ISSN 2063-5346

A NOVEL ROAD TRAFFIC ACCIDENTS PREDICTION MODEL WITH RANDOM CLASSIFIER AFTER HYPER-PARAMETER TUNED USING GRIDSEARCHCV

Syeda Sadiya Sultana[1] Dr. Anitha Patil[2]

Article History: Received: 10.05.2023 Revised: 29.05.2023 Accepted: 09.06.2023

Abstract

Accidents are occurring increasingly frequently and at an alarming rate as a result of the quickly increasing number of vehicles on the road. certain the increased number of traffic incidents and fatalities today, the ability to anticipate the number of accidents over a certain time period is essential for the transportation department to make wise decisions. In this case, it will be advantageous to examine the accident frequency so that we may make use of this knowledge to build tactics to minimise them. Even if the majority of accidents have ambiguous characteristics, a certain amount of regularity can be seen when incidences are observed in one place over time. This regularity can be used to make precise predictions about the probability of accidents happening in a specific region and to develop accident prediction models. In this project, we looked into how road conditions, environmental factors, and traffic accidents are related to one another. We have created a data mining-based accident prediction model using GridsearchCV and the Random Forest Classifier. A road traffic accident dataset that was publicly available online was used for this inquiry. The study's findings can be used to improve the construction of roads and automobiles by a variety of parties, including but not limited to the government's public works agency, contractors, and other automobile businesses.

***Index Terms* —** Industries, Road accidents, Government, Predictive models, Prediction algorithms, Data models, Road safety.

[1]Research Scholar, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

[2]Professor, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

DOI:10.48047/ecb/2023.12.9.32

I. INTRODUCTION

The alarming rate of increase of accidents in India is now a cause for serious concern. According to some recent statistics [1], India accounts for roughly six percent of global road accidents while owning only one percent of the global vehicle population. There are a lot of accident cases reported due to the negligence of two-wheelers, whereas over-speeding is also another contributing factor. Accidents caused while under the influence of alcohol or during general traffic violations are also common. In spite of having set regulations and the highway codes, the negligence of people towards the speed of the vehicle, the vehicle condition and their own negligence of not wearing helmets has caused a lot of accidents. While the major cause of road accidents is attributed to the increasing number of vehicles, the role played by the condition of the roads and other environmental factors cannot be overlooked. The number of deaths due to road accidents in India is indeed a cause for worry. The scenario is very dismal with more than 137,000 people succumbing to injuries from road accidents. This figure is more than four times the annual death toll from terrorism.

Accidents involving heavy goods vehicles like trucks and even those involving commercial vehicles used for public transportation like buses are some of the most fatal kind of accidents that occur, claiming the lives of innocent people. Weather conditions like rain, fog, etc., also play a role in catalysing the risk of accidents. Thus, having a proper estimation of accidents and knowledge of accident hotspots and causing factors will help in taking steps to reduce them. This requires a keen study on accidents and development of accident prediction models. To implement a well-designed road framework management system for looking into road security aspects, it is often desired to have an optimized accident prediction model which can analyze potential issues arising due to infrastructure fallbacks and to estimate the effect of existing models in reducing the occurrence of accidents.

The main challenges involved in the creation of such a model include the evaluation of the weight that can be attributed to the impact of each variable in contributing to the accident and assessing how the model can be best designed to incorporate the effects of all such variables. Data mining techniques and models have in the past been found useful for the purpose of data interpretation in a variety of domains including but not limited to credit risk management, fraud detection, healthcare informatics, recommendation systems and so on. Approaches involving artificial intelligence and machine learning have further helped to augment these studies. For this project, we have investigated the inter-relationship between the occurrences of road accidents and the roles played by the underlying road conditions and environmental factors in contributing to the same. Since such a study requires us to cover several aspects affecting accidents, we can make use of data mining techniques to analyze this data to extract relevant details from them, as these huge volumes of data would otherwise be meaningless without the right interpretation applied to them.

In this project, we are discussing the effects of such an accident prediction model in identifying the risks involved in road accident scenarios. The next section discusses the prior works done with respect to analyzing the different accidents that have taken place over the years. This is followed by a summarized description of the methodology used in this work. Further, the different components of implementation including the system architecture, software and languages used, simulation, user interface and screenshots of the developed application are discussed. Finally, the discussion and conclusions derived from the present study and the future scopes are outlined in the last two sections. The results from this study have been used to propose a model that can be used as a tool to estimate the possibility of road accidents in a particular area chosen by the user.

II. SYSTEM ANALYSIS

Problem Statement

Due to the exponentially increasing number of vehicles on the road, the number of accidents occurring on a daily basis is also increasing at an alarming rate. With the high number of traffic incidents and deaths these days, the ability to forecast the number of traffic accidents over a given time is important for the transportation department to make scientific decisions. There is no automated system that can achieve this activity and needs human intervention.

Objective

The primary aim of this project is to predict the severity of accidents. We have used the Road Traffic Accident data set on kaggle for this purpose. This data set is preprocessed, analyzed and fed to multiple classification algorithms and its metrics are compared.

Technical Approach

Below is the technical approach to address the problem:

1. Identification of dataset
2. Data Pre-processing and Exploratory data analysis
3. Feeding the dataset to multiple algorithms and finding the best algorithm that suits the scenario
4. Training the final classifier and creating a model for the final classifier
5. Testing the final classifier and saving the results.

Proposed System:

The proposed system architecture of the accident prediction model. This architecture consists of three main phases such as pre-processing, modelling and result analysis. For predicting controlling the road accidents, different techniques are used and performance estimation of classifiers are discussed in the following parts. Before construction of each model data preparation was performed. All missing values are removed and all the numerical values are converted to nominal values according to data dictionary. Also the removing of unnecessary attributes are performed. In Modeling, to show the basic characteristics of the accidental deaths the statistical are calculated. Then classifying the different accident rates in different areas. Based on the dataset collected, future accident rates are predicted. Later comparison of the accident rates in different states with distinct datasets.

In the result analysis, the accuracy of different techniques are analyzed. In the proposed work, An hybrid technique is developed using the data mining techniques such Linear Regression, K-Means Clustering, Association Rule and Naive Bayes algorithms.

Advantages of Proposed System

- Able to predict the accident occurrence accurately.
- Minimize loss and maximize life saving.

ALGORITHM/ TECHNIQUE USED

After Feature Transformation and Data Pre-processing, the dataset is fitted to a model. The algorithm receives the training set in order to learn how to forecast values. After creating a target variable to predict, testing data is provided as input. The models are created using a technical implementation for the Drug Review dataset for the Drug Recommendation System.

- Splitting the data set into test and training data;
- Data cleaning and visualisation;
- Feature Extraction;
- Data preparation using BOW, TF-IDF, and Word2Vec;
- Data modelling using Sequential and XGBoost classifier;
- Data evaluation and prediction;
- Building the Drug recommendation system.

The technical method for the Drug Recommendation System using the Drug Review Dataset is shown below.

1. Cleaning and displaying data
2. Secondly Feature Extraction
3. Data Preparation Using TF-IDF, Word2Vec, and BOW
4. Dividing the data set into training and test data
5. Data modelling with the XGBoost classifier and sequential LSTM
6. Evaluation and Prediction of Data
7. Developing a mechanism for drug recommendations

III. PROPOSED MODULAR IMPLEMENTATION

Data Pre-processing:

1. Detect the outliers
2. Impute missing details using simple imputer
3. Encode categorical data
4. Split train data and test data
5. Perform feature selection using Chi-Squared method
6. Feed the dataset to multiple machine learning algorithms.
7. Below are the accuracy results:

Algorithm	Accuracy in %
Decision Trees	69.19
Random Forest	84.54
Suport Vector Machine	84.38
Decision Trees	73.13
LogisticRegression	84.37
KNN	82.63
GradientBoostingClassifier	84.9
AdaBoost	84.25

Model Creation :

Create a random forest classifier after hyper parameter tuning.

We achieve accuracy of 91.68% for the hyper tuned random forest classifier.

Modular Implementation

Below is the proposed modular implementation of the project. It consists of two modules:

1. Admin

Admin Module:

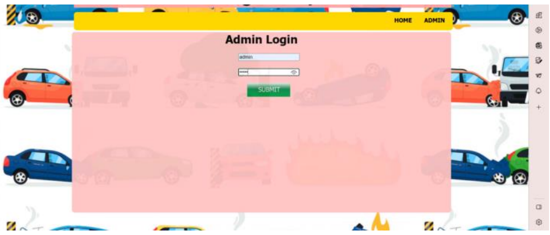
1. Login
2. Upload Road Traffic Accidents dataset that was downloaded from Kaggle
3. Exploratory Data Analysis
4. Data Preprocessing
   a. Check for duplicates in the dataset.
   b. Fill missing values

   c. Transform Categorical features using label encoding.
   d. Drop unnecessary features
   e. Balance the dataset using SMOTE
   f. Split the data into Training and Testing Datasets.
5. Feeding the dataset to multiple classification algorithms
   a. Random Forest
   b. Decision Trees
   c. Support Vector Machine
   d. Logistic regression
   e. K-Nearest Neighbour
   f. Gaussian Naïve Bayes
   g. Adaboost
   h. Gradient Boosting
6. Creation of model using Gradient Boosting Classifier

IV. PROJECT EXECUTION

Admin Login:

This is the login page for the admin module. The admin need to login into the system with his credentials in order to perform operations like uploading the dataset, Training the dataset, Exploratory data Analysis of the dataset, Feeding the dataset to different Machine learning Algorithms to find the Algorithm that can meet the best accuracy and Create a model that can be hosted on the Flask Application to be used by the users.



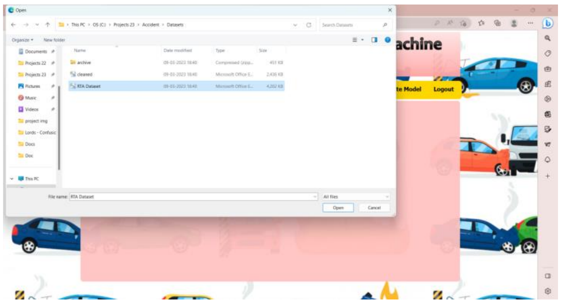
Admin Login

admin

Fig: Admin Login

Upload Dataset:

On this page, the administrator of the system can upload datasets that are used for training the machine learning models. The admin has to select the file by clicking on the Choose file button and click on the upload button to upload the file to the server. Once the upload is complete, a success message would be displayed that the file is successfully uploaded. For this project we are using RTA Dataset.csv as a dataset.

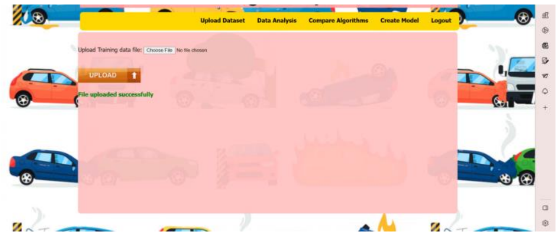


Name	Date modified	Type	Size
archive	09-03-2023 18:40	Compressed (zipp...	451 KB
cleaned	09-03-2023 18:40	Microsoft Office E...	2,436 KB
RTA Dataset	09-03-2023 18:40	Microsoft Office E...	4,202 KB

File name: RTA Dataset

All files

Fig: Upload Dataset & File Uploaded Successfully.



Upload Training data file: No file chosen

File uploaded successfully

Data Analysis:

Exploratory Data Analysis is performed on the dataset in order to clean the dataset for any missing data, identify patterns, identify the relationships of various parameters of the outputs with the help of graphs, statistics etc.

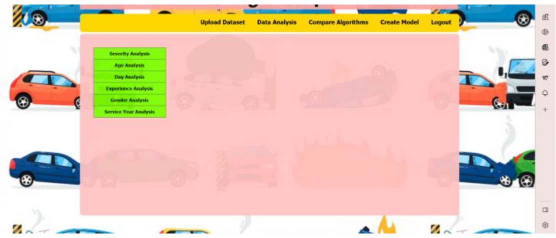


Fig: Data Analysis

Severity Analysis:

The below graph shows the Severity Analysis over data present in the dataset.

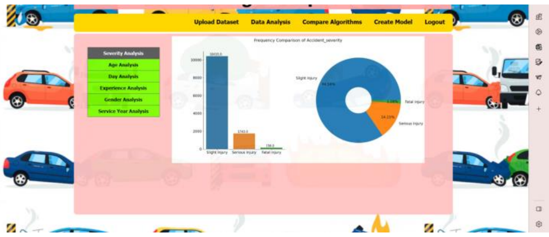


Fig: Severity Analysis

Age Analysis:

The below graph shows the Age Analysis over data present in the dataset.

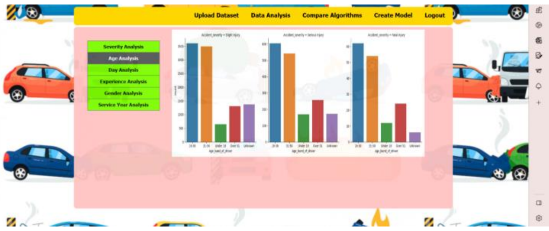


Fig: Age Analysis

Day Analysis:

The below graph shows the Day Analysis over data present in the dataset.

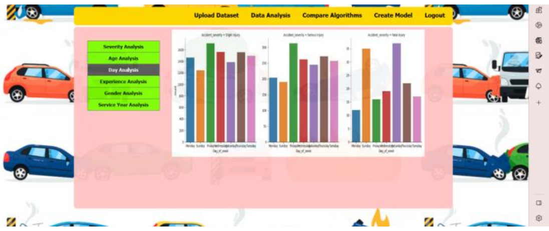


Fig: Day Analysis

Experience Analysis:

The below graph shows the Experience Analysis over data present in the dataset.

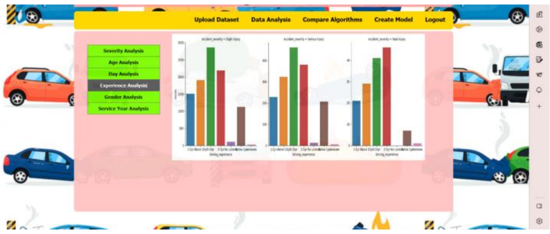


Fig: Experience Analysis

Gender Analysis:

The below graph shows the Gender Analysis over data present in the dataset.

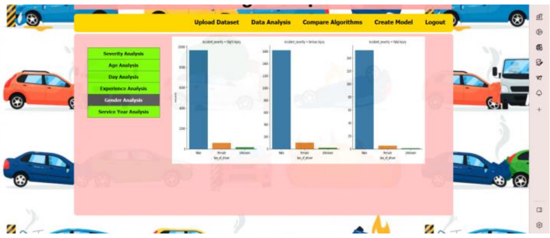


Fig Gender Analysis

Service Year Analysis:

The below graph shows the Service Year Analysis over data present in the dataset.

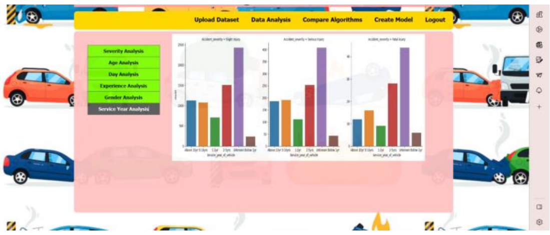
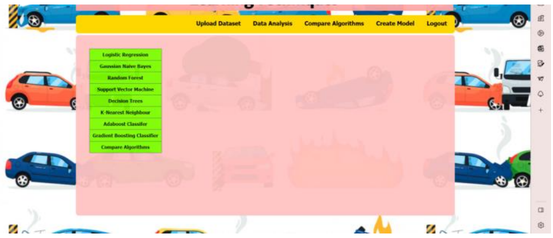


Fig: Service Year Analysis

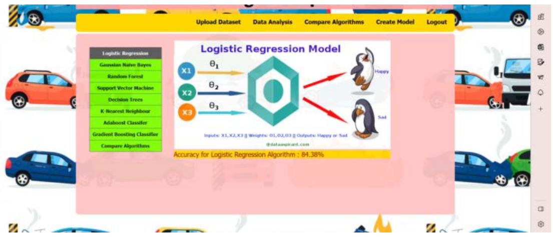
Compare Algorithms:

On this page, the admin can feed the dataset to various Algorithms to train them and get the test accuracy for each algorithm. When the dataset is feed to various algorithms to evaluate the situation with some parameters like Accuracy, F1-Score , Recall…



Logistic Regression Classifier:

When the dataset is feed to Logistic Regression algorithm we observe that the test accuracy is 84.38%.

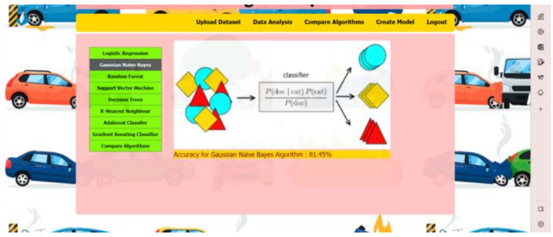


Accuracy for Logistic Regression Algorithm : 84.38%

Fig: Logistic Regression Classifier

Gaussian Naive Bayes Classifier:

When the dataset is feed to Gaussian Naive Bayes algorithm we observe that the test accuracy is 81.45%.

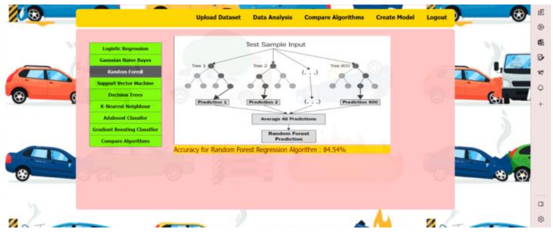


Accuracy for Gaussian Naive Bayes Algorithm : 81.45%

Fig: Gaussian Naive Bayes Classifier

Random Forest Classifier:

When the dataset is feed to Random Forest Classifier algorithm we observe that the test accuracy is 89.54%.

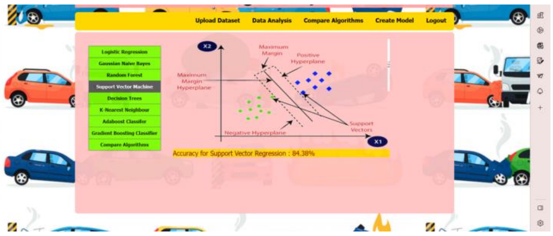


Accuracy for Random Forest Regression Algorithm : 84.54%

Fig: Random Forest Classifier

Support Vector Machine Classifier:

When the dataset is feed to Support Vector Machine Classifier algorithm we observe that the test accuracy is 84.38%.

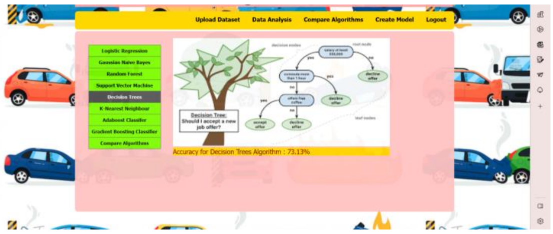


Accuracy for Support Vector Regression : 84.38%

Fig: Support Vector Machine Classifier

Decision Tree Classifier:

When the dataset is feed to Decision Tree Classifier algorithm we observe that the test accuracy is 73.13%.

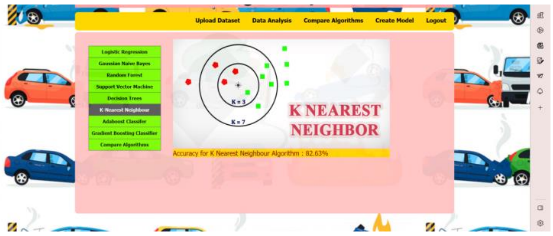


Accuracy for Decision Trees Algorithm : 73.13%

Fig: Decision Tree Classifier

k nearest neighbor Classifier:

When the dataset is feed to k nearest neighbor Classifier algorithm we observe that the test accuracy is 82.63%.

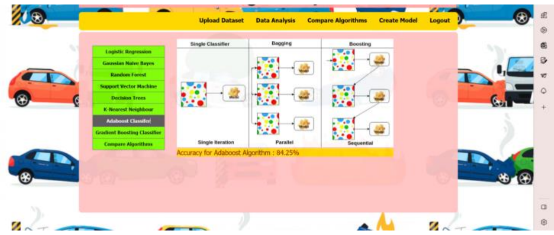


Accuracy for K Nearest Neighbour Algorithm : 82.63%

Fig: k nearest neighbor Classifier

Adaboost Classifier:

When the dataset is feed to Adaboost Classifier algorithm we observe that the test accuracy is 84.25%.

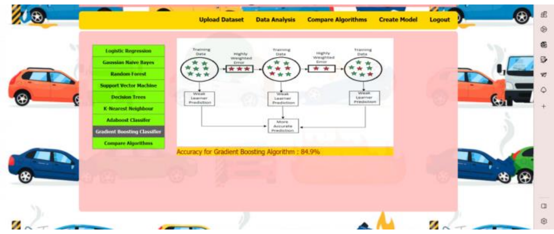


Accuracy for Adaboost Algorithm : 84.25%

Fig: Adaboost Classifier

Gaussian Naive Bayes Classifier:

When the dataset is feed to Gaussian Naive Bayes Classifier algorithm we observe that the test accuracy is 84.9%.



Accuracy for Gradient Boosting Algorithm : 84.9%

Fig: Gaussian Naive Bayes

Compare Algorithm Summary:

On this page, the admin can feed the dataset to various Algorithms to train them, get the test accuracy for each algorithm and their accuracies are summarized here.

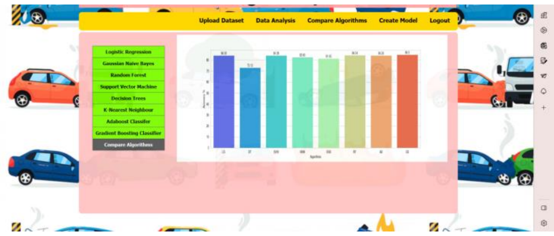
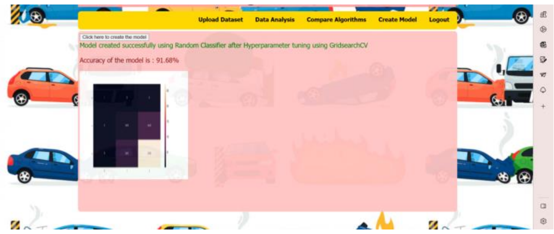


Fig: Compare Algorithm Summary

Create Model:

This screen shows the Accuracy of the Final Classifier Model is 91.68%.

Fig: Create Model



Model created successfully using Random Classifier after Hyperparameter tuning using GridsearchCV

Accuracy of the model is : 91.68%

CONCLUSION

An accident can change the lives of many people. It is up to each of us to bring down this increasing number. This can be made possible by adopting safe driving measures to an extent. Since all instances of accidents cannot be attributed to the same cause, proper precautionary measures will also need to be exercised by the road development authorities in designing the structure of roads as well as by the automobile industries in creating better fatality reducing vehicle models. One thing within our capability is to predict the possibility of an accident based on previous data and observations that can aid such authorities and industries. An accident has the power to change the lives of several people. To lower this rising number, we must all cooperate. Adopting safe driving habits can help with this to some extent. The road building authority must take sufficient safeguards when developing the layout of roads because not all accidents can be attributed to the same reason, and the automobile industry must create better car models that decrease fatalities. We can help these authorities and businesses by predicting the likelihood of a disaster based on historical data and observations.

The model can be improved further in the future to incorporate a number of limitations that were not considered in the current study. The government may effectively use these optimised models to lower traffic accidents and execute regulations for road safety. The creation of a smartphone application that will assist drivers in deciding on a route for a ride is another aspect of this endeavour. It is also possible to implement a call-out to the driver using the mapping service, which would also declare the likelihood of risk along a selected route in addition to the instructions. In the future, service provider businesses like Uber, Ola, and others may implement this. Additionally, this will help in improving the surveillance of regions that are prone to accidents and in providing emergency assistance in the event of one. The dangers identified by this algorithm can be used to improve the road safety signs that are posted along roadways.

REFERENCES

[1] George Yannis, Anastasios Dragomanovits, Alexandra Laiou, Thomas Richter, Stephan Ruhl, Francesca La Torre, Lorenzo Domenichini, Daniel Graham, Niovi Karathodorou, Haojie Li . "Use of accident prediction models in road safety management – an international inquiry". Transportation Research Procedia 14, pp. 4257 – 4266.

[2] Srivastava AN, Zane-Ulman B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In Aerospace Conference, IEEE. IEEE 3853-3862.

[3] Ghazizadeh M, McDonald AD, Lee JD. Text mining to decipher free-response consumer complaints: Insights from the nhtsa vehicle owner’s complaint database. Human Factors 56(6): 1189-1203.

[4] Chen ZY, Chen CC.. Identifying the stances of topic persons using a model-based expectationmaximization method. J. Inf. Sci. Eng 31(2): 573-595.

[5] Williams T, Betak J, Findley B. Text mining analysis of railroad accident investigation reports. In Joint Rail Conference. American Society of Mechanical Engineers V001T06A009- V001T06A009.

[6] Suganya, E. and S. Vijayarani. “Analysis of road accidents in India using data mining classification algorithms.” International Conference on Inventive Computing and Informatics (ICICI): 1122-1126.

[7] Sarkar S, Pateshwari V, Maiti J. Predictive model for incident occurrences in steel plant in India. In ICCCNT, IEEE, pp. 1-5.

[8] Stewart M, Liu W, Cardell-Oliver R, Griffin M. An interactive web-based toolset for knowledge discovery from short text log data. In International Conference on Advanced Data Mining and Applications. Springer, pp. 853-858.

[9] Zheng CT, Liu C, Wong HS. Corpus based topic diffusion for short text clustering. Neurocomputing 275: 2444-2458.

[10] ArunPrasath, N and Muthusamy Punithavalli. “A review on road accident detection using data mining techniques.” International Journal of Advanced Research in Computer Science 9: 881-885.