International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 524-530 March 2026-Sept 2026 www.csjournals.com

Machine Learning Model for Hate speech detection in Bengali social media text

Mukesh[1], Dr. Naveen[2] , Dr. Anupam
MCA Student, Asst Professor, Associate Professor
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract :

Social media platforms are commonly used for communication and information sharing, they also include a lot of insuitable and dangerous information. Online text is informal and contains a lot of data, which makes it challenging. The goal of the current project is to use machine learning techniques to automatically find hate speech in Bengali social media posts. First, URLs, mentions, hashtags, and special characters are removed from the text data.

After preprocessing, text is converted into number format utilizing word and character n-grams and the TF-IDF (Term Frequency Inverse Document Frequency) to better capture text patterns. To identify between hate speech and regular language, a Logistic Regression model is employed. The model has an accuracy of about 65% after being trained and checked on labelled data. The results of this method is easy to use and useful for detecting hate speech. But, with the help of advanced deep learning techniques can give better results.

1. Introduction:

Social media now plays a major role in daily life. People share their ideas, views, and facts on Facebook, Twitter, and other sites. It helps easy communication and personal connections. But it's challenging to manually keep an eye on everything because so much things is exchanged every second.

On social media, hate speech and abusive material are common. Such information has the possible to hurt people, confuse people, and sometimes lead to major social issues. Posts written by users in informal words, informal language, and a different typesof styles make it more difficult to properly find dangerous text. Users are harmed by the harmful online surroundings that has been created by an increase in harmful language on social media [1].

It is not feasible to carefully review every post due to the large volume of data published daily. As a result, an automated method is required to quickly and effectively find hate speech. Machine learning algorithms that find patterns in data and provide predictions can solve this issue.

In order to find hate speech in Bengali social media posts, this project creates a machine learning-based algorithm. To find if the given text is hate speech or regular speech, the system uses text processing methods and classification algorithms. This improves online safety.

Social media has changed how individuals exchange information and communicate. Users share messages, images, videos, and comments on many platforms on a daily basis. On the internet, a lot of bad and damaging remarks are posted including helpful information. These remarks have the possible to harm people, strain personal bonds, and support a dangerous online atmosphere.

It is getting harder to manually keep an eye on all online information as the number of social media users keeps increasing . To quickly and exactlyfind dangerous texts, an automated method is required. Large volumes of text may be effectively checked and classified as either hate speech or non-hate speech using machine learning. dangerous information may be identified more quickly and effectively by employing text preparation and grouping tools.

This study focuses on create a simple and fast and effective machine learning model for identifying hate speech in Bangla. By limiting the share of bad information, the suggested tries to support text monitering and improve online conversation.

2. Literature review

Given the fast growth of dangerous and abusive information on online platforms, the identification of hate speech in Bangla social media text has started as a important study subject. A number of academics have improved internet safety by automatically identifying hateful remarks using machine learning and deep learning techniques.previous research has mostly focused on building datasets, extracting features, understanding meaning, and comparing model result. These investigations helped us analyze many other studies in the same field, highlighting and providing profound cultural insights on this subject [2].

The HS-BAN dataset was created by Romim et al. (2021) for Bangla hate speech detection. It includes 50k+ Bangla comments from social media. The comments are classified into hateful and non-hateful. The models used by the researchers are BiLSTM, Naïve Bayes, and Support Vector Machine (SVM). They tested all these models on their dataset . The results of this comparison indicate the better results of the BiLSTM over the other models. It performed well and achieved an F1-score of 86.78%. The model was better at detecting hate speech. Deep learning models could better capture the meaning of Bangla text. Complex text and modern writing patterns proved more difficult fo basic machine learning models. The research found that BiLSTM is an appropriate model for Bangla hate speech detection. This model works well and many researchers have used it in their research.

In this research multiple models were used including SVM, Linear Regression, and Naïve Bayes. The study reveals that the BiLSTM model performed better than the other models. The BiLSTM model achieved higher accuracy than the other models [3].

The results indicate that attention-based deep learning approaches are useful for hate speech detection. The model performed less well on social media content, due to spelling errors, informal language, etc. These problems made it hard for the model to understand some text correctly. The study suggested that improving data quality and text cleaning can help in better model results.

Das et al. [4] utilised attention-based recurrent neural networks for the detection of hate speech in Bangla. They used an attention-based LSTM model to understand the sense and the meaning of social media comments. The researchers compared different approaches and found that the LSTM-based model was better than traditional machine learning models. They found that knowing the meaning and context of the text helps to detect abusive content better.

Fayaz et al. (2025) introduced BIDWESH dataset. It has Bangla hate speech data from different local languages and social media sites. The study aimed to enhance current datasets by adding common words, idioms, and different spellings used by Bangla speakers in their daily life.

The dataset made it easier for deep learning and machine learning models to understand the different types of regional languages. But there were still issues with processing vast amounts of multilingual and code-mixed text.

In contrast to current investigations, the suggested study employs RandomOverSampler for data balancing and a basic Logistic Regression model with TF-IDF feature extraction. in place of using difficult deep learning models, this work focuses on building a easy to use and useful, and low-cost system for Bangla hate speech detection. The suggested is easier to implement, requires less computational power, and provides good result for basic hate speech classification.

2.1 Research gap:

- Small Bengali hate speech datasets.
- Models cannot handle informal and misspelled words well.
- Bengali-English mixed text is difficult to detect.
- Traditional machine learning models still need better accuracy.
- More robust models are needed for real-world social media data.

3. Research Methodology

3.1 Data collection:

The text data is cleaned using preprocessing methods after it has been gathered. In this stage, extra spaces, URLs, hashtags, special characters, and unnecessary symbols are removed from the text. After the text has been cleaned, it is formatted properly for additional processing and analysis.

3.2 Data cleaning and feature engineering:

Many important steps are included in the research methodology of the suggested hate speech detection system. The Bangla social media data is first collected using Kaggle's online resources. Social media comments, both hateful and non-hateful, are included in the dataset.

First, URLs, hashtags, mentions, special characters, emojis, and extra spaces are removed from the gathered Bangla social media data. Following cleaning, both word-level and character-level TF-IDF feature extraction are used to transform the text into numerical features.

Label Encoding is then used to translate the class labels into numerical values. The RandomOverSampler is used to balance the hate and non-hate text samples in order to address the class imbalance issue.

3.3 Dataset Splitting:

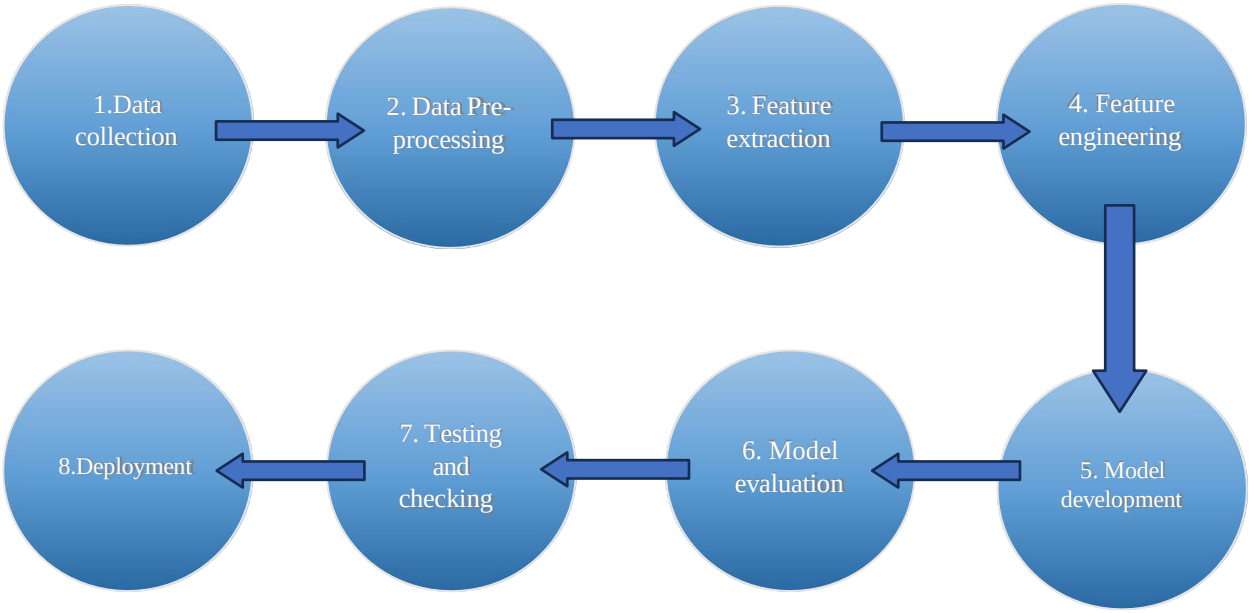
Unseen data is used to test the model after training. Lastly, Accuracy, Precision, Recall, and F1-score are used to assess the model's result. The trained model may be used to find hate speech in fresh Bangla social media posts following successful testing and checking.

3.4 Machine Learning Model:

Logistic Regression is a machine learning algorithm used for classification. It predicts whether a text belongs to a specific class, such as hate speech or non-hate speech. It is simple, fast, and effective for text classification tasks.

3.5 Mathodological workflow:

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 524-530 March 2026-Sept 2026 www.csjournals.com



Steps for Research Methodology

4. Result Analysis:

In this study, we used the Logistic Regression machine learning model to test the suggested hate speech detection system using Bangla social media text. Data was pre-processed and transformed into number formatusing TF-IDF feature extraction techniques. To address the class imbalance and improve the model's result, RandomOverSampler was employed.

The test results found that the proposed model can correctly identify between comments that are harmful and those that are not. The system was able to find important text patterns and

information in Bangla social media text with the help of word-level and character-level TF-IDF features.

This approach also discovered that data balance and preprocessing methods increased the system's total prediction accuracy.

Although the model worked well, it struggled to handle spelling changes in Bangla social media comments, mixed language, loud text, and comman terms.

Still, the proposed approach performed better than traditional methods in identifying hate speech.

5. Finding and future work:

5.1Findings:

The findings show that the proposed machine learning model can effectively detect hateful and non-hateful Bengali social media comments. The use of TF-IDF features helped improve the

classification performance and achieved good accuracy on the dataset. But the model had some difficulty with comman language, spelling mistakes, and Bengali-English mixed text. The proposed approach is simple, effective, and suitable for Bengali hate speech detection.

5.2 Limitations:

This machine learning model has some limitations. It was tested on a limited Bengali dataset, so its performance may change on larger or different datasets. The model also finds it difficult to handle local words, spelling mistakes, and Bengali-English mixed text. In addition, it may not correctly detect indirect or text-based hate speech.

5.3 Future work:

To improve accuracy, more modern models like LSTM, BiLSTM, BERT, and transformer models can be applied. These models are better at understanding social media material. It is also possible to train on a bigger dataset that includes local languages, comman words spelling errors, emoticons, and mixed language. This may improve the model's performance on actual social media data.

The technology can be improved in the future to include languages other than Bengali. Also, it may be connected to social media platforms to directly identify hate speech and check text.

Improved text preparation methods might be applied, and other deep learning and machine learning algorithms could be checked for more improvement s. These improvement can increase the system's precision, reliability , and usefulness to help to a more safe online environment.

References:

[1] K. J. Keya, A. J. Jannat, et al., “G-BERT: An fast and effective method for identifying hate speech In Bengali texts on social media,” IEEE Access, vol. 11, pp. 79697–79709, 2023.

[2] G. So-2, “Analyzing the result of deep learning models for detecting hate speech on social media platforms,” MIJST, 2024

[3] N. Romim, et al., “HS-BAN: A benchmark dataset for hate speech detection in Bangla,” arXiv preprint arXiv:2112.01902, 2021.

[4] A. K. Das, et al., “Bangla hate speech detection using attention-based recurrent neural network,” arXiv preprint arXiv:2203.16775, 2022.

[5] A. H. Fayaz, et al., “BIDWESH: A Bangla regional hate speech dataset,” 2025