Bengali Review Analysis for Predicting Popular Cosmetic Brand Using Machine Learning Classifiers

Chapter · November 2022

DOI: 10.1007/978-981-19-4960-9_21

Tapasy Rabeya
Daffodil International University
9 PUBLICATIONS **79** CITATIONS
SEE PROFILE

Mst. Eshita Khatun
Louisiana State University
12 PUBLICATIONS **46** CITATIONS
SEE PROFILE

Sheak Rashed Haider Noori
Daffodil International University
67 PUBLICATIONS **343** CITATIONS
SEE PROFILE

Sharmin Akter
Daffodil International University
11 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Bengali Review Analysis for Predicting Popular Cosmetic Brand Using Machine Learning Classifiers

Tapasy Rabeya, Eshita Khatun, Sheak Rashed Haider Noori, Sharmin Akter, and Israt Jahan

Abstract Nowadays, online platform has become one of the most popular media to express people's thought of all ages. That made the online platform a precious source for getting almost every kinds of information. As online shopping is rising in no time in recent years, as a result millions of comments are generating every single day. These users generated opinions on social media and different websites has made it easier for the people choosing the right product for them. Hence, sentimental analysis is a sought-after research topic nowadays. Our research paper has portrayed an experimental study on different cosmetics products review. To do so, we have selected ten popular cosmetic brands for analyzing their product review and chosen to analyze Bengali comments or sentences. The main focus of our work was to get out the most popular cosmetic brands among ten chosen brands. We have applied four classification algorithm such as naive Bayes, random forest, decision tree, and support vector machine for analyzing the final outcome and found vaseline and clear are the most popular brands.

Keywords Sentiment analysis · Experimental study · Opinion mining · Cosmetic product review · Bengali comment · NLP

T. Rabeya (✉) · E. Khatun · S. R. H. Noori · S. Akter · I. Jahan
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
e-mail: tapasyrabeya.cse@diu.edu.bd

E. Khatun
e-mail: eshita.cse@diu.edu.bd

S. R. H. Noori
e-mail: drnoori@daffodilvarsity.edu.bd

S. Akter
e-mail: sharminakter.cse@diu.edu.bd

I. Jahan
e-mail: isratjahan.cse@diu.edu.bd

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 383,
https://doi.org/10.1007/978-981-19-4960-9_21

1 Introduction

Sentiment analysis is a major segment in natural language processing. It provides the highest possibility to make a decision for an ongoing process. The sentiment means an attitude, conceptions, and mental feeling toward something. Analysis is the pathway to break a complex object into smaller pieces for getting better understanding. So, sentiment analysis is the process to classify sentiment of people either positive, negative or neutral. Social platforms data help us to predict and discover various events or in total reflection of the mind of the community.

In recent times, people love to live along with the progressive Internet-based information society. Following the time flow, people are habituated with e-commerce around the world. Customer reviews are a reflection of customer satisfaction. Users share their reviews and feedback by posting on social media such as Facebook pages and e-commerce sites. These reviews produce a huge volume of data which can be used for making an effective decision. However, measuring analysis from Bengali text is not so easy. Bengali text data is the main element for data preprocessing such as contraction add, punctuation remove, and stop words. We have tokenized processed text data and also employed various machine learning algorithms in order to get the best possible result.

Our method mentioned in this paper can predict popular cosmetic brands based on the customer's review. For preparing the dataset, we have used 800 comments for ten cosmetic brands from different websites. This paper presents the work on sentiment analysis of Bengali reviews for predicting popular cosmetic brands and identifying the user acceptance accuracy rate from various machine learning algorithm aspects.

The following paper is stated as follows. In Sect. 2, we have presented a literature review on sentiment analysis of the public review on different aspects. In Sect. 3, we have discussed data sources. Methodology is described in Sect. 3. Finally, conclusion is illustrated in Sect. 4.

2 Literature Review

In recent years, researchers are showing interest in the field of natural language processing (NLP). Some researchers have been emphasized on solving problems in e-commerce business [1]. This study represents some of literature on NLP related to Bangla text.

In [2], prominent authors proposed a system that can classify the customer's feedback into two categories, positive, and negative in an online restaurant. This research has used 1000 Bengali text reviews and attain 80.48% accuracy using multinomial naive Bayes.

Clara et al. conducted aspect based analysis of beauty products review. Random forest is used for classification and multiaspect sentiment analysis attain highest accuracy 90.48% [3].

Product reviews enhance the communication between buyers and customers. Data pre-processing, feature extraction, and sentiment classification. These three phases have been used to predict customer's sentiment polarity [4]. Authors of [5], proposed a system that can suggest product and shop on the basis of customers review analysis. As discussed in [6], is to understand the marketing strategies of recent times, *K*-nearest neighbor (KNN) is used to analysis of reviews.

Dr. Rajesh Bose et al. explored that the customer's two sentiments and eight emotions such as happiness, sadness, anger, trust, disgust, fear, surprise, and anticipation with NRC emotion lexicon. From October 1990 to October 2012, they worked on 568,454 reviews, 74,258 products from more than 256,059 users from amazon [7]. Samina Kausar et al. worked with three polarity features—verb, adverb, and adjective along with these they are divided into five classes and those classes are—strongly positive, neutral positive, negative, and strongly negative. They have got better results from other's previous work through their work [8]. Hang Cuireferred et al. that two classifications of customer reviews which are positive and negative. They also discussed various machine learning algorithms to assess numerous trade-offs using an average of 100k reviews from different websites. They have also shown through their work that their high-order *n*-gram feature with discrimination classification gives better results than academic papers [9].

Najma Sultana et al. proposed an approach which is worked with combination verbs, adverbs, and adjectives using machine learning. They are using a document-level approach to detect positive and negative sentiments in sentences [10]. Apoorve et al. wanted to show in their result that a significant improvement over a majority baseline and a more difficult baseline consisting of lexical *n*-gram. To capture the effect of context, they augmented lexical scoring with *n*-gram analysis [11].

Upma Kumara et al. suggested a model by using SVM on a dataset of smartphone reviews to find out the sentiment which is mainly good, bad, and super hit, and by using the SVM algorithm they got 90.99% accuracy [12]. Jorge Carrillo et al. proposed a model which is the feature-driven approach which is given a better result than the then previous approach. They also said that about a product the user opinion varies for different impacts for different product features [13]. A work was done on detecting emotion from Bengali sentences, and they proposed a new algorithm called backtracking approach. And their accuracy was more than 70%. The same algorithm was used to analyze the sentiment of a YouTube channel comment [14, 15].

3 Methodology

In this section, we illustrate our model which can predict the popular cosmetics brand on the basis of product reviews. We maintain few steps to create our working model. After collecting product reviews we have done text cleaning and data labeling for data pre-processing. Finally, applied four classification machine learning algorithms to obtain desire result. In Fig. 1 shows the working flowchart.

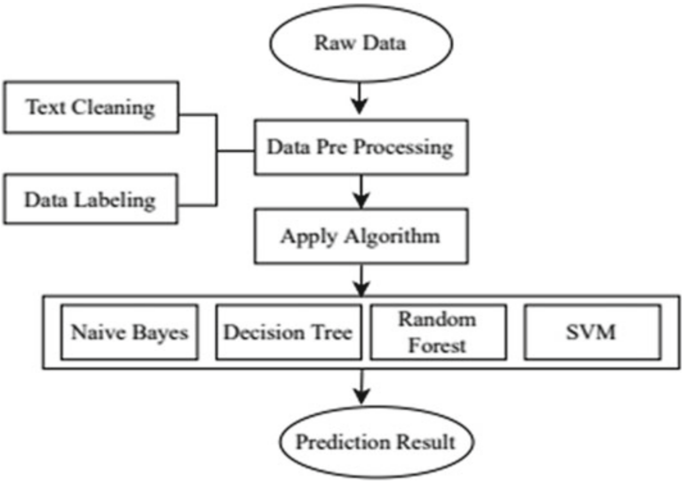

Fig. 1 Work flowchart

3.1 Data Collection Procedure

The data are collected from different online websites for this experimental analysis. As our main focus was to analyze the Bengali reviews to predict the popular cosmetic brand, firstly we have selected ten brands to collect their products review such as sunsilk, dove, tresemme, clear, ponds, simple, vaseline, fogg, parachute, and nivea. We have collected Bengali review from different website like Daraz, shajgoj, stylin, etc. Table 1 shows the row Bengali comments collected from different website.

3.2 Data Pre-processing

In the first step of data pre-processing, we eliminate all the emotional icon and secondly, divided the collected reviews into two categories such as positive and negative. And assigned a 1 or a 0 to each positive and negative comment.

Table 2 presents the data levelling according to their sentiment.

3.3 Experimental Results and Analysis

We have applied four algorithms to do our experimental analysis: SVM, decision, random forest, and multinomial naive Bayes tree to see how they perform with the

Table 1 Raw data collected from different websites

Company name	Product name	Comments
ইউনিলিভার	ক্লিয়ার শ্যাম্পু মেন কুল স্পোর্ট মেনথল এন্টিড্রান্ড্রফ	১.মানসম্মত ২.অথেনটিক ৩.ভালো পণ্য ৪. অরিজিনাল প্রোডাক্ট। ভালো লাগছে ৫. আসাধারন ৬.সত্যিই ভালো পণ্য ৭.প্রোডাক্ট ভালো ছিলি ৮.অরিজিনাল প্রোডাক্ট হাতে পেয়েছি ৯.ভালো পণ্য
ইউনিলিভার	ট্রেসেমি শ্যাম্পু বোটানিক ন্যারিসিং রিপ্লেস	১.প্রোডাক্ট ভাল, আগে ব্যবহার করেছি ২. প্রডাক্ট অনেক ভালো ১০/১০ ৩.শ্যাম্পুটা ভালো। আমার কোঁকড়া চুল হলেও নরম আর জটমুক্ত হয়ে গেছে একদম ৪. অরিজিনাল প্রোডাক্ট ৫.এখনও ব্যবহার করিনি ৬.ভালো প্রোডাক্ট ৭. ভাল ছিলি পণ্য টা ৮.অসাধারণ ৯. চুল স্মুথ করে, ভালো প্রোডাক্ট ১o.আমি প্রোডাক্টটিকে ভালোবেসে ফেলেছি

Table 2 Data pre-processing

Sentiment types	Label	Example
হ্যাঁ বোধক	1	১. পন্যটি ভালো
		২. অরিজিনাল পন্য
		৩. অসাধারন প্রোডাক্ট
না বোধক	0	১.পন্যটি ভালো না
		২.এর মান ভালো না
		৩.একদম ভালো লাগেনি

same data set. We had collected total 800 comments for ten cosmetic brands from different websites. After applying these four algorithms, we have found that some products have 100% user acceptance, some other have very low user acceptance. Table 3 shows the uses acceptance rate from different algorithms perspective.

According to algorithm performance analysis on our data set, vaseline and clear are the most popular cosmetic brands among the ten cosmetic brands. Our four algorithms showed 100% acceptance rate for vaseline and clear. Additionally, our analysis showed that the less popular cosmetic brand among our chosen ten brands is parachute of 62.5% acceptance rate of user.

Figures 2, 3, 4 and 5 are representing the pie charts of the acceptance measurement of naïve Bayes, random forest, decision tree, and SVM. They are representing all of the cosmetic brands acceptance rate within 100% band.

Figure 2 represents the user acceptance rate of all brands in between 100% band using multinomial naïve Bayes classification algorithm. Where sunsilk, clear, and

Table 3 User acceptance rate using naïve Bayes, random forest, decision tree, and SVM

Brand	Multinomial Naive Bayes (%)	Random forest (%)	Decision tree (%)	SVM (%)
Sunsilk	100.0	100.0	87.5	100.0
Dove	50.0	87.5	75.0	87.5
Tresemme	100.0	80.0	80.0	80.0
Clear	100.0	100.0	100.0	100.0
Ponds	75.0	87.5	75.0	75.0
Simple	66.66	66.67	100.0	66.66
Vaseline	100.0	100.0	100.0	100.0
Fogg	66.66	100.0	83.3	66.66
Parachute	62.5	62.5	62.5	62.5
Naivea	60.0	100.0	80.0	80.0

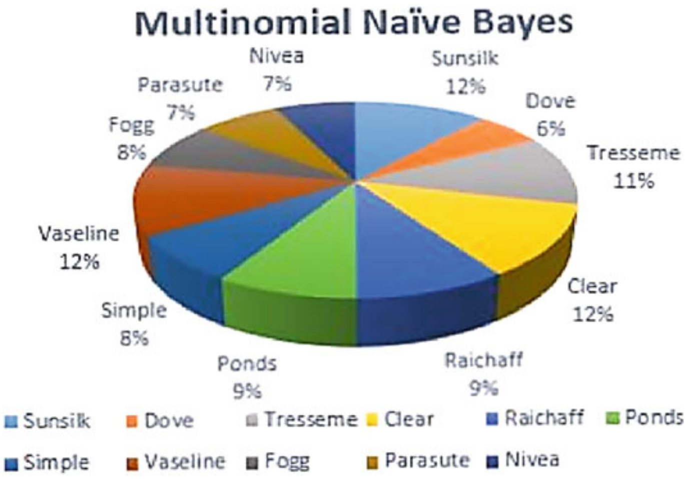

Fig. 2 Accuracy of multinomial Naïve Bayes

vaseline has 12% acceptance rate that is highest. Tresseme get the second highest user acceptance, 11%, Raichaff and ponds get 9% acceptance rate, simple and fogg 8%, Parachute and nivea 7%, and finally dove 6%.

Figure 3 shows the user acceptance rate of all brands in between 100% band using decision tree classification algorithm. Where simple, clear, and vaseline has 11% acceptance rate that is highest. Sunsilk, rainchaff, fogg, and nivea got the second highest user acceptance, 9%, dove, tresemme, and ponds got 8%, and finally para-chute got 7% user acceptance rate.

Figure 4 illustrates the user acceptance rate of all brands in between 100% band using random forest classification algorithm. Where clear, fogg, ivea and vaseline has 11% acceptance rate that is highest. Sunsilk get the second highest user

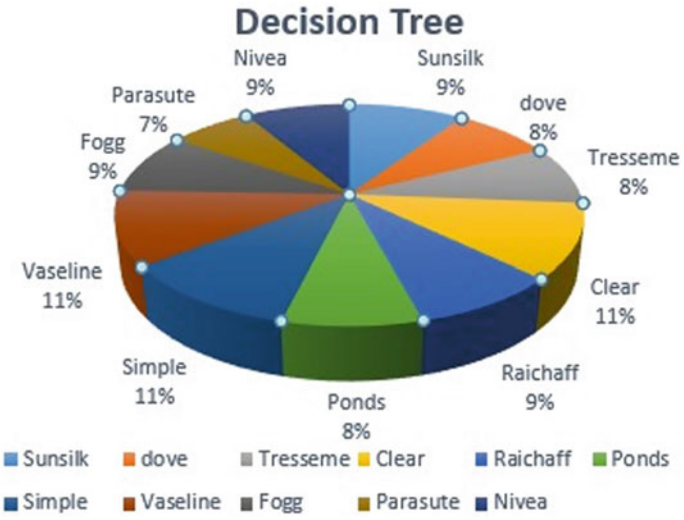

Fig. 3 Accuracy of decision tree

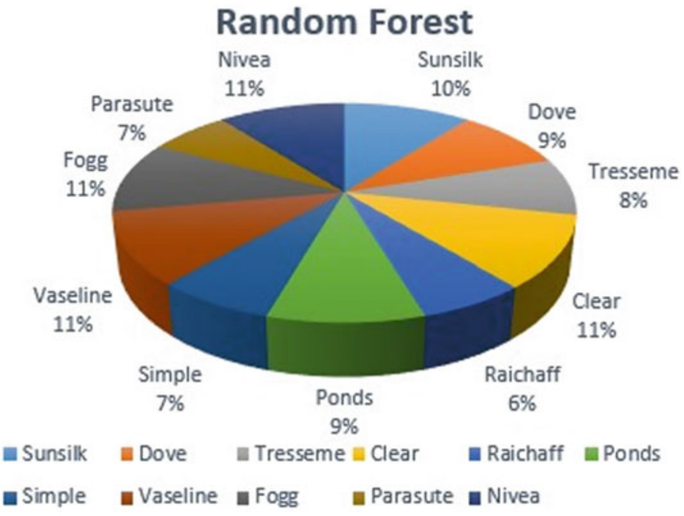

Fig. 4 Accuracy of random forest

acceptance 10%, dove and ponds get 9%, and finally parachute gets 7% user acceptance rate, tresemme gets 8%, and simple and parachute get the lowest acceptance rate 7%.

Figure 5 states the user acceptance rate of all brands in between 100% band using SVM classification algorithm. Where sunsilk, clear, and vaseline have 11%

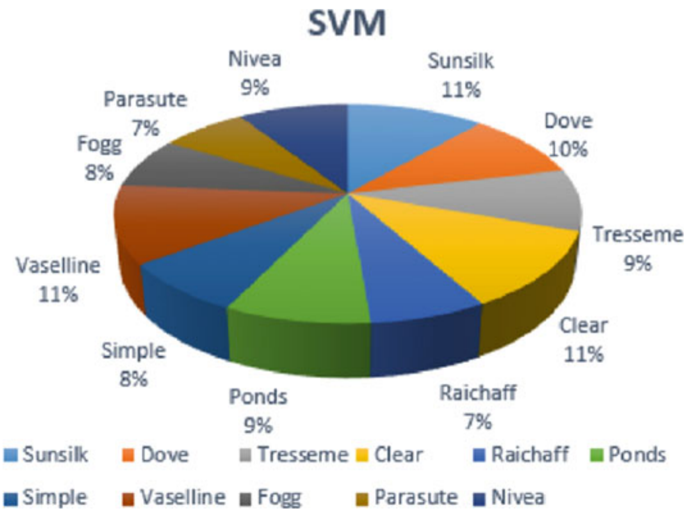

Fig. 5 Accuracy of SVM

acceptance rate that is highest. Dove got the second highest user acceptance, 10%, tresemme, nivea, and ponds get 9% acceptance rate, simple and fogg get 8%, parachute gets 7% the lowest acceptance rate.

3.4 Comparative Analysis

As we have applied four algorithms with same data set, so results vary with the algorithms. Table 4 presents a comparative analysis of the algorithm's performance similarity with each other.

Analysis showed that naïve Bayes gave four times similar result with random forest and five times similar result with decision tree and seven times similar results with SVM. That depict that, the naïve Bayes and SVM algorithms analysis result are quite closer than random forest and decision tree. But, similarity in result analysis is greater between random forest and SVM, among 11 brands they gave the same result for eight times. Based on our experimental work, it is clear that random forest and SVM performance is much similar to each other compared to other algorithms.

Additionally, for clear, vaseline, and parachute all the algorithms result were same. On the other hand, for sunsilk, ponds and simple naïve Bayes, random forest, and SVM's performances were same.

Table 4 Performance similarity among algorithms

	Naive Bayes	Random forest	Decision tree	SVM
Naive Bayes	–	4	5	7
Random forest	4	–	4	8
Decision tree	5	4	–	6
SVM	7	8	6	–

4 Conclusion

In online shopping, while buying a product people are basically influenced by the products reviews. Research showed that people generally scroll down the comments at least one before buying a product [16]. That makes sentimental analysis a sought-after research topic now a days. We have found the Bengali comment analysis quite tough as people tend to express their emotions in a mixture language. Like, Bengali comment written in English or using English word in the Bengali comments. For example:

"Amar product ta onek valo legeche" (Bengali comment written in English)
"Product ta onek awesome" (using English adjective)

In our empirical work, we were intended to find the most popular cosmetic brand that the majority of people care about while selecting their cosmetics. In this connection, we have applied four established classification algorithms on 800 collected user generated Bengali reviews from different online sites. And results showed, vaseline and clear are the most popular brands with 100% user acceptance rate among ten chosen cosmetic brands and parachute is the less popular brand with 62.5% user acceptance rate. And additionally, we found that random forest and SVM performance is much similar to each other compared to other algorithms. Among ten brands they gave the same result for eight times. For a strong result, a large data set with different dimension is highly required.

References

1. Munna MH, Rifat Md RI, Badrudduza ASM (2020) Sentiment analysis and product review classification in e-commerce platform. In: 2020 23rd International conference on computer and information technology (ICCIT). IEEE
2. Sharif O, Hoque MM, Hossain E (2019) 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). IEEE
3. Clara AY, Adiwijaya A, Purbolaksono MD (2020) Aspect based sentiment analysis on beauty product review using random forest. J Data Sci Appl 3(2):67–77
4. Jindal K, Aron R (2022) A hybrid machine learning approach for sentiment analysis of beauty products reviews. J Inf Syst Telecommun (JIST) 1(37):1

5. Yi S, Liu X (2020) Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. Complex Intell Syst 6(3):621–634
6. Kirana YD, Al Faraby S (2021) Sentiment analysis of beauty product reviews using the K-nearest neighbor (KNN) and TF-IDF methods with chi-square feature selection. J Data Sci Appl 4(1):31–42
7. Bose R, Dey RK, Roy S, Sarddar D (2020) Sentiment analysis on online product reviews. In: Information and communication technology for sustainable development. Springer, Singapore, pp 559–569
8. Kausar S, Huahu X, Ahmad W, Shabir MY (2019) A sentiment polarity categorization technique for online product reviews. IEEE Access 8:3594–3605
9. Cui H, Mittal V, Datar M (2006) Comparative experiments on sentiment classification for online product reviews. In: AAAI, vol 6, no 1265–1270, p 30; Fang X, Zhan J (2015) Sentiment analysis using product review data. J Big Data 2(1):1–14
10. Sultana N et al (2019) Sentiment analysis for product review. ICTACT J Soft Comput 9(03)
11. Agarwal A, Biadsy F, Mckeown K (2009) Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In: Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009), pp 24–32
12. Kumari U, Sharma AK, Soni D (2017) Sentiment analysis of smart phone product review using SVM classification technique. In: 2017 International conference on energy, communication, data analytics and soft computing (ICECDS). IEEE, pp 1469–1474
13. De Albornoz JC, Plaza L, Gervás P, Díaz A (2011) A joint model of feature mining and sentiment analysis for product review rating. In: European conference on information retrieval. Springer, Berlin, pp 55–66
14. Rabeya T, Ferdous S, Ali HS, Chakraborty NR (2017) 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). IEEE, pp 1–7
15. Rabeya T, Chakraborty NR, Ferdous S, Dash M, Al Marouf A (2019) Sentiment analysis of Bangla song review-a lexicon based backtracking approach. In: 2019 IEEE International conference on electrical, computer and communication technologies (ICECCT). IEEE, pp 1–7
16. Indhuja K, Reghu RPC (2014) Fuzzy logic based sentiment analysis of product review documents. In: 2014 First international conference on computational systems and communications (ICCSC). IEEE, pp 18–22; Elissa K (unpublished) Title of paper if known