II. Related Work

To detect abusive speech on social media, a supervised learning-based method is often used. Sazzed et al. [2] focused on spotting abusive language in internet forums. The authors compiled 8,600 comments from Facebook and YouTube into five categories. Various deep learning and transformer models were evaluated using statistical measures to detect and remove malicious content. The BERT model [3] achieved the highest accuracy of 80% using the new dataset and 97% with an existing dataset of 30,000 records. The study faced challenges due to the lack of public datasets and limited research on hate speech in languages other than English.

Das et al. [4] used an attention-based recurrent neural network to detect hate speech in Bengali language [5] on social media. Into a training and a testing set,the dataset was divided and various machine learning algorithms were compared and analyzed. The attention-based decoder achieved the best accuracy of 77%, and the model showed high precision and recall in classifying hate speech categories. However, more research is needed in this specific domain, and the paper did not provide details about the dataset's representativeness or computational resources used. Additionally, the study proposes the use of the AHP technique to evaluate the effectiveness and efficacy of monitoring patients.

Mona et al. [6] introduced a novel stacked ensemble method for hate speech recognition, incorporating a two-stage process with diverse base classifiers and logistic regression for meta-level classification. This approach outperforms traditional stacking techniques [7], achieving remarkable F1-scores of 92.5% (Davidson dataset), 88.0% (HatEval dataset), and 85.5% (COVID-HATE dataset). It successfully addresses challenges posed by limited labeled data and the evolving nature of hate speech. Key terminologies in this context include hate speech, stacking, and logistic regression.

Torki et al. [8] addressed the classification of toxic comments in user-generated online content [9]. They develop a multi-label classification [10] scheme employing an ensemble of CNN, LSTM, and GRU neural networks. The model achieves impressive F1-scores of 0.828 for toxic/nontoxic classification and 0.872 for identifying toxicity types on the Wikipedia talk edits dataset. This research contributes significantly to content moderation, especially for imbalanced data. Key terms encompass toxic comments, deep learning, and F1-score.

Sayem et al. [11] confronted the challenge of classifying toxic comments in Bangla online conversations. Their objective is to develop a classification system capable of detecting various forms of toxicity in comments. The study explores multiple classification methods, including SVM, Naive Bayes, and neural networks like BP-MLL. Preprocessing steps involve punctuation removal and text tokenization. The best-performing model, a BP-MLL Neural Network, achieves a 60.00% accuracy on the test set, with the lowest hamming loss and log loss. This research significantly advances toxicity classification in Bangla conversations, relying on metrics such as accuracy and hamming loss. Notable terms encompass toxic comments, multi-label classification, and BP-MLL Neural Network. Ghosal et al. [12] presented mBERT uncased + FSVMCIL + HS, a potent hate speech detection model that surpasses other models on the NNTI dataset, achieving a remarkable 2.35% increase in F1-score and an impressive 9.11% boost in accuracy. This underscores its efficacy in detecting Bengali hate speech on social media.

Mahmud et al. [13] present a novel machine-learning approach for detecting abusive Bangla social media comments. Incorporating formal justifications and semantic meaning, Logistic Regression achieves an impressive 97% accuracy, emphasizing the importance of context in classification.

Therefore, we suggest a transformer-based method that employs the BERT-CNN hybrid deep learning strategy to identify hate speech in Bengali. We present the hybrid model, which combines the two algorithms, to detect abusive language in Bengali.

III. Methodology

Following are some portions that illustrate the technique that has been proposed:

A. Outline of the Architecture

We examined a sizable and varied dataset of Bangla text that had been classified as either hate speech or non-hate speech for our study. Fifty thousand comments gathered from various social media platforms make up the dataset. All languages also have a distinct vocabulary of their own. Therefore, a unique approach is required to define the Bengali language. We suggest a brand-new hate speech detection model for Bangla that combines CNN and BERT, two powerful deep learning algorithms. Our research work's outline is shown in Fig. 1.

B. Data Analysis and Prepossessing

Fifty thousand comments gathered from various social media platforms make up the dataset. Human annotators label the comments, and a rigorous process was put into place to guarantee consistency and accuracy in the labeling process.

The dataset is formed of three subsets: the training set, test set, and validation set. The training set has a total of 40,225 comments, while the test set and validation set both have 5,000 comments. 19,325 hate speech and 20,900 non-hate speech comments make up the training dataset, and the targets are broken down by gender, individual, and group. Hate speech is classified as slander, gender, religion, and incitement to violence. The Kaggle dataset we used to be accessible to the general public[1].

[1]https://www.kaggle.com/datasets/naurosromim/bdshs/

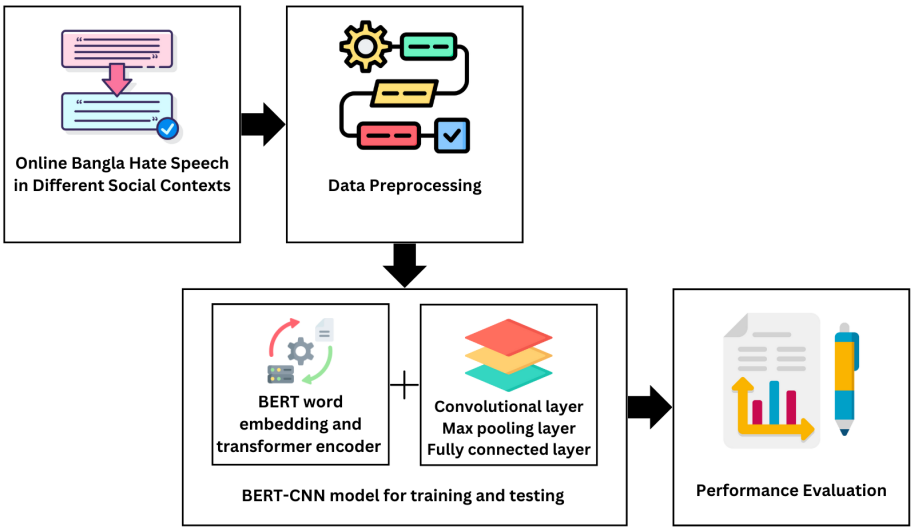

Fig. 1. Workflow of the proposed architecture for identifying hate speech in Bengali

C. Experimental Setup and Data Visualization

Python is frequently used for gathering data, pre-processing, testing, and evaluation at different phases of the construction of deep learning models. To perform simple mathematical computations, NumPy was utilized. TensorFlow was also used to implement the GPU performance of the neural network. We used the validation dataset to create the deep learning model's baseline, and we assessed the test data to finish things off.

Once more, the text contains a wide variety of data, such as words, emoticons, and emojis, so processing it could be difficult. Many users frequently use different emoticons and emojis when leaving comments on social media platforms. In actuality, they work incredibly well for expressing your feelings and ideas. That being said, these opinions might also be offensive. As such, a large portion of abusive speech detection relies on emoticons and emojis. We display the dataset in this section according to its type and target from the training and testing datasets. We can see the counts against the corresponding attributes in the following figures. Figures 2 and 3 graphically depict the type and target of the data set visualization:

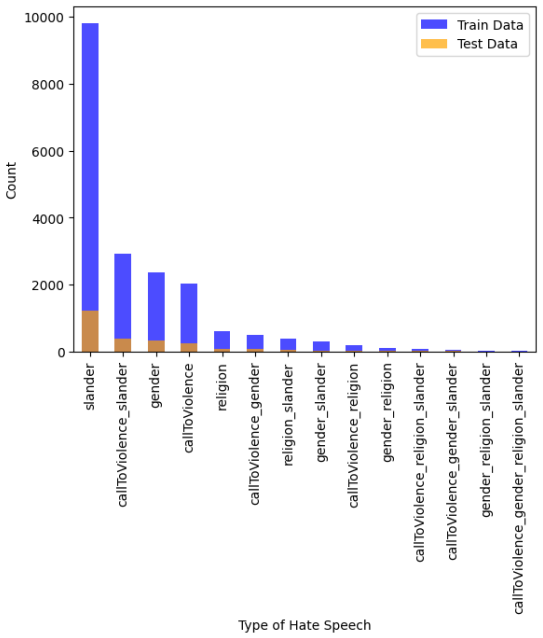

Fig. 2. Type of the Hate Speech

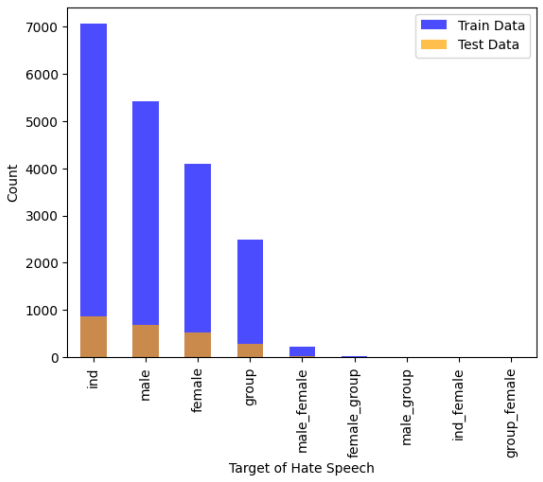

Fig. 3. Targeted victim of the Hate Speech

D. Model Development

Hate speech is a major issue on social media, and it is particularly common in languages with a big internet user base, such as Bangla [5].

The difficulty of identifying hate speech in Bangla stems from the language's intricacy and the paucity of resources. However, efficient hate speech detection models can already be created because of recent developments in natural language processing (NLP). In this study, we introduce a unique model that combines the strengths of two powerful Deep learning techniques: CNN and BERT, for the purpose of identifying abusive messages in Bangla.

BERT: BERT is a language model that Google's artificial intelligence has developed and has already undergone training. The Transformer architecture, a kind of neural network that excels at natural language processing tasks, is the foundation of BERT. Through extensive training on a vast corpus of text and code, BERT acquires the ability to represent words and phrases in a way that accurately conveys their context. Because of this, BERT is an effective tool for many tasks related to natural language processing, such as text classification. This facilitates the computer's understanding of the meaning of the text's unintelligible language. With the text in the sidebar, the context is established. Transformers are the basis of its operation. A two-way self-monitored model is denoted by the acronym BERT [14] and is founded on deep learning. Depending on the link between the entities involved in the bidirectional self-supervised model—more specifically, the encoder and the decoder—the weight between them is dynamically determined.

CNN: One kind of neural network that works well for extracting local features from data is a convolutional neural network (CNN) [15]. CNNs can be used for text classification in addition to the common application of CNNs for image classification. The feature extraction method that utilizes the CNN model's convolution process has several advantages over the tf-idf methodology. Using a trainable neural network architecture that considers multiple hidden layers to extract the hidden characteristics from tweets is one of its benefits. Contextual features can be extracted on their own without requiring the assistance of human feature engineers. CNN is additionally capable of recognizing a phrase that is semantically identical [16].

Proposed BERT-CNN: The proposed architecture consists of four levels: the output, pooling, convolutional, and embedding layers. The BERT model generates both word vectors of the input sequence and a main input matrix. The input vector matrix is then used by the convolutional layer to produce feature maps. The pooling layer takes the feature vectors with the highest values and extracts them from the convolutional layer's output. The feature vectors are finally received by a fully connected layer, which classifies the data using them. Because local characteristics are extracted using a Convolutional Neural Network, BERT-CNN reduces the dimensionality of the data. The convolutional layer is essential to convolutional neural networks (CNNs). Recapitulating the main goals of the pooling layer are to extract the most representative portion of the feature maps and the output of the Convolutional layer. The feature maps that are produced are determined by the filter sizes, and vectors of a predefined size are produced by a pooling function. In this study, the pooling function is also utilized to extract the important features from feature maps. The SoftMax function, which has the following formula, is the model's output function after that.

$$M_i = \frac{\exp(x)}{\sum_{j=1}^{c} \exp\left(x_j\right)} \tag{1}$$

The formula specifies the difference between the obtained distribution and the real classification distribution:

$$L_{\text{loss}} = \sum_{S \in T} \sum_{i=1}^{L} \hat{M}_i(C) \log\left(M_i(C)\right) \tag{2}$$

Where n, and L indicate training set and class respectively.

IV. Empirical Evaluation

A. Performance Evaluation-Metric

The Performance Evaluation-Metric is used in order to assess the model.Precision, recall and F1 score are the commonly used metric in automated text classification to determine how well a model classifies any offensive text from a large sample.

B. Results and Discussion

TABLE I
PERFORMANCE COMPARISON OF BERT-CNN ON BD-SHS

Model	Precision(%)	Recall(%)	F1-score(%)
CNN	86.4	87.6	85.5
LSTM	86.3	86.7	87.0
GRU	88.3	88.4	88.3
BERT	89.1	89.0	89.2
SVM + U	88.9	88.7	88.7
BiLSTM + MFT	89.2	89.2	89.1
BiLSTM + BFT	89.9	89.9	89.9
BiLSTM + RE	90.1	90.0	90.0
SVM + U + C	90.8	90.7	90.8
SVM + C	91.1	90.9	90.9
BiLSTM + IFT	91.0	91.0	91.0
BERT-CNN	**93.55**	**92.67**	**94.44**

In Table I, we compare our proposed model BERT-CNN with some other basic deep learning models such as CNN, LSTM, GRU, and BERT and with some hybrid models such as BiLSTM + MFT, BiLSTM + BFT, BiLSTM + RE, BiLSTM + IFT, SVM + U + C, and SVM + C. where our proposed BERT-CNN model fared better on the Bengali Hate Speech from Social Platforms (BD-SHS) dataset than the majority of baseline architectures with 92.67%, 93.55%, recall, and precision as well as F1-score 94.44%, in that order. The weighted average precision, recall, and F1-score we got from

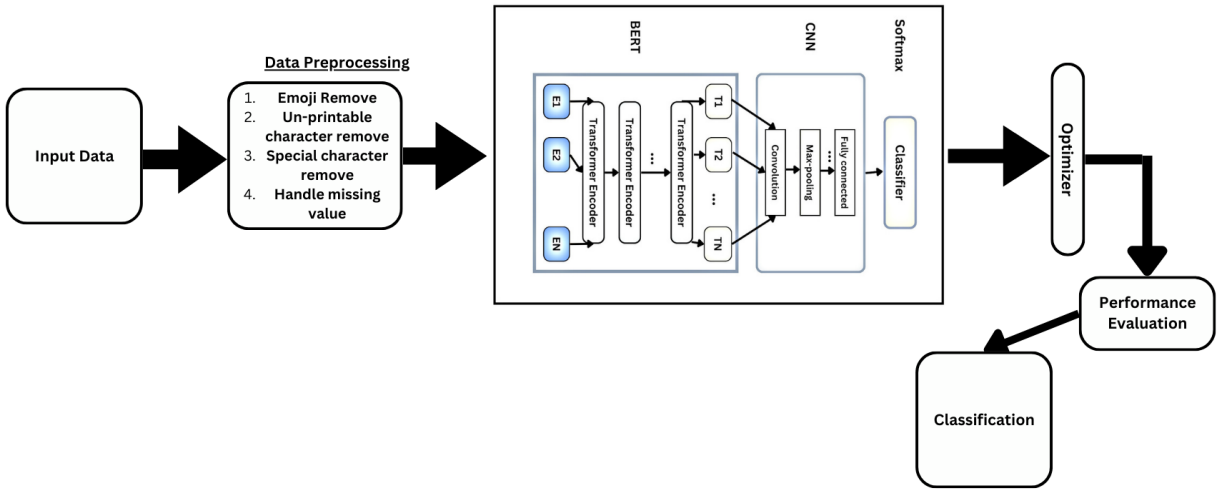

Fig. 4. Architecture of our proposed model BERT-CNN for identifying hate speech in Bengali

the previous research [17] is less then BERT-CNN model which is depicted in Table I.

V. Conclusion

This research has the potential to identify hate speech and provide mitigation strategies to improve the safety and inclusivity of online communities. A sophisticated hybrid model that we have built uses CNNs to categorize the context and BERT to control the Bangla language. Our model's sophisticated BERT-powered contextual awareness makes it extremely sensitive to even the smallest details and relationships in text. This feature enables our model to recognize even the subtlest examples of hate speech, which could be challenging to find with more conventional methods. Further, our model recognizes word and phrase combinations frequently linked to hate speech by using Convolutional Neural Networks (CNNs) to extract local attributes and patterns. After a thorough testing process on multiple datasets, our hybrid model achieves 93.55% Precision, 92.67% Recall, and 94.44% F1-score, outperforming existing models in hate speech recognition.

Our findings demonstrate that integrating CNN and BERT [18] can enhance hate speech detection systems' functionality and enable them to process content at varying intensities. We must continue refining our strategies to combat the increasing problems posed by hate speech on online platforms. Our research offers a solid foundation for creating increasingly sophisticated systems for detecting hate speech, which can improve the safety and inclusivity of online communities.

VI. Future Work

Future research could use temporal patterns and user data to accomplish this, and transfer learning could be used to improve model performance. The accuracy and adaptability of the model can be increased by using these techniques. We're dedicated to enhancing our output and creating more sophisticated technologies to identify hate speech and improve online communities. We understand how critical it is to improve our model's efficacy and utility for Bengalis worldwide. We have chosen to use English-written Bangla text to train the model to accomplish this goal. This will make it possible for the model to precisely comprehend and interpret the subtleties of the language as it is spoken in nations where English is the primary language.

Furthermore, we'll make sure the model is trained using a variety of Bangla dialects that are spoken by Bengalis all over the world. This will support the model's accurate interpretation and comprehension of hate speech. and ensuring that it works for all varieties of Bengalis.

References

[1] E. Balt, S. Mérelle, J. Robinson, A. Popma, D. Creemers, I. van den Brand, D. Van Bergen, S. Rasing, W. Mulder, and R. Gilissen, "Social media use of adolescents who died by suicide: lessons from a psychological autopsy study," *Child and adolescent psychiatry and mental health*, vol. 17, no. 1, p. 48, 2023.

[2] S. Sazzed, "Abusive content detection in transliterated Bengali-English social media corpus," in *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Online:

Association for Computational Linguistics, Jun. 2021, pp. 125–130. [Online]. Available: https://aclanthology.org/2021.calcs-1.16

[3] K. M. Hasib, N. A. Towhid, K. O. Faruk, J. Al Mahmud, and M. Mridha, "Strategies for enhancing the performance of news article classification in bangla: Handling imbalance and interpretation," *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106688, 2023.

[4] A. K. Das, A. A. Asif, A. Paul, and M. N. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578–591, 2021. [Online]. Available: https://doi.org/10.1515/jisys-2020-0060

[5] S. Thapa, A. Maratha, K. M. Hasib, M. Nasim, and U. Naseem, "Assessing political inclination of bangla language models," in *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, 2023, pp. 62–71.

[6] M. K. A. Aljero and N. Dimililer, "A novel stacked ensemble for hate speech recognition," *Applied Sciences*, vol. 11, no. 24, p. 11684, 2021.

[7] K. M. Hasib, N. A. Towhid, and M. R. Islam, "Hsdlm: a hybrid sampling with deep learning method for imbalanced data classification," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 11, no. 4, pp. 1–13, 2021.

[8] M. Ibrahim, M. Torki, and N. El-Makky, "Imbalanced toxic comments classification using data augmentation and deep learning," in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2018, pp. 875–878.

[9] A. Amira, A. Derhab, S. Hadjar, M. Merazka, M. G. R. Alam, and M. M. Hassan, "Detection and analysis of fake news users' communities in social media," *IEEE Transactions on Computational Social Systems*, 2023.

[10] K. M. Hasib, A. Tanzim, J. Shin, K. O. Faruk, J. Al Mahmud, and M. Mridha, "Bmnet-5: A novel approach of neural network to classify the genre of bengali music based on audio features," *IEEE Access*, vol. 10, pp. 108 545–108 563, 2022.

[11] A. Jubaer, A. Sayem, and M. A. Rahman, "Bangla toxic comment classification (machine learning and deep learning approach)," in *2019 8th international conference system modeling and advancement in research trends (SMART)*. IEEE, 2019, pp. 62–66.

[12] S. Ghosal, A. Jain, D. K. Tayal, V. G. Menon, and A. Kumar, "Inculcating context for emoji powered bengali hate speech detection using extended fuzzy svm and text embedding models," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.

[13] T. Mahmud, S. Das, M. Ptaszynski, M. S. Hossain, K. Andersson, and K. Barua, "Reason based machine learning approach to detect bangla abusive social media comments," in *International Conference on Intelligent Computing & Optimization*. Springer, 2022, pp. 489–498.

[14] S. Quadri *et al.*, "Encoder/decoder transformer-based framework to detect hate speech from tweets," in *Intelligent Data Analytics, IoT, and Blockchain*. Auerbach Publications, 2024, pp. 195–207.

[15] K. Md. Hasib, M. Oli Ullah, M. Imran Nazir, A. Akter, and M. Saifur Rahman, "Icdp: An improved convolutional neural network model to detect pneumonia from chest x-ray images," in *International Conference on Big Data, IoT and Machine Learning*. Springer, 2023, pp. 467–479.

[16] V. S. Raj, C. N. Subalalitha, L. Sambath, F. Glavin, and B. R. Chakravarthi, "Conbert-rl: A policy-driven deep reinforcement learning based approach for detecting homophobia and transphobia in low-resource languages," *Natural Language Processing Journal*, vol. 6, p. 100040, 2024.

[17] N. Romim, M. Ahmed, M. S. Islam, A. Sen Sharma, H. Talukder, and M. R. Amin, "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*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 5153–5162. [Online]. Available: https://aclanthology.org/2022.lrec-1.552

[18] K. M. Hasib, M. A. Rahman, M. I. Masum, F. De Boer, S. Azam, and A. Karim, "Bengali news abstractive summarization: T5 transformer and hybrid approach," in *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2023, pp. 539–545.