Sentiment Classification of Tourist Place Reviews

[1]Galipelli Saisreeja, [2] Dr. D. Srinivas Reddy

[12] Vaageswari College of Engineering, Telangana, India

[1]saisreejagalipelliknr@gmail.com [2]srinivasreddydhava@gmail.com

ABSTRACT

In today's society, using social media is very common. Tourism websites get millions of reviews and ratings from users. These testimonials can lead to best analysis that leads to reveal a destination's general level of popularity among tourists. Tourists can get the vacation spot from the data in websites. In this work, a machine-learning strategy for sentiment analysis is presented. The information in the dataset comes from a wide range of travel review websites. We have analyzed the best method in uprooting techniques, Count Vectorization, TFIDF-Vectorization and compared their performance. Besides the well-known NB (Naive Bayes), SVM (Support Vector Machine), and RF (Random Forest) classification techniques. Several metrics including accuracy, recall, precision, and f1-score, have been used in evaluating the algorithm's relative performances. In our experiments, in terms of classification accuracy for the test dataset, we found that the TFIDF Vectorization feature extraction method performed better than the count Vectorization methodology. TFIDF Vectorization, RF has shown the highest accuracy (86%) in a study classifying the sentiment of reviews written about tourism destinations.

INDEX TERMS Classification, TFIDV (Term Frequency-Inverse Document Frequency), Sentiment Analysis, SVM (Support Vector Machine), Random Forest, Machine learning.

I. INTRODUCTION

Social networking websites are becoming increasingly popular. Millions of individuals use travel review websites every day to express their thoughts and experiences about various tourist spots. All of these testimonies can be understood using sentiment analysis. Through rigorous research and reviews, a pattern in a location's appeal among tourists can be found. The compiled findings of sentiment analysis will help travelers make destination decisions and plan their ensuing itineraries. The Count Vectorization algorithm and the TFIDF Vectorization approach are two different feature extraction strategies used in this work to accomplish its objectives. The three classification techniques used for sentiment analysis are Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF). Only a few of the measures used to assess the effectiveness of various feature extraction and classification algorithm combinations include execution time, accuracy, recall, precision, and f1-score.

II. RELATED WORKS

Various sentiment analyses are used in this work [1]. Various sentiment analyses are used in this work [1]. methodologies are examined and contrasted. Document, sentence, and aspect are the three different sentiment levels that have been established. This study uses lexical, rule-based, and machine learning-based techniques for sentiment analysis. Among the machine-learning-based techniques listed are Support Vector Machine (SVM), Naïve-Bayes, Maximum Entropy, K-NN, Weighted K-NN, Multilingual Sentiment Analysis, and Feature Driven Sentiment Analysis. The advantages and disadvantages of various sentiment analysis techniques have been compared. Various metrics, such as performance, efficiency, and accuracy, have shown that machine learning is the most successful approach. A study on Twitter sentiment analysis of movie reviews is described in [2] article. Numerous supervised machine learning algorithms have been used in conjunction with a variety of feature extraction techniques (including unigrams, bigrams, and a hybrid of the two, unigrams plus bigrams) (SVMs, NBs, and MEs, to name a few). The study's findings demonstrate that SVM with a hybrid feature extractor is superior to the rival approaches.

According to [3], study, a survey has been done on both the principles of sentiment analysis as well as its use in a variety of fields and the various methodologies that are used

for sentiment analysis. Machine learning-based and lexicon-based approaches can be used to approach the topic of sentiment analysis. Dictionary-based and corpus-based lexicon-based systems are two subcategories of these systems. The two halves of a corpus-based method are statistics and semantics. The statistical approach looks for instances of the term, as opposed to the semantic methodology, which is focused on word similarity. Machine learning is divided into two categories: supervised and unsupervised. Only a few of the supervised algorithms described include support vector machines, neural networks, Bayesian networks, maximum entropy, and Naïve-Bayes. The author of this paper uses the Count Vectorizer and the TFIDF Vectorizer to extract features from reviews. These features are then applied to machine learning algorithms, and accuracy, precision, recall, and F1SCORE are calculated to compare the performance of the two feature extraction algorithms.

III. DATASET DESCRIPTION

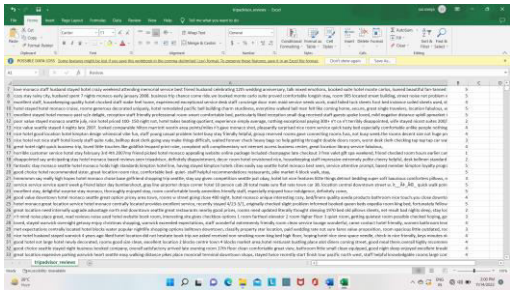


Fig 2: Dataset

We collected dataset from the Kaggle website

IV. METHODOLOGY

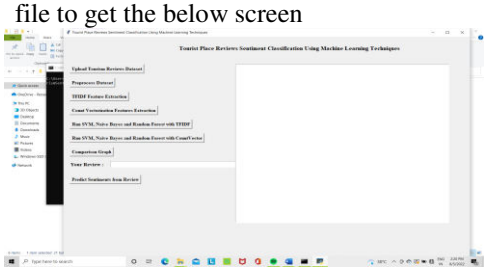
In this study, the author predicts attitudes from a dataset of traveler reviews using Machine Learning-methods including SVM, Naïve Bayes, and, Random-Forest, then measures how well Count-Vectorizer and TFIDF-Vectorizer do at extracting features. The author of this research uses Count-Vectorizer and TFIDF Vectorizer to extract features from reviews. These features are then applied to machine learning methods, and the accuracy, precision, recall, and F! SCORE of the two feature extraction algorithms are calculated.

A. IMPLEMENTATION

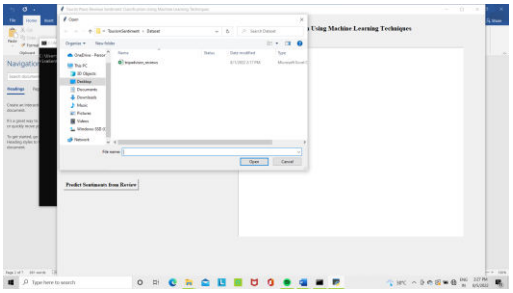
- **Upload Dataset**: Using this module we will upload reviews dataset to the application
- **Data Preprocessing**: Using this module we will read all reviews and then eliminate stop words and special symbols and apply NGRAM techniques to review clean text.
- **Count Vectorization**: using this module clean text will be converted to a count vector where each word count will be calculated and then a features vector will be generated.
- **TFIDF Vectorization**: Using this module Term-frequency (TF) and IDF (Inverse Document-Frequency) will be calculated and then generate a features vector.
- **RUN SVM, Naïve Bayes, and Random Forest with TFIDF**: Using this module we will train all 3 algorithms with TFIDF features and then calculate execution time, accuracy, precision, Recall, and, F1SCORE.
- **RUN SVM, Naïve Bayes, and Random Forest with Count Vector**: Using this module we will train all 3 algorithms with Count Vector features and then calculate execution time, accuracy, precision, Recall, and F1SCORE.
- **Comparison Graph**: using this we will visualize the performance graph of both feature extraction algorithms with various machine learning algorithms
- **Predict Sentiments from Review**: using this module user can enter his review and then the application calculates sentiments from that review

V. EXPERIMENT, RESULTS, AND ANALYSIS

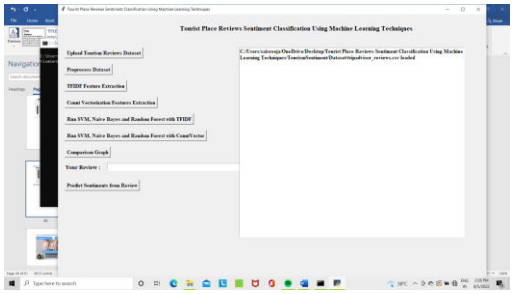
To run the project double click on 'run.bat' file to get the below screen



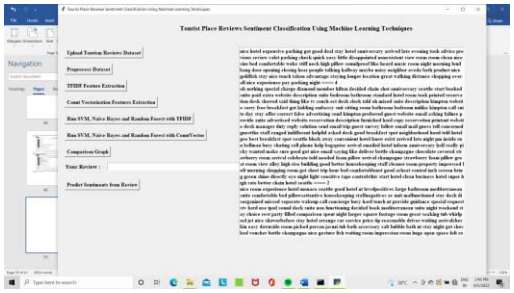
To run the project, double-click the 'run.bat' file to receive the screen shown below.



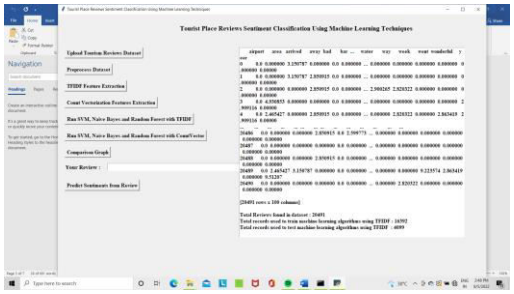
Choose the "TripAdvisor reviews.csv" file from the above screen, upload it, and then click "Open" to load the dataset and display the screen below.



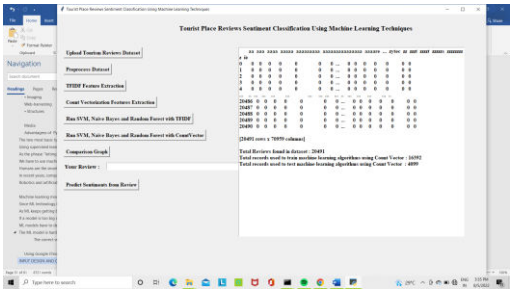
On the preceding screen, a dataset is loaded; now, click the 'preprocess Dataset' button to read and clean the dataset.



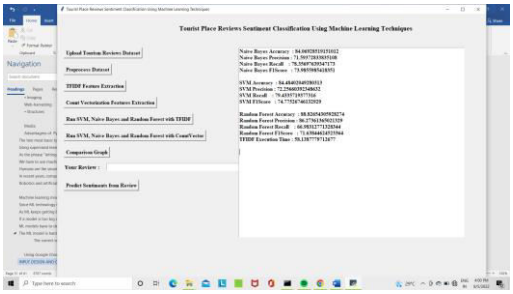
The dataset's reviews are read and displayed in the text area of the screen above. To begin extracting features using the TFIDF technique, click the "TFIDF Feature Extraction" button.



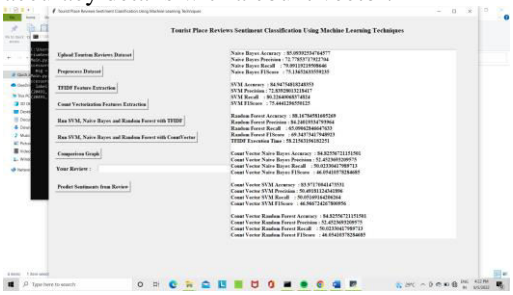
For each word, we can observe the total number of reviews found in the dataset and application, with 80% utilized to train and 20% used to test machine learning techniques. Select 'Count Vectorization' now. The TFIDF for each word was calculated in the above screen application, and the numerical TFIDF values can be viewed. Use the Feature Extraction button to count each word.



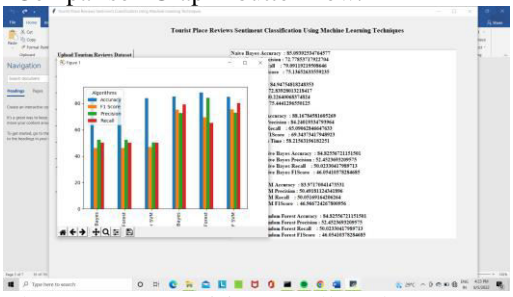
Click the "Run SVM, Naïve-Bayes, and Random Forest with TFIDF" button to train three machine learning algorithms with TFIDF features and obtain accuracy and other information. The above screen has a count vectorizer, and both the TFIDF and the count vector are now available.



TFIDF took 35.77 milliseconds to execute in the above screen's selected text for all three algorithms, and you can now click the button that says "Run SVM, Naïve-Bayes and Random Forest with Count Vector" to see accuracy details with a count vector.

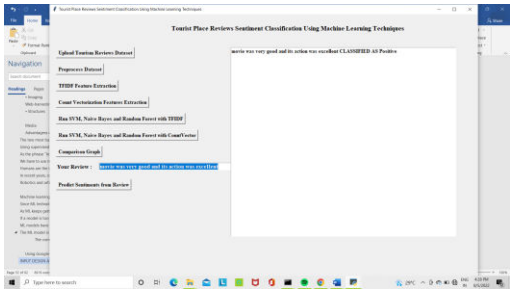


The count vector in the screen above took 51 milliseconds, and it also had lower accuracy, precision, recall, and F1-SCORE than TFIDF did. To access the graph below, click the "Comparison Graph" button now.

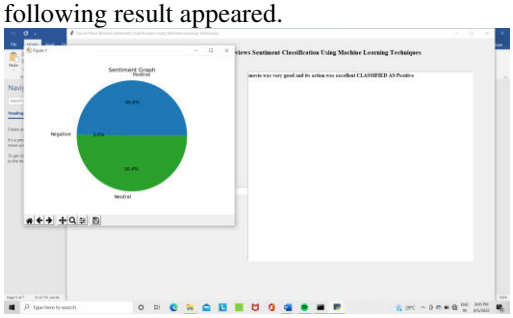


The accuracy, precision, recall, and F1SCORE are displayed on the y-axis in the graph above, while the method name is displayed on the x-axis. The accuracy line is shown in blue, the precision line in yellow, the recall line in

green, and the F1SCORE line in red in the graph above. We can see that TFIDF performed better at making predictions than count. To examine the results, type any review into the text field and choose "Predict Sentiments from Review."



I typed some feedback into the text field on the screen above, clicked the last button, and the following result appeared.



Positive, negative, and neutral values are displayed on the graph and the text area of the above screen, respectively.

VI. CONCLUSION AND FUTURE WORK

Based on the results of the investigation, TFIDF-Vectorization appears to be a more effective feature extraction approach than Count Vectorization. However, the TFIDF Vectorization technique takes more time to execute than the Count Vectorization approach does when it comes to feature extraction. Some examples of classification algorithms used in studies include the Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF). Using metrics like accuracy, precision, recall, and f1-score, TFIDF Vectorization + RF was found to be superior to other techniques.

The research study for machine learning-based review classification of tourist destinations has the potential to handle multilingual review classification in the future. Additionally, in an effort to increase classification accuracy, we will test using alternative feature selection techniques such as recursive feature elimination with cross-validation. For better performance in upcoming work, we'll strive to apply deep learning-based algorithms for feature extraction and categorization.

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, 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 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