

Ecommerce Product Rating System Based on Senti-Lexicon Analysis



Bahar Uddin Mahmud, Shib Shankar Bose, Md. Mujibur Rahman Majumder, Mohammad Shamsul Arefin, Afsana Sharmin

Abstract: E-commerce is one of the popular systems for buying and selling the products. In comment section of products that they have purchased, customer express their opinion based on the quality of product, the attitude of vendor, the delivery of product etc. This information acts as a reference for the new customers, whether they have bought the product or not. To evaluate the users' comments, sentiment analysis is played important roles where this approach not only focuses on the product itself but also the features of product itself. In this work, We have calculated the score /rating of user's sentiment for Amazon products i.e. Mobile phone; by taking the comments from the review section of product which is implied by some words or phrases, are very significant and meaningful to express users' opinion. This approach performs sentiment analysis using lexicon based approach with the help of Natural Language Toolkit (NLTK) and compare the result with the Amazon's own product rating. The experimental results prove the effectiveness of the approach.

Keywords : Product rating, E-commerce, Sentiment analysis, Lexicon, Polarity-text, NLTK.

I. INTRODUCTION

People often rely on others opinion when taking any decision and it is more critical to take decision when those choices involve resources like money and time. In that case they prefer to use previous experiences of others. Social media allows us to efficiently create and share ideas with everyone connected to the World Wide Web via forums, blogs, social networks, and content-sharing services. Public opinion about various topics are the main resource of opinion mining and sentiment analysis although these data are mostly unstructured. When an individual want to make a decision about buying a product or using a service, they have access to a huge number of user reviews, but reading and analyzing all of them is a tedious task.

Revised Manuscript Received on June 30, 2020.

* Correspondence Author

Bahar Uddin Mahmud*, Computer Science & Engineering, Feni University, Feni-3900, Bangladesh. Email:mahmudbaharuddin@gmail.com

Shib Shankar Bose, Computer Science & Engineering, NIT-Silchar, Silchar, India. Email: shibshankarbose71@gmail.com

Md. Mujibur Rahman Majumder, Computer Science & Engineering, Feni University, Feni-3900, Bangladesh. Email:mujib.contact@gmail.com

Dr. Mohammad Shamsul Arefin, Computer Science & Engineering, Chittagong, Bangladesh. Email: sarefin@cuet.ac.bd

Afsana Sharmin, Computer Science & Engineering, Chittagong University of Science & Technology, Chittagong, Bangladesh. Email: afsana.cuet@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

An organization can be benefited by obtaining the public opinion .Several services like to market company's products, identify opportunities, predict their sales and survey their reputation in the market. Heavy volume of data need to deal with in this process. With sentiment analysis techniques, it is possible to analyze a large amount of available data, and extract opinions from them that may help both customers and organization to achieve their goals. In the field of computational study, Sentiment analysis is the process that analyzes people's opinions expressed in written language. It is the field where focus of research is mostly given on the processing of text in order to identify opinionated information. Most of the existing research in natural language processing and text analysis is based on mining and retrieval of factual information which is different from this approach. There are basically two approaches available for sentiment analysis first one is called machine learning approach and the other one is called lexicon-based approach [1]. In machine learning approach a training data set is provided to the machine and it understands by setting predictions without being explicitly programmed. Where as in lexicon-based approach instead of training data dictionary or set of lexicons are used and it is assumed that the final orientation of sentence is based on the polarity of individual words in the sentence [2]. The increased popularity of E-commerce resulted in a huge accumulation of user generated data on the internet in the form of reviews, opinions and comments on different services, events and products and this trend is continually growing day by day. Various users express their views and emotions in the comments section of the ecommerce websites after buying a product or a service. 32 percent of customers on social media expect a response from companies within 30 minutes and 42% within 60 minutes [3]Propelled by rising Smartphone penetration, the launch of 4G networks and increasing consumer wealth, the e-commerce market is expected to grow rapidly in future. So this huge chunk of data generated everyday gives us an opportunity to look into a particular section of data i.e. the data generated by the user reviews on ecommerce websites and apply the technique of sentiment analysis using lexicon based approach to get some useful results to develop a recommendation system based on that for other users. This paper performs the sentiment analysis of 4 smart phone products of Amazon i.e. 5T, iPhone 6, Nokia6 and MotoGs plus, whereas for each smartphone 990 data reviews are considered and compute rating of each product as well as compare the result with Amazon's own rating.

The data is extracted using a web scrapper that is built using python followed by storing data in json format .

II. LITERATURE REVIEW

Various studies have been done on various approaches from machine learning based to lexicon-based approach and even hybrid approach which internally uses both lexicons based as well as machine learning approach. And these various studies gave different results according to the technique used. In recent time most of the data for sentiment analysis gather from several social media. It is related to user responses to particular events or products. Several approaches have introduced to analysis sentiment which includes supervised and also unsupervised approaches. Sometimes both approaches are combined with some other concepts like fuzzy concepts or neural network for improved results. D. Mrs. Sayantani Ghosh et. al. [4] have used sentiment analysis with hardware level concept to give little bit parallelism to overall process. It combined scheduling algorithm with the process to give speed over multiple processors. But it showed the drawback related to system complexity E. HuLi et. al. [5] have defined the lexicon based model for sentiment analysis of several data of social media. F. Christopher D. Manning et. al. [6] have used twitter data that is related to security for normalized lexicon based sentiment analysis. The analysis showed that it did not use universal data for the experiment. G. Mumtaz et.al. [7] have used senti-lexicon algorithm to find polarity of a review as positive, negative and neutral. They apply senti-lexicon algorithm for analyzing movie review. Major drawbacks of their work were problem in variation of spelling, opinion faking and sarcastic sentences. H. Nasim et. al. [8] have analyzed sentiments of student feedback by using machine learning and lexicon-based approaches. They have proposed a hybrid approach for analyzing the student sentiments. Drawback of their approach was that they didn't give proper explanation on multilingual content and sarcastic sentences. I. Gutam et.al. [9] performed sentiment analysis over twitter data by using machine learning approaches and sentiment analysis. In their analysis they have used Naive Bayes, Maximum entropy and SVM along with the Semantic Orientation based WordNet for extracting synonyms and similarity for the content feature. They have found accuracy of 89.9%. Aruna Sathish [10] has done sentiment analysis in E-Commerce and Information Security in order to perform this he has taken helps Natural Language Processing (NLP) to classified text as positive and negative. He has used Recurrent Neural Network and Long Short-Term Memory (LSTM) to analyze user sentiment and compared his result with Naive Bayes Classifier. He has shown that the accuracy of RNTN with increases with the size of data. K.. Kolekar et. al. [11] have analyzed sentiment and classified sentiment by using lexicon-based approach and addressing polarity shift problem. They have built their model based on antonym dictionary machine learning approach. They have built system for sentence level sentiment classification. Their proposed system also addresses and solve polarity shift problem to provide feasible solution to the BOW model in sentiment classification. They have achieved this by Detecting, Eliminating, and Modifying negation polarity shifter from a given text. Natural Language Toolkit (NLTK) is a platform of Python for programming human's natural language data in English. NLTK provides vast libraries for

users that provide tokenization, tagging, parsing, stemming, semantic reasoning and wrappers for industrial strength etc. [12]. NLTK also includes graphical demonstration of data as well as sample the data sets that aiding to users in decision making process [13]

III. RESEARCH METHODOLOGY

In this work, we have followed the following steps to figure out product rating-

- Collection of real time data for a particular product from Amazon's website (amazon.com).
- Data extraction using Web Scraper of Python.
- Preprocessed the extracted data.
- Applied sentiment analysis using NLTK on the data set.
- Rating on the scale of 0-10 and compare with Amazon's own product rating.

The process of this research is explained below-

A.Data Extraction

The data is first collected from the ecommerce site for smart phones. In our case, we have collected data from Amazon's website (www.amazon.com). Usually data of a webpage store on paragraph $\langle p_i \rangle$ tag. First, we have collected reviewer's comments on those specific products which were stored on paragraph tag.

B. Data Set Collection

Then the extracted data are stored into json file based on following parameters:

- "review header"
- "review text"
- "review comment count"
- "review posted date"
- "review rating"
- "review author"

```
Algorithm: Data extraction
Input: Received URL
begin
foreach page_no=1:99
if there is no comment
then return
else
foreach comment_no=1:10
store comment based on data collection
parameters
convert and save comment parameters into
json file
end inner foreach
end if
end outer foreach
end
```

Fig. 1. Algorithm for data extraction & collection

Data extraction algorithm is given at figure 1, where 2 foreach loops is used. Outer foreach loop used to control page number and inner foreach loop control the number of comments per page.

As we have considered 990 reviews, 10 per page, outer foreach loop run 99 times.

C. Tokenization and Score calculation

The process of converting text into tokens before transforming it into vectors is called tokenization. Tokenization helps to filter out unnecessary tokens from the sentence. The processed algorithm initializes the stop words, final score, rating, and count to zero. Now the file which we have obtained from the data extraction is given as an input to the system and the processing starts. It checks whether there are reviews present or not if not present then go the end. If review is present take the first review and tokenize it i.e. divide it into smallest independent part or word. Now filter it with the help of stop word dictionary which is feeded in the system. If the word is stop word filter it out increment count and now calculate the polarity of first word. Now apply this algorithm for all words on the list. The process is given at figure2.

```

Algorithm: Removal of stop word
Input: List of words
Output: Preprocessed-text
Begin
for each word
    search the stop word list
    if word is found
        remove the word from the sentence
    else
        do nothing
    end if
end for
end
    
```

Fig. 2. Stop word removal algorithm.

The preprocessed (filter) text after removing the stop words is further passed through algorithm given at figure 3. Figure 3 explained the process of removing emoticons, leading-trailing punctuations as well as explain the process of calculate the polarity score of sentence.

At first, we have removed the emoticons and leading and trailing punctuations from the pre- processed (filtered) text form senti-text. Then, it is checked for the intensifiers like hardly, barely etc which lead to stronger emotions followed by checking polarity by matching it in VADER LEXICON dictionary (N. Hutto et.al,2014).

After that, we check whether the word is written in Capital letters as it again counts in the polarity as it leads to stronger emotion. Furthermore, we have checked the negation term of the sentence and calculate the polarity value.

Finally it gives the sentiment value by updating polarity value based on conjunctions (i.e. and) and or destructive conjunction (i.e. but) and also normalize the score to get final score.

```

Algorithm: Removal emoticon, leading -
trailing punctuations and calculate score of
review
Input: Preprocessed text
Output: Score of review
Begin
for each preprocessed- text
    search for the emoticons and leading -
trailing Punctuations.
    if emoticons and leading-trailing punctuation
found remove these from the preprocessed-
text which form senti-text.
    else
        do nothing
    end if
end for
for each word in senti-text
    Search for stronger emotion intensifier, If
found Match this word VADER lexicon
dictionary
    If word matched with VADER dictionary
    Check the word polarity (positive negative or
neutral)
    and set the score based on polarity.
    end if
    Check for all words of a sentence and
compute sentence polarity score
    end for
Update polarity score value of sentence based
on conjunction and destructive conjunction
end
    
```

Fig. 3. Stop word removal algorithm.

D. Rating calculation

- In order to calculate final rating of product, at first, score of each review’s of a particular product are partitioned into a range which ranges from 0 to 10. Partitions are:
 - 0-2 (Poor rating)
 - 2-4 (Below average rating)
 - 4-6 (Average rating)
 - 6-8 (Good rating)
 - 8-10 (Excellent rating)
 Then calculate the rating by dividing the final score by total no. reviews (i.e. count) as like as follows:

$$\text{Rating} = \text{Round} (\text{Final- Score} / \text{No-Of-Reviews}). (1).$$

Figure 4 shows the overall system architecture. At first, reviews of product are collected from particular e-commerce site (In our case, we have chosen amazon.com).

From which, we have extracted user comments from the website, followed by storing on json file based on comments parameters.

After that, we have tokenized the sentence into words and remove stop word and emoticons from the sentence. Then, negation parts of sentence are checked, if have, and calculate polarity score based on Natural Language Toolkit (NLTK). Finally score of sentence is calculated and also computed overall product rating.



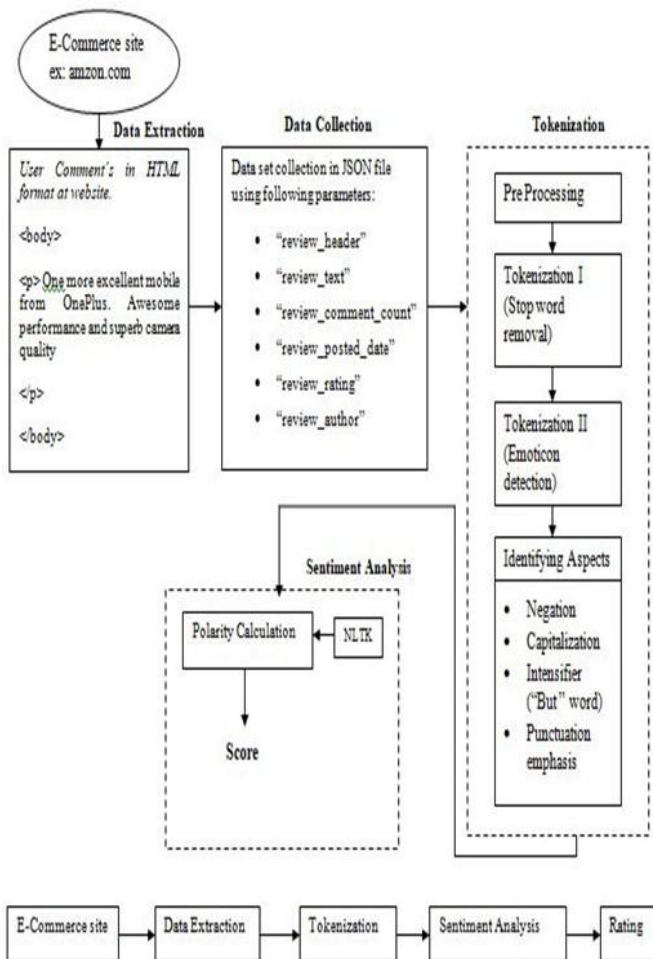


Fig. 4. System architecture of senti-lexicon analysis.

IV. RESULT AND DISCUSSION

We have collected the data set for 4 different latest smartphones from amazon i.e. OnePlus 5T, iPhone 6, Nokia6 and MotoGs plus. We have collected data set of 990 reviews for each smartphone and applied the sentiment analysis to produce the rating based on it. Table 1 shows the calculation of senti-lexicon model for oneplus 5T model exact polarity score is 7.20750154639 by rounding the value final score is obtained which is 7.0 whereas, 726 review's have polarity greater than 8, 155 comments between 6-8, 79 between 4-6, 20 between 2-4 and 10 between 0-2 ranges. Overall, For Iphone 6, Approximate rating was 7.0; 8.0 for MotoGs plus; 8.0 for Nokia 6 and 8.0 for One plus 5T. Figure 5 represent the comparison of rating between this model and amazon.com. From figure 6 it is shown that, rating of amazon for Iphone6 is 8.0 but rating 7.0 when applying this senti-lexicon approach. For OnePlus5T, 9.0 in amazon and 8.0 in senti-lexicon approach. For Nokia6, 7.0 in amazon and 8.0 in senti-lexicon approach. For MotoG5sPlus, 8.0 in both amazon and senti-lexicon approach. After analyzing the sentiment of customers, we have found that rating found from this analysis is close to the amazon rating; thus we conclude that, this approach is acceptable to compute product rating.

Table- I: RESULT FOR ONEPLUS 5T, IPHONE 6, MOTO PLUS AND NOKIA 6

Parameters	Iphone 6	MotoGs plus	Nokia 6	One plus 5T
Exact score	7.20750154639	7.6299010101	7.53837626263	8.40200151515
Final score	7.0	8.0	8.0	8.0
Total Comments	970	990	990	990
Comments >8	373	517	429	726
Comments between 6-8	380	320	395	155
Comments between 4-6	169	115	108	79
Comments 2-4	46	30	40	20
Comments 0-2	2	8	18	10

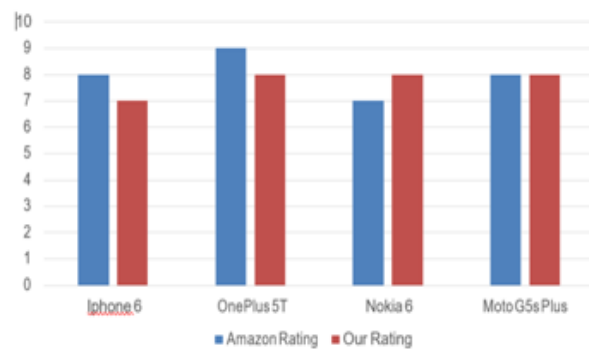


Fig. 5. Comparison of Our Results vs. Amazon Rating.

V. CONCLUSION

In this paper, after taking reviews for smart phones from Amazon and applying sentiment analysis using lexicon-based approach with the help of Natural Language Processing Toolkit (NLTK). From result section it is shown that several smart phones show several rating based on their polarity calculation. Senti-lexicon model calculates the polarity of comments based on their analysis of words. And with the help of NLTK lexicon based sentiment analysis performed on reviews and rate them between 1 to 10. This method would be applicable alternative of popular machine learning approach in which, this senti-lexicon based approach matched the extracted word with dictionary of sentiment or VADER lexicon dictionary. Whereas, machine learning approach uses previously labeled data to determine the sentiment of never-before-seen sentences. However, the excellent things about machine learning approach is that greater the volume of data, better the accuracy.

REFERENCES

1. Nigam, Nitika, Yadav, Divakar, "Lexicon-Based Approach to Sentiment Analysis of Tweets Using R Language", 2018, Second International Conference, ICACDS 2018, Dehradun, India, April 20-21, 2018, Revised Selected Papers, Part I. 10.1007/978-981-13-1810-8 16
2. Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, ch.1, 2012

3. Jay Baer, "42 Percent of Consumers Complaining in Social Media Expect 60 Minute Response Time", [Online], Access date: 14/06/2019
4. Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K., "A tutorial review on Text Mining Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Pages: 223-233, Vol. 1, Issue 4, June 2012.
5. HuLi, Yong Shi "WordNet based lexicon model for CNL" 2009 IEEE proceeding at 2009 International Conference on Test and Measurement.
6. Christopher D. Manning, Prabhakar Raghavan, Hinrich Sch'utze, "Introduction to Information Retrieval", ISBN-13 978-0-511-41405-3, 2013.
7. Mumtaz, Deebha, and Bindiya Ahuja. "Sentiment analysis of movie review data using Senti-lexicon algorithm." Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on. IEEE, 2016.
8. Zarmeen Nasim, Quratulain Rajput and Sajjad Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches", 2017, International Conference on Research and Innovation in Information Systems (ICRIIS). doi:10.1109/icriis.2017.8002475 .
9. Geetika Gautam and Divakar yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis", 2014, Seventh International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2014.6897213 .
10. Aruna Sathish. "Sentiment Analysis in E-Commerce and Information Security", 2017, International Journal of Innovative Research in Computer and Communication Engineering.
11. Nilam V. Kolekar, Prof.Gauri Rao, Sumita Dey, Madhavi Mane, Veena Jadhav And Shruti Patil, "Sentiment analysis and classification using lexicon-based approach and addressing polarity shift problem", 2016, Journal of Theoretical and Applied Information Technology 90.1.
12. S. Bird, I. S. Bird, E. Klein, E. Loper "Natural Language Processing with Python", Sebastopol: O'Reilly Media, Inc., pp. 504, 2009 .
13. Steven Bird, "NLTK: The natural language toolkit", 2006, 21st International Conference on Computational Linguistics, Proceedings of the Conference.
14. Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule- based model for sentiment analysis of social media text." Eighth international AAAI conference on weblogs and social media. 2014.



Afsana Sharmin, is working as a Software Developer. She has graduated from Chittagong University of Engineering and Technology, Bangladesh in 2017. Now she is pursuing her Master degree from same university. Her research interests are Machine Learning, Data Mining and Data

AUTHORS PROFILE



Bahar Uddin Mahmud, Lecturer, Department of Computer Science & Engineering, Feni University, Bangladesh. He received his Bachelor degree in CSE from Maulana Azad National Institute of Technology, Bhopal, India in 2013 and earned Master degree in Information Technology from Dhaka University, Bangladesh in 2016.

He has couple of conference and journal paper. His research interests are Machine Learning, Data Science and Deep Learning .



Shib Shankar Bose, pursued his Bachelor degree from NIT, Silchar and now working in Wipro Ltd as a Software developer. His research interests are Machine Learning, Data Science



Md. Mujibur Rahman Majumder, Lecturer, Department of Computer Science & Engineering, Feni University, Bangladesh. He received his Bachelor degree and Masters degree in ICT from Comilla University ,Bangladesh. His research interests are Data mining, Machine Learning



Dr. Mohammad Shamsul Arefin, is affiliated with the Department of Computer Science and Engineering (CSE), Chittagong University of Engineering and Technology, Bangladesh. Earlier he was the head of the department. He received his Doctor of Engineering Degree in Information Engineering from Hiroshima

University, Japan with support of the scholarship of MEXT, Japan. As a part of his doctoral research, Dr. Arefin was with IBM Yamato Software Laboratory, Japan. His research includes privacy preserving data publishing and mining, distributed and cloud computing, IoT and big data management, multilingual data management, semantic web, object oriented system development and IT for agriculture and environment.