Indian Railways Tweets Classification System using Naive Bayes Classifier

Rutuja Samant[1], Gaurang Yadav[2,] Diksha Poojary[3], Guruprasad Tandlekar[4] and Manisha Ahirrao[5]
[1,2,3,4] *Student, Department of Electronics and Telecommunication, Rajiv Gandhi Institute of Technology*
[5]*Faculty, Department of Electronics and Telecommunication, Rajiv Gandhi Institute of Technology*

***Abstract*—With the advent of social media like Twitter, Facebook, Instagram it has been easier for people to voice their opinions, thoughts to a wider audience in a concise manner. Twitter has been increasingly used by organizations to address customer reviews, complaints and feedback. Many users resort to using social media for immediate response to their queries. But due to exceedingly large amount of data it becomes difficult for organizations to respond to every customer query or complaint. Some of these complaints might be urgent queries demanding immediate attention which due to vast data would be responded late. The proposed system uses Machine learning algorithm-Naïve Bayes to classify the tweets based on their urgency in order to respond to emergency tweets on a priority.**

***Index Terms*—Machine learning, Naïve Bayes Classifier, Twitter API.**

I. INTRODUCTION

The past few years has witnessed increasing number of social media users with about 3 billion users worldwide. Due to social media, connectivity has been ameliorated to a huge extent. Social media has a profound effect on how people communicate over the world. Users resort to social media since it has made communication fast-paced, immediate and brief. Not only individual users but also many local and national authorities resort to social media platforms for the same advantages. These organizations use social media to respond to customer users in a quick manner. These authorities use social media for complaint resolution and feedback. There has been a huge surge in the number of complaint tweets received by Indian Railways Twitter (RailwaySeva) in the past few years. On a daily basis their Twitter account receives thousands of tweets tagging their account and requesting resolution. Now from these thousands of tweets there are many which require attention on a priority i.e., tweets informing about accidents, happenings, threats and medical emergency. Some tweets received by Indian Railways would be about feedback on service, general queries, or reviews which don't require an immediate response and can be dealt with on a low priority. The proposed system classifies tweets using Naïve Bayes classifier. Naïve Bayes classifiers are classifiers which make use of probability and apply the Bayes theorem in order to classify them.

II. REQUIREMENTS AND TECHNOLOGY USED

A. Requirement

The most important requirement is to obtain access to Twitter streaming data. So as to induce access to Twitter data, Twitter API credentials are required. With these credentials we are able to use Twitter API to stream real time tweets. Dataset is another crucial requirement for any processing system. Most typically a knowledge set corresponds to the contents of one database table, or one statistical data matrix, where every column of the table represents a specific variable, and every row corresponds to a given member of the info set in question. Twitter account is required by the user to post a complaint which the Indian Railway will retort to.

B. Technology used

1. Apache Kafka

Apache Kafka is a dispersed event store and stream-processing platform. It is an open-source system which uses zookeeper to manage configuration and implement various protocols on Hadoop clusters. Proposed system uses Apache Kafka to process big data from twitter efficiently. It also a delivers a distributed queuing service.

2. Apache Spark

Apache Spark is considered to be a system that supports stream processing for the processing of real-

time data. These processing systems provide the choice for continuous computations, as data is continuously flowing through them. Instances of such functionalitiesare the regular polishing and accumulation of the incoming data before storage. Apache Spark has two ways to work with streaming data; one being the structured streaming method and the other non-structured streaming method. The structured streaming method is preferred for streaming data real-time.Apache Spark supports Machine Learning libraries.

3. Apache ZooKeeper

Apache ZooKeeper is server used for various cloud applications. ZooKeeper is actually a service for distributed systems offering a hierarchical key-value store, which is employed to produce a distributed configuration service, synchronization service, and naming registry for giant distributed systems.

III. METHODOLOGY

Training Dataset-

Machine learning models are heavily dependent on data. If not provided with an exceptional training data, the model will not produce accurate results. Many machine learning models can fail if they are trained on meagre datasets in the beginning phases.

The dataset will include classified tweets to be fed to the Machine Learning Algorithm (Naïve Bayes).

	A	B
19	0	pnr 4512791357 mosquitoes flis hoverng , pls do some cleanup, tt not responding positively
20	0	torn seat pnr 4512791357 filthy environment need some action to be taken
21	0	copassenger playing loud music , not obeying the tt too, sort it out pnr 4512791357
22	0	tt declining to accept the college id a true id , higher authorities do act pnr 4512791357
23	0	found a fun, we passengers couldn't differentiate b/w a real and a toy pnr 4512791357
24	0	pantry officials billing us double than printed . All complaints in vain pnr 4512791357
25	0	many passengrs pnr 4512791357 feeling stomach upset after dinner frm railways pantry
26	0	train diverted from its actual route , no knowhwdge of where the fuck are we being taken.track us pnr 4512791357
27	0	train is at halt for the last 4.5 hrs amidst of jungle, wt the fkk is happening??? pnr 4512791357
28	0	pregnant lady needs the help of some lady doctor immediately, her pnr 4512791357
29	0	charger points of the complete bogie malfunctioning, need immediate attention pnr 4512791357
30	0	a sweeper supposed to be thief but the matter is out of the passengers cntrl pnr 4512791357
31	0	window pane jammed #utter winter, bone freezing cold pnr 4512791357
32	0	fan's speed not decreasing , we need help pnr 4512791357
33	0	acs not working , pnr 4512791357, tt not responding positivelt, what do we pay for
34	0	berth env is qute unhygenic, insects wandering here n there pnr 4512791357
35	0	unremoved heavy bedrolls , stinky smell pnr 4512791357, emergency cleanup reqd
36	0	two copassengers fought brutally, 1 got a head injury.emergency!! His pnr 4512791357
37	1	@SBKULAL @Bkugne @riyaj_ali @niraj1712 @FAZALALAM234 @phanipeddapalli @ratneshthakur86 @railmitraa @drmned @drmgt6â€¦ https://t.co/ZisjnoOQLd
38	0	@drmkx25 @RailMinIndia @RailwayNorthern Kindly send anyone railway staff to attend 14203
39	0	@RailMinIndia plz continue train no 14307-14308
40	1	RT @IndiaBTL: Why is @RailMinIndia procuring Congress mouthpiece National Herald (priced a hefty Rs. 20/piece) & providing it toâ€¦
41	0	@RailMinIndia Dear Sir, my father is a senior citizen and is a patient too.Travelling alone and his seat is not confirmed but is having RAC.
42	1	@NitinSubudhi @RailMinIndia @PiyushGoyal 4 paise legally sahi daam pe bech ke bhi kamay jaa sakte hai.
43	1	@RailMinIndia @PiyushGoyal If a person buys ticket online and if its not confirmed then why that ticket is not a vâ€¦ https://t.co/fgalcVcM6C
44	1	@RailMinIndia @drmnfr_lmg @IR_ENHM till now no action has been taken

Fig.1 Training Dataset

Parsing of raw tweets-

Parsingis the process of analyzing a string of symbols and adhering them to the rules of grammar. Preprocessing helps reduce the complexity of classification or any type of data handling. In this process the Tweets are formatted to remove unnecessary URLs, special symbols, usernames, emoticons, Hashtags and blank spaces. This ensures that the Tweets are ready to be classified based on their sentiments alone.

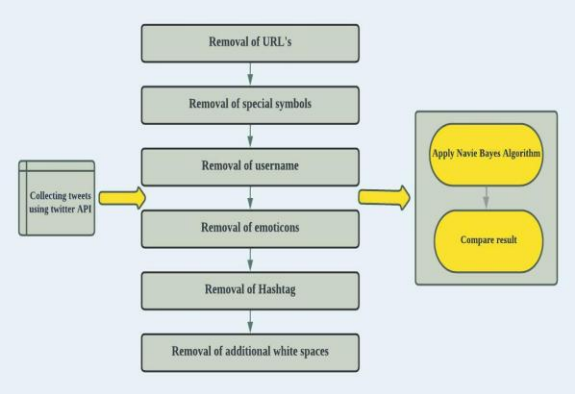

Fig.2 Parsing of Tweets

TF-IDF-

Term frequency is the number of times a term occurs in a document.

Inverse Document frequency is a measure of how common a word is.

The TF-IDF measure is simply the product of TF and IDF.[1]

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

Naïve Bayes classifier-

After data is parsed URL, special symbols, user name and hashtag get eliminated. This data is now ready for classification by the Naïve Bayes Classifier which classifies it into two categories-Emergency and Feedback. In the Naive Bayes each word is contributing to the sentiment, which can be calculated by the ratio of the probability of occurrence of the word for Emergency and Feedback class. The equation used for classification is[2]:

$$P(c|x) = \frac{P(x|c) * P(c)}{P(x)}$$

Where P(c|x) is posterior probability,

P(x|c) is the likelihood,

P(c) is the prior probability of a class

P(x) is the prior probability of predictor

Using this formula individual word's probability is determined for their appearance in category 1 (Emergency) or category 2 (Feedback). Sentiment analysis involves extracting subjective emotions and feelings from text. The primary goal of sentiment analysis is to determine if a text expresses negative or

positive feelings. Naive Bayes is a desired algorithm for classifying text. It is preferred for solving multi-class prediction problems. Naive Bayes is preferred for categorical input variables than numerical variables. Naive Bayes is super-fast learning classifier which makes it suitable to be used for making predictions in real time. Naive Bayes classifier is applied successfully for various applications likeclassification of mails, text classification and sentiment analysisWhen used for textual data analysis itproduces good results. As it has the assumption of the "Naive" features it performs much better than other algorithms.

ER Diagram-

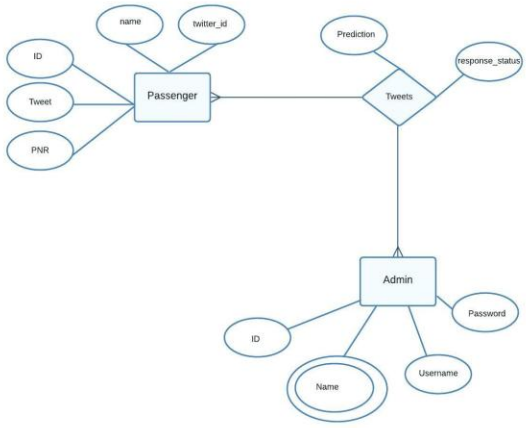

Fig.3 ER Diagram

ER diagram is used to represent relationship between entities.Here the diagram represents two entities Passenger and Admin both of these are allowed to tweet. Passenger and Admin have many attributes like ID, Name, Username etcetera.

Level 0 Data Flow diagram-

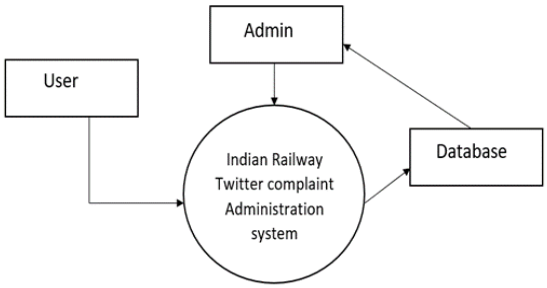

Fig.4 Level 0 DFD

Level 0 Data Flow Diagram represents the function of the proposed system in relationship to external entities.

Level-1 Data Flow Diagram

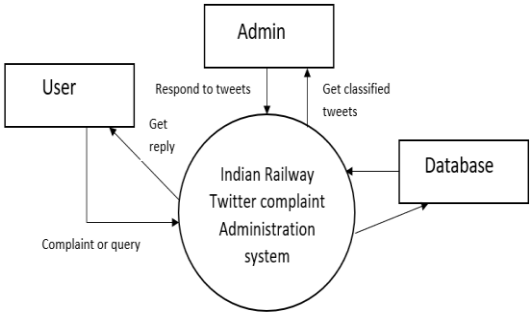

Fig.5 Level 1 DFD

In this Data Flow Diagram, the main functions of the proposed system are highlighted.

Level-2 Data Flow Diagram

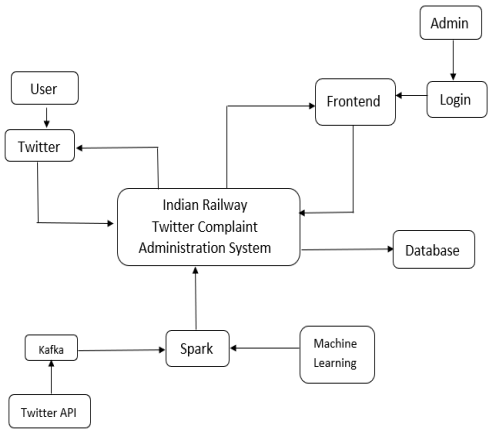

Fig.6 Level 2 DFD

In this Data Flow Diagram more details of the function and entities are displayed. It is used to highlight the working and tools of the process.

User Case Diagram

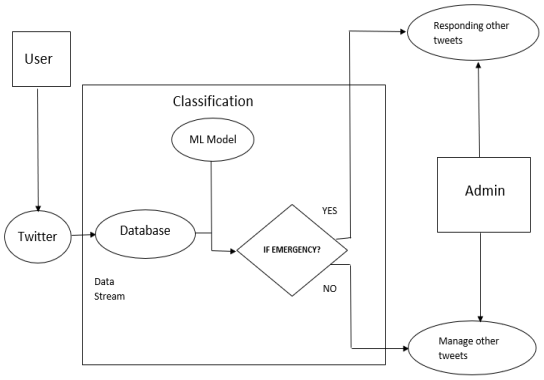

Fig.7 User Case Diagram

This Diagram is used to represent how a user interacts with the system and explains about the process flow. It is used to show the relationship between the user and the different use cases in which the user is involved.

IV. RESULTS

Accuracy of the Classification model

For the proposed system Naïve Bayes Machine Learning Model was selected from a plethora of Machine learning models used for text classification since it provided the highest accuracy, classifying the tweets in the dataset accurately with an efficiency of 85%.

```
In [36]: from sklearn.metrics import accuracy_score

         score_naive = accuracy_score(predicted_naive, y_test)
         print("Accuracy with Naive-bayes: ",score_naive)

         Accuracy with Naive-bayes:  0.85
```

Fig.8 Efficiency of Naïve Bayes

The Real Time Model

The website contains a login page and a homepage. Only Admins with the right credentials are allowed to access the tweet viewing and responding page. Multiple admins can login simultaneously on their systems.

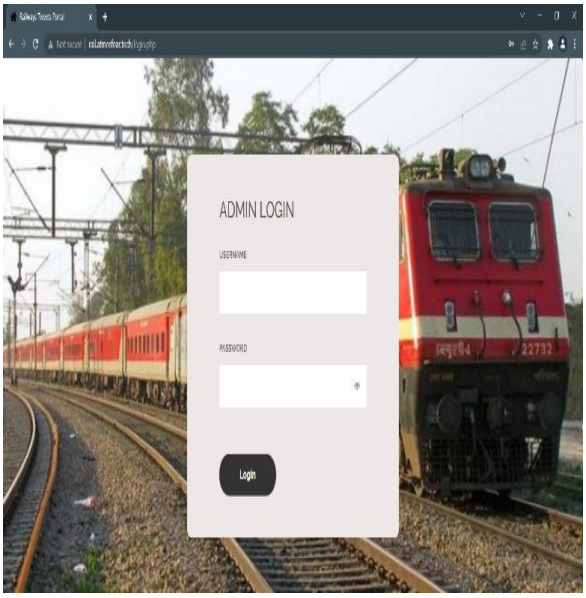

Fig.9 Login Page of the website

The classified tweets are streamed on the front-end (server hosted website). There is a button to switch from emergency tweets to feedback tweets and vice versa. The admin can also reply to these tweets from the website instead of twitter application. The tweets which are inserted into the database from Kafka stream data using spark get updated here within 2-3 seconds of delay. Tweets are live streamed on the front end of the proposed project using Twitter API and Apache Kafka. Kafka is used to buffer the tweets before processing.Tweets are inserted into the database from Kafka stream data using Spark.

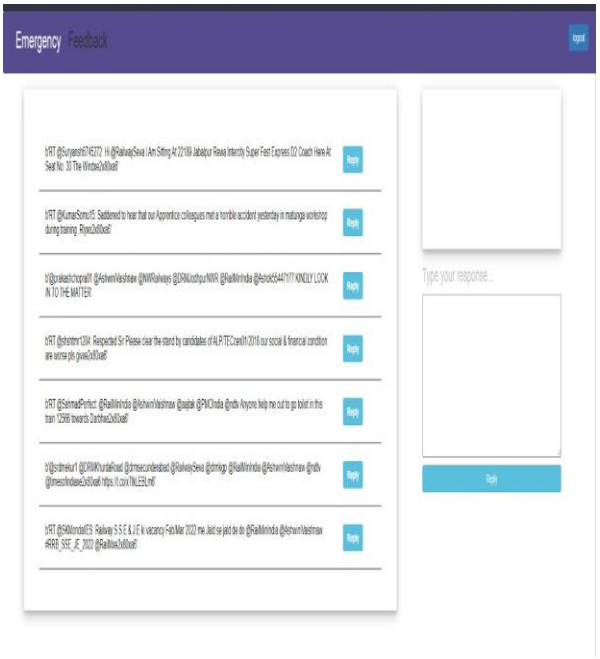

Fig.10 Classified real time tweets on website

IV. CONCLUSION

By implementing this Machine Learning based system instead of the one currently in use it will reduce the manual work and complexity of scanning through a plethora of tweets. It will resolve the issue of unanswered tweets as this system will classify every single tweet into the two categories emergency and feedback. The existing complaint handling system requires plenty of manual labor and hours of scanning through tens of thousands of tweets to find the ones which require immediate response. This 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. The categories for

the tweets to be classified into can be determined by the authority. The accuracy of this system is 85% when trained on a smaller training data set but it is expected to reach 93-95% with the aid of a vast training data set. It is proven that for text classification or sentiment analysis Naïve Bayes algorithm is the fastest, most efficient and reliable.

Since Naïve Bayes calculates conditional probability of occurrence of two events based on probability of occurrence of individual event, it helps generate an output with much higher accuracy than by any other method like linear regression.Similar text classification systems using Linear regression have shown an average accuracy of about 70-73%. This is due to the fact that in Linear regression the predicted value is continuous and not probabilistic. Linear regression is also sensitive to unbalanced data which reduces the overall efficiency of classification. Considering all the limitations of other Machine learning model Naïve Bayes is the most efficient model.

APPENDIX

- DFD – Data Flow Diagram
- API- application programming interface
- URL- Uniform Resource Locator
- TF-IDF-Term frequency-inverse document frequency
- ER diagram- entity relationship diagram

REFERENCES

[1] Kim, SW., Gil, JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum. Cent. Comput. Inf. Sci. 9, 30 (2019).

[2] Chen, H., Hu, S., Hua, R. et al. Improved naive Bayes classification algorithm for traffic risk management. EURASIP J. Adv. Signal Process. 2021, 30 (2021)

[3] M. Wadera, M. Mathur and D. K. Vishwakarma, "Sentiment Analysis of Tweets- A Comparison of Classifiers on Live Stream of Twitter," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 968-972, doi: 10.1109/ICICCS48265.2020.9121166.

[4] M. Assefi, E. Behravesh, G. Liu and A. P. Tafti, "Big data machine learning using apache spark MLlib," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3492-3498, doi: 10.1109/BigData.2017.8258338

[5] Wongkar, Meylan & Angdresey, Apriandy. (2019). Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. 1-5. 10.1109/ICIC47613.2019.8985884.