IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105304 |

Tourist Place Reviews Sentiment Classification Using TF-IDF and Count Vectorization

Dr.Vilas Joshi[1], Yaswanth Battina[2], Nayan Narke[3], Priyesh Thakre[4], Arya Dixit[5]

Assistant Professor, Dept. of Computer Engineering, ISB&M College of Engineering, Pune, India[1]

UG Students, Dept. of Computer Engineering, ISB&M College of Engineering, Pune, India[2,3,4,5]

ABSTRACT: Social media is growing trend now a days. Every day millions of user review and rate tourist places on tourism websites. Sentiment analysis can be performed over these reviews which will be helpful to find tourist place popularity. Based on sentiment analysis result, tourist can easily decide tour destination to be visited. In this paper sentiment analysis has been implemented using machine learning approach. The Dataset has been collected from various tourism review websites. Here we have performed comparative study of feature extraction algorithms i.e. CountVectorization,TFIDFVectorization. Along with classification algorithms Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF). Performance of algorithms has been compared using various parameters like accuracy, recall, precision and f1- score. From experiment we found thatTFIDFVectorization feature extraction algorithm has improved accuracy of classification algorithm as compare to CountVectorization for given review dataset. In sentiment classification of tourist place reviews TFIDFVectorization+RF has given highest accuracy 86% for a research dataset used.

KEYWORDS: Count Vectorization, TFIDF Vectorization, Naive Bayes, Support Vector Machine, Random Forest

I. INTRODUCTION

Social media is rapidly growing now a days. Millions of users post reviews and rate tourist place on a daily basis over tourism websites. For analyzing this reviews sentiment analysis can be performed. Proper analysis of reviews will able to find a trend of tourist place popularity. Summarized results from sentiment analysis will help tourist to decide the tour destination and tour planning.

In this research paper two feature extraction algorithms have been used i.e. CountVectorization and TFIDFVectorization algorithm. Also three classification algorithms Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) has been used for sentiment classification. Comparison of performance has been performed for combination of fea- ture extraction and classification algorithms on the basis of parameters like execution time, accuracy, recall, precision and f1-score.

The content of this paper is structured as follows. Literature survey on sentiment analysis are reviewed in Section II . Section III defines Basic concept of Machine Learning. Section IV describes our Methodology of sentiment analysis for tourist place review classification its visualization and performance evaluation. Section V presents the experimental implementation using machine learning algorithms for tourist place popularity distribution calculation. Section VI contains the results of experiment executed. Section VII presents the com- parative analysis of sentiment analysis using machine learning algorithms used in research study. Section VIII concludes this research paper. Section IX describes future scope of research paper.

II. RELATED WORK

In this paper [1] various techniques of sentiment analysis has been studied and compared. Different levels of sentiments are document level, sentence level, aspect level which has been elaborated Approaches used for sentiment analysis in this paper are machine learning based, Rule based and lexical based. Inside machine learning approach various techniques are SVM (Support Vector Machine), NB (Naive Bayes), Maximum Entropy, K-NN and Weighted K-NN, Multilingual Sentiment Analysis also feature driven sentiment analysis has been described in detailed. Various approaches of sentiment analysis has been compared its corresponding advantages and disadvantages are described in

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105304 |

detail. From Various parameters of comparison like performance, efficiency, and accuracy it has been found that machine learning approach gives best result. As described in
[2] paper twitter sentiment analysis has been performed on movie reviews. They have used various supervised machine learning algorithms such as support vector machine, naive bayes and maximum entropy using various feature extraction techniques like unigram, bigram and hybrid i.e. unigram + bigram. From research study they have concluded that SVM using hybrid feature extractor outperforms over other techniques.
As elaborated in [3] paper survey on basics of sentiment analysis has been performed its application in various domain has been elaborated also various techniques used for senti ment analysis has been studied. There are two approaches of sentiment analysis lexicon based and machinelearning based. Lexicon based is further categorized into 2 types dictionary based and corpus based. Corpus based consists of 2 ways statistical and semantic. Statistical approach find occurrence of term whereas semantic approach based on similarity of words. Machine learning is categorized into 2 types supervised and unsupervised They stated that supervised algorithm consists of various algorithms like support vector machine, neural net- work, bayesian network, maximum entropy and naive bayes.

As given in paper [4], author performed detailed sur- vey on text mining. Author states various applications and approaches of text mining. Also several steps involved in text preprocessing. He has described vector space model in detailed. Various classification algorithms like naive bayes, support vector machine, decision tree, nearest neighbor has been elaborated in detail. Also various clustering algorithms like kmeans, hierarchical clustering, topic modeling has been explained. Role of text mining in information extraction has been elaborated. Also use of text mining in biomedicine and healthcare has been explained.

In [5] text feature extraction approaches has been used for classifying short sentences and phrases into classes. Author has used Term Frequency Inverse Document Frequency (TF- IDF) approach and its two modifications using different dimensionality reduction techniques Latent Semantic Analysis (LSA) and Linear Discriminant Analysis (LDA). It found that TF-IDF has outperformed over other techniques used.

Author [6] has performed news classification into 5 groups. They have used TF-IDF as feature extraction algorithm and Support Vector Machine (SVM) as classification algorithm. They got 97.84% accuracy for BBC news dataset and 94.93% accuracy for 20 Newsgroup dataset.

As described in [7], author has performed sentiment analysis over movies review dataset. He stated that previous research focused more on SVM, Naive Bayes, and Maximum Entropy algorithms for classification. In this paper he has performed classification of review's sentiment using random forest classifier which has gave best accuracy 90%.

In this paper [8] author has performed sentiment analysis over movies review dataset using various feature like uni-gram, bigram, unigrams+bigrams, POS, adjectives, top 2633 unigrams, unigrams+position also various classification algo- rithms like maximum entropy, naive bayes and SVM used whose accuracy comparison has been done. From research it is found that naïve Bayes gives worst accuracy where as SVM gives highest accuracy.

As presented in [9], various techniques of opinion mining like trend based, aspect based and sentence based. Author has proposed aspect based opinion mining, tourist place related aspects has been extracted from tourist reviews and then categorized the reviews into positive and negative sentiments with respective aspects. They has adopted POS tagger and WordNet for aspect extraction and opinion trend extraction based on that they has performed tweet classification into positive, negative and neutral class. System performance can be improved using machinelearning approach.

In this paper [10] sentiment analysis has been performed on smart phone product review dataset using SVM (Support Vector Machine) clustering of features using TF-IDF to improve result of traditional machine learning.

Paper [11] author has performed sentiment analysis on

amazon product review data using POS tagging and Negation Phrases Identification algorithm, classification

algorithms used in research study are Naive Bayes, Support Vector Machine and Random Forest.

Contribution in research paper are outlined as below:

As 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.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105304 |

III. METHODOLOGY

In this paper sentiment analysis has been performed by following steps. System architecture is as shown in a Fig 1.

A.DATASET

The research uses review data from various tourism websites.[12][13] 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 steps such as tokenization, Stop word removal, stemming and lemmatization.

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

II. 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.

III. 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.

IV. 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 Naive 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 reviews to 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 of positivity, 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.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105304 |

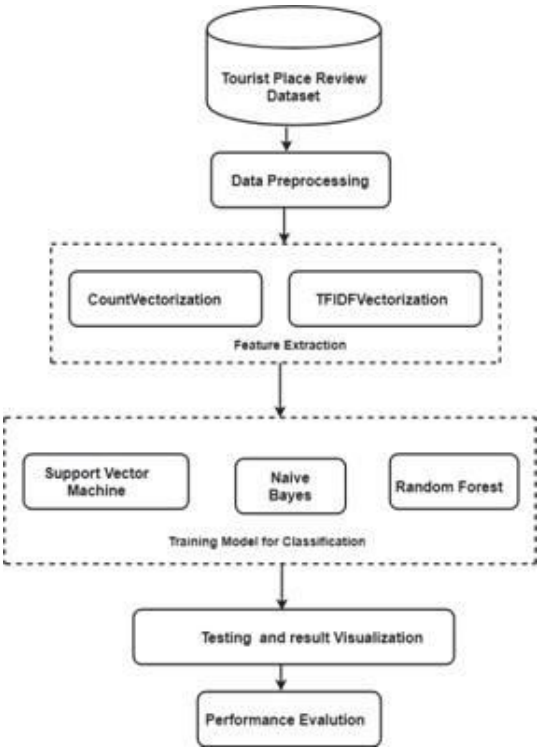

fig1.System Architecture

IV. SIMULATION RESULTS

A.EXECUTION TIME

Feature extraction algorithms has been compared on bases of execution time. Fig 3 shows execution time comparison. Figure shows that TFIDFVectorization require more time than CountVectorization feature extraction algorithm.

B. Accuracy Score

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Accuracy comparison graph as shown in Fig 4. From Fig 4 it infer that TFIDFVectorization+RF has better accuracy over other algorithms used.

C.Precision

It is also called positive predictive value. Precision comparison graph is as shown in Fig 5. From Fig 5 we can infer that TFIDFVectorization+RF has better precision over other algorithms used.

D.Recall

It is also called sensitivity. Recall comparison graph is as shown in Fig 6. From Fig 6 it infer that CountVectorization+RF and TFIDFVectorization+RF has better recall over other algorithms used

E.F1-Score

F1-Score is indicates balance between precision and recall. F1-score Comparison graph is as shown in Fig 7. From Fig 7 shown it infer that TFIDFVectorization+RF has best F1-score.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105304 |

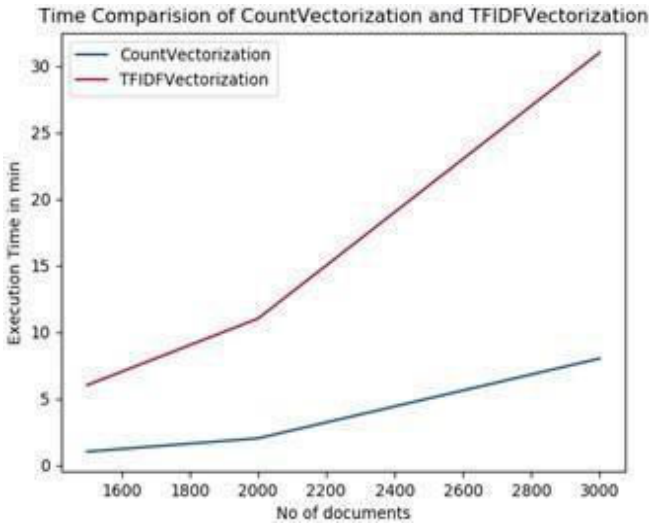

Fig. 3. Execution Time Comparison of Feature extraction algorithms

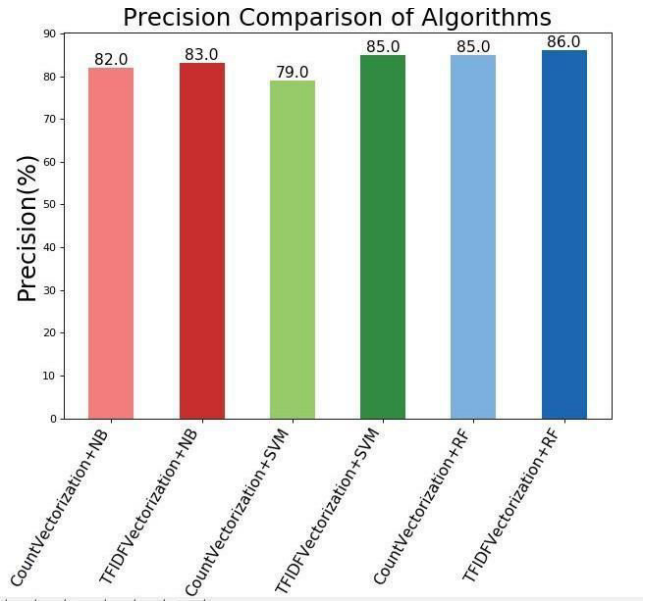

Fig.5.PrecisionComparisonofalgorithms

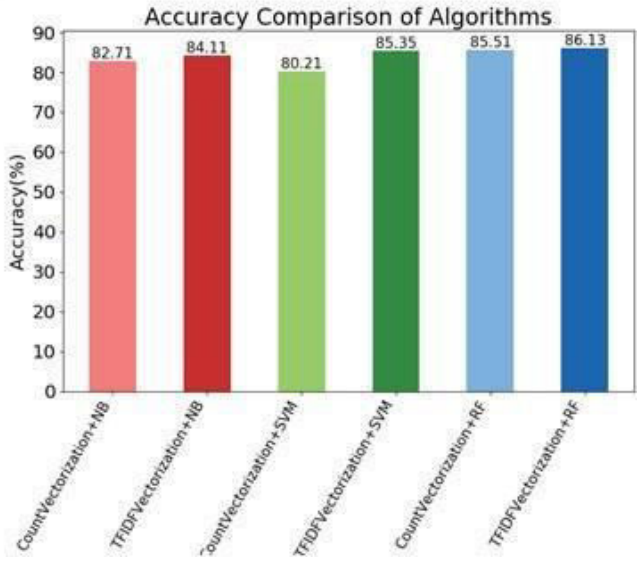

Fig. 4. Accuracy Comparison of algorithms

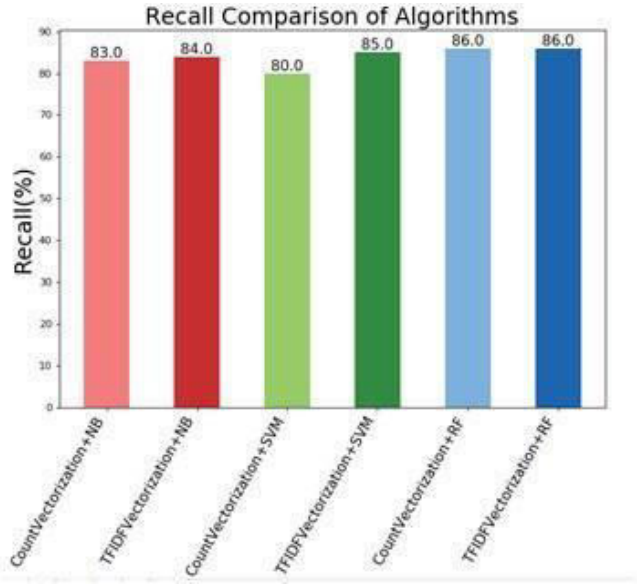

Fig. 6. Recall Comparison of algorithms

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105304 |

V. CONCLUSION AND FUTURE WORK

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), Naive Bayes(NB), Random Forest(RF) have been used. It has found that TFIDFVectorization+RF outperformed other algorithms used on bases of several evaluation parameters like accuracy, precision, recall and f1-score.

REFERENCES

1. M.D.Devika, C.Sunitha, Amal Ganesh "Sentiment Analysis: A Comparative Study on Different Approaches" ScienceDirect Fourth International 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, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", arXiv:1707.02919 [cs.CL], July 2017
5. Robert Dzisevicˇ , Dmitrij Sˇesˇok "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, Shivakumar Vaithyanathan "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 onEnergy, 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. https://www.tripadvisor.in/
13. https://www.mouthshut.com

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details