JES
Journal of Engineering Sciences
Vol 13, Issue 06, JUNE/2022
ISSN NO:0377-9254

Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques

E. Sai Kumar #1, K. Navya #2, K. Archana #3, K. Vandana#4, V. Anusha #5
#1 Assistant Professor, #2,3,4,5 B.Tech., Scholors
Dept of Information Technology, QIS Institute of Technology, Ongole, Prakasam(Dt)

Abstract:
Nowadays, social media is becoming more popular. On tourism websites, millions of users review and rate tourist attractions every day. These may be subjected to sentiment analysis. Reviews that can assist you in determining the popularity of a tourist attraction. Tourists may simply make decisions based on the results of sentiment analysis. To be visited as part of a tour. In this study, we look at how people feel has been implemented with the help of a machine learning algorithm. The data set was compiled from a variety of tourism reviews websites. We have conducted a comparative analysis of CountVectorization, for example, is a feature extraction approach. TFIDFVectorization. In addition to classification algorithms, Support Vector Machine (SVM), Naive Bayes (NB), and Forest of Chance (RF). Algorithm performance has been improved compared utilising multiple metrics such as recall, accuracy, and precision precision and f1-score are two terms that come to mind while discussing precision and f1-score We discovered via experimentation that TFIDFVectorization has a feature extraction approach called TFIDFVectorization better classification algorithm accuracy when compared to For a particular review dataset, use CountVectorization. In terms of feeling Reviews of tourist attractions are categorised.

1. Introduction

Nowadays, social media is quickly expanding. Hundreds of millions On a daily basis, people publish reviews and rank tourism attractions. Websites dedicated to tourism For the purpose of analysing the sentiment of these reviews, It is possible to do research. Reviews will be properly analysed if they are properly analysed possible to identify a tourist attraction's popularity trend Summarized Tourists will be able to make better decisions based on the findings of sentiment analysis Tour location and itinerary. Two feature extraction methods are presented in this study. CountVectorization and CountVectorization

TFIDFVectorization is a vectorization algorithm developed by TFIDF. There are three classifications as well. Support Vector Machine (SVM), Naive Bayes (NB) For sentiment analysis, SVM and Random Forest (RF) were utilized classification. Performance has been compared conducted for the purpose of combining fea- ture extraction and classification algorithms based on variables like as Execution time, accuracy, recall, precision, and f1-score are all factors to consider.

2. Literature Survey

Various strategies of sentiment analysis are discussed in this study [1] has been investigated and contrasted. Various degrees of Document-level feelings, sentence-level emotions, and aspect-level sentiments which has been refined Sentiment analysis methods The study in this research is based on machine learning and rules as well as lexical. Various approaches to machine learning are used. SVM (Support Vector Machine) and NB (Naive Bayes) are two approaches. Maximum Entropy, K-NN, and Weighted K-NN are all examples of Bayesian models. Sentiment Analysis in Multiple Languages is also feature-rich. The process of sentiment analysis has been well explained. Various sentiment analysis methodologies have been contrasted associated benefits and drawbacks are detailed in detail. Various comparison criteria, such as accuracy, efficiency, and performance It was discovered that The strategy based on machine learning yields the greatest results as outlined in On the [2] paper, a sentiment analysis of Twitter was carried out movie critiques They've utilised a number of monitored machines Support vector machines and naive Bayes are examples of learning algorithms. Using a variety of feature extraction techniques, Bayes and maximum entropy were calculated approaches such as unigram, bigram, and hybrid (unigram + bigram) bigram. They came to the conclusion that SVM is effective based on their research use the hybrid feature The extractor outperforms the competition techniques.

As described in [3], a paper survey on the fundamentals of sentiment The analysis has been completed, and it has been used in a variety of situations Various approaches for this topic have also been developed. The concept of sentiment analysis has been investigated. There are two of them lexicon-based and machine-learning methods to sentiment analysis based on learning There are two types of lexicon-based systems. There are two sorts of dictionaries: dictionary-based and corpus-based. Based on a corpus There are two ways to look at it: statistical and semantic. Statistical The recurrence of a phrase is determined by the approach, but the semantic approach is determined by the approach based on a word's resemblance The term "machine learning" refers to the process of two sorts of classifications supervised and unsupervised environments They According to the statement, supervised algorithms are made up of a variety of components Support vector machines and neural networks are examples of algorithms. Maximum entropy, Bayesian network, and naive Bayes As stated in article [4], the author conducted a thorough investigation relating to text mining The author mentions a number of uses and techniques for text mining There are numerous processes involved in this process as well preprocessing of text In his paper, he describes the vector space model detailed. Various classification methods, such as naive bayes, are available. Nearest neighbour has a support vector machine, a decision tree, and a support vector machine. Detailed explanations have been provided. Various clustering techniques are also available. Topic modelling, like kmeans and hierarchical clustering, has been used explained. The importance of text mining in the extraction of information has been discussed. Text mining is also used in biomedicine. The topic of healthcare has been discussed.

Text feature extraction algorithms were employed in [5] for dividing short sentences and phrases into categories Author has made use of Inverse Document for Term Frequency The frequency (TF- IDF) method and its two variations using various strategies for dimensionality reduction Latent Linear Discriminant Analysis (LDA) and Semantic Analysis (LSA) (LDA). It was discovered that TF-IDF outperformed other models procedures that were employed Author [6] classified the news into five categories groups. They employed TF-IDF to extract features as well as the Support Vector Machine (SVM) as an algorithm Algorithm for classification For BBC, they obtained 97.84 percent accuracy 94.93 percent accuracy for 20 Newsgroups in a news dataset.

As stated in [7], the author has undertaken sentiment analysis a study of the movie review dataset He remarked that in the past the emphasis of study was on SVM, Naive Bayes, and Classification techniques based on maximum entropy. In this article, we'll look at He used sentiment classification to classify the reviews. The most accurate classifier was the random forest classifier, which had a 90% accuracy rate. The author of this study [8] used sentiment analysis employing several features such as unigram, over a movie review dataset top 2633, bigram, unigrams+bigrams, POS, adjectives numerous classification techniques, as well as unigrams and unigrams+position Maximum entropy, naive bayes, and SVM were all used a comparison of whose accuracy has been made It has been discovered via study It was discovered that naive bayes has the lowest accuracy, while SVM has the most provides the greatest level of accuracy.

As shown in [9], there are a variety of opinion mining strategies such as trend-driven, aspect-driven, and sentence-driven. Author has proposed aspect-based opinion mining, as well as a tourist destination Visitor reviews have been mined for information on connected topics The reviews were then divided into two categories: good and negative feelings in relation to several issues They've gone with POS. For aspect extraction and opinion trending, use tagger and WordNet. They've done a tweet extraction based on it.Positive, negative, and neutral classifications are used. Machine learning may help enhance system performance approach to learninghas been developed Text mining is also used in biomedicine.The topic of healthcare has been discussed.

Text feature extraction algorithms were employed in [5] for dividing short sentences and phrases into categoriesAuthor has made use of Inverse Document for Term FrequencyThe frequency (TF- IDF) method and its two variationsusing various strategies for dimensionality reduction LatentLinear Discriminant Analysis (LDA) and Semantic Analysis (LSA)(LDA). It was discovered that TF-IDF outperformed other models procedures that were employed

Author [6] classified the news into five categories groups. They employed TF-IDF to extract features as well as the Support Vector Machine (SVM) as an algorithmAlgorithm for

classification For BBC, they obtained 97.84 percent accuracy 94.93 percent accuracy for 20 Newsgroups in a news dataset.

As stated in [7], the author has undertaken sentiment analysis a study of the movie review dataset He remarked that in the pastemphasis of study was on SVM, Naive Bayes, andClassification techniques based on maximum entropy. In this article, we'll look atHe used sentiment classification to classify the reviews.The most accurate classifier was the random forest classifier, which had a 90% accuracy rate.The author of this study [8] used sentiment analysis employing several features such as unigram, over a movie review datasettop 2633, bigram, unigrams+bigrams, POS, adjectivesnumerous classification techniques, as well as unigrams and unigrams+positionMaximum entropy, naive bayes, and SVM were all utilized a comparison of whose accuracy has been made It has been discovered via studyIt was discovered that naive bayes has the lowest accuracy, while SVM has the most provides the greatest level of accuracyAs shown in [9], there are a variety of opinion mining strategies such as trend-driven, aspect-driven, and sentence-driven.

3. Existing System

Data sparsity is main problem in tourism domain We have tried to collect large amount of data from heterogeneous tourism websites. From literature survey we have infer that machine learning can improve classification accuracy over lexicon based approach So sentiment analysis using machine learning techniques has been adopted for research. Result of reviews sentiment classification using different machine learning techniques has been compared and analysed.

Drawbacks:

Classification is slower and costlier with respect to time and memory

4. Proposed System

In this paper author using machine learning algorithms such as SVM, Naïve Bayes and Random Forest to predict sentiments from tourist reviews dataset and then evaluating performance of CountVectorizer and TFIDFVectorizer features extraction algorithms. In this paper author is extracting features from reviews by using both CountVectorizer and TFIDFVectorizer and then applying this features on machine learning algorithms and then calculating accuracy, precision, recall and F!SCORE between both feature extraction algorithms.

A.Dataset

The research uses review data from various tourism websites.[18][19] Data has been collected in CSV format which consists of review text and associated rating. From rating we calculated sentiment whether positive, negative or neutral. If rating is greater than 3 then it is considered as

positive if less than 3 then it is considered as negative and if equal to 3 then it is considered as neutral.

B. Data Preprocessing

Social media data is highly raw, so there was a need of data cleansing. Data preprocessing involves various stepssuch as tokenization, Stop word removal, stemming and lemmatization.

 Tokenization: splitting sentence into words has been performed. Each word is a token, so process called as tokenization.

 Stop word Removal: In documents, words which occur very frequently such as a, the, this, you, in, is, was.. etc are stop words which should not get passed to text mining algorithms. In research study, We have used customized stop word list which contains words which was irrelevant occurring with very high frequency in corpus. This has reduced feature vector size as well as improved performance of system.

 Lemmatization: tokens in past or future tenses get converted into present tense. also tokens in third person form gets converted into first person form. For ex: token mountains converted to mountain.

 Stemming: Finding root word of token is called stemming. For ex: token trekking converted to trekk. along with above steps Short word removal, Punctuation mark removal, Numeric and Special character removal, lower case conversion has been performed for better performance of machine learning algorithms.

C. Feature Extraction

There are various Feature extraction algorithms in natural language processing. We have used CountVectorization and TFIDFVectorization algorithm for feature extraction from reviews data. A CountVectorization is identical to Bag of word (BoW) approach. It is an indication of text occurrence along with its frequency of occurrence within a particular document. Whereas TFIDFVecrorization is an extension of CountVectorization where inverse document frequency also taken into consideration in parallel with term frequency.

D. Training Model

For training model 80% data has been used. The research uses Linear Support vector machine, Multinomial Naïve Bayes, Random Forest as classification algorithms for training reviews dataset.

E. Testing Model

From total reviews data 20% data has been used for testing. Testing has been performed on new unseen reviewsto predict polarity of sentiment. Trained model will classify review's sentiment into 3 classes positive, negative, neutral.

F. Visualization

Predicted sentiment has been visualized through matplotlib by plotting pie chart for percentage distribution ofpositivity, negativity and neutrality of reviews of each tourist place.

G. Performance Evaluation

Performance evaluation is one of the crucial step in machine learning. Performance evaluation has been performed using parameters like accuracy score, precision, recall, f1-score and execution time measurement.

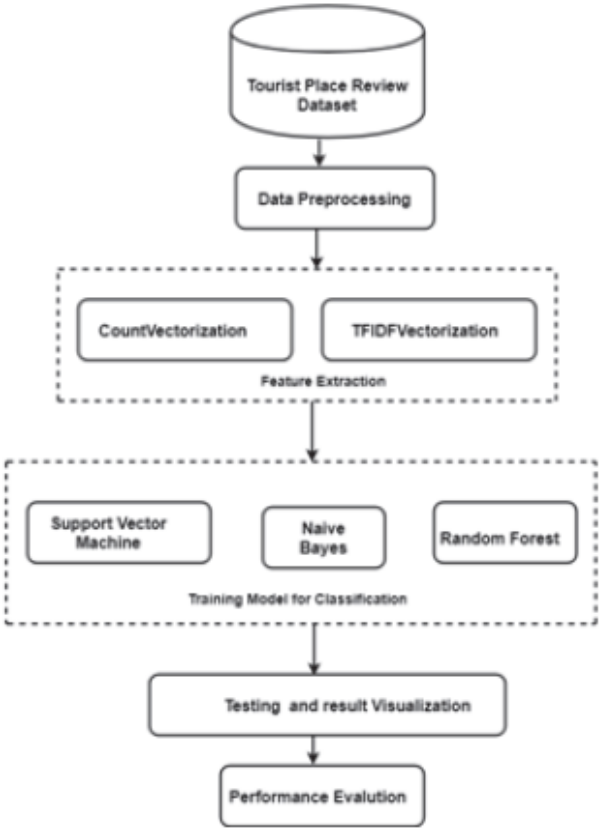

Fig. 1. System Architecture

5. Conclusion

From research study, we can infer that TFIDFVectorization has outperformed over CountVectorization feature extraction algorithm by increasing accuracy of classification. But feature extraction using TFIDFVectorization requires more execution time than CountVectorization algorithm. In research, classification algorithms Support Vector Machine(SVM), Naïve Bayes(NB), Random Forest(RF) has been used. It has found that TFIDFVectorization+RF outperformed over other algorithms used on bases of several evaluation parameters like accuracy, precision, recall and f1-score.

6. Future Scope

The research study for Tourist Place review classification using machine learning algorithm has future scope of handling multilingual review classification. Also we will try to use different feature selection method like Recursivefeature elimination with cross-validation to improve accuracy of classification. In future work we will try to use deep learning based techniques for feature extraction and classification for better performance.

References

[1] M.D.Devika, C.Sunitha, Amal Ganesh "Sentiment Analysis: A Comparative Study on Different Approaches" ScienceDirect Fourth Interna- tional Conference on Recent Trends in Computer Science Engineering https://doi.org/10.1016/j.procs.2016.05.124

[2] Rohit Joshi , Rajkumar Tekchandani " Comparative analysis of Twitter data using supervised classifiers" 2016 International Conference on Inventive Computation Technologies (ICICT) DOI: 10.1109/INVENTIVE.2016.7830089

[3] Harpreet Kaur, Veenu Mangat, Nidhi "A Survey of Sentiment Analysis techniques " 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) DOI: 10.1109/ISMAC. 2017.8058315

[4] Mehdi Allahyari, SeyedaminPouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, KrysKochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", arXiv:1707.02919 [cs.CL], July 2017 [5] Robert Dzisevicˇ , DmitrijSˇesˇok "Text Classification using Different Feature Extraction Approaches Text Classification using Different Feature Extraction Approaches" 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)

[6] Seyyed Mohammad Hossein Dadgar, Mohammad Shirzad Araghi, Morteza Mastery Farahani "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification" 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th 18thMarch 2016, Coimbatore, TN, India. \

[7] Rasika Wankhede, Prof. A.N.Thakare "Design Approach for Accuracy in Movies Reviews Using Sentiment Analysis". International Conference on Electronics, Communication and Aerospace Technology ICECA 2017 [8] Bo Pang and Lillian Lee, ShivakumarVaithyanathan "Sentiment Classifi- cation using Machine Learning Techniques " Proceedings of the Confer- ence on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86. Association for Computational Lin- guistics.

[9] Muhammad Afzaal, Muhammad Usman "Novel Framework for Aspect- based Opinion Classification for Tourist Places" The Tenth International Conference on Digital Information Management (ICDIM 2015)

[10] Upma kumari, Dr. Arvind K Sharma, Dinesh Soni "Sentiment analysis of smart phone product reviews using SVM classification techniques" 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)

[11] Xing Fang and Justin Zhan "Sentiment analysis using product review data " Springer an Journal of Big Data (2015) 2:5 DOI 10.1186/s40537- 015-0015-2

[12] C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery, vol. 2, pp. 121–167, 1998.

[13] Leo Breiman ”RANDOM FORESTS” Statistics Department University of California Berkeley, CA 94720

[14] C. Sheppard, Tree-based Machine Learning Algorithms: Decision Trees, Random Forests, and Boosting. CreateSpace Independent Publishing Platform, 2017.

[15] Kamal Sarkar ”Using Character N-gram Features and Multinomial Naive Bayes for Sentiment Polarity Detection in Bengali Tweets” 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)

[16] Text Classification and Naive Bayes https://web.stanford.edu/ jurafsky/ slp3/slides/7 NB.pdf

[17] Dixa Saxena, S. K. Saritha, PhD , K. N. S. S. V. Prasad ”Survey Paper on Feature Extraction Methods in Text Categorization” International Journal of Computer Applications (0975 – 8887) Volume 166 – No.11, May 2017

[18] https://www.tripadvisor.in/

[19] https://www.mouthshut.com/