Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)
IEEE Xplore Part Number: CFP21K25-ART

A Road Accident Prediction Model Using Data Mining Techniques

Dhanya Viswanath
Assistant Professor, Department of Information Science and Engineering
CMR Institute of Technology
Bengaluru, India
dhanyaarun0705@gmail.com

Preethi K
UG Scholar, Department of Information Science and Engineering
CMR Institute of Technology
Bengaluru, India

Nandini R
UG Scholar, Department of Information Science and Engineering
CMR Institute of Technology
Bengaluru, India

Bhuvaneshwari R
UG Scholar, Department of Information Science and Engineering
CMR Institute of Technology
Bengaluru, India

2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418336

***Abstract*— 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. In this scenario, it will be good to analyze the occurrence of accidents so that this can be further used to help us in coming up with techniques to reduce them. Even though uncertainty is a characteristic trait of majority of the accidents, over a period of time, there is a level of regularity that is perceived on observing the accidents occurring in a particular area. This regularity can be made use of in making well informed predictions on accident occurrences in an area and developing accident prediction models. In this paper, we have studied the inter relationships between road accidents, condition of a road and the role of environmental factors in the occurrence of an accident. We have made use of data mining techniques in developing an accident prediction model using Apriori algorithm and Support Vector Machines. Bangalore road accident datasets for the years 2014 to 2017 available in the internet have been made use for this study. The results from this study can be advantageously used by several stakeholders including and not limited to the government public work departments, contractors and other automobile industries in better designing roads and vehicles based on the estimates obtained.**

Keywords*— *Accident prediction, Data mining, Apriori algorithm, Rule mining, Classification

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 paper, 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 paper, 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. LITERATURE SURVEY

The steady increase in the rate of accidents in India have prompted many researchers to look into the factors affecting road accidents and study about it. Since data mining techniques do not require certain assumptions between dependent and independent variables which are required in traditional statistical techniques, various categories of data mining techniques have been made use of in creating prediction models for road accidents in the past. Researchers have focussed on different sets of attributes in developing such models. Srivastava et al. [2] and Ghazizadeh et al. [3] have mainly concentrated on studying the accidents occurring at intersection points. While the former has looked into categorizing the accidents based on their levels of seriousness using a Multi-layered perceptron (MLP) technique that was found to be more effective, the latter has made use of a feed forward MLP that utilizes back propagation learning to analyse the effect of several factors such as day or night, traffic conditions etc. on accidents. Chen et al. [4] in their studies have found highways to be the common area where a majority of accidents had been reported to occur.

Williams et al. [5] have found through their studies that the age and experience of a driver also play a major role in the occurrence of accidents. Suganya, E. and S. Vijayarani [6] in their paper have analysed the road accidents in India and compared the performance of different classification algorithms such as linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm using accuracy, error rate and execution time as a measure of performance. They have found the performance of KNN to be better than that of the others. Sarkar et al. [7] have done a comparative study on the type of roads that are prominent in accidents. While exploring the other components associated with accidents, they have found that the occurrence of accidents in highways is more common than in a normal road similar to [4]. Stewart et al. [8] have utilized original data in building a neural network model to predict accidents. They found that this model was able to give quicker results than those being used in the models built on Indian roads.

Zheng et al. [9] have studied the range of injuries that come forth in a motor vehicle accident and have also analyzed the emotions of the drivers involved in the accidents that could have been a causal factor. Arun Prasath N and Muthusamy Punithavalli [10] have conducted an extensive survey on the different techniques used in road accident detection over the years, the approaches implemented in them and discusses their merits and de-merits.

George Yannis et al. [11], in their paper, have discussed about the current practices used in the development of accident prediction models on an international level. Detailed information on various models have been collected with the help of questionnaires and they have made use of this data to identify which could be the most useful model that can be applied for accident prediction.

Anand, J. V [12] has developed a method to determine the effect of different variables in the detection and prediction of atmospheric deterioration all over the world. Fuzzy C means clustering, R-studio, and the ARIMA frame work have been made use of in creating this method. A similar approach can also be tried in evaluating the impact of various factors on road accidents. Analyzing the original cause of accidents is important because this will tell us the impact factor and contribution of each attribute towards road accidents. Tiwari et al. [13] have made use of self-organizing maps, K-mode clustering techniques, Support Vector Machines, Naïve Bayes and Decision tree to classify the data from road accidents based on the type of road users.

An analysis of the historical data will be useful in identifying accident hotspots. This was done by N. Singh et al. [15] to create a conceptual framework for the development of a model to detect the accident prone areas. On another note, Kaur, G et al. [14] have utilized correlation analysis and other visualization techniques with the help of R tool to study road accidents and traffic collision data and created an accident prediction model, focusing on state highways and normal district roads.

The key takeaway from all the previously mentioned researches done in the past is that if we are able to provide information to the people regarding the possibility of an accident, it will act as a guide to the inexperienced or new travelers in an area and make the people more cautious while driving. Government authorities will get to know the causes of accidents, the dominant factors such as weather conditions, transport infrastructure etc. that are inflicting upon the various accident prone regions and will help to provide assistance in sketching out the association between the various factors that directly or indirectly have a part in causing the accident. Information can be provided to the Regional Transport Office regarding the accident prone regions where alcoholic intoxication, distracting circumstances like making a phone call while driving, aggressive and careless driving, and disregard towards safety rules, fatigue of driver are the main causes of accidents. This information will help the RTO to impose strict actions such as checking the license of the driver, conducting alcohol check or even placing of additional traffic policemen in such areas. Our aim is also to assist them in organizing the traffic.

III. Proposed Methodology

In this paper, we have built an application that is capable of predicting the possibility of occurrence of accidents based on available road accident data. Data pre-processing is done on this road accident data to obtain a dataset. The data pre-processing step includes cleaning to remove the null and garbage values, and normalization of the data, followed by feature selection, where only relevant features from the original dataset are selected to be included in the final dataset. The dataset is then subjected to different data mining techniques. Clustering is performed on this dataset. The clusters are then subjected to other algorithms like Support Vector Machines (SVM) and Apriori. Since the data being used for the study has an unknown distribution and we need to sort out the frequent and infrequent items in the dataset, the former (SVM) is used to predict the probable risk of accidents while the latter (Apriori) is applied to perform rule mining, that is, to generate a frequent item set based on given support and confidence values. Rules have been set considering different combinations of factors which have caused accidents of varying nature and severity in different road types and weather conditions. For the frequently occurring item sets, the chosen support and confidence values imply the higher probability of the particular combination of attributes in leading to an accident. For example, based on the rule mining done, the probable risk of an accident occurring even during fine weather in a junction on account of over-speeding is high and could prove to be fatal based on the training dataset. SVM classification has been used to characterize each accident event into a high or a low risk category. Various data mining techniques and exploratory visualization techniques are applied on the accident dataset to get the interpreted results. The architecture of the model implemented in this paper is shown in Fig. 1.

The estimation of various factors involved in road accidents help to determine their contribution towards causing accidents. These analyses and research helps in providing solutions in order to reduce the accident rate and decrease the fatality in the number of deaths.

IV. Implementation

A. Dataset

The dataset used in this study was obtained from the Open Government Data (OGD) Platform, India. Datasets pertaining to accidents in Bangalore region over the years 2014 to 2017 were made use of in developing the model. This dataset covers details including date, time and location of accidents, the nature of the accident, whether it was a head-on collision or caused due to over-speed, skidding or other causes, the type of the road – straight road or a curved road, how many lanes were there, whether it was a junction of multiple roads, the number of fatalities, and so on. It is the combination of these factors that can be modelled for the study in hand. But this cannot be modelled using a simple deterministic model, instead it would need a stochastic model to deliver the expected results. This necessitates the need of supporting machine learning algorithms to be added to the data mining techniques.

B. System Architecture

The raw road accident data obtained is pre-processed to form the dataset which will be input to the model. The model is further trained using the training data and made to predict the possible risk of accidents for an area that will be input by a user. A graphical representation is also shown to the user based on the obtained statistics.

The working of this model can be divided into four modules – Rule Mining, Risk Prediction, Graph Plot and New Data Entry.

Rule mining is done using Apriori Algorithm, where we generate frequent item set based on the dataset provided as input. Risk Prediction is done using SVM (Support Vector Machine) Algorithm, which is primarily used for classification. It takes the sample data set as input and performs classification. This module predicts the risk of accidents in a particular area. Plot graph generates the bar charts based on weather, previous accidents and causes for the accidents. The New Data Entry module is used to report the new cases of accidents.

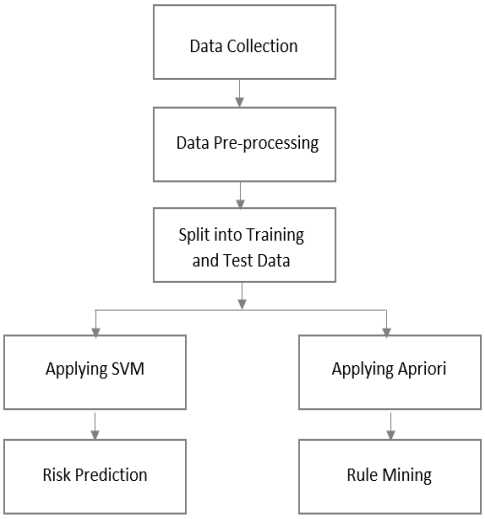

Fig.1. Architecture of the implemented model

C. Software and Languages Used

The application has been developed in Python language using Anaconda Spyder software for the implementation.

D. Simulation

The simulation is performed by using R tools. Various data mining techniques and exploratory visualization techniques are applied on the accident dataset to get interpreted results. The R tools help to develop an interactive user interface. Thus we can analyse the various factors contributing for the accidents by plotting various graphs, charts and other statistical and graphical representations.

1) User Interface

In the final User Interface of the application based on this model, we have four buttons, each corresponding to a particular module in the model. They have been named as

Rules, Plot Graph, Risk Prediction and New Data Entry (Fig. 2).

- Rules – This button will generate the frequent item sets based on the support and confidence values using Apriori Algorithm (Fig. 3).
- Plot Graph – This generates four graphs for the given area (Fig. 4). The first graph is based on the attributes such as over speed, skidding, etc which have been the causes of accidents (Fig. 5). The second is based on the weather conditions such as fine, cloudy, rainy, etc during the time of reported accidents (Fig. 6). The third graph is generated based on number of accident cases reported to nearby hospitals (Fig. 7) and the fourth graph predicts the number of accidents due to heavy vehicles such as trucks or lorries over the years in a particular area (Fig. 8).
- Risk Prediction – For a particular area submitted, this button predicts the possibility of accidents in the area as either HIGH or LOW (Fig. 9).
- New Data Entry – New accidents can be reported through this option. Information about the accident such as date, time, location, type of accident, etc. can be entered here which will be used for the collection of data sets in future (Fig. 10).

E. Screenshots of Application

To illustrate the working of the application based on the prediction model, the following screenshots from the application have been included below.

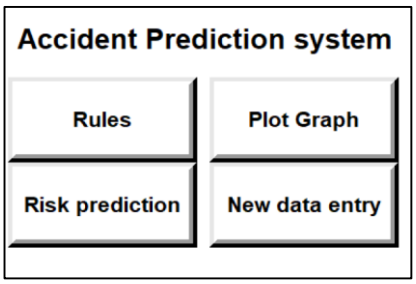

Fig.2: Application Interface

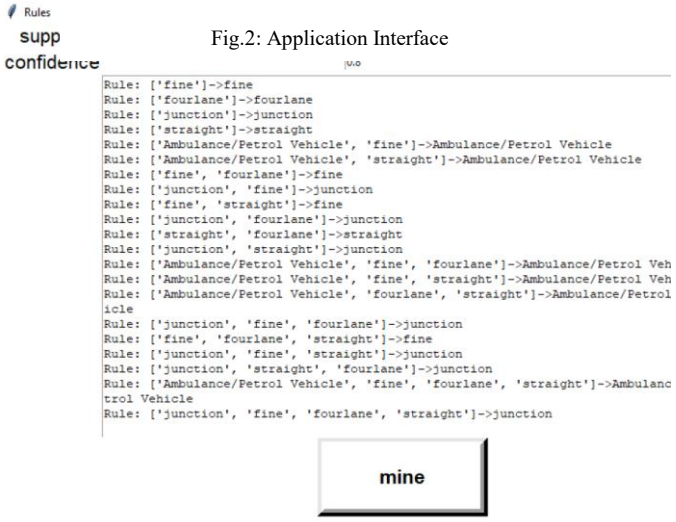

Fig.3: Rule Mining - View

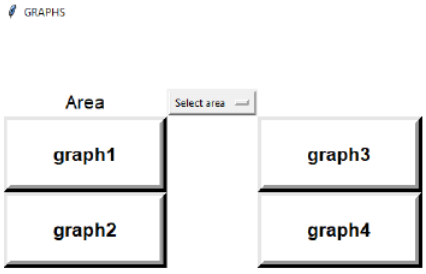

Fig.4: Graphical plot of risk related to accident - View

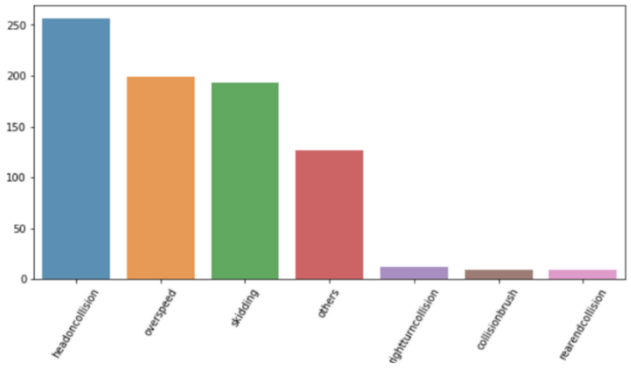

Fig.5: Plot graphs – Graph 1

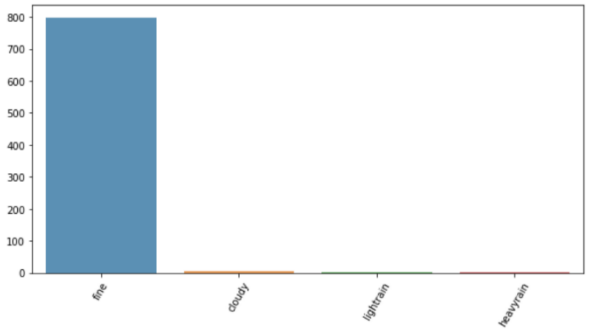

Fig.6: Plot graphs – Graph 2

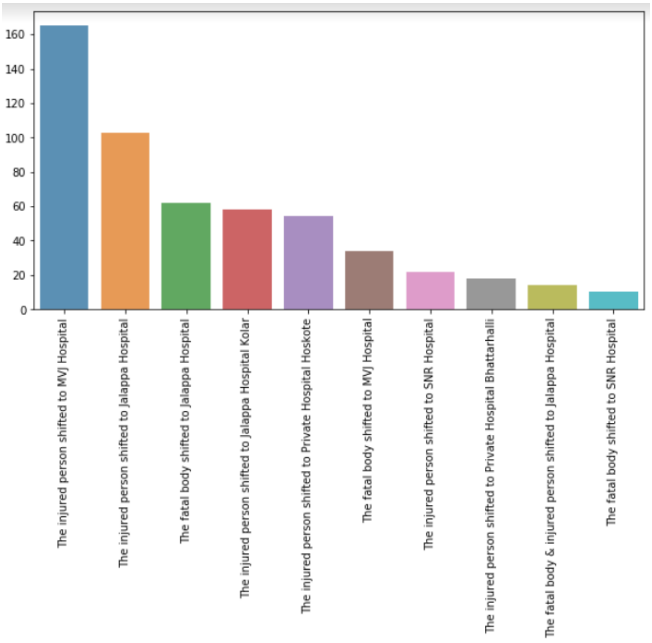

Fig.7: Plot graphs – Graph 3

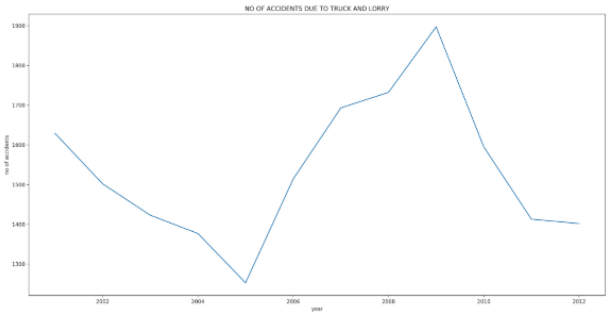

Fig. 8: Plot graphs – Graph 4

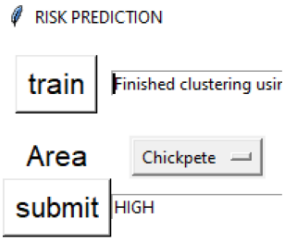

Fig.9: Risk Prediction for the chosen area - View

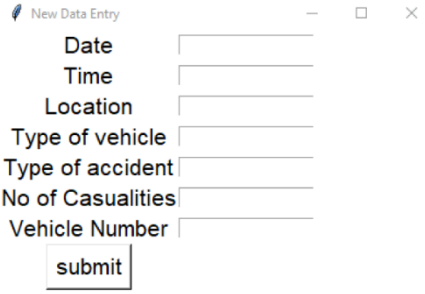

Fig.10: New data entry - View

V. DISCUSSIONS

In this paper, a road accident prediction model has been developed and implemented, taking into consideration different possible causative factors. The range of factors chosen for the study are limited to mainly the condition of the road, weather influences and the nature of accident cause. The emotional state of mind and experiential influence of the driver have not been considered as in past literature. Figures 5, 6, 7 and 8 are indicative of the various parameters that have been used in the study and creation of the prediction model. Figure 5 shows a comparative analysis of the number of accidents reported for each type of accident, such as head-on collision, over-speed, skidding and so on. Figure 6 shows the weather type observed for the reported accidents whereas figure 7 indicates the activity initiated in response after each reported accident. Figure 8 shows a graphical representation of the increasing number of accident cases reported due to heavy duty vehicles. All these data have been included in the dataset used for this study. This model has been used in the creation of an application that can be used to predict the probability of risk of accident over an area inputted by the user.

The user interface of the model based application outputs a graphical visualization of the factors that have been responsible for causing accidents relative to a specified area in the past. Based on this, a categorical prediction as high or low risk relative to accident occurrences is made for an area chosen by the user. The overall model has helped to give an understanding of the combinations of factors that have proven fatal in accident scenarios. A provision to further improve the dataset for future use has also been made in the form of an option to enter details of new accident cases.

VI. 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. This project was successful in creating such an application that can help in efficient prediction of road accidents based on factors such as types of vehicles, age of the driver, age of the vehicle, weather condition and road structure, so on. This model was implemented by making use of several data mining and machine learning algorithms applied over a dataset for Bangalore and has been successfully used to predict the risk probability of accidents over different areas with high accuracy.

The model can be further optimized in future to include several constraints that have been left out in the current study. These optimized models can be efficiently utilized by the government to reduce road accidents and to implement policies for road safety. Another scope of this work would be to develop a mobile app that will help the drivers in choosing a route for a ride. A call out to the driver through the maps service can also be implemented that would also announce the risk probability in a chosen route along with the directions. This can then be implemented by service provider companies such as Uber, Ola and so on in future. This will also be useful in having a better surveillance of accident prone areas and providing emergency services in the event of an accident. Better road safety instructions can also be installed along the highways taking into account the risks obtained from this model.

REFERENCES

[1] https://www.statista.com/topics/5982/road-accidents-in-india/

[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. (2014). Text mining to decipher free-response consumer complaints: Insights from the nhtsa vehicle owner's complaint database. Human Factors 56(6): 1189-1203. http://dx.doi.org/10.1504/IJFCM.2017.089439.

[4] Chen ZY, Chen CC. (2015). Identifying the stances of topic persons using a model-based expectationmaximization method. J. Inf. Sci. Eng 31(2): 573-595. http://dx.doi.org/10.1504/IJASM.2015.068609.

[5] Williams T, Betak J, Findley B. (2016). Text mining analysis of railroad accident investigation reports. In 2016 Joint Rail Conference. American Society of Mechanical Engineers V001T06A009-V001T06A009. http://dx.doi.org/10.14299/ijser.2013.01.

[6] Suganya, E. and S. Vijayarani. "Analysis of road accidents in India using data mining classification algorithms." 2017 International Conference on Inventive Computing and Informatics (ICICI) (2017): 1122-1126.

[7] Sarkar S, Pateshwari V, Maiti J. (2017). Predictive model for incident occurrences in steel plant in India. In ICCCNT 2017, IEEE, pp. 1-5. http://dx.doi.org/10.14299/ijser.2013.01.

[8] Stewart M, Liu W, Cardell-Oliver R, Griffin M. (2017). 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. http://dx.doi.org/10.1007/978-3- 319-69179-4_61.

[9] Zheng CT, Liu C, Wong HS. (2018). Corpus based topic diffusion for short text clustering. Neurocomputing 275: 2444-2458. http://dx.doi.org/10.1504/IJIT.2018.090859.

[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 (2018): 881-885.

[11] George Yannis, Anastasios Dragomanovits, Alexandra Laiou, Thomas Richter, Stephan Ruhl, Francesca La Torre, Lorenzo Domenichini, Daniel Graham, Niovi Karathodorou, Haojie Li (2016). "Use of accident prediction models in road safety management – an international inquiry". Transportation Research Procedia 14, pp. 4257 – 4266.

[12] Anand, J. V. "A Methodology of Atmospheric Deterioration Forecasting and Evaluation through Data Mining and Business Intelligence." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 02 (2020): 79-87.

[13] Prayag Tiwari, Sachin Kumar, Denis Kalitin (2017). "Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques". International Conference on Computational Intelligence, Communications, and Business Analytics. 10.1007/978-981-10-6430-2_31.

[14] Kaur, G. and Er. Harpreet Kaur. "Prediction of the cause of accident and accident prone location on roads using data mining techniques." 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (2017): 1-7.

[15] Irina Makarova, Ksenia Shubenkova, Eduard Mukhametdinov, and Anton Pashkevich, "Modeling as a Method to Improve Road Safety During Mass Events", Transportation Research Procedia 20 (2017) 43.