Bengali Hate Speech Detection Using Deep Learning Technique

Chandan Senapati[1], Utpal Roy[1]

[1]Visva-Bharati University, Department of Computer and System Sciences, Siksha-Bhavana, Visva-Bharati, Santiniketan 731235, W.B., India

Abstract

Social media has become a part of life and a great platform to communicate with each other and share ideas. With the proliferation of online platforms, and social media, sharing of ideas, and posting comments on different issues, mainly posting abusive comments on religion, gender, political ideology, race, and other issues has become a significant concern in the digital era. These negative messages are collectively called hate speech. Hate speech promotes discrimination, hostility, or violence towards individuals or groups based on attributes such as race, religion, ethnicity, gender, etc. Hate speech in various languages has made a surge, including Bengali. Detecting hate speech in Bengali presents unique challenges due to the language's linguistic complexity, diversity, and the absence of comprehensive resources. Social media is a place of interest for researchers in the fields of Natural Language Processing, Machine Learning and Deep Learning due to its huge collection of data. In this paper, we implement the deep learning model Long Short Term Memory (LSTM), a powerful recurrent neural network (RNN) architecture to automatically learn intricate patterns and contextual information from text data and detect hate speech. LSTM networks are well-suited for sequence modeling, making them particularly effective in capturing the context and nuances of natural language. We fine-tune the LSTM model to optimize its performance for Bengali text, considering factors such as word embeddings, tokenizing, stop words, architecture, etc. To evaluate the effectiveness of our approach, extensive experiments are conducted on the given dataset, employing various evaluation metrics such as precision, recall, Macro F1-score, and accuracy. The test dataset is labeled using our model. The results demonstrate the robustness and efficiency of our LSTM-based hate speech detection system in identifying offensive or hate content in Bengali text.

Keywords

Hate speech, Natural Language Processing, Machine Learning, Deep Learning, LSTM, RNN

1. Introduction

In recent years, the proliferation of online communication platforms along with high-speed internet and smart devices, has given rise to a concerning phenomenon: hate speech. The offensive, discriminatory, or harmful language aimed at individuals or groups based on their race, religion, ethnicity, gender, political ideologies or other attributes, has become a pervasive issue on the internet. It poses a significant threat to social cohesion, balance, and the mental well-being of internet users. The social impact of hate speech and the huge collection of data have drawn the interest of researchers. Many works have been done on sentiment analysis[1], fake news detection[2], cyberbullying[3]. While various solutions have been proposed for hate

Forum For Information Retrieval Evaluation, December 15-18, 2023, Goa, India

senapatichandanglg@gmail.com (C. Senapati); roy.utpal@gmail.com (U. Roy)
0009-0006-4840-9982 (C. Senapati)

speech detection in English and other widely spoken languages, there is a need to develop effective tools for detecting hate speech in low-resource languages like Bengali. The Bengali language, spoken by over 230 million people worldwide, is one of the 22 scheduled languages of India and the official language of Bangladesh. Despite its widespread use, the detection and mitigation of hate speech in Bengali remain less-studied and underdeveloped. Addressing this gap is crucial, as online hate speech in Bengali can have real-world consequences, including inciting violence, perpetuating discrimination, and fostering division within communities. Some researchers have implemented machine learning algorithms to detect hate speech in social networks[4, 5]. This paper focuses on the development of a hate speech detection system tailored to the Bengali language, utilizing the power of deep learning techniques, specifically Long Short-Term Memory (LSTM) networks. Deep learning has proven to be highly effective in natural language processing tasks, including sentiment analysis, machine translation, and text classification. LSTM, a variant of recurrent neural networks (RNNs), is particularly well-suited for sequence modeling, making it an ideal choice for the nuanced and context-dependent nature of natural language.

Keeping this scenario in mind the organizers of HASOC[1] (2023) Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages propose 4 tasks:

Task 1- ”focus on identifying hate speech, offensive language, and profanity in different languages using natural language processing techniques”

Task 2- ”known as the Identification of Conversational Hate Speech in Code-Mixed Languages (ICHCL), addresses the challenge of identifying hate speech and offensive content in code-mixed conversations on social media. Code-mixed text includes multiple languages within a single conversation. The task is divided into two subtasks”

Task 3- ”aims to detect the various hateful spans within a sentence already considered hateful. A hate span is a set of continuous tokens that, in tandem, communicate the explicit hatefulness in a sentence”

Task 4- ”aims to detect hate speech in Bengali, Bodo, and Assamese languages[6, 7]. It is a binary classification task. Each dataset (for the three languages) consists of a list of sentences with their corresponding class (hate or offensive (HOF) or not hate (NOT)). Data is primarily collected from Twitter, Facebook, or YouTube comments”

This paper attempted to identify hate speech content in Task 4 to detect hate speech in Bengali text. The LSTM model is used for this work. The rest of the paper is structured as follows: Section 2 is the work related to hate speech detection in Bengali and other languages. Section 3 describes the Methodology, including the dataset, preprocessing steps, and LSTM model. Section 4 shows the results and findings from the experiments. Finally, it is concluded in Section 5.

2. Related Works

Different techniques of detecting hate speech have been implemented by various researchers for different languages. Machine learning models and deep learning models have been widely used in English and other widely used languages including code mixed languages. While the

[1] https://hasocfire.github.io/hasoc/2023/index.html

majority of researches have focused on widely spoken languages, there is a growing interest in extending these efforts to languages like Bengali. In this section, we review some of the related work in the field of hate speech detection, with a particular emphasis on Bengali and other non-English languages and deep learning approaches.

2.1. Hate Speech Detection in English:

Numerous studies [8, 9, 10, 11] have explored hate speech detection in English, often relying on deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These models have achieved promising results in identifying hate speech on platforms like Twitter and Facebook.

2.2. Hate Speech Detection in Multilingual Contexts:

Some researchers [12, 13, 14, 15] have extended their work to multilingual hate speech detection, acknowledging the importance of addressing the issue across different languages. These studies have employed cross-lingual techniques and multilingual datasets to develop models that can detect hate speech in multiple languages simultaneously.

2.3. Hate Speech Detection in Non-English Languages:

Several studies have been done in English language on hate speech but lesser studies have been done in non English languages specifically in Indian languages due to non availability of sufficient datasets. Some researches have been done on Marathi[16, 17], Assamese[18], Hindi[19] and Bengali[20] languages.

2.4. Deep Learning Approaches:

Deep learning techniques[21], including LSTM and its variants, have proven to be effective for hate speech detection due to their ability to capture context and sequential patterns in text. Researchers have applied LSTM-based models for hate speech detection in various languages, including English, Spanish, and German.

2.5. Bengali Hate Speech Detection:

Studies related to natural language processing (NLP) in Bengali have gained attraction in recent years. Researchers have developed Bengali language models, word embeddings, and sentiment analysis tools. These resources can be leveraged for hate speech detection in the Bengali language. While hate speech detection in Bengali is relatively underexplored, there have been efforts to address this issue. Some researchers have initiated the collection and annotation of Bengali hate speech datasets, laying the foundation for future research in this area. Some Deep learning and machine learning models[22, 23] have gained success instead of having low resources in the Bengali language.

2.6. Online Safety and Social Media Platforms:

Research in the realm of online safety and content moderation on social media platforms has emphasized the need for effective hate speech detection tools. These tools are essential for maintaining a healthy online environment and ensuring the well-being of users.

In summary, the related work encompasses a wide range of studies, from hate speech detection in English to the adaptation of deep learning models for multilingual contexts. As the field evolves, there is a growing recognition of the need to address hate speech in languages like Bengali. This research aims to contribute to this emerging area by developing a specialized hate speech detection system for the Bengali language using LSTM-based deep learning techniques. In this paper, we built a deep learning model using a Bengali training dataset to classify the test data into two categories, hate or offensive(HOF) and not hate(NOT). The Bengali training dataset and test dataset containing social media text were taken from HASOC[2] (2023).

3. Methodology

In this paper, we proposed to classify the text using a Long Short Term Memory (LSTM), is a powerful Recurrent Neural Network (RNN) architecture with persistent memory. Figure 1 shows the steps of the LSTM model through a Flow chart.

3.1. Long Short Term Memory(LSTM):

This special type of neural network is designed to work well when one has a sequence data set and there exists a long-term dependency. These types of networks can be useful when one needs a network to remember information for a longer period. This feature makes LSTM suitable for processing textual data. Figure 2 shows a typical structure of an LSTM.

There are three parts of an LSTM unit, known as gates. They control the flow of information in and out of the memory cell called the LSTM cell. The first gate is called the Forget gate, the second one is called the Input gate, and the last one is the Output gate. An LSTM unit that consists of these three gates and a memory cell or LSTM cell can be treated as a layer of neurons in a traditional feed-forward neural network, where each neuron has a hidden layer and a current state.

3.2. Data Preprocessing:

We collected datasets from HASOC(2023)[3]. The training dataset contained hate or offensive text and non-offensive text with labels from social networking sites like X(formerly Twitter), Facebook, etc for training the model. Another dataset containing text data was provided for prediction. The labeled data set which was approximately balanced contained two classes namely HOF (Hate or Offensive) and NOT (Non-hate speech). Table 1 shows the number of hate speech and non-hate speech from the training dataset. All data preprocessing works like

[2] https://hasocfire.github.io/hasoc/2023/index.html
[3] https://hasocfire.github.io/hasoc/2023/dataset.html

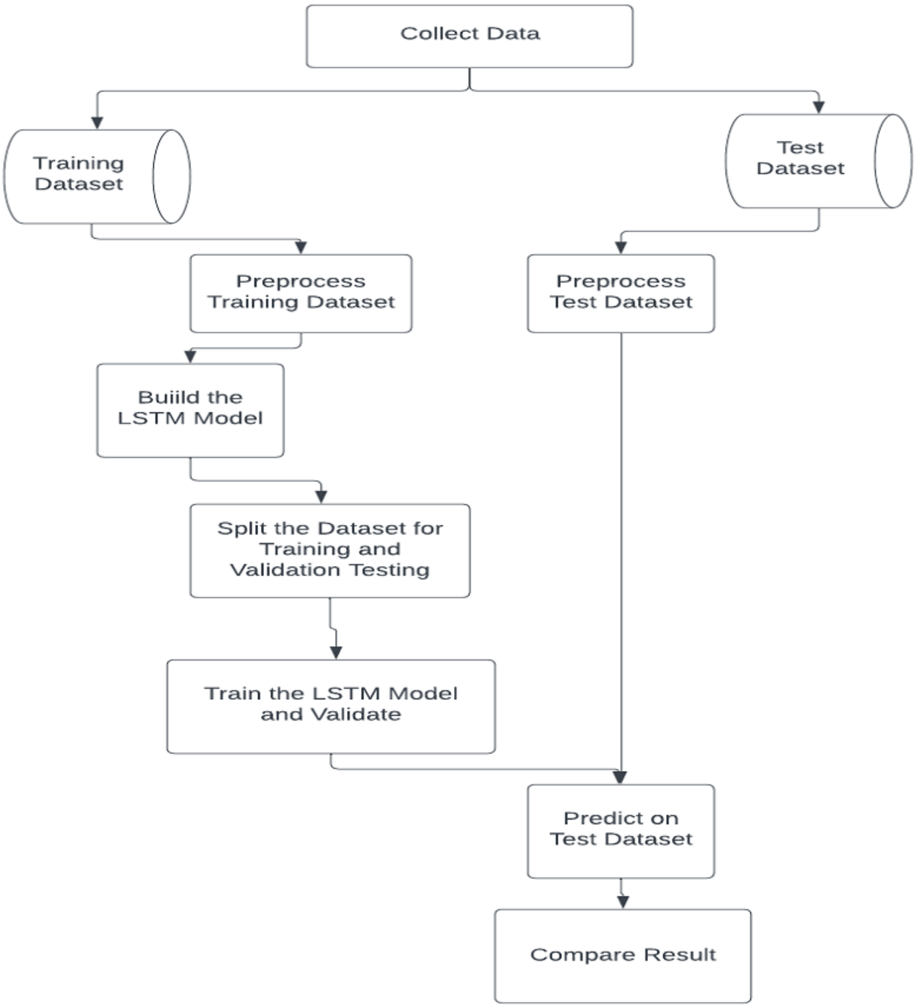

Figure 1: Flow chart for LSTM model

Class Name	Number of sentences
HOF	515
NOT	766

Table 1
Training data statistics

tokenizing, removing stop words, symbols, URLs, stemming, label encoding, etc. were done on the Bengali text data before training the model.

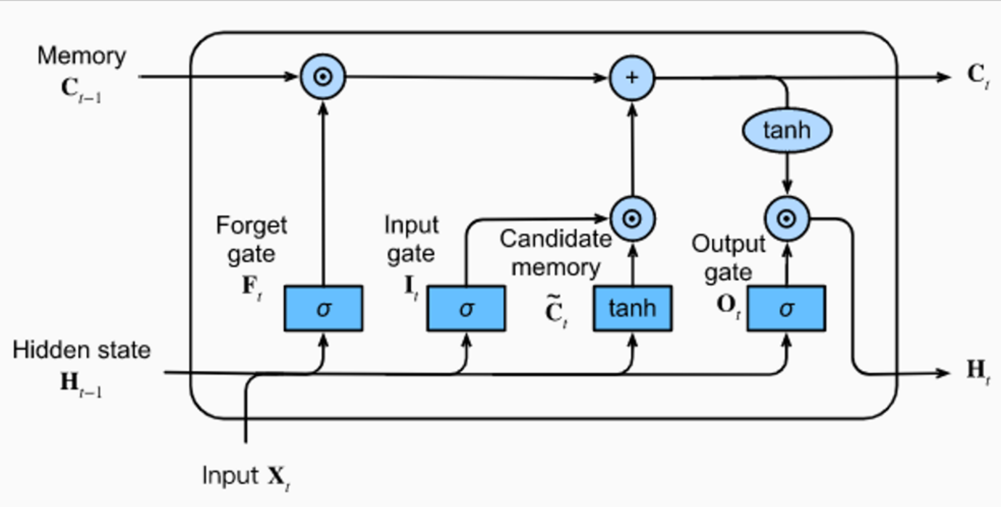

Figure 2: Structure of LSTM Unit

3.3. Model Development:

We harness the capabilities of LSTM networks to build a robust hate speech detection model for the Bengali language. Fine-tuning the LSTM architecture, optimizing hyperparameters, and experimenting with word embeddings and padding we try to maximize the accuracy and efficiency of our system. For that, We divided the data set into training and testing keeping a ratio of 80:20.

4. Result

Numerous teams participated in the HASOC (2023), Task 4. The training and test datasets[4] have been supplied, as detailed in a previous section. Among them, 22 teams submitted their system output in compliance with their specified format. The HASOC (2023) team assessed the results using the Macro F1 score as the performance metric. The summarized outcomes of these systems can be found in the accompanying Table 2. The table demonstrates the presence of dominant scores, with my own result (Serial Number 19) falling among the lower end of the scores. The results indicate that our system's performance falls short of expectations. To identify the system's weaknesses, we conducted an investigation at the macro level. Our analysis revealed that a significant number of errors can be attributed to inadequate pre-processing. Additionally, factors such as the chosen system architecture, fine-tuning parameters, and the size of the test dataset also play a role in our system's performance.

Several commonly employed performance metrics include:

- ***Accuracy:***

[4] https://hasocfire.github.io/hasoc/2023/dataset.htm

Sl no.	Participating Team	Score
1	Sanvadita	0.77025
2	FiRC-NLP	0.76422
..	....	...
19	**CHANDAN SENAPATI**	**0.50625**
20	Team +1	0.47094
21	CIT TEAM	0.37548
22	InclusiveTechies	0.35831

Table 2
Result: Evaluated by HASOC (2023, Task 4) Team

Accuracy is one of the most widely used performance measures. It is used when the target variable class of data is approximately balanced.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- ***Precision:***

Precision is another performance measure that is used to overcome the limitations of Accuracy. It provides information about the performance of a classifier with respect to false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- ***Recall:***

Recall is the proportion of positive observations that were successfully detected. It provides information about the performance of a classifier with respect to false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- ***f1 Score:***

If both Precision and Recall are important for evaluation then f1 Score can be calculated as

$$\text{f1 Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

- ***Specificity:***

Specificity measures the proportion of true negatives that are successfully identified by the model among all actual negatives.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

5. Conclusion

In conclusion, this research tries to contribute to mitigating the harmful impact of Bengali hate speech used in social networks and to provide a safer and more inclusive online environment for Bengali-speaking communities, by developing a reliable and efficient deep learning LSTM network. We also explore potential applications of our model, including content moderation on social media platforms, early detection of hate speech trends, and support for online safety initiatives. This research contributes to the ongoing efforts to combat hate speech in the Bengali language. Furthermore, we discuss the challenges faced in the context of the Bengali language. Converting emojis and emoticons to text helps to increase performance. More experiments on preprocessing of Bengali text are needed to increase the model's performance.

References

[1] S. Muthukumaran, P. Suresh and J. Amudhavel, Sentimental analysis on online product reviews using LS-SVM method: Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 12, pp. 1342-1352, 2017

[2] P. Nakov, G. D. S. Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 Check That!lab on detecting check-worthy claims,previously fact-checked claims ,and Fake News, in :Experimental IR Meets Multilinguality, Multimodality, and Interaction :12th International Conference of the CLEF Association, CLEF Virtual Event, September 21-24,volume 12880 of Lecture Notes in Computer Science , Springer, 2021,pp.264–291.URL: https://doi.org/10.1007/978-3-030-85251-1-19.

[3] W. N. H. W. Ali, M. Mohd, F. Fauzi, Identification of profane words in cyberbullying incidents within social networks: Journal of Information Science Theory and Practice 9 (2021)24–34.doi:https://doi.org/10.1633/JISTaP.2021.9.1.2.

[4] C. Paul, D. Sahoo and P. Bora, Aggression In Social Media(2020): Detection Using Machine Learning Algorithms: International Journal of Scientific and Technology Research, vol. 9, no. 4, pp. 114-117, 2020.

[5] A. Gaydhani, V. Doma, S. Kendre and L. Bhagwat, Detecting hate speech and offensive language on twitter using machine learning: An ngram and tfidf based approach, arXiv preprint arXiv:1809.08651., pp. 1-5, 2018.

[6] K. Ghosh, A. Senapati, A. S. Pal, Annihilate Hates (Task 4, HASOC 2023): Hate Speech Detection in Assamese, Bengali, and Bodo languages: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[7] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha and S. Satapara, Overview of the (HASOC) Subtracks at (FIRE) 2023: Hate Speech and Offensive Content Identification in Assamese, Bengali, Bodo, Gujarati and Sinhala, Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023, ACM, 2023.

[8] G. K. Pitsilis, H. Ramampiaro and H. Langseth, Effective hate-speech detection in Twitter

data using recurrent neural networks: Applied Intelligence, vol. 48, no. 12, p. 4730–4742, 2018

[9] M. L. Williams, P. Burnap, A. Javed, H. Liu, S. Ozalp, Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime, The British Journal of Criminology 60 (2019) 93–117. URL: https://-doi.org/10.1093/bjc/azz049. doi:10.1093/bjc/azz049. arXiv:https://academic.oup.com/bjc/articlepdf/ 60/1/93/31634412/azz049.pdf

[10] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[11] Md. A. H. Wadud, Md. M. Kabir, M. F. Mridha, M. A. Ali, Md. A. Hamid and M. M. Monowar, How can we manage Offensive Text in Social Media-A Text Classification Approach using LSTM-BOOST :International Journal of Information Management Data Insights, 2(2), 100095. https://doi.org/10.1016/j.jjimei.2022.100095

[12] K. Ghosh, A. Senapati: Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation, Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, URL:https://aclanthology.org/2022.paclic-1.94, pages:853-865

[13] D. Vani V, B. Bharathi, Hate Speech and Offensive Content Identification in Multiple Languages using machine learning algorithms , Forum for Information Retrieval Evaluation, CEUR,2022

[14] K. Sreelakshmi , B. Premjith and K. P. Soman: Detection of Hate Speech Text in Hindi-English Code-mixed Data, Procedia Computer Science, Trivandrum, 2020.

[15] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini and A. K. Jaiswal, Overview of the HASOC Subtrack at FIRE 2021 : Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, Forum for Information Retrieval Evaluation, December 13-17, 2021, India

[16] K. Ghosh , A. Senapati and U. Garain, Baseline BERT models for Conversational Hate Speech Detection in Code-mixed tweets utilizing Data Augmentation and Offensive Language Identification in Marathi, CEUR-WS.org/Vol-3395

[17] A. Velankar, H. Patil, R. Joshi, Mono vs multilingual bert for hate speech detection and text classification : A case study in Marathi, arXiv preprint arXiv:2204.08669(2022).

[18] K. Ghosh, D. Sonowal, A. Basumatary, B. Gogoi and A. Senapati: Transformer-Based Hate Speech Detection in Assamese, 2023 IEEE Guwahati Subsection Conference (GCON), doi: 10.1109/GCON58516.2023.10183497

[19] M. A. Bashar, R. Nayak, Qutnocturnal@hasoc'19: CNN for hate speech and offensive content identification in hindi language, CoRR abs/2008.12448 (2020). URL: https://arxiv.org/abs/2008.12448. arXiv:2008.12448.

[20] M. R. Karim, B. R. Chakravarthi, J. P. McCrae and M. Cochez, Classification benchmarks for under-resourced Bengali language based on multichannel convolutional-lstm network, 2020 IEEE 7th International Conference on Data Science and advanced Analytics(DSAA),IEEE,pp.390-399

[21] C. Paul and P. Bora, Detecting Hate Speech using Deep Learning Techniques,(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 12, No. 2, 2021

[22] M. Das, S. Banerjee, P. Saha, A. Mukherjee, Hate Speech and Offensive Language Detection in Bengali, AACL-IJCNLP 2022

[23] S. Ahammed, M. Rahman, H. M. Niloy and S. M. H. Chowdhury, Implementation of Machine Learning to Detect Hate Speech in Bangla Language, International Conference on System Modeling & Advancement in Research Trends, Moradabad, India, 2019.