Sentiment Analysis of Assamese Text Reviews: Supervised Machine Learning Approach with Combined n-gram and TF-IDF Feature

Chandana Dev[1], Amrita Ganguly[2]

[1,2] Department of Electrical Engineering, Assam Engineering College
Jalukbari, Guwahati- 781013, Assam, INDIA.
[1]*chandanaaec@gmail.com**, [2]*aganguly.ele@aec.ac.in*

Abstract: ***Sentiment analysis (SA) is a challenging application of natural language processing (NLP) in various Indian languages. However, there is limited research on sentiment categorization in Assamese texts. This paper investigates sentiment categorization on Assamese textual data using a dataset created by translating Bengali resources into Assamese using Google Translator. The study employs multiple supervised ML methods, including Decision Tree, K-nearest neighbour, Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine, combined with n-gram and Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction methods. The experimental results show that Multinomial Naive Bayes and Support Vector Machine have over 80% accuracy in analyzing sentiments in Assamese texts, while the Unigram model performs better than higher-order n-gram models in both datasets. The proposed model is shown to be an effective tool for sentiment classification in domain-independent Assamese text data.***

Keywords: Assamese; Machine Learning; n-gram; NLP; Sentiment Analysis; TF-IDF.

Article history- Received: 5 July 2023 and Accepted: 13 September 2023.

1. Introduction

With the enormous growth of information technology, the internet has become an indispensable component of people's daily lives. It has become increasingly common for individuals to express their thoughts on a range of matters such as products, films, recent news, and services via natural language on web discussion forums, blogs, and various social media outlets. These platforms provide useful information regarding a wide range of domains, including commercial, political and social applications [1]. It is challenging to manually process this large volume of data for analysis. In this context, sentiment analysis (SA) has proven to be an extremely useful tool. It has emerged as a popular research topic in Natural Language Processing (NLP). Sentiment analysis, also known as opinion mining, explores feelings and thoughts towards specific entities [2-9]. It is used for a variety of other applications, including user views analysis and prediction, marketing strategies and stock market prediction [10]. As the Internet and other digital platforms have rapidly developed, natural language processing and opinion extraction specialists have put greater focus on analyzing large databases and text archives. Extracting sentiment from this plethora of data, which involves labelling a text as a positive or a negative appraisal of a target object (movie, book, product, service etc.) has become a challenging task for researchers [11].

Many studies on sentiment analysis have been carried out in several languages, including Chinese [12,13] Spanish [14], French [15] and many Indian languages like Bengali [16], Hindi [17], Tamil [18] Telugu [19], Malayalam [20,21] etc. However, no SA has been carried out in the Assamese language, which is a popular language spoken in the Northeastern region of the country. Resources available for Assamese are also meagre. As such, Sentiment Analysis in low-resource languages, such as Assamese, is hampered by a lack of adequate data sets. To date, no literature has attempted to specifically analyze Assamese text data for sentiment.

For a domain-specific sentiment classifier to predict accurate results, the system must be trained with a large amount of labelled data. The creation of this large labelled data is costly and time-consuming [22]. Like any other language, sentiment analysis in Assamese has grown in popularity as a result of the growing need to interpret sentiments in social media, customer reviews and other text sources. Traditional machine learning methods have been used effectively to do such research. The first

step for this research work is to create a corpus of labelled data in Assamese. This research work presents an empirical study of sentiment analysis on the newly created labelled Assamese text data. To achieve this a supervised learning model is developed to classify text data in terms of positive and negative sentiments written in Assamese. This study aims to determine the optimal combination of TF-IDF and n-grams for different data sets. The most efficient technology for sentiment analysis is the conventional ML approach because it offers a dynamic platform for experimentation and analysis. Support vector machine (SVM), Naive Bayes (NB), Decision tree (DT), k-nearest neighbours (kNN), Random Forest (RF) classifier, etc. are some of the traditional ML classifiers that are trained on the labelled dataset to predict sentiment polarity. This research work mainly focuses on using distinctive features and multiple machine learning algorithms on two different domain datasets.

2. Literature review

The sentiment analysis of review data has become a popular area of research in recent times. Related work on sentiment categorization has been presented in this section. Sentiment classification has been studied in a variety of contexts, including film reviews, product reviews, and customer feedback reviews [23,17]. Till now, most of these research efforts have focused on training machine learning algorithms to classify reviews [11]. In [11], the authors experimented with a dataset of movie reviews for the classification of sentiments where they classified sentiments into two classes positive and negative. Three machine learning techniques (NB, SVM and DT) have been used to perform SA using the n-gram technique. It is observed that in comparison to the other two algorithms, NB underperforms in classifying sentiments in the texts. Authors in [24] have used SVM to classify sentiment in a dataset of movie reviews. Here, SA using a unigram-based feature model has been performed and the results are compared to previous works using k-fold cross-validation with k = 3,10. Researchers in [25] used SVM to perform SA on movie reviews while taking context valence shifters into account. It is found that including context valence shifters increases the system's accuracy. To increase the number of words in the term count method, authors have also used General Inquirer and Choose to Right Word. They implemented unigrams, bigrams, and adjectives as features for the machine learning algorithm and achieved an 86.2% accuracy using context valence shifters. In [26], authors used three machine learning techniques NB, SVM and the n-gram model to perform SA on travel blog reviews. Using three-fold cross-validation in the experiments, it is observed that SVM and n-gram models outperform NB. Authors in [27] used blogs, forums, and online reviews in English, French, and Dutch to perform SA. Machine learning techniques such as Maximum Entropy (ME), SVM, and Multinomial Naive Bayes using unigrams, bigrams, and adjectives as features have been implemented in that work. Using unigrams as features, for English, French, and Dutch, an accuracy of 83%, 68%, and 70% respectively have been reported. In [28] the authors performed document-level sentiment classification on product and movie reviews, using Artificial Neural Network (ANN) and SVM machine learning algorithms. Here, using a bag-of-words model, sentiment has been classified into positive and negative sentiments. In [29] authors used Twitter data sets such as the Obama-McChain Debate, SentiStrength Twitter data set, Sanders, and SemEval to perform sentiment classification. They used four popular machine learning algorithms to perform sentiment analysis which are Naïve Bayes (NB), Decision Tree, K-Nearest Neighbor (kNN) and Support Vector Machine (SVM). For text-to-feature vector conversion, n-gram, Medical Research Council (MRC), Linguistic Inquiry and Word Count (LIWC), Apache OpenNLP toolkit and Stanford Part-Of-Speech (POS) tagger have been used. Tweet SA was performed by [30], where the feature set for the implementation of NB and SVM is comprised of information gain, diagrams and object-oriented extraction techniques. Researchers in [31] performed SA on tweets at the global and aspect levels in the Spanish language. A graph-based algorithm to extract the features and polarity lexicons to determine the sentiments has been used in this paper. Authors in [32] discuss a sentiment analysis approach to movie reviews in Hindi. Also, in [33] the authors created a Bangla sentiment analysis model using a Bengali tweet dataset. They classified the sentiment polarity expressed in tweets using Multinomial Naive Bayes and SVM classifiers, with a feature set including N-gram and SentiwordNet features. The developed model demonstrated that an SVM classifier trained with unigram and SentiwordNet features outperforms other classifiers on the Bengali tweet dataset. Aspect-based mining identifies a sentence aspect and the user's comments on these aspects to generate a positive and negative review [34]. The authors provide a novel approach to creating TF-IDF vectors for Assamese text. Authors in [35] developed a sentiment polarity classification model for the low-resource Assamese language within the news domain. Their model incorporates lexical features such as adjectives, adverbs, and verbs. The developed model demonstrates enhanced performance compared to the baseline, as indicated by an improved F1-score on the standard dataset. Like any other language, SA in Assamese has a high demand as a result of the growing need to interpret sentiments in social media, customer reviews and other text sources. Researcher [36] focuses on

conducting SA on Assamese Texts. Their study leverages the well-known sentiment analyzer "Vader" and adapts it for Assamese by building upon the "Bengali-Vader" framework.

3. Classification Techniques

Different classification techniques used in this research work are briefly discussed in the following.

3.1 Decision Tree (DT)

Decision Tree analysis is a supervised learning technique used for both classification and regression. It works by hierarchically splitting data based on conditions set on its attributes, recursively dividing it until the leaf nodes have the smallest number of records. The class label with the majority in the leaf node is then used for classification. [17,34].

3.2 Logistic Regression (LR)

Logistic regression, also known as logit regression, is used to estimate the parameters of a logistic model, such as the coefficients in the linear combination. This type of regression analysis is composed of binary logistic regression, wherein the independent variables can take the form of either continuous or binary values. The dependent variable, labelled "0" and "1" respectively, is also a binary indicator. As such, the corresponding probability of the "1" value will lie between 0 and 1, due to the logistic function converting the log-odds into a probability [37,38].

3.3 Multinomial Naïve Bayes (MNB)

Multinomial Naïve Bayes (MNB) is a probabilistic classifier based on Bayes' theorem which assumes that the terms occur independently. In Bayesian classification, a hypothesis is established that given data belongs to a sentiment analysis system for a specific class and the probability of the hypothesis being true is calculated [39]. Collection of N sentences $(S)_{i=1}^{N}$ is given in review data, where each sentence consists of 'T' terms such that $S_i = \{t_1, t_2, \ldots\ldots, t_T\}$, the probability of S_i to occur in class C_k is given by the following equation

$$P(C_k|S_i) = P(C_k) \prod_{j=1}^{T} P(t_j|C_k) \qquad \ldots\ldots\ldots\ldots(1)$$

Here, $P(C_k|S_i)$ is the conditional probability of the term t_j to occur in a sentence of class C_k, and $P(t_j|C_k)$ is the prior probability of a sentence occurring in a class C_k. $P(t_j|C_k)$ and $P(C_k)$ are calculated from training data.

3.4 Support Vector Machine (SVM)

Following the rule of Structural Risk Minimization, this strategy operates by finding a dividing surface, or hyperplane, that separates two classes of data points in the best possible way. This is especially pertinent in binary classification problems, where the hyperplane must ensure the greatest amount of space between the two classes [40]. SVM's decision surface for linearly separable space is a hyperplane. Since SVM requires input in the form of a vector of numbers, text file reviews for classification must be converted to numeric values. After converting the text file to a numeric vector, it may go through a scaling process to help manage the vectors and keep them within the range of [1, 0].

3.5 K Nearest Neighbor (kNN)

Since the kNN classifier is more suitable for large datasets and sentiment analysis is a binary classification, hence kNN is chosen for the work. In this case, the classifier is trained using a manually generated training set. Within the training set, there is an X:Y relationship provided in which the score of an opinion word is represented by 'X' and the score of whether the word is positive or negative is represented by 'Y' [41-42]. The kNN classifier is fed a score of the opinion word related to a feature in the review.

3.6 N-gram model

The N-gram model is a method for checking 'n' continuous words or sounds from a given text or speech sequence. This model aids in the prediction of the next item in a sequence. The n-gram model aids in sentiment analysis by analyzing the sentiment of a text or document. Unigram is an n-gram of size 1, Bigram is an n-gram of size 2, and Trigram is an n-gram of size 3. Four-gram, five-gram, and so on are examples of higher n-grams [43].

A typical example of a sentence may be considered as "*The movie is not a good one*".

- Its unigram: "'The', 'movie', 'is', 'not', 'a', 'good', 'one'" where a single word is considered.
- Its bigram: "'The movie', 'movie is', 'is not', 'not a', 'a good', 'good one'" where a pair of words are considered.
- Its trigram: "'The movie is', 'movie is not', 'is not a', 'not a good', 'a good one'" where a set of words having a count equal to three is considered

4. Methodology

This section describes the sentiment classification system methodology that has been used in this research work. The dataset comprises restaurant and movie reviews in text format. However, numerical matrices are required as input for sentiment classification using machine learning algorithms. Hence, different methods are used to convert text data in reviews into numerical matrices. To do this initially, Unigram, Bigram and Trigram features have been extracted from the cleaned texts, and then TF-IDF vectorizer is used to convert a text document into a numerical vector, which is then fed into a supervised machine learning algorithm. The input dataset needs to be labelled when thinking about supervised machine learning algorithms for classification. In this work, different supervised machine learning techniques such as Decision Tree (DT), Logistic Regression (LR), Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), and K Nearest Neighbor (kNN) have been used. A variety of statistical measures such as Confusion matrix, precision, recall, and f1 score are evaluated [44]. The confusion matrix's parameters used to evaluate the performance of the proposed model are briefly discussed in Table 1.

Table 1: Confusion Matrix Parameters and their meanings

TP	It represents the positive reviews, and the classifier also classifies them as positive
FP	It represents positive reviews, but the classifier classifies them as negative
TN	It represents the negative reviews, and the classifier also classifies them as negative
FN	It represents negative reviews, but the classifier classifies them as positive

As shown in Table 1, a comparison of labels of classes is done using four terms which are, true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Based on the values of parameters retrieved from the confusion matrix, other performance measures such as Accuracy, F-measure, precision, and recall are calculated as given in Table 2.

Table 2: Machine Learning performance parameters

Performance measure	Description	Formula
Accuracy	Accuracy measures the fraction of correctly classified samples.	$\frac{TP+TN}{TP+FP+TN+FN}$
Precision	It is described as the proportion of correct predicted positive cases.	$\frac{TP}{TP+FP}$
Recall	It represents the proportion of correctly identified positive cases.	$\frac{TP}{TP+FN}$
F1 score	It is given by harmonic mean and computes the average of precision and recall.	$\frac{2\times precision\times recall}{precision+recall}$

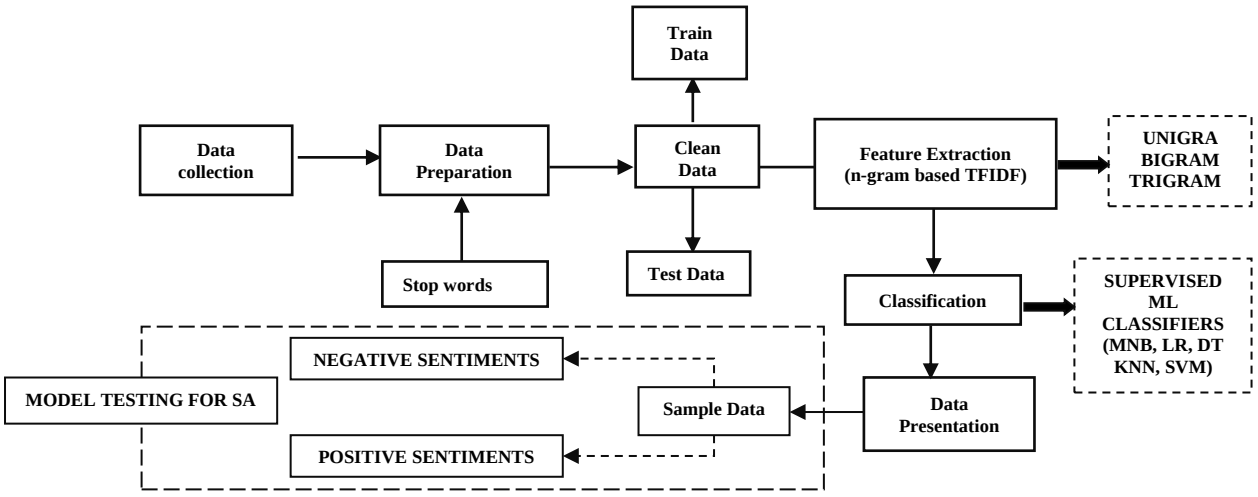


Figure 1: Proposed sentiment analysis model

Figure 1 above shows the operation of a proposed SA system for the Assamese language. It is divided into the following Sections: 'Data collection', 'Data preparation and Data Cleaning', 'Feature Extraction', 'Classification' and finally 'Model Testing'.

Step 1: Data Collection

The unavailability of the Assamese standard dataset makes research work in this language more challenging. However, it has been observed that recently researchers are compiling their dataset for their respective objectives. As there are no standard datasets available in Assamese reviews, it is difficult to develop a corpus that contains Assamese texts. In this work initially, a labelled dataset of Bengali restaurant reviews of around 1400 reviews is gathered from various social media groups where customers provide food and restaurant reviews [43]. Also, a labelled data set of around 2,000 movie reviews [44] have been collected for this research work which have been then translated through Google translator.

Step 2: Data Preparation & Data Cleaning

Data are pre-processed for further processing during this phase. In the pre-processing step, tokenization, stopword removal, removal of punctuations etc. have been done. Pre-processed data is then cleaned by removing unnecessary symbols, tokens and numbers from the texts.

A training set R = {r_1, r_2, r_3, ..., r_n} is made up of 'n' training reviews. Each review contains either positive or negative sentiment, denoted by the letters C_p and C_n respectively. A word vector W[ ] = {w_1, w_2, w_3, ..., w_l} represents a review ' r_i ' with 'l' words. To remove inconsistencies from the dataset, all reviews are preprocessed. To remove the words 'w_i', which have no contribution in determining whether a review 'r_i' conveys positive (C_p) or negative (C_n) sentiment, a stopwords list S[ ] = {s_1, s_2, s_3, ..., s_t} with 't' stopwords has been developed. Removal of stopwords from a review is done by removing stopwords s_1, s_2, s_3, ..., s_t which are in the stopwords set S. Conjunctions, prepositions, interjections, pronouns, suffixes, and prefixes are all examples of stopwords. Some sample stopwords in the Assamese language are given in Table 3.

Table 3: Sample stop words

Stopwords	Type	examples
s_1	Pronoun	আপুনি
s_2	Preposition	ওপৰত
s_3	Interjection	বাহ্
s_4	Conjunctions	আৰু

A sample of cleaned data is shown below:

Movie Dataset:

Original: নাটকখন ভাল লাগিল, দৰ্শকক এনেকুৱা সুন্দৰ নাটক দিয়াৰ বাবে ধন্যবাদ।

Cleaned: নাটকখন ভাল লাগিল দৰ্শকক সুন্দৰ নাটক ধন্যবাদ

Here, stop words and punctuations like 'এনেকুৱা', 'দিয়াৰ', 'বাবে' and, respectively which are removed from the original review.

Restaurant Dataset:

Original: গ্ৰিলটো ঠাণ্ডা আছিল, ৰুটিখন শুকান আৰু কঠিন আছিল, ষ্টকটো পুৰণি আছিল। অতি নিম্নমানৰ খাদ্য হতাশ।

Cleaned: গ্ৰিলটো ঠাণ্ডা ৰুটিখন শুকান কঠিন ষ্টকটো পুৰণি নিম্নমানৰ খাদ্য হতাশ

Here, stop words like 'আছিল', 'আৰু', 'আছিল', 'অতি', respectively are removed from the original review. The final dataset prepared thus consists of two columns namely 'Reviews' and 'Sentiment', where reviews are listed under the 'Reviews' column and on the other hand sentiments are listed under the 'Sentiment' column labelled as positive and negative. The sample dataset template is shown in Table 4 and Dataset statistics are summarized in Table 5.

Table 4: Dataset Template

Index	Reviews	Sentiment
1	Review 1	Positive
2	Review 2	Positive
⋮	⋮	⋮
N	Review N	Negative

Table 5: Dataset Summary

Sl. No.	Reviews	Total Reviews	No. Reviews		No of words		Unique words		No. Documents	
			Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
1	Movie dataset	1990	977	1013	6432	6416	1470	2007	974	995
2	Restaurant dataset	1401	637	764	5669	14262	1449	3984	561	745

Step 3: Feature selection method

TF-IDF is a commonly utilized feature extraction technique in various Natural Language Processing (NLP) tasks, including text processing, information retrieval, and opinion mining. By calculating the probability of a word appearing in a document, it is possible to assess the relevance of a given word to that particular document. In essence, the higher the TF-IDF value of a word, the greater its importance in the text [35]. In this work, a vocabulary of Assamese words is created by tokenizing reviews of the corpus. "TF IDF": the term frequency-inverse document frequency statistics along with n-gram are used as features of the proposed model. Here, used "TF IDF" statistic is written as

$$tf\ idf\ (w,r) = tf(w,r) \log N|\{r€R : w€r\}|.. (3)$$

Here, $tf\ idf\ (w,r)$ = value of word w in review r.
$tf(w,r)$ = frequency of word w in review r.
N = total number of reviews.
$|r€R : w€r\}|$ = number of reviews containing w.

Step 4: Machine Learning Methods

Different machine learning techniques, MNB, SVM, and DT, LR, k-NN have been used in this research experiment to classify the sentiments of reviews written in Assamese. Movie reviews and Restaurant reviews are two different types of data used for the same. Since the objective of the proposed model is to categorize the reviews, classification is the most important part of the system. The extracted features from the reviews are used to train the proposed model, which can categorize reviews as positive or negative sentiments.

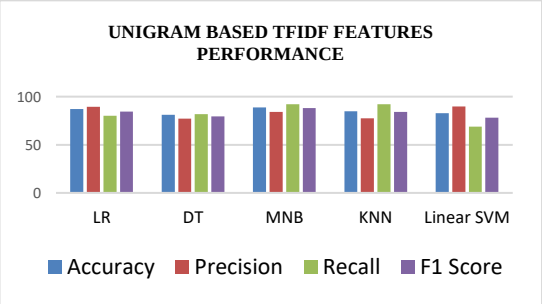
5. Experimental Design

In this work, the movie reviews and restaurant reviews are analyzed using various machine learning classifiers such as MNB, SVM, LR, DT, and kNN. The following section presents a comparison of the performance of these classifiers in terms of different performance evaluation attributes. Analysis results on various evaluation aspects for two different datasets are depicted in Fig. 2 and Fig. 3.

Classifier Report for N-gram based TFIDF Feature extracted model:

The following graphs depict the performance of different classifiers on (i) restaurant review and (ii) movie review datasets using unigram-based TFIDF features respectively.

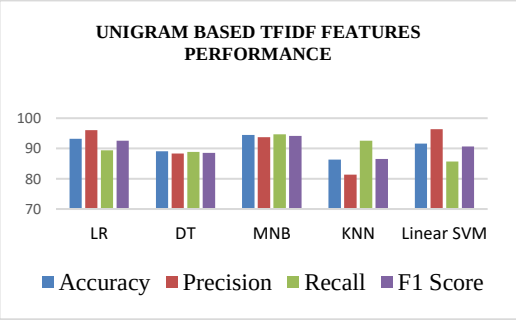
Restaurant Review dataset:



Classifier	Accuracy	Precision	Recall	F1 Score
LR	87.02	89.42	80.17	84.55
DT	81.3	77.24	81.9	79.5
KNN	84.73	77.54	91.24	83.83
MNB	88.93	84.25	92.24	88.07
Linear SVM	82.82	89.89	68.97	78.05

Figure 2: Unigram-based TF-IDF features performance for restaurant dataset

Movie Review Dataset:



Classifier	Accuracy	Precision	Recall	F1 Score
LR	93.15	96	89.36	92.56
DT	89.09	88.36	88.83	88.59
MNB	94.42	93.68	94.68	94.18
KNN	86.29	81.31	92.55	86.57
Linear SVM	91.62	96.41	85.64	90.7

Figure 3: Unigram-based TF-IDF features performance for movie dataset

As illustrated in the results shown in Figures 2 and 3 above, the following observations are obtained:

Multinomial Naïve Bayes classifier performs with the highest accuracy with an average of more than 85% (i.e. 88% for restaurant data and 94% for movie data) whereas, Support Vector Machine (SVM) also performs well with an accuracy of more than 80% (i.e 83% for restaurant data and around 92% for movie data).In terms of the evaluation parameter "Precision", Support Vector Machine (SVM) performs with the highest value of 89 % for restaurant data and 96 % for movie data compared to all other classifiers. The unigram-based TFIDF feature extraction technique has achieved maximum accuracy for both classifiers. For both the datasets used in this work, the model has shown similarity in performance in terms of machine learning classifier and feature extraction techniques.

The experimental design indicates that both Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) classifiers using Unigram-based TFIDF feature extraction method can perform with more than 80% accuracy level for Assamese text data irrespective of domains.

6. Results and Analysis

The proposed model has also been tested with all the mentioned machine learning classifiers. It is observed that out of the classifiers, MNB and SVM classifiers give the better performance. Classifier reports of MNB and SVM are illustrated in Table 6 for both datasets. It shows that as the dataset size increases the accuracy increases, showing the efficacy of both the classifiers. Also, a higher value of F1 score indicates better classification, the values obtained show that MNB and SVM have performed better as compared to the rest of the classifiers for both datasets. Confusion matrices of MNB and SVM classifiers have been observed for the validation of correct predictions of the sentiments for both datasets as shown in figures 4 to 7 below. Figure 8 shows the performance of the MNB and SVM classifiers for different N-gram features, where N = 1,2,3.

Table 6: Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) Classifier report

CLASSIFIER	DATASET	ACCURACY	PRECISION	RECALL	F1-SCORE
MNB Classifier	RESTAURANT (1400 reviews)	0.89	0.90	0.89	0.90
	MOVIE (2000 reviews)	0.93	0.97	0.88	0.92
SVM Classifier	RESTAURANT (1400 reviews)	0.82	0.89	0.69	0.78
	MOVIE (2000 reviews)	0.91	0.96	0.86	0.90

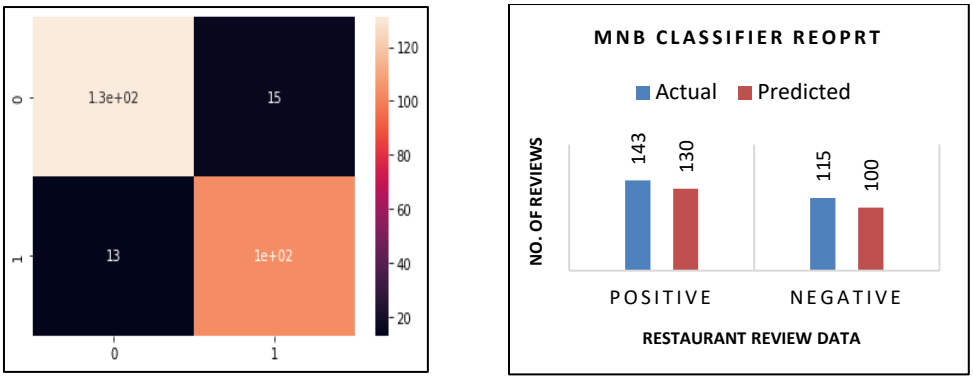


Figure 4: Confusion matrix evaluation of MNB classifier for Restaurant data

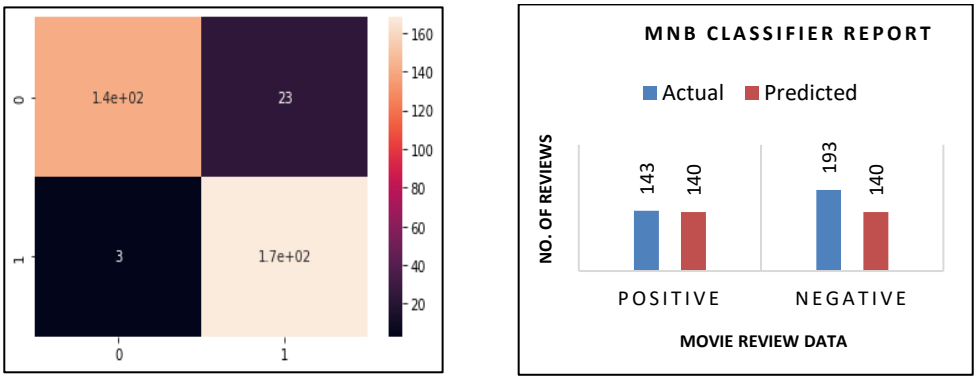


Figure 5: Confusion matrix evaluation of MNB classifier for Movie data

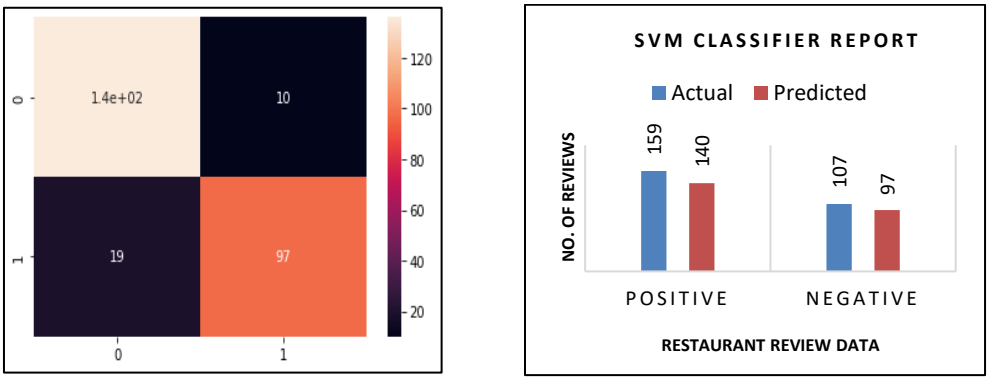


Figure 6: Confusion matrix evaluation of SVM classifier Restaurant data

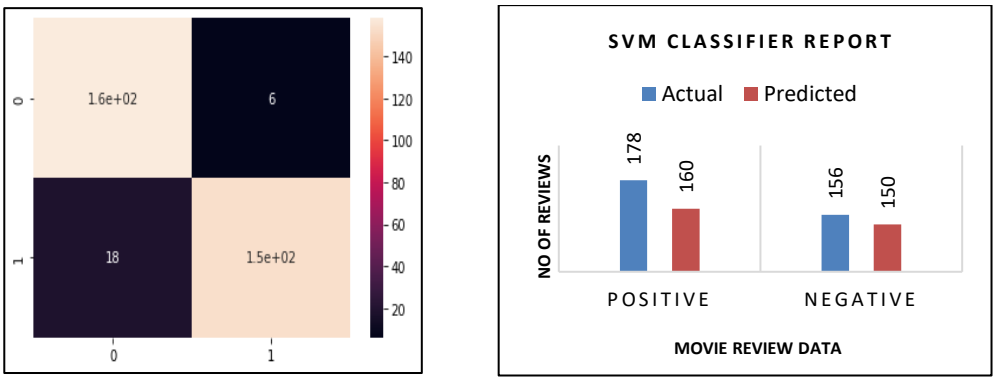


Figure 7: Confusion matrix evaluation of SVM classifier Movie data

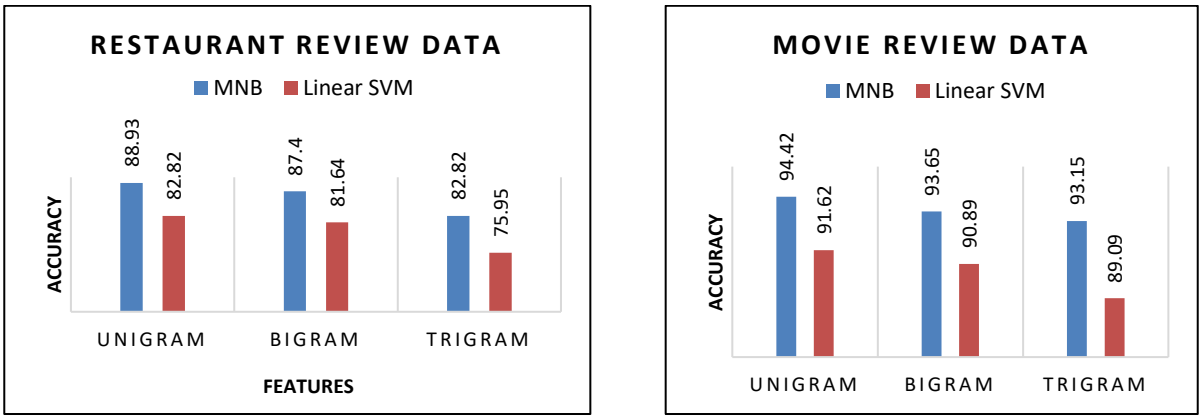


Figure 8: Comparative analysis of N-gram feature-based methods for MNB and SVM Classifier

The effectiveness of the system has been examined using K-fold cross-validation with K = 4, 6, 10. The performance of these algorithms is compared in Table 7 and Table 8 for various K values. According to the findings in the cross-validation report for the K values, MNB and SVM perform better than the rest of the classifiers for both datasets. Both the classifiers show an accuracy of more than 85% for both the datasets considered for this experiment in the validation accuracy test. Finally, all the classifiers used in this research work for the experimental design have been tested on a few sample reviews chosen randomly. To verify the scope and limitations of the developed model commonly annotated reviews on restaurant and movie reviews have been used with all possible common comments generally posted in online platforms. These review texts from both datasets have predicted accurately in terms of sentiment categorization by the proposed model. Additionally, to verify the efficiency of the model performance the same sample texts have been tested with the existing standard Natural Language Tool Kit (NLTK) sentiment analyzer. A brief overview of this comparative analysis is shown in Table 9. The comparison results among our proposed model, NLTK and human interpretation of sentiments in Assamese text data have been illustrated in Figure 9. This shows the effectiveness of the proposed model in sentiment analysis of Assamese text data obtained from different domains.

Table 7: Cross-Validation Report for Restaurant Data

K-VALUE	CLASSIFIER	ACCURACY	PRECISION	RECALL	F1-SCORE
4	MNB	0.86	0.85	0.88	0.87
	SVM	0.82	0.89	0.71	0.79
6	MNB	0.87	0.87	0.87	0.87
	SVM	0.84	0.91	0.76	0.82
10	MNB	0.88	0.87	0.89	0.87
	SVM	0.85	0.91	0.77	0.83

Table 8: Cross-Validation Report for Movie Data

K-VALUE	CLASSIFIER	ACCURACY	PRECISION	RECALL	F1-SCORE
4	MNB	0.89	0.83	0.99	0.89
	SVM	0.90	0.90	0.95	0.90
6	MNB	0.91	0.85	0.99	0.91
	SVM	0.91	0.91	0.95	0.91
10	MNB	0.89	0.83	0.99	0.90
	SVM	0.90	0.90	0.92	0.91

Table 9: Comparative analysis of sample texts with NLTK and Human Score

Sl. No.	SAMPLE TEXTS	CLASSIFIER	SENTIMENT CLASS		
			OUR MODEL	NLTK	HUMAN OUTPUT
1	Assamese: খাদ্য বেয়া, পৰিৱেশ ভাল English: Food is bad, atmosphere is good	MNB	POSITIVE	POSITIVE	POSITIVE
		LINEAR SVM			
2	যদি আপুনি এটা সাধাৰণ সোৱাদ ভাল পায় ,এয়া এটা বিকল্প হ'ব পাৰে English: If u like a simple taste, it may be an option	MNB	NEGATIVE	NEGATIVE	NEGATIVE
		LINEAR SVM			
3	ভাল সেৱা পাইছিলোঁ, খাদ্য সুন্দৰকৈ পৰিবেশন কৰা হৈছিল English: Got good service, food was nicely served	MNB	POSITIVE	POSITIVE	POSITIVE
		LINEAR SVM			
4	চিনেমাৰ চিত্রনাট্য আছিল অতি উত্তম,এইটো এখন সুন্দৰকৈ চিত্রিত কৰা চিনেমা হৈছে English: The script of the movie was excellent; it is a beautifully portrayed movie	MNB	POSITIVE	POSITIVE	POSITIVE
		LINEAR SVM			
5	চিনেমাখন বৰ সাধাৰণ আছিল,চিনেমাৰ অভিনেতা বেয়া আছিল English: movie was very average; movie cast was bad.	MNB	NEGATIVE	NEGATIVE	NEGATIVE
		LINEAR SVM			

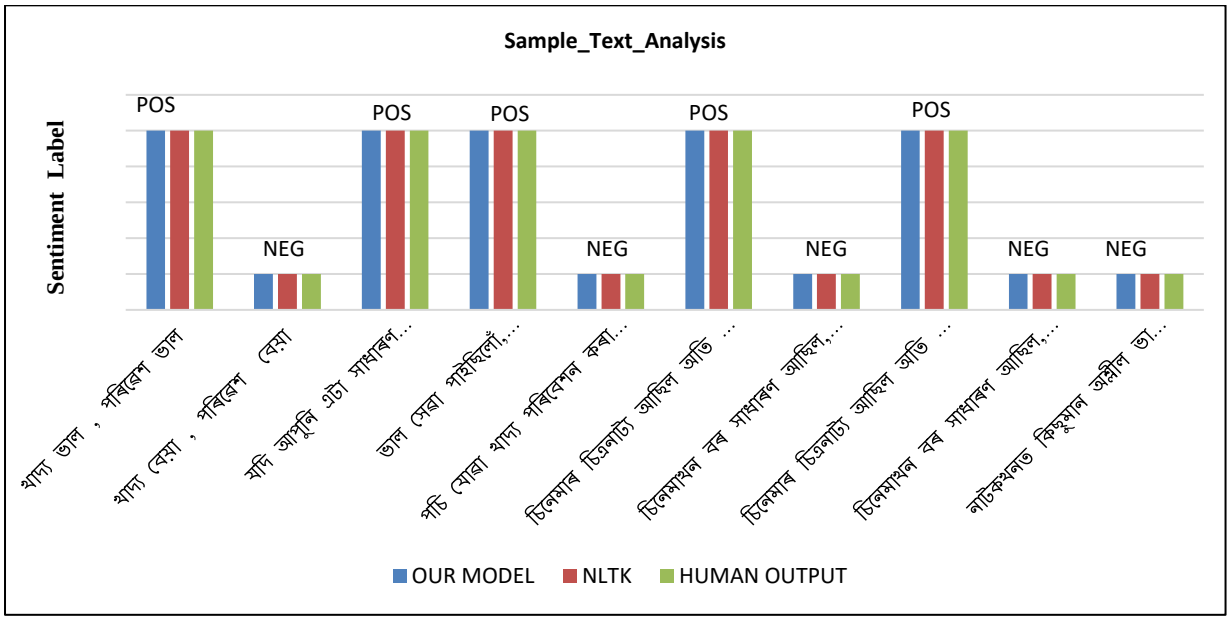


Figure 9: Comparison of sentiment labels for sample texts

This study uses multiple machine learning classifiers to test a proposed model for sentiment analysis in restaurant and movie reviews. Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) classifiers consistently outperform other classifiers, with accuracy rates of over 85% across different dataset sizes. The model's efficiency is demonstrated by precise sentiment analysis of test texts from Assamese-language restaurant and movie reviews, comparing it with the Natural Language Toolkit sentiment analyser and human interpretation. The model's efficacy is further verified through cross-validation and comparison with other sentiment analysis techniques.

7. Conclusion

This study focused on experiments that were designed to predict the proper sentiment in two distinct categories positive and negative with randomly created review data from two different domains (movie and restaurant). It is the first of its kind to analyze Assamese text data. The developed model demonstrated impressive results on randomly chosen reviews from both domains. It is plausible that the accuracy and precision of the classification can be augmented with advanced feature extraction methods and even more with the amalgamation of machine learning techniques. For the future, this research intends to investigate deep learning models and incorporate larger datasets from various domains such as social media, product reviews, and news.

References:

[1] O. Almatrafi, S. Parack, and B. Chavan, "Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014," in *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, Bali Indonesia: ACM, Jan. 2015, pp. 1–5. doi: 10.1145/2701126.2701129.

[2] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th International Conference on Computational Linguistics - COLING'04*, Geneva, Switzerland: Association for Computational Linguistics, 2004, pp. 1367-es. doi: 10.3115/1220355.1220555.

[3] B. Liu, "Sentiment Analysis and Subjectivity" in *Handbook of Natural Language Processing*, N. Indurkhya and F. J. Damerau, Eds., 2nd edition, New York: Chapman and Hall/CRC, 2010. Available: https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf [Accessed: May 19, 2023]

[4] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the Web," in *Proceedings of the 14th International Conference on World Wide Web - WWW'05*, Chiba, Japan: ACM Press, 2005, p. 342. doi: 10.1145/1060745.1060797.

[5] V. N. Patodkar and S. I.R, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 12, pp. 320–322, Dec. 2016, doi: 10.17148/IJARCCE.2016.51274. Available: http://ijarcce.com/upload/2016/december-16/IJARCCE%2074.pdf [Accessed: May 19, 2023]

[6] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL'04*, Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 271-es. doi: 10.3115/1218955.1218990.

[7] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *FNT in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[8] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 417. doi: 10.3115/1073083.1073153.

[9] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, Bremen Germany: ACM, Oct. 2005, pp. 625–631. doi: 10.1145/1099554.1099714.

[10] S. Rani and P. Kumar, "A Sentiment Analysis System to Improve Teaching and Learning," *Computer*, vol. 50, no. 5, pp. 36–43, May 2017, doi: 10.1109/MC.2017.133.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP'02*, Stroudsburg, United States: Association for Computational Linguistics, 2002, pp. 79–86. doi: 10.3115/1118693.1118704.

[12] S.-J. Wu, R.-D. Chiang, and Z.-H. Ji, "Development of a Chinese opinion-mining system for application to Internet online forums," *J Supercomput*, vol. 73, no. 7, pp. 2987–3001, Jul. 2017, doi: 10.1007/s11227-016-1816-6.

[13] Z. Li, L. Liu, and C. Li, "Analysis of customer satisfaction from Chinese reviews using opinion mining," in *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China: IEEE, Sep. 2015, pp. 95–99. doi: 10.1109/ICSESS.2015.7339013.

[14] C. Henriquez Miranda and J. Guzman, "A review of Sentiment Analysis in Spanish," *TECCIENCIA*, vol. 12, no. 22, pp. 35–48, Dec. 2016, doi: 10.18180/tecciencia.2017.22.5.

[15] A. Rhouati, J. Berrich, M. G. Belkasmi, and T. Bouchentouf, "Sentiment Analysis of French Tweets based on Subjective Lexicon Approach: Evaluation of the use of OpenNLP and CoreNLP Tools," *Journal of Computer Science*, vol. 14, no. 6, pp. 829–836, Jun. 2018, doi: 10.3844/jcssp.2018.829.836.

[16] N. Banik and Md. Hasan Hafizur Rahman, "Evaluation of Naïve Bayes and Support Vector Machines on Bangla Textual Movie Reviews," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sylhet: IEEE, Sep. 2018, pp. 1–6. doi: 10.1109/ICBSLP.2018.8554497.

[17] S. Rani and P. Kumar, "A sentiment analysis system for social media using machine learning techniques: Social enablement," *Digital Scholarship in the Humanities*, vol. 34, no. 3, pp. 569–581, Sep. 2019, doi: 10.1093/llc/fqy037. Available: https://academic.oup.com/dsh/article/34/3/569/5146723. [Accessed: Jun. 19, 2023]

[18] S. Thavareesan and S. Mahesan, "Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation," in *2019 14th Conference on Industrial and Information Systems (ICIIS)*, Kandy, Sri Lanka: IEEE, Dec. 2019, pp. 320–325. doi: 10.1109/ICIIS47346.2019.9063341.

[19] R. Naidu, S. K. Bharti, K. S. Babu, and R. K. Mohapatra, "Sentiment analysis using Telugu SentiWordNet," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai: IEEE, Mar. 2017, pp. 666–670. doi: 10.1109/WiSPNET.2017.8299844.

[20] D. S. Nair, J. P. Jayan, R. R. Rajeev, and E. Sherly, "SentiMa - Sentiment extraction for Malayalam," in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, New Delhi: IEEE, Sep. 2014, pp. 1719–1723. doi: 10.1109/ICACCI.2014.6968548.

[21] S. S. and P. K.V., "Sentiment analysis of Malayalam tweets using machine learning techniques," *ICT Express*, vol. 6, no. 4, pp. 300–305, Dec. 2020, doi: 10.1016/j.icte.2020.04.003.

[22] R. Das and T. D. Singh, "A multi-stage multimodal framework for sentiment analysis of Assamese in low resource setting," *Expert Systems with Applications*, vol. 204, p. 117575, Oct. 2022, doi: 10.1016/j.eswa.2022.117575.

[23] M. Gamon, "Linguistic correlates of style: authorship classification with deep linguistic analysis features," in *Proceedings of the 20th International Conference on Computational Linguistics - COLING'04*, Geneva, Switzerland: Association for Computational Linguistics, 2004, pp. 611-es. doi: 10.3115/1220355.1220443.

[24] V. S and T. S. N, "Breast Cancer Diagnosis and Classification Using Support vector machines With Diverse Datasets," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 4, pp. 442–446, Apr. 2019, doi: 10.26438/ijcse/v7i4.442446. Available: http://www.ijcseonline.org/full_paper_view.php?paper_id=4054. [Accessed: May 20, 2023]

[25] A. Kennedy and D. Inkpen, "sentiment classification of movie reviews using contextual valence shifters," *Computational Intell*, vol. 22, no. 2, pp. 110–125, May 2006, doi: 10.1111/j.1467-8640.2006.00277.x.

[26] P. De Pelsmacker, S. Van Tilburg, and C. Holthof, "Digital marketing strategies, online reviews and hotel performance," *International Journal of Hospitality Management*, vol. 72, pp. 47–55, Jun. 2018, doi: 10.1016/j.ijhm.2018.01.003.

[27] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Inf. Retrieval*, vol. 12, no. 5, pp. 526–558, Oct. 2009, doi: 10.1007/s10791-008-9070-z.

[28] S. Al-Natour and O. Turetken, "A comparative assessment of sentiment analysis and star ratings for consumer reviews," *International Journal of Information Management*, vol. 54, p. 102132, Oct. 2020, doi: 10.1016/j.ijinfomgt.2020.102132.

[29] A. C. E. S. Lima, L. N. De Castro, and J. M. Corchado, "A polarity analysis framework for Twitter messages," *Applied Mathematics and Computation*, vol. 270, pp. 756–767, Nov. 2015, doi: 10.1016/j.amc.2015.08.059.

[30] B. Le and H. Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques," in *Advanced Computational Methods for Knowledge Engineering*, H. A. Le Thi, N. T. Nguyen, and T. V. Do, Eds., Cham: Springer International Publishing, 2015, pp. 279–289. doi: 10.1007/978-3-319-17996-4_25.

[31] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, Jul. 2017, doi: 10.1016/j.eswa.2017.02.002.

[32] C. Nanda, M. Dua, and G. Nanda, "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning," in *2018 International Conference on Communication and Signal Processing (ICCSP)*, Chennai: IEEE, Apr. 2018, pp. 1069–1072. doi: 10.1109/ICCSP.2018.8524223.

[33] K. Sarkar and M. Bhowmick, "Sentiment polarity detection in Bengali tweets using multinomial Naïve Bayes and support vector machines," in *2017 IEEE Calcutta Conference (CALCON)*, Kolkata: IEEE, Dec. 2017, pp. 31–36. doi: 10.1109/CALCON.2017.8280690.

[34] H. Borkakoty, C. Dev, and A. Ganguly, "A Novel Approach to Calculate TF-IDF for Assamese Language," in *Electronic Systems and Intelligent Computing*, P. K. Mallick, P. Meher, A. Majumder, and S. K. Das, Eds., Singapore: Springer Singapore, 2020, pp. 387–393. doi: 10.1007/978-981-15-7031-5_37.

[35] R. Das and T. D. Singh, "A Step Towards Sentiment Analysis of Assamese News Articles Using Lexical Features," in *Proceedings of the International Conference on Computing and Communication Systems*, A. K. Maji, G. Saha, S. Das, S. Basu, and J. M. R. S. Tavares, Eds., Singapore: Springer Singapore, 2021, pp. 15–23. doi: 10.1007/978-981-33-4084-8_2.

[36] C. Dev, A. Ganguly, and H. Borkakoty, "Assamese VADER: A Sentiment Analysis Approach Using Modified VADER," in *2021 International Conference on Intelligent Technologies (CONIT)*, Hubli, India: IEEE, Jun. 2021, pp. 1–5. doi: 10.1109/CONIT51480.2021.9498455.

[37] J. Tolles and W. J. Meurer, "Logistic Regression: Relating Patient Characteristics to Outcomes," *JAMA*, vol. 316, no. 5, p. 533, Aug. 2016, doi: 10.1001/jama.2016.7653.

[38] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, Third edition. in Wiley series in probability and statistics, no. 398. Hoboken, New Jersey: Wiley, 2013.

[39] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[40] M. Rushdi Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. Ureña-López, "Experiments with SVM to classify opinions in different domains," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14799–14804, Nov. 2011, doi: 10.1016/j.eswa.2011.05.070.

[41] A. Naresh and P. Venkata Krishna, "An efficient approach for sentiment analysis using machine learning algorithm," *Evol. Intel.*, vol. 14, no. 2, pp. 725–731, Jun. 2021, doi: 10.1007/s12065-020-00429-1.

[42] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117–126, Sep. 2016, doi: 10.1016/j.eswa.2016.03.028.

[43] 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)*, Dhaka, Bangladesh: IEEE, May 2019, pp. 1–6. doi: 10.1109/ICASERT.2019.8934655.

[44] A. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," in *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, San Antonio Texas: ACM, Oct. 2012, pp. 1–7. doi: 10.1145/2401603.2401605.