Bangla Sentiment Analysis

1st Md. Tahiadur Rahman
dept. of CSE
Ahsanullah University of
Science and Technology
Dhaka, Bangladesh
tahiad.dhk@gmail.com

2nd Adiba Amin
dept. of CSE
Ahsanullah University of
Science and Technology
Dhaka, Bangladesh
adiba4@gmail.com

3rd Md. Rafiu Alam Rafi
dept. of CSE
Ahsanullah University of
Science and Technology
Dhaka, Bangladesh
rafiu@gmail.com

***Abstract*—This study investigates the use of Convolutional Neural Networks (CNNs) for sentiment analysis in Bangla text, addressing linguistic challenges and the scarcity of resources in Bangla Natural Language Processing (NLP). The research involves comprehensive preprocessing of annotated Bangla text data, including the removal of non-Bangla elements, followed by tokenization and padding. The CNN model, designed with embedding, convolutional, and pooling layers, was trained and evaluated on a dataset, achieving a test accuracy of approximately 83%. While the results demonstrate the effectiveness of CNNs in capturing sentiment from Bangla text, the study acknowledges limitations such as data imbalance and potential overfitting, highlighting the need for further validation with larger datasets. This work contributes to the growing field of Bangla NLP and lays the groundwork for future sentiment analysis research.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Sentiment analysis in Bangla is underdeveloped due to linguistic complexities and a lack of resources. This study addresses these challenges by applying Convolutional Neural Networks (CNNs) to Bangla text for sentiment analysis. Using annotated data and extensive preprocessing, the CNN model achieved a test accuracy of approximately 83%. The research contributes to Bangla Natural Language Processing (NLP) by demonstrating the effectiveness of CNNs in this context. However, the study also notes limitations such as data imbalance and potential overfitting, highlighting the need for further validation with larger datasets to improve model robustness.

II. Motivation

Our research aims to fill the gap in Bangla sentiment analysis by leveraging Convolutional Neural Networks (CNNs). This initiative is critical for stakeholders in Bangla-speaking regions and addresses several key motivations:

- Unique Linguistic Challenges: Bangla's linguistic nuances require specialized sentiment analysis tools.
- Resource Scarcity: There is a lack of annotated datasets and NLP tools for Bangla, hindering research.
- Impact Potential: Accurate sentiment analysis can benefit businesses in customer feedback analysis and public opinion monitoring.
- Advancing Bangla NLP: Our study pioneers CNN-based models for Bangla NLP, fostering a vibrant research community.

III. Literature Review

Recent studies in Bangla NLP have demonstrated the effectiveness of these models. The shift from traditional machine learning to deep learning has led to significant improvements in Bangla sentiment analysis. CNNs, RNNs, LSTMs, and hybrid models have shown superior performance, despite challenges such as overfitting and hyperparameter tuning.
This study [1] explores the application of deep learning models for Bangla sentiment analysis, demonstrating the effectiveness of hybrid architectures combining CNNs and RNNs.
The paper [2] presents a deep learning-based sentiment analysis approach for both Bangla and Romanized Bangla text, showing the potential of neural networks in handling Bangla text sentiment analysis.
The study [3] focuses on using LSTM networks for Bangla sentiment analysis, emphasizing the effectiveness of RNN architectures in understanding contextual information.
This paper [4] investigates the use of CNNs with pre-trained word embeddings for Bangla sentiment analysis, demonstrating significant improvements in classification accuracy.

Methodology

The study utilized annotated Bangla text data sourced from an Excel file for sentiment analysis. Initial preprocessing involved cleaning the data by removing rows with missing values and irrelevant columns. The text was further refined through several preprocessing steps: emojis and special characters were removed using regular expressions, English words and alphanumeric characters were stripped out to focus exclusively on Bangla content, and punctuation marks were eliminated to simplify the text. The cleaned text was then tokenized into sequences of integers using the Keras Tokenizer, with sequences padded to a fixed length for uniform input size. The model architecture included a 128-dimensional embedding layer to convert integer sequences into dense vectors, followed by two 1D convolutional layers with 128 filters and a kernel size of 5, each activated by ReLU. Dimensionality was reduced using a MaxPooling1D layer and a GlobalMaxPooling1D layer. Fully connected layers comprised 128 units with ReLU activation, a Dropout layer with a 0.5 rate to prevent overfitting, and a final Dense layer with two units and a softmax activation function for binary classification. The data was split into 80% for training and 20% for testing, and the model was trained for

10 epochs with a batch size of 32, using 20% of the training data for validation. Performance metrics, including accuracy and loss, were monitored for both training and validation sets, with the model achieving a test accuracy of approximately 83%. Visualization of training and validation accuracy and loss was also performed to analyze model performance and detect potential overfitting.

IV. Limitations and Justifications

A. Limitations

- Data Imbalance: If the dataset is not balanced between positive and negative sentiments, it could bias the model.
- Overfitting: Despite dropout layers, the model might still overfit if the training set is not sufficiently large or diverse.

B. Justifications

- Choice of CNN: Convolutional Neural Networks are effective in capturing local patterns in text data, making them suitable for sentiment analysis tasks.
- Preprocessing Steps: Comprehensive text cleaning ensures that the model focuses on relevant features, enhancing its performance.
- Evaluation Metrics: Using both accuracy and loss provides a balanced view of the model's performance and helps in identifying overfitting. Limitations.

V. Result Analysis

The sentiment analysis model applied to the Bangla text dataset achieved significant results through a CNN-based architecture. The model demonstrated a test accuracy of approximately 83%, with training and validation accuracy metrics closely aligned, indicating minimal overfitting. Decreasing training and validation loss curves further suggest a well-behaved learning process and good generalization to unseen data. Although the confusion matrix, precision, and recall were not explicitly plotted, evaluating these metrics would provide additional confirmation of the model's robustness. High precision and recall values for both positive and negative sentiments would indicate balanced performance across sentiment classes. The study highlights that CNNs are effective in capturing local text patterns essential for Bangla sentiment analysis, with comprehensive preprocessing steps playing a key role in enhancing model performance. This research contributes to the broader field of Bangla NLP by demonstrating the potential of deep learning models. However, it also acknowledges potential challenges, such as data imbalance, which might affect real-world model performance, and the risk of overfitting, although this was mitigated by regularization and dropout layers. The study underscores the need for further validation with larger datasets to ensure the model's applicability in diverse real-world scenarios.

References

[1] Akter, S., & Azad, S. (2020). Deep Learning Approach for Sentiment Analysis in Bangla Text. Journal of Intelligent & Fuzzy Systems, 38(4), 4513-4521.

[2] Islam, M. S., Hasan, M. R., & Rahman, M. S. (2016). Sentiment Analysis on Bangla and Romanized Bangla Text Using Deep Learning. 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 831-836.

[3] Islam, M. A., Khan, M. R., & Das, P. (2020). Sentiment Analysis in Bangla Texts Using Long Short-Term Memory (LSTM). International Journal of Advanced Computer Science and Applications, 11(4), 342-347.

[4] Noor, M. S., Hossain, M. S., & Rahman, M. A. (2019). Sentiment Analysis for Bangla Texts Using Convolutional Neural Network with Pretrained Word Embeddings. Proceedings of the International Conference on Computer and Information Technology (ICCIT), 1-6.