Impact of Features on Sentiment Analysis of Bangla E-Commerce Products using Machine Learning

Tokey Ahmmed, Ipshita Tasnim Raha, Nakib Aman Turzo

Abstract— In Bangladesh, E-commerce is gaining more popularity day by day, especially during the COVID 19 pandemic. Among many e-commerce sites in Bangladesh, Daraz is one of the most popular e-commerce sites among general people. But biased reviews and comments often make it challenging to choose the best product. In this paper, we will focus on finding the most significant features for classifying the positive and negative reviews. To achieve this goal, we have used various machine learning classifiers using Python. Data cleaning was done and afterward, Term Frequency-Inverse Data Frequency (TF-IDF) was applied. Among the classifier used Extreme Gradient Boosting algorithm achieved the highest accuracy followed by the Extra Trees Logistic Regression algorithm. It was found that most classifiers predicted the same features as most significant.

Index Terms— Bangla Text mining, E-Commerce review, Machine learning, NLP, Sentiment analysis, Feature selection

1 Introduction

Sentiment analysis is a technique of classifying texts based on the sentiment orientation of opinions they contain. It is a part of Natural Language Processing. The main purpose of sentiment analysis is to detect the contextual polarity of the text which can be positive, negative, or neutral. Since these types of text can reflect the opinion of the user sometimes it is called as opinion mining. [1]

As a subfield of Natural Language Processing (NLP) the sentiment analysis has drawn the interest of many researchers. The main reason is the vast amount of data on the internet. Nowadays people in social media sites, newspapers, blogs, etc., express their reviews on a specific product or item. There are also forum discussions, opinions on specific posts, comments, and emotions. There are many observative opinions in them, those can be classified into binary classes such as positive or negative opinions. To analyze the sentiment lexicon-based dictionary approach can be used to determine the positive or negative polarity of the sentiments. But to analyze this huge amount of data using machine learning techniques has brought the significant interest of many researchers. The reason is these types of machine learning models can consider the huge amount of versatile features for predicting the positive-negative polarity. Various feature extraction process such as Bag of Word (BoW), term frequency-inverse document frequency (TF-IDF) is applied to the text to convert it into feature vector for the machine learning algorithms. Since the importance of these extracted features may vary in these different types of machine learning algorithms, we will be working on finding the important features for various machine learning algorithms and comparing them. [2]

Before buying a product online, a consumer usually wants to make a decision about the product, its service, price and etc. For this, they may have to go through a very large number of user reviews. But reading and analyzing all of them is a difficult task. Also, the organizations are interested in mining the reviewers' opinions so that they can gain benefit by identifying new opportunities, predicting sales trends, and so on. But it needs to deal with an overwhelming number of available customer comments. With the sentiment analysis techniques, opinions can easily be extracted by analyzing this huge amount of available data which may help both customers and organizations to achieve their goals. [8]

Sentiment analysis can be done by 2 different types of approaches.

i. Lexicon based approaches
ii. Machine learning approaches

Lexicon Based Approach: Lexicon Based techniques work on an assumption that the collective polarity of a sentence or document is the sum of polarities of the individual phrases or words. This method is based on emotional research for sentiment analysis dictionaries for each domain. [1]

Machine Learning Approach: Sentiment analysis using machine learning strategies works by training an algorithm with a

- *Tokey Ahmmed is currently working as a lecturer in the department of Computer Science & Engineering in Varendra University, Bangladesh, PH-+8801717099321. E-mail: tokey.ahmmed@gmail.com*
- *Ipshita Tasnim Raha is currently working as a lecturer in the department of Computer Science & Engineering in Varendra University, Bangladesh, PH-+8801715673436. E-mail: ipshita032@gmail.com*
- *Nakib Aman Turzo is currently working as a research coordinator in National Computer Training and Research Academy, Bangladesh, PH-+8801762910933. E-mail: nakibaman@gmail.com*

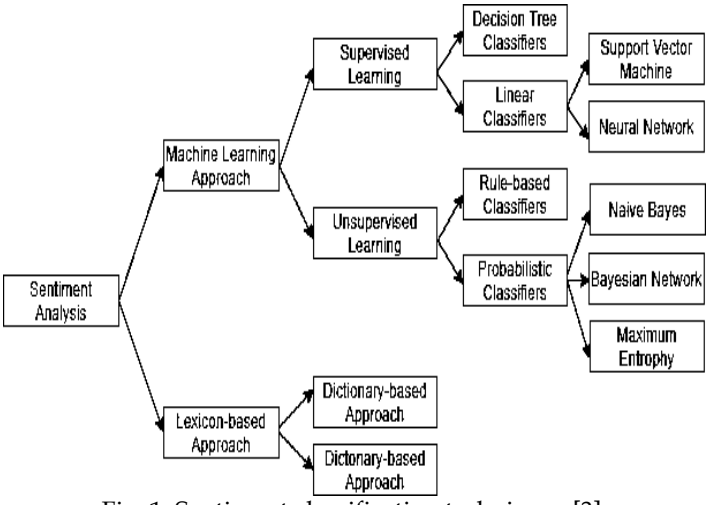

Fig. 1. Sentiment classification techniques [3].

training data set before applying it to the actual data set. Machine learning techniques first train the algorithm with some particular inputs with known outputs so that later it can work with new unknown data. [2]

2 Literature Review

A large number of people have focused on solving the problem of sentiment classification. They used different techniques for classifying the sentiments of individuals at various levels.
In 2016 Devika M D, Sunitha C, Amal Ganesh published a paper showing a comparative study on various approaches of sentiment analysis. The paper shows a detailed comparison between three main categories of classification approaches (machine learning approach, rule-based approach and lexicon-based approach) with both advantages and disadvantages. According to their research, the machine learning approach gives the highest accuracy but the result may mostly depend on trained data. On the other hand, the rule-based approach fully depends on how the rules are being set. Whereas the unsupervised lexicon-based may not need any labeled data but it requires a lot of linguistic resources which may not be always available. [1]

In 2021 Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, M. Rubaiyat Hossain Mondal, M. S. Islam published a paper in which they proposed an extended lexicon-based supervised machine learning approach to analyze sentiment from the Bangla text. They developed a rule-based algorithm termed as Bangla Text Sentiment Score (BTSC) for detecting sentence polarity. This algorithm can generate a score for each word thus of a whole sentence. After generating the BTSE score and then applying the vectorization technique to populate the feature matrix the data set was applied to the supervised classifier to predict the outcome of the sentences. According to their result, they have found the best result for the bigram result with the support vector machine classification algorithms. [2]

In 2020 Muhammad Marong, Nowshath K Batcha, Raheem Mafas, published a paper on reviewing sentiment analysis techniques and algorithms in E-commerce data. The paper was mentioned with various types of approaches such as lexicon-based approaches and machine learning-based approaches. Lexicon usually works on two types of approaches dictionary-based approach and the corpus-based approach. On the other hand machine learning, approaches can be classified into two broad categories supervised learning and unsupervised learning. These supervised learning can be computed in various ways. such as linear classifiers, probabilistic classifiers, rule-based classifiers, and decision tree classifiers. [3]

In 2019 another work has been done by Ravinder Ahujaa, Aakarsha Chuga, Shruti Kohlia, Shaurya Guptaa, and Pratyush Ahuja. In that paper, they experimented on the different types of feature extraction process before applying into the machine learning process. They have experimented with Term Frequency-Inverse Document Frequency (TF-IDF) and N-Gram feature extraction techniques on twitter data. After extracting the features they analyze it using six machine learning classifier algorithms (Decision Tree, Support Vector Machine, K-Nearest Neighbour, Random Forest, Logistic Regression, Naïve Bayes). They have found that the performance of sentiment analysis using the TF-IDF feature extraction method is 3-4% higher than other methods. [4]

In 2019 a comparison paper has been published by Monika Kabir, Mir Md. Jahangir Kabir, Shuxiang Xu & Bodrunnessa Badhon. The research was about comparing the performance of machine learning approaches (Maximum entropy, SVM, Bagging, Boosting, Random forest, Decision tree) in sentiment classification. They have work around 3000 instances of data from amazon, yelp and IMBD websites reviews. According to their experiments, they have found that maximum entropy has the overall highest performance. They have also mentioned all the machine learning approaches nearly achieved 80% accuracy for all kinds of datasets, which gives the idea that machine learning approaches are capable of the most feasible outcome in sentiment analysis techniques. [5]

In 2020 another E-commerce review paper by M. J. Ferdous, P. Sarker and N. A. Turzo. In this paper, the various machine learning approaches (Ridge Classifier, Logistic Regression, SVM – Linear Kernel, Extreme Gradient Boosting, CatBoost Classifier, Light Gradient Boosting Machine, Gradient Boosting Classifier, Extra Trees Classifier, K Neighbors Classifier, Random Forest Classifier) was used to classify and compare the performance of the algorithms with Bangla e-commerce reviews. According to the paper, ridge classification gave the highest accuracy followed by the logistic regression and SVM classifiers. [6]

3 METHODOLOGY

The dataset is made of bangla text data which are collected from the website of Daraz. It contents a total 900 e-commerce reviews those were scraped from this website. Among them, 593 positive comments were labeled as 0's and 307 negative comments were labeled as 1's.

খুব খারাপ কাপড়	1
অনেক সুন্দর হয়েছে কোয়ালিটি ভালো	0
ফাউল	1
কাপড়ের ডিজাইন ঠিক নাই। দুইটি আলাদা ডিজাইনের। কোয়ালিটি ও খুব ভালো না.....😔	1
দাম হিসাবে ভালই। কাপড়ের কোয়ালিটি ভালই।	0
ভাল লেগেছে, thanks	0
দাম হিসেবে কাপড়ের মান এভারেজ	0

Fig. 2. Review samples with label.

TABLE 1
SUMMARY OF USED DATA

Data Source	Category	No. of reviews (positive/negative)
Daraz	Multiple types (Mobile, Camera, Dress etc.)	900 (593/307)

Sentiment analysis process can be done using both lexicon based approch or machine learning approches. But whole process can be is performed in several steps. Each of its steps crutial and can be done differently. These steps include:

1) Text cleaning,
2) Tokenization,
3) Stop word removal,
4) Text to vector transformation,
5) Feature selection, and finally
6) Classification.

3.1 Text cleaning

Text cleaning is the first step where the basic cleaning of text data is done. Such as removing any unwanted symbol, emoji, URL, etc. Case transformation is also done in this stage to avoid the redundancy of the words. We have used python's Natural Language Toolkit (NLTK) for this purpose.

3.2 Tokenization

Tokenization is the process to divide textual information into individual words. Generally, the initial data for sentiment analysis is just a set of characters. But to process these using classification algorithms the machine needs to analyze the word individually. Tokenization is responsible for this action. Various open-source tokenization tools such as text blob, NLTK word tokenizer, Nlpdotnet tokenize can be used for this step. [7]

3.3 Stop word removal

Stop word refers to the common words that usually don't have or have very low value for text analysis. This usually includes conjunction, preposition, pronoun, etc [5]. Some customized keywords also can consider as stop words which entirely depend on the language. Since these stop words don't put any value removing stop words is important for sentiment analysis. It can be done by checking the tokens with the stop word library of the specific language. We have used the Bangla stop word list and python's NLTK to complete this step.

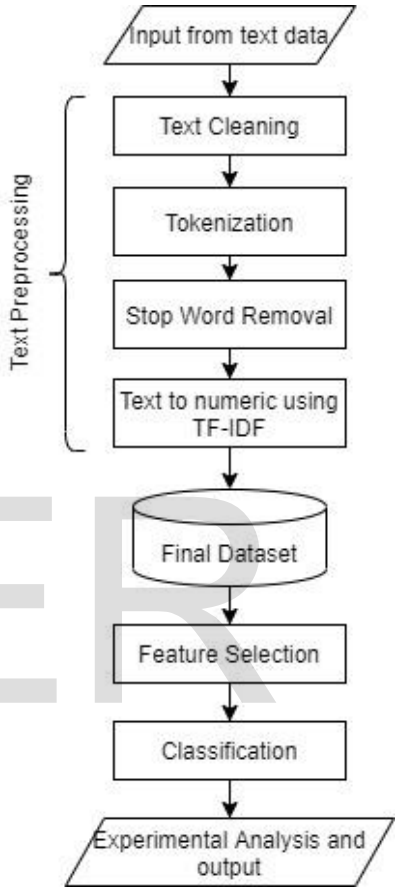

Fig. 3. Methodological workflow

3.4 Text to vector Transformation

Since none of the machine learning classifiers can't direct process textual data the text needs to be converted into the vector form. For further processing of the text, it is converted into some types of numerical vector format [6]. There are several techniques for the process. In our case, we will be working with Term Frequency-Inverse Document Frequency (TF-IDF) since it works relatively well due to its performance and computational time.

Term Frequency (TF): The TF is simply the frequency of a word divided by the gross number of words in the document.

$$tf_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{i,j}} \quad (1)$$

Inverse Document Frequency (IDF):
IDF is the log of the documents number divided by word 'w' containing documents. IDF determines the weight of rare words across all documents in the corpus.

$$idf(w) = \log(\frac{N}{df_i}) \quad (2)$$

The TF-IDF is simply the TF multiplied by IDF

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (3)$$

where,

$tf_{i,j}$= number of occurrences of i in j
df_i = number of documents containing
N= total number of documents

3.5 Feature Selection

In this feature selection process, important features are selected. It is done so that the features that contribute the most in predicting the target variable can be distinguished for further processing. Working with selected features instead of all the features reduces the risk of over-fitting, improves accuracy, and decreases the training time. We are using the feature selection parameter from the PyCaret library to achieve this. It uses a combination of several supervised feature selection techniques to select the subset of features that are most important for modeling.

3.6 Classification

Machine learning techniques tend to be the most preferred for sentiment analysis. These machine learning algorithms provide better accuracy, speed and capacity for classifying huge amounts of data rather than traditional techniques. For our classification problem, we will be using given machine learning classifiers.

3.6.1 Ridge Classifier

The Ridge Classifier is basically based on the Ridge regression method. It converts the label data into [0, 1] and solves the problem with the regression method. The highest value in pre-diction is accepted as a target class.

3.6.2 Logistic Regression

Logistic Regression is a classification algorithm that is used to predict a binary outcome (1 /0) given a set of independent variables. It can calculate as given logit function:

$$p = \frac{e^y}{1+e^y} \quad (4)$$

3.6.3 Support Vector Machine

Support Vector Machine (SVM) is one of the dominant machine learning classification algorithms for solving multiclass classification problems from ultra large data sets [5]. In sentiment analysis usually, SVM performs better than any other machine learning classifiers. It works by constructing a separator to separate the input data into output classes. The separator is constructed on the hyperplane based on the training data [9]. The linear kernel SVM can be defined as follows:

$$yi(w.xi + b) \geq 1 \text{ for all } 1 \leq i \leq n \quad (5)$$

where,

x = input vector,
w = weight vector w and
b = bias

3.6.4 Boosting

Boosting classifier is an ensemble machine learning algorithm that combines a number of weak classifiers to creates a strong classifier. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model. [10]
The main boosting formula can be defineds as follows:

$$f(x) = f_0(x) + \sum_{n=1}^{N} \theta_n \emptyset_n(x) \quad (6)$$

where,

f_0 = initial predicted value,
θ_n = represent weight for nth iteration, and
$\emptyset_n(x)$ = base estimator

There are several types of boosting algorithms such as gradient boosting, extreme & light gradient boosting, catBoosting classifier, etc.

3.6.5 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is another popular implementation of gradient boosting. XGBoost uses a more regularized model formulization to control overfit-ting. For this, it provides better accuracy. Rather than other boosting algorithms extreme gradient boosting is also chosen for its performance and speed.

3.6.6 K Neighbors Classifier

K-Nearest Neighbor (KNN) is one of the simplest supervised algorithms that classifying objects based on learning data that are closest to the object [11]. The KNN can predict the test object x_0 either belong to class j or not, by overserving k-nearest neighbor as follows:

$$\Pr(Y = j|X = x_0) = \frac{1}{k}\sum_{i \in N_0} I(y_i = j) \quad (7)$$

where,

X = Matrix of features
Y = Class label
N_0 = Set of k -nearest observations

3.6.7 Random Forest Classifier

Random forest classifier simple supervised algorithm that creates a set of decision trees from a randomly selected subset of the training set. Then it averages the prediction value of its trees is used for the final prediction. It can be defined as follows,

$$F(x) = \frac{1}{j}\sum_{j=1}^{J} f_j(x) \qquad (8)$$

3.6.8 Extra Trees Classifier

Extra Trees Classifier also known as Extremely Randomized Trees is another type of ensemble learning technique that collects the results of multiple randomized decision trees on various sub-samples of the dataset and uses averaging for predictive accuracy and control over-fitting.

4 Experimental Results

Each machine learning algorithm used for the classification works in a different manner. But all of them used a set of training data to build a predictive model to classify new data. The preset values that were set before classification are shown in Table 2.

TABLE 2
PRESETS VALUES BEFORE CLASSIFICATION

Preset Description	Value
Target Type	Binary
Numeric Imputer	Mean
Categorical Imputer	Constant
Normalize	False

During training with various features, some features can be very important than other features. The following charts show the top important features for the various classification algorithms.

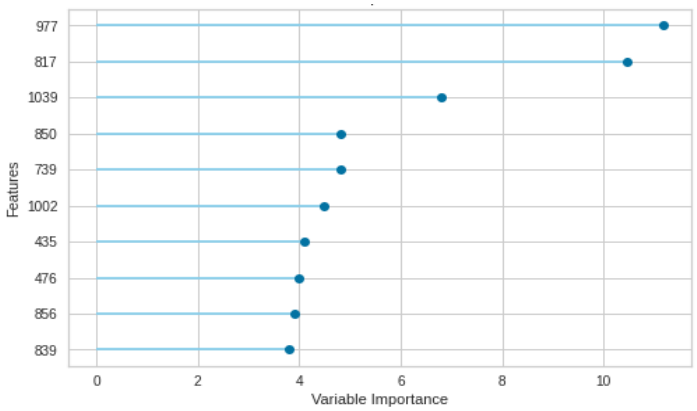

Fig 4. Feature Importance Plot of Support Vector Machine.

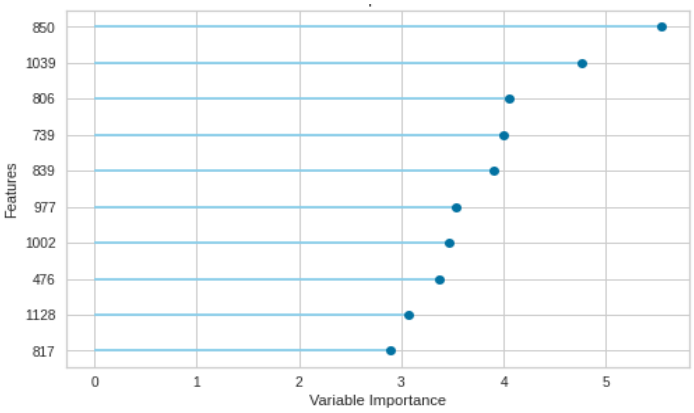

Fig. 5. Feature Importance Plot of Logistic Regression.

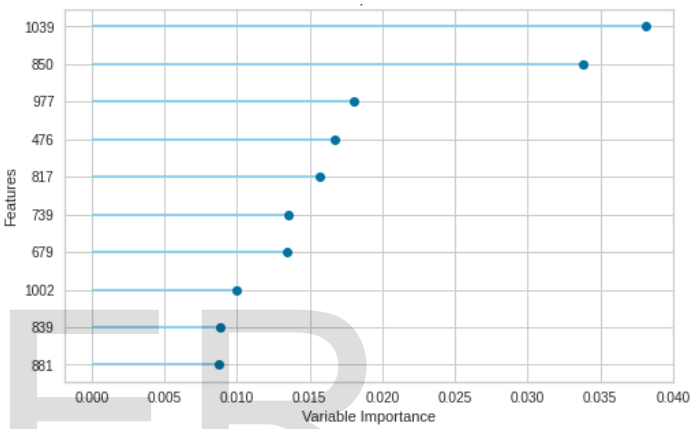

Fig. 6. Feature Importance Plot of Extra Trees classifier.

From the above result of our experiment we can conclude that 850th number feature has the most important role for classifying sentiment using Logistic Regression, Ridge classifier, XGBoost classifier and has the 2nd highest important role in Extra Trees classifier. Furthermore, 1039th feature has the position in the importance plot of all the classifiers algorithms. So we can say that all the classifier algorithms have a similar reaction for the important features of the final dataset.

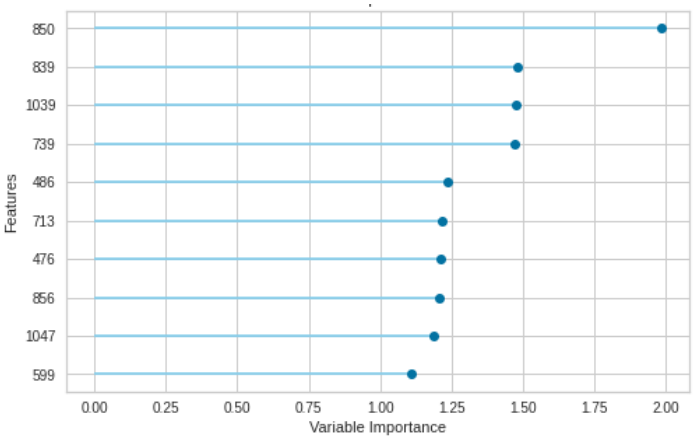

Fig. 7. Feature Importance Plot Ridge classifier.

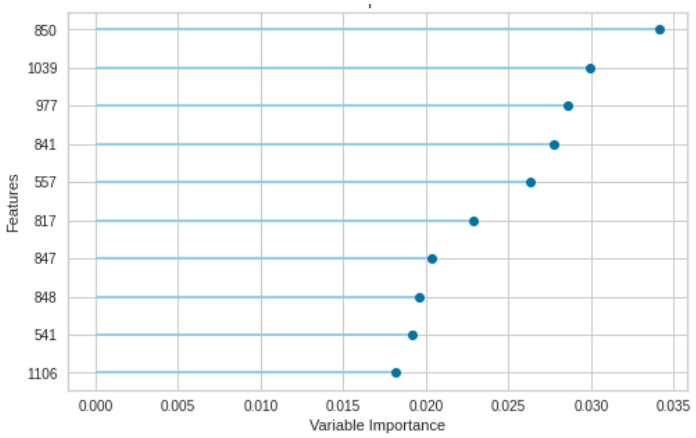

Fig. 8. Feature Importance Plot Extreme Gradient Boost classifier.

TABLE 3
SUMMARY OF VARIOUS CLASSIFIERS

Classifier Name	Accuracy	AUC	Recall	Prec.
Support Vector Machine	0.6980	0.6512	**0.5182**	0.6374
Logistic Regression	0.7187	**0.7913**	0.2426	0.7827
Extra Trees	0.7376	0.7735	0.4208	0.6953
Ridge Classifier	0.7154	0.7423	0.2426	**0.7827**
Extreme Gradient Boosting	**0.7569**	0.7862	0.4344	0.7540

The results in Table 3 shows the Accuracy, AUC, Recall and Precision of Support Vector Machine, Logistic Regression, Extra Trees, Ridge Classifier and Extreme Gradient Boosting algorithms. In our experiments, the Extreme Gradient Boosting achieved the highest accuracy of 75.69% followed by Extra Trees 73.76% and Logistic Regression 71.87%. The difference is due to the characteristics of different datasets and processes.

5 CONCLUSION

In this study, an attempt was made to classify sentiment from several customer reviews of Bangla text. From the customer reviews, the positive and negative sentiments were classified using Support Vector Machine, Logistic Regression, Extra Trees, Ridge Classifier and Extreme Gradient Boosting algorithm. From them, Extreme Gradient Boosting classified the sentiment with the highest accuracy. We have also analyzed the output feature importance for each machine learning algorithm. Based on our result we can conclude all the important features were identified almost identically by most of the machine learning algorithms.

In the future, we plan to experiment with more datasets and apply the proposed feature selection strategy into revising other existing feature selection metrics. We also consider exploring how to better determine the importance of a feature in realtime to give customers feedback from live chat features.

ACKNOWLEDGEMENT

The research was supported by the National Academy of Computer Training and Research (NACTAR), Bogra, Bangladesh.

REFERENCES

[1] J M.D. Devika, C. Sunitha, Amal Ganesh, Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science, Volume 87, 2016, Pages 44-49, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.05.124.

[2] Nitish Ranjan Bhowmik Mohammad Arifuzzaman M. Rubaiyat Hossain Mondal M. S. Islam, Bangla Text Sentiment Analysis Using Supervised Machine Learning with Extended Lexicon Dictionary Natural Language Processing Research Volume 1, 2021, Pages 34-45, ISSN 2666-0512, https://doi.org/10.2991/nlpr.d.210316.001.

[3] Marong, Muhammad & Batcha, Nowshath & Raheem, Mafas. (2020). Sentiment Analysis in E-Commerce: A Review on The Techniques and Algorithms. 6.

[4] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, Pratyush Ahuja, The Impact of Features Extraction on the Sentiment Analysis, Procedia Computer Science, Volume 152, 2019, Pages 341-348, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.05.008.

[5] Monika Kabir, Mir Md. Jahangir Kabir, Shuxiang Xu & Bodrunnessa Badhon (2019): An empirical research on sentiment analysis using machine learning approaches, International Journal of Computers and Applications, DOI: 10.1080/1206212X.2019.1643584

[6] M. J. Ferdous, P. Sarker and N. A. Turzo, "Assortment of Bangladeshi E-commerce Site Reviews using Machine Learning Approaches," 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, pp. 1-6, DOI: 10.1109/STI50764.2020.9350526.

[7] Vijayarani, S & Janani, Ms. (2016). Text Mining: open Source Tokenization Tools – An Analysis. Advanced Computational Intelligence: An International Journal (ACII). 3. 10.5121/acii.2016.3104.

[8] Ranga, S., & Raghavendra, S. N. (2018). E-commerce Product Review Analysis using Data Analytics. International Journal of Pure and Applied Mathematics, 120(6), 65-73.

[9] Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), 1138–1152. DOI:10.1016/j.ins.2010.11.023

[10] Kaur, P., & Gurm, R.K. (2016). Design and Implementation of Boosting Classification Algorithm for Sentiment Analysis on Newspaper Articles.

[11] Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018). Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes. 2018 International Conference on Orange Technologies (ICOT). DOI:10.1109/icot.2018.8705796