Unraveling Hate Speech: A Clear Approach for Bengali Language

1st S M Samin
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
samin.cse.200104112@aust.edu

2nd Asad Chowdhury
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
asad.cse.200104127@aust.edu

3rd Aman Bhuiyan
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
aman.cse.200104102@aust.edu

4th Tanzilur Rahman
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
tanzilur.cse.200104103@aust.edu

I. INTRODUCTION

Social media is like a big public space where people can share their thoughts and feelings. Platforms like YouTube and Facebook are used by many different people. But, sometimes, people get bullied or see mean things about topics like sexism, racism, and politics. There's also more cyberbullying and blackmail happening, making the online environment not very friendly.

Finding and stopping hate speech on social media has its own difficulties, like having small and uneven collections of examples, choosing the right computer models, and picking the best ways to analyze the information. For Bengali speakers, these challenges are even bigger. Bengali is a big language spoken by lots of people in Bangladesh and India, but there aren't enough tools for doing research on this topic.

Because Bengali doesn't have many resources, it's hard to create and use computer programs that can help with real-life problems, especially when there are not many examples and tools for classifying Bengali text. So, it's really important to study and stop hate speech on social media in Bengali.

In conclusion, it's very important to do research and take steps to prevent hate speech on social media in Bengali. The lack of examples and tools for classifying Bengali text shows how urgent it is to address this problem.

II. RELATED WORK

In recent years, there has been an increased interest in identifying hate speech on social media platforms, particularly in languages other than English. Detecting hate speech in Bengali is crucial due to the widespread use of social media in this language. However, existing methods for identifying hate speech in Bengali require improvement in terms of accuracy and interpretability.

In the study titled "DeepHateExplainer: Explainable Hate Speech Detection in Under-resourced Bengali Language" (Paper 1), a diverse set of models, including SVM, KNN, CNN, Bi-LSTM, and Conv-LSTM, is employed for explainable hate speech detection. Notably, precision scores across these models range from 0.67 to 0.79, underscoring their efficacy.

The study "Hate Speech Detection in the Bengali language: A dataset and its baseline evaluation" (Paper 2) presents a valuable dataset comprising 30,000 comments collected from Facebook pages. SVM, LSTM, and Bi-LSTM are utilized, achieving commendable accuracy rates ranging from 81.52

In the exploration of hate speech detection on public Facebook pages, "Hateful Speech Detection in Public Facebook Pages for the Bengali Language" (Paper 3) investigates this specific domain. The paper employs SVC, Linear SVC, Naive Bayes, and Random Forest, reporting varied accuracy levels. This underscores the challenges associated with detecting hate speech in the context of public Facebook pages.

Collectively, these papers contribute significantly to the understanding and advancement of hate speech detection in the Bengali language, offering valuable insights into model selection, explainability, and dataset creation.

III. DATASET REVIEW

Our dataset comprises a total of 5888 data entries. It consists of three columns: "text," "label," and "target." The "text" column likely contains the actual textual data, while the "label" column may indicate the level or category of hate speech. The "target" column represents the target label or class for the hate speech classification task. Specifically, 0 denotes the Personal category, 1 indicates the Political category, 2 represents the Religious category, and 3 signifies the Geopolitical category.

This dataset plays a crucial role in the training, testing, and validation of proposed hate speech detection models. It forms the foundation for evaluating the effectiveness of the developed approach in identifying and categorizing hate speech in the Bengali language.

Label	Target	Amount
Personal	0	2561
Political	1	727
Religious	2	908
Geopolitical	3	1691

TABLE I

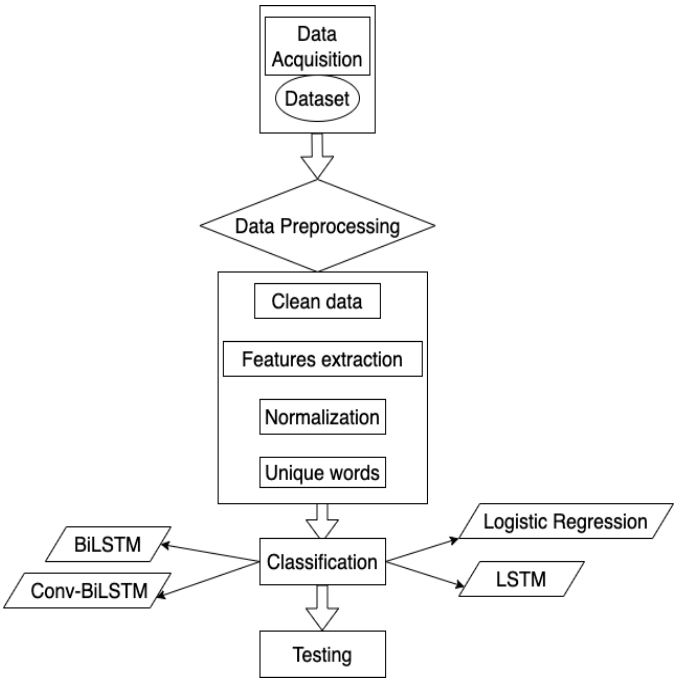

Fig. 1. Proposed Methodology

IV. METHODOLOGY

Numerous computational models have been explored to formulate a system for detecting hate speech in Bengali. In this section, we elaborate on our proposed approach, encompassing aspects such as data acquisition, data preprocessing, classification, and evaluation. The graphical representation of our proposed methodology is depicted in Figure 1.

A. *Data Acquisition*

The dataset employed in this study is referred to as the "Bengali Hate Speech Dataset." This dataset categorizes observations into distinct classes, namely political, personal, geopolitical, and religious abusive hates.

Approximately 80% of the data has been utilized for training our models, while the remaining 20% has been reserved for testing these models.

B. *Data Preprocessing*

Preprocessing plays a pivotal role in every Natural Language Processing (NLP) study. Its primary objective is to ensure that the preprocessed data does not introduce any bias or skewness into the experiments. To achieve this, we initiated the preprocessing by cleaning our dataset, wherein we removed Bengali stop words and punctuations. This step significantly reduced the size of our dataset. Subsequently, for feature extraction, we employed the TF-IDF technique. Following feature extraction, we normalized the data to ensure that all features, specifically words in this context, have a similar scale. This normalization helps prevent any skewness in the performance of our models.

C. *Classification*

In our dataset classification process, we utilized four distinct classifiers to categorize the data into personal, political, religious, and geopolitical segments. The employed models are as follows:

i. Logistic Regression: Logistic Regression is a math tool used to figure out if something belongs to one of two or more groups. It looks at the chance of something being in a group and uses a math function that turns numbers into a range from 0 to 1. This method calculates a special sum of the things we know about our data and uses that to guess the chance of each group. The group with the biggest chance is the one we say our data belongs to.

ii. LSTM (Long Short-Term Memory): LSTM is a special kind of computer network that's good at understanding and remembering information in a sequence, like words in a sentence. It's better than regular networks because it doesn't forget important things easily. This type of network is really helpful for looking at sequences of data, like sentences, and figuring out the meaning behind them. In our case, it's great for understanding and classifying text data in our hate speech dataset.

iii. BiLSTM (Bidirectional Long Short-Term Memory): BiLSTM is like an upgraded version of LSTM. It looks at sequences of data in two ways: from the start to the end and from the end to the start. This helps the model understand the context of information better by considering what happened before and after each point in the data. So, for our dataset, BiLSTM helps improve how well we can classify and understand the context of the information.

iv. ConvBiLSTM (Convolutional Bidirectional Long Short-Term Memory): ConvBiLSTM is like a mix of two different types of tools. It uses one tool to look at small patterns in the information, and another tool to understand the order of things happening. The first tool is good at finding small details, and the second one helps to understand the bigger picture. So, for our dataset, ConvBiLSTM uses both tools to do a better job at figuring out and classifying hate speech.

V. RESULTS ANALYSIS

This section delineates the outcomes derived from the experiments conducted on four distinct models: Logistic Regression, LSTM (Long Short-Term Memory), BiLSTM (Bidirectional Long Short-Term Memory), and ConvBiLSTM (Convolutional

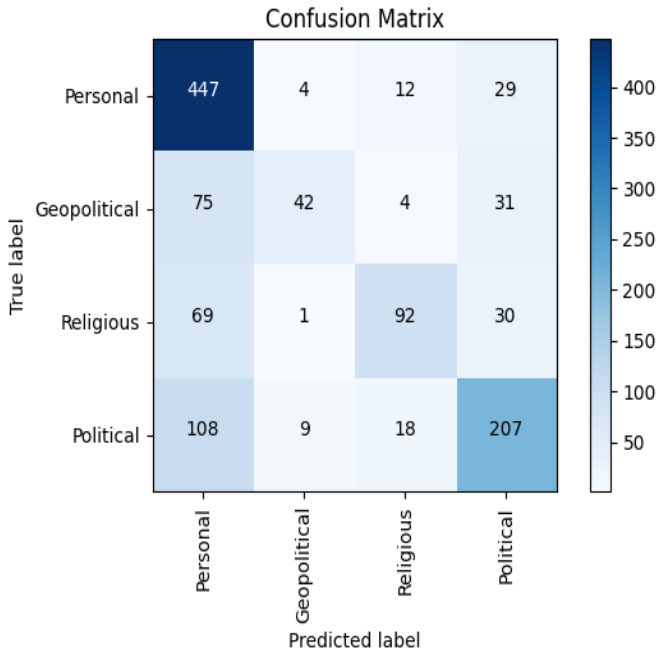

Fig. 2. Confusion Matrix

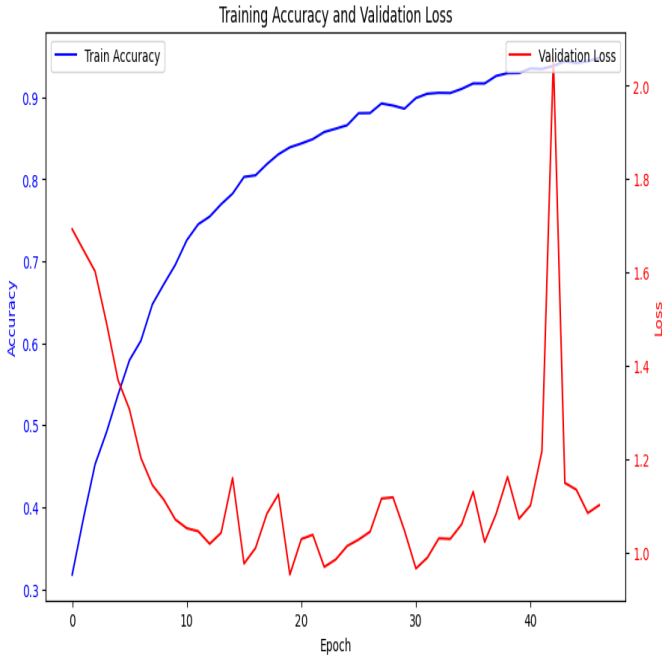

Fig. 3. LSTM: Accuracy VS Loss

Bidirectional Long Short-Term Memory). The ensuing subsections provide a detailed presentation of the obtained results.

A. *Logistic Regression:*

When we used the Logistic Regression model on our dataset, we got some numbers that show how well it performed. The Accuracy is like 67%, Precision is around 69%, Recall is about 67%, and F1 Score is 65%. We also made a picture called a confusion matrix that shows the real rates for different categories. For example, it says we correctly identified 447 cases for personal, 207 cases for political, 92 cases for religious, and 42 cases for geopolitical, as shown in Figure 2.

B. *LSTM (Long Short-Term Memory):*

When we taught the LSTM (Long Short-Term Memory) model using our data, we found that it made some mistakes. The loss value was 1.0229, which means it got some things wrong during learning. But when we tested it, it got things right about 77.41% of the time. We made a picture, called a graph, to show how it learned over time, and you can see it in Figure 3. The model learned from a total of 5887 examples, with 2561 for personal, 727 for political, 908 for religious, and 1691 for geopolitical.

C. *BiLSTM (Bidirectional Long Short-Term Memory)):*

When we taught the BiLSTM (Bidirectional Long Short-Term Memory) model using our data, it made some mistakes during the learning process. The loss value was 1.1992, showing the errors it made. But when we tested it, it got things right about 77.24% of the time. We made a picture, called a graph, to show how it learned over time, and you can see it in Figure 4. The model learned from a total of 5887 examples, with 2561 for personal, 727 for political, 908 for religious, and 1691 for geopolitical.

D. *ConvBiLSTM (Convolutional Bidirectional Long Short-Term Memory):*

When we taught the ConvBiLSTM (Convolutional Bidirectional Long Short-Term Memory) model using our data, it made some mistakes during the learning process. The loss value was 1.2604, showing the errors it made. But when we tested it, it got things right about 75.29% of the time. We made a picture, called a graph, to show how it learned over time, and you can see it in Figure 5. The model learned from a total of 5887 examples, with 2561 for personal, 727 for political, 908 for religious, and 1691 for geopolitical.

A comparative analysis of the four models is presented in Table 2. This table provides a systematic examination of the performance metrics associated with each model, facilitating a comprehensive understanding of their respective strengths and weaknesses. The metrics considered encompass accuracy, precision, recall, and F1 Score, providing a holistic evaluation of the models' effectiveness in hate speech detection for the Bengali language.

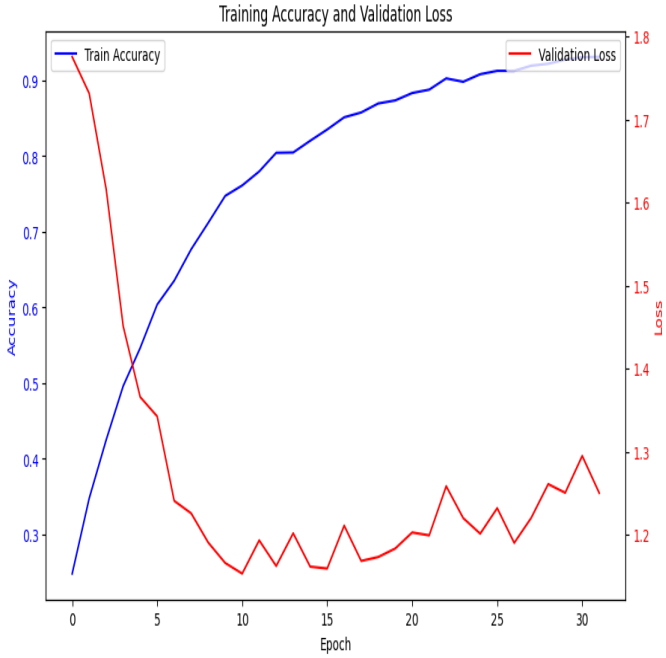

Fig. 4. BiLSTM: Accuracy VS Loss

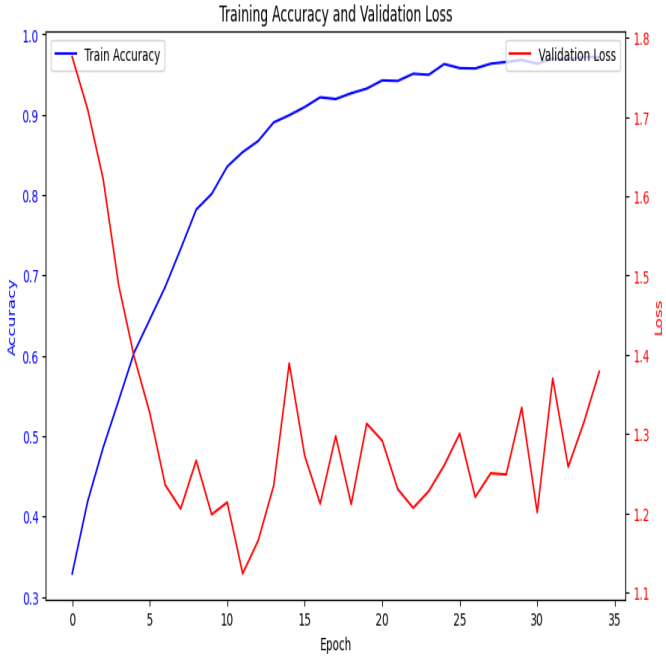

Fig. 5. ConvBiLSTM: Accuracy VS Loss

Model	Accuracy	Loss
LSTM	77.41%	1.0229
BiLSTM	77.24%	1.1992
ConvBiLSTM	75.29%	1.2604

TABLE II

VI. CONCLUSION

In short, this paper talks about how important it is to find and stop hate speech on social media, especially in languages like Bengali that don't have many resources. We suggest a way to find hate speech in Bengali using a dataset that puts examples into different categories: political, personal, geopolitical, and religious abusive hates. The way we do this involves getting data, preparing it, figuring out what category it belongs to, and checking how well our method works, using four different tools—Logistic Regression (LR), LSTM (Long Short-Term Memory), BiLSTM (Bidirectional Long Short-Term Memory), and ConvBiLSTM (Convolutional Bidirectional Long Short-Term Memory). Out of these, LSTM did the best with an accuracy of 77.41%.

We also say it's important to keep studying and finding better ways to stop hate speech in different languages and on different social media sites. In conclusion, this paper shares important ideas about the difficulties and possibilities of dealing with hate speech in today's digital world.

VII. REFERENCE

[1] DeepHateExplainer: Explainable Hate Speech Detection in Under-resourced Bengali Language (Md. Rezaul Karim, Sumon Kanti Dey , Tanhim Islam , Sagor Sarker , Mehadi Hasan Menon , Kabir Hossain, Bharathi Raja Chakravarthi , Md. Azam Hossain , Stefan Decker)

[2]Hate Speech detection in the Bengali language: A dataset and its baseline evaluation (Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam

[3]Hateful Speech Detection in Public Facebook Pages for the Bengali Language (Alvi Md Ishmam and Sadia Sharmin)