NLP based Grievance Redressal System

Alok Pratap Singh
Computer Science & Engineering Department
Meerut Institute of Engineering and Technology Meerut, India

Ankur Goel
Computer Science & Engineering Department
Meerut Institute of Engineering and Technology
Meerut, India

Aakansha Goel
Computer Science & Engineering Department
Meerut Institute of Engineering and Technology Meerut, India

Diksha Arya
Computer Science & Engineering Department
Meerut Institute of Engineering and Technology Meerut, India

ABSTRACT
Internet is almost accessed by every individual and for expressing themselves and their thinking about the Politics, Country, Sports, and various other topics. Analyzing these trends of the public, can yield various result for variety of purposes. Social Media platforms are also used by several government ministries, mostly Twitter, as its main purpose is data sharing and complaint accumulation. By this, one can collect various data, sentiments, knowledge, and requirements of citizens by applying analyzing citizen sourcing ideas to provide better public service. It is hard to search for the complaint tweets, as these tweets have high velocity and are unstructured in nature. The study provides a framework that helps the Railway Ministry to classify the tweets into complaints/suggestions and compliments. The research shows the usage of Natural Language Processing (NLP) and sentiment analysis for the classification of tweets as the data set is written as general spoken language. The accuracy of the framework is 95.8%.

Keywords
Natural Language Processing, Sentiment Analysis, Twitter Analysis, Naïve Bayes, Decision Tree, Random Forest, Correlation and Regression

1. INTRODUCTION
Government agencies use social media platforms to allow public involvement. Twitter is most widely used platform for this communication medium. Government to Citizen (G2C) can help in improvising the government services by increasing their availability [1]. The Indian government runs many accounts on Twitter to receive a variety of complaints, opinions, suggestions, and grievances for various departments like UP Police, Delhi Police, Rail Ministry of India, Income Tax Department.

The paper focuses on the identification and analysis of tweets related to railway complaints. As the usage of Railway Twitter service has increased, it has become complex for Indian railways to comprehend the working of Twitter services in real-time [8]. Analyses of railway twitter service tweets and the citizen's reactions on train disruptions are done. The results provide insights in Indian Railways by showcasing their views and reaction to various events and sharing data of Complaint Tweets with Railway on Twitter.

Indian Railways uses an account named @RailMInIndia to accumulate all the complaint tweets from the public. These Citizens are diverted to the respective department for further questioning if needed. The tweet can be a complaint, suggestion or an Appreciation. The various complaints and suggestions tweets can be related to a many topics such as unhygienic environment, bad ventilation services, poor food etc. It is very hard and slow process to manually go through all the tweets and categorize them as the number of tweets is very high.

This paper represents tweets classification using various algorithms like – Naïve Bias and Decision Tree. Also, the techniques like – Support Vector Machine (SVM), Random Forest, Regression, and Correlation are used. The presented work is different from the initials works under the same problem. Each tweet is identified and analyzed individually. The work comprises of the construction of various non-textual elements to elaborate and analyze the complaint tweets.

2. RELATED WORK
Many researchers have worked on Sentiment Analysis of social media and this paper is inspired from these researches. Here is a brief review of the papers that are referred. Omar Adwan et al, [4], have given an approach based on machine learning by using SVM to classify the tweets based on thoughts and opinions in multiple domains. He has reviewed many other approaches such as Lexicon- Based approach, Hybrid- Based approach and various Graph-Based approaches. Sanjay Rai et al, [5], have proposed an idea for categorization of twitter data in positive and negative categories. He has briefed a hypothesis review cycle to rapidly order unstructured twitter data. His research gathers information present online and order emotions accordingly. Sachin Kumar et al, [3], have used various approaches to evaluate the hidden.

Sentiments from tweets data using machine learning techniques like Back Propagation Artificial Neural Network (BPANN) and Random Forest. According to his study BPANN provided more accuracy with high training on the data. Nadeem Akhtar et al, [8], have used social network graph to model complaint identification problem. He has used Graph Filtering and Complaint Sub graph Selection approach to identify whether the tweets are suggestive or complaint in nature. The accuracy achieved by this approach for complaints and suggestions is 92% and 70% respectively.

3. METHADOLOGY
This section provides a complete description and working of the proposed framework. Tweet's classification has become the latest approach by many sectors for gathering the complaints and analyzing the latest trends of public. This paper proposed an approach that classifies the tweets into positive and negative; and the negative tweets are further classified into food problems, hygiene problems and train late problems using various Machine learning techniques. The overall working is shown in Fig 1.

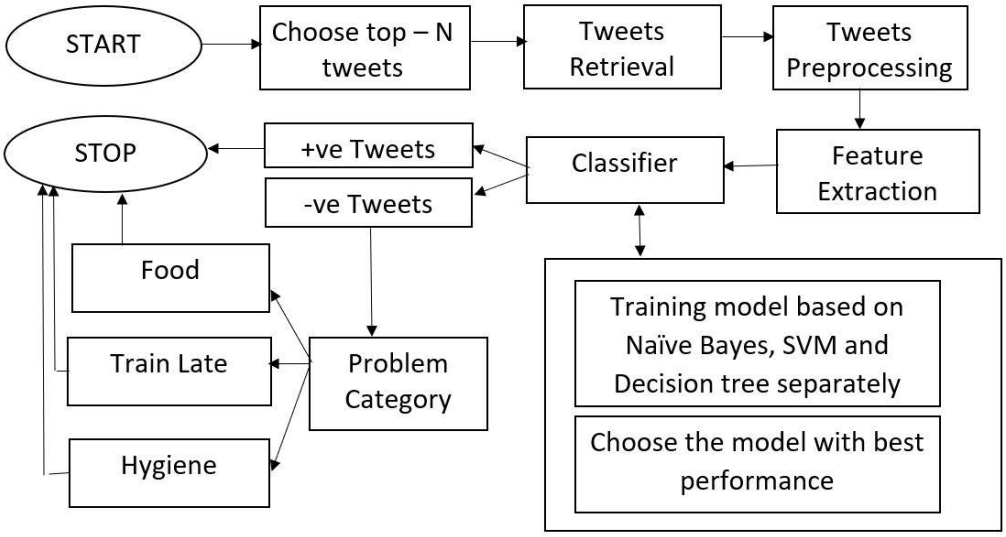

Fig 1: Flow Chart

3.1 Data Set

The dataset used for this framework is tweets as shown in Fig 2. The library used for collecting the data is Tweepy. Twitter provides its python library (Tweepy) which helps in extracting the tweets from the platform in real time. It uses some selected keywords to select the tweets and download them. To extract the tweets from twitter access key, access secret key, consumer key and consumer secret are required for authentication. These keys are easily available for all the users.

Steps for obtaining keys:

1. Login to twitter developer section
2. Go to "Create an App"
3. Fill the required data in the application.
4. Click "Create your Twitter Application"
5. The consumer key and consumer secret will be provided along with the details of the application.
6. For access token, click "Create my access token"

@GM_CRly
Sir,even after repeated requests serious attention has not been paid to keep compartments of the PA-LNU EMU rakes clean. In the skeleton services too compartments appear to be filled with dust. Why should the commuters travel in unhealthy conditions ater paying fare?

Fig 2: Sample Data

3.2 Data Description

The tweets are downloaded using Tweepy and are transmitted in JavaScript Object Notation (JSON) file format. These tweets are selected on the basis of some chosen tags that can be changed by the admin only. Python 3.6 is used for all the programming. Only the latest tweets are selected, and the number is chosen by the user. Some tags used for selection in this project are @CentralRailway, @RailMinistryIndia, @PiyushGoyal, @PiyushGoyalOffc , @SCRailwayIndia . A total of 100 tweets are sent to the server for analysis.

3.3 Data Preprocessing

JSON format is used to issue the tweets. Hence the tweets have been extracted and converted to Comma Separated Values (CSV) Format as shown in Fig 3. For Database Management a class label is provided to each tweet and MySQL is used for maintenance of database. The name of labels provided is positive or negative. Additionally, the cleaning and sorting of data is done.

1,2022-01-26 11:27:06+00:00,1486299654161858560,"@RailwaySEVA SUR/Madam,
PNR No 8747305430, train no 11039. Train is currentl...https://t.co/4dvnRUuWp1"
2,2022-01-21 06:01:10+00:00,1486217630939639809,@GM_CRly @drmpune Thanks
The train is standing at Miraj Junction from last 40-45 Minutes.
3,2022-01-25 20:38:05+00:00,1486075927230443522, @RailwaySeva @drumpune
HAPPY REPUBLIC DAY Madam/Sir https://t.co/eoeMvFTuw7"

Fig 3: CSV File

3.4 Feature Extraction

The pre-processed data achieved cannot be used directly for classification. Classification techniques use certain aspects to determine the polarity of a sentence which defines the opinion of the public. These aspects are extracted from the pre-processed data and are as follows-

1. **Syntax** - The syntax of the sentence can be used to determine the overall subjectivity pattern.
2. **Position of a Term** - The sentiment of a sentence can be changed by changing the position of some specific terms.
3. **Negation** - If a negation is present in the sentence then it changes the polarity of the opinion.
4. **Parts of Speech** - The parts of speech can usually indicate the sentiments of the opinion.

3.5 Classification

The aspects gained from feature extraction are subjected to various machine learning techniques so that the final sentiment of the tweet can be recognized [5]. And the tweets are then classified in respective classes (positive and negative) using Naïve Bayes, Regression and Correlation. Further the negative tweets are classified into sub-classes(Food problems, Hygiene problems, Train late problems) using Decision Tree, Random Forest and SVM.

4. MACHINE LEARNING TECHNIQUES

Naive Bayes: It is a Machine learning technique mostly used for data collection. It has simple probabilistic classifiers and coupled with kernel density estimation [6]. Mathematically, this theorem finds the probability of event of occurring and gives the probability of another event that has occurred. This algorithm is used to check the sentiment of a particular tweet. It considers a dataset that describes whether the railways tweet condition is positive or negative, each classifies the conditions as (Appreciation) or (Complaint). In this "scikit-learn" library for Gaussian Naive Bayes classifier is used.

Regression & Correlation: It is a set of statistical process and is used to find the relationship between independent variables & dependent variable. The best relation between these variables is known as the best fit or curve fitting. If the relationship between the dependent variable and independent variables is linear then, it is known as Linear Regression. It is used to identify and analyze the similarity between the old and new tweets. For that purpose, a relationship is established between them. The relation established is done using correlation and the value of the relation is predicted using regression. This is mainly used for the classification of tweets into positive and negative alongside Naïve Bayes algorithm.

SVM: In this research SVM is used for classification. It creates hyper plane in a multi-dimensional space which is used to classify the data points [4]. The quantity of features used defines the dimensions of the hyper plane. E.g., If there are two input features then a line is formed as a hyper plane, for three input features hyper plane becomes two-dimensional. It is difficult to imagine the plane if the number of features exceed three. SVM is used asa secondary method for random forest to classify the negative tweets into sub-classes. It is used to provide an extensive training on data.

Decision Tree: It is supervised learning technique which is used for the classification of the dataset [6]. Decision Nodes and Leaf Nodes are the two types of nodes that are used in Decision tree [9]. It branches the long tweets into words and then analyze each single word as shown in Fig 4. Then moving from bottom to top system analyzes the word to be in one of the three sub-classes. By this the whole tweet is categorized in a sub-class. Pseudo Code 1 represent decision tree.

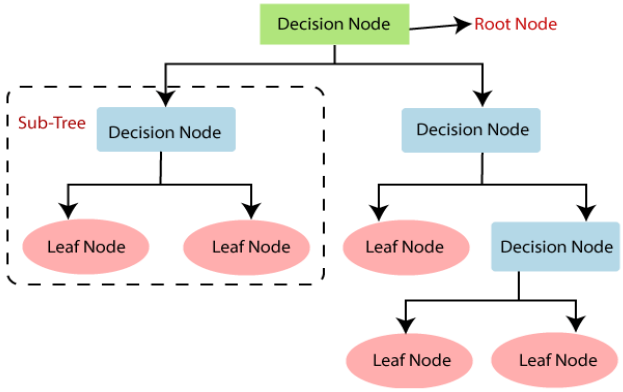

Fig 4: Decision Tree

Algorithm : Generating a decision tree from data partition (D).

INPUT :

1. Training data (D) with the class labels
2. List of attributes and the set of candidate attributes

OUTPUT: Final decision tree.

Method:

1. create a node N
2. if data in D are of same class, then
   return N as leaf
3. if attribute_list is empty then,
   return N as leaf for major class
4. apply selection_methd for splitting_attribures acc.
5. label node N with splitting_criterion
6. if splitting_attricute is discrete-valued and multiway splits allowed then
7. attribute_list <- attribute_list – splitiing_attribute
8. for result I of splitting_criterion
9. let D(i) be the data set in D satisfying result i
10. if D(i) is empty then,
    attach a leaf label with majority class in D
11. else
    attach the node returned by Generate_decision_tree
12. endfor
13. return N.

Pseudo Code 1: Decision Tree Algorithm

Random Forest: It is used for the classification of numerous decision trees on various datasets and then finding the average to precise the output prediction [7]. In this large number of decision trees that are formed are considered as forest and are used for the classification of tweets [10]. It is used to make the output more accurate as represented in Equation 1.

More number of Decision Trees = More Precise Output ~(Eq – 1)

5. RESULT

This section discusses the final result presented by the framework. Several python libraries are used for the implementation purpose; some of them are Tweepy, Pandas, Textblob, JSON and Natural Language Toolkit. Only those tweets are selected in which these personals are tagged, and the ratio of tweets is shown in Figure 5 and its categorization in Fig 6.

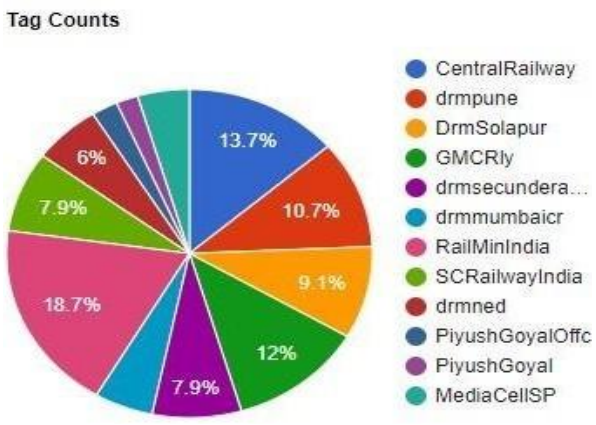

Fig 5: List of Tags

These tweets are classified into positive and negative tweets and the negative tweets are further classified into multiple sub-classes. These classifications are done with the help of multiple machine learning techniques mentioned above in the paper.

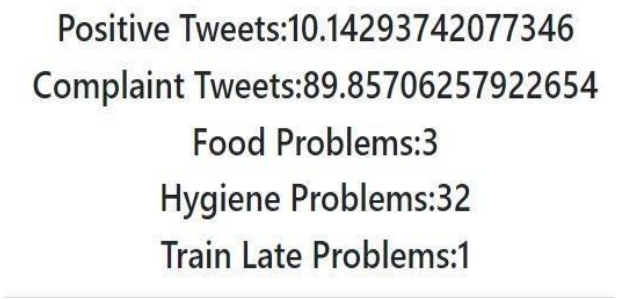

Fig 6: No. of Classified Tweets

First Confusion Matrix – This confusion matrix represents the result of Naïve Bayes, Regression and Correlation algorithms. The accuracy calculated from the matrix of Table 1 comes out to be 95% as shown in Equation 2.

Accuracy = ((TP + TN)*100)/100
Accuracy = ((8+87)*100)/100
Accuracy = 95% ~ (Eq -2)

Table 1: First Confusion Matrix

Input Tweets(N=100)	PositiveTweets	NegativeTweets
Positive Tweets	8	3
NegativeTweets	2	87

Second Confusion Matrix – This confusion matrix of Table 2 represents the result of Random Forest and SVM. The accuracycalculated from the matrix comes out to be 96.6% as shown in Equation 3.

Accuracy = ((TP + TN)*100)/100
Accuracy = ((2+31+1+53)*100)/100
Accuracy = 96.6% ~ (Eq - 3)

Table 2: Second Confusion Matrix

Input Tweets N=90	Food	Hygiene	Late	Rest
Food	2	1	0	0
Hygiene	1	31	0	1
Train Late	0	0	1	0
Rest	0	0	0	53

These results clearly show that the framework provides more accuracy than any other framework that was used before. The average final accuracy is 95.8%. The final accuracy is the result of both the accuracies provided above.

6. CONCLUSION

Various machine learning approaches are used to classify and categorize tweets on the basis of positive and negative using Naïve Bayes, Regression and Correlation, and further classified into sub-classes i.e. (food problems, hygiene problems, train late problems) using Decision Tree, Random Forest and SVM. Considerable work has been done in the field of sentiment analysis using various machine learning techniques. The research shows the usage of NLP and sentiment analysis for the classification of tweets as the data set is written as general spoken language. This research focuses on increasing the accuracy of tweets classification as multiple algorithms are used to train dataset. All these algorithms are used to increase the accuracy of their corresponding work in training the dataset. The final accuracy comes out to be 95.8% which is the average accuracy of both the confusion matrices.

6.1 Future Scope

- Integration of PNR detection can be done in future as it will be helpful to railways in gaining information about the problem of a specific train.
- This process can be automated to show the results in an application.
- A mobile based application can be developed for this process.
- This system can be calibrated to detect tweets of different languages.
- This system can be designed to detect anti-national tweets and collect the data of the accounts that are being used for these purposes.

7. REFERENCES

[1] Lee, G., & Kwak, Y. H. (2012). **"An open government maturity model for social media- based public engagement"**. Government Information Quarterly, 29(4), 492-503

[2] Mukta Goyal, Namita Gupta, Ajay Jain & Deepa Kumari (2020). **"Smart Government E- Services for Indian Railways Using Twitter"**, Micro-Electronics and Telecommunication Engineering (pp.721-731), DOI:10.1007/978-981-15-2329-8_73

[3] Sachin Kumar & Marina I. Nezhurina (2020). **"Sentiment Analysis on Tweets for Trains Using Machine Learning"**. Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018) (pp.94-104), DOI:10.1007/978-3-030- 17065-3_10

[4] Omar Adwan, Marwan Al-Tawil, Ammar M Huneiti, Rawan Shahin, Abeer Abu Zayed & Razan Al-Dibsi (2020). **"Twitter Sentiment Analysis Approaches"**, International Journal of Emerging Technologies in learning (iJET) 15(15):79, DOI:10.3991/ijet.v15i15.14467

[5] Sanjay Rai, S. B. Goyal & Jugnesh Kumar (2020). **"Sentiment Analysis of Twitter Data"**, International Research Journal on Advance Science Hub, e-ISSN: 2582 -4376

[6] Kadda Zerrouki, Reda Mohamed Hamou & Abdellatif Rahmoun(2020)**. "Sentiment Analysis of Tweets Using Naïve Bayes, KNN, and Decision Tree"**, International

Journal of Organizational and Collective Intelligence 10(4):35-49, DOI:10.4018/IJOCI.2020100103

[7] Rajeev Kumar & Jasandeep Kaur(2020) **"Random Forest-Based Sarcastic Tweet Classification Using Multiple Feature Collection"**, Multimedia Big Data Computing for IoT Applications (pp.131-160), DOI:10.1007/978-981-13-8759-3_5

[8] Nadeem Akhtar & M. M. Sufyan Beg(2021). **"Railway Complaint Tweets Identification. Data Management, Analytics and Innovation"** (pp.195-207), DOI:10.1007/978-981-15-5616- 6_14

[9] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[10] https://www.javatpoint.com/machine-learning-random-forest-algorithm