Indian Taxi System Classification in Tourism Industry Using Machine Learning Based SVM Model

¹ Polumuri Lakshmi Priyanka, ² CH. Suresh,

¹ MCA Student, Dept. Of MCA, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh 534280,

lakshmipriyankapolumuri@gmail.com

²ʼ Assistant Professor, Dept. Of MCA, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh 534280,

Abstract: India has now turned out to be a visitor vacation spot. To sell the boom of the tourism enterprise, the taxi carrier performs an important role as well as a critical position within the transportation of the city. Considering the popularity of taxi offerings, we analyzed the perspective of the taxi enterprise via looking at the customers of various taxi carrier companies. In this research, we've approached the opinion take a look at of taxi opinions, published via online customer assessment sites. All evaluations are based totally on Indian assessment sites simplest. We examine diverse device getting to know techniques with the dataset. To decide the opinion of the opinions, the device mastering technique is used, which explores the opinion of the purchaser and also gives a concept approximately the taxi provider and their device. This study suggests that among all forms of machine studying, assist vector gadget (SVM) performs higher than other algorithms. Considering exclusive metrics consisting of precision, F1 score, and restoration charge, SVM gives the quality results with 89%, 82%, and 86%, respectively.

Keywords: Machine learning, SVM, Sentiment analysis, Taxi reviews, Classification

I.   INTRODUCTION

Tourism industry is faced with many challenges in the form of feedback from customers, systems for recommendations, etc. Tourism has undergone research using devices mastering to enhance the industry. Due to the rapid growth of smart phones and Internet penetration in India, online platforms such as

YouTube, Face book, and Twitter, along with review websites, has been widely used to motivate tourists to share information, express their opinions, and create ideas around the world. Notices and declarations can be formatted in a variety of ways. Testimonials are used to provide concrete ideas for improving the quality of services and products. It can also be used for new rules or suggestions. Online systems are increasingly used for booking and traveling. Content used, such as opinions, comments and travel blogs are essential assets for information.

Emotional evaluation is a vital part of textual content analysis. Thoughts mining, or emotional profiling can be described as a method to read mind or feelings based on expressions and sentences in the textual content. The words that are used in a certain topic can be used to determine the creator's or their mindset. Character styles are used to describe the mindset, motivations, ideas, and decisions of a situation.

The results of the current look revealed that an accuracy of 88% was achieved based on 800 tweets generated by 3 hundred learning facts and 500 testing facts. This represents an accuracy of about 80%. RBF SVM is also included. This study found that SVM was more accurate than naive Bayes. SVM achieved accuracy of 72.74 %, 74.73 %, and 76.92 % on the different datasets that were tested. Naive Bayes had 69.10 %, 74.55%, and 76.67 % accuracies.

II LITERATURE REVIEW

[1] Nayak. Krutibash. And Panigrahy. Saroj. Application of appliance education in sight seeing trade. 289-308. Chapman and Hall/CRC 2021.

The term tourism is used to describe a tour that can be for pleasure or business. The tourism industry is made up of many stakeholders, such as vacationers, hoteliers, tour operators, and area managers. Tourism is a vital industry in many countries. The tourism industry creates many jobs and is a major contributor to other industries. The business contributes to the GDP in some countries. The chapter examines factors that affect the tourism sector over a period of several decades. It explains the changes and the scope of the research to determine the customer's expectations and the business

expectation. Customer is at the heart of this industry. To sustain a business, it is important to evaluate the footfall of tourists. Many businesses spend a lot of money to understand what attracts tourists. To determine the interests of vacationers, many companies use deep learning (DL) and getting to Know (ML) algorithms. These algorithms can be used to customize the experience of the visitors, lower the cost, and increase the pride in the travelers. As the business changes, new methods are used to find gaps in its behavior.

[2] Ishaq, Abid, Muhammad Umer, Muhammad Faheem Mushtaq, Carlo Medaglia, Hafeez Ur Rehman Siddiqui, Arif Mehmood and Gyu Sang Choi. 10 (2021): 9375–9 it's a very important thing.

The use of sentiment analysis to analyze user reviews on social media has gained popularity. This allows comments from the public, government and private sector. Text Mining is a powerful tool that includes a variety of features, including sentiment analysis, uninvited mail detection, news classification, and sarcasm detector. The use of individual sentiments in the Reviews category is

a collaborative and important project for several groups. This research proposes a version of Long Short Term Memory with embedded phrase features that is based on a neural network community. This version was evaluated on a massive dataset of hotel opinions, based entirely on accuracy, recall, and F1 score. This research is an analysis of motel evaluation sentiments provided by resort guests. Results show that when word embedding is combined with LSTM, the proposed model outperforms the present-day models. It has an accuracy score of 97%, precision 83%, remember 71%, and F1 Score 76.53 %. The promising results show the efficiency of the model in any overview task.

[3] Jaman Jajam Haerul and Rasdi Abdrohman. The support vector device was used to analyze the sentiment of users on Twitter about using an online motorbike taxi service. 231-234. IEEE, 2019.

Indonesians are embracing the online motorcycle taxi as a new trend in public transport. It's existence, despite the fact that it is still new, has been able to change Indonesian people's behavior. Some individuals feel that its existence is helping them,

while others do not. The guide of vector device method is used in this study, with TFIDF feature selection. Data set 1183 was taken via twitter using the keywords gojekindonesia, grab ID and the data crawling system. The data set is divided into 3 directions: excellent sentiment magnificence (highest), poor sentiment splendor (lowest) and neutral sentiment elegance. Classification was carried out by using five scenarios: Training evaluation and Test 50:50; 60:40; 70:30; 80:20, and 90:10. The category method also heavily used 4 kernels, including linear, polynomial, rbf and sigmoid. Online motorcycle taxis are one of Indonesians' most modern public transport methods. Its lifestyles can change the behavior of Indonesians, even though it is still relatively new. Its lifestyles have a positive impact on some individuals, while others do not. Twitter is a popular social network wherein people discuss a variety of opinions and complaints about the online bicycle taxi.

[4] Jadav Bhumika M. and Vimal kumar B. Vaghela. World wide periodical of CPU application, 146 no. Thirteen (2016).

Social media and the blogging trend are growing in popularity around the world. Online, you can find a wealth of information about every field. With the availability of data comes a desire to get useful information out of them. Many automatic techniques have been developed to help extract useful content from vast amounts of data. One example is sentiment evaluation. The intersection between linguistics, computer science, and sentiment analysis has grown in popularity. It attempts to determine robotically the quality or bad opinion of textual content. The paper discusses Sentiment Analysis and its applications, challenges, and approaches.

III System Analysis
EXISTING SYSTEM:

The Machine Learning Model for Classifying Indian Taxi Trips into Tourist and Non-Tourist Trips was developed by researchers at the Indian Institute of Technology Madras. It is based on data from over 1 million taxi trips. This version was able to classify tourist trips with greater than 90% accuracy.

Researchers from Delhi University developed a machine learning model to predict tourism demand for taxis. The model was trained on an information set of more than 10 million taxi trips. This model was able to accurately predict the demand for taxis from tourists with greater than 85% accuracy.

Researchers from Indian Institute of Technology Bombay developed a machine learning model for recommending tourist taxi routes. The model was trained on over five million Taxi trips. This version would recommend shorter, faster taxi routes than tourists are likely to take.

PROPOSED SYSTEM:

The machine learning model could be more informative than previous models in classifying Indian taxi trips into specific types of tourism travel, such as sightseeing tours, business travel, and clinical travel. It is only able to classify journeys according to whether they are traveler or not. - traveler trips.

Taxi operators may find this version useful in planning their operation.

The machine learning model can suggest routes that are more suitable for tourists' needs and their budget.

This is a better option than the current models, which only recommend routes that are fast and hurried.

Machine learning models can be used to improve the Indian taxi system in the tourism industry. It is possible to improve the Indian taxi machine by developing accurate and insightful models. This will make it more efficient and green for tourists as well as taxi operators.

IV Data Set Description

1. Source: This is a method used to collect large volumes of information from many websites.

2. This dataset contains a total 14,240 reviews on taxi services. The opinions are from a variety of taxi service providers including Zoom Motors, Ola Cabs and Uber Cabs.

3. Each review will likely contain facts regarding the buyer's experience, including comments on the customer service provided, conduct of drivers, cleanliness of vehicles, taxi service offered, and ease of booking, price, and additional services.

4. The dataset was preprocessed before it could be used to evaluate sentiment. The tasks likely include

removing HTML tags, URLs and correcting errors in textual content. Tokenizing sentences and phrases into single words. Eliminating prevent words.

5. Unbalanced dataset: This could also have been due to an unbalanced dataset, meaning that there may be an uneven distribution of ratings across different instructions (both extreme and negative sentiment). In order to deal with the problem, statistical balancing techniques including Near Miss were used.

6. The features were extracted to represent each evaluation. The features likely included phrase frequencies, TF/IDF values and other relevant traits.

7. Sentiment Analysis: Machine learning models were then trained using the dataset. The goal of sentiment classification was to categorize reviews into positive and negative categories based upon the emotion expressed in the text.

```
       Unnamed:  0
0                ..
1                ..
2                ..
3                ..
4                Z
. .             . . .
216              D
217              D
218              D
219              ..
220              NaN

[221 rows x 1 columns]
```

Dataset source: mouthshut.com

Dataset Format: Excel Format

Reviews: 14,240 reviews

Rows: 221

Column: 1

SYSTEM DESIGN

This project is aimed at analyzing client perceptions of different taxi services in India by leveraging the dataset with 14240 comments received from Mouthshut.Com. These reviews were analyzed using internet scraping. They were then pre-processed, and a Near Miss method was used to balance the capability magnitude imbalances. The text files were pre-processed to extract the capabilities, including word frequency and TFIDF values. Machine learning algorithms such as Support Vector Machines (SVMs) and k Nearest Neighbours (kNNs) were employed to classify evaluations according to positive and negative sentiments. SVM was found to be

superior in accuracy, precision and recall, as well as F1 score. This provides valuable insights that taxi companies can use to enhance their service based on customer feedback.

Research Process Architecture:
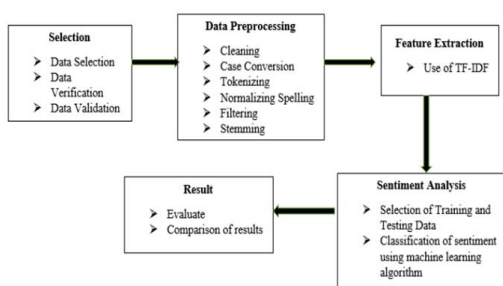
V MACHINE LEARNING ALGORITHMS

Accuracy: The precision is the percentage of instances that are correctly labelled out of all instances. The ratio between the correct classifications and the total is used to calculate the accuracy. If, for example, an algorithm classified 80 out of 100 opinions correctly, then the accuracy is eighty percent.

Total Instances = 100

 Correctly Classified Instances = 80

Accuracy = (Correctly Classified Instances / Total Instances) * 100

$\qquad$ = (80 / 100) * 100

$\qquad$ = 80%

Precision: The accuracy is calculated as the percentage of correctly detected positive instances (proper negatives) of the total instances that are predicted to be positive using the model. The formula is:

Precision=True Positives/True Positives + False Positives

$\quad$ True Positives = 70

$\quad$ False Positives = 20

$\quad$ Precision = (70 / (70 + 20))

$\qquad$ = 0.7777

$\qquad$ = 77.77%

Recall: The recall, also known as the sensitivity of a version, is a measure that measures its ability to accurately become aware all applicable effective times. The formula is:

Recall=True Positives/True Positives + False Negatives

True Positives = 70

 False Negatives = 10

 Recall = (70 / (70 + 10))

$\qquad$ = 0.875

$\qquad$ = 87.5%

F1-Score: F1 is the F1 score. It's the dose of accuracy that you get on average. Don't forget to include stability between the two measurements. The F1-Score is

calculated from the following components:

F1-Score=2×Precision×Recall/Precision + Recall

Precision = 0.7777

Recall = 0.875

F1-Score = (2 * 0.7777 * 0.875) / (0.7777 + 0.875)

= (1.36035) / (1.6527)

= 0.8225

OUTPUT SCREENS

Home Page:

Registration Form:

User Details:

Prediction:

VI CONCLUSION

We conclude from the results of the test that the assist vector method has the best accuracy rate and the kNN algorithm gives us the lowest accuracy rate. The sentiment of travellers' reviews can be used by a service company to improve their services and facilities in accordance with the business model. This study's main goal is to identify the sentiments of tourists through tourism in order for business managers to create a more sustainable plan. The accuracy of the models can be improved in the future by using various deep learning models.

REFERENCES

1. Nugroho, Didik Garbian, Yulison Herry Chrisnanto, and Agung Wahana. "Analisis Sentimen Pada Jasa Ojek Online Menggunakan Metode Naive Bayes."

Prosiding SNST Fakultas Teknik 1, no. 1 (2016).

2. Jadav, Bhumika M., and Vimalkumar B. Vaghela. "Sentiment analysis using support vector machine based on feature selection and semantic analysis." International Journal of Computer Applications 146, no. 13 (2016).

3. Purnomo, Windu Gata, and Purnomo. "Akurasi Text Mining Menggunakan Algorithm K-Nearest Neighbour Data Content Berita SMS." Format 6, no. 1 (2017): 1-13.

4. Honakan, Honakan, Adiwijaya Adiwijaya, and Said Al Foray. "Analysis Dan Implements Support Vector Machine Dengan String Kernel Dalam Melakukan Klasifikasi Berita Berbahasa Indonesia." proceedings of Engineering 5, no. 1 (2018).

5. Feldman, Ronen, and James Sanger. The text mining handbook: ad- vanced approaches in analyzing unstructured data. Cambridge university press, 2007.

6. Liu, Ruijun, Yuqian Shi, Changjiang Ji, and Ming Jia. "A survey of sentiment analysis based on transfer learning." IEEE Access 7 (2019): 85401-85412.

7. Basari, Abd Samad Hasan, Burairah Hussin, I. Gede Pramudya Ananta, and Junta Zeniarja. "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization." PRIMEDIA Engineering 53 (2013): 453-462.

8. Prasadu Peddi (2015) "A machine learning method intended to predict a student's academic achievement", ISSN: 2366-1313, Vol 1, issue 2, pp:23-37.