Bangla Social Media Text for Improved Emotion Detection

I. Introduction

In recent years, the field of natural language processing (NLP) and machine learning has witnessed significant advancements, enabling researchers and developers to explore various applications related to text analysis. [**?**] Emotion detection from text, a subfield of NLP, has gained substantial attention due to its potential to extract sentiments and emotions expressed in written content. The ability to automatically detect emotions from text has applications in areas such as social media sentiment analysis, customer reviews analysis, mental health assessment, and more. [**?**] [**?**]

The complexity of Bengali text and the limited resources available in comparison to English make emotion detection in this language a challenge. Establishing reliable Bengali emotion detection models can improve our knowledge of human communications and sentiment analysis across multiple languages. [**?**]

The motivation to investigate Bengali emotion detection is due to the linguistic diversity, the cultural significance of Bengali, the technological implications, the limited resources available, and the practical applications of emotion detection algorithms. The linguistic complexities of different languages enhance our comprehension of intercultural communication. The cultural importance of Bengali makes it an attractive area for emotion detection research, and emotion detection algorithms have a wide range of applications, such as market trend analysis and assessment of mental health. Despite the limited resources, the exploration of Bengali emotion detection may open the door to future research and tools.

Bangla is one of the most widely spoken languages in the world. However, due to its complexity, it is classified as a Grammatical and Verbal Expression and Vocals. Researchers are working on Bangla take analysis such as cyberbullying detection and emotion detection. However, there is very few research sir that can give a Conventions result and accuracy. [**?**]

In this paper our's aim to enhance data preprocessing, investigate algorithm suitability, develop and fine-tune models, evaluate performance, achieve fine-grained emotion classification, and offer insights into challenges and opportunities in Bengali emotion detection. And make enhanced preprocessing pipeline, algorithm suitability assessment, model development and fine-tuning, and comprehensive performance evaluation using Logistic Regression(LR), and Support Vector Machine(SVM). These aspects advance Bengali text emotion detection using machine learning, offering practical insights for real-world applications and enriching the understanding of multilingual sentiment analysis.

Identify applicable funding agency here. If none, delete this.

II. Literature Review

In [**?**] using a Naive Bayes classifier Azmin et al present a method for **detecting emotions from Bangla text corpus**. Their dataset consists of 1,500 comments from Facebook users, which is very samll and have been manually annotated with one of three emotion labels: happy, sad, or angry. The pre-processing steps applied to the dataset include removing stop words, stemming, and lemmatization. Three feature selection techniques are evaluated: term frequency-inverse document frequency (TF-IDF), n-grams, and a combination of TF-IDF and n-grams. The best results are obtained using a combination of TF-IDF and n-grams, with an accuracy of 78.6% based on a single fold of cross-validation.

In [**?**] Alam et al proposed **a novel framework for sentiment analysis of Bangla sentences using convolutional neural networks (CNNs)**.The framework consists of four steps including Preprocessing, Word embedding, CNN, and Classification. The authors evaluated the performance of their framework on a dataset of 1,000 Bangla sentences.The framework achieved an accuracy of 99.87%, which is the highest accuracy reported in the literature for sentiment analysis of Bangla sentences.Also, the framework is simple and efficient, and it can be easily adapted to other languages.

In [**?**] the paper surveys the current state of the art in emotion detection from text, with a focus on lexicon-based approaches. Rabeya et al propose a new **lexicon-based approach for**

detecting emotion from Bengali text, which they evaluate on a dataset of Facebook statuses. Their approach achieves an accuracy of 77.16%. They then focus on lexicon-based approaches, which they argue are the most effective for detecting emotion from text. They discuss the different types of lexicons that can be used for emotion detection, as well as the different ways in which these lexicons can be used.

In [**?**] About a new dataset containing Bangla documents with annotation of three emotions- Happy, Sad and Angry Sadia Afrin Purba et al used two major feature extraction techniques - Bag of Words(BoW) and Word Embedding is used to extract features from the documents. BoW is used by Logistic Regression and Multinomial Naive Bayes classifiers. Word Embedding is used by Artificial Neural Network(ANN) and Convolutional Neural Network(CNN) classifiers. Among all, Multinomial Naive Bayes classifier has given the best performance on the test set and the accuracy is 68.27%.

In [**?**] on Bangla Microblog Posts Shaika Chowdhury et al present a new method for performing sentiment analysis on Bangla text. The authors evaluate their method on a single classifier and test set of 300 tweets and achieve an accuracy of 75%. The authors first collect a dataset of Bangla tweets and then preprocess the tweets by removing special tokens, normalizing the text, and performing POS tagging. They then construct a Bangla sentiment lexicon by exploiting existing English lexical resources. Finally, they use SVM and MaxEnt classifiers to classify the sentiment of the tweets.It constructs a Bangla sentiment lexicon, which can be used by other researchers for sentiment analysis.

In [**?**] The paper "Emotion Detection Based on Bangladeshi People's Social Media Response on COVID-19" by Md. Rumman Hussain Khan Rahib et al. [**?**] investigate the use of social media to detect the emotions of Bangladeshi people in response to the COVID-19 pandemic. It is one of the first studies to use social media to detect the emotions of people in Bangladesh. The authors collected a dataset of Facebook posts and YouTube comments, and used a variety of machine-learning methods to analyze the data. LSTM models are typically more complex and require more training data than other machine learning models They found that the LSTM model was the most effective in detecting emotions.

In [**?**] Abdhullah-Al-Mamun et al. collected a dataset of Bangla text from social media platforms and labelled it as either bullied or not bullied. They then used different machine learning algorithms to train and test their models. Using support vector machines (SVMs) achieved the best performance, with an accuracy of 97%. This suggests that SVMs are a promising approach for cyber bullying detection on Bangla text.

In [**?**] the paper "BanglaSenti: A Dataset of Bangla Words for Sentiment Analysis" by Hasmot Ali et al.presents a dataset of Bangla words with sentiment polarity and labels. The dataset was created by collecting words from a variety of sources, including social media, news articles, and books. The authors then manually labeled each word as positive, negative, or neutral. The authors argue that their dataset is more comprehensive than existing datasets of Bangla words for sentiment analysis. They also provide a statistical analysis of the dataset, which shows that it is well-balanced in terms of positive, negative, and neutral words.The model achieved an accuracy of 85%.

In [**?**] Rashedul Amin Tuhin et al introduce the paper on sentiment analysis techniques for Bangla using supervised learning. Comparing Naïve Bayes and Topical Approaches on sentence and article levels, the study demonstrates the latter's superiority with over 90% accuracy. The contribution lies in bridging the gap in Bangla sentiment analysis, highlighting machine learning's effectiveness over rule-based methods. Acknowledging data challenges, the paper suggests future exploration of unsupervised methods. This work not only advances Bangla sentiment analysis but also provides insights for other languages with limited resources.

Tuhin et al. [**?**] proposed An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques.The main goal is Find out The expected data from random data through the data set include lots of missing val-ues and noisy data, and it is clear Here why tra-ditional algorithms like Na ıve Bayes fail to extract information.The research paper utilizes supervised learning techniques, specifically Na¨ıve Bayes Classification Algorithm and the Topical approach, to extract emotions from Bangla text. These methods are applied at different levels of scope and are compared to determine their performance in sentiment analysis.Best accuracy attained by SVM (93%).

Gupta et al. [**?**] proposed Toward Integrated CNN-based Sentiment Analysis of Tweets for Scarce-resource Language—Hindi.This article first explores the machine learning-based approaches—Na¨ıve Bayes, Support Vector Machine, Decision-Tree, and Logistic Regression—to analyze the sentiment contained in Hindi language text derived from Twitter.The proposed approach in the paper is an integrated convolutional neural network (CNN)-Recurrent Neural Network and Long Short-term Memory (RNN-LSTM) for sentiment analysis of Hindi language tweets.The final result of the sentiment analysis on Hindi language tweets using the proposed integrated CNN-RNN-LSTM approach was an accuracy of 85%.

FA Acheampong et al. [**?**] proposed a survey on Text-based emotion detection: Advances, challenges, and opportunities.The objective is to discuss open issues and future research directions for text-based ED.language representation, classification, and contextual information extraction.In the language representation phase, the extraction of contextual information is crucial for improving classification accuracies. This involves techniques such as rule-based approaches, keyword recognition (KR),and lexical affinity methods. The

classification phase involves the use of machine learning techniques, including K-NN, MLP, Decision Tree, SVM, and NB, to detect and classify emotionsin textual data.

Saad Bin Ahmed et el. [**?**] proposed a Sentence Continuation Inference of Urdu Text by BERT Technique. The study improves a sentence continuation inference model for Urdu text using pre-trained BERT models like Multilingual-BERT and Arabic-BERT. The model splits sentences into segments using tokens and padding tokens. The researchers used a 53k character synthetic URL dataset for good results, demonstrating the effectiveness of these models in recognizing cursive Urdu text.

MD. RAJIB HOSSAIN et el [**?**]. proposed a Authorship Classification in a Resource Constraint Language Using Convolutional Neural Networks. The paper presents a method for authorship classification in resource-limited languages like Bengali using Convolutional Networks. It evaluates 90 embedding models using Word2Vec, GloVe, and FastText, selecting 9 best performing models. The optimized CNN with GloVe achieves the highest accuracy in benchmark datasets BAAD-18, BAAD16, and LD, with a classification accuracy of 93.45%, 95.02%, and 98.67%.

ASHFIA JANNAT KEYA et el. [**?**] proposed a G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media. The research introduces the G-BERT model, a new method for detecting hate speech on Bengali social media. This model, which combines BERT architecture and GRU model, achieves high accuracy, precision, recall, and F1-score, making it effective in reducing online hate speech in Bengali. The model outperforms other classification algorithms, achieving 95.56% accuracy, 95.07% precision, 93.63% recall, and 92.15% F1-score, respectively.

Sara Azmin et al. [**?**] proposed a system for Emotion Detection from Bangla Text Corpus Using Naïve Bayes Classifier. This paper seeks to explore the detection of various types of emotions in Banglayan text using a Natural Language Processor (NB) classifier, as well as other features such as Stemmer, POS Tagger, Ngrams, TF-IdF, and more. This paper will focus on the current state of emotion detection in Bangladesh, which is yet to be fully explored, as well as the analysis of human emotions from Banglayan text data. The proposed method of emotion detection from Banglayan text was found to be accurate in classifying emotions into three categories of happy, sad, and angry, with an accuracy of 78.6.

III. Goal of the Study

The goal of this study is to improve Bengali text emotion detection by improving preprocessing techniques, evaluating algorithms, developing models, and measuring performance. The preprocessing techniques are improved to make the input data cleaner. The machine learning algorithms are tested to see if they're suitable for detecting Bengali emotions. Then, emotion detection models are developed and tested to see how well they work. The model performance is evaluated to see how well it can classify the text. The goal of the study is to go beyond just traditional sentiment analysis and make it possible to classify Bengali emotions in a more detailed way. The goal is to help people understand sentiment analysis better in multilingual settings, so they can use it to detect emotions in Bengali.

IV. Methodology

This chapter presents the methodology employed in achieving the research objectives outlined in the previous chapter. The study comprises several key phases, each contributing to the development and evaluation of emotion detection models for Bengali text.

A. Data Collection

The initial step consists of assembling a broad collection of Bengali texts containing various emotive expressions. This data is sourced from various sources, such as Kaggle [**?**] social media, customer feedback, and academic literature. The data is then subjected to a thorough preprocessing process to improve its quality. The preprocessing process removes noise, symbols, emoticons, and other spaces from the text using tools provided by the NLTK library and other related Python packages. The input data is then purified and optimized for further analysis.

B. Dataset Analysis

From these multiple source, we have amassed a total of 10001 data related to people's emotions from various perspectives on daily occurrences. The data we have collected is not classified as data; it contains unnecessary spaces symbol strings or emoticons. Additionally, some of the comments are too small to be taken into account. Therefore, prior to commencing our work, we must work with the data set to make it more efficient and redundant for a better outcome.

C. Data Pre-processing

Data preprocessing is essential for analyzing and training your data. In this section, we'll cover the steps you need to take to make sure your data is ready for analysis and training. We'll cover things like removing stopwords, getting rid of unnecessary symbols, making sure punctuation is in place, and getting rid of stemming. We'll also look at how to get rid of stopwords that don't have any meaningful content. We'll use a set of Bengal stopwords to make sure our analysis is as good as it can be. We'll also focus on words that are rich in content, which are important for spotting emotions. Finally, we'll get rid of any unnecessary symbols like emojis or characters that don't have much meaning, so your data is clean and clear. Punctuation is important for grammar, but it is often not relevant for emotion analysis. All punctuation marks are extracted methodically from the text, keeping the core semantic content intact and allowing for feature extraction. Stemming truncates words to their basic form, standardizes vocabulary and improves model efficiency. Aligning words with linguistic roots concentrates the model's attention on the

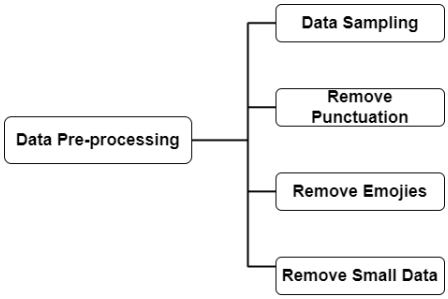

Fig. 1.

Data Pre-processing Stage

substance of text. Data summarization uses frequency distribution analysis to categorize Bengali words based on occurrence, extracting unique terms that show significant variations across emotion classes. This approach improves model discrimination by highlighting emotionally relevant terms. In short, data preprocessing fine-tunes raw data into a structured format that is optimized for the development of emotion detection models. Noise reduction, simplification and emphasis on emotionally relevant components enhance the accuracy and effectiveness of emotion detection in Bengal text.

D. Feature Extraction

In order to construct a reliable machine learning model, it is essential to extract essential features from the dataset. The following feature extraction techniques are employed in this work:

Count Occurrence: Count the number of times a word occurs as a feature. We have used this method because the important words occur over and over again. The number of occurrences indicates how important the word is, and more occurrences means more important features.

Term Frequency - Inverse Document Frequency(TF-IDF):TF-IDF is a numerical representation that quantifies a term's importance in a document relative to a corpus. The formula for calculating TF-IDF for a term t in a document d is given by:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$$

Here, TF denotes the Term Frequency of term t in document d, and IDF represents the Inverse Document Frequency of term t in the corpus.

E. Model

In our paper, model selection turns out to be one of the most important stages in our methodology. It plays an important role because it involves the identification and implementation of machine learning algorithms specially designed for the complex task of detecting emotions in Bengali texts. Because Bengali text is full of complex and nuanced expressions, the selection of algorithms has a significant impact on the accuracy and effectiveness of our emotions detection efforts. In order to do this thoroughly, we perform a thorough evaluation. We look at algorithms such as multinomial naive Bayes (MBA), support vector machines (SVM), logistic regression (logistic regression), and K-means cluster (K-means cluster). We look at a number of important metrics such as accuracy, precision (Q), recall (RQ), and F1 (F1-score). By looking at these metrics, we will be able to identify the algorithm which not only has superior overall performance, but also has an improved ability to detect the unique range of emotions that are intricately woven in Bengali text content. These algorithms are then implemented, refined, and integrated in the model development stage, leading to the development of highly optimized emotion detection models specifically designed for the nuances of Bengali language.

F. Our Methodology

Here is an overall summary of how our research works:

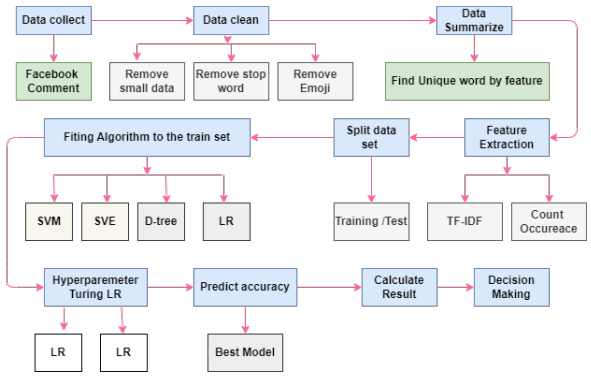

Fig. 2.

Our Proposed Methodology

G. Workflow of Our Methodology

Here is an overall summary of how our research works:

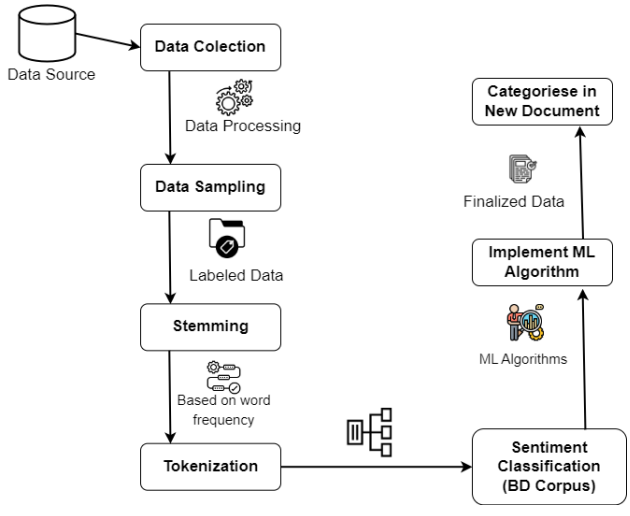

Fig. 3.

Workflow of Symentic Analysis

As previously mentioned, we work with text data to analyze Bangla emotions. Web collect cheater from different data

source like Kaggle Facebook comment YouTube comment Twitter tweets and so on and we have collected a total number of 10,000 data. We collected the data from different social media platforms so the data was not labeled, There were many unnecessary things like extra spaces images a string, and extra small sentence or even their null values.

First, we sample the data. In this state, Will labels the unlabeled data and the data from the unlived data as if there are no null values. For data sampling, we have used the over-sampling method here to create a perfect data set.

After sampling, We start streaming data. in this state with classified data on the basis of the grammar and use the base form of the word and after that, we go through the tokenization phase. In this state, we distribute data according to their frequency count and use the numeric value afterward implementation. Here this too face steaming and Tokenization is under feature extraction.

After this, we have classified our data according to our national learning algorithm. We have applied Logistic regression, support vector machine, decision tree, and other machine learning algorithms and finalized the output.

REFERENCES

[1] Deng, L. and Liu, Y., 2018. A joint introduction to natural language processing and to deep learning. Deep learning in natural language processing, pp.1-22.

[2] Ekman, P., 1992. An argument for basic emotions. Cognition emotion, 6(3-4), pp.169-200.

[3] Lind, F., Eberl, J.M., Galyga, S., Heidenreich, T., Boomgaarden, H.G., Jiménez, B.H. and Berganza, R., 2019. A bridge over the language gap: Topic modelling for text analyses across languages for country comparative research. University of Vienna: Working Paper of the REMINDER-Project.

[4] Ghosh, A., Das, A. and Bandyopadhyay, S., 2010, August. Clause identification and classification in bengali. In Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (pp. 17-25).

[5] Avraamidou, L., 2020. Science identity as a landscape of becoming: Rethinking recognition and emotions through an intersectionality lens. Cultural Studies of Science Education, 15(2), pp.323-345.

[6] Hardeniya, N., Perkins, J., Chopra, D., Joshi, N. and Mathur, I., 2016. Natural language processing: python and NLTK. Packt Publishing Ltd.

[7] Bonny, J.J., Haque, N.J., Ulla, M.R., Kanungoe, P., Ome, Z.H. and Junaid, M.I.H., 2022, April. Deep learning approach for sentimental analysis of hotel review on bengali text. In 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) (pp. 1-7). IEEE.

[8] Naithani, K. and Raiwani, Y.P., 2023. Realization of natural language processing and machine learning approaches for text-based sentiment analysis. Expert Systems, 40(5), p.e13114.

[9] Azmin, S. and Dhar, K., 2019, December. Emotion detection from bangla text corpus using naive bayes classifier. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-5). IEEE.

[10] Alam, M.H., Rahoman, M.M. and Azad, M.A.K., 2017, December. Sentiment analysis for Bangla sentences using convolutional neural network. In 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[11] Rabeya, T., Ferdous, S., Ali, H.S. and Chakraborty, N.R., 2017, December. A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from Bengali text. In 2017 20th international conference of computer and information technology (ICCIT) (pp. 1-7). IEEE.

[12] Purba, S.A., Tasnim, S., Jabin, M., Hossen, T. and Hasan, M.K., 2021, February. Document level emotion detection from bangla text using machine learning techniques. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp. 406-411). IEEE.

[13] Chowdhury, S. and Chowdhury, W., 2014, May. Performing sentiment analysis in Bangla microblog posts. In 2014 International Conference on Informatics, Electronics Vision (ICIEV) (pp. 1-6). IEEE.

[14] Rahib, M.R.H.K., Tamim, A.H., Tahmeed, M.Z. and Hossain, M.J., 2022. Emotion detection based on Bangladeshi people's social media response on COVID-19. SN Computer Science, 3(2), p.180.

[15] Akhter, S., 2018, December. Social media bullying detection using machine learning on Bangla text. In 2018 10th International Conference on Electrical and Computer Engineering (ICECE) (pp. 385-388). IEEE.

[16] Ali, H., Hossain, M.F., Shuvo, S.B. and Al Marouf, A., 2020, July. Banglasenti: A dataset of bangla words for sentiment analysis. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-4). IEEE.

[17] Tuhin, R.A., Paul, B.K., Nawrine, F., Akter, M. and Das, A.K., 2019, February. 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). IEEE

[18] Tuhin, R.A., Paul, B.K., Nawrine, F., Akter, M. and Das, A.K., 2019, February. 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). IEEE.

[19] Gupta, V., Jain, N., Shubham, S., Madan, A., Chaudhary, A. and Xin, Q., 2021. Toward integrated CNN-based sentiment analysis of tweets for scarce-resource language—Hindi. Transactions on Asian and Low-Resource Language Information Processing, 20(5), pp.1-23.

[20] Acheampong, F.A., Wenyu, C. and Nunoo-Mensah, H., 2020. Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7), p.e12189.

[21] Ahmed, S.B., 2022. Sentence Continuation Inference of Urdu Text by BERT Technique. Available at SSRN 4144163.

[22] Hossain, M.R., Hoque, M.M., Dewan, M.A.A., Siddique, N., Islam, M.N. and Sarker, I.H., 2021. Authorship classification in a resource constraint language using convolutional neural networks. IEEE Access, 9, pp.100319-100338.

[23] ASHFIA JANNAT KEYA et. Ai. proposed a G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media.

[24] Azmin, S. and Dhar, K., 2019, December. Emotion detection from bangla text corpus using naive bayes classifier. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-5). IEEE.

[25] Tuhin, R.A., Paul, B.K., Nawrine, F., Akter, M. and Das, A.K., 2019, February. 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). IEEE.