Turkish Journal of Computer and Mathematics Education *Vol.11 No.01 (2020), 901-907*
DOI: https://doi.org/10.17762/turcomat.v11i1.13588

Research Article

Quantitative Evaluation of Sentiment Analysis on E-Commerce Data

Aditya Verma
Department of Computer Science & Information Technology, Graphic Era Hill University, Dehradun Uttarakhand India 248002

Abstract
With the emergence of technology and Big Data in the modern world, it is important to govern such vast amounts of data. Amazon and other e-commerce websites play a vital role in the delivery of goods services to the user. Nonetheless, such services are frequently accompanied by consumer reviews and ratings of the things they sell. Such reviews and ratings provided in the form of textual feedback serve to improve service delivery and product quality in the event that a consumer is dissatisfied. Thus, the process of evaluations and ratings is regarded as a significant component of the customer satisfaction process. Thus, it is vital to analyse them so that useful insights can be gleaned from them and the correct judgements can be made regarding the enhancement of the product. With emphasis on this concept, the authors of the proposed paper intend to develop a model capable of performing sentiment analysis on customer-provided product reviews and analysing products with positive and negative comments. On the basis of these reviews, three machine learning algorithms, primarily Decision Trees, XGBoost, and AdaBoost, are utilised to undertake sentiment analysis on products. It was noticed that XGBoost produced a greater detection accuracy of 84.45%.
Keywords: Amazon, Aspect level, machine learning, sentiment analysis

Introduction

In the past decade, there has been a significant paradigm shift as individuals have adopted the practise of shopping online. These services are supplied by commercial websites that aid traders, sellers, and consumers in the exchange of their products. This involves the process of user reviews and ratings, which may inspire other consumers to purchase similar products. This has caused the Internet to surpass all other information sources and become reliant on client reviews. Thus, it can be concluded that this procedure has in a sense transformed the E-Commerce environment and made the marketplace more accessible to buyers [1]. Amazon, as the most popular e-commerce website, features customer reviews and ratings for the desired products. Such customer reviews are regarded as valuable feedback and opinions that may influence other users' decisions to purchase the same goods. This not only encourages people to purchase the products, but also helps shops understand the specifics of each product and create it properly. Amazon's evolution has also lowered prices, allowing them to remain competitive with other e-commerce platforms. Consumers will now be able to compare the cost prices of products and consequently make purchases in accordance. This had led to an expansion of the product rating system as a whole. Conversely, the reviews, ratings, and feedback gathered from customers assist vendors to strengthen their products and services. As hundreds of reviews are posted on many products, this facilitates the decision-making process for buyers who are hesitant to purchase anything online [2].

When the amount of reviews and ratings increases, it becomes increasingly difficult for legitimate customers to acquire a product. Moreover, good and negative reviews of the same product mislead customers who are deciding whether or not to purchase. With such a vast expansion of data on the internet, it becomes vital to analyse this dynamically growing data of evaluations and recommend to clients what they should purchase in the most straightforward manner. In such a scenario, the existing customer evaluations help estimate what a potential new client will likely purchase. Yet, it is important to highlight that Amazon is generating an exponential amount of data online, which must be managed efficiently so that the customer reviews and ratings may be used to provide insight to other customers.

The topic of sentiment analysis has been widely used on E-Commerce platforms and has assisted several customers in making informed judgements regarding goods purchases. Since every customer reads reviews and comments of a product before making a purchase, this information is seen as vital. Companies and businesses utilise these reviews to gain important insights into client behaviour and to offer them products that will compel them to purchase. However, such comments and feedback can often be deceptive, since the review may comprise service reviews rather than product reviews, so giving the user a false impression and discouraging

him from purchasing the goods. In such a situation, it is necessary to conduct a comprehensive investigation of a customer's feelings regarding the acquisition of a product. The influence of emotions, feedback, and reviews is directly proportionate to a customer's likelihood of making a purchase. The explicit connection between user emotions and the product or service must be analysed so that the aforementioned issues can be resolved. The major objective of a recommender system, on the other hand, is to provide a list of products that would be relevant to clients with similar preferences. This technique of analysing client feelings and providing responses in the form of recommendations is crucial to the overall profitability of an e-commerce website such as Amazon.

The objective of this study is to do experimental analysis on Amazon product data. Customer reviews play a significant part in influencing the purchasing decisions of other consumers who are considering purchasing a comparable product. So, we are developing a model capable of reading client evaluations and classifying them as good or negative feedback. In addition, we propose implementing an automatic recommendation system that might suggest products with similar preferences to different clients. But, the purpose of the research is to adopt machine learning techniques to anticipate the sentiments gleaned from product reviews and then to execute these strategies.

Related Works

The authors of [3] created a methodology capable of identifying the sentiments associated with a product evaluation. The fundamental concept was to do sentiment analysis on textual inputs so that the obtained words could be classed as positive or negative. The major purpose of the study was to draw conclusions from lengthy texts and then categorise them as good or negative thoughts. The author also suggested the use of machine learning methods to detect the same anomalies. An experimental investigation of the Naive Bayes and Support Vector Machine (SVM) algorithms was conducted; the analysis was conducted using the dataset received from Amazon. The procedure involved removing stop words and punctuation so that the algorithm could achieve the needed level of precision. From the dataset, a total of 4,783 negative words and 2,006 positive terms were identified and categorised.

Similar research was seen in [4], in which the authors conducted an Aspect-Based Sentiment Analysis on words collected from Internet customers. Customers provided ratings and opinions of the products they purchased as part of their feedback. In the presented study work, the authors briefly described the identification, detection, and classification processes. The study also offered information regarding the pre-processing phases of tokenization, lemmatization, and bag of words. However, the authors' methodology involved analysing attitudes at the phrase level and not the document level. The dataset utilised for the execution of the thesis was retrieved from the Kaggle repository and consisted of reviews from six distinct classes.

Occasionally, sentiment analysis is also known as opinion mining. In a research report published in [5], the authors recommended extracting the emotions of users who offered feedback in textual formats. The authors presented extremely capable tools and algorithms for producing the requisite precision. However, the planned project was implemented for business ventures in which input from consumers of different genders and ages was collected. The objective of this exercise was to determine the corporate working environment and the feelings of employees regarding their job satisfaction. The authors in [6] employed machine learning-based supervised algorithms capable of predicting the evaluations and ratings offered by end users. Implementation was performed such that the obtained text could be transformed to numerical representations and binary data could be obtained. The conversion of text to binary format was required so that the machine could interpret the user's intent within the textual format. The full data set was divided into a 70:30 ratio, and evaluation metrics such as the accuracy and recall factors were utilised.

In a research article presented in [7] a survey of collaborative filter approach was undertaken. Throughout the study, kinds of filtering as well as their merits and downsides were discussed in depth. Also detailed were the functionalities combined with the predictive model. The author has highlighted a quick summary of mathematical formulas related to the system's functionality.

Proposed Methodology

The primary objective of the research paper is to conduct the process of sentiment analysis on customer reviews being provided on Amazon. The reviews of the customer are used to categorise and classify the customer sentiments as positive or negative with respect to the products they purchase on the E-Commerce site. Through this categorisation of product reviews being made by the customer; recommendation of similar products with customer likes and dislikes are recommended to other customers while they make a purchase. This enhances the overall system model to achieve and calculate sentiments thus attached to the customer reviews and thereby recommend other customers with respect to the same. In the initial phase; a data is collected from a repository and further pre-processed. The stage if pre-processing is done so as to remove unnecessary data and redundancy from the dataset thus obtained. This process also involves removal of noise and irrelevant information so as to clean the dataset from redundancy. This stage is primarily carried out so that the overall performance of the system model can be enhanced and various phases of this research paper could be carried out in an efficient manner. This stage of data pre-processing is further followed by removal of errors, altering the punctuation of the words and sentiments thus obtained. It also includes a process of lemmatization and stemming wherein tokens of words are used to reduce the original word and later perform NLP techniques into it to process the data. The utilization of NLP is done so as to improve the overall system performance of the model and thereby increase the accuracy thus obtained through the implementation of machine learning algorithms which are to be executed. In addition to the execution of NLP toolkit; the usage of Snowball as a medium to perform word stemming and lemmatization also occurs in the stage of pre-processing a data. Once the data is obtained in a respective format; the further process interpreting takes place; wherein the conceptual theory of TF-IDF is used. TF-IDF primarily converts the textual format of the sentiments into its respective numerical figure so that the machine can further interpret and classify them as either positive or negative. In addition to the implementation of TF-IDF; the authors have also proposed the usage of VADER so as to calculate the sentiment score of the product reviews given by the customer. The obtained and calculated sentiment score is used to classify the sentiments as either positive or negative so that further labelling of the process can be done in an effective manner. The value of sentiment score tends to range and fall between 0 to 0.5. If the value exceeds more than 0.5; the sentiment is classified as 1 and further labelled as a negative comment; whereas on the other hand; if the value thus calculated lies in less range of 0.5; the score is calculated to be 0 which is further labelled as a positive comment. Based in this numbering and libelling technique the sentiments thus attached to the customers being made on Amazon for the respective products they purchase is made. After this stage; the dataset undergoes the process of training and testing wherein the data is run on three machine learning algorithms; namely; decision trees, XGBoost and AdaBoost. After this stage; the respective experimental analysis of the same is performed using evaluation parameters of confusion matrix and the model which generates the highest accuracy is thus declared as the optimised model.

Figure 1 below explains the architecture diagram of the proposed methodology:

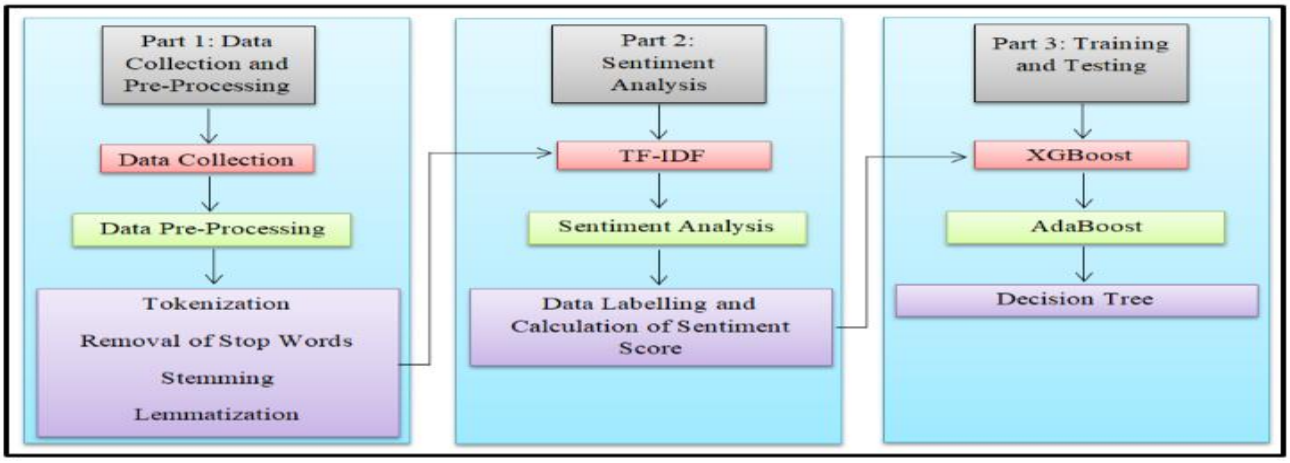

Figure 1: Architecture of the Proposed Methodology

Implementation Details

This section of the paper briefs on the steps involved to accomplish the proposed research study.

A. Data Collection

The process of data collection is done by obtaining the dataset from the Kaggle repository and performing machine learning algorithms on it to process and execuet the data. The dataset thus obtained contains of two csv files as train and test wherein cusomer reviews with their respective sentiments thus attached is present. The sentiments of customer available on Aamazon has been collected in the dataset and labelled as the "baby.json" dataset. An entire analysis of sentimental calculation of score using VADER is published in the dataset which includes of 160,763 customer reviews being posted about the same. Figure 1 below depicts the "baby.json" dataset obatined from Kaggle repository.

	reviewerID	asin	reviewerName	helpful	reviewText	overall	sum
0	A1HK2FQW6KXQB2	097293751X	Amanda Johnsen "Amanda E. Johnsen"	[0, 0]	Perfect for new parents. We were able to keep ...	5	Awe
1	A19K65VY14D13R	097293751X	angela	[0, 0]	This book is such a life saver. It has been s...	5	Shou require all par
2	A2LL1TGG90977E	097293751X	Carter	[0, 0]	Helps me know exactly how my babies day has go...	5	Grandm wat

Figure 2: Dataset Used

B. Data Pre-Processing

Once the dataset is collected from the respective repository; the obtained data undergoes a series of pre-processing wherein the data is refined for irrelevant and redundant data. This is done so as to filter unnecessary columns and data from the dataset so that it does not cause a hindrance while experimental analysis of the same using machine learning algorithms. The raw data thus collected is cleaned and a filtered format of textual reviews is obtained. This process helps to derive insights from historical data and establish certain patterns so that the respective noise from the same can be discarded. However, the process of removal of this noise is established by cleaning the dataset using crawlers so that the overall performance of the system model can be further enhanced. This stage of data pre-processing also involves removal of unnecessary data, elimination of redundant words, discarding NULL values, removing punctuations and outliers which are thereby present. In order to carry out this process of execution, a conceptual theory of polarity is thus developed wherein the obtained data is performed on the csv files thus obtained from the repository. The csv files contain train and test data and the step of pre-processing is thereby performed on the same. Once this stage of data pre-processing is achieved; training and testing of the dataset is done on the machine learning algorithms thus selected. For the purpose of implementation of the proposed research paper; decision trees, XGBoost and AdaBoost are used for the same.

C. Data Labelling

The process of data labelling is performed after the data is sent of pre-processing as seen in the previous stage. Labelling the dataset in the system model is an essential step since the obtained data is categorized as either positive or negative by depending on the labels thus assigned to them. This process is majorly carried out through binarization wherein the obtained data is either categorized as 0 or 1. The 0 label refers to the positive reviews thus given by the customer on Amazon and the 1 label thus refers to the negative reviews thus gives by

the customer on Amazon. However; the process of analysing whether the reviews must be binarized as 1 or 1 is done by carrying out the process of calculating sentiment scores using VADER as the technique. This sentiment score is calculated through mathematical operation of determining the values in the range of 0.5. If the obtained mathematical value tends to be greater than 0.5; the review is rated as 1 and if the mathematical value tends to lesser than 0.5; the review is further rated as 0. This calculation of 0 and 1 is done to categorise the obtained sentiments as either positive or negative. However; the entire execution of the process is carried out through the implementation of VADER as the NLP toolkit.

D. Data Visualization

The process of data visualization is primarily done to visualize the obtained data from the repository. This is done so as to get an overview and define whether the obtained data is balanced or not. For the implementation of the proposed research paper; it has been observed that the obtained data from the dataset was imbalanced in nature; and hence for that reason the authors extended their work by using SMOTE. SMOTE is a technique that is primarily used to balance an imbalanced dataset. Once the process of data balance is performed; the data is further sent for training and testing purpose; wherein the respective algorithms are used to perform the same.

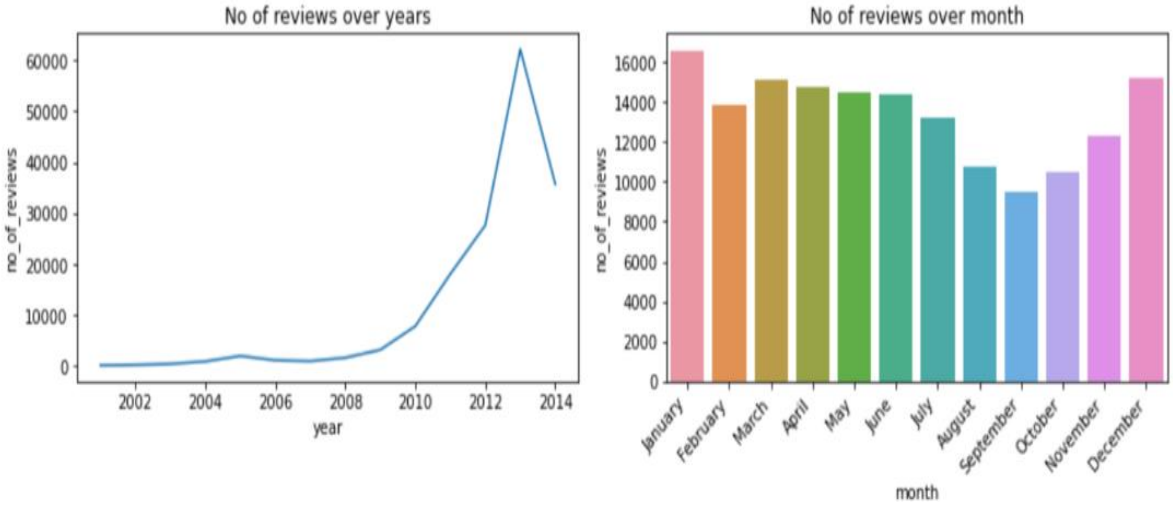

Figure 3: Data Visualization

Results

This section of the research paper briefs on the results thus obtained on performance of three machine learning algorithms to detect E-Commerce data being analysed on Amazon through sentiment analysis. The product reviews thus submitted to the machine for further processing is trained and tested over decision trees, XGBoost and AdaBoost. A confusion matrix is thus generated and values of TP, FP, TN and FN are calculated. The confusion matrix gives information on the number of values thus correctly predicted by the machine during the testing phase with respect to the actual values thus given in the initial stage. Figure 4 below depicts the generation of confusion matrix of three machine learning based algorithms.

DOI: https://doi.org/10.17762/turcomat.v11i1.13588

Research Article

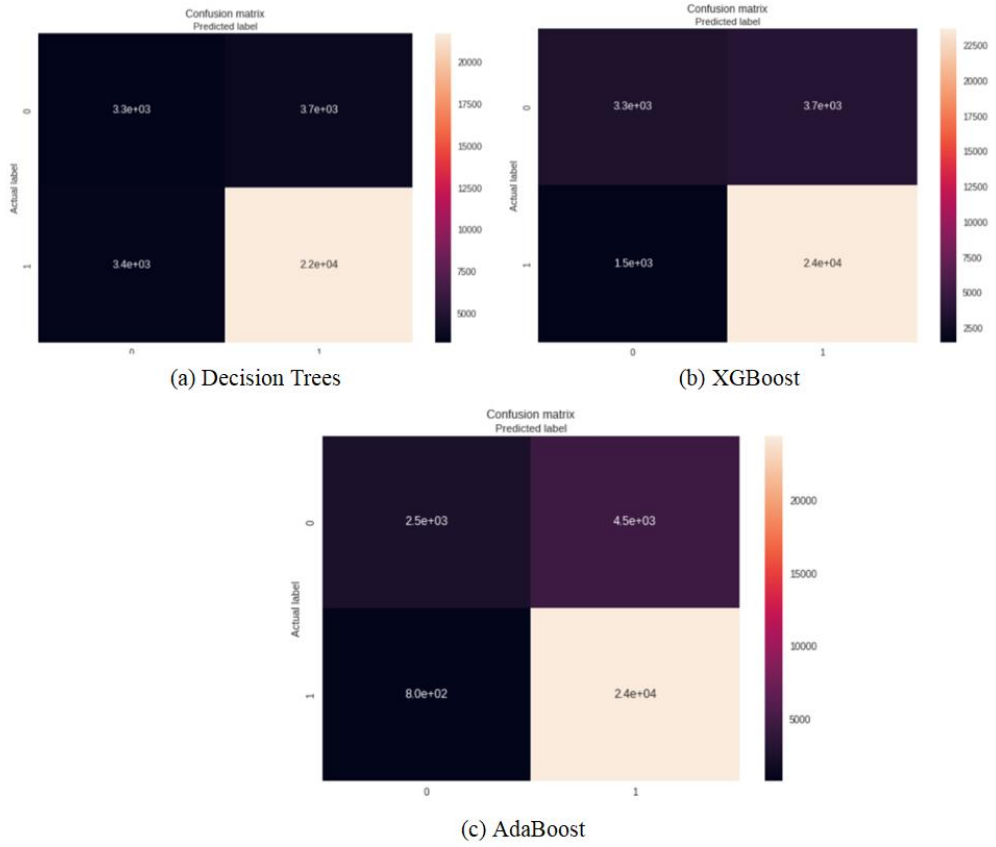

Figure 4: Confusion Matrix

Conclusions

The primary objective of the research paper is to perform the process of sentiment analysis on customer reviews being provided on an E-Commerce platform such as Amazon. For this purpose, the dataset of amazon reviews was collected and gathered. The customer reviews were then filtered for redundant data and further pre-processed for next stages. The process also included a step of data visualization wherein the obtained data was visualized on the basis of graphs thus obtained. Through sentiment analysis; the reviews given by customer on Amazon were binarized and labelled as 0 and 1 wherein 0 represented positive comments and 1 represented negative comments. The comments were further used to train and test the data on three machine learning based algorithms. Decision trees, XGBoost and AdaBoost were used for the purpose of implementation. Through experimental analysis it was observed that the executional implementation of XGBoost generated an optimised accuracy of 84.45 percent and was thereby declared to be as the highest in comparison to the algorithms thus used.

References

[1] Ahlgren, Oskar (2016). “Research on Sentiment Analysis: The First Decade”. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)

[2] Gupta, Pankaj, Ritu Tiwari and Nirmal Robert (2016). “Sentiment analysis and Text Summarization of Online Reviews: A Survey”. In: 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 0241–0245

[3] Jagdale RS, Shirsat VS, Deshmukh SN. Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing, Springer, Singapore, pp. 639–647, 2019

[4] Vamshi KB, Pandey AK, Siva KA. Topic model based opinion mining and sentiment analysis. In International Conference on Computer Communication and Informatics (ICCCI), IEEE, pp. 1–4, 2018

[5] S. Tan and J. Zhang, “An empirical study of sentiment analysis for chinese documents,” Expert Systems with applications, vol. 34, no. 4, pp. 2622–2629, 2008

[6] A modern approach for plant leaf disease classification which depends on leaf image processing CG Dhaware, KH Wanjale 2017 International Conference on Computer Communication and Informatics (ICCCI)
[7] Social re-ranking of image based on visual and semantic information P Phursutkar, K Wanjale 2017 8th International Conference on Computing, Communication and Networking (ICCCNT)
[8] Shivadekar, S., Abraham, S. R., & Khalid, S. (2016). Document validation and verification system. Int. J. Adv. Res. Comput. Eng. Technol.(IJARCET), 5(3).
[9] Jeysudha, A., Muthukutty, L., Krishnan, A., & Shivadekar, S. (2017). Real Time Video Copy Detection using Hadoop. International Journal of Computer Applications, 162(9), 42-45.
[10] Kokare, R., & Wanjale, K. (2015). A natural language query builder interface for structured databases using dependency parsing. International Journal of Mathematical Sciences and Computing, 1(4), 11-20.