Analytical Study of Transliterated Bengali-English Social Media Comments using Machine Learning Models

Tonmoy Banik
Department of CSE
Ahsanullah University of Science Technology
Dhaka, Bangladesh
tonmoybtonoy2@gmail.com

Sanjida Akter
Department of CSE
Ahsanullah University of Science Technology
Dhaka, Bangladesh
190204076@aust.edu

Piyal Datta
Department of CSE
Ahsanullah University of Science Technology
Dhaka, Bangladesh
190204080@aust.edu

Dipanwita Bala
Department of CSE
Ahsanullah University of Science Technology
Dhaka, Bangladesh
190204087@aust.edu

***Abstract*—Sentiment analysis is helpful to understand how people feel about certain words. Understanding sentiment for computers is a tricky task because of its different nature, especially when it is written in Banglish. In this paper, a "Banglish" balanced dataset was used. We used four different machine learning models, Logistic Regression, Multinomial Naive Bayes, Support Vector Machine, and K-Nearest Neighbour to classify if a sentence is slang or sexually abusive or other and made a comparative analysis. The accuracy of all models came almost same though differences in Precision, Recall, and F1-Score are seen.**

I. Introduction

Due to the widespread use of social media platforms like Facebook, Instagram, and Twitter, more people than ever are interacting online. Large volumes of textual data are produced as a result, creating challenging NLP issues. Much research is now being done on automatically recognising a wide range of vocal expressions, such as aggression, sarcasm, hate, and irony. NLP specialists are also interested in figuring out how to deduce a person's emotions from their messages. To be specific, analyzing the reactions by users accumulated from social media contents and posts lead to categorize them into several labels i.e. sad, angry, love, sexual etc. [1]

In this research, we analyzed Bengali texts which are written in English from Mendaly Data using Banglish (Bengali in English letter) hate speech detection. In classical approaches, we used Logistic Regression, Support Vector Machine, Multinomial Naive Bayes and K-Nearest Neighbour.

II. Related work

In this section, we have focused on reviewing some techniques related to sentiment analysis. Over time, various machine learning and deep learning approaches are used for classification.

In the paper [2], two approaches are used for sentiment analysis using NLP, SVM, Naive Bayes and hybrid models. The maximum accuracy for the used architecture is 96.8%. An analytical study of Singh [3] shows 68% of maximum accuracy for logistic regression among used models(SVM, KNN, Decision Tree, Logistic Regression and Naive Bayes) where MTf-IDF, one hot and count vectorizer were used for tokenization.

In a comprehensive analysis of machine learning techniques applied With Doc2vec for Classifying Sentiment of Bengali Natural Language [4], various algorithms were assessed, with Bi-Directional Long Short-Term Memory (BiLSTM) emerging as the top performer, achieving 77.85% accuracy, accompanied by precision, recall, and F-1 scores of 78.06%, 77.39%, and 77.72% respectively. The work of Rezaul and Naimul [1] introduces a novel CLSTM architecture, achieving a significant improvement in multi-class sentiment classification on Bengali social media comments with an accuracy of 85.8%.

An investigation targeting depression detection from Bengali text [5] implemented RNN with LSTM architecture, leveraging Adam optimizer and achieved an approximate 98% accuracy. In the work of Monjoor and Faruk [6], similar kind off model were used to perform Comparative Study for Sentiment Analysis which explores sentiment analysis in Bengali, revealing higher accuracy for raw Bengali text (83%) compared to translated text (65.3%). A research study focusing on sentiment analysis of Bengali Facebook data [7], with deep learning models (LSTM, CNN) outperforming classical approaches. The deep learning approach performed better than the traditional approach, with an accuracy of 96.95% for LSTM where Support Vector Machine and Random Forest classifier achieved the accuracy of 78.23% and 78.37%, respectively.

In the research [8], they have showed analysis using BERTBSA, a multi-lingual BERT model, achieving competitive accuracy for two-class (71%) and three-class (60%) sentiment classification in Bengali and demonstrating effectiveness in analyzing public comments from an daily online newspaper. Another work using multilingual BERT and XLM-RoBERTa model [9] shows 95% accuracy for two-class sentiment analysis for Bengali language. In [10], HAN-LSTM, D-CAPSNET-Bi-LSTM and BERT-LSTM model achieved accuracy values of 78.52%, 80.82%, and 84.18% respectively.

III. Dataset Property

The dataset used for this paper is a multi class dataset where class number is three. The three classes are sexual, slang and others. It contains 1675 short comments of three categories which are collected from different social medias [11]. The ratio of data from three categories are quite similar in this dataset.

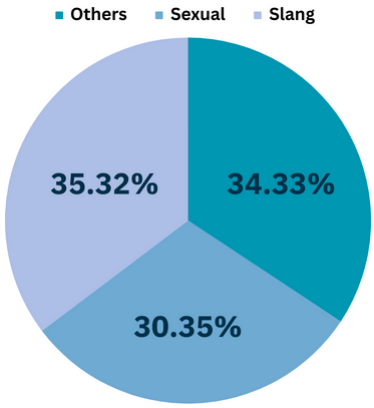

Fig. 1. Different Class Ratio

The used dataset consists of 510 sexual texts, 575 slang texts and 590 others texts.

TABLE I
Banglish Balanced Dataset

Text	Class
Potitar Cheye o oke beshi ghrina kore	Sexual
Botkalab Jao porimoni kanki	Sexual
Eshb Bokachondro re niye bangla tribune ekhn news kore chi	Slang
Wait amare tag marailo kun chudir vai!!!	Slang
Ajke match here bangladesh final a uthe gelo	Others
Bangladesher player gulare gorom dim therapy dawa hok	Others

IV. Methodology

Our proposed methods for comment classification are described in this section. This paper proposes K Nearest Neighbour, Support Vector Machine and Logistic Regression machine learning models for comment classification.

A. Dataset Pre-processing

We pre-processed the dataset so that it can be viably utilized for machine learning models. This step expels any non-alphanumeric characters from the content, but for white spaces. This step helps clean the text by getting rid of symbols, punctuation, etc. Now, the pre-processed contents are used for further analysis or feature extraction.

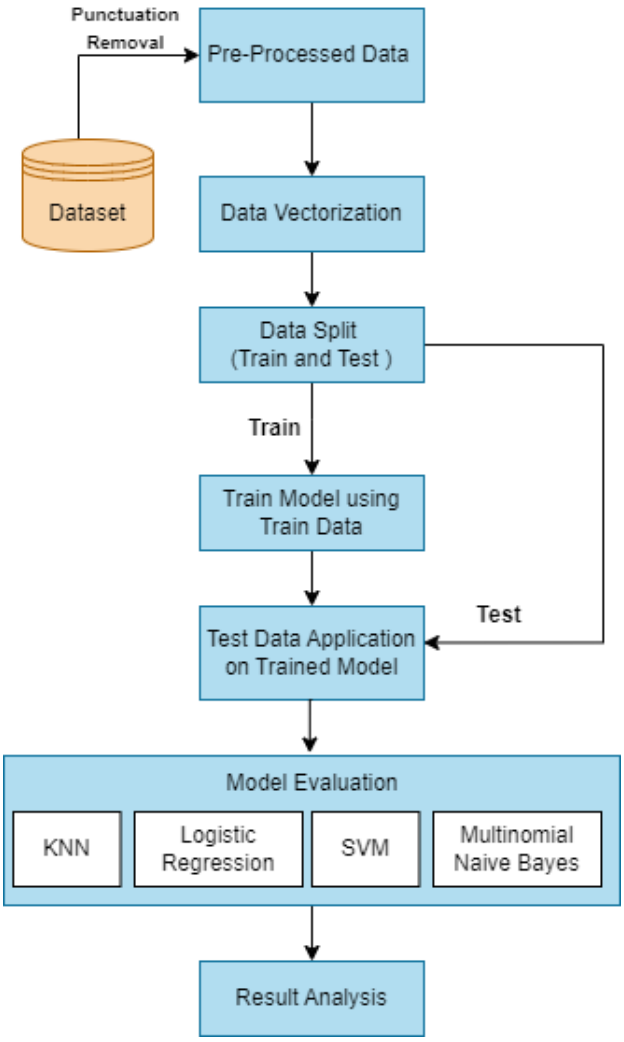

Fig. 2. Workflow Diagram

B. Vectorization

The pre-processed text data is converted into a suitable numeric vectors format for machine learning models. The pre-processing is done using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique. This technique assigns numerical values to words based on their frequency in each document and across the entire dataset. Raw documents are converted to a matrix of TF-IDF features using TfidfVectorizer. A document is represented in each row and a unique word in the entire dataset is shown in each column.

C. Model Architecture

The model architecture involves the utilization of three different machine learning algorithms: Logistic Regression, KNN, MNB, and SVM.

1) Logistic Regression: It is a mathematical data analysis technique which identifies relationships between two factors and uses this insight to predict the value of one factor based on the other, often resulting in finite outcomes like ”yes” or ”no”. This technique can be regularized to prevent over-fitting, thereby enhancing its performance on unseen data.

2) K-Nearest Neighbour (KNN): KNN algorithm, a non-parametric, supervised learning classifier widely employed for classification and regression tasks. It predicts the classification of a new sample point based on its proximity to existing data points within distinct classes.

3) Support Vector Machine (SVM): A supervised learning algorithm adept at solving intricate classification, regression, and outlier detection problems by determining optimal data transformations that establish boundaries between data points according to predefined classes or labels. SVMs excel in binary classification scenarios, where the task involves categorizing elements into two distinct groups.

4) Multinomial Naive Bayes (MNB): The term “multinomial” is used to denote the type of data distribution which is assumed by the model. The features in text classification are typically word counts (term frequencies). This distribution is used to estimate the likelihood of seeing a specific set of word counts in a document.

V. Experimental Results

The performance of four classification models (MNB, KNN, SVM, Logistic Regression) are focused on this section. These three classifiers were used on data obtained after data cleaning and Tf-Idf Vectorizer was used as the vectorization technique.The outcomes of our experiments are presented in Table II and Table III

TABLE II
Experimental Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**62.09%**	**62.54%**	**62.08%**	**61.88%**
KNN	**57.91%**	**62.54%**	**62.08%**	**61.87%**
SVM	**61.79%**	**62.54%**	**62.08%**	**61.88%**
MNB	**62.38%**	**62.01%**	**63.50%**	**62.0%**

Among the models, 62.38% test accuracy is highest achieved by MNB. MNB model has performed well from the training data by showing 96.40% training accuracy. SVM, KNN and Logistic Regression shows 61.79%, 57.91% and 62.09% of testing accuracy respectively.

TABLE III
Experimental Results (Translated)

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**50.14%**	**52.14%**	**50.15%**	**49.10%**
KNN	**30.15%**	**52.14%**	**50.15%**	**49.15%**
SVM	**49.26%**	**52.12%**	**50.15%**	**49.14%**
MNB	**51.04%**	**54.0%**	**59.0%**	**57.0%**

After translation from Banglish to Bangla, the accuracy and other scores reduced by a certain percentage.

VI. Result Analysis

This section focuses on the analysis of the results obtained from the experiments on our dataset. Evaluation Matrix is used to compare the performance of four models which are Logistic Regression, KNN, SVM, and MNB.

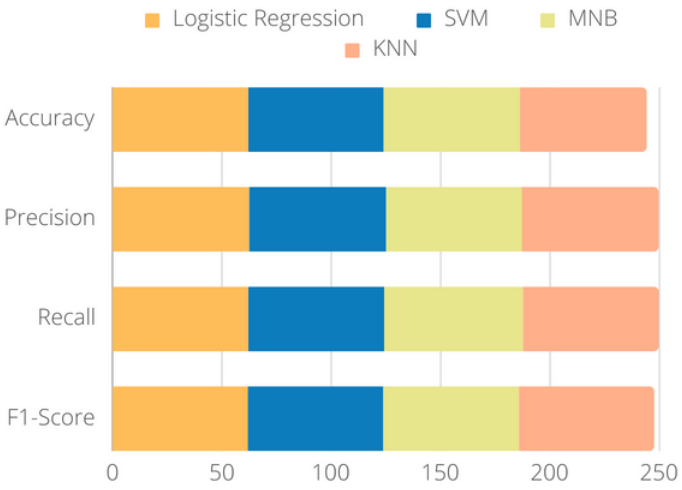

Fig. 3. Evaluation Scores

During the training phase SVM gives the best accuracy which is almost 97% and KNN gives lower accuracy which is about 75% among the models.

In the testing phase, SVM, MNB and logistic regression perform almost the same. They give about 62% and almost the same precision, recall, and f1-score. These results can be increased if we can manage a large amount of dataset. By the training and testing comparisons, we can see that MNB learnt our dataset very well and it gave the best results of all.

VII. Conclusion

In this paper, a balanced banglish dataset collected from [11] is used to detect sentiment (slang, sexual and others) using Machine Learning models. A comparative analysis is done on this dataset using Logistic Regression, Support Vector Machine, K-Nearest Neighbour, and Multinomial Naive Bayes.

After comparison, it is clear that Multinomial Naive Bayes performed better than other three models for our dataset. Deep Learning Models can be a good choice along with Machine Learning Models. In future, deep learning models will be implemented to analyze the sentiments.

References

[1] R. Haque, N. Islam, M. Tasneem, and A. K. Das, “Multi-class sentiment classification on bengali social media comments using machine learning,” *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 21–35, 2023.

[2] V. Thakur, R. Sahu, and S. Omer, “Current state of hinglish text sentiment analysis,” *SSRN Electronic Journal*, 01 2020.

[3] G. Singh, “Sentiment analysis of code-mixed social media text (hinglish),” 02 2021.

[4] M. T. Hoque, A. Islam, E. Ahmed, K. A. Mamun, and M. N. Huda, "Analyzing performance of different machine learning approaches with doc2vec for classifying sentiment of bengali natural language," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–5, IEEE, 2019.

[5] M. Rafidul Hasan Khan, U. S. Afroz, A. K. M. Masum, S. Abujar, and S. A. Hossain, "A deep learning approach to detect depression from bengali text," in *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 2*, pp. 777–785, Springer, 2021.

[6] S. N. Monjoor, O. Faruk, K. M. Mahmudul Haque, F. B. Iqbal, M. A. Mitu, M. A. Islam, and M. S. H. Mukta, "A comparative study for sentiment analysis of raw and translated text," in *The 2nd International Conference on Distributed Sensing and Intelligent Systems (ICDSIS 2021)*, vol. 2021, pp. 220–231, 2021.

[7] 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, Springer, 2022.

[8] K. I. Islam, M. S. Islam, and M. R. Amin, "Sentiment analysis in bengali via transfer learning using multi-lingual bert," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5, IEEE, 2020.

[9] A. Bhowmick and A. Jana, "Sentiment analysis for bengali using transformer based models," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pp. 481–486, 2021.

[10] N. R. Bhowmik, M. Arifuzzaman, and M. R. H. Mondal, "Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms," *Array*, vol. 13, p. 100123, 2022.

[11] Faisal Bin Ashraf, "Banglish (bengali in english letter) hate speech dataset," 2022.