A Machine Learning Approach for Sentiment Analysis of Book Reviews in Bangla Language

Conference Paper · April 2022

DOI: 10.1109/ICOEI53556.2022.9776752

2 authors:

Mst. Eshita Khatun
Louisiana State University
12 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Tapasy Rabeya
Daffodil International University
9 PUBLICATIONS **77** CITATIONS
SEE PROFILE

A Machine Learning Approach for Sentiment Analysis of Book Reviews in Bangla Language

Mst. Eshita Khatun
Sr. Lecturer, CSE
Daffodil International University
Email: eshita.cse@diu.edu.bd

Tapasy Rabeya
Lecturer, CSE
Daffodil International University
Email: tapasyrabeya.cse@diu.edu.bd

***Abstract*— With the advent of technology, Sentiment polarity detection has recently piqued the interest of NLP researchers. Sentiment analysis determines the profound meaning of an article. Due to COVID- 19 pandemic, online shopping is the safest way of shopping. Moreover, there are product quality and service issues. Our target is to analyze the book reviews which provide positive and negative reviews in Bangla language. For this, a total of 5500 user generated Bengali reviews are collected from various book review pages of social media. In order to get the best possible result, sentiment analysis is used. Thereafter, five different algorithms are applied to predict with almost high accuracy. Among them, the Random Forest provides us the maximum accuracy which is 98.39%.**

Keywords— *Book Review, Sentiment Analysis, NLP, Opinion Mining, Random Forest, SVM, AdaBoost, Decision Tree, LightGBM.*

I. INTRODUCTION

Sentiment analysis (SA), also known as information extraction, extracts people's feelings, attitudes, and emotions in order to represent the polarity of public opinion in textual form or other[1]. Facebook, Twitter etc. different social sites are now the most comfortable place for people to share their feelings. Data is generally unstructured in social media [2]. Sentiment detection is a technique for estimating a user's point of view on a given topic. It assigns a positive, neutral, or negative polarity to text material, it can be reviews, postings, comments, tweets or bulletins. On the other hand, Sentiment detection is a method for detecting a user's point of view on a given topic.

As a result of large amount of e-commerce sites, people's shopping habits have changed dramatically in recent years. That makes book as one of the most popular online items for the book lovers. Books are crucial in every human life because they provide information about the outside world, improve their reading, writing, and speaking abilities, and improve memory and intellect. Sometimes, buying books from online stores creates difficulties if the reader is unfamiliar with the books or with the bookstore itself. However, product reviews on the Internet have become an essential source of information for customers making purchasing decisions. Our goal is to analyze user generated Bengali reviews about different bookstore and books and therefore, provide some valuable information to the consumers to find out better online bookstores. This sentiment analyzed information will also help the shop owners to enhance their service quality.

When technology advancement brings people closer together, it becomes simpler to deceive or lie on the Internet. People are hesitant to buy books after seeing adverts for bookstores on the internet. They also pay attention to past consumers' comments and evaluations on these items and service providers. Nevertheless, analysis of the Bengali dataset is difficult. Data preprocessing as like to add contraction, remove punctuation it carries the most significant section for analyzing the Bengali text. After tokenizing the processed data, different machine learning algorithms were embedded for awaited output [3].

Our stated process in this paper can identify whether a certain book has received negative or positive comments on a rational basis. The review comments are collected from social sites for preparing the datasets. It's quite difficult to collect a large amount of data. 5500 Bengali book reviews are categorized into positive and negative attitudes. And then our model was trained with these datasets for more accurate accuracy.

The organization of our paper is designed as follow. In section I the introduction is described. The literature review is explained in section II, where all papers are chosen based on sentimental analysis related works. The methodology of our experimental analysis is presented in section III. Data analysis is also described in this section. Results are elaborately discussed in section IV. Finally, the paper ended with the conclusion, that is described in section V.

II. LITERATURE REVIEW

In the recent era, prediction results are the most often used applications of machine learning. These studies targeted specific issues and utilized a range of machine learning approaches to solve them. In this section, highlights the actions that several experts in the preceding region successfully carried out. Sentiment Analysis (SA) is a mixture of opinions, feelings and textual subjectivities. SA is the most difficult natural language processing job at present. Social networking sites such as Facebook are often used to share views on a single life entity. Newspapers published news about a specific incident, and in news comments the user shared his input. The amount of feedback received from online items is increasing every day. As a result, reviews and opinions play an important part in determining people's levels of satisfaction. Opinion mining can help you find out more information.

Chowdhury et.al. presented an approach that extracted the sentiments of user from Bangla blog post either it is positive or negative. SVM performed 93% with unique characteristics from 1300 col-selected data in its proposed process [5]. Aspect-based or feature-based analysis is a type of sentimental analysis that examines the sentiments surrounding a certain issue. Bengali language analysis, lexicon as part of the speech tagger, and other tasks are challenging due to the scarcity of resources such as a well annotated data collection. Their focus was on a restaurant review and the application of aspect-based research to get cricket opinions. SVM has the maximum validity for extracting and discovering polarity in insects and restaurants, respectively, with 71% and 77% [6].

In online shopping, understanding what the client wants is critical, but firms may not be as informed as they should be. In order to validate their ratings, a machine learning algorithm has been used by C. Chauhan et al. to distinguish between negative and positive comments from potential customers. They examined a variety of publications and found that Naive Bayes provided positive results, the results varied depending on the environment, strategy, and goals [7]. Modeling constructed with this sort of network and its variations recently proved excellent performance in various downstream NLP applications, particularly in resource-rich languages like English. However, these models have not been fully studied for categorization challenges. In Bangladesh, they fine-tune the multilingual text classification transformer model. In order to describe the text-based sensations given by analyzing in Bangla, a Convolution Neural Network (CNN) was introduced by Alam et al. that achieves 99.87% accuracy and around data points of 850. Among the data set 500 were positive and 350 were negative [8].

Two different methods were proposed by Tuhin et al. for identifying and classifying emotions in various dimension from all around Bangladesh. These were ecstatic, enraged, sorrowful, terrified, enthusiastic, and sensitive. In Naive Bayes, the topical solution and the method of grouping are strategies. A data collection of 7400 Bangladesh phrases was employed, and achieve 90% accuracy. They then compared their article to two others, both of which scored 93% for SVM and 83% for document frequency. The emotional parameter in each of the three articles was different [9].

Mittal et al. showed a technique for analyzing language Hindi that yields 82.89% and 76.59% positive and negative validity, respectively. They opted to assess emotions and increase the database's coverage in order to improve the database's consistency. This article describes an educational program that analyzes the Roman Urdu people's emotions through the genres of sports, software, cuisine and recipes, theatre, and politics. It contains 10,021 sentences culled from 566 internet discussions. Basically, the project's goal was designed in two steps: (1) developing a Roman Urdu human-annotated corpus for emotional analysis; and (2) evaluating feeling analysis approaches based on Rule-based, N-gram (RCNN) models [4].

To identify positivity and negativity from the book reviews Hossain et al. introduced a machine learning based method. They used 2000 reviews on Bengali books and applied logistic regression, naive Bayes, SVM, and SGD machine learning approaches. They found better accuracy 84% by using multinomial Naive Bayes [10]. Kuhamanee et al. has done an analysis of foreign tourist's reviews of Bangkok for their visiting purpose [11]. They classified the used English sentiment words into five categories. They found the highest percentage for travelling purposes 71.93%. Along with ANN, SVM, Naïve Bayes and decision tree were also used for their analysis. And finally come up with highest accuracy of 80.33% given by ANN.

The comparative analysis has been done on movie reviews from the data of twitter. And a Senti-lexicon algorithm was introduced by them. They have analyzed 300 twitter reviews with 70% accuracy [12]. Another work has been done for product review. They have used 1000 feedbacks and implemented Logistic Regression, Decision Tree, KNN, SVM and Random Forest classification algorithms. And the highest accuracy was performed by SVM that was 88.81% [13].

A lexicon-based backtracking approach was introduced to detect emotion from Bengali text by analyzing the sentiment first. With more than 70 percent accuracy they showed that people most of the time express their real emotion whenever talking and writing [16]. The same algorithm was also used for analyzing Bengali song review from a specific YouTube channel [17]. In social networks, the interactive users act is weighted automatically along with the user's interest [18].

III. METHODOLOGY

In this section, our research methodology, how the collected Bengali reviews are analyzed has been described. Our main focus is to protect the new customer from online scammed and increase the buying and selling rate for the service providers. To create our model, three steps are maintained. Firstly, Bengali reviews are collected for analysis. Then data preprocessing step is done, where the comments are being preprocessed based on certain criteria. After preprocessing, tokenization is performed and finally 5 different classification algorithms are applied. The work methodology flow chart is presented by Fig.1.

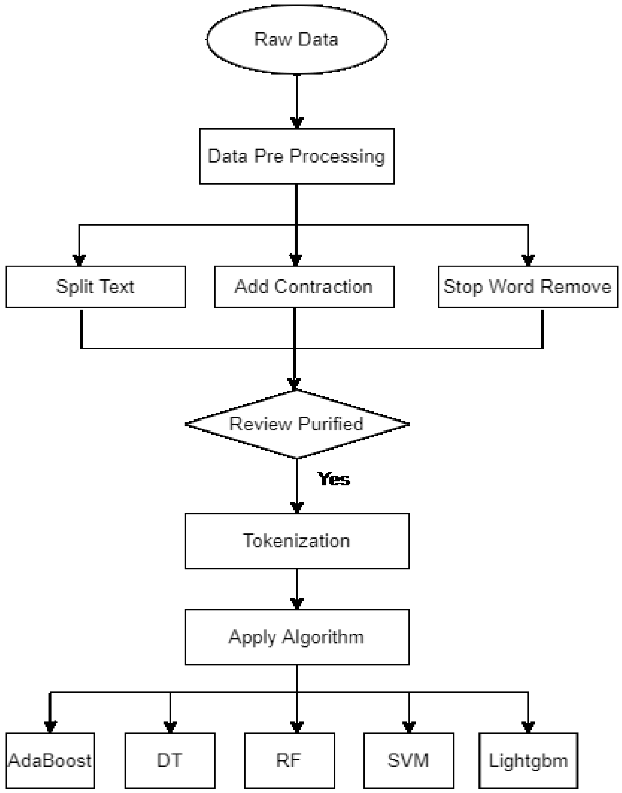
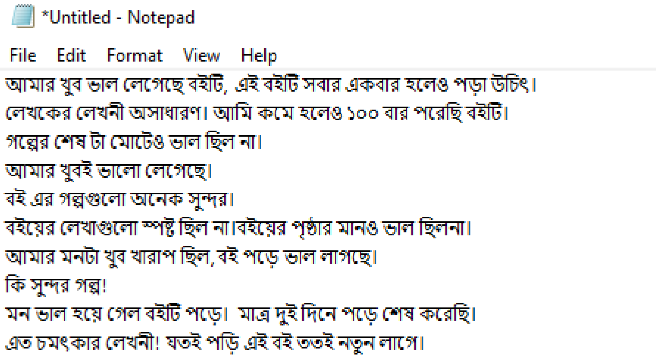


Fig.1. Flowchart of our proposed model

A. Data Collection

A book review is a delicate piece of information that is essential to the success of any book. A large amount of data is required to implement machine learning algorithms. Analysis results depend on the volume and quality of data. A total 5500 user generated Bengali review are collected from various book review pages of social media. While selecting a comment, the sentiment of a sentence is emphasized on. In case of neutral sentiment, those sentences are skipped. Only two parameters are there for the dataset, one for text and paragraph and another for classification. Fig. 2 represents the collected raw data from the online media.



আমার খুব ভাল লেগেছে বইটি, এই বইটি সবার একবার হলেও পড়া উচিৎ।
লেখকের লেখনী অসাধারণ। আমি কমে হলেও ১০০ বার পরেছি বইটি।
গল্পের শেষ টা মোটেও ভাল ছিল না।
আমার খুবই ভালো লেগেছে।
বই এর গল্পগুলো অনেক সুন্দর।
বইয়ের লেখাগুলো স্পষ্ট ছিল না।বইয়ের পৃষ্ঠার মানও ভাল ছিলনা।
আমার মনটা খুব খারাপ ছিল, বই পড়ে ভাল লাগছে।
কি সুন্দর গল্প!
মন ভাল হয়ে গেল বইটি পড়ে। মাত্র দুই দিনে পড়ে শেষ করেছি।
এত চমৎকার লেখনী! যতই পড়ি এই বই ততই নতুন লাগে।

Fig.2. Collected raw Bengali data from online platform

B. Data PreProcessing

Data preprocessing is a data mining tool that converts raw data into a usable and efficient format. Information preprocessing is critical for knowledge acquisition. Our model is based on KDD (Knowledge Discovery Database). According to Kamiran et al. [14], the four most important data pre-processing procedures are elimination, data massaging, weighting, and same poling. To develop accessible data sets, data messaging methodologies are predominantly used in the work. 3 steps are used in the pre-processing phase. Firstly, the long text is split into small text. Secondly, contraction is added in the comments if necessary and deleted unneeded points and phrases from the Bangla stop. And thirdly the stop and exclamation sign are removed from the text. TABLE I represents the data preprocessing technique.

TABLE I PREPROCESSING OF RAW DATA

Raw Comment	Splitting Text	Adding Contraction	Removing Stop Words
লেখকের লেখনী অসাধারণ।	লেখকের লেখনী অসাধারণ।	---	লেখকের লেখনী অসাধারণ
আমি কমে হলেও ১০০ বার পরেছি বইটি।	আমি কমে হলেও ১০০ বার পরেছি বইটি।	---	আমি কমে হলেও ১০০ বার পরেছি বইটি
কি সুন্দর গল্প!	---	---	কি সুন্দর গল্প
আমার মনটা খুব খারাপ ছিল,বই পড়ে ভাল লাগছে।	---	---	আমার মনটা খুব খারাপ ছিল বই পড়ে ভাল লাগছে

C. Data Tokenization

Tokenization is defined by Pinto et al. [15] as a method of separating flag phrases, which can be words or signals. In our database, there are a lot of phrases. The task was not completed using a phrase mark rather than a word label. It's also crucial to tokenize. Tokenization divides our entire phrase into terms. TABLE II shows the tokenization approach.

TABLE II PREPROCESSED DATA TOKENZIATION

Raw Data	**Type**	**Tokenized data**
বই এর গল্পগুলো অনেক সুন্দর	Positive	‘বই’ ,‘এর’ , ‘অনেক’, ‘গল্পগুলো’, ‘অনেক’, ‘সুন্দর ‘
বইয়ের লেখাগুলো স্পষ্ট ছিল না	Negative	‘বইয়ের’ , ‘লেখাগুলো’, ‘স্পষ্ট,‘ ছিল’, ‘না’
বইয়ের পৃষ্ঠার মান ভাল ছিলনা	Negative	‘বইয়ের’, ‘পৃষ্ঠার ‘ ,’ মান’, ’ভালো ‘, ’ ছিলনা’

D. Model Implementation

There are some significant implemented classification algorithms are available for performing analysis In Natural Language Processing. Among them, these five algorithms like: AdaBoost, Decision Tree, SVM, LightGBM, and Random Forest are chosen to train the model.

IV. EXPERIMENT AND OUTPUT ANALYSIS

In analytical research, this portion mostly depends on empirical evidence and test results. In two modules our data were classified: “positive” and “negative”.

TABLE III POSITIVE NEGATIVE MODULES FOR TESTING

Actual sentiment	Negative (Predict)	Positive (Predict)
Negative (Actual)	True Negative	False Positive
Positive (Actual)	False Negative	True Positive

Based on TABLE III the proposed model performance is assessed against four parameters: Accuracy, Precision, Recall, and F1-Score.

$$Accuracy = \frac{True\ Positive + True\ Negative}{(Total\ Example)} \quad (1)$$

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (2)$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (3)$$

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)} \quad (4)$$

5 classifications algorithms are used to train our dataset along with 30% to 70% of test data to determine which item works the best. Experiment results shown in Fig. 2. Here AdaBoost attain 86.67%, Decision Tree reach 90.06%, SVM attain 97.45%, Random Forest attain 98.39%, LightGBM attain 92.67% accuracy.

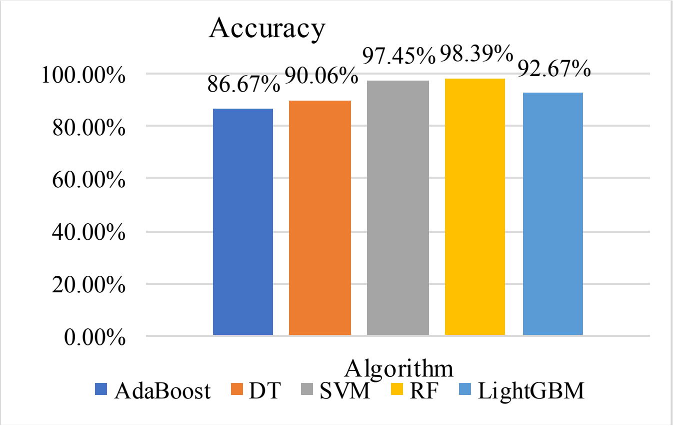


Fig. 2. Bar chart of the Algorithms accuracy

Fig. 3. shows the accuracy of variety algorithms. Random Forest reaches the highest accuracy compared to other algorithms. SVM is also noticeable which is mostly nearest to Random Forest.

TABLE IV F1-SCORE, RECALL, PRECISION, SPECIFICITY VALUES

Score Matrix	**Algorithms**				
	AdaBoost	**Decision Tree**	**SVM**	**Random Forest**	**LightGBM**
F1 Score	0.8663	0.8998	0.9746	0. 9849	0.9413
Recall	0.8213	0.8935	0.9734	0. 9811	0.9049
Precision	0.9165	0.9062	0.9761	0. 9887	0.9608
Specificity	0.8505	0.9037	0.9661	0. 9819	0.9187

TABLE IV displays the Score Matrix of algorithms used.

V. CONCLUSION

SA is dependent on a dataset of specific material due to the fast growth of Internet users. Using multiple feature extraction approaches, this research portrays an experimental analysis on machine learning-based sentiment classification system using Bengali book reviews. 5500 consumer reviews from 55 different book categories are analyzed. Book information and descriptions are gathered, as well as screening book reviews is done for key feature phrases. As a consequence, five primary characteristics that exist in virtually all customer evaluations and have an impact on them are discovered: pricing, transportation, quality, design, and satisfaction. Following that, five common algorithms like AdaBoost, Decision Tree, RF, SVM and LightGBM are employed. Random Forest provides the best perforation 98.39%, in terms of accuracy. Bookshop owners and customers both may learn which books are worth looking at and which ones aren't, and potential purchasers can identify which books have good or awful characters. Our system has sometimes failed to detect exact sentiment from sentences. Spelling structure is a major problem in the Bengali language. Context meaning can be changed by adding or missing any vowel. In the future, better predictions can be propagated.

REFERENCES

[1] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015): 1-14.

[2] Y. Ko and J. Seo, ―Automatic text categorization by unsupervised learning,‖ in Proceedings of the 18th conference on Computational

linguistics-Volume 1. Association for Computational Linguistics, 2000, pp. 453–459.

[3] A. K. Mohammad Masum, S. Abujar, M. A. Islam Talukder, A. K. M. S. Azad Rabby and S. A. Hossain, "Abstractive method of text summarization with sequence to sequence RNNs," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-5

[4] Mittal, Namita, et al. "Sentiment analysis of hindi reviews based on negation and discourse relation." Proceedings of the 11th Workshop on Asian Language Resources. 2013.

[5] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), 2014, pp. 1-6, doi: 10.1109/ICIEV.2014.6850712.

[6] M. Rahman and E. Kumar Dey, “Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation,” Data, vol. 3, no. 2, p. 15, May 2018.

[7] Chauhan, Chhaya and Smriti Sehgal. “Sentiment analysis on product reviews.” 2017 International Conference on Computing, Communication and Automation (ICCCA) (2017): 26-31.

[8] M. H. Alam, M. -M. Rahoman and M. A. K. Azad, "Sentiment analysis for Bangla sentences using convolutional neural network," 2017 20th International Conference of Computer and Information Technology (ICCIT), 2017, pp. 1-6, doi: 10.1109/ICCITECHN.2017.8281840.

[9] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter and A. K. Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019, pp. 360-364, doi: 10.1109/CCOMS.2019.8821658.

[10] Hossain, Eftekhar, Omar Sharif, and Mohammed Moshiul Hoque. "Sentiment polarity detection on bengali book reviews using multinomial naive bayes." Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, 2021. 281-292

[11] Kuhamanee, Taweesak, Nattaphon Talmongkol, Krit Chaisuriyakul, Wimol SanUm, Noppadon Pongpisuttinun, and Surapong Pongyupinpanich. "Sentiment analysis of foreign tourists to Bangkok using data mining through the online social network." In Industrial Informatics (INDIN), 2017 IEEE 15th International Conference on, pp. 1068-1073. IEEE, 2017.

[12] Mumtaz, Deebha, and Bindiya Ahuja. "Sentiment analysis of movie review data using Senti-lexicon algorithm." In Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on, pp. 592-597. IEEE, 2016.

[13] Shafin, Minhajul Abedin, et al. "Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language." 2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE, 2020.

[14] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." Knowledge and Information Systems 33.1 (2012): 1-33.

[15] Pinto, Alexandre, Hugo Gonçalo Oliveira, and Ana Oliveira Alves. "Comparing the performance of different NLP toolkits in formal and social media text." 5th Symposium on Languages, Applications and Technologies (SLATE'16). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[16] Rabeya, Tapasy, Sanjida Ferdous, Himel Suhita Ali, and Narayan Ranjan Chakraborty. "A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from Bengali text." In *2017 20th international conference of computer and information technology (ICCIT)*, pp. 1-7. IEEE, 2017.

[17] Rabeya, Tapasy, Narayan Ranjan Chakraborty, Sanjida Ferdous, Manoranjan Dash, and Ahmed Al Marouf. "Sentiment analysis of bangla song review-a lexicon based backtracking approach." In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1-7. IEEE, 2019.

[18] Sivaganesan, D. "Novel influence maximization algorithm for social network behavior management." *Journal of ISMAC* 3.01 (2021): 60-68.