

Automatic Classification of Railway Complaints using Machine Learning

Tulasi Sathivika Roy, G. Vasukidevi, TYJ Naga Malleswari*, *S. Ushasukhanya*, Nayani Namratha

Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

*Corresponding author: nagamalt@srmist.edu.in

Abstract - People may now express their thoughts and ideas with a wider audience because of the popularity of social media sites like Twitter, Instagram, and Facebook. Businesses now utilise Twitter to reply to client comments, reviews, and grievances. Every day, millions of individuals discuss a wide range of issues on Twitter by sharing their ideas and interests. Sentiment analysis is a useful method for analysing such data, which involves identifying the sentiment of the source text and classifying it as positive, neutral, or negative. However, due to the vast amount of data, it can be challenging for businesses to address every customer's question or complaint in a timely manner. Some issues may be urgent but delayed due to the volume of information. In order to prioritize emergency tweets, a system is proposed that utilizes machine learning algorithms such as Random Forest, Support Vector Machine, Logistic Regression, and Naïve Bayes to identify tweets based on their urgency. The proposed system gathers and preprocesses unstructured data, performs feature extraction, trains, assesses and compares multiple machine learning models to determine the best classifier with the highest accuracy, and uses vectorization via a pipeline to determine the sentiment of a new tweet provided as input.

Keywords- Machine learning, Random Forest classifier, support vector machine, Logistic Regression, Naïve Bayes Classifier, Twitter API.

1 Introduction

India has recently surpassed China to become the world's most populous country. Train travel is a popular choice for long journeys in India, thanks to its ease, comfort, and affordability. However, various factors such as the quality of food, coach cleanliness, seat quality, and number of passengers can make train travel less enjoyable [1]. Indian Railways (IR) is one of the largest railway systems in the world, with 7,112 stations and 67,312 kilometres of track. As per the data published in 2017, IR is owned and operated by the

Indian government and comes under the control of the Ministry of Railways. IR operates 20,000 passenger trains daily and also handles freight and mail operations.

Over the past few years, the number of social media users worldwide has increased significantly, and connectivity has improved dramatically as a result. Social media has revolutionized how people communicate worldwide, as it allows for fast, immediate, and brief communication. Not only individual users, but many local and national authorities have turned to social media platforms for the same benefits. These organizations use social media to respond quickly to customer complaints and feedback.

There has been a significant increase in the number of complaint tweets received by Indian Railways Twitter (RailwaySeva) over the past few years. Their Twitter account receives thousands of tweets daily, tagging their account and requesting resolution. However, from these thousands of tweets, many require priority attention, such as tweets regarding accidents, incidents, threats, and medical emergencies. Some of the tweets received by Indian Railways are about service feedback, general queries, or reviews that do not require an immediate response and can be dealt with on a low priority.

To address this issue, an automatic classification system based on machine learning techniques is proposed. The proposed system can classify railway complaints into different categories, such as delay, cancellation, infrastructure, and customer service, among others. The system can help railway companies to automate the complaint handling process, reduce response time, and improve customer satisfaction.

In this paper, we present the design and implementation of an automatic classification system for railway complaints. We collected a dataset of railway complaints from various sources, including social media platforms, online forums, and customer service emails. We pre-processed the data to extract relevant features, such as keywords, sentiment, and context. We then trained and evaluated several machine learning models, including Support Vector Machine, Naive Bayes, Decision Tree, and Random Forest, to classify the complaints into different categories.

Sentiment analysis focuses on identifying and categorizing the attitudes or feelings expressed in the text. When finding the general consensus, user-generated data sentiment analysis is quite helpful. The inclusion of slang terms and misspelt words makes Twitter sentiment analysis more difficult than traditional sentiment analysis [6]. The knowledge base methodology and the machine learning strategy are the two techniques used to extract emotions from text. It is feasible to ascertain the influence of domain knowledge on sentiment classification by undertaking sentiment analysis in a particular domain. Future research in this area could include the development of a system that can prioritize tweets based on their urgency and automatically generate appropriate responses to each tweet, saving time and improving efficiency. Additionally, research could be conducted on the use of sentiment analysis to improve customer satisfaction and loyalty.

2 Literature survey

Indian Railway and social media [2], this paper discusses how social media can be used to improve business-to-consumer interaction for Indian Railways. It proposes the use of a new approach called Universal Language Model Fine-Tuning (ULMFiT), which involves training a language model and transferring its knowledge to a final classifier. To increase reproducibility, the ULMFiT model is fine-tuned by optimizing the parameters and trained in a deterministic approach. The study performs multi-class classification using various techniques such as Fine-Tuned ULMFiT, Naive Bayes, SVM, Logistic Regression, Random Forest, and K-Nearest Neighbors on the Twitter US Airline dataset from Kaggle. The results of the study can be used to improve Indian Railways' customer service by

identifying customer complaints and feedback on social media and addressing them promptly.

Twitter Sentiment Analysis using Machine Learning Techniques [3] In this paper, the authors explore the use of machine learning techniques for sentiment analysis on Twitter data. The main objective of the study is to determine the sentiment of tweets and classify them into positive, negative, or neutral categories. One of the key challenges in sentiment analysis is selecting an appropriate algorithm to accurately classify the sentiment of tweets. To address this issue, the authors examine several foundational classifier techniques, including Logistic Regression, Naive Bayes, Random Forest, and SVMs.

In particular, the authors focus on using the Naïve Bayes classifier and Logistic Regression for sentiment analysis, based on their superior accuracy in categorizing tweets. The results of the study indicate that the Naive Bayes classifier outperforms the other techniques in this context. By utilizing machine learning techniques for sentiment analysis, this paper demonstrates the potential for automated analysis of social media data, which can provide valuable insights for businesses and organizations seeking to understand customer sentiment and improve their overall customer experience.

Performance Evaluation of Learners for Analyzing the Hotel Customer Sentiments Based on Text Reviews [4,10] The study aims to tackle the challenge of analyzing the enormous amount of hotel reviews and opinions available on the internet. With the availability of large datasets containing text reviews, it has become possible to automate sentiment profiling and opinion mining. The researchers collected over 800 hotel reviews from travel information and review aggregator site Trip Advisor. They performed pre-processing on the collected raw text reviews and extracted various features using unigram, bigram, and trigram methods. The extracted features were labeled and used to train binary classifiers. The performance of ensemble classifiers, support vector machines, and linear models was compared and contrasted using accuracy, F-measure, precision, and recall measures.

One of the critical requirements for future research in this area is access to Twitter streaming data. This would allow researchers to apply similar techniques to analyze customer sentiments about hotels on social media platforms, such as Twitter.

Twitter Sentiment Analysis using Naive Bayes Algorithm [5] in this paper we have discussed that, Sentiment analysis searches for opinions, attitudes, and sentiments on social media platforms like Twitter. It is currently a well-liked academic subject. The classic sentiment analysis method places the greatest focus on textual data. Extraction of sentiment via Twitter, a popular microblogging platform where users submit their opinions and thoughts. We conducted sentiment analysis on tweets in order to make some business intelligence predictions. For processing a movie data set made up of reviews, comments, and feedback that can be found on the Twitter website, we employ the Hadoop Framework. The outcomes of the sentiment analysis on the data from Twitter will be presented in parts that represent positive, negative, and neutral sentiments.

Indian Railways Tweets Classification System using Naive Bayes Classifier [6], in this paper the proposed system will classify them and increase the rate of response to the complaints. It will improve data security as only verified admins will be allowed to access the twitter data. This model can be used by other authorities to classify tweets based on their requirement. This model can be adapted to do so by using different dataset for training. Naïve Bayes is a classification algorithm that is based on Bayes' theorem and calculates the conditional probability of an event based on the probability of occurrence of individual events, it helps generate an output with much higher accuracy than by any other method like linear regression.

Sentiment Analysis in Twitter using Machine Learning Techniques [7] this paper discusses the challenges of identifying emotional keywords from tweets, especially when dealing

with slang and misspelled words. To address these issues, the paper proposes an effective feature vector that includes Twitter-specific characteristics. The feature extraction is performed in two phases - first, the Twitter-specific features are extracted and added to the feature vector, and then these features are removed and feature extraction is carried out as usual. The resulting feature vector is used to train several classifiers, including Naive Bayes, SVM, Maximum Entropy, and Ensemble classifiers. The accuracy of all these classifiers is found to be similar when using the proposed feature vector.

3 Experimental setup

3.1 Motivation

The Indian Railways is a massive transportation network in India that requires an efficient system for resolving customer complaints. Compared to manual processing, machine learning can automate the process of classifying tweets related to railway complaints. In this study, multiple machine learning algorithms such as Decision Tree, Random Forest, Logistic Regression, and Naive Bayes are applied and tested on railway Twitter datasets. The best model is selected using an ensemble algorithm (voting classifier) for prediction [5]. A web application is developed with three modules: admin, complaint management, and user, to demonstrate the process of complaint solving.

3.2 Design

An activity diagram is a type of UML (Unified Modeling Language) diagram used to visualize the flow of activities or actions in a system or process. It shows the sequence of activities, decision points, and branching possibilities in a clear and easy-to-understand way. In the below fig 1, the user first logs in by registering with the system, then accesses the website to submit their issue. The admin then authorizes the complaint and uses machine learning algorithms to predict the issue.

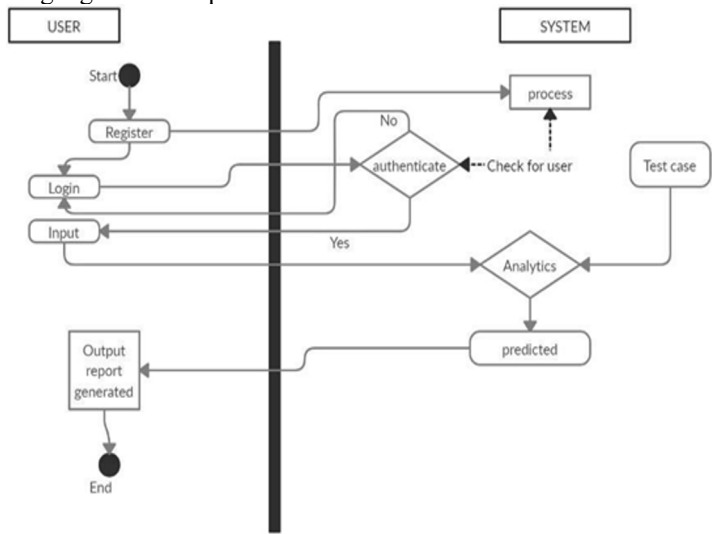


Fig. 1.Sequence Diagram

3.3 Requirements

The ability to access Twitter streaming data is essential for any system that performs sentiment analysis on tweets. To gain access to this data, a user must have valid Twitter API credentials. These credentials enable users to use Twitter API and stream real-time tweets[7]. Additionally, having a well-curated dataset is crucial for any processing system.

Typically, a dataset corresponds to the contents of one database table or one statistical data matrix, with each column representing a specific variable and each row corresponding to a given member of the dataset. In the case of the Indian Railways complaint management system, Twitter accounts are required for users to post complaints, which can then be addressed by the railway authorities.

3.4 Machine Learning

Machine learning is a subset of artificial intelligence that involves the use of algorithms and statistical models to analyze and interpret data, identify patterns, and make predictions or decisions without explicit instructions. It has become increasingly popular and relevant in recent years, as businesses and organizations generate vast amounts of data that can be leveraged for insights and decision-making. The inputs to the machine learning algorithms are railway tweets and sentiment, and the output is the predicted sentiment class. Multiple classification algorithms, including Naive Bayes, are experimented with to determine the best performing one. As shown in fig 2 The first step would involve collecting data related to railway complaints, which may be done using various sources such as social media, customer feedback forms, and other online forums. The collected data would then undergo several preprocessing steps such as text cleaning, removal of stop words, stemming, and other normalization techniques. The extracted features would then be used to train and test various classification algorithms such as Naive Bayes, Support Vector Machines (SVM) [8,9], Random Forest, and others.

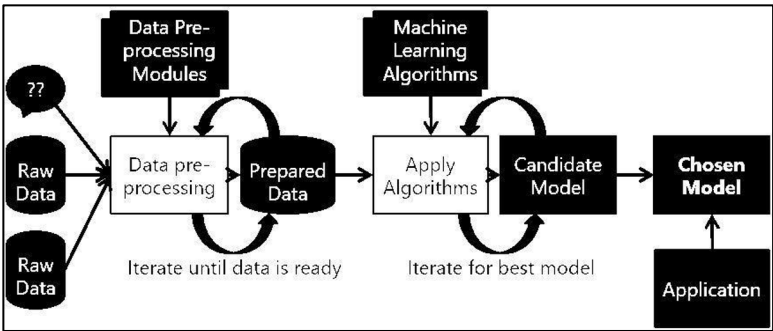


Fig. 2. Machine learning process

3.5 Dataset

Assuming the user wants to utilize a program for training a classifier using machine learning to analyze sentiments with three categories: positive, negative, and neutral. A CSV file containing two columns is used as the training data, with the first column (A) containing the sentiment (positive, negative, neutral) and the second column (B) containing tweets of up to 140 characters. We can have a large amount of training data.

To begin, we read the CSV file using the csv reader and create a list of datasets named Tweets. Each element in the list contains either emergency or feedback and a tweet. We then split each row into two parts: the first part as sentiment and the second part as tweet. For example, the first row contains emergency as the sentiment and "we are struck in the train" as the tweet.

In the pre-processings step, we remove all unwanted characters that do not help us determine the tweet's sentiment. Based on Figure 3, some of the actions performed in the preprocessing module are:

Lowercase conversion: We convert the tweets to lowercase.

URL's: We eliminate the URL's via regular expressions and replace it with the word URL.

@username: We eliminate "@username" and replace it with the word "AT_USER".

#(hashtag): We eliminate “#tag word” and replace it with the word “tag_word”. Additional whitespaces: We eliminate all the additional white spaces.

The function that generates the feature vector takes the processed tweets as input and outputs a list of meaningful words that have sentiment. It refines the processed tweets using several functions to generate the feature vector. These functions are as follows:

Stop words: Words that do not indicate any sentiment are removed from the tweets.

Repeating letters: Sometimes, people repeat letters in tweets to stress the emotion. To handle this, two or more repetitive letters in words are replaced by two of the same letter.

Punctuation: Punctuation marks such as colons, semicolons, commas, apostrophes, and question marks are removed from the tweets.

Words starting with an alphabet: All words that do not start with an alphabet are removed.

Words that start with a number or symbol usually do not have any sentiment and hence, are not relevant for sentiment analysis.

Item ID	Sentiment	SentimentText	processed_tweet	
0	1	0	@RailMinIndia My PNR is 8348062961. I am in way but there is no water in toilet and mess is everywhere in coach. please provide basic facil.	my pnr is am in way but there is no water in toilet and mess is everywhere in coach please provide basic facil
1	2	0	@sureshprabhu @RailMinIndia AC not working in prayag lucknow intercity today which departed from prayag at 3.45 pm. staff not helping us	ac not working in prayag lucknow intercity today which departed from prayag at pm staff not helping us
2	3	0	@RailMinIndia I'm traveling to chennai by train 16102 coach S4-46. My berth is very dirty(some one vomitted on the birth).	traveling to chennai by train coach my berth is very dirty some one vomitted on the birth
3	4	5	@RailMinIndia irctc is not responding at the time of tatkal booking.	irctc is not responding at the time of tatkal booking
4	5	7	@DRMBhopal @RailMinIndia @sanjaygupta2012 @dmrcrald Matter notified to concerned official @BhusavalDivn	matter notified to concerned official
5	6	6	@RailMinIndia If you can't give justice to graduate engineers then just do one thing at least. Stop recruiting B. Tech. as SSE.	if you can give justice to graduate engineers then just do one thing at least stop recruiting tech as sse
6	7	4	@sureshprabhu @RailMinIndiaPlz wrkout smthng tht cn hlp trains to run on right time Malwa exprs delayed more than 3 hrs @ Mathura junction	wrkout smthng tht cn hlp trains to run on right time malwa exprs delayed more than hrs mathura junction
7	8	0	@RailMinIndia @mumbairailusers Dirty water flowing frm blower of local coach aft 1st rain. 8.08 Thane Vashi https://t.co/mBZUOFkLXQ	dirty water flowing frm blower of local coach aft st rain thane vashi
8	9	6	Enter to win \$150 Amazon Gift Card! #Books #PNR #UrbanFantasy #Romance https://t.co/jB0xZe7P2	enter to win amazon gift card books pnr urbanfantasy romance
9	10	3	@RailMinIndia still vendors are selling local local products like water etc. inside the train https://t.co/wdFyLimMek	still vendors are selling local local products like water etc inside the train

Fig. 3.Dataset

3.6 Algorithms

To forecast the category class labels, it is employed. It utilises the training set and the values of a classifying attribute to categorise the class data and then applies that classification to fresh data, Model Use and Model Creation are the two steps in the procedure [5]. Both continuous and categorical values characteristics may be solved with it.

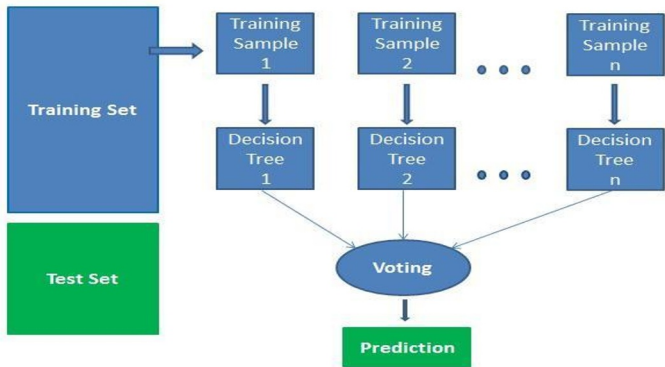


Fig.4.Random Forest algo

The Bayes theorem determines the likelihood that a new event will occur based on the likelihood that a prior event will occur.As we can see in figure 4, Using the provided dataset, select random samples.
For each sample, make a decision tree, and predict the outcome. Aggregate the predicted outcomes through voting. Select the outcome with the highest number of votes as the final prediction.
From fig 5, The SVM (support vector machine) is a popular method applicable to both classification and regression problems. Its main objective is to identify the optimal decision boundary or hyperplane that separates data points into different categories in n-dimensional space. The hyperplane is chosen to enable new data points to be easily classified in the future. The SVM method identifies extreme points and vectors that contribute to the creation of the hyperplane. This approach is based on support vectors, which represent the extreme situations and are used to build the SVM model.

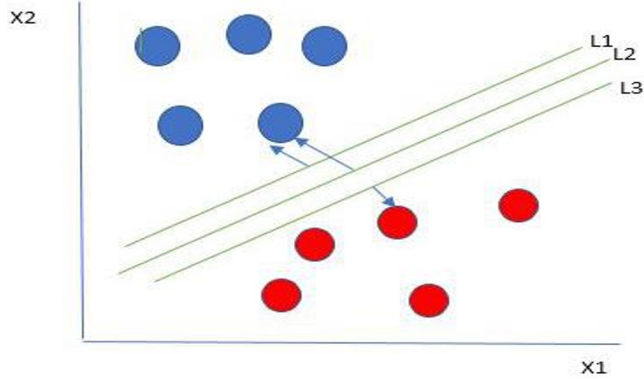


Fig.5. SVM Graph

Logistic regression(LR) is a method used to predict the output when the dependent variable is categorical. It is used for classification tasks where the output is discrete or categorical, such as Yes or No, True or False, or 0 or 1 from fig 6. Rather than providing exact numbers provides probabilistic values between 0 and 1. Because it can classify new information from both constant and exclusive sets of data, logistic regression is an important machine learning technique. Logistic regression fits a "S"-shaped logistic function that forecasts two maximum values rather than a straight regression line.

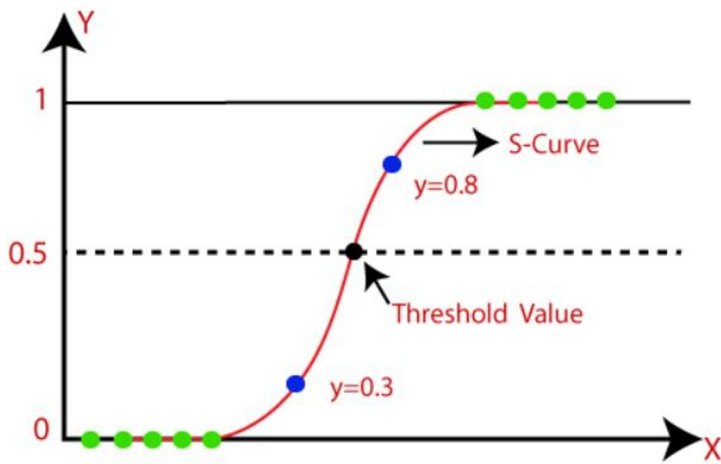


Fig. 6.LR Graphical representation

4 Implementation

The given code reads a file using csv reader and creates a list named "Tweets" which contains the sentiments and tweets for each element in the list. Each row is then split into two halves - the first half contains the sentiment, and next part contains tweet. For example, consider the first row of the dataset where the sentiment is "positive" and the tweet is "@sathwika \I can't afford flight tickets of airlines but I can uy rail ticket.#sad".

The "extract features" operation assigns a true or false value to each word in the feature list for each class separately. For each word in the phrase, the probability is calculated using this list. A term is marked as true for the positive class if it appears in any phrase of the positive class; if it does not, it is marked as false. This process is repeated for all classes. The output of this process is called "Features".

The apply feature function from NLTK (Natural Language Processing Tool) is made use to create Training set from Features.

The function train() in the NLTK library's of Naive Bayes Classifier class uses the training set produced in Fig. 7 as a parameter and outputs a trained classifier. The train() method from the sklearn collection of several machine learning classes is used to generate the other six classifiers Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, Stochastic Gradient Descent, Support vector machine and Linear Support vector machine)

As you can observe fig 7, trained classifiers are stored in the system to avoid training classifier every time. Whenever it is required to classify the tweet or sentence then the classifier is loaded from pickle and used to give the output without training once again.

```
print("X_train_shape : ",X_train.shape)
print("X_test_shape : ",X_test.shape)
print("y_train_shape : ",y_train.shape)
print("y_test_shape : ",y_test.shape)
```

```
X_train_shape : (1092, 16252)
X_test_shape : (274, 16252)
y_train_shape : (1092,)
y_test_shape : (274,)
```

Fig. 7. train() function

Finally, we have used voting classifier as our main algorithm for efficient classification of tweet and to incorporate in our website, as you can observe in the fig 8 that we got approximately 75% efficiency.

```
from sklearn.ensemble import VotingClassifier

#create a dictionary of base learners
estimators=[('rfc', clf), ('NB', model_naive)]
#create voting classifier
majority_voting = VotingClassifier(estimators, voting='soft')

#fit model to training data
majority_voting.fit(X_train, y_train)
#test our model on the test data
vacc=majority_voting.score(X_test, y_test)
print(vacc)
```

```
0.7554744525547445
```

Fig.8.VC Efficiency value

Finally, A functioning model that we implemented after incorporating the trained algorithm to our website fig 9.After logging in, the user can access the complaint page where they can add details about the complaint they have faced. This information includes the category of the complaint, the description of the complaint, and the railway zone where the incident occurred (fig 10). Once the complaint is submitted, it is processed using the sentiment analysis algorithm and the corresponding railway zone is notified about the complaint. The user can also view the status of their complaint in the complaint status page.

The screenshot shows a web application titled "Railway Feedback Using Machine Learning". On the left is a dark blue sidebar with a "Railway Feedback" header and navigation links: Home, Admin, User, Complaints, and Analyst. Below these are social media icons for Facebook, Twitter, LinkedIn, and YouTube. The main content area is light blue and contains a "user Register form". The form fields are: User Name, Password, Email, Select Gender (a dropdown menu), Dob (a date picker showing "dd-mm-yyyy"), Mobile, and Address. A "Register" button is at the bottom of the form.

Fig.9.webpage of this project

The screenshot shows the same web application as Fig.9, but on the "Add post form" page. The sidebar is identical. The main content area has a title "Add post form" and two input fields: "Name" and "post". Below these fields is an "Add" button. There are also two small square buttons, one blue and one dark blue, to the right of the "Add" button.

Fig.10.complaint page

5 Conclusion and future scope

As Artificial Intelligence continues to advance, it is likely that AI Assistants will soon be capable of performing most human tasks with greater efficiency and accuracy. In our own project, we have employed a machine learning model to classify posts into two categories: Feedback and Emergency. By modelling a website, users are able to log in and directly report any railway-related issues, receiving a real-time response. By reducing the need for manual work, complaints can be resolved more quickly, and actions can be taken faster. The implementation of this project has the potential to save time and resources, replace traditional stationery materials with electronic devices, and improve the overall efficiency of the complaint resolution process. we can integrate it with chatbots in the future, Chatbots can be employed to reply to customer concerns and offer immediate solutions. With the help of Natural Language Processing (NLP), chatbots can understand the user's message and provide a relevant response. Another potential scope for this project is to provide multilingual support. With the railway industry being a global sector, users from different parts of the world can have complaints in different languages. Thus, providing multilingual

support can help cater to a wider audience. Finally, we have gained valuable knowledge about complaint management within the railway system throughout the course of this project. Ultimately, we believe that this project will provide efficient customer service to users.

REFERENCES

1. Kumar, Sachin & Nezhurina, Marina. (2020). Sentiment Analysis on Tweets for Trains Using Machine Learning.
2. Shrivastva, Chitresh. (2017). Indian Railway and Social Media – The way forward to improving Business to Consumer Interaction.
3. Bac le and Huy Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques", Advanced Computational Methods for Knowledge Engineering, Advances in Intelligent Systems and Computing 358, DOI: 10.1007/978-3-319-17996-4_25
4. Sisodia, D.S., Nikhil, S., Kiran, G.S., Shrawgi, H. (2020). Performance Evaluation of Learners for Analyzing the Hotel Customer Sentiments Based on Text Reviews. In: Pant, M., Sharma, T., Basterrech, S., Banerjee, C. (eds) Performance Management of Integrated Systems and its Applications in Software Engineering. Asset Analytics. Springer, Singapore. https://doi.org/10.1007/978-981-13-8253-6_20
5. P. Mishra, S. A. Patil, U. Shehroj, P. Aniyeri and T. A. Khan, "Twitter Sentiment Analysis using Naive Bayes Algorithm," 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2022, pp. 1-5, doi: 10.1109/IISEC56263.2022.9998252.
6. Rutuja Samant, Gaurang Yadav, Diksha Poojary, Guruprasad Tandlekar, Manisha Ahirrao "Indian Railways Tweets Classification System using Naive Bayes Classifier," April 2022| IJIRT | Volume 8 Issue 11 | ISSN: 2349-6002.
7. M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, 2013, pp. 1-5, doi: 10.1109/ICCCNT.2013.6726818.
8. Jamwal, G. Vasukidevi, T. N. Malleswari, T. Vijayakumar, L. C. S. Reddy and A. S. A. L. G. G. Gupta, "Real Time Conversion of American Sign Language to text with Emotion using Machine Learning," 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Dharan, Nepal, 2022, pp. 603-609, doi: 10.1109/I-SMAC55078.2022.9987362.
9. Mithra, S., Nagamalleswari, T. An analysis of deep learning models for dry land farming applications. Appl Geomat (2022). <https://doi.org/10.1007/s12518-022-00425-3>
10. T. S. Roy, N. Namratha and T. Y. J. Naga Malleswari, "Voice E-Mail Synced with Gmail for Visually Impaired," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 802-807, doi: 10.1109/ICAIS56108.2023.10073879.