Sentiment Analysis of Bengali Texts on Online Restaurant Reviews Using Multinomial Naïve Bayes

Omar Sharif[1], Mohammed Moshiul Hoque[2] and Eftekhar Hossain[3]
Department of Computer Science and Engineering[1,2]
Department of Electronics and Telecommunication Engineering[3]
Chittagong University of Engineering and Technology[1,2,3]
Chittagong-4349, Bangladesh
omar.sharif1303@gmail.com[1], mmoshiulh@gmail.com[2], eftekhar.13ete@gmail.com[3]

***Abstract*—Recently, determining the customer impression is considered one of the prominent factors on the success of the restaurant businesses. Due to the rapid growth of digital contents related to restaurant or foods in the web, people are more inclined on reviews before going to any restaurant so the significance of customer review is inevitable. In order to selects a restaurant customer needs to check thousands of feedback's to understand the restaurant quality or services. Therefore, classification of a significant amount of reviews into a sentimental category is required to attain meaningful insights so that the customer can choose restaurants based on their preferences. This classification can be done by sentiment analysis. This paper proposes a system that can classify customer reviews into positive and negative classes based on their sentimental feedback. We have tested the proposed system with 1000 restaurant reviews text written in Bengali. The experimental result shows that the proposed the system can classify restaurant reviews with 80.48% accuracy using multinomial Naïve Bayes.**

***Keywords*—Bangla language processing, Sentiment analysis, Customer reviews, Machine learning.**

I. INTRODUCTION

The amount of people has increased significantly who are taking their foods or meals in restaurants due to rise their financial capability, miscellaneous food habits and improved lifestyle. Therefore, restaurant business have achieved remarkable popularity in recent years. Due to the rapid growth of Internet contents, customers has affinity to check the reviews of a restaurant before visit it. Thus, reviewing a restaurant from text comments become a common scenario.

The key objective of this paper is to build a supervised learning model to categorize the customer's feedback's in terms of positive and negative sentiments which are written in Bengali sentences. Any restaurant with plenty of positive reviews can attain a symbol of faith, ensure quality of foods and services among the customers. Conversely, it is very difficult to attract new customers as well as retain existing customers by a restaurant without positive reviews or feedback. Usually, more positive reviews conveys trustworthiness to customers. Thus, people appreciate the experience of the consumer and consent to others and the review on a restaurant is the only way to comprehend others opinion on the restaurant. Opinions collected from users' experiences regarding specific criteria of a restaurant like food quality, environment and service quality, straightforwardly make an impact on the future customer decisions. Likewise, negative reviews often yield disinclination on the customer choices. Therefore, restaurants needs to take into account the customer's feedback for future improvements and attracts much attention of customers. This might totally unusual to check every customer reviews one by one as it requires ample amount of money and innumerable manpower to lead such a surveys. Taking consideration the rapid growth of visitors and users requirements, this imposes a concealment for a system that can understand the attitude of a reviewer comments written in various social media communities or blogs. In that case, a term sentiment analysis come uppermost that aim is to uncover how customer sense about a certain foods.

The sentiment analysis is used to attain a grasp of sentimental note behind the person's reviews. It is also known as opinion mining and it determines whether the sentiment of a person comment is positive, neutral or negative based on their writing. Many researches in sentiment analysis have been done in many languages like Chinese [1][2], Vietnamese [3], Arabic [4], Thai [5], Myanmar [6] and Bengali [7]. Research on Bengali text based opinion mining of restaurant reviews is in rudimentary stage still now. This paper presents an automated system for sentiment analysis of Bengali text for restaurant reviews to classify an opinion into two classes: positive or negative sentiment. Three machine learning algorithms such as decision tree, random forest and multinomial naïve Bayes classifier are implemented to classify the sentiment from Bengali review text.

II. LITERATURE REVIEW

The sentiment analysis of customer reviews has become one of the interested research area. Some of the research work by noted scholars on restaurant reviews on different languages are studied and discussed below. Aye et.al developed a lexicon-based sentiment analysis model in Myanmar text. On their work, the customer reviews on food and restaurant quality was the prime insight for the sentiment analysis. The sentiment classification was done by using a sentiment dictionary which contains a list of sentiment words. Their system successfully distinguished data for the reviews on Myanmar Language at

978-1-7281-3445-1/19/$31.00

96% overall accuracy over 500 customers' reviews on food and restaurant domain [6].

Kaviya et.al developed a sentiment analysis model where the rating system of a restaurant was done depending upon the user reviews. Their system splits user comments to find sentiment keywords and associates the comment with a sentiment rank. The reviews were categorized into positive and negative class and an overall rating was assigned to the restaurant. The Yelp dataset was taken for the testing of their developed system.

Perera et.al adopted an aspect-based opinion mining for restaurant reviews. The aspect-based mining finds a sentence aspect and the opinion of the user's comments on these aspects to make a positive and negative review. SentiWordNet a lexical dictionary was used to determine the positive and negative polarity value for each opinion word. Dependency parser was used for the opinion word extraction. The evaluation of their developed system was done on both manually and systematically. From the comparison of the two approaches they found that their developed system provided a satisfactory accuracy of 70% [8].

Doan et.al developed a sentiment analysis model for customers' reviews where a variant of online random forests was taken to perform the analysis. Their proposed model was an ensemble of tree-based classifiers and a portion of the restaurant reviews from the Yelp dataset was taken for the sentiment analysis. For the experiment, they set up an ensemble learning model for 100 trees using Online Random Forests (ORF) and incremental random forest (proposed approach) (iRF). Their model has longer run time than iNB because incremental learning was adopted in their design that induces trees instance by instance [9].

Sarkar et.al developed a Bangla sentiment analysis model where Bengali tweet dataset released for SAIL contest 2015 was used for classification. They used Multinomial Naïve Bayes and SVM classifiers for the classification of sentiment polarity expressed in tweets. N-gram and SentiWordnet features were used as their feature set.Their devleloped model showed that SVM classifier trained with unigram and SentiWordNet features performs best on Bengali tweet dataset [10]. Gan et.al designed a model for restaurant rating based on sentiment analysis. In their work, in addition to food, ambience and service, the overall rating of the restaurant also focuses on the pricing and special contexts but there was no emphasizing done on the scores [11]. Lak et.al developed a model to compare sentiment analysis results with star ratings in three different domains [12]. Xing Fang et.al developed a system for sentiment analysis of product reviews which used the sentiment polarity categorization process. They selected naïve bayesian, support vector machine and random forest classification methods for categorization. For the sentiment polarity classification a sentiment score formula was taken to generate the feature vector for a sentiment and based on this score they performed the sentiment polarity categorization [13]. Lada Banic et. al designed a hotel review mining system using natural language processing and machine learning. Number of negative and positive terms or phrases are used for the generation of cumulative information on the level of final evaluation. A grade from 1 to 5 in which 1 and 5 is referred as bad and excellent respectively was assigned in final evaluation[14]. Singla et.al done sentiment analysis on unstructured data of Mobile phone reviews. They used decision tree, naïve Bayes, SVM, and cross validation for the purpose of classification. Among these classification models they concluded that svm provides highest accuracy in opinion mining of product reviews [15].

III. Dataset Accumulation

Success of any machine learning system heavily depends on the dataset used to train that system. But unfortunately, it is really difficult to build a corpus which contains a large amount of Bangla text reviews on restaurant because there is lack of resources in Bengali language as well as researchers do not publish their dataset. For the purpose of research, Bengali benchmark dataset is not yet published. So an English benchmark dataset have taken [12] and translated into Bengali to build the corpus. Fig.1 shows sample positive and negative restaurant reviews.

	Sample Reviews	Sentiment
English	Preferably hot and spicy chicken fry. Excellent environment and hospitality. I'll give it a rating in the top restaurant in Dhaka.	Positive
Corresponding Bengali	গরম এবং মসলাযুক্ত চিকেন ফ্রাইটা পছন্দনীয়। চমৎকার পরিবেশ এবং আতিথেয়তা। আমি এটিকে ঢাকায় শীর্ষস্থানীয় রেস্তোরাঁর মধ্যে রেটিং দিব।	
English	The grill was cold, the bread was dry and strong, the stock was old ... very bad food ... frustrating ...	Negative
Corresponding Bengali	গ্রিলটা ঠান্ডা ছিল, রুটিও শুকনো এবং শক্ত ছিল, স্টকটা পুরাতন ছিল ... খুবই খারাপ মানের খাবার ... হতাশ...	

Fig. 1. Dataset Sample

Besides, different public pages [16], groups [17][18][19] of Facebook are used to collect data. As most part of the dataset collected manually some negligible inconsistencies may occur in it. Collected text documents are used to train the system which will be used for further classification of reviews. Dataset statistics is summarized in Table I.

TABLE I: Data Summary

	Training set	Testing set
Number of documents	800	200
Number of sentences	2559	628
Number of words	8236	2497
Total unique words	1069	358

IV. Proposed System Framework

Fig. 2 shows an abstract view of proposed system. Different statistical machine learning techniques are used to train the system. The overall framework is divided into three phases: training set preparation, feature extraction and classification respectively.

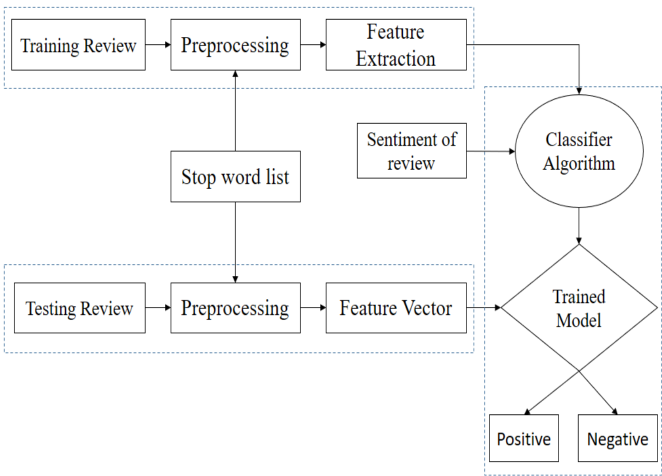


Fig. 2. Proposed System Framework

A. *Training Set Preparation*

Training set $R = \{r_1, r_2, r_3, ..., r_n\}$ consists of n training reviews. Each review either contains positive of negative sentiment where C_p and C_n are used to denote positive and negative class respectively. A review r_i with l words is represented by a word vector $W[] = \{w_1, w_2, w_3, ..., w_l\}$. All reviews are preprocessed to remove inconsistencies from dataset. Punctuation's are also removed in this step. Table II represents word vector$W[i]$ of a random review r_i.

TABLE II: Word vector of a review

r_i	w_1	w_2	w_3	...	w_l

A list $S[] = \{s_1, s_2, s_3, ..., s_t\}$ with t stopwords have been developed to remove the words w_i which has no contribution in deciding whether a review r_i conveys positive (C_p) or negative (C_n) sentiment. Stop words $s_1, s_2, s_3, ..., s_t$ are removed from a review by finding the match from stop word list $S[]$. Conjunctions, prepositions, interjections, pronouns, suffixes and prefixes are considered as stop words.

	Type	Example
s_1	pronoun	সে
s_2	conjunction	এবং
s_3	preposition	থেকে
...	...	...
s_r	interjection	সাবাশ

Fig. 3. Sample stop words

B. *Feature Extraction*

A vocabulary of Bengali words is created by tokenizing reviews of the corpus. "$tfidf$" : the term frequency-inverse document frequency statistic is used as features of the proposed model. Our used "$tfidf$" statistic is,

$$tfidf(w, r) = tf(w, r) \log \frac{N}{|\{r \epsilon R : w \epsilon r\}|} \quad (1)$$

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

TABLE III: A small fragment of feature matrix

r \ c	w_1	w_2	w_3	w_4	...	w_v
r_1	0.21	0.13	0.20	0.16	...	0.10
r_2	0.18	0.36	0.17	0.21	...	0.31
r_3	0.17	0.13	0.14	0.15	...	0.19
r_4	0.24	0.40	0.48	0.47	...	0.41
...	...	...	...	...	...	...
r_u	0.25	0.32	0.46	0.71	...	0.48

Table III shows the feature matrix of proposed system. Features $(F[][])$ of the reviews are represented by a two dimensional$(U \times V)$ feature matrix. Here, rows represent the reviews $r_1, r_2, r_3, ..., r_u$ and columns represents unique words $w_1, w_2, w_3, ..., w_v$ respectively. Each cell of the feature matrix $(F[][])$ represent the "$tfidf$" value of word w_v occurs in a specific review r_i. Each row represents the feature vectors $F[1], F[2], F[3], ..., F[n]$ for the reviews $r_1, r_2, r_3, ..., r_n$ of the corpus.

C. *Classification*

As, the goal of the proposed system is to classify the reviews thus classification is the most significant part of the system. Extracted features from the reviews are used to train the proposed model that could classifies reviews into positive and negative sentiment. Mainly, the proposed system is implemented by using multinomial naïve bayes.

Multinomial naïve bayes [20] performs really well when multiple occurrences of the words have significant impact in the classification problem. It can be defined by bayes theorem where all the variables $V_1, V_2, V_3, ..., V_n$ in a given class C are conditionally independent with each other given C. Using bayes rule for a text review (R) and class (C),

$$P(R|T) = \frac{P(R|C)P(C)}{P(R)} \quad (2)$$

Final equation,

$$C_{MAP} = argmaxP(R_1, R_2, ..., R_n|C)P(C) \quad (3)$$

A **decision tree** [21] has external and internal nodes. Decision classes are represented by external nodes and internal nodes correspond to attribute which are used for making decision. Entropy calculation is used to build a decision tree,

$$E(S) = \sum_{c \epsilon X} P(C)E(C) \tag{4}$$

Random forest is mainly collection of decision trees with some differences. Decision trees make predictions by formulating some set of rules but in random forest algorithm decision trees are built by selecting features randomly and finally acquired results are averaged. But for real time predictions a large number of trees can make this algorithm slow and ineffective.

Trained classifier model is tested with a test set, $TR = \{tr_1, tr_2, tr_3, ..., tr_x\}$ with x test documents. Feature extraction method is used to extract features $(F[][] = F[1], F[2], ..., F[x])$ from the test set TR. Proposed model use these features to classify a review tr_i as positive(C_p) or negative (C_n).

V. Evaluation Measures

A number of statistical and graphical measures have been adopted for evaluating the performance of the proposed system. Statistical measures such as confusion matrix, precision, recall, f_1 score and graphical measures such as precision-recall curve, receiver operating characteristics (ROC) curve have used for the measurement.

- A **confusion matrix** is used for evaluating the performance of a classification model. As the proposed system is designed for binary classification, thus the confusion matrix of the system must have two columns and two rows. This matrix indicates the number of true positives, false positives, true negatives and false negatives.
  - True Positive (TP): Reviews that conveys positive sentiment and also classified as positive.
  - True Negative (TN): Reviews that conveys negative sentiment and also classified as negative.
  - False Negative (FN): Reviews that conveys positive sentiment but classified as negative.
  - False Positive (FP): Reviews that conveys negative sentiment but classified as positive.

  This rates are used to calculate other evaluation measures.
- **Precision** refers as positive predictive value. Precision can be obtained form the following equation.

  $$Precision = \frac{TP}{FP + TP} \tag{5}$$

  If precision is high then the algorithm is doing well.
- **Recall** is the ratio of correctly classified positive reviews to the total number of positive reviews. Recall can be obtained from the following equation.

  $$Recall = \frac{TP}{TP + FN} \tag{6}$$

  High value of precision and Recall are vital for a model.
- f_1**-score** of algorithms need to be calculated for choose a particular learning algorithm between several algorithms. f_1-score can be obtained from the following equation,

  $$f_1 - score = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

VI. Results and Analysis

Proposed model has been tested with random forest, decision tree and multinomial naïve bayes algorithms. K-fold cross validation has been used to test the performance of the system. Table IV presents a comparison of performance of this algorithms for different values of K.

TABLE IV: Performance Comparison

K-value	Classifier	Precision	Recall	f_1 score	Accuracy
5-fold	Random Forest	0.75	0.76	0.75	75.51
	Decision Tree	0.71	0.70	0.71	70.40
	Proposed	0.81	0.80	0.79	79.59
6-fold	Random Forest	0.79	0.78	0.77	78.04
	Decision Tree	0.70	0.70	0.70	69.51
	Proposed	0.82	0.80	0.80	**80.48**
7-fold	Random Forest	0.70	0.70	0.70	70.00
	Decision Tree	0.71	0.72	0.71	71.42
	Proposed	0.80	0.79	0.78	78.57

Same number of training and test reviews used for all algorithms. Table IV shows that multinomial naïve byes with ”tfidf” feature outperforms other algorithms for every fold size K. Different value of K has been used for finding the optimum value of K which gives maximum accuracy. Initially, 2-fold cross validation has been selected and then by increasing the value of k we observed that after a certain value of K the accuracy become turns down. The reason is that, the increasing the number of folding size will decrease the number of testing data in the individual fold that effects the accuracy. From this discussion it can be concluded that, proposed system gets maximum accuracy of 80.48% for 6-fold cross validation.

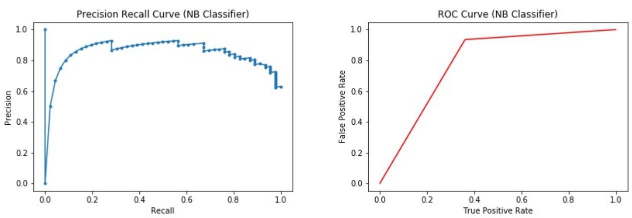
Precision, recall and f_1-score of each class can be presented by a classification report. This report helps for further analysis of an algorithm. In table V classification report of proposed algorithm is presented.

Classification report shows value of f_1-score is 0.74 and 0.84 for positive and negative class respectively. That means proposed system can correctly classify 74% of negative and 84% of positive reviews correctly. For the further evaluation

TABLE V: Classification Report (Proposed algorithm)

Class (C)	Precision	Recall	f_1-score	Support
Negative (C_n)	0.88	0.64	0.74	87
Positive (C_p)	0.77	0.93	0.84	113
avg./total	0.82	0.80	0.80	200

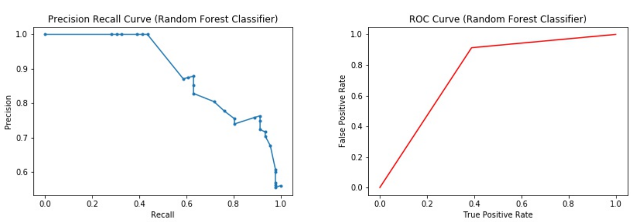
of proposed model graphical measures such as precision-recall curve and receiver operating characteristics curve have been used. Fig. 4 to Fig. 6 show these measures for the tested algorithms.



(a) precision-recall **(b)** ROC

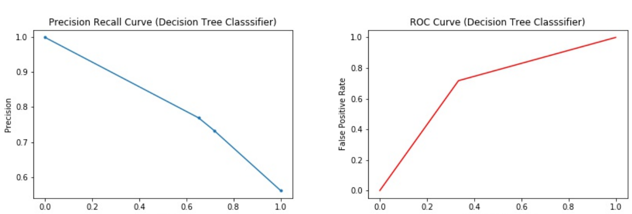
Fig. 4. Result of Multinomial Naïve Bayes

Fig. indicates the precision value of the multinomial naïve bayes is increasing for a certain margin of recall value. Fig. 5(a) and Fig. 6(a) show increase in recall make an adverse effect in precision.



(a) precision-recall **(b)** ROC

Fig. 5. Result of Random Forest



(a) precision-recall **(b)** ROC

Fig. 6. Result of Decision Tree

Fig. 4(b), 5(b) and 6(c) show that true positive rate linearly increase for certain value of false positive rate after that rate of increase becomes almost constant. The trade-off between

Sample Text = > এরাবিয়ান মাস্টারের অভ্যন্তরীণ সজ্জাটা অনন্য। এদের খাবার সুস্বাদু এবং একটু মসলাযুক্ত .. তাদের কর্মীরা খুব বন্ধুত্বপূর্ণ।
System output: Positive
Desired output: Positive

Sample Text = > অপূর্ব খাবার, গ্রেট সার্ভিস... !!
System output: Positive
Desired output: Positive

Sample Text = > আমার জীবনের খাওয়া সবচেয়ে খারাপ খাবার। পুরো টাকাটাই গোল্লায় গেল।
System output: Negative
Desired output: Negative

Sample Text = > পিজার স্বাদটা খুবই জঘণ্য ছিল।
System output: Negative
Desired output: Negative

Sample Text = > খাবারের পরিমাণ কম কিন্তু দাম খুব বেশী। খাবারটা আন্তর্জাতিক মানের নয় কিন্তু স্থানটা ভাল ছিল।
System output: Negative
Desired output: Neutral

Fig. 7. Sample Output

the true positive rate and the positive predicted value is summarized by precision-recall curve and receiver operating characteristics curve summarizes the trade-off between the true and false positive rate for different probability thresholds. Fig. 7 shows some sample restaurant reviews classified by the proposed system.

VII. Conclusion

In this paper a sentiment analysis scheme is presented for restaurant reviews on Bangla language. In this work, different machine learning algorithms have been employed to classify the underlying sentiment of the customer reviews for Bangla text. Besides this, an accuracy comparison of the used models has also been showed where multinomial naïve bayes obtained maximum 80.48% accuracy for 6-fold cross validation among all other used algorithms such as Random forest and Decision Tree. Sentiment analysis has become very effective for the restaurant owners as it can helps owner by providing an essence that what customers think about his restaurant. For sentiment analysis of restaurant reviews many works have been done in English. On the contrary, very few works are done in Bengali. This work is a little effort to compensate the deficiency. In future, more powerful algorithms can be used to find the semantic connection between words of a text that would help us to provide sentiment accurately for more sentiment category. To achieve this goal, increasing the size of dataset would be our prime concern.

References

[1] S. J. Wu, R. D. Chiang, and Z. H. Ji, "Development of a chinese opinion-mining system for application to internet online forums," *The Journal of Supercomputing*, vol. 73, no. 7, pp. 2987–3001, 2017.

[2] 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)*, IEEE, 2015, pp. 95–99.

[3] H. S. Le, T. Van Le, and T. V. Pham, “Aspect analysis for opinion mining of vietnamese text,” in *2015 international conference on advanced computing and applications (ACOMP)*, IEEE, 2015, pp. 118–123.

[4] S. Atia and K. Shaalan, “Increasing the accuracy of opinion mining in arabic,” in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, IEEE, 2015, pp. 106–113.

[5] T. Chumwatana, “Using sentiment analysis technique for analyzing thai customer satisfaction from social media,” 2015.

[6] Y. M. Aye and S. S. Aung, “Sentiment analysis for reviews of restaurants in myanmar text,” in *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, IEEE, 2017, pp. 321–326.

[7] S. Chowdhury and W. Chowdhury, “Performing sentiment analysis in bangla microblog posts,” in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, IEEE, 2014, pp. 1–6.

[8] I. Perera and H. Caldera, “Aspect based opinion mining on restaurant reviews,” in *Computational Intelligence and Applications (ICCIA), 2017 2nd IEEE International Conference on*, IEEE, 2017, pp. 542–546.

[9] T. Doan and J. Kalita, “Sentiment analysis of restaurant reviews on yelp with incremental learning,” in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2016, pp. 697–700.

[10] K. Sarkar and M. Bhowmick, “Sentiment polarity detection in bengali tweets using multinomial naıve bayes and support vector machines,” in *Calcutta Conference (CALCON), 2017 IEEE*, IEEE, 2017, pp. 31–36.

[11] Q. Gan and Y. Yu, “Restaurant rating: Industrial standard and word-of-mouth–a text mining and multi-dimensional sentiment analysis,” in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, IEEE, 2015, pp. 1332–1340.

[12] P. Lak and O. Turetken, “Star ratings versus sentiment analysis–a comparison of explicit and implicit measures of opinions,” in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, IEEE, 2014, pp. 796–805.

[13] X. Fang and J. Zhan, “Sentiment analysis using product review data,” *Journal of Big Data*, vol. 2, no. 1, p. 5, 2015.

[14] L. Banić, A. Mihanović, and M. Brakus, “Using big data and sentiment analysis in product evaluation,” in *Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on*, IEEE, 2013, pp. 1149–1154.

[15] Z. Singla, S. Randhawa, and S. Jain, “Sentiment analysis of customer product reviews using machine learning,” in *Intelligent Computing and Control (I2C2), 2017 International Conference on*, IEEE, 2017, pp. 1–5.

[16] *Facebook page foodbank bd*. [Online]. Available: https://www.facebook.com/foodbankbd/.

[17] *Food bloggers bd*. [Online]. Available: https://www.facebook.com/groups/foodbloggersbd/.

[18] *Food lovers bd*. [Online]. Available: https://www.facebook.com/groups/230918710613193/.

[19] *The food guide chittagong*. [Online]. Available: https://www.facebook.com/groups/1879200579030221/.

[20] J. Y. Yoo and D. Yang, “Classification scheme of unstructured text document using tf-idf and naive bayes classifier,” in *Computer and Computing Science 2015*, 2015, pp. 263–266.

[21] G. S. Chavan, S. Manjare, P. Hegde, and A. Sankhe, “A survey of various machine learning techniques for text classification,” *International Journal of Engineering Trends and Technology*, vol. 15, no. 6, 2014.