TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 5703-5708

Machine Learning Techniques for Hotel Reviews Classification

B.Aparna, A.Nagaratnam

Article Info
Volume 83
Page Number: 5703 - 5708
Publication Issue:
May - June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *17 May 2020*

Abstract

In this advanced age where the internet is developing quickly, the presence of the internet can make it simpler for the traveler to discover any data. In the hotel tourism sector, the use of the internet is very helpful for the promotion of their business. There is a vast number of hotel reviews available on the internet. Based on these reviews, hotel owners take advantage of improving facilities. With the accessibility of a large number of reviews, it is difficult to see everyone's review to identify whether they contain positive or negative reviews. In this paper, we proposed machine learning techniques for hotel review classification. The dataset is taken from Kaggle, which contains 10000 reviews. We applied ML natural language processing techniques and classification algorithms on the dataset and achieved an accuracy of 94%. Python language was used for implementation of all experiments.

Keywords: *Hotel Reviews, Machine Learning, Python.*

I. INTRODUCTION

At the point when the travelers need to pick an agreeable lodging for their excursion, they will search for surveys from different explorers. Online visitor audits are basic to what's to come of their property. Reviews accessible on the internet are increasingly significant, genuine and nitty-gritty than the audits found in inn leaflets. In any case, with a large number of reviews, hotel owners can't comprehend and abridge all the available reviews whether they contain positive sentiments or negative sentiments. Some survey sites just give appraisals that are viewed as not objective and thusly can't be utilized as correlations between reviews. The rating framework valuation is unique concerning the composed surveys so it can't be utilized for lodging examination because the data gave isn't clear.The hotel customers can express their emotions and reviews on online-surveys. By playing out the opinion mining and sentiment analysis on these subtleties we can foresee the rating of that hotel. A good recommender framework is required for producing the ratings clearly and exactly. For a lodging business, reviews about different viewpoints like Maintenance, Food, Accommodation, Room neatness, Response from the staff are major items for the recommender framework. The Customer's feelings with respect to a lodging depend upon the facilities he/she got from that hotel, much the same as tidiness, area of the lodging, administrations gave by the inn like free Wi-Fi, multilingual staff, bar/relax, keeping an eye on, and wheel seat and so on. The assessments can be communicated as fantastic, great, normal, poor, terrible, etc. Generally, customers need to express their feelings with rating values. The hotel surveys are given only by clients who have reserved a spot at a specific lodging. A recommendation system can give recommendations to a solitary thing or a succession of things. A solitary thing (single thing) is largely utilized at web-based business sites to discover the deal and result of one single thing while the succession recommender framework for a succession of things is utilized to anticipate the matter of any association; for example, lodging business, tidiness, administrations, and these qualities are utilized. The recommender framework is considered as a decent one that can distinguish the fitting clients, their inspiration, desires and the objectives. Machine learning doesn't allude to only certain something; it's an umbrella term that can be applied to a wide range of ideas and strategies. Understanding AI implies being acquainted with various types of model investigation, factors, and calculations.Machine Learning models work with mathematical operations. So, the datasets used for

May – June 2020
ISSN: 0193-4120 Page No. 5703-5708

machine learning must be numeric. Even though, the dataset contains categorical features they are converted to numerical values before applying ML techniques. But hotel reviews purely contain textual data, not numeric data. For handling textual data, Machine learning provides Natural Language Processing (NLP) techniques. After applying various NLP techniques, finally, textual data is converted to numeric.Then, data is ready for applying ML techniques.

II. LITERATURE SURVEY

Various researchers applied machine learning techniques for the classification of hotel reviews. As input for the machine learning model is in the form of text, NLP techniques must be applied before proceeding with ML algorithms. Han-Xiao Shi [1] et.al proposed a sentiment analysis technique for hotel reviews using supervised machine learning methods. They applied TFIDF (term frequency-inverse document frequency) technique and support vector machines. With their model, they achieved a recall of 89% and a f-score of 87%.Eivind Bjorklund [2] et.al explained how the outcome of the sentiment analysis reviews can be visualized with Google Maps, which are very helpful for the new customers to detect good/bad hotels to stay in.

They studiedopinion mining applied on data from travel review sites. Tushar Ghorpade [3] et.al proposed sentiment classification for hotel reviews by Bayesian classifier. They applied feature-based sentiment classification and achieved an accuracy of 96% and recall of 98% with their model.Shelly Gupta [4] et.al proposed a decision tree classification method for hotel reviews. They applied the C4.5 decision tree model and achieved good results. After applying the decision tree, they derived 13 rules and made some conclusions.

Dietmar [5] et.al proposed a lexicon-based method to classify customer reviews. They applied a corpus consists of reviews of TripAdvisor, which is a major web 2.0platform. In this method, the rating contains five levels "terrible" to"excellent". A separate overall rating summarizes customer feelings.StanimiraYordanova [6] et.al proposed classification model for hotel reviews in social media with decision tree approach. They applied three classification techniques on three different datasets with two schemes and achieved good results. George Markopoulos [7] et.al proposed a supervised machine learning model for text mining. They applied Support Vector Machines on hotel reviews written in Modern Greek. By applying a unigram model, they compared two separate methodologies and achieved the emerging results.HarshitSanwal [9] et.al proposed Support Vector Classifier with PSO (Particle Swarm Optimization) technique for opinion mining in Hotel reviews.

III. RESEARCH METHODOLOGY

Figure 1 shows the framework of proposed model for classification of hotel reviews. First, we collected hotel reviews dataset from Kaggle.com [8].The dataset name is "Datafiniti_Hotel_Reviews_Jun19". It contains 26 features. Although it has 26 features namely address, categories, city, country, latitude, longitude, name, postal Code, province, reviews. date, reviews. Date Added, reviews. do Recommend, reviews.id, reviews. rating, reviews. Text, reviews. title, reviews. user City, reviews. Username, reviews. user Province. But we considered only two important features namely rating and review. Ratings are given between 1 to 5 levels. But we converted this level into positive and negative reviews by considering rating levels 1, 2 as negative review and rating levels 3,4,5 as positive reviews. So finally, the dataset contains two features, one is review and another is 0 (negative review) or 1(positive review).After that, we applied ML NLP techniques to convert data from textual notation to numeric notation. Then, the given problem becomes classification task. So, we applied various classification algorithms.

Algorithm for proposed work:

Step 1: Collection of dataset

Step 2: Data Preprocessing (Feature Selection)

Step 3:Apply stemming using python porterstemmer and remove stop words

Step 4: Apply count vectorizer/TFIDF vectorizer

May – June 2020
ISSN: 0193-4120 Page No. 5703-5708

methods

Step 5: Divide the data into training and testing sets

Step 6: Apply Machine Learning Algorithms and find best

Algorithm for the given dataset.

Proposed method:

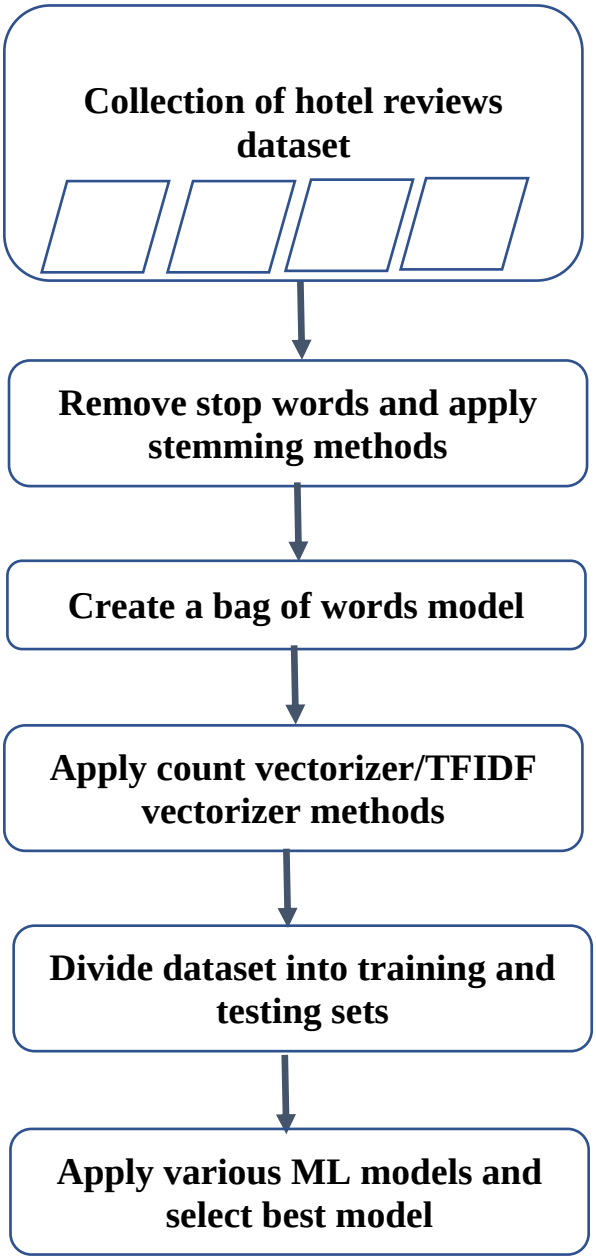


Figure 1: Proposed Model

Classification algorithms:

In machine learning, classification task is a supervised learning method. Supervised learning methods are very useful for solving many real world problems [10]. There is large number of algorithms available for classification problems. The hotel reviews dataset contains a dependent feature called as "liked/not liked". So, this is a two-class problem. The algorithm needs to divide the data sample into either liked category or 'not liked' category. Some of the classification algorithms are SVM, Adaboost, Logistic Regression, Naive Bayes model,K-NN model, random forest, decision tree classifier etc..

Classification accuracy can be measured by using following notations:

Classifier accuracy= (True Positives+True Negatives)/Total number of samples

True positives (TP):The actual data label is True and the algorithm predicted as TRUE

False Positives (FP): The actual data label is FALSE and the algorithm predicted as TRUE

False Negatives (FN):The actual data label is TRUEand the algorithm predicted as FALSE

True Negatives (TN):The actual data label is FALSE and the algorithm predicted as FALSE

But accuracy is not only the measure used for comparing ML techniques. Sometimes accuracy misleads us. So, there is a need to use other measures like precision, recall,f1-score. Precision is nothing but a Positive Predictive Value. Recallis True Positive rate and F-score is the harmonic mean of precision and recall.

Precision = Number of TPs / (Number of TPs + Number of FPs)

Recall = Number of TPs / (Number of TPs + Number of FNs)

IV. EXPERIMENTS

Machine Learning algorithms can be implemented in different platforms like R programming, python programming, weka tool, etc. We selected python language for implementing ML models because Python provides rich libraries for implementing all machine learning models. For applying Natural Language Processing techniques on text data, we installed NLTK (Natural Language Processing Toolkit). NLTK provides different options for text mining. We used stopwords class for deleting

TEST Engineering & Management

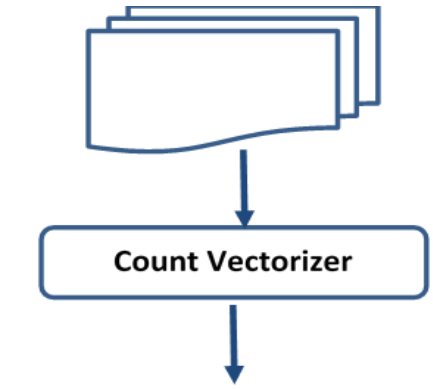
May – June 2020
ISSN: 0193-4120 Page No. 5703-5708

stopwords in the text review. For example, consider a review “rooms are nice”, in which “are” is a stop word, because it is non-informative. Similarly, we applied “PorterStemmer” class for stemming process.Hotel reviews are in the form of English text. These text reviews are made up of words that are derived from one another. Stemming reduces inflection in words to their root words. For example, the words "playing", "plays", "played" are formed from a root word "play”. Stemming maps, a group of words to the same stem. After that, we applied two feature extraction techniques namely CountVectorizer, TFIDF for bag of words model. We selected maximum features as 5000.

Next, we applied machine learning techniques by dividing the dataset into training and testing parts. Count vectorizer easily build a vocabulary of words. It can also encode new documents using vocabulary.

Count Vectorizer:

We created an instance of CountVectorizer class. We called fit() method to learn a vocabulary from text reviews. Finally, we called transform() method on text to encode each as a vector. This vector is come back with a length of the whole jargon and it likewise contains a number, which shows the no.of times each word showed up in the reviews.

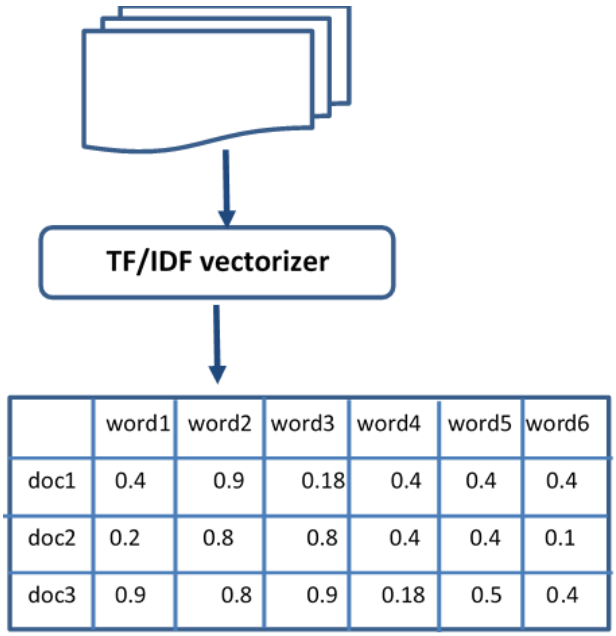


	word1	word2	word3	word4	word5	word6
doc1	1	2	1	0	1	2
doc2	1	2	1	0	1	2
doc3	1	2	1	0	1	2

Figure 2: Count Vectorizer

TF/IDF:

Count vectorizer s based on word counts.One problem with this method is that words like "the" occurs many number of times and their high count is not meaningful in encoded vector.An alternative to word count is word frequency.TFIDF(Term Frequency - Inverse Document frequency")is based on word frequency.Term Frequency is utilized to abridge how a given word appears in a document, whereas Inverse Document Frequency downscales words that seem a lot across documents.



	word1	word2	word3	word4	word5	word6
doc1	0.4	0.9	0.18	0.4	0.4	0.4
doc2	0.2	0.8	0.8	0.4	0.4	0.1
doc3	0.9	0.8	0.9	0.18	0.5	0.4

Figure 3: TF/IDF Vectorizer

IDF(word)=log((#documents)/#(documents containing word))

V.RESULTS

We applied eight classification algorithms Logistic Regression classifier, Support Vector classifier, K-NN, Adaboost classifier, Decision Tree Classifier, Random Forest Classifier, Extra Tree Classifier, Gradient Boosting classifier. All these models applied with two different methods count vectorizer,TF/IDF separately. We observed that, there is no more large difference between count vectorizer and TF/IDF for our dataset. For both techniques, Logistic Regression classification model achieved high accuracy value.All these ML algorithms given good precision, recall and accuracy values. Out of eight algorithms, fivealgorithms (Logistic Regression, Adaboost, DT

May – June 2020
ISSN: 0193-4120 Page No. 5703-5708

classifier, Random Forest model, ETF,GBC) given more than 90% accuracy. The remaining three algorithms also given accuracy greater than 85%.All the models showed a good balance among precision, recall and accuracy values. All the results are tabulated.

Results with count vectorizer:

Algorithm	Precision	Recall	Accuracy
Logistic Regression	93%	93%	93%
K-NN	86%	89%	89.2%
AdaBoost	91%	92%	92%
Decision Tree	88%	88%	87%
Random Forest	91%	91%	91%
Extra Tree Classifier	92%	92%	92%
Gradient Boosting classifier	92%	93%	92%
SVM	90%	90%	89%

Table 1:comparison of ML models with count vectorizer

Accuracy comparison chart with count vectorizer:

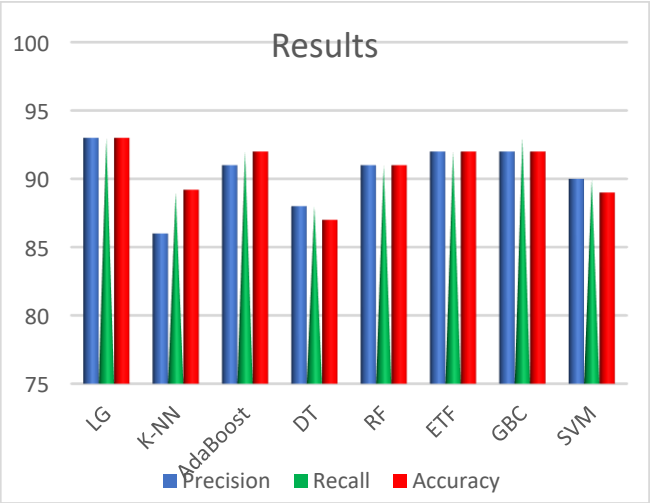


Figure.4: comparison of ML models for count vectorizer

Results with TF/IDF:

Algorithm	Precision	Recall	Accuracy
Logistic Regression	93.5%	94%	94%
K-NN	86%	89%	89.2%
AdaBoost	91%	92%	92%
Decision Tree	88%	88%	87%
Random Forest	91%	91%	91%
Extra Tree Classifier	92%	92%	92%
Gradient Boosting classifier	92%	93%	92%
SVM	90%	90%	89%

Table 2: comparison of ML models with TF/IDF

Accuracy comparision chart with TF/IDF:

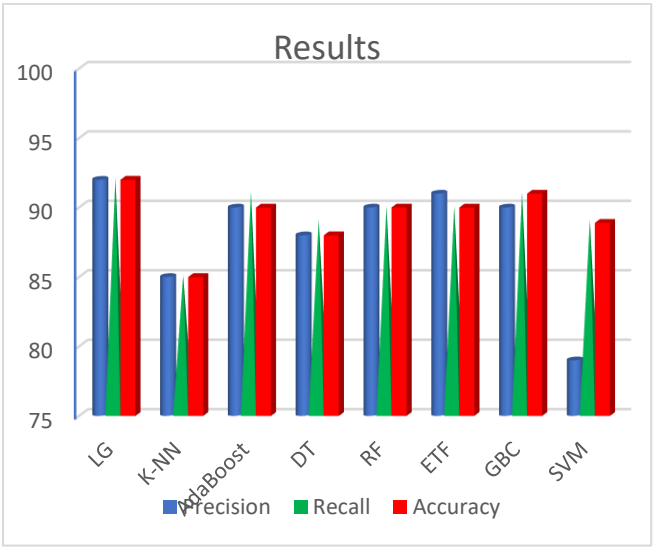


Figure.5: comparison of ML models for TF/IDF

VI.CONCLUSION

In this paper, we proposed a machine learning approach for the classification of hotel reviews. We applied machine learning techniques on a hotel reviews dataset with 10000 samples. As hotel reviews contain textual data, we applied natural language processing techniques on the given dataset to convert textual data into numeric vector form. We applied two techniques count vectorizer, TF/IDF vectorizer methods. We achieved the best results with all the ML models. All the ML models given more than 85% accuracy. We achieved an accuracy of 94% with Logistic Regression and

May – June 2020
ISSN: 0193-4120 Page No. 5703-5708

TF/IDF technique.

REFERENCES

[1] Machine Learning and Cybernetics, Guilin, 10-13 July, 2011W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Han-Xiao Shi, Xiao-Jun Li, "A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning", Proceedings of the 2011 International Conference on Wadsworth, 1993, pp. 123–135,978-1-4577-0308-9/11, 2011 IEEE

[2] EivindBjørkelund, Thomas H. Burnett, KjetilNørvåg, "A Study of Opinion Mining and Visualizationof Hotel Reviews", iiWAS2012, 3-5 December, 2012, Bali, Indonesia.

[3] Tushar Ghorpade, LataRagha, "Featured Based Sentiment classification for Hotel Reviews using NLP and Bayesian Classification," 2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India

[4] Shelly Gupta,Shubhangi Jain,Shivani Gupta, "Opinion Mining forHotel Rating Through Reviews Using DecisionTree Classification Method", International Journal of Advanced Research in Computer Science. Volume 9, No. 2, March-April 2018.

[5] DietmarGrabner, Gunther Fliedl, Markus Zanker, "Classification ofcustomer reviews based on sentiment analysis", M. Fuchs et al. (eds.), Information and Communication Technologies in Tourism 2012
Springer-Verlag/Wien 2012

[6] StanimiraYordanova,DorinaKabakchieva,"*Sentiment Classification of Hotel Reviews in Social Media with Decision Tree Learning",International Journal of Computer Applications (0975 – 8887), Volume 158 – No 5, January 2017.*

[7] George Markopoulos, George Mikros, Anastasia Iliadi, and MichalisLiontos, "Sentiment Classification of Hotel Reviews in Social Media with Decision Tree Learning", Springer International Publishing Switzerland 2015, Springer Proceedings in, Business and Economics, DOI 10.1007/978-3-319-15859-4_31.

[8] Available online:https://www.kaggle.com/datafiniti/hotel-reviews

[9] George Markopoulos, George Mikros, Anastasia Iliadi, and MichalisLiontos, "Design Approach for Opinion Mining in Hotel Review using SVM With PSO (Particle Swarm Optimization)", International Journal of Engineering Research & Technology (IJERT),Vol-8 Issue 09, Sep-2019.

[10] F.Y.Osisanwo,J,Akinsol,O.Awodele,Hinmikaiye, Olakanmi,Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison",International Journal of Computer Trends and Technology (IJCTT),Vol 48 No 3 June-2017.

AUTHORS PROFILE

Mrs. B.Aparna, Assistant Professor fromBABA Institute of Technology & Sciences, Visakhapatnam, completedBTech in Computer Science Engineering and MTech in Computer Science Engineering. Her areas of interest are Machine Learning, Cyber Security and Deep Leaning.

Mrs. A.Nagaratnam, Assistant Professor from BABA Institute of Technology & Sciences, Visakhapatnam, completed Bachelor of Engineering in CSE and M.Tech in Computer Scinece& Engineering. Her areas of interest are Data Mining, Machine Learning, Deep Learning.