I.J. Information Engineering and Electronic Business, **2023, 6, 48-63**
Published Online on December 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2023.06.04

Binary vs. Multiclass Sentiment Classification for Bangla E-commerce Product Reviews: A Comparative Analysis of Machine Learning Models

Shakib Sadat Shanto
Department of Computer Science, American International University Bangladesh, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
Email: shakibsss080@gmail.com
ORCID iD: https://orcid.org/0009-0009-8798-9010

Zishan Ahmed
Department of Computer Science, American International University Bangladesh, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
Email: zishanahmed599@gmail.com
ORCID iD: https://orcid.org/0009-0004-9598-917X

Nisma Hossain
Department of Computer Science, American International University Bangladesh, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
Email: nisma.hossain.41982@gmail.com
ORCID iD: https://orcid.org/0009-0001-7633-9559

Auditi Roy
Department of Computer Science, American International University Bangladesh, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
Email: royauditi808@gmail.com
ORCID iD: https://orcid.org/0009-0002-8009-3990

Akinul Islam Jony*
Department of Computer Science, American International University Bangladesh, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
Email: akinul@aiub.edu
ORCID iD: https://orcid.org/0000-0002-2942-6780
*Corresponding Author

Received: 23 March, 2023; Revised: 18 May, 2023; Accepted: 17 July, 2023; Published: 08 December, 2023

Abstract: Sentiment analysis, the process of determining the emotional tone of a text, is essential for comprehending user opinions and preferences. Unfortunately, the majority of research on sentiment analysis has focused on reviews written in English, leaving a void in the study of reviews written in other languages. This research focuses on the understudied topic of sentiment analysis of Bangla-language product reviews. The objective of this study is to compare the performance of machine learning models for binary and multiclass sentiment classification in the Bangla language in order to gain a deeper understanding of user sentiments regarding e-commerce product reviews. Creating a dataset of approximately one thousand Bangla product reviews from the e-commerce website 'Daraz', we classified sentiments using a variety of machine learning algorithms and natural language processing (NLP) feature extraction techniques such as TF-IDF, Count Vectorizer with N-gram methods. The overall performance of machine learning models for multiclass sentiment classification was lower than binary class sentiment classification. In multiclass sentiment classification, Logistic Regression with bigram count vectorizer achieved the maximum accuracy of 82.64%, while Random Forest with unigram TF-IDF vectorizer achieved the highest accuracy of 94.44%. Our proposed system outperforms previous multiclass sentiment classification techniques by a fine margin.

Index Terms: Sentiment analysis, E-Commerce, Binary classification, Multiclass classification, Natural language processing, Feature extraction, Accuracy.

1. Introduction

Sentiment analysis is the method of understanding a text's emotional tone, such as whether it is positive or negative [1]. Multiclass sentiment analysis implies categorizing text into more than two categories, such as very positive, positive, neutral, negative, and very negative [2]. The seventh most extensively spoken language in the world is Bangla, which is also the official language of Bangladesh and the Indian states of West Bengal, Tripura, and Assam. So, need for automated systems to determine the sentiment of Bangla text is increasing. Many studies on sentiment analysis and opinion mining have been conducted, and after using NLP approaches, different sentiments were categorized using machine learning models [3]. Despite extensive study on sentiment analysis in English, there is little research on the topic in Bangla. Lack of annotated data, the difficulty of the Bangla language, and the requirement for a thorough understanding of the cultural background all pose obstacles to the development of a Bangla sentiment analysis system [4]. Applications for a powerful Bangla multiclass sentiment analysis system include tracking public opinion on social media and evaluating user comments on various e-commerce sites like 'Daraz', 'Bikroy', 'Evaly', and Chaldal.com. As the number of online e-commerce platforms in Bangladesh grows fast, millions of consumers provide numerous Bangla product reviews for a variety of products. It is challenging to compare and classify binary and multiclass sentiments since there are insufficient balanced annotated datasets in Bangla for E-commerce product reviews. There is an obvious need for more study in this field given the significance of sentiment analysis in comprehending human behavior and decision-making as well as the rising need for automated sentiment analysis systems in Bangla.

The primary objective of this work is to create and test a machine learning-based technique for sentiment analysis in Bangla text. The study's specific goal is to assess user sentiments across binary (positive, negative) and multiclass (very positive, positive, negative, very negative) classes. The purpose of this work is to address the lack of research and resources in the field of sentiment analysis that are especially designed for the Bangla language. The research aims to evaluate the effectiveness of machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest classifier, Linear SVM, RBF SVM, Multinomial Nave Bayes, KNN, and Stochastic Gradient Descent (SGD) in classifying Bangla product reviews into appropriate sentiment categories. The study objective also includes increasing the accuracy of existing multi-class sentiment categorization algorithms for Bangla language. This study attempts to determine the most effective way for sentiment analysis in the Bangla language by analyzing and comparing the performance of several machine learning models and vectorization approaches.

The outcome of our study is as follows:

- Comparison of various machine learning models for binary and multi-class sentiment classification on Bangla product reviews.
- Improved accuracy of existing research on multi-class Bangla text classification.

The work contributes to the creation of a pure Bangla annotated dataset on e-commerce product reviews and a strong Bangla sentiment analysis system by fulfilling these key research objectives. This may be used to follow public opinion on social media platforms and evaluate user comments on e-commerce sites in Bangladesh, where the number of online platforms and Bangla product reviews is quickly increasing. The study additionally acknowledges the need of balanced annotated datasets in Bangla for sentiment analysis emphasizing the importance of additional research in this area and difficulty in classifying multiclass sentiments for Bangla texts.

The following sections are included in the paper; Section 2 reviews important related work in detail. This section examines and synthesizes earlier research, exposing the intellectual landscape of the subject. Section 3 discusses the research dataset. This section describes the data sources, their features, and the dataset preprocessing steps. Section 4 discusses the study's methodology. This section describes the methods used to achieve research goals. Section 5 analyzes the results extensively. This section analyzes the findings, explaining their consequences and relevance. This section also contains a comparative analysis with prior studies and discussion of the results obtained. Section 6 concludes with a well-structured conclusion and a consideration of future research. This section summarizes the important results and their possible implications. This section also highlights knowledge gaps, recommending future research paths.

2. Literature Review

The biggest difficulty in performing Bangla text-based preprocessing is the lack of NLP tools for this language [4]. Bengali language study was done by researchers to overcome these limitations. Hasan *et al.* [5] used Amazon reviews as web data to generate an English corpus. After that, they converted the text into Bengali by using Google Translate. The dataset was divided into two categories: positive with high ratings (4, 5) and negative with low ratings (1, 2).

Reviews with a rating of 3, which are regarded as impartial, were removed. They used the Naive Bayes technique, which on the dataset had an accuracy of 85 %. Yet, unlike machine translation, which lacks human interactions, sentiment analysis heavily depends on the context and the language's structure.

Researchers are currently concentrating more on gathering data in native Bangla for machine learning for this reason. Using restaurant reviews as case study, Haque *et al*. [6] created a manual dataset without the use of a translator in order to establish a review analysis system for Bangla and Phonetic Bangla. The procedure starts with preprocessing the raw data, then feature extraction using a variety of N-gram techniques. The data is then vectorized using many types of vectorizers, including count vectorizers, hash vectorizers, and TF-IDF vectorizers. Eventually, a variety of machine learning-based methods were used to categorize evaluations into three categories: bad, good and excellent. Lastly, they compared the vectorizers for the various classifiers, and found that SVM offers greater accuracy 75.58 %. Three different categories of user sentiment were used by Akter *et al*. [7] for the implementation of sentiment analysis on Bengali opinions or reviews (positive, negative & neutral). They gathered evaluations from "Daraz," an online retailer, and used the TF-IDF vectorizing approach to extract features. They then used a variety of machine learning techniques, with KNN having the greatest accuracy 96 %. Ahmed *et al*. [8] used a combined dataset for their research. They categorized 12628 texts according to six feelings: happiness, sadness, anger, disgust, surprise, and fear. Three emotion detection methods for Bangla text were evaluated. Three different statistical methods—Logistic Regression, Multinomial Naive Bayes, and Multi-layer Perceptron—arrived to the same conclusion of six distinct emotional classes. Two blended datasets are used to compare the TF-IDF, count vectorizer, and their combination. LR with TF-IDF had the highest accuracy of 44 %. Mahtab *et al*. [9] created a dataset of three sentiment classes—Positive, Negative, and Neutral—based on the opinions of actual individuals on Bangladesh Cricket. There are 1601 records in the collection. In their method, they classified sentiments using the TF-IDF Vectorizer and SVM. Using their dataset, they suggested a model employing an SVM classifier, which had an accuracy of 64.59 %.

In their work, Chowdhury *et al*. [10] suggested a method of sentiment analysis for Bangla-language movie reviews. The social media websites' publicly accessible comments and posts served as the source of the dataset that was manually compiled and labelled for this experiment. There were around 4000 samples total, each with a positive or negative label. The model has an accuracy of 88.90 % on the test set when using the Support Vector Machine technique. Sharif *et al*. [11] suggested a mechanism that may divide customer evaluations primarily into two categories, positive and negative, based on an assessment of their emotional response. 1000 restaurant reviews in Bengali were used to evaluate the recommended method. For their suggested project, they employed the TF-IDF approach for feature extraction. According to the experimental results, the proposed system can categorize restaurant evaluations using multinomial Naive Bayes with an accuracy of 80.48 %. Shafin *et al*. [12] employed TF-IDF vectorizer for preprocessing while they examined customer reviews from a number of e-commerce websites from “Daraz”, “Bikroy”, “Evaly”, and “Chaldal”. Then, several classification methods were used, including KNN, Decision Tree, SVM, RF, and LR. SVM fared best with 88.81 % accuracy in this situation. Khatun and Rabeya [13] gathered 5500 user-generated Bengali reviews from several book review platforms on social media in total, after which they created a model that can assess user sentiment, classified as either positive or negative. Five distinct machine learning algorithms were used after the requisite data analysis and tokenization. The Random Forest among them has the highest accuracy, which is 98.39 %. M. Hassan *et al.*[14] gathered around 1,141 data with two classifications, Good and Negative. They used the TF-IDF approach for feature extraction. The most successful techniques for selecting the ideal classifier for their dataset was SVM with the accuracy of 85.59 %.

According to our background study, there aren't enough balanced annotated datasets in Bangla for E-commerce product reviews, making it difficult to compare and categorize binary and multiclass sentiments. So, the goal of our work is to create a dataset of e-commerce product reviews and measure performance of several machine learning models to compare user sentiment across binary and multiclass categories.

3. Dataset

For our study, we first collected the data, labeled it ourselves, preprocessed it, and then conducted an analysis of the preprocessed data. This is a description of each of the aforementioned steps.

The dataset consists of more than 1000 pure Bangla comments on multiple product types. All the comments were collected and annotated manually. There are 1011 total comments and 2 attributes or features; Comment and Sentiment.

3.1 Data Collection

The e-commerce industry in Bangladesh has grown tremendously in recent years. Each product receives hundreds of comments each day from users and buyers. Every review expresses a different set of feelings. ‘Daraz’ provides a wide range of products and reviews, thus we gathered trustworthy and comprehensible Bangla reviews from there. For our study, we have gathered more than 1000 comments in pure Bangla.

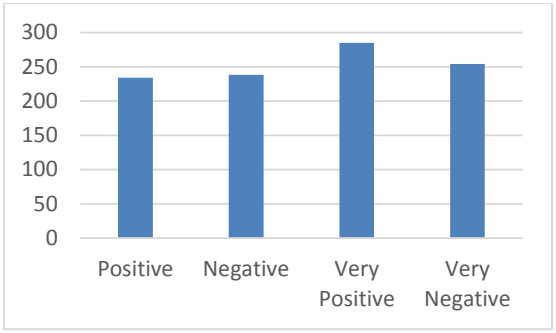

Fig. 1. Sentiment Category Distribution in Multiclass Classification Dataset

Fig 1 shows the initial dataset distribution for multiclass classification, very positive and very negative have the highest frequencies, with counts of 285 and 254, respectively. The sentiment categories of positive and negative have slightly lower frequencies, with counts of 234 and 238, respectively.

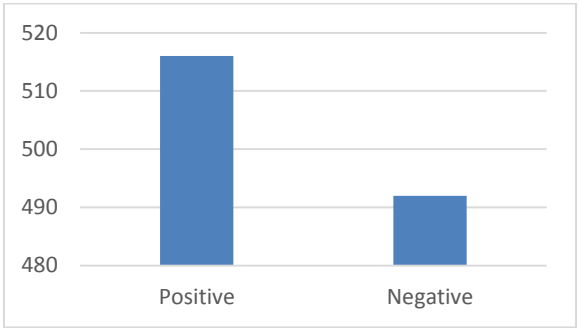

Fig. 2. Sentiment Category Distribution in Binary Classification Dataset

Fig 2 shows the initial dataset distribution for binary classification, there are 519 Positive reviews and 492 negative reviews. Here, positive reviews have higher frequency than negative reviews.

3.2 Data Labelling

For our dataset, we labeled the reviews with four types, which are: positive, very positive, negative and very negative according to their sentiments. According to the Table 1.

Table 1. Multiclass Labelling

Comment	Sentiment
অসাধারণ একটা আয়না। আমার মেয়ের খুব পছন্দ হয়েছে।।	Very Positive
ইয়ারফোন কোয়ালিটি ভালো	Positive
হেয়ারজেল ভালো ছিল না|	Negative
বালিশ গুলো নেতায় গেছে এক মাসেই। এতো বাজে কোয়ালিটি	Very Negative

For conducting binary classification, we merged all the very positive and positive comments to ‘Positive’ and all the very negative and negative comments to ‘Negative’. According to Table 2.

Table 2. Binary Labelling

Comment	Sentiment
শাড়ি ভালো হয়েছে সবাই নিতে পারেন	Positive
পাতলা সিঙ্গেল একটি কম্বল দিয়েছে ফালতু	Negative

3.3 Data Preprocessing

First, we removed any unnecessary punctuation, numbers, emoji, pictorial icons, and alphabets from collected comments, using a list of our own stop words. For both sentiment classification and product category classification, we employed separate stop-word lists. For better filtering, we sought to create our stop-word list using the words that were used the least frequently in the dataset. We next combine all the words into sentences using the stemming strategy. We

also removed minor reviews from the dataset. The preprocessed comments for sentiment classification are displayed in Table 3.

Table 3. Preprocessed Comment

Original Comment	Cleaned Comment
জুতা ভালই লাগল। সাইজ ঠিক দিয়েছে। কালার মোটামুটি। জুতার মুখ খুবই ছোট। পা ভিতরে ঢুকাতে এবং বের করতেখুব কষ্ট হয়। রাবার বা জিহবা কিছুই নেই। উপরের ফিতারও কোন কাজই নাই। হুদাই	জুতা সাইজ ঠিক দিয়েছে কালার খুবই ছোট নেই
কোয়ালিটি অনেক ভাল। পায়ে আরামদায়ক। তবে বাম পায়েরটা ডান পা থেকে সামান্য একটু বড়। তবে বেশি সমস্যা না।	কোয়ালিটি অনেক ভাল

3.4 Data Analysis

After preprocessing our dataset, we then analyzed the statistic of our data. Fig.3 and Fig.4 shows the distribution percentage for binary class sentiment analysis and multiclass sentiment analysis respectively. For each of the mentioned analytical activities, we additionally looked at the data statistics in Fig.5 and Fig.6. In order to determine which length of comment is most common in this context, we also looked at the length-frequency distribution in Fig.7 for sentiment analysis.

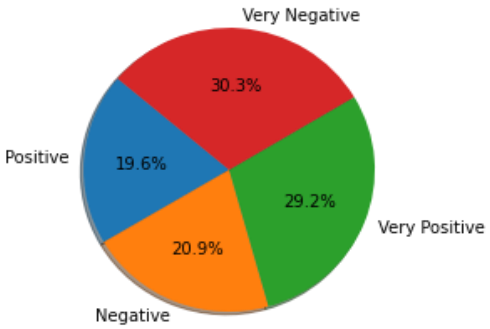

Fig. 3. Pie chart Representation of comment distribution for multiclass sentiment analysis

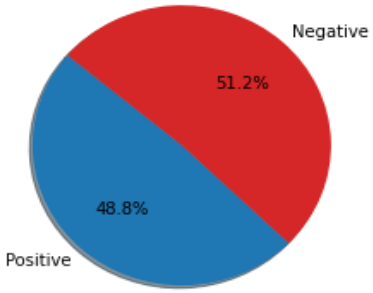

Fig. 4. Pie chart Representation of comment distribution for binary class sentiment analysis

Fig.3 shows the distribution of multiclass comments using a pie chart. It shows that 30.3% of comments were very negative, 29.2% were very positive, 19.6% were positive, and 20.9% were negative. Fig.4 represents the distribution of comments for binary class sentiment analysis. The chart reveals that 48.8% of the comments were positive and 51.2% were negative. The pie charts demonstrate that each of the categories or sections representing the various classes resemble being of equal size, indicating a fairly even distribution of sentiment classes in the dataset after preprocessing.

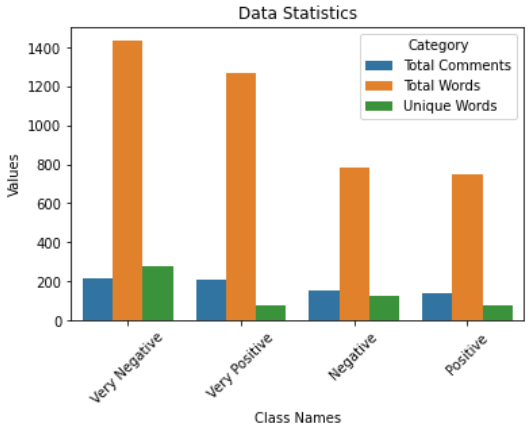

Fig. 5. Data statistics for multiclass sentiment classification

Fig. 6. Data statistics for binary class sentiment classification

Fig.5 and Fig.6 depict the data statistics for multiclass and binary sentiment classifications, respectively. It can be seen that the number of total words for multiclass very negative and very positive is greater than that of negative and positive classes. Both numbers of total words for the binary class are comparable. In both types of classifications, the number of unique words is less than the total number of words for each of the classes. Reduced lexical diversity can be useful when a corpus contains numerous misspelled or inept words. By having fewer unique words, it may be simpler for the models to recognize and deal with such variations in the text. With a lesser vocabulary, the dimensionality of the vector space representation can be decreased, resulting in more efficient computation and memory utilization. However, the number of unique words in the very positive and positive classes is lower than in the very negative and negative classes. In this case, a lack of lexical diversity can result in an increase in ambiguity, particularly if multiple words are assigned to an identical vector representation.

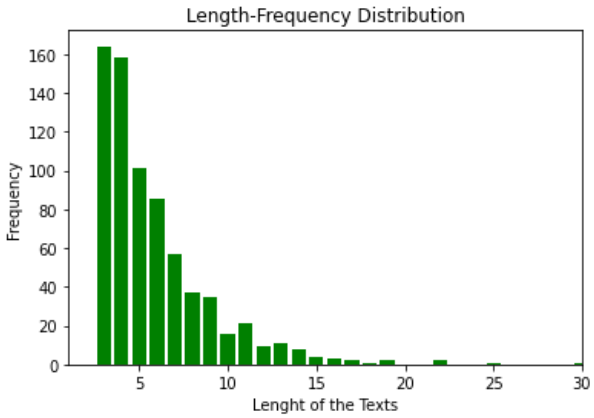

Fig. 7. Length-Frequency distribution for sentiment classification

Fig.7 depicts the distribution of comment lengths relative to number of comments. The majority of the comments range in length from two to nine words. There are significantly fewer lengthy comments in the dataset.

4. Research Methodology

In this work, sentiment analysis model for both binary and multiclass sentiment classification for Bangla-language e-commerce product reviews based on supervised machine learning is presented. The following procedures are used throughout research methodology and the complete structure of proposed system is shown in Fig.8.

- Feature Extraction
- Classifiers
- Evaluation measures

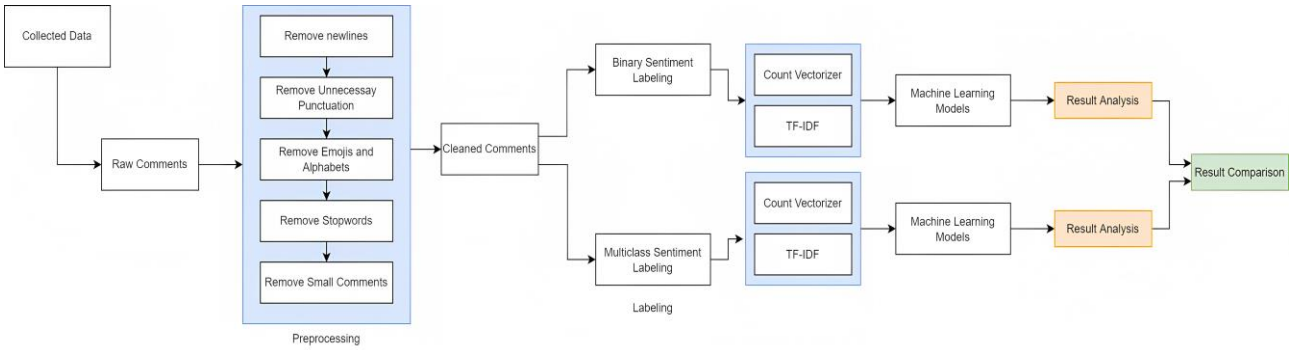

Fig.8. Proposed system diagram

4.1 Feature Extraction

Due to the fact that multiple dimensionalities require a lot of computational resources to analyze, we must extract the feature from our data. Data dimensionalities may be effectively reduced by feature extraction, which also gets our data ready for learning. With the help of the Count Vectorizer and TF-IDF features, we assessed our processed data. To get the highest possible result, we further integrated these approaches.

4.1.1 Count Vectorizer

A group of textual documents are converted into something resembling a matrix of word/token frequencies using the Count Vectorizer [8] method. In addition, textual data pre-processing is possible before creating matrix forms. It has a very flexible textual feature display architecture as a result. The fact that the essential terms occurred often led us to choose this approach. A word's frequency of use illustrates its significance, with a greater frequency signifying more important characteristic.

4.1.2 TF-IDF

TF-IDF [15] is a better approach than counting occurrences because it considers the possibility that high frequency may not provide meaningful information advantage. In other words, unusual terms could give the model more credence. TF-IDF algorithm is represented as,

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \tag{1}$$

In the TF-IDF equation (1) $W_{i,j}$, is a TF-IDF score. $tf_{i,j}$, represents the number of occurrences of i in j. The quantity of documents is represented by N. Where df_i is the number of documents that contain i.

We have finally configured the n-gram [16] value of (1, 2) to assess differently in TD-IDF and Count Vectorizer. The application of unigrams and bigrams to Bangla in our methodology is illustrated by Fig.9.

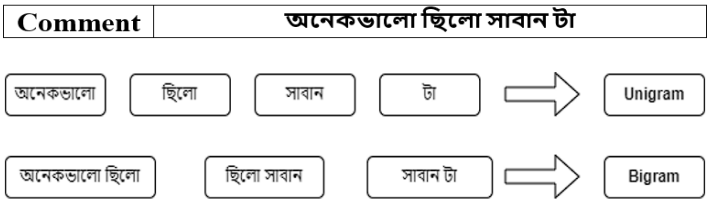

Fig. 9. Unigram and Bigram for proposed model

4.2 Classifiers

For classifying binary and multiclass sentiments and also multiple classes of products, we used several machine learning algorithms.

The Logistic Regression [17] model calculates the probability of the binary result as a function of the independent factors by fitting an S-shaped logistic curve that maps the independent variables to the likelihood of the binary outcome. The logistic regression is represented by the following equation:

$$y = \frac{e^{(a_0 + a_1 x)}}{1 + e^{(a_0 + a_1 x)}} \tag{2}$$

Where, x is input value, y is predicted output, a_0 is bias or intercept term and a_1 = coefficient for input (x)

Decision tree [18] is a tree-like model used for making decisions or predictions. It consists of nodes and branches that represent decisions and their outcomes, respectively. The general form of the equation for a decision tree can be represented as follows:

$$f(x) = C_k \text{ for x} \in R_k \tag{3}$$

Where f(x) is the predicted output (or decision) for a given input x. Rk represents the region of the input space that is assigned to the kth leaf node of the decision tree. Ck is the output (or decision) associated with the kth leaf node.

For training, Random Forest [19] builds several decision trees and delivers the mean (for regression) or the mode (for classification) of the individual tree results. We frequently utilize the Gini index or entropy while running Random Forests on classification data.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{4}$$

This algorithm calculates the Gini for each branch off a node, revealing which is more probable based on class and probability. Whereas c is the entire number of classes in the collection, pi denotes the frequency with which this class occurs.

$$Entropy = \sum_{i=1}^{C} -p_i \times \log_2 p_i \tag{5}$$

To determine which branch the node should follow, entropy evaluates the likelihood of a specific outcome.

Multinomial Naive Bayes [20] method assumes that the frequency of words in a text follows a multinomial distribution, meaning that the likelihood that a word will appear in a document is inversely proportional to the number of times the word occurs there. The Bayes theorem may be used to describe it, which states that all variables in a given class A are conditionally independent of one another given A. Bayes rule application to a remark sentiment (S) and class (A),

$$P(S \mid T) = \frac{P(S \mid A)P(A)}{P(S)} \tag{6}$$

$$\text{Final Equation} = \arg\max P(S_1, S_2, S_3, ..., S_n \mid A)P(A) \tag{7}$$

K Nearest Neighbor [21] is referred to as KNN. It is a supervised machine learning method that categorizes the fresh text by comparing it to the training data's closest matches to provide predictions.

A version of the SVM technique known as a Linear SVM [22] divides the data points of two classes using a linear hyperplane. A linear SVM allocates new data points to the class that is on the same side of the hyperplane as the training data after measuring the distance between each new data point and the hyperplane.

A variation of the support vector machine technique known as the radial basis function support vector machine, or RBF SVM, finds a hyperplane that may divide the two classes of data points by transforming the data into a higher-dimensional space using a non-linear kernel. The gamma parameter, which controls the breadth of the Gaussian distribution used to estimate the similarity between data points, is a parameter of the RBF kernel.

Stochastic gradient descent or SGD [23] is an iterative optimization algorithm commonly used in machine learning to minimize the loss function of a model during training. The learning rate is an important parameter in SGD, as it controls the size of the step taken in the direction of the negative gradient. The update rule for stochastic gradient descent (SGD) is as follows:

$$w(t+1) = w(t) - \alpha \times \nabla(w(t), X(i:t)) \tag{8}$$

Where w(t) is the weight vector at iteration t, α is the learning rate, and ∇(w(t), X(i:t)) is the gradient of the loss function with respect to the weights w(t), computed using a randomly selected mini-batch of training examples X(i:t) at iteration t.

4.3 Evaluation Measures

Machine learning model performance is measured using evaluation metrics. For the purpose of our study, we evaluated our model using accuracy, precision, recall, and F1 score.

A performance metric called accuracy [24] is used to assess how well a machine learning model is doing. Which is defined as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{9}$$

Precision [24] quantifies the ratio of true positives (positive occurrences that were accurately predicted) to the total number of cases that actually fall into the positive category. Which is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{10}$$

Recall [24] is the calculation of how many positive real portrayals by our criterion are marked as positive (true positive). Which is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{11}$$

F1-Score [24], which is the harmonic mean of accuracy and recall, is frequently employed as a single metric to assess the effectiveness of a model. The F1-score is defined as:

$$\text{F1- Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision+Recall}} \tag{12}$$

5. Result Analysis and Discussion

For each type of feature extraction approach, we have constructed and analyzed multiple machine learning methods using various machine learning evaluation metrics, including accuracy, precision, recall, and F-1 score. We used 80% of our data for training and 20% for testing.

5.1 Model Evaluation of Multiclass Sentiment Classificaion

The outcomes of our multiclass sentiment classification are displayed in Table 4. Here, the models were trained and evaluated for four distinct classes, as previously described.

Table 4. Multiclass Sentiment Classification Evaluation metrics

Feature Extraction	Machine Learning Models	Accuracy	Precision	Recall	F1-Score
Unigram TF-IDF	LR	79.17%	77.95	76.50	76.63
	Decision Tree	70.14%	70.32	69.58	67.99
	Random Forest	77.08%	75.93	73.48	74.35
	MNB	77.08%	77.18	76.79	75.23
	KNN	56.94%	59.12	57.31	55.43
	Linear SVM	70.83%	73.31	65.38	64.42
	RBF SVM	75.69%	74.95	72.40	72.29
	SGD	79.17%	77.08	78.60	77.54
Bigram TF-IDF	LR	78.47%	78.03	74.51	74.93
	Decision Tree	70.14%	69.37	70.52	68.38
	Random Forest	79.86%	79.50	78.04	77.54
	MNB	75.00%	73.32	73.46	72.01
	KNN	56.25%	63.03	59.69	55.92
	Linear SVM	70.83%	74.98	62.83	62.35
	RBF SVM	70.83%	74.98	62.83	62.35
	SGD	81.94%	80.56	80.11	79.64
Unigram CountVec	LR	79.86%	78.51	80.72	79.13
	Decision Tree	69.44%	69.58	70.74	68.37
	Random Forest	77.08%	77.27	76.67	75.30
	MNB	77.08%	76.19	76.60	75.17
	KNN	61.81%	62.79	62.38	60.12
	Linear SVM	75.69%	74.77	76.17	74.14
	RBF SVM	77.08%	76.11	78.13	75.89
	SGD	75.69%	74.09	76.43	74.45
Bigram CountVec	**LR**	**82.64%**	**81.65**	**82.70**	**81.66**
	Decision Tree	68.06%	67.62	65.55	65.73
	Random Forest	75.69%	75.48	76.39	74.27
	MNB	71.53%	71.40	70.63	68.77
	KNN	60.42%	61.46	60.49	58.47
	Linear SVM	77.08%	77.04	78.15	75.34
	RBF SVM	76.39%	75.71	77.10	74.65
	SGD	79.17%	76.92	78.67	77.26

We can see that when logistic regression was paired with the bigram CountVec, the maximum accuracy for multiclass sentiment classification was 82.64%, with generally good precision, recall, and F-1 score. However, it is important to remember that the performance of the other models fell short in contrast to that of the logistic regression model. The Unigram TF-IDF with SGD and the Bigram CountVec with Linear SVM both perform decently. In general, the KNN method performs less well than other models, notably Unigram TF-IDF and Bigram TF-IDF. Across many feature extraction methodologies, the Linear SVM and RBF SVM models produce comparable performance.

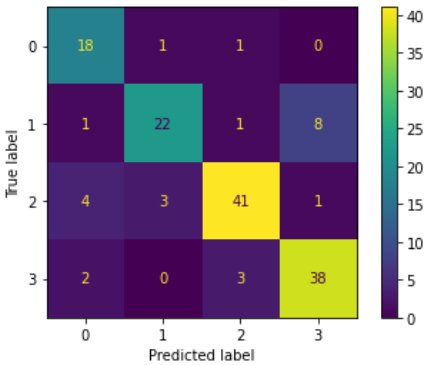

Fig.10. Confusion Matrix for Logistic Regression for multiclass sentiment classification

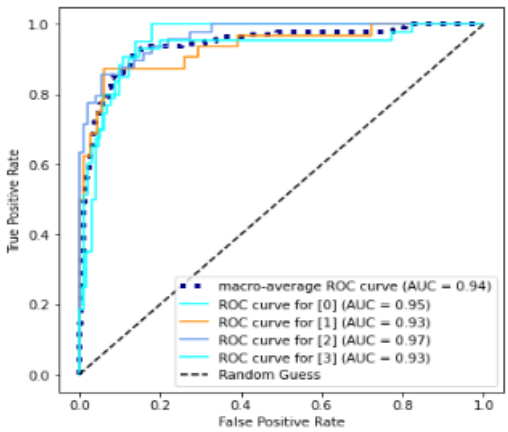

Fig. 11. ROC curve for Logistic Regression for multiclass sentiment classification

The Fig.10 and Fig.11 depicts the confusion matrix and the ROC curve of the most accurate sentiment classification algorithm, Logistic Regression.

In the confusion matrix, Negative, positive, very negative, and very positive are denoted as 0, 1, 2, 3 respectively. The majority of positive and very positive classes and the majority of negative and very negative classes were misclassified to each other. This suggests that distinguishing between these two pairs of sentiment classes proved to be more challenging for our model.

The macro-average ROC curve, which considers the overall performance across all classes, has an AUC of 0.94. This indicates that the model has good discriminatory power and performs well in distinguishing between positive and negative instances on average. The ROC curve for positive class, has an AUC of 0.93. While slightly lower than the ROC curve for negative class. The ROC curve for very negative class, has the highest AUC value of 0.97. This indicates that the model performs exceptionally well in correctly identifying very negative instances and distinguishing them from positive instances. The ROC curve analysis shows that the model generally performs well across all classes, with particularly high performance for the negative and very negative classes. The model demonstrates good discrimination ability and is successful in differentiating between positive and negative instances for each class.

5.2 Model Evaluation of Binary Sentiment Classification

The results of our categorization of sentiments into binary classifications are shown in Table 5. In this phase, the models were trained on and assessed in relation to two separate classes.

Table 5. Binary Sentiment Classification Evaluation metrics

Feature Extraction	Machine Learning Models	Accuracy	Precision	Recall	F1-Score
Unigram TF-IDF	LR	88.89%	89.46	89.16	88.88
	Decision Tree	86.11%	86.26	86.26	86.11
	Random Forest	**94.44%**	**94.42**	**94.49**	**94.44**
	MNB	92.36%	92.48	92.26	92.33
	KNN	84.03%	84.22	83.86	83.93
	Linear SVM	88.89%	89.46	89.16	88.88
	RBF SVM	90.97%	91.22	91.16	90.97
	SGD	93.06%	93.03	93.10	93.05
Bigram TF-IDF	LR	86.81%	87.50	87.10	86.79
	Decision Tree	86.11%	86.09	86.14	86.10
	Random Forest	88.19%	88.64	88.43	88.19
	MNB	92.36%	92.48	92.26	92.33
	KNN	81.94%	84.35	82.49	81.77
	Linear SVM	88.19%	88.91	88.49	88.18
	RBF SVM	88.19%	88.92	88.54	88.21
	SGD	92.36%	92.45	92.49	92.36
Unigram CountVec	LR	90.97%	90.97	91.04	90.97
	Decision Tree	87.50%	87.54	87.59	87.50
	Random Forest	90.97%	90.95	90.99	90.96
	MNB	91.67%	91.72	91.59	91.64
	KNN	83.33%	83.35	83.25	83.28
	Linear SVM	89.58%	89.82	89.77	89.58
	RBF SVM	89.58%	89.67	89.71	89.58
	SGD	88.19%	88.65	87.97	88.10

	LR	90.97%	91.06	91.10	90.97
	Decision Tree	87.50%	88.36	87.83	87.48
	Random Forest	90.97%	91.06	91.10	90.97
	MNB	91.67%	91.86	91.54	91.63
Bigram CountVec	KNN	81.25%	81.29	81.13	81.18
	Linear SVM	91.67%	92.01	91.88	91.67
	RBF SVM	90.97%	91.06	91.10	90.97
	SGD	89.58%	89.82	89.77	89.58

We are able to observe that the sentiment classification for binary classes yielded significant advances for every feature extraction technique. With an accuracy of 94.44%, the Random Forest with unigram TF-IDF vectorizer model produced the most remarkable results of all the models. It is essential to observe that the results of the remaining models for the remaining models were also quite impressive. Despite not being able to match the performance of Random Forest using the unigram TF-IDF vectorizer, they were still able to exhibit significant improvements over the multiclass sentiment classification. The Random Forest model consistently achieves high levels of accuracy, precision, recall, and F1-score across various feature extraction techniques. The performance of the Unigram TF-IDF feature extraction method combined with the SGD model is consistently high across all metrics. The efficacy of the Bigram Countvec feature extraction method combined with the linear SVM model is also impressive. Overall, the accuracy rates of the models range from 81.25 to 94.44%, indicating effective classification abilities.

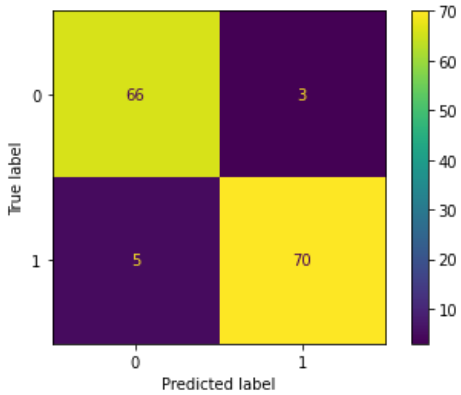

Fig.12. Confusion Matrix for Random Forest for binary class sentiment classification

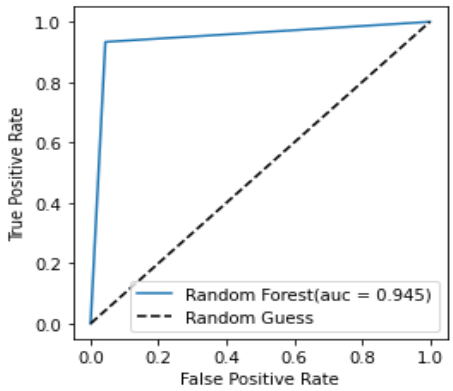

Fig.13. ROC curve for Random Forest for binary class sentiment classification

Fig.12 and Fig.13 shows the confusion matrix and ROC curve of the highest accurate algorithm; Random Forest for binary class sentiment classification.

In the confusion matrix, the value 0 represents the negative class, while the value 1 represents the positive class. It is clear that the majority of the occurrences that belonged to the positive class were incorrectly classified.

With an AUC of 0.945, the ROC curve demonstrates that the Random Forest model achieves a high level of discrimination ability. It means the model can effectively differentiate between positive and negative instances, resulting in a high true positive rate while maintaining a low false positive rate. The shape of the ROC curve provides additional insights into the model's performance. A curve that hugs the top-left corner of the graph suggests excellent performance, indicating a high true positive rate and a low false positive rate across various classification thresholds.

5.3 Comparison with previous works

Let's now explore how our machine learning-based classification of Bangla text compares to previous approaches.

Table 6. Comparative Analysis of other's ML models for sentiment classification

Author	Review Topic	Dataset	Classification Type	Best Model	Best Accuracy
[6]	Restaurant	1500	Multiclass	SVM	75.58%
[11]	Restaurant	1000	Binary	MNB	80.48%
[7]	E-Commerce	7905	Multiclass	KNN	96%
[12]	E-Commerce	1020	Binary	SVM	88.81%
[25]	Horoscope	6000	Binary	SVM	98.7 %
[13]	Book	5500	Binary	RF	98.39%
[8]	Social-Media	12628	Multiclass	LR	44%
[14]	Movies	1141	Binary	SVM	85.59%
[9]	Bangladesh Cricket	1601	Multiclass	SVM	64.59%
[26]	Trip advisor	337	Binary	Naive Bayes	72.04
[10]	Movies	4000	Binary	SVM	88.90%
Ours	**E-Commerce**	**1011**	**Binary** **Multiclass**	**RF** **LR**	**94.44%** **82.64%**

In Table 6, we can examine how our classification of Bangla text using machine learning models compares to that of others. Our primary objective was to evaluate the accuracy of binary and multiclass sentiment classification using machine learning algorithms. We discovered in the literature that the majority of past work mostly concentrated on binary class classification, with a smaller portion delving into multiclass classification. To give a thorough study, we decided to do both binary and multiclass classification, allowing us to directly compare and assess their performance in sentiment analysis tasks for Bangla text. Furthermore, we were able to attain substantially better results for multiclass classification.

5.4 Discussion

The accuracy of multiclass sentiment classification is substantially lower than that of binary classification. Multiclass models are intrinsically more complicated than binary models. As the number of classes increases, the decision boundary gets increasingly complex and challenging to interpret. Additionally, for multiclass sentiment classification, there was no significant distinction between the positive and very positive classes and the negative and very negative classes. With binary classification, however, the distinctions were far more obvious; hence, the outcomes were also significantly improved.

Compared to the few studies that have investigated multiclass sentiment analysis in Bangla, our research demonstrated significantly higher levels of accuracy. Compared to [6,8,9] where the authors conducted multiclass sentiment classification for the Bangla language using machine learning and attained 75.58%, 44%, and 64.59% using SVM, our best model outperformed them by a fine margin. Using a variety of machine learning algorithms and vectorization techniques, we were able to accurately classify Bangla text into very positive, positive, negative, and very negative sentiment classes.

In terms of binary sentiment classification, our findings showed notable improvements. We were able to significantly increase the accuracy with which Bangla text was classified as either positive or negative by employing numerous machine learning algorithms and vectorization techniques.

This improvement relates to the detailed evaluation and comprehensive study conducted on the performance of each classifier, which allowed us to identify the most appropriate method for multiclass sentiment analysis in Bangla.

6. Conclusion and Future Works

In this study, we analyzed the application of sentiment analysis for Bangla product reviews, concentrating on binary and multiclass sentiment classification. Our analysis involved a dataset containing roughly one thousand product reviews derived from the 'Daraz' website. The primary objective was to compare the sentiment classification accuracy of various machine learning models. Using the bigram count vectorizer, Logistic Regression attained an accuracy rate of 82.64 % for multiclass sentiment classification on our dataset, according to our findings. Regarding binary sentiment classification, Random Forest employing the unigram TF-IDF vectorizer produced the highest accuracy of 94.44%.

Based on the findings of this study, there are a number of potential avenues for future research. Primarily, it is essential to increase the scale and diversity of the dataset by integrating a wider variety of e-commerce sources. By incorporating a larger and more diverse dataset, it is possible to increase the accuracy of sentiment classification models, allowing for more reliable and generalized results. In order to improve the accuracy of the sentiment classification

model, it may be worthwhile to investigate deep learning models. In natural language processing tasks, including sentiment analysis, deep learning models, such as recurrent neural networks (RNNs) or transformer-based models such as BERT, have demonstrated strong performance. Incorporating such models into the analysis of Bangla product reviews could result in more precise and nuanced findings. Moreover, extending the research beyond sentiment classification to other aspects of opinion mining, such as aspect-based sentiment analysis, could provide a more comprehensive comprehension of consumer sentiments toward particular product attributes or features. This expansion would enable businesses to acquire a deeper understanding of consumer preferences and accordingly enhance their products and services.

This study contributes to the field of Bangla sentiment analysis by comparing the efficacy of machine learning models for binary and multiclass sentiment classification. There is room for improvement in multiclass classification despite obtaining high accuracy in binary classification.

References

[1] W. Medhat, A. Hassan, and H. J. A. S. e. j. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng J, vol. 5, no. 4, pp. 1093-1113, April. 2014, doi: 10.1016/j.asej.2014.04.011.

[2] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in Proceedings of human language technology conference and conference on empirical methods in natural language processing, 2005, pp. 579-586, doi: 10.3115/1220575.1220648.

[3] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in Proceedings of the first ACM conference on Online social networks, 2013, pp. 27-38, doi: 10.1145/2512938.2512951.

[4] K. A. Hasan, S. Islam, G. Mashrur-E-Elahi, and M. N. Izhar, "Sentiment recognition from bangla text," in Technical Challenges and Design Issues in Bangla Language Processing: IGI Global, 2013, pp. 315-327, doi: 10.4018/978-1-4666-3970-6.ch014.

[5] K. A. Hasan, M. S. Sabuj, and Z. Afrin, "Opinion mining using naive bayes," in 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Dec. 2015, pp. 511-514, doi: 10.1109/WIECON-ECE.2015.744398.

[6] F. Haque, M. M. H. Manik, and M. Hashem, "Opinion mining from bangla and phonetic bangla reviews using vectorization methods," in 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Dec. 2019, pp. 1-6, doi: 10.1109/EICT48899.2019.9068834.

[7] M. T. Akter, M. Begum, and R. Mustafa, "Bengali sentiment analysis of E-commerce product reviews using K-Nearest neighbors," in 2021 International conference on information and communication technology for sustainable development (ICICT4SD), April. 2021, pp. 40-44, doi: 10.1109/ICICT4SD50815.2021.9396910.

[8] T. Ahmed, S. F. Mukta, T. Al Mahmud, S. Al Hasan, and M. G. Hussain, "Bangla Text Emotion Classification using LR, MNB and MLP with TF-IDF & CountVectorizer," in 2022 26th International Computer Science and Engineering Conference (ICSEC), Dec. 2022, pp. 275-280, doi: 10.1109/ICSEC56337.2022.10049341.

[9] S. A. Mahtab, N. Islam, and M. M. Rahaman, "Sentiment analysis on bangladesh cricket with support vector machine," in 2018 international conference on Bangla speech and language processing (ICBSLP), Sep. 2018, pp. 1-4, doi: 10.1109/ICBSLP.2018.8554585.

[10] R. R. Chowdhury, M. S. Hossain, S. Hossain, and K. Andersson, "Analyzing sentiment of movie reviews in bangla by applying machine learning techniques," in 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), Sep. 2019, pp. 1-6, doi: 10.1109/ICBSLP47725.2019.201483.

[11] O. Sharif, M. M. Hoque, and E. Hossain, "Sentiment analysis of Bengali texts on online restaurant reviews using multinomial Naïve Bayes," in 2019 1st international conference on advances in science, engineering and robotics technology (ICASERT), May. 2019, pp. 1-6, doi: 10.1109/ICASERT.2019.8934655.

[12] M. A. Shafin, M. M. Hasan, M. R. Alam, M. A. Mithu, A. U. Nur, and M. O. Faruk, "Product review sentiment analysis by using NLP and machine learning in Bangla language," in 2020 23rd International Conference on Computer and Information Technology (ICCIT), Dec. 2020, pp. 1-5, doi: 10.1109/ICCIT51783.2020.9392733.

[13] M. E. Khatun and T. Rabeya, "A Machine Learning Approach for Sentiment Analysis of Book Reviews in Bangla Language," in 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Apr. 2022, pp. 1178-1182, doi: 10.1109/ICOEI53556.2022.9776752.

[14] M. Hassan et al., "Sentiment analysis on Bangla conversation using machine learning approach," Int J Elec & Comp Eng, vol. 12, no. 5, p. 5562, Oct. 2022, doi: 10.11591/ijece.v12i5.pp5562-5572.

[15] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing Management, vol. 39, no. 1, pp. 45-65, Jan. 2003, doi: 10.1016/S0306-4573(02)00021-3.

[16] M. Garg, "UBIS: Unigram bigram importance score for feature selection from short text," Expert Systems with Applications, vol. 195, p. 116563, Jun. 2022, doi: 10.1016/j.eswa.2022.116563.

[17] M. Maalouf, "Logistic regression in data analysis: an overview," International Journal of Data Analysis Techniques Strategies, vol. 3, no. 3, pp. 281-299, Jul. 2011, doi: 10.1504/IJDATS.2011.041335.

[18] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," Shanghai archives of psychiatry, vol. 27, no. 2, p. 130, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.

[19] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, pp. 197-227, Apr. 2016, doi: 10.1007/s11749-016-0481-7.

[20] L. Jiang, S. Wang, C. Li, and L. Zhang, "Structure extended multinomial naive Bayes," Information Sciences, vol. 329, pp. 346-356, Feb. 2016, doi: 10.1016/j.ins.2015.09.037.

[21] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," Neurocomputing, vol. 195, pp. 143-148, Jun. 2016, doi: 10.1016/j.neucom.2015.08.112.

[22] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: a review," Artificial Intelligence Review, vol. 52, no. 2, pp. 803-855, Jan. 2019, doi: 10.1007/s10462-018-9614-6.

[23] P. Netrapalli, "Stochastic gradient descent and its variants in machine learning," Journal of the Indian Institute of Science, vol. 99, no. 2, pp. 201-213, Jan. 2019, doi: 10.1007/s41745-019-0098-4.
[24] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," Procedia Computer Science, vol. 152, pp. 341-348, 2019, doi: 10.1016/j.procs.2019.05.008.
[25] T. Ghosal, S. K. Das, and S. Bhattacharjee, "Sentiment analysis on (Bengali horoscope) corpus," in 2015 Annual IEEE India Conference (INDICON), Dec. 2015, pp. 1-6, doi: 10.1109/INDICON.2015.7443551.
[26] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes," in 2019 12th international conference on information & communication technology and system (ICTS), Jul. 2019, pp. 49-54, doi: 10.1109/ICTS.2019.8850982.

Authors' Profiles

Shakib Sadat Shanto an undergraduate currently studying Bachelor of Science in Computer Science and Engineering at American International University Bangladesh. He is extremely passionate about Artifical Intelligence and Data Science domain. He wants to do further research on Computer Vision, Human Computer Interaction, Natural Language Processing and Deep Learning.

Zishan Ahmed an enthusiastic undergraduate pursuing a Bachelor of Science in Computer Science and Engineering. He is captivated by the potential of data to alter the world we live in. In data science, natural language processing (NLP), and machine learning, he sees the greatest potential for innovation and influence. His knack for mathematics and programming has been refined throughout his academic career. He is well-versed in several programming languages, including Python, Java, and C++, and is always keen to acquire new tools and technologies. His education has included data structures and algorithms, database management, artificial intelligence, and computer vision.

Nisma Hossain is studying at American International University Bangladesh in Computer Science and Engineering. She has interests in Software Engineering, Web Development and Data Science, Data processing, Data mining algorithms, etc. She wants to do further research in Machine Learning and Data Science.

Auditi Roy is studying at American International University Bangladesh in Computer Science and Engineering. She is strongly captivated to continue research and open to work in any area of Machine Learning especially Natural Language Processing. Besides this, Big Data and Sentiment Analysis are also the realm of her interest.

Dr. Akinul Islam Jony currently works as an Associate Professor of Computer Science at American International University- Bangladesh (AIUB). He was a recipient of a Doctoral Grant for doing his Doctor of Philosophy (PhD). He has a number of publications in different peer reviewed journals. Also, he has presented several research papers at well-known international conferences. He is the author of several book chapters published in Springer Proceedings of Complexity. His current research interest includes e-Learning, machine learning, big data, and issues in data science.c

How to cite this paper: Shakib Sadat Shanto, Zishan Ahmed, Nisma Hossain, Auditi Roy, Akinul Islam Jony, "Binary vs. Multiclass Sentiment Classification for Bangla E-commerce Product Reviews: A Comparative Analysis of Machine Learning Models", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.15, No.6, pp. 48-63, 2023. DOI:10.5815/ijieeb.2023.06.04