PREDICTION OF ONLINE PRODUCTS RATING FROM CUSTOMER FEEDBACK USING RANDOM FOREST CLASSIFIER

NALLANUKALA.SAI MRUDULA[1], CHILUKURI.JYOTHI NAGA SANDHYA[2], GUNTUPALLI.BHAGYA LAKSHMI[3], LAKKARAJU.JAYA DURGA[4], Student, Department of CSE, NRI INSTITUTE OF TECHNOLOGY, Vijayawada, A.P., India.

B.VENU GOPAL Associate professor, Department of CSE, NRI INSTITUTE OF TECHNOLOGY, Vijayawada, A.P., India.

ABSTRACT

An online product's rating is an essential metric to understand the acceptability of that product to users. Shoppers use the rating to measure the quality andexcellence of the online product. It helps an online shopper to decide whether to buy a product or not. It also helps the producer to further modification of that product during reproduction. Sometimes people buy products through online and give a text review but apathy towards giving a number rating, commonly a star rating. But producers need to know the rating ofproducts for analysis of their business. Producers can use this rating for business analysis and can drive better revenue to their business. We have used supervised machine learning approach to predictratingbased on customer text review by applying Random Forest classifier, with TF-IDF vectorizer from extensive and series of experiments. We applied the abovealgorithm on the dataset named "Amazondataset".

We proposed to use the Random Forest Classifieralgorithm with TF-IDF Vectorizerfortheclassificationofrating.

Thefirststepof project is here we use methodistopre-processthedataofthedataset.Inthesecond step,weappliedthefeatureextractiontechnique, whichisTF- IDF. In the third step, we used three supervised machine learningalgorithmsforpredictingratingbasedont extreview. Finally, we compared the results provided by the machine learningapproachandfoundoutwhichoneproduc esabetter result in the applied dataset. In the dataset, user rating varies user to user. Same text review can get a different numbered rating from theuser.

INTRODUCTION

The fast development of Internet business and administrations is changing the manner in which individuals live, including posting their perspectives, considerations, remarks, and surveys on different sorts of sites or discussions, like web based shopping sites (Amazon.com, Bestbuy.com), voyaging sites (TripAdvisor.com, Expedia.com), film rating sites (IMDb.com, RottenTomatos.com), administration rating sites (Airbnb.com, Booking.com), and food proposal sites (Yelp.com, OpenTable.com), and so forth.

Rating prediction refers to predicting star or numbered rating from the analysis of text review given by online shoppers. Day by day, online shopping is becoming an integral part of our life. As the days go by, people are getting accustomed to online shopping. The most important thing people rely on when buying something online is product reviews and product ratings. Business reviews are very important and help companies to assess quality and excellence in a number of directions that contribute to predicting the

value of a business on the market. A producer can also understand the acceptability of his product to the customer through these reviews and ratings. So the rating andreview are significant to both the shoppers and producers. However, sometimes shoppers give text reviews afteronline shopping but forget or apathy towards giving number ratings. Here we machine learning algorithm to predict rating based on review text.

MOTIVATION

Big data analytics plays a major role in various industries using computingapplications such as Ecommerce and real-time shopping. Big data are used for promoting products and provide better connectivity between retailers and shoppers. Nowadays, people always use online promotions to know about best shops for buying better products. This shopping experience and opinion about the shopper's shop can be observed by the customer-experience shared across social media platforms. A new customer when searching a shop needs information about manufacturing date (MRD) and manufacturing price (MRP), offers, quality, and suggestions which can only be provided by the previous customer experience. The MRP and MRD are already available in the product cover or label.

These days, individuals are accustomed to evaluating the help they experience and rating how great/terrible it is as per their opinion on it. For example, a visitor in a café could give a survey of the supper he/she requested, concerning the food, how the server/server serves, the air, and whatever else connected with the experience. The supportiveness score of an audit is casted a ballot by watchers who think the survey is useful. As the quantity of surveys continues developing quick, assessing the support scores by watchers becomes incomprehensible; in this manner, the undertaking of assessing the accommodation score of a particular thing or subject is vital for

the specialist co-op and the clients. Positioning the surveys in view of their supportiveness will incredibly further develop the client experience, at the end of the day, it will be a lot simpler for clients to get the most valuable data about the assistance or the item they are searching for.

Web based shopping has been expanding throughout recent many years. The solid requirement for online items likewise makes a staggeringly huge number of item surveys. Shoppers might depict explicit elements of the item they buy and express their feelings, suppositions, and contemplations in the surveys. Truth be told, item surveys enormously affect item deals, as buyers will generally peruse highest level audits prior to settling on their last buy choices. This happens on the grounds that internet based shoppers regularly observe they miss the mark on information and time to settle on the most ideal choice out of millions of existing and contending items on different sites. For internet shopping organizations, item surveys give potential purchasers important assets of data on the item to settle on more normal and productive buy choices, which will help online organizations consequently.

EXISTING SYSTEM

Various algorithms had implemented for review rating prediction in which they provide the results where the accuracy of the model changes from one another.Sometimes the dataset is imbalanced where we need to balance the dataset to achieve higher accuracy.

The existing system shows various related works which performs on business review through opinion mining and sentiment analysis.Theyachievedarobustresultusing both binary and multiclass from their dataset. The authors used two feature extraction techniques named word2vec and GlobalVector.Latertheyusedthisfeatureextra ction result with the Multinomial Naive

Bayes algorithm, Deep learning algorithm and Convolution Long Short Term Memory algorithm. They got 84% accuracy on CLSTM multiclassusingword2vecandwithGlove.InS hahproposed an abstractlevel sentimental analysis of user review. The n-gram classification and POS tagging with MaxEnt, Naïve Bayes, SVM and Random Forest classifier wasusedintheirwork.TheyfoundMaxEntand NaïveBayes showed better performance than the other two algorithms. Their algorithm achieved accuracy. Hajiproposed an approach of the combination of lexicon and machine learning for rating prediction based on a textual review of a restaurant dataset. Their system improves the Naive Bayes classifier accuracy of by 5% to10%.

The intial process of dataset contains various columns on different electronic products, movies,etc.If in case there is an imbalance in the data set then we need to balance the dataset by using the technique called oversampling.Oversampling copies the data until all the distribution are not equal to the highest one.And then the dataset undergoes various preprocessing techniques like Puncuation removal, Lower case conversion,Tokenization,Stop word removal,Stemming etc,. For the better accuracy purpose in order to fit for the algorithms applied.

PROPOSED SYSTEM

We proposed the supervised machine learning algorithm Random Forest Classifier respectively for predicting rating based on text review of products. The reason for the proposed machine learning algorithm is to achieve higher accuracy than other models.

In the proposed system along with the text reviews we can also use the image which consists of mostly repeated words for the rating prediction we can obtain the overall
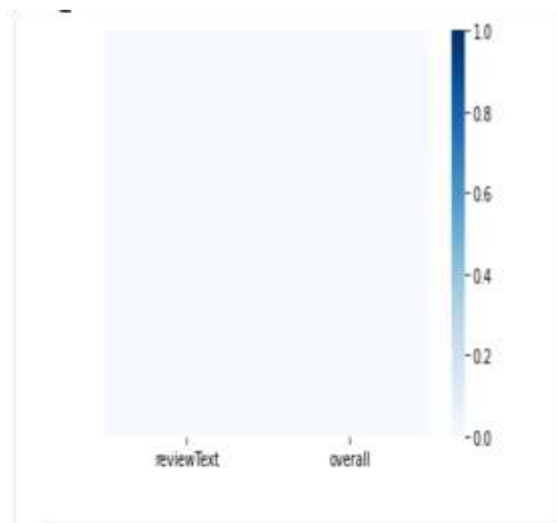
rating based on either the image or the text review. Here in the dataset we consider the attributes like the reviewrID, asin, reviewer name, helpful,review text,overall rating, summary,unix review time ,review time.

After the preprocessing the very next step is to extract the features which is also known as feature extraction from our text review by using a technique called TF-IDF Vectorizer (term frequency-Inverse document frequency), a method to convert aplain text into a vector with a n-gram value.Here the value of TF describes the number of words that appears in a document which divides by the total number of words in the document and then the IDF is used to downscale the words that appears more number of times in a document.At first the technique tokenizes the document to learn the vocabulary and then it inverses the document frequency weightings and allows to encode new documents.

Words repeated (major)
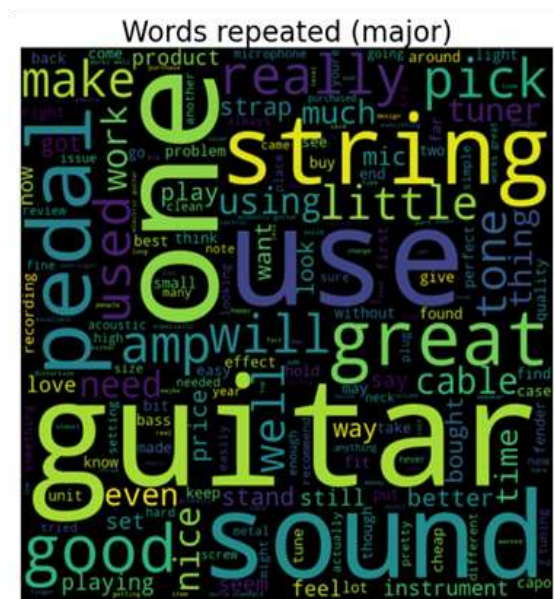
Random forest classifier is the algorithm for supervised learning which can be used for classification as well as regression.Trees constitute a forest.The more trees it has,it is said,and the more resilient a forest is.On randomly selected data samples,Random Forests generates decision trees,get prediction from all the trees and chooses the

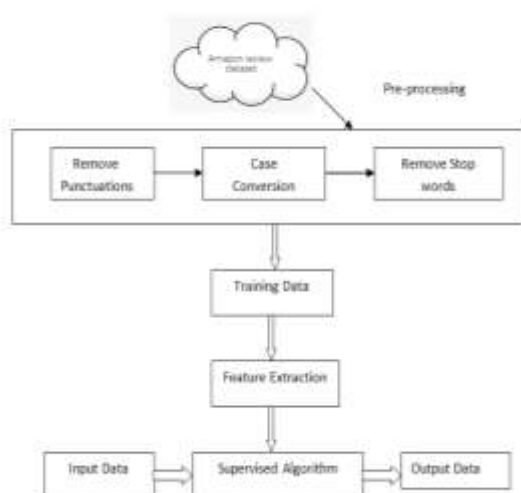best solution.It also gives a good indicator of the value of the function.

SYSTEM ARCHITECTURE

The system architecture describes about the structure of the entire system in which each step of the project is clearly explained where we get a detailed and a complete view of the process flow of our project.

The dataset undergoes various preprocessing techniques as shown in the figure. Later for the training data the dataset will be given and then the feature extraction process takes place. From there the input data is given to the supervised algorithm and we get the required output.

The overall body diagram shows the cloud platform in which we get the data by extracting and then we make a dataset which later performs preprocessing and then followed by the dataset balance technique and then the splitting of data as the train and test data where the train data is of 70% and the test data is of 30% and then the feature extraction followed by the vector representation. After applying the machine learning model the train model is set for evaluation based on which values of accuracy, precision, recall and F1-score is generated.

CONCLUSION

In this paper, we have proposed a framework to predict the best product in the e-commerce website by taking all the important aspects and opinions given by various customers.We utilize the Pros and Cons opinions for improve the feature identification and opinion classification on text reviews. We then developed an aggregate ranking algorithm to summarize the importance of various features of a product from numerous users' reviews. The algorithm simultaneously inspects the aspect frequency and the influence of customer opinions are given to each feature over collected opinions.

Thispaperusedpopularandwidelyusedsupervised machine learning algorithm, Random Forest classifier, to solve the rating prediction problem. The problem went through several stages, including pre-processing, feature extraction and algorithmtrainingtosolveit.Bythisalgorithm, theRandomForestachievedanaccuracy.So theRandomForestachievedbetterresults. In future we will enhance this work to implement in comparing the product in different websites in order to get the best

product with good quality, cost and more to satisfy the user requirements.Future study is to predict rating based on synonym and combine it with proposedalgorithm.This prediction data will help eventual buyers to have a better knowledge of the product.

REFERENCES

[1]Kumar,Y.Mahendra,andRoy, "Fusion of EEG response and sentimentanalysisofproductsreviewtopredictc ustomersatisfaction,"Information Fusion, vol. 52, December 2019, pp.41-52.

[2]Z.Cheng,,D.Ying,Z.Lei,andK.Mohan,"As pectawarelatentfactor algorithm:Ratingpredictionwithratingsandrev iews,"inProceedings oftheIntern.WorldWideWebConf.,April23-27,2018,pp.639-648.

[3]A. Rafay, M. Suleman, and A. Alim, "Robust review rating prediction algorithmbasedonmachineanddeeplearning:Ye lpdataset,"inIntern. Conf.onEmergingTrendsinSmartTechnologies ,March26-27,2020, pp.8138-8143.

[4] Lamb AF, Varadarajan M, Tran R,Vandier N, Doshi BL, Bear C (2012) Thevertica analytic database: Cstore 7 years later.

[5] Betru,B. T., Onana, C. A., &Batchakui,B.(2017). Deep learning method.

[6] Data Preparation,"2019. [Online].Accessed on: June 25, 2020.Available:https://towardsdatascience.c om/feature-engineering-formachinelearning-3a5e293a5114.

[7] TF-IDFVectorizer,"Accessedon:June30,2020.htt ps://towardsdatascience.com/natural-language processingfeatureengineering-using-tf-idf-e8b9d00e7e76.