iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.40760

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

Indian Railway Tweet Monitoring System

Sanyam Aware[1], Sarvesh Pathak[2], Satish Shukla[3], Sasha Kumar[4], Prof. Gurunath Waghale[5]

***Abstract:** In today's fast paced and competitive world, India is home to one of the world's largest railway networks. According to statistical facts, approximately 22 million passengers travel via trains every day. A majority of them are on various social networking site including Twitter, adding to this, more than 200 million users are active daily on this application, tweeting more than 500 million tweets every day, out of them, at least 600 are based on the sole topic of complaining about railways, about 5% are suggestions and etc. To help out the Indian Railway Ministry to quickly classify and/or clarify the tweets as soon as possible and in a much more efficient manner, we would provide a solution by helping them classify the types of posts such as if it is a complaint, or just feedback, or some other type. Since almost everyone on this planet wishes to get a swift and quick response for any query they have, this application would allow the whole team behind the scenes to efficiently handle all the tweets, give a response to the posts, and allows them enough time to get working on an issue.*

I. INTRODUCTION

Through this project, we focus to decrease the complex job of having to scan through thousands of useless data just to find some particular information. With this web application, we can easily identify the tweets that need a prompt reply, which in turn also means that priority wise scanning is an advantage of it. In order to provide a solution to this problem, we will be using a machine learning model called "Naive Bayes". It is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. In machine earning, Naïve Bayes classifiers is one of the simple "probabilistic classifiers", which is as the name states based on Bayes' Theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models. Naïve Bayes has been studied extensively and was introduced into the text retrieval community in the early 1960s. It still remains as a popular method for text categorization.

Practically, it help to distinguish between the category of all documents such as if it is related to promotional items, travel, social messages, or even if it is just a spam. This model can also be made flexible and can be applied on various other local authorities like Nagar Nigam, State electricity board, etc. To do that we just have to filter the tweets on the basis of that particular authority. It will reduce the operational time.

A. Background

The Indian Railways currently receives tonnes of tweets in a day. The tweets are continuously checked and the response is provided according to their level of seriousness. Therefore, a life-threatening complaint such as sexual harassment and medical emergencies feature at the top of the list and are dealt with instantly. Complaints featuring the railway management, hygiene or matters that need immediate attention within trains or in stations can be classified as second on the list. The third on this list can be the complaints and requests that are not needed to be attended immediately and can be solved later.

II. LITERATURE SURVEY TABLE

Title of Paper	***Short Description***
1.Micro-Electronics and Telecommunication Engineering - Smart Government E-Services for Indian Railways Using Twitter	*The Paper shows us a methodology of analyzing tweets which is done by first collecting the data, where all the tweets addressed to the Official Railway Ministry of India are collected. Then using the Micro Post Enrichment Algorithm which is used to clean the tweets and ensuring that the tweets are apt for the various classifiers which are to be applied on it. It is a multistep iterative process. It will take a raw tweet as an input and provide a syntactic and semantically enriched tweet. Next come the Feature Extraction step that chooses the top N unigrams and bigrams from frequency distribution of the unigrams and bigrams present in the dataset we also perform a comparison, between Naïve Bayes, SVM, decision tree where we reach at a conclusion that the best accuracy is achieved using Naïve Bayes.*

2. Benefits of AWS in Modern Cloud	***The paper shows us the benefits of using AWS, some of them being:*** ***Data Protection –*** *Data Protection is the major concern of most organizations and applications. The article states various ways in which we can use AWS for data protection and storage and prevent data leakage or any other damage to sensitive data. Some of them being Implementation of data hashing on the device and server, use of remote wipe APIs, etc.* ***Increased Productivity –*** *Cloud has solved most of the time-consumption issues of software installation, maintenance of the product and backing up on a regular basis. Cloud computing provides more flexibility. The users can access files from anywhere at any time using web-enabled devices such as laptops, smartphones, notebooks etc* ***Backup options – With automated backup helps in avoiding manual errors and also helps in time and money and ensures*** *all application data is properly backed up.* ***It speeds up and minimizes the workload in my ways and hence by*** *AWS CodeDeploy, it easier for to rapidly release new features, and avoid downtime during application deployment.*
3. Using TF-IDF to Determine Word Relevance in Document Queries	*This Paper gives us information about the results that are achieved when Term Frequency Inverse Document Frequency (TF-IDF) is applied to a document. It calculates values for every word by an inverse proportion of the frequency of the word in a particular document to the percentage of documents that the word appears in. This paper also provides proof of how simple algorithm can be used to categorize relevant words that can enhance query retrieval. The paper goes ahead to show us how TF-IDF returns documents that are relevant to a particular query. Once a user inputs a query for a particular topic, TF-IDF can find documents that contain relevant information on the query. So we can use it in our project to measure the importance of a word in data.*
4. Understanding Regression Testing Techniques	*The paper informs us about the types of regression testing and also tells us why it is important to include regression testing in order to determine all defects in our project.* *In the maintenance phase of the software development life cycle we need to retest the software for the modifications it underwent and for this purpose we use various types of regression testing techniques. The paper shows us the complete structure of Regression Testing, also further classifying each one of them as explained by various authors, explaining Regression Test Selection and Test Case Prioritization in detail with Search Algorithms for Test Case Prioritization.*
5. A framework for massive Twitter data extraction and analysis	*This study illustrates the capabilities of twitter with two study cases in Spanish, one related to a high impact event and another one related to regular political activity on Twitter. It provides a framework designed mainly to ease the process of DATA EXTRACTION AND ANALYSIS which is followed a sentiment analysis which is divided into three subsystems: Trainer, classifier, and tester and applies in on the respective two case studies. For the analysis it performs both sentiment analysis as well as quantitative analysis. We also understand how supervised classification is very expensive and time consuming.* *We could use it as a tool which could analyse the data in real time and use it to predict trends or mass opinion.*

6. An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation	*This paper is focused on testing the TF-IFD(Term Frequency-Inverse Document Frequency) algorithm, which is a technique to quantify words in a set of documents. A score is computed for each word to signify its importance in the document and corpus. In order to proceed with our project we have to be able to identify the tweets also based on their semantic analysis which in-turn can help classify the tweet.* *So what we would be extracting from this paper is how we can use TF-IDF to our advantage by successfully identifying the importance of the post as quickly as possible and take the appropriate actions.*
7. Response Time analysis for XAMPP Server based on Different Versions of Linux Operating System	*Taking in consideration the technologies we would be using for the implementation of this project, XAMPP is a software distribution which provides the Apache web server, MySQL database, Php and Perl (as command-line executables and Apache modules) all in one package.* *The activity performed in this paper is the calculation of the response time of XAMPP for various LINUX OS. For our project, we are aiming to deploy it on AWS and for that, XAMPP instance will be the perfect solution for an all-in-one software.*
8. Security and Safety in Amazon EC2 Service – A Research on EC2 Service AMIs	*After going through this paper, we learnt about functionalities of the AWS EC2 Service, including what and how the safety and security are kept in this system. We also learnt about how we could use 3 EC2 systems for our project by keeping 1 Master and 2 Slave machines.* *This paper does not contain a research gap as this reference paper will be used solely to grasp the features of the service.*
9. A Review Study of Apache Spark in Big Data Processing	*This paper sheds light on how Apache Spark is a fast-growing new generation technology in the field of Big Data Analytics and due to its features such as lightning-fast cluster computing which is specifically designed for fast computation. It also compares between spark and hadoop, which has led us to believe that each has its own advantages, but since we would like our application to have a high processing speed we would go with Apache Spark.*
10. Research Paper on AWS Cloud Infrastructure vs Traditional On-Premise	*This research paper is solely based on the advantages of AWS Cloud such as, avoiding direct capital expense for variable capital expense, less complication to run and maintain data centres, an upper hand for global deployment, and many more. So, for us to build a financially feasible web application we can use this service provided by Amazon for an efficient and flexible working. The author of this paper has also conducted surveys on the awareness of people about cloud computing, AWS and a few other factors. This survey helped us even more in determining the best technology for our project.*
11. Learning Applications Exploring the Cost-benefit of AWS EC2 GPU Instances for Deep Learning Applications	*This paper presents the implications, in terms of runtime and cost, of running two different deep learning problems on AWS GPU-based instances, and it proposes a methodology, based on the previous study cases, that analyses instances for deep learning algorithms by using the information provided by the Keras framework. Our experimental results indicate that, despite having a higher price per hour, the instances that contain the NVIDIA V100 GPUs (p3) are faster and usually less expensive to use than the instances that contain the NVIDIA K80 GPUs (p2) for the problems we analysed. Also, the results indicate that the performance of both applications did not scale well with the number of GPUs and that increasing the batch size to improve scalability may affect the final model accuracy. Finally, the proposed methodology provides accurate cost and estimated runtime for the tested applications on different AWS instances with a small cost. Hence, on observing the whole experiment it is clear that AWS EC2 instance would be the best choice for our project.*

12. Understanding Users' Satisfaction towards Public Transit System in India: A Case-Study of Mumbai	*In this paper we got to know the quality of public transit system in resource constrained regions using user-generated contents. With growing urban population, it is getting difficult to manage travel demand in an effective way. Due to resource constraints, developing cities have limited infrastructure to monitor transport services. To improve the quality and patronage of public transit system, authorities often use manual travel surveys. But manual surveys often suffer from quality issues. To do this, we assumed that, if a tweet is relevant to public transport system and contains negative sentiment, then that tweet expresses user's dissatisfaction towards the public transport service. From this paper we can learn the thing which we can implement is that we can got to know whether the tweet is a general tweet or a complaint tweet related to cleanliness or other services and take the action accordingly.*
13. Comparison of Flutter with Other Development Platforms	*From this paper we got to know this flutter is a useful toolkit enables easy ways of creating new applications. The basic results in this report indicates flutter has a slight edge as compared to native application development platforms but further more conclusive tests still needs to be carried out to come to a final conclusion.* *Flutter has a great potential in future but further tests are needed to be performed as it is easy to use and getting popular day by day. And so, when we start to implement a mobile application we can use Flutter, as it grants the ability to design mobile applications in both iOS and Android with a single codebase, furthermore it also has the potential to make the app in all respects flexible and fluent.*
14. Kafka: a Distributed Messaging System for Log Processing Jay Kreps LinkedIn Corp.	*From this paper we learnt Kafka is used for processing huge volumes of log data streams.* *Unlike typical messaging systems, a message stored in Kafka doesn't have an explicit message id. Instead, each message is addressed by its logical offset in the log. This avoids the overhead of maintaining auxiliary, seek-intensive random-access index structures that map the message ids to the actual message locations.* *As we are developing a real time application, we will be dealing with heavy streaming data, hence we can use Kafka to stream the tweets.*
15. Comparing Database Management Systems: MySQL, PostgreSQL, SQLite	*This paper introduces the two types of databases, i.e. relational and non-relational. We came to know about three of the databases, which were SQLite, MySQL, and PostgreSQL and also witnessed the advantages and disadvantages of both the database management systems. This also helped us in understanding which database would be much easier to connect and work with the AWS RDS and hence, we came to a conclusion that MySQL will be much more efficient than the other two, so, we will use MySQL for the EC2 Slave servers.*
16. Building a replicated logging system with Apache Kafka	*This paper instructs us about the online data integration task with the help of event pipelining platform used in Apache Kafka which allows to do many tasks like data streaming, stream storage and message queue.* *So in our project we are using the data of twitter for our Indian Railways Tweet monitoring system and, the Apache Kafka provides the real time streaming of data to collect big data to do real time analysis. So , this technology will help us to analyse the tweet which are tweeted by the passengers for any complaint in the service and with the help of Apache Kafka and the system can help the passenger to get proper and good facilities during travelling.*

17. ZOOKEEPER: Wait-free coordination for Internet-Scale Systems	*In this paper we learnt about the Zookeeper which provides a simple and high-performance Kernel for building more complex coordination primitive at the client. It provides guarantee FIFO client ordering. In our project it works based on the timing of the Tweet tweeted by the passengers. Tweet will be checked and implemented in the manner of FIFO which means First in First Out. So, every passenger will get preference with the timing of their tweet. It also uses simple pipelined architecture that allows us to have hundreds or thousands of requests outstanding while still achieving low latency which naturally enables the execution of operation from a single client in FIFO order.*
18. A Collaborative Filtering Approach Based on Naïve Bayes Classifier	*This paper shows us about the naïve Bayes classification Algorithm which uses the application of Bayes rule . Bayes rule states about the use of Venn diagram to classify the two types of things and then use of the part which we need in our technology. The Naïve Bayes Algorithm uses the same theory but it is used to differentiate the two types of messages which are coming to the system.* *In our project we are using this algorithm to classify the tweets coming from the passengers of the Indian Railways for any negative complaint or positive review so that we can observe all the negative complaint and provide proper solution to that complaint.*
19. Measuring user influence on Twitter: A survey Using Twitter API	*In this paper we have seen that the Twitter API works to analyse all types of text such as short, long, use of bad words, misspelling and emotions are analysed. It also analyses the real time information from twitter or any social media platform. So, using this information and almost simultaneously observing other social networking applications, we were able to analyse the amount of impact tweets have throughout the whole social media.* *So, for our project we came to a decision to use Twitter, above other application because as we observed the influence of a tweet in the world of internet, there are a number of tweets that broke the internet and was instantly the most viral topic around the whole world. Apart from this, the Indian Railways doesn't have an official account on Instagram, they do have it on Facebook, however, if both the applications are to be compared we could process the posts from Twitter much fluently and take the necessary actions.*
20. Web Scraping versus Twitter API: A Comparison for a Credibility Analysis	*The study of this paper sheds light on the difference between data extraction models, there are various data extraction techniques such as web scraping, extraction using API and manual extraction. This authors of this paper focus on the differentiation between web scraping and drawing out information using API. As simple as it may sound, both of them have their own set of limitations, yet they both are very useful depending on the type of task they are chosen for.* *As for the credibility, Twitter API was made to go through a real-time credibility model where we found out that through an API we could collect more data, web scraping is definitely a faster method but the amount of information it provides won't be enough for our project. Hence, we came to a decisive statement to use the API of Twitter rather than the web scraping method.*

III. REQUIREMENT ANALYSIS

A. Requirement Gathering

The requirement gathering phase involves an exploratory process of generating a list of requirements that need to be fulfilled by our proposed system and gather insights the aspects that project needs to address or include.

B. Information Gathering

In the information gathering, on understanding our problem statement, we need to identify the tasks and goals and for that we collect all the required information. In case of our project the majorly required information includes:

1) *Tweets (Railway Complaints)*
2) *Basic Information about our user (user id, PNR number of the train)*

C. Functional Requirements

1) *User: It is necessary that the users have a twitter account. In case the user doesn't have a twitter account, they can easily signup by filling in their basic information like their email-id, name, etc. On having a complaint or in order to provide feedback, they can tweet at @Railminindia which the twitter account of the ministry of railways*
2) *Dataset: It refers to a file which contains one or more records. A Data set is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable. And each row corresponds to a given member of the data set, as per the given question*
3) *App Module: It consists of three modules-*

D. Non – Functional Requirements

1) *Essential Requirements*

The system should have the following –

a) *97% of reliability*
b) *A backup scheduled every-day*
c) *Should provide data-security using encryption method.*
d) *Should update the scheme availabilities in Real-time*

2) *Optional Requirements*

The System should –

a) *Process request within 2 seconds*
b) *Be able to retrieve the data since created.*

IV. SYSTEM ANALYSIS

A. Existing System

When a traveller, posts a tweet with the tags necessary to contact the Indian Railways, the application over there receives a request. This request is automated by twitter and is sent to Indian Railways.

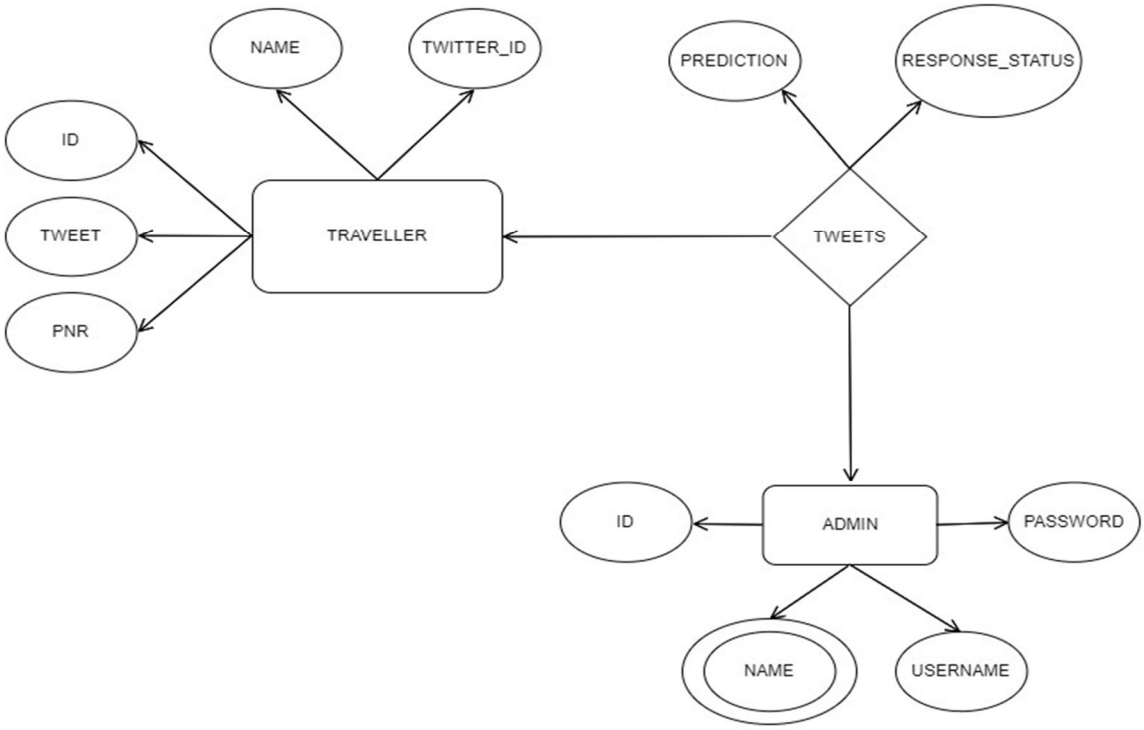
Once the application receives this request, it generates a 'Ticket' for the support team. A huge team of technicians and supporting team is seated behind the scenes to efficiently resolve all the queries. Now, this ticket gets assigned to one of the team members and it pops up on their computer. As soon as the member receives the ticket, he has to act on it accordingly, within a particular amount of time.

B. Proposed System

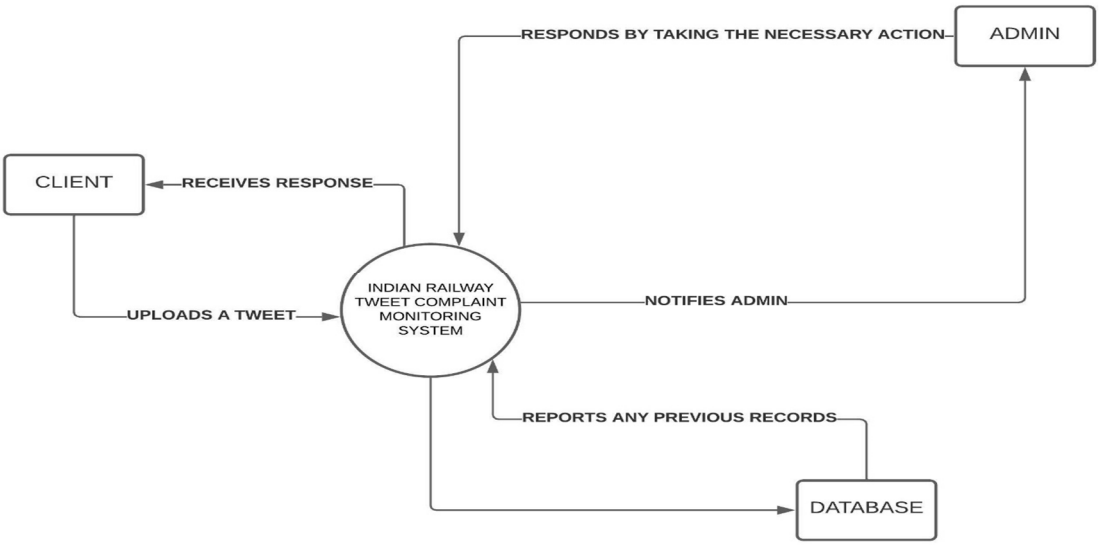
The project reduces the work complexity of scanning through thousands of useless data to find particular information but from here we can directly find the relevant tweets that needs attention. To solve the problem, we will be using a Machine Learning (Naive Bayes) model. In machine learning, Naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. The proposed system will divide the tweets in two category that is emergency and feedback. The manual work done by the employees will be reduced .The tweets will be replied in the real time.

C. Diagrams

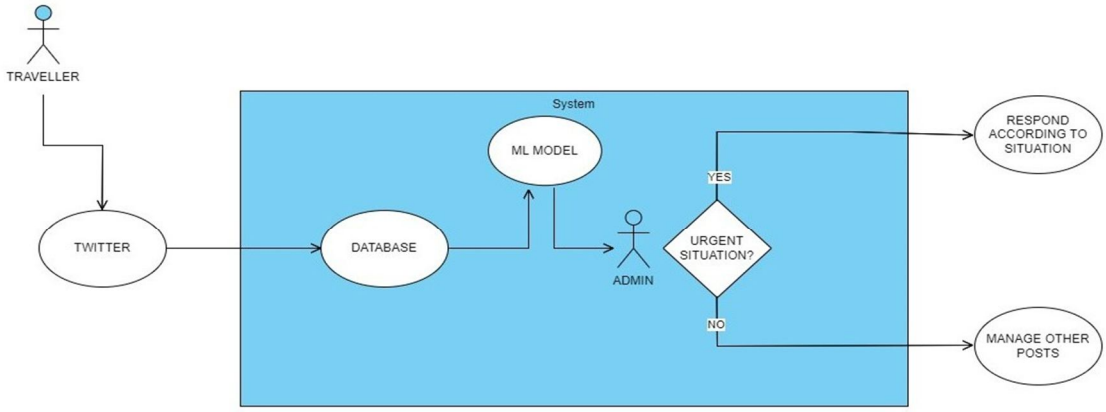
1) *Entity-Relationship Diagram:* An Entity–relationship model (ER model) describes the structure of a database with the help of a diagram, which is known as Entity Relationship Diagram (ER Diagram). An ER model is a design or blueprint of a database that can later be implemented as a database. The main components of E-R model are: entity set and relationship set.



2) *Data Flow Diagram:* A data flow diagram (DFD) is a graphical representation of the flow of data through an information system without any indication of time. DFDs are commonly used to provide an initial top-down analysis of a system, identifying the processes to be carried out and the interactions and data exchanges between them. DFDs can be either logical, providing an implementation-independent description of the system, or physical describing the actual entities (devices, department, people, etc.) involved.



3) *Use Case Diagram:* A use-case model is a model of how different types of users interact with the system to solve a problem. As such, it describes the goals of the users, the interactions between the users and the system, and the required behaviour of the system in satisfying these goals. A use-case model consists of a number of model elements. The most important model elements are: use cases, actors and the relationships between them.

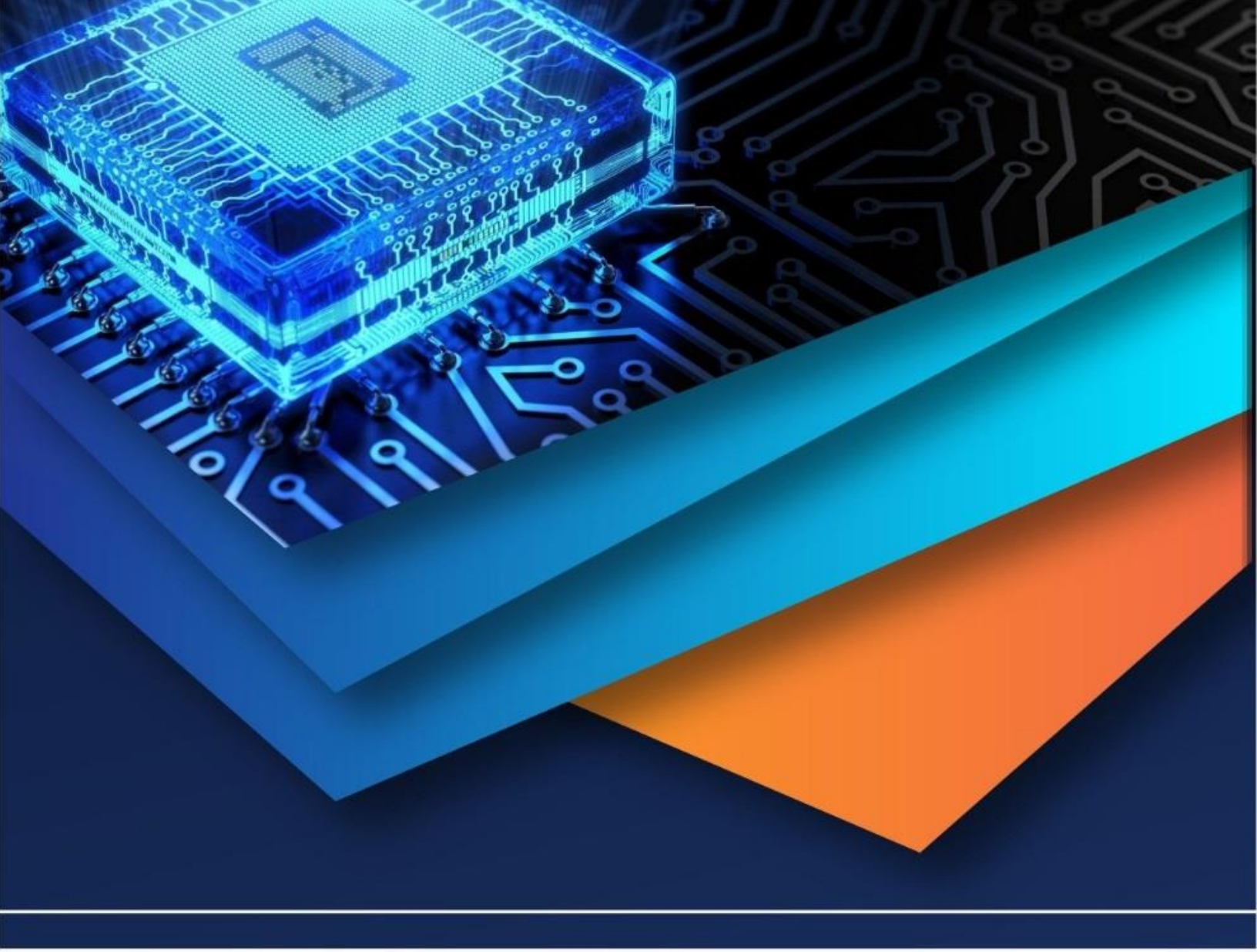


V. CONCLUSION

This project is basically focused on increasing the efficiency of the existing system and even if this project does not replace the existing system of Indian railways, it can still be depicted to show that such sort of applications can be developed to help the government bodies working at different levels.

REFERENCES

[1] https://www.researchgate.net/profile/Sandip_Vijay2/publication/340403235_Secure_Intelligent_Optimized_Link_Heuristic_in_Cross-Network_Handover_for_IoT/links/5e91b676a6fdcca7890a6621/Secure-Intelligent-Optimized-Link-Heuristic-in-Cross-Network-Handover-for-IoT.pdf#page=697
[2] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3415956
[3] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf
[4] https://www.researchgate.net/profile/Bharti-Suri/publication/228943618_Understanding_Regression_Testing_Techniques/links/5580f92f08aea3d7096e5842/Understanding-Regression-Testing-Techniques.pdf
[5] http://ajba.um.edu.my/index.php/MJCS/article/view/6793/4466
[6] https://arxiv.org/ftp/arxiv/papers/1806/1806.06407.pdf
[7] https://www.researchgate.net/profile/Asan-Baker/publication/348277400_Response_Time_analysis_for_XAMPP_Server_based_on_Different_Versions_of_Linux_Operating_System/links/5ff5dfa8299bf14088759157/Response-Time-analysis-for-XAMPP-Server-based-on-Different-Versions-of-Linux-Operating-System.pdf
[8] https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F11490486S419.pdf
[9] http://www.ijcstjournal.org/volume-4/issue-3/IJCST-V4I3P16.pdf
[10] https://www.irjet.net/archives/V7/i1/IRJET-V7I131.pdf
[11] https://dl.acm.org/doi/abs/10.1145/3344341.3368814
[12] https://www.mdpi.com/2220-9964/10/3/155
[13] https://www.ijcrt.org/papers/IJCRT2102147.pdf
[14] http://notes.stephenholiday.com/Kafka.pdf
[15] https://www.irjet.net/archives/V7/i6/IRJET-V7I6418.pdf
[16] https://dl.acm.org/doi/abs/10.14778/2824032.2824063
[17] https://www.usenix.org/legacy/event/usenix10/tech/full_papers/Hunt.pdf
[18] https://ieeexplore.ieee.org/abstract/document/8787761
[19] https://www.sciencedirect.com/science/article/abs/pii/S03064573163005894
[20] https://www.researchgate.net/profile/Irvin-Dongo/publication/348813187_Web_Scraping_versus_Twitter_API_A_Comparison_for_a_Credibility_Analysis/links/603b2c14299bf1cc26f7ab16/Web-Scraping-versus-Twitter-API-A-Comparison-for-a-Credibility-Analysis.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY