Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques

[1]Prof. Gayatri Naik,[2]Mr. Sushant Wani, [3]Mr. Rushikesh Pawar, [4]Mr. Rohan Randhir
[1]Asst.Professor,[2,3,4]UG Student,[1,2,3,4] Computer Engg. Dept. Shivajirao S.Jondhle College of Engineering & Technology, Asangaon, Maharshatra, India. [1]krishngita123@gmail.com, [2]sushantwani65@gmail.com, [3]rushikeshpawar1804@gmail.com, [4] rohanrandhir22@gmail.com,

Abstract- Social networking is getting increasingly popular these days. On tourism websites, millions of users evaluate and rate tourist attractions every day. These reviews may be subjected to sentiment analysis, which will aid in determining the popularity of tourism destinations. A ML approach was used to implement sentiment analysis in this research The information came from a number of different travel review websites. The feature extraction techniques CountVectorization and TFIDF-Vectorization were compared in this study and the classification techniques naive bayes, Support Vectors Machines, and Random Forest too. Various measures such as accuracy, recall, precision, and f1-score were used to compare algorithm performance. TFIDFVectorization feature extraction approach offers better classification accuracy than CountVectorization for the provided review dataset, according to the experiment.It has delivered the greatest accuracy of 86 percent for a study dataset utilised in sentiment categorization of tourist site reviews.

Keywords-Supervise algorithm,RF(random forest).

I. INTRODUCTION

Nowadays, social media is quickly expanding. On a daily basis, millions of users publish-reviews & evaluate tourist attractions on tourism websites. This review may be analysed using sentiment analysis. A proper examination of evaluations will reveal a pattern in the famous places of tourist locations. The results of the sentiment research will be summarised to assist travellers in deciding on a holiday destination and itinerary. The CountVectorization and SVM Vectorization techniques were employed in this research study to extract features. For sentiment categorization,three classification algorithms were used: naive bayes (NB), Support-Vectors-Machines (SVM), & Random Forest (RF). Performance for a feature combination is calculated using variables such as output time, accuracy, recalls, precision, and scores.

The algorithms for extraction and classification have been compared . This paper's material is organised as follows. In this section This study examine at sentiment analysis literature surveys.

It outlines the fundamental notion of machine learning ,our sentiment analysis methodology for tourist site review categorization, visualisation, and performance evaluation are also described. It also shows how to calculate the popularity distribution of tourist destinations using machine learning methods. It shows the outcomes of the experiment that was carried out & offers a comparison of sentiment analysis employing machine learning methods that were employed in the research project & brings this study report to a close. Also includes a narration of the research paper's future scope of tourist review classification using machine learning.

II. AIMS AND OBJECTIVE

a) Aim

The project aim is, to identify if the material is good, negative, or neutral. This research intends to investigate and illustrate the value of tourist place analytics using sentiment analysis in order to better understand crucial concerns, such as the link between the sites where tourists go the most and their satisfaction ratings.

b) Objective

The objective of planning input is to make information section more straightforward and to be liberated from mistakes. The data segment screen is set up accordingly that all of the data controls are accessible.It similarly gives record seeing workplaces. It is accomplished by making easy to use evaluates for the information passage to deal with enormous volume of information.

III. LITERATURE SURVEY

Sentiment-analysis can be applied to these reviews, which will aid in determining the popularity of tourism destinations. Tourists can easily choose a tour destination based on the sentiment analysis results. Sentiment analysis has been implemented utilising a ML approach in this work. SVM (Support Vector Machine), NB (Naive Bayes),

 DOI : 10.35291/2454-9150.2022.0161

Maximum Entropy, K-NN, and Weighted K-NN are examples of machine learning approaches.

Paper 1: Sentiment Analysis: A Compar-ative Study On Different Approaches:

Sentiment-analysis (SA) is a method of eliciting a user's feelings and vibes using logic.It is one of most gains fields of (NLP). The expansion of Internet-based apps has resulted in a flood of personalised evaluations for a variety of internet resources. Sentiment Analysis is a sophisticated technique that allows users to extract necessary information as well as mixed reviews' collective sentiments .

Paper 2: A Novel Frame-work for aspect-based opinion categorization in tourist destinations:

Before visiting a city or nation, tourists want to know the advantages and drawbacks of the tourist attractions. They usually utilise social media sites to read what previous guests have to say. The analysis in this work is based on three types of opinion mining techniques: trend-based opinion minings, aspect-based opinion mining, and sentence-based opinion mining. has collected useful information from tweets, reviews, and travel blogs.

Paper 3: A Comparative analysis of Twitter data using supervise-classifiers:

Online social media microblogging is used to share opinions on specific topics in very short messages.There are some famous microblogs such as Twitter, Facebook, etc.where Twitter gets the most attention in areas like product research, movie reviews, stock market, etc. The applied supervise ML algorithms are (SVM), maximum-entropy & naïve-bayes for classification of data using the unigram, bigram and hybrid.

IV. EXISTINGSYSTEM

Customers can become active users by providing reviews of various products/services that other potential customers may find useful, but there are thousands or even more product/service reviews on the Internet, and reading all of these reviews is a very important task. And it's hard for clients. However, there has been almost little study done on tourism assessments to define their feelings. To classify product evaluations based on their sentiment polarity, The findings of a unique feature vector generating approach were unreliable. The study's purpose is to determine how comfortable people are with technology. This involves the process of instructing the user on how to utilize the system effectively. Rather then being fearful of the system, the user should accept it as a need. The amount of acceptability by consumers is totally determined by the tactics used to teach them about it. As a result, the produced system was also developed within the plan, which was accomplished because the majority of the technologies used were freely viewable. Tourism industry is also based nature calamities like floods, storms, tsunamis, volcanic eruption etc so weather forecast will be one of the important factor in tourist places as tourists safety is prior first in the tourism industry.

V. COMPARTIVE STUDY

SR NO.	PAPER TITLE	AUTHOR NAME	METHOD	ADVANTAGE	DISADVANTAGE
1.	Sentiment Analysis: A Comparative Study on Different Approaches	M. D. Devika, Sunitha Amal Ganesh	support vectors machines (SVM)	Sentiment Analysis is a sophisticated technique that allows users to extract relevant information as well as assemble the reviewer's combined sentiments.	Time Consum-ing
2.	A Novel Frame-work for aspect-based opinion categorization in tourist destinations	Muhammad Afzaal, Muhammad Usman	Trend based, aspect-based, and sentence based opinion mining	Good Approach Explained	Difficult to understa-nd
3.	A Comparative analysis of Twitter data using supervise classifiers	Rohit Joshi, Rajkumar Tekchandani	SVM maximum entropy and naive bayes	Best Approach Explained	Little bit Time Consum-ing

VI. PROBLEM STATEMENT

This project works around the entire tourism industry. Due to global pandemic like covid-19, entire tourism industry faced a lot of loss.That loss affected the services tourists get, those free & paid services are becoming costlier, so the new datasets will include service prices and charges.Currently only limited datasets are trained and accuracy is up to 80%.This project's updated version has the accuracy of up to 80% and big datasets will be trained easily. As a result, the customer will be held to unrealistic standards. The planned system should have a minimal need since very small or no adjustments are required to deploy this system..

VII. PROPOSED SYSTEM

Various strategies of sentiment-analysis have been researched and compared in the suggested method. This plan is essential to keep away from messes up in the information input affiliation and show the right scrambling toward the association for getting right data from the electronic situation.There are numerous levels of sensations that have been created, including document, phrase, and aspect levels.

 DOI : 10.35291/2454-9150.2022.0161

Machine learning, rule-based, and lexical approaches were employed in this research for sentiment-analysis. SVM (naïve bayes),& feature-driven sentiment analysis are few of the techniques used in machine learning methodologies. Various techniques to sentiment analysis have been examined, and the benefits and drawbacks of each have been thoroughly documented. Various comparative metrics, such as performance, efficiency, and accuracy, have disclose that the ML ideas are produces the greatest results. Product reviews, movie reviews, restaurant reviews, blog entries, and other comparable areas have all gotten a lot of attention. Sentiment-analysis in the tourism area Researchers have looked into approaches like Naive bayes and SVM for it.

VIII. ALGORITHM

The working of proposed system algorithm is given as follow:

Step.1: Start

```
#importing libraries
from admins import views as admins
from sklearn.ensemble import
RandomForestClassifier from
Sklearn. from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features = 1420)
```

Step.2: Processing dataset

```
import files from section google colab #Only
use for Google Colab uploaded ←
files.upload() #Only use for Google Colab
df ← pd.read_csv("tourist_places.csv")
```

Step.3: Feature selection process

```
 - feature selection method implement
through univariate selection and correlation
matrix and feature importances
```

Step.4: Classification algorithm implement

```
1. Random forest
clf←Model(model←RandomForestClassifie
r(),X←X,y←y) clf.crossValScore(cv←10)
clf.crossValScore(cv←10)
clf.accuracy()
clf.confusionMatrix()
clf.classificationReport()
st_x= StandardScaler()
x_train← st_x.fit_transform(x_train)
x_test ← st_x.transform(x_test)
2 .Naïve bayes
load the iris dataset
from sklearn.datasets import load
 # store the feature matrix (X) and response vector (y)
X←iris.data
y←iris.target
# splitting X and y into training & testing sets
From_sklearn.model_selection
import train_test_split
X_train, X_test, y_train, y_test ← train_test_split
(X, y, test_size←0.4, random_state←1)
```

Step.5: training the model on training set

```
 from sk-learn.naive_bayes
import GaussianNB
gnb = GaussianNB()
gnb.fit(X_train, y_train)
```

Step.6: Making predictions on the testing set

```
y_pred ←gnb.predict(X_test)
 # comparing actual response values (y_test) with predicted response values
(y_pred)
from sklearn import metrics
print("Gaussian Naive Bayes model accuracy(in%):",
metrics.accuracy_score(y_test, y_pred)*100)
dt←Model(model←DecisionTreeClassifier(
),X←X,y←y)
dt.crossValScore(cv←10)
dt.accuracy()
dt.confusionMatrix()
dt.classificationReport()
```

Step.7:Visualizing_result_from_above algorithm.

Step.8: Exit.

IX. MATHEMATICAL MODEL

1. Feature Extraction

In machine learning, features-extraction is a crucial stage. Because the data in NLP is in textual format, the word is a feature. The word called as token. For the machine learning classifier to operate, the document must be in vector format. Tokens must be converted into characteristics., which is known as feature extraction. Textual feature extraction may be done using a variety of approaches.

CountVectorization:

As name indicates Count-Vectorization feature extraction algorithm counts occurrence of word i.e token in document and transform whole document into document term matrix. Its analogous to bag-of-word (BOW) approach. Below is formula for Count of term.

Count(word) = No of occurrence of a word in particular document.

TFIDFVectorization:

It is term frequency inverse document frequency. It calculate weight of term in particular document by considering occurrence of term in document as well as in whole corpus then it transform whole document into document term matrix. Following are formulas for TF-IDF calculation.

$$TF=\frac{No\ od\ times\ particular\ word\ w'occurs\ in\ a\ documnet}{Total\ no\ of\ words\ in\ a\ document}$$

$$IDF=log\left(\frac{Total\ no\ of\ document}{No\ of\ documents\ contains\ particular\ word}\right)$$

$$TF\text{-}IDF=TF\times IDF$$

Equation1-Evaluates term frequency of particularword.Equation2-determinesinverse document frequency of word Equation 3 finds TF-IDF i.e term frequency .

2. Supervised Algorithms

Supervised-machine algorithms ideas are used to train a well-labeled dataset. It implies that, similar to learning with a supervisor, input and output are delivered in the structure of a dataset. A supervised algorithm technique may be split into two categories. They're called classification and regression, respectively. In this research project categorization of some classification algorithms are listed below.

Multinomial Naive Bayes &

 DOI : 10.35291/2454-9150.2022.0161

Linear Support Vector Machine.

3. Random Forest

Random-forests are a mixes of several decision tree predictors in which each tree is reliant on the merit of a random-vector collected independently-data & with the same grouping across the forest. It's termed Ensemble approach since it makes a final forecast based on a collection of findings. More information may be established in the references.For each decision-tree, Scikit-learn compute a node importance for each decision-tree, assuming minimum two child nodes of binary tree:

When solving regression problems with the Random Forest Algorithm, used the mse to figure out how each node's data forks.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

Where N is the number of data points,
fi is the value returned by the model and
yi is the actual value for data point i.

This formula calculates the-mean square error of every branch upon a node based on the class and probability, deciding which branch is more certainly to occur. The class's relative frequency This dataset having at in the dataset is represented by pi, and the amount of classes is represented by c.

X. SYSTEM ARCHITECTURE

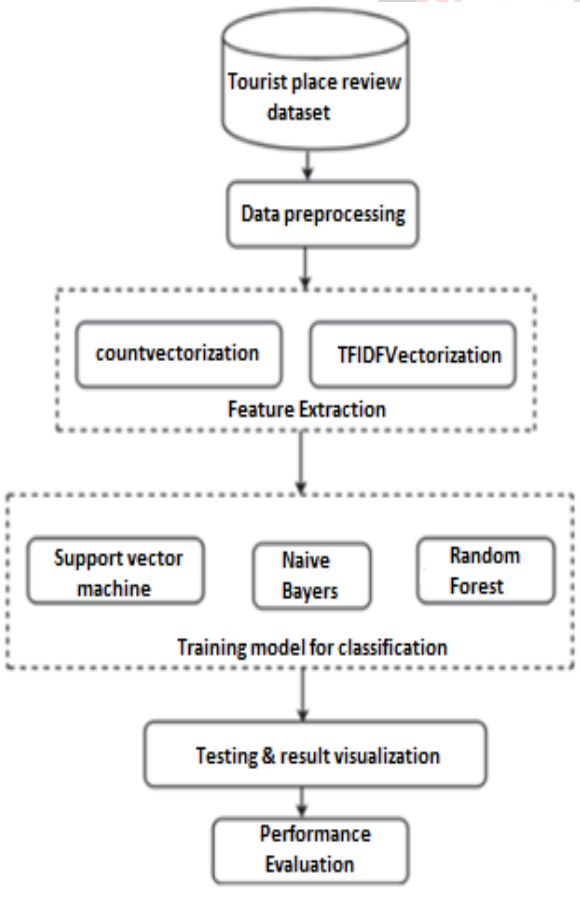

Fig.1: System Architecture

Description:

Data collected from various tourism websites was raw so need of data preprocessing to eliminate irrelevant and unuseful words from reviews. In data preprocessing stop words, punctuation mark, short word has been removed. Also tokenization, lemmatization, stemming has been performed. Data preprocessing is crucial step which help to feature reduction and better performance of machine learning algorithms. After data cleaning, Feature extraction algorithm has been implemented from scratch. CountVectorization and TFIDFVectorization has used for feature-extract.

XI. ADVANATGES

1) The interface is simple,user friendly and attractive It will help tourists to choose right place for them.

2) TFIDF-Vectorization gives accurate prediction upto 86% which can be useful in other industries also.

XII. DESIGN DETAILS

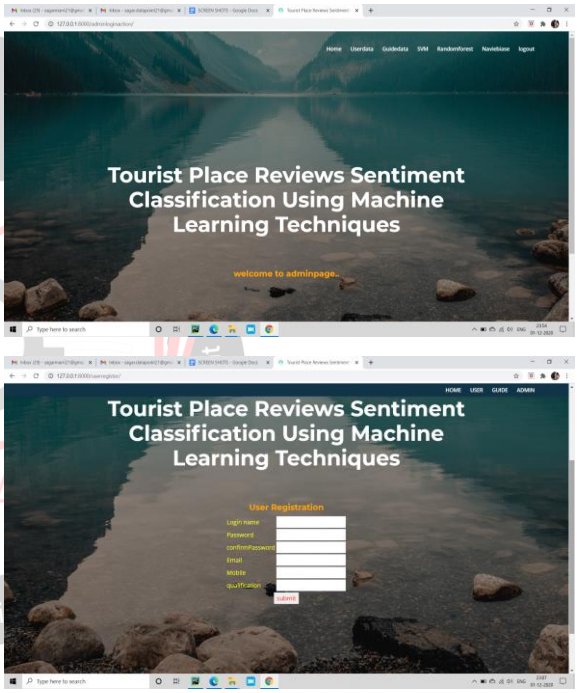

Fig 2: Result

XIII. CONCLUSION

Thus, we have tried to implement the paper " Apeksha Arun Wadhe, Shraddha S. Suratkar", "Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques", February-2020 I4Tech,IEEE organization & According to this paper TFIDF-Vectorization 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 Random Forest, Support Vector Machine and Naïve Bayes has been used. It has found that TFIDFVectorization+RF outperformed over other

 DOI : 10.35291/2454-9150.2022.0161

algorithms used on bases of several evaluation parameters like accuracy, precision, recall and f1-score.

REFERENCE

[1] Wadhe, Arun. Apeksha., & Suratkar, Shraddha. S. (2020, February). Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques. In 2020 International Conference on Industry 4.0 Technology (I4Tech) (pp. 1-6). IEEE.

[2] Amal Ganesh, M D Devika, C-Sunitha ScienceDirect Fourth International Conference on Recent Trends in Computer Science Engineering, "Sentiment Analysis: A Comparative Study on Different Approaches."

[3]Rohit-Joshi,Rajkumar-Tekchandani Comparative analysis of Twitter-data using supervised classifiers" 2016 International Conference on Inventive Computation.

[4] "A Survey of Sentiment Analysis Techniques," by Harpreet Kaur,-Veenu Mangat, and Nidhi, presented at the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud).

[5] "A Brief Survey of Text Mining: Classification, Clustering, and Extraction Techniques,"by_MehdiAllahyari,Seyedamin-Pouriyeh, M Assefi, Saied S, E D. Trippe, Juan B-Gutierrez..

[6] Dzisevic-Dmitrij-Sesok-Robert Open Conference of Electrical, Electronic, and Information Sciences "Text Classification Using Different Feature Extraction Approaches Text Classification Using Different Feature Extraction Approaches"

 DOI : 10.35291/2454-9150.2022.0161