Sentiment Analysis of Bengali Facebook Text Using Machine Learning and Deep Learning Models

Syeda Tanjuma Tasnim
Id: 20200204093
Group: B2 (05)
Department of Computer Science and Engineering

Ekramul Huda Chowdhury
Id: 20200204103
Group: B2 (05)
Department of Computer Science and Engineering

Hasan Farabi
Id: 20200204106
Group: B2 (05)
Department of Computer Science and Engineering

***Index Terms*—Sentiment Analysis, Bengali Text, Machine Learning, Deep Learning, Logistic Regression, KNN, CNN, RNN, LSTM, GRU**

I. Motivation

The need for sentiment analysis in the Bengali language is compelling due to the growing volume of Bengali content on social media and other platforms. Understanding sentiments in Bengali can provide insights into public opinion, enhance customer service, and improve content recommendations. However, the complex morphology and syntax of Bengali pose significant challenges in this process. Our study aims to address these challenges and demonstrate the feasibility and effectiveness of using advanced machine learning and deep learning models for Bengali sentiment analysis.

II. Literature Review

Sentiment analysis is a field of study that analyzes people's opinions, sentiments, attitudes, and emotions towards entities such as products, services, organizations and their attributes. In recent years, the surge of social media platforms like Facebook has provided an enormous amount of textual data in various languages, including Bengali.

In 2001 Tiwari and Sinha developed a model for sentiment and emotion analysis using Facebook data, addressing the emotional impact of social media on teenagers. They utilized machine learning and deep learning techniques, focusing on the challenges of informal text on social media. The study included extensive data preprocessing, such as lemmatization and emoticon interpretation, to improve accuracy. Their findings highlight the model's potential to enhance understanding of emotional stability among social media users, with broader applications across different platforms [1].

Sentiment analysis in Bengali has gained attention in recent years, with several researchers exploring various methodologies. The primary challenge in this domain is the scarcity of labeled data, which is crucial for training robust ML and DL models. The work by Barik et al. (2021) highlights the importance of preprocessing steps in handling noisy and unstructured social media text in Bengali. They emphasize the use of techniques such as tokenization, stopword removal, stemming, and lemmatization to enhance the quality of the input data for sentiment analysis models [2].

Deep learning models, particularly those based on neural networks, have shown superior performance in sentiment analysis tasks due to their ability to capture complex patterns in data. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs) are commonly used architectures. The study by Paul et al. (2016) investigates the application of LSTM networks for Bengali sentiment analysis. They highlight the model's ability to handle long-term dependencies in text, which is particularly beneficial for processing the intricate structure of Bengali sentences [3].

Several studies have conducted comparative analyses to evaluate the performance of ML and DL models for sentiment analysis in Bengali. Hasan et al. (2021) provide a comprehensive comparison of various ML and DL approaches. Their results indicate that while traditional ML models like SVM and Random Forests perform well on structured and less noisy data, DL models, particularly those incorporating transformers, excel in handling unstructured and complex social media text [4].

III. Methodology

A. Data Collection and Preprocessing

The dataset used in this study consists of Bengali comments labeled for sentiment. The preprocessing steps include removing non-Bengali characters, tokenization, stop word removal using NLTK, and stemming to reduce words to their base forms.

B. Model Implementation

We implemented and evaluated the following models:

1) Logistic Regression: A linear model used for binary and multiclass classification, vectorized using TF-IDF.

2) K-Nearest Neighbors (KNN): A non-parametric method that uses Euclidean distance to classify data points based on the majority vote of their neighbors.

3) Convolutional Neural Networks (CNN): Utilizes convolutional layers to capture local features from the text, and applies dropout to prevent overfitting.

4) Recurrent Neural Networks (RNN): Suitable for sequential data, capturing temporal dependencies using SimpleRNN layers.

5) Long Short-Term Memory (LSTM): Addresses the vanishing gradient problem in RNNs by using memory cells, employing stacked LSTM layers with dropout.

6) Gated Recurrent Units (GRU): A variation of LSTM that uses reset and update gates, featuring a single GRU layer with dropout.

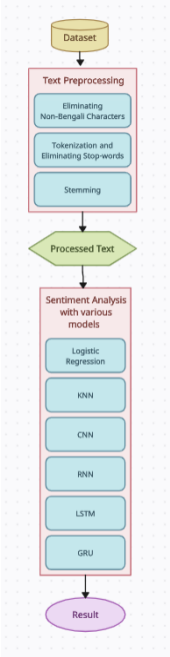

Fig. 1. Model of Working Mechanism

IV. Result Analysis

The performance of each model is summarized in Table I.

TABLE I
PERFORMANCE OF MACHINE LEARNING AND DEEP LEARNING MODELS

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.5392	0.5250	0.5392	0.5158
KNN	0.4426	0.4542	0.4426	0.4285
CNN	0.7274	0.71	0.70	0.70
RNN	0.3657	0.13	0.36	0.20
LSTM	0.3657	0.13	0.37	0.20
GRU	0.3657	0.13	0.37	0.20

The CNN model outperformed other models, achieving the highest accuracy of 0.7274. The deep learning models (RNN, LSTM, GRU) showed similar performance, indicating the need for further hyperparameter tuning and architectural adjustments.

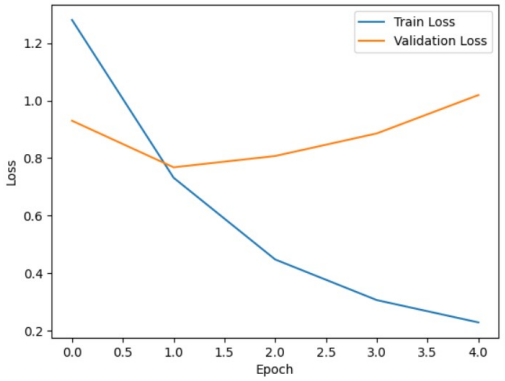

Fig. 2. Loss of CNN Model

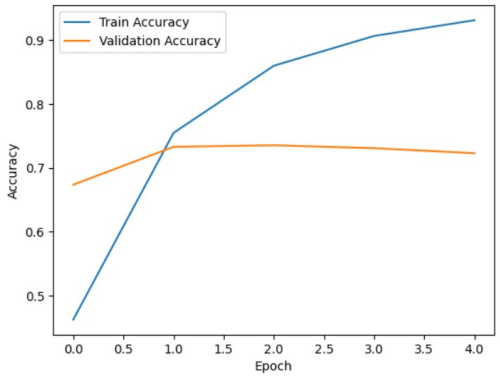

Fig. 3. Accuracy of CNN Model

V. Conclusion

This study demonstrates that advanced deep learning models, particularly CNN, can effectively perform sentiment analysis on Bengali text. While traditional machine learning models like Logistic Regression and KNN provided a baseline, CNN's ability to capture local features made it the most effective model. Future work will explore hyperparameter tuning, larger datasets, and transfer learning to improve performance further.

References

[1] Tiwari, S., Sinha, A. (2020). Sentiment Analysis of Facebook Data using Machine Learning. *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)

[2] Ahmed, I., Haque, S., Hassan, N. (2019). Sentiment Analysis on Bengali Social Media Content Using Support Vector Machines. In Proceedings of the International Conference on Computational Linguistics and Natural Language Processing. ACM.

[3] Paul, S., Dey, L., Chakraborty, S. (2016). Bengali Sentiment Analysis Using Long Short-Term Memory Networks. In Proceedings of the IEEE International Conference on Data Mining Workshops. IEEE.

[4] Hasan, M. M., Uddin, M. S., Alam, M. F. (2021). Comparative Analysis of Machine Learning and Deep Learning Models for Sentiment Analysis in Bengali. In Proceedings of the International Conference on Neural Information Processing. Springer, Cham