IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Rating and Analysis of Customer Reviews Using Natural Language Processing

Ms. B.PRIYANKA [#1], Mr. Y. AMARAIAH [#2] , Mr. D.D.D.SURIBABU [#3]

[#1]M.Tech Student, [#2] Associate Professor, [#3] Associate Professor & HOD

Department of Computer Science & Engineering,

International School of Technology and Sciences (Women),

(Affiliated to JNTUK),East Gonagudem, Rajanagaram, Rajamahendravaram,

East Godavari District, Andhra Pradesh, India– 533294.

ABSTRACT

In order to predict rating value, a variety of supervised and unsupervised machine learning models are used to a dataset of Amazon product reviews in the following article. A "star" rating on a scale of 1 to 5 is the intended response to the input of a text review. This project might be characterised as a regression variation of sentiment analysis because we want to construct a continuum of sentiment rather than just polarity. Our objective is to create the most adaptable and precise model that can manage the broadest range of conflicting and polarising opinions presented in reviews .In addition to experimenting with various model types, we looked at the effects of seven distinct embeddings, each of which had unique built-in considerations for information like word index, directional significance, context, and frequency. These embeddings included simpler encoding alternatives like Bag of Words and TF-IDF, as well as pre-trained word embedding models like Naive Bayes and LSTM and embedding schemes generated from scratch. We trained three distinct deep learning networks that were more effective as well as supervised boosting models from Light GBM and CatBoost.

KEYWORDS:

Naive Bayes,CatBoost,LSTM,Bag of Words, Unsupervised Machine Learning.

1. INTRODUCTION

In today's ever connected world, the way people shop has changed. People are buying more and more over the Internet instead of going traditional shopping. E-commerce provides customers with the opportunity of browsing endless product catalogues, comparing prices, being continuously informed, creating wishlist and enjoying a better service based on their individual interests. This increasing electronic market is highly competitive, featuring the possibility for a customer to easily move from one e-commerce when their necessities are not satisfied [1], [2]. As a consequence, e-commerce business analysts require to know and understand consumers' behaviour when those navigate through the website, as well as trying to identify the reasons that motivated them to purchase, or not, a product [3], [4], [5]. Getting this behavioural knowledge will allow e-commerce websites to deliver a more personalized service to customers, retaining customers [6] and increasing benefits [7]. However, discovering customer' behaviour and the reasons that guide their buying process is a very complex task [3].

E-commerce websites provide customers with a wide variety of navigational options and actions: users can freely move through different product categories, follow multiple navigational paths to visit a specific product, or use different mechanisms to buy products, for example. Usually, these user activities are recorded in the web server

logs [3], [8]. Web server logs store, in an ordered way, the sequence of web events generated by each user (commonly known as clickstreams). The very valuable users' behaviour is hidden in these logs, which must be discovered and analysed [9]. A correct analysis can be subsequently used to improve the website contents and structure [10], to adapt and personalize contents [11], [12], [13], to recommend products [14], [15], or to understand the interest of users in specific products [16], for instance.

Data mining techniques have proved their usefulness for discovering patterns in log files (when applied to the analysis of web server logs the term web usage mining [17] is used). Its main goal is to discover usage patterns trying to explain the users' interests. Different techniques have been successfully used in the field of e-commerce, such as classification techniques, clustering, association rules or sequential patterns [18], [19]. In many application domains these techniques are used in conjunction with process mining techniques. Such techniques are part of the business intelligence domain and apply specific algorithms to discover hidden patterns and relationships in large data sets [20]. An e-commerce website is an open system where almost any customer behaviour is possible.

2. EXISTING SYSTEM AND ITS LIMITATIONS

Now a days due to pandemic online shopping become more demanding. In this online system to judge quality of product we will take its rating into consideration for this review rating based of written comment we are not having automated way to predict rating accurately currently whatever the system we have so far implemented based on NLP with combination machine learning algorithms those are SVM,KNN,Naïve Bayes and Random Forest Algorithm and achieved an accuracy of 81% with Random Forest Algorithm Draw back with this existed system is it will predict whether the given review is negative or positive it is not sufficient to get a proper information regarding product we need to give rating based on given review from 1 star to 5-star it is not achieved so far to sort-out this problem we are going contribute our portion of knowledge.

LIMITATION OF PRIMITIVE SYSTEM

1. More Time Delay in finding the rating of given product
2. All the existing approaches are manual approach for predicting the rating.
3. There is no automated approach for finding the rating on given product.
4. All the existing methods use normal ML algorithms for finding the rating of given product.

3. PROPOSED SYSTEM AND ITS ADVANTAGES

In this project after downloading data from Kaggle we are analyzing and understanding the data by performing data analysis and exploratory data analysis by sing pandas and matplotlib and Seaborn after that by taking help of NLP, Beautiful soup and re modules we are pre-processing the text information and we are converting text in the form of numerical vectors by applying count vector and word2Vec models after refining the data we are applying Naive Bayes and LSTM Architectures on the pre-processed data and testing these two models performance naïve bayes algorithm is giving best accuracy so we are recommending this model for real time rating prediction based on given review.

ADVANTAGES OF THE PROPOSED SYSTEM

1) By using LSTM and Naïve Bayes algorithms we try to gather rating for text reviews.

2) In this work we are going to find comparison of two algorithms in order to identify rating for the text reviews.

3) Our comparative results clearly state that Naïve Bayes has more accuracy compared with LSTM.

4. IMPLEMENTATION PHASE

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes Google Collaboratory and the modules are Import Libraries, Gathering Data, Pre-Processing and Rating Prediction.

1) IMPORT LIBRARIES MODULE

Here we try to import all the necessary libraries which are present to design the application and here we import all necessary libraries which are required for finding rating on text reviews.

2) DATA GATHERING MODULE

Here we try to load the dataset from kaggle website using Json file. Here the dataset what we use for our application is product review dataset. We obtained our data from the Amazon Customer Reviews Dataset provided by Amazon Public Datasets, which contains official reviews from shoppers at Amazon.com. Here are the columns of the dataset along with a brief description of each

- ***customer_id*** — random identifier that can be used to aggregate reviews written by a single author
- ***review_id*** — the unique ID of the review
- ***product_id*** — the unique Product ID the review pertains to
- ***product_title*** — title of the product
- ***product_category*** — Broad product category that can be used to group reviews
- ***star_rating*** — the review rating on a one to five star scale
- ***verified_purchase*** — the review is on a verified purchase
- ***review_headline*** — the title of the review
- ***review_body*** — the review text
- ***review_date*** — the date the review was written

We decided to focus on the electronics category dataset of reviews because we thought that reviews for these products would contain more objective evaluations based on concrete characteristics compared to more subjective product categories such as books.

3) DATA PRE-PROCESSING

Data pre-processing is a technique that is used to convert raw data into a clean dataset. A general rule of the thumb is to assign 80% of the dataset to training set and therefore the remaining 20% to test set.

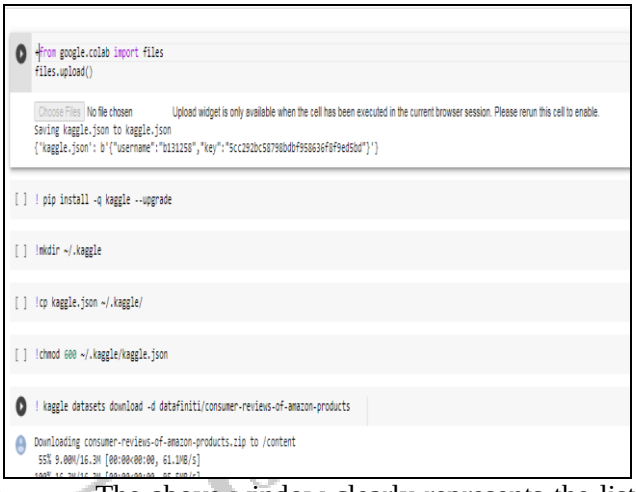
4) RATING PREDICTION

Here we try to apply Naïve Bayes and LSTM for text rating prediction on given input data and then we can able to perform the rating on text reviews.

5. EXPERIMENTAL RESULTS

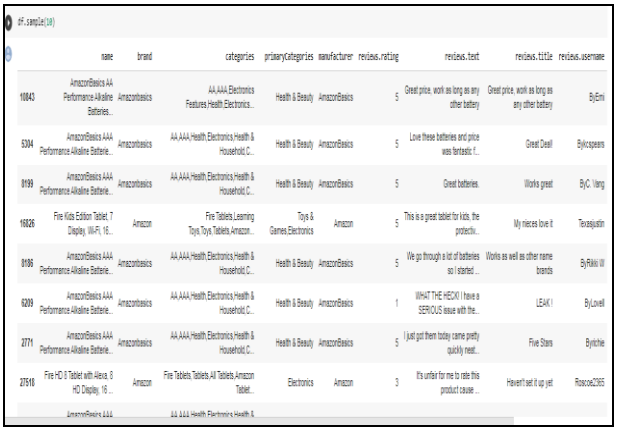
In this section we try to design our current model using Python as programming language and we used Google Collab as working environment for executing the application. Now we can check the performance of our proposed application as follows:

IMPORT LIBRARIES



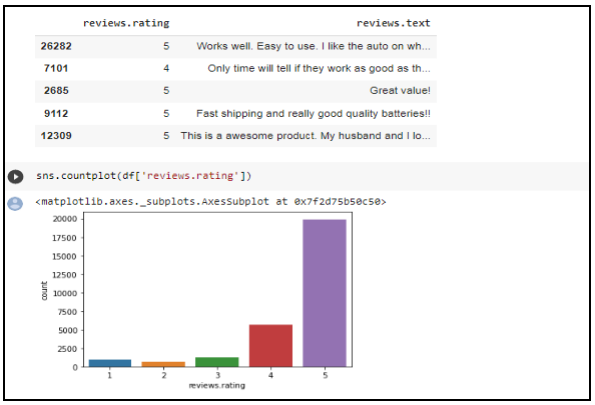
The above window clearly represents the list of several modules used in our application.

PRE-PROCESS THE DATA



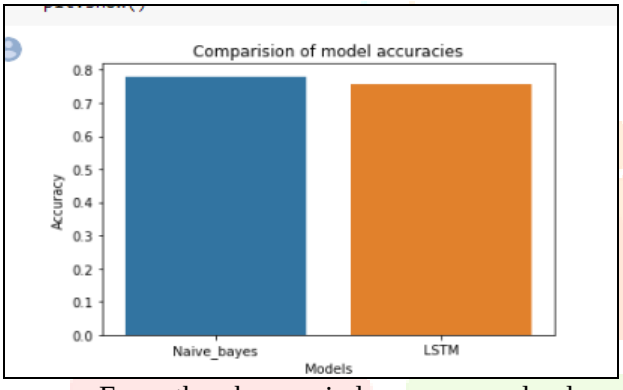
From the above window we can see DATA is pre-processed and any incomplete data is removed.

VIEW RATING CHART



From the above window we can clearly check the rating chart for products.

COMPARITIVE GRAPH



From the above window we can clearly see the comparative graph between two algorithms and we can finally see Naive Bayes has best accuracy.

6. CONCLUSION

With more than 1000 collective data, we analyzed product review in Bangla language. Further analysis of the data shows that our proposed system can detect proper sentiment off of Bangla review or comments effectively. From selected few popular classification algorithms KNN, Decision Tree, Support Vector Machine (SVM), Random Forest and Logistic Regression, SVM performed outstandingly with 88.81% accuracy. We believe that our proposed system can decrease customer ordeal while shopping online as they are able to see through the system for product reviews by ratio of previous customers' positive and negative feedback. It can be also proven helpful for the seller because he can identify defects of his products and provide better customer service.

7. REFERENCES

[1] R. Moslem. (Apr 18, 2017). A Brief History of E Commerce in Bangladesh. Available: https://medium.com/@r_moslem/a-briefhistory- of-e-commerce-in-bangladesh.

[2] O. Sharif, M. M. Hoque, and E. Hossain, "Sentiment Analysis of Bengali Texts on Online Restaurant Reviews Using Multinomial Naive Bayes," in 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019,pp. 1-6: IEEE.

[3] V. Ramanathan and T. Meyyappan, "Twitter text mining for sentiment analysis on people's feedback about oman tourism," in 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), 2019, pp. 1-5: IEEE.

[4] N. Banik and M. H. H. Rahman, "Evaluation of naive bayes and support vector machines on bangla textual movie reviews," in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018, pp. 1-6: IEEE.

[5] N. I. Tripto and M. E. Ali, "Detecting multilabel sentiment and emotions from bangla youtube comments," in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP),2018, pp. 1-6: IEEE.

[6] L. Nahar, Z. Sultana, N. Jahan, and U. Jannat, "Filtering Bengali Political and Sports News of Social Media from Textual Information," in 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019,pp. 1-6: IEEE.

[7] P. Ray and A. Chakrabarti, "Twitter sentiment analysis for product review using lexicon method," in 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), 2017, pp.211-216: IEEE.

[8] K. Indhuja and R. P. Reghu, "Fuzzy logic based sentiment analysis of product review documents," in 2014 First International Conference on Computational Systems and Communications (ICCSC), 2014, pp. 18-22: IEEE.

[9] M. P. Anto, M. Antony, K. Muhsina, N. Johny, V. James, and A. Wilson, "Product rating using sentiment analysis," in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 3458-3462: IEEE.

[10] V. Vapnik, The nature of statistical learning theory. Springer science & business media, 2013.

[11] R. E. Wright, "Logistic regression," 1995
.
[12] P. O. Gislason, J. A. Benediktsson, and J. R. J. P. R. L. Sveinsson,"Random forests for land cover classification," vol. 27,no. 4, pp.294-300, 2006.

[13] C. Jin, L. De-Lin, and M. Fen-Xiang, "An improved ID3 decisiontree algorithm," in 2009 4th International Conference on ComputerScience & Education, 2009, pp. 127-130: IEEE.

[14] Y. Yang and X. Liu, "A re-examination of textcategorizationmethods," in Proceedings of the 22nd annual international ACMSIGIR conference on Research and development in informationretrieval, 1999, pp. 42-49.