2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December 2023, Cox's Bazar, Bangladesh

# Automatic Nostalgia Detection from Bengali Text

Ameya Debnath*, Bipul Karmokar†, and M. Shahidur Rahman‡
*Department of Computer Science & Engineering*
*Shahjalal University of Science and Technology*
Sylhet, Bangladesh
*ameyadebnath@gmail.com, †bipul27@student.sust.edu, ‡rahmanms@sust.edu

***Abstract*—The rise of user-generated content on digital platforms has led to various types of communication, including comments on social media, articles, and videos. In this context, detecting nostalgic feelings in Bengali comments is a unique challenge. This study focuses on creating a dataset on "Nostalgic Bengali comments" and will check which methods perform better to automatically spot nostalgic Bengali comments. We will use techniques from natural language processing, sentiment analysis, and machine learning to differentiate between comments that contain nostalgia and those that don't. Our dataset includes diverse Bengali comments collected from YouTube. Detecting nostalgic Bengali comments could help us understand user feelings, cultural discussions, and how people engage with content. This research contributes to the field of sentiment analysis by specifically addressing nostalgia in the Bengali language, enhancing our understanding of emotions in digital communication.**

***Index Terms*—Nostalgia Detection, Bengali Text Analysis, Bengali Comment Collection from YouTube, Deep Learning, Machine Learning ,**

## I. Introduction

The profound influence of nostalgia on human emotions and experiences has long been acknowledged across cultures and societies. Nostalgia, often defined as a bittersweet yearning for the past, carries immense significance in understanding the emotional resonance of various textual content. Its recognition and analysis can provide valuable insights into sentiment and emotional landscapes within the context of written communication.

Despite the extensive research on sentiment analysis and emotion detection in the field of Natural Language Processing (NLP), there remains a noteworthy gap concerning the automated detection of nostalgia in Bengali text. Nostalgia, a complex sentiment characterized by a mix of happiness and longing, presents a unique challenge for computational linguistics due to its nuanced nature.

In this pioneering study, we present a novel approach to address this gap by introducing the first-ever Bengali Nostalgia Detection dataset, a collection of 10,089 Bengali comments extracted from YouTube videos. Each comment in this dataset has been meticulously annotated as either nostalgic or non-nostalgic, providing a rich resource for training and evaluating machine learning (ML) and deep learning (DL) models.

Our research encompasses a wide array of traditional ML algorithms, including Logistic Regression, Support Vector Machines (SVM), Decision Trees (DT), Random Forest, and K-nearest Neighbors (KNN). These models serve as valuable benchmarks in evaluating the efficacy of more advanced DL techniques.

The DL models in our study encompass Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Bidirectional LSTMs (BiLSTM), Attention-based BiLSTMs, and BERT Nostalgia Classifier model. These deep learning architectures offer the capacity to capture intricate linguistic patterns and contextual information, essential for detecting nostalgia in Bengali text.

Through this research endeavor, we aim to not only bridge the existing gap in Bengali language sentiment analysis but also contribute to the broader field of nostalgia detection. Our findings are anticipated to offer new perspectives on the emotional underpinnings of Bengali textual content, thereby enriching the understanding of language, culture, and sentiment in the digital age.

To summarize, this paper represents a significant milestone as it pioneers the exploration of nostalgia detection in Bengali text, providing a robust dataset and a comprehensive evaluation of both traditional ML and state-of-the-art DL models. We believe that our work holds the potential to unlock novel applications in sentiment analysis, cultural studies, and human-computer interaction within the Bengali-speaking community.

## II. Related Works

Bengali is the sixth most commonly spoken native language globally. However, there hasn't been a lot of research carried out on this topic. Currently, numerous researchers are focusing their efforts on studying this language.

Abinash Tripathy [1] compared the results from two methods, Naive Bayes (NB) and SVM, to figure out if a sentiment review is positive or negative.

Tuhin et al [2] identified emotion in Bengali text by employing the Naive Bayes Algorithm along with the Topical Approach.

Chakraborty et al [3] collected 10,819 data points from Facebook to perform sentiment analysis. They used both classical algorithms and Deep Learning algorithms. From classical algorithms, they obtained the best result from Random forest. And LSTM gives better accuracy among Deep Learning algorithms.

979-8-3503-5901-5/23/$31.00 ©2023 IEEE

Menke et al [6] employed a hybrid research methodology to investigate how populists leverage nostalgia within their communication strategies and the persuasive influence that nostalgic rhetoric holds in garnering support for their assertions

Frischlich et al [4] used the Supervised Learning Approach to identify Nostalgia. They used 4,022 Facebook posts.

Clever et all [5] collected data from the questionnaire. They totally collected 285 questionnaires. For data representation and feature extraction, they mainly followed the bag-of-word(BOW) approach. For their dataset, Multinomial Naive Bayes worked better than any other approach.

We got inspiration from Clever et all [5].

## III. Data Collection and Preprocessing

As there was no established Bengali dataset on Nostalgia detection during our research. Thus we set out to create our own dataset.

We began scraping YouTube comments using YouTube API. To get API access a project needs to be created in the Youtube Data API platform and this will provide the API key and other necessary authentication. Next, we had to specify our search parameters to get appropriate Bengali videos that make people nostalgic. Keywords like "Old Bangla Song" and specific old Bangla song names were included and then the video IDs needed for our task were gathered. After that, we send an API request to fetch the comments associated with the video IDs that were collected previously. The comments along with their metadata will be returned by the API.

The comments were preprocessed and transformed into embedded tensors before they were used. The collected comments were not clean as they contained English comments, Hindi comments, Arabic comments, Special characters, different HTML tags, short comments, and other irrelevant things. So we programmatically handled this situation and these irrelevant contents were removed. Also, we found some comments that had no relation to our work or the associated video. We removed it manually by checking all the comments. We removed all comments that have less than 3 words because they don't make any proper sense. Although there can be some exceptions, for most cases they don't make any sense.

After cleaning the data we got 10089 comments related to the topic.

TABLE I
Overview of comment classes

| | |
|---|---|
| **Non-nostalgic Comments** | 7461 |
| **Nostalgic Comments** | 2628 |
| **Total** | 10089 |

We performed EDA on our dataset. We have seen that the maximum character length of a comment is 1909 and the minimum is 11. According to Fig 1, we can say most comments' lengths lie within the range of 11 to 250.

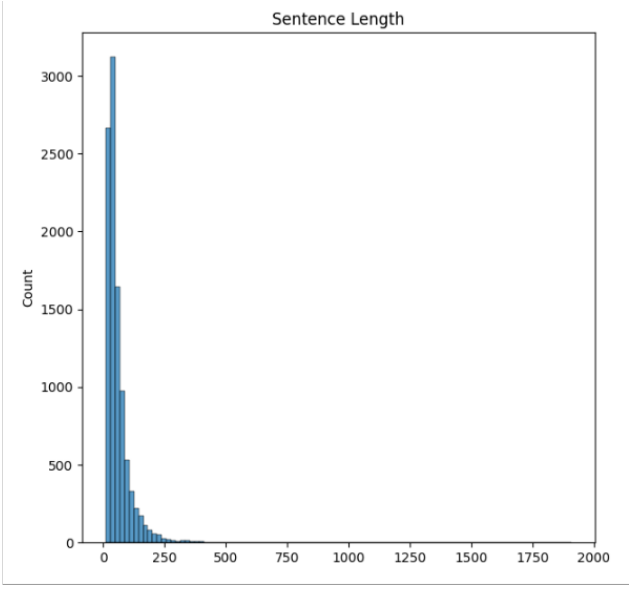


Fig. 1. Comments length

An organized process was used to annotate YouTube comments gathered from old Bengali songs. We manually annotated all the comments. Two annotators have annotated comments, and later merged them. We sat together and gave it the appropriate label if there was any conflict. We labeled "1" if the sentence expresses a nostalgic reaction and "0" if it is just a normal sentence. Table II shows several examples of annotated Comments in our dataset.

TABLE II
Sample of Nostalgia Bengali comments Dataset

| Comments | Nostalgic |
|---|---|
| কৈশোরের স্মৃতি মনে করিয়ে দিচ্ছে। | 1 |
| একদম ঠিক বলেছেন সঙ্গীত আমাদের মনে গভীর ছাপ ফেলে। | 0 |
| যখন একদম ছোট্ট ছিলাম বড়দের মুখে ও ঠাম্মার বড় রেডিওতে গানগুল শুনতে শুনতে কবে যেন গানগুলকে মনের মধ্যে গেঁথে ফেলেছি। | 1 |
| মন প্রাণ ভরে গেল ফিরে গেলাম ছোটবেলায়। | 1 |
| চ্যানেলের নাম কি বকবক | 0 |

## IV. Baseline Models

### A. *Classical Approaches*

We used 5 classical Algorithms on our dataset. These are :

- **Logistic Regression:** As logistic regression uses probability $0 < x < 1$ it is used in producing output values between 0 and 1. That is why it can be used in the stance detection task.
- **KNN:** In the realm of natural language processing (NLP), particularly in tasks like stance detection, a direct and effective machine learning technique known as K-Nearest Neighbors (KNN). This method operates on classification tasks and aims to forecast the outcome of a

target variable for a novel data point by assessing the attributes of its k-closest neighbors within the feature space.

- **Support Vector Machines (SVM):** The Support Vector Machine (SVM) is a potent method of machine learning that has come into use in a number of areas, including stance detection and natural language processing (NLP). SVM provides a useful method for categorizing text into various stances in the context of stance identification based on expressed attitudes and opinions.
- **Decision Tree:** Decision trees are useful for classification problems, such as stance detection, since they offer a structured manner to make decisions based on the characteristics. It is actually a divide-and-conquer approach for classification.
- **Random Forest:** Random forest is a method used to categorize groups, where instead of making a single classification, it creates a classification group. This group is then employed to classify new data points by leveraging the predictions made by the classification. It gives better performance than the Decision Tree. That's why we will use it in our classification task.

### *B. Deep Learning Approaches*

We used 5 Deep Learning Algorithms on our dataset. These are

- **CNN:** Convolutional Neural Networks (CNNs) are well-suited for classification tasks due to their innate ability to learn features autonomously. Particularly effective in scenarios involving intricate data like images, they identify patterns via convolutional and pooling layers. These acquired characteristics subsequently facilitate precise classification through fully connected layers. The efficiency with which CNNs extract and handle vital information solidifies their role as a robust option for tasks involving classification.
- **LSTM:** Long Short-Term Memory (LSTM) is valuable for text classification due to its capability to grasp context from sequential data. It retains information about words further back in a sentence, which aids in understanding nuanced meanings. LSTMs are adept at capturing dependencies within text, making them suitable for tasks where word order matters. By considering the context and sequence of words, LSTMs excel at tasks like sentiment analysis and language translation, enhancing the accuracy of text classification tasks without requiring extensive manual feature engineering.
- **BiLSTM:** Bidirectional Long Short-Term Memory (BiLSTM) is beneficial for text classification because it considers both past and future words in a sentence. This enables it to capture context and nuances effectively. BiLSTM is particularly useful for tasks where understanding word order is crucial, as it comprehends dependencies in both directions. It enhances tasks like sentiment analysis and language translation by leveraging the full context of the text. Utilizing BiLSTM can improve text classification accuracy without extensive manual feature work.
- **Attention-based BiLSTM:** The attention-based Bidirectional Long Short-Term Memory (BiLSTM) is a neural network architecture extensively used in natural language processing. By combining Bidirectional LSTMs and attention mechanisms, it excels at capturing contextual information from text data, making it invaluable for language-related research and applications. It consists of an Embedding Layer for vectorizing input sequences, a Bidirectional LSTM Layer with 64 units for comprehensive sequence processing, and an Attention Layer that enhances the model's focus on relevant information in the input data.
- **BERT Nostalgia Classifier:** In the context of this research endeavor, we have devised a binary text classification model, under the nomenclature 'BERT-Nostalgia-Classifier'. We have implemented the 'BERT-Nostalgia-Classifier' model leveraging the 'bert-base-uncased' architecture. We have maintained the confidentiality of the precise internal layer configuration and hyperparameter values. The model's architecture features a customized classification head, housing a single dense layer designed for binary prediction. It has undergone rigorous standard procedures in the realm of machine learning, encompassing data tokenization and training processes."

## V. Experimental Setup

### *A. Implementation Details*

**Data Pre-processing** : For all ML models, cleaning and vectorization are performed on our dataset. All models are trained in conventional processes. For the DL models, preprocessing is done as per model requirements.

- In **CNN** there is a Keras Tokenizer for text preprocessing, followed by padding to a 100 length. It consists of an Embedding layer, Conv1D layer, MaxPooling1D layer, Dropout layers, and a GlobalMaxPooling1D layer, concluding with an output layer for the binary classification.
- In this **LSTM-based** model there is a fixed vector size of 300 for word embeddings. It comprises an Embedding layer that's non-trainable, followed by a single LSTM layer with 200 units and a 20% dropout rate. The output layer has a sigmoid activation for binary classification. The model is compiled with the Adam optimizer, binary cross-entropy loss, and accuracy as the evaluation metric.
- In this **Bidirectional-LSTM** model with 200-dimensional embeddings incorporates a non-trainable Embedding layer. The Bidirectional LSTM layer has 200 units with 40% dropout on input and recurrent connections. The output layer uses sigmoid activation for binary classification.
- In **Attention-based BiLSTM** there is a tokenizer with a vocabulary size of 20,000 words, tokenizes and pads text sequences to a fixed length of 3,000 tokens. The neural network architecture consists of an Embedding layer, a Bidirectional LSTM layer with 64 units, and an attention

layer. The output layer employs a sigmoid activation for binary classification.

- **BERT Nostalgia Classifier** uses the "bert-base-uncased" model to create a binary classification neural network. Its learning rate is set to 2e-5 and uses binary cross-entropy as the loss function.

### B. *Evaluation Metrics*

To evaluate the performance of all the ML and DL models' we used the Accuracy and macro avg F1 scores. Our dataset isn't balanced that's why we counted the F1 score. We have run all the models several times to validate their performance. We have seen that the outcome remains the same.

TABLE III
ML MODELS PERFORMANCE TABLE

| ML Model | Accuracy | macro avg F1-score |
|---|---|---|
| Logistic Regression | 0.81 | 0.72 |
| KNN | 0.76 | 0.59 |
| SVM | 0.80 | 0.70 |
| Decision Tree | 0.75 | 0.67 |
| Random Forest | 0.80 | 0.70 |

TABLE IV
ML MODELS PERFORMANCE TABLE

| DL Model | Accuracy | F1-score |
|---|---|---|
| CNN | 0.73935 | 0.4251 |
| LSTM | 0.73934 | 0.4251 |
| BiLSTM | 0.78691 | 0.4251 |
| Attention Based BiLSTM | 0.73736 | 0.4251 |
| BERT Nostalgia Classifier | 0.81417 | 0.76 |

## VI. RESULTS AND DISCUSSION :

From Table III we can see the Accuracy and macro avg F1-score of every ML model. According to accuracy, we can say Logistic Regression performed better than all other ML models on our dataset whose accuracy is 81%. After Logistic Regression, the second position holds both SVM and Random Forest. Their accuracy is 80%. KNN holds the 3rd position with an accuracy of 76%. The worst performance we got from Decision Tree which has 75% accuracy. However according to the macro avg. F1-score Decision Tree holds the 3rd position with a score of 0.67. And KNN holds the last position.

Also according to the ROC curve in Fig 2, L.R. performs better than all other ML models. It also indicates that the verdict of Accuracy and F-1 score is correct.

It's clear from Table IV that our BERT Nostalgia Classifier is the best-performing model among all DL models on our dataset. Both Accuracy and F-1 scores agree on that. Accuracy is 81.417% and macro avg F1-score is 0.76. According to Accuracy 2nd, 3rd, 4th, and 5th place holders are BiLSTM, CNN, LSTM, and Attention-Based BiLSTM. Their accuracies are 78.691%, 73.935%, 73.934% and 73.736%. But if we take a look at the F1 score of the models, we see that the F1 score of all 4 models except the BERT Nostalgia Classifier is the same and that is 0.4251.

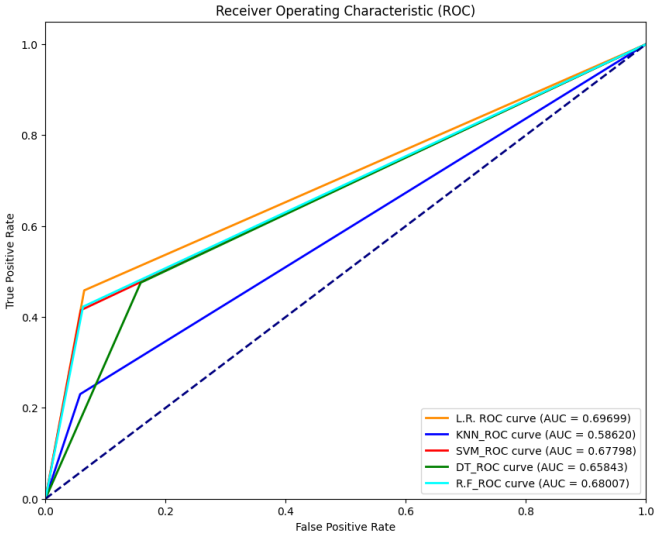


Fig. 2. ROC curve of ML models

## VII. CONCLUSIONS AND FUTURE WORK

In this work, we have constructed a dataset named "Nostalgic Bengali comments". We have run multiple baseline models and the BERT Nostalgia Classifier model to evaluate our dataset. They gave decent results. For further research on Nostalgia Detection, this dataset can be used. We hope this will be a stepping stone for future researchers in this domain. Our work has some business scope. One of them is, If we are able to determine what people actually want, then for sure there will be a huge business scope by providing that kind of service. In the future, we will enlarge our dataset. We also wish to explore other transfer learning and self-supervised approaches.

## REFERENCES

[1] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentimental reviews using machine learning techniques," in *Procedia Computer Science*, vol. 57, pp. 821-829, 2015.

[2] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, and A. K. Das, "An automated system of sentiment analysis from Bangla text using supervised learning techniques," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pp. 360-364, 2019, IEEE.

[3] P. Chakraborty, F. Nawar, and H. A. Chowdhury, "Sentiment analysis of Bengali Facebook data using classical and deep learning approaches," in *Innovation in Electrical Power Engineering, Communication, and Computing Technology: Proceedings of Second IEPCCT 2021*, pp. 209-218, 2022, Springer.

[4] L. Frischlich, L. Clever, T. Wulf, T. Wildschut, and C. Sedikides, "Populists' Use of Nostalgia: A Supervised Machine Learning Approach," in Proc. IEEE International Conference on Communications (ICC), 2022, pp. 1-24.

[5] L. Clever, L. Frischlich, H. Trautmann, and C. Grimme, ”Automated Detection of Nostalgic Text in the Context of Societal Pessimism,” in *Multidisciplinary International Symposium on Disinformation in Open Online Media*, pp. 48-58, 2019, Springer.
[6] M. Menke and T. Wulf, ”The dark side of inspirational pasts: An investigation of nostalgia in right-wing populist communication,” IEEE International Conference on Communications (ICC), 2021, pp. 237-249.