A Model for Traffic Management based on Text Mining Techniques

Ahmed Ibrahim Naguib[1], Hala Abdel-Galil[2], Sayed AbdelGaber[3]
Faculty of Computers and Artificial Intelligence
Helwan University, Helwan
Egypt

***Abstract*—It is very important for traffic management to be able to correctly recognize traffic trends from large historical traffic data, particularly the congestion pattern and road collisions. This can be used to reduce congestion, improve protection, and increase the accuracy of traffic forecasting. Choosing the correct and effective text mining methodology helps speed up and reduces the time and effort needed to retrieve valuable knowledge and information for future prediction and decision-making processes. Modeling collisions or accident risk has also been an important aspect of traffic management and road safety, as it helps recognize problems and causes that contribute to a higher risk of accidents, promotes treatment delivery, and reduces crashes to save more lives and avoid road congestion. Therefore, this work-study proposed a model that relies on the different text mining methodology to determine clearly what circumstances affect and who is involved more in an accident. Using different classification and machine learning techniques applied to get the optimum classifiers used in this model. The experimental results on real-world datasets demonstrate that the proposed models outperform Prayag Tiwari's Research Work related to the Leeds UK Dataset.**

***Keywords*—Classification; machine learning; text mining; traffic management**

I. Introduction

Traffic management is a major problem which almost daily affects us. Usage of technologies such as the Internet of Things (IoT) and image processing will result in a smooth traffic management system. The main cause of traffic congestion is the lack of an appropriate mechanism for prioritizing traffic. The IoT is an infrastructure network. There are switches, sensors, actuators, and circuits in the embedded systems. With software and connectivity locally or over the internet helps in the transfer of data which can be provided by ThingSpeak API that can get the data in CSV text format [1]. As outlined in different IoT applications, the deployment of incident notification systems is one of the most common and important technologies in the smart transport field [2].

The implementation of sensors and IoT devices in Smart Traffic Network helps collect user expectations and contextual details that may be in the form of travel time, weather conditions, or real-life driving patterns Once the congestion situation is expected, alternative congestion-free routes are proposed that can be propagated by text according to the desired criteria are suggested that can be propagated through text messages or e-mails to the users [3].

There are two forms of congestion: structural or incidental. Structural congestion arises when traffic demand is greater than available, while incidental congestion results from irregular circumstances such as accidents, bad weather, or road work that alters traffic flow [4]. The ability to predict the impact of an incident immediately after its occurrence is crucial to advanced traffic management and significantly improves the system's performance [5].

Present traffic flow management strategies may not be adequately successful to monitor the changing and continuing traffic as transportation departments face the possibility of being lost in the increasing volume of traffic details they handle. For example: assume the crash is detected on the highway that is being tracked using sensors installed on roadside poles by the transport authority. The directions should be given to traffic controllers to open an emergency exit road after investigating. This will require traffic to travel across the crash to prevent any road blockage. But the traffic controllers do not have advanced knowledge of any successful incident expected to take place at the emergency exit soon [3].

Text mining, also known as text analysis, is the process of transforming unstructured text into meaningful and actionable information. By identifying topics, patterns, and relevant keywords, text mining allows us to obtain valuable insights without needing to go through all data manually. Machine learning is an AI-derived discipline that focuses on developing algorithms that allow computers to learn from examples-based tasks. Machine learning models need to be equipped with input data, after which they can automatically predict with some degree of precision. The automatic text processing is proper when data mining and machine learning are merged.

The first thing should train a classifier subject model, by importing and marking a series of examples manually. The model should learn to distinguish topics and start creating correlations as well as its own predictions after being fed on many examples. To achieve reasonable standards of precision, you can feed a wide number of examples of the models that are representative of the problem you are trying to solve [6, 7].

Heterogeneity is the fundamental concern with accident data investigation is to recognize the most persuasive feature influencing accident recurrence and seriousness of the accident. The real issue with accident dataset analysis is its heterogeneous behavior. Heterogeneity in accident data is exceedingly undesirable and unavoidable [7]. With the growing amount of traffic information obtained from floating car data, it is highly beneficial to identify useful traffic patterns from the cumulative vast historical data collection, such as congestion patterns. Nevertheless, owing to the immense scale of the data collection, and the complexities and dynamics of traffic phenomena, it is difficult [8].

The accuracy of using traditional statistical analysis methods greatly relies on the size of the data. However, in many situations, data can be limited. The issue of small sampling has been always a problem in using crash data [9].

The proposed model integrated Variant algorithms in the preprocessing and text mining phases to enhance accuracy. It is not affected by the data size or quality. It uses efficient methods in preprocessing that enables the mining phase to work on any sample as presented the model applies on more one datasets (Maryland U.S. and Leeds UK).

The rest of this paper is organized as follows. Section 2 surveys the related works. Section 3 introduces the research objective. Section 4 discusses methods and techniques concerning text mining of the proposed model. Section 5 gives a description of the used datasets. Section 6 extends the experimental studies and results. Finally, Section 7 concludes the paper and presents suggestions for future work.

II. Related Work

Theoflatos et al. (2019) compared numerous machine learning approaches for real-time crash prediction for estimation of crashes. (including K-Nearest Neighbor, Naïve Bayes, Decision Tree, Random Forest, SVM, and Shallow Neural Network) and the Deep Feedforward Neural Network (DFNN) and found that the DFNN had more robust results in terms of different output parameters relative to other models [10].

Poch and Mannering used a seven-year incident dataset from 63 intersections in Bellevue, Washington (all based on procedural changes), this work tests a pessimistic binomial approximation of the recurrence of accidents at the approaches to crossing points. The calculation comes about disclosing crucial intersections between variables related to traffic and geography, and crash rates. The purpose of work-study offers exploratory analytical and objective data that may stimulate a way to work with the gage of the crash minimizing the benefits of multiple planned improvements on intersections that are operationally lacking [11].

Karlaftis and Tarko used the inquiry to aggregate the data and subsequently sorted the crash dataset into different categories and clustered output of the examined dataset by using Negative Binomial (NB) to classify the cause of the accident by a driver's centering age, which could have a few results [12].

Kwon OH used Naive Bayes and Decision Tree classification methodology to examine road safety-related factor dependencies [13]. Youthful Sohn utilized an alternative algorithm to enhance the accuracy of different types of classifiers for two severity categories of a traffic accident and every classifier utilized the neural network and decision tree [14].

Junhua Wanga used Two modeling approaches are implemented, including a conditional logistic regression model and a support vector machine model, and contrasted with forecasting crashes. The technique is evaluated based on the data obtained from the Shanghai Middle Ring Expressway, one of the major rings in Shanghai city, China [15].

Logistic regression was applied to collect crash-related information from traffic police reports, keeping in mind the end aim of examining the role of various situations in the incident severity. The demanded probity model was used to measure the effect on accident severity of pedestrian collision in a rural area of the highway and zone sort factors [16].

The model with the best fit and most elevated prescient capacity was used to classify the severity-related circumstances of the highway, an ecological problem, vehicle, and driver. Gadget usage, travel speed, impact intent, use of drugs and liquor, person condition, regardless of whether the driver is to blame, age, curve/grade, and rural/urban nature presence in the crash area were established as crucial elements for having an adverse effect on older drivers muddled in single-vehicle accidents [17].

Machine learning techniques such as the Support Vector Machine Algorithm (SVM) and the Fuzzy Clustering Models provide a modern and theoretically more reliable way to investigate the conventional safety issue. For example, in modeling the crash frequency on freeways, the negative binomial regression and the artificial neural network were compared and found the artificial neural network in modeling the crash frequency on freeways and found that the artificial neural network slightly outperformed the negative binomial regression [9].

In forecasting accident injury frequency on a mountainous highway based on real-time traffic and weather data, the analysis compared the fixed-parameter logistic model, the SVM model, and the random parameter logit model Other research suggested a new approach based on clustering algorithm and SVM model get a 78.0% accuracy [18].

Via Virtual Crash training tools, Oana Victoria Oțăt set out to educate the technological skills and analytical reasoning of the students in assessing the impacts of the pedestrian-vehicle. Therefore, familiarizing students with the methods offered by the program and to assist them in the study of virtual simulation of traffic accidents. Using the Simulated Crash program, students can learn how to assess the vehicle trajectory, based on vehicle movements during the pre-and post-impact phase [19].

III. Research Objective

The main objective of this work-study is to establish a model can achieve the optimum accuracy and identify the factors behind crashes or accident that could be helpful to reduce accident ratio in near future and could be helpful to save many lives, support in constructing the roads infrastructure, deteriorate wealth destruction as well as many other things. These outcomes will impact the urban movement police authorization measures, which will change the improper conduct of drivers and secure the minimum experienced street users. And give a fast response in case of crashes to get immediate support from EMS (Emergency Medical Service) unit for treatment. All the stakeholders and emergency teams will be automatically informed if an emergency is detected.

Proposes an IoT-based traffic management system for smart cities where traffic flows can be remotely managed by on-site traffic officers via their smartphones or centrally tracked or managed by the Internet [20]. There are circumstances that require defining policy objectives to constantly supply traffic information to the drivers, to keep the traffic moving.

IV. Proposed Model

The proposed Traffic Management Model discusses the main five characteristics as shown in Fig. 1 which implemented through this work-study using the Text Mining and Machine learning techniques.

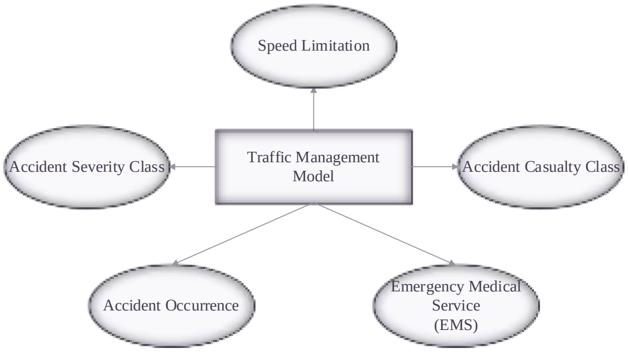

Fig. 1. Traffic Management Model Attributes.

The amount of data created is very huge and the amount of traffic information is collected through floating car data, so semantic models for data fusion and efficient algorithms for artificial intelligence are required to organize and process these data for extracting meaningful information especially from the accumulated massive historical dataset. Hence, data analyzing, and processing consider the big challenge for the data generated by IoT applications [2, 8].

Machine learning is a set of algorithms and statistical models that are used by computers to perform a required task. Machine learning can be used in traffic prediction. The data collected could be used in the construction of an idea display current traffic in the city and could be used in the future in making predictions of traffic & a congestion analysis can be done [21]. The explorer of knowledge sources that contain text or unstructured data is called “text mining”.

Text mining (also referred to as text analytics) is an artificial intelligence (AI) technology that transforms free (unstructured) text in documents and databases using natural language processing (NLP) into standardized, organized data suitable for analysis or to drive algorithms for machine learning (ML) [6].

For text mining, there are distinct approaches and techniques. The method of extracting essential patterns to discover information from textual data sources is text mining. Text mining is a multidisciplinary discipline focused on data retrieval, data mining, machine learning, digital linguistics, and statistics. It is possible to apply several text mining techniques to retrieve information, such as grouping, description, clustering, etc. Text mining deals with text in natural language and is encoded in semi-structured and unstructured format [22].

The structured data created by text mining can be integrated into databases, data warehouses, or business intelligence dashboards and used for descriptive, prescriptive, or predictive analytics. And the workflow detailed as following:

A. *Data Source and Gathering*

To make an effective model generate an accurate result, a real dataset from a trusted source must be provided, so the used datasets generated from official governmental source belongs to the US and UK government and published on the governmental official websites. A data preparation method, involving crash data filtering, floating car data filtering, and data matching on the road network, is introduced for the safety analysis purpose. And it contains:

- Dataset selection: The performance of real-time crash prediction relies greatly upon, in addition to the performance of the prediction model applied, a time-efficient and reliable data collection method [15].
- Data filtration and Cleansing: to define the most effective inputs (independent variables) compared to the traffic attributes in the related studies such as Weather condition, lighting, road conditions, … etc. and remove unwanted data which effect negatively on the accuracy and performance.

B. *Data Preprocessing*

Data preprocessing is performed on the road accident data to give it the proper shape required for analysis. Several attributes are transformed into a suitable form using data transformation methods such as Natural Language Toolkit (NLTK) and genism libraries.

Text clean-up: Removing any unnecessary or unwanted information such as ads from pages as the dataset (Text or CSV files) has a lot of noise and to perform machine learning algorithms efficiently as shown in Fig. 2.

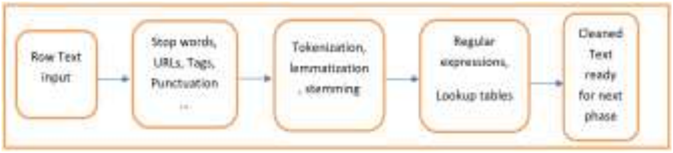
Row Text input
Stop words, URLs, Tags, Punctuation ...
Tokenization, lemmatization , stemming
Regular expressions, Lookup tables
Cleaned Text ready for next phase

Fig. 2. Text Clean-Up.

1) Feature extraction (also called the collection of attributes): This is the method of characterizing the text to achieve a quantitative measurement package. The number of words in a text, word forms, syntactic details, for example. It is possible to use these features for further processing.

2) Categorical values handling: Text Mining can turn text into numbers in the most general words, many machine learning or data science operations may include text or categorical values in the dataset (basically non-numerical values). With numerical inputs, most algorithms perform better. Therefore, translating text/categorical data into numerical data and still making an algorithm/model to make sense of it is the biggest obstacle faced by analyst principal methods: One-Hot-Encoding and Label-Encoder. And are used to convert text or categorical data into numerical data which the model expects and performs better with.

3) Data transformation (feature scaling): (Data Normalization and standardization) since both the features have different scales, there is an opportunity that higher weightage is given to features with higher magnitude. this can impact the performance of the machine learning algorithm and clearly, no have to algorithm to be biassed towards one feature. The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a spread of [0,1] Xchanged=(X−Xmin)/(Xmax−Xmin). Standardization rescales data to own a mean (μ) of 0 and variance (σ) of 1 (unit variance) Xchanged=(X−μ)/σ. Using sklearn.preprocessing and StandardScaler libraries.

C. Index

Creating an index of certain terms, their locations, and numbers. This allows quick access to and structuring of the processed data (Structured Database).

- Define the input features (independent variables) = X, and output labels (dependent variables) = y.
- Splitting the Dataset to (Training_Dataset) and (Testing_Dataset) to be X_train, X_test, y_train, y_test using sklearn.model_selection and train_test_split Libraries then define test_size data Percentage.

D. Mining

At this step, the text has been properly pre-processed and can now be ‘mined’. For that, applying different data exploration techniques to reveal new knowledge with high accuracy and performance.

- Dimensionality Reduction technique: PCA (Principal Component Analysis) is an unsupervised machine learning technique that attempts to derive a set of low-dimensional set of features from a much larger set while still preserving as much variance as possible. PCA can be thought of as a clustering algorithm. Usually, the original data is normalized before performing the PCA. It can be used for feature selection and visualizing higher-dimensional data where the feature pp is large. It is an unsupervised learning technique that can be used to identify patterns, clusters, and perhaps any latent information. Using sklearn.decomposition and PCA libraries. There are many techniques for the selection (reduction) such as wrappers but in choosing the most desirable subset of characteristics, PCA was concluded to be accurate by obtaining more variables of highest variance and low collinearity. Thus, in analysis and prediction, the right choice of algorithms offers performance. And it is more commonly used to reduce the dimensionality of a large dataset such that implementing machine learning where the original data is essentially high dimensional data becomes more realistic [23].
- Classification Techniques: Classification could be a form of supervised learning. It specifies the category to which data elements belong and is best used when the output has finite and discrete values. It predicts a category for an input variable yet. Classification predictive modeling involves assigning a category label to input examples.

E. Result Analysis

The mining steps produce raw results. These need to be evaluated and visualized so that they can be interpreted with respect to the questions the text-miner wants to investigate.

- Measurement of Accuracy: The classification accuracy is one of the important measures of how correctly a classifier classifies a record to its class value. The confusion matrix is an important dataa structure that helps in calculating different performance measures such as precision, accuracy, recall, and sensitivity of classification technique on some data using sklearn.metrics and confusion_matrix libraries. As shown in Table I.

TABLE I. CONFUSION MATRIX

	Negative	Positive
Negative	TN (True Negative)	FN (False Negative)
Positive	FP (False Positive)	TP (True Positive)

$$\text{Accuracy} = (TP + TN)/ (TP + TN + FP + FN) \quad (1)$$

$$\text{False Positive Rate} = FP/(TN+FP) \quad (2)$$

$$\text{Precision} = TP/(FP+TP) \quad (3)$$

$$\text{Sensitivity} = TP/(FN+TP) \quad (4)$$

$$\text{Recall} = TP/(TP+FN) \quad (5)$$

Using sklearn.metrics and accuracy_score, precision_ score, recall_score libraries in python to define the detailed results according to the previous functions.

F. Emergency Medical Service

To improve the possibilities of survival for passengers involved in car accidents, it is desirable to scale back the latent period of rescue teams and to optimize the medical and rescue resources needed. A faster and more efficient rescue will increase the probabilities of survival and recovery for injured victims. Thus, once the accident has occurred, it is crucial to managing the emergency rescue and resources efficiently and quickly. An Automatic Crash Notification system will automatically notify the closest emergency unit when a vehicle crash. These units will determine the character of the crash and, making it possible to predict the severity of injuries, data from vehicular sensors will allow the unit to judge if the vehicle has been involved during a collision. By using vehicular communications, cars involved in an accident can send an alert and other important information about the accident to nearby vehicles and to the closest wireless base station. Soon, a community-based effort involving the state departments, public organizations, and the industry is required to deploy the specified technology and infrastructure to attach all the vehicles on the road and therefore the emergency services. Basically, the data to be sent after an accident should include the following: (a) the time when the accident has occurred, (b) the placement of the vehicle to see the placement of the injured consistent with (longitude and latitude), (c) the characteristics of the vehicle and collision severity (allowing rescue services to send appropriate equipment to the accident site, and to warn them about the amount of complexity and dangers). All this information helpful in determining the severity of the impact, making it possible to avoid wasting lives, manage resources efficiently, and enable crashed vehicles to be far from the positioning, restoring traffic flow quickly.

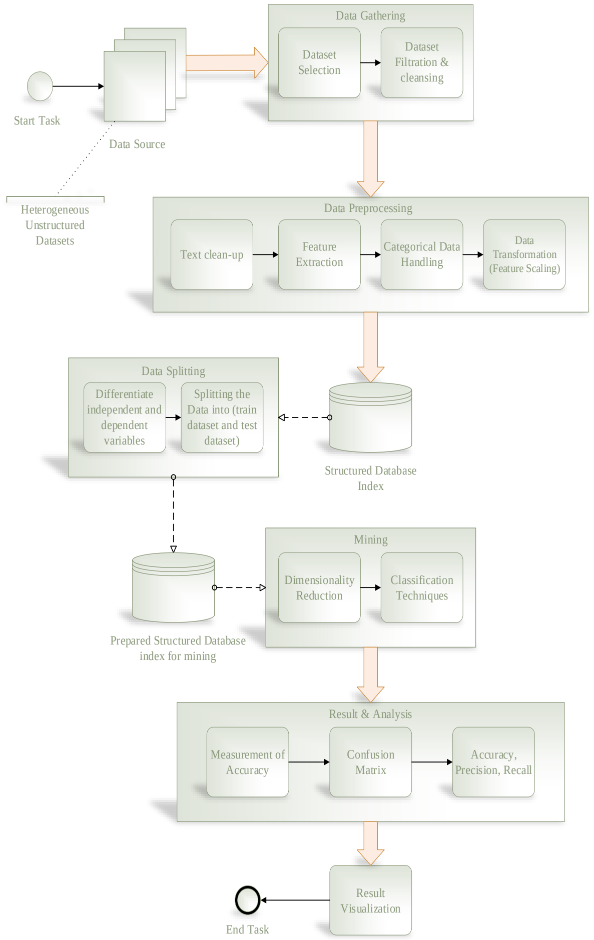
Start Task
Data Source
Heterogeneous Unstructured Datasets
Data Gathering
Dataset Selection
Dataset Filtration & cleansing
Data Preprocessing
Text clean-up
Feature Extraction
Categorical Data Handling
Data Transformation (Feature Scaling)
Data Splitting
Differentiate independent and dependent variables
Splitting the Data into (train dataset and test dataset)
Structured Database Index
Prepared Structured Database index for mining
Mining
Dimensionality Reduction
Classification Techniques
Result & Analysis
Measurement of Accuracy
Confusion Matrix
Accuracy, Precision, Recall
Result Visualization
End Task

Fig. 3. Traffic Management Model Workflow.

Accident notification systems will be specially designed for post-collision rescue services. New Intelligent Transportation Systems will emerge with the capability of improving the responsiveness of roadside emergency services. So, database to store the needed data must be designed, in used database the counties splitted to determine all the rescue units in the county related to the longitude and latitude, the rescue way according to the point either ground or air ambulance, the notified unit according to the nearest point in the crash location, and the treatment point which responsible to treat the injured persons according to injured severity. Using MySql with Python to implement that output. The detailed Traffic Management Model Workflow is shown in Fig. 3.

V. Description of Datasets

The traffic accident data is obtained from the online data source for Leeds UK. This data set comprises a massive number of accidents that occurred during a specific duration. Initial preprocessing of the data results in a set of attributes that found to affect the crash and the level of severity The attributes selected for analysis are several vehicles, time of the accident, road surface, weather conditions, lighting conditions, casualty class, the gender of casualty, age, type of vehicle, day, and month of the accident [24].

Crash data for Maryland U.S. from 2016 through 2019. Data is accurate as of the creation of the data. Only Approved Crash reports have been included in the file. Related datasets include Statewide Vehicle Crashes. This dataset provides general information about each collision and details of all traffic collisions occurring on the county and local roadways within Montgomery County, as collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police, and reported by the Montgomery County Police, Gaithersburg Police, Rockville Police, or the Maryland-National Capital Park Police. All this data is unstructured in text format located in CSV files, the data contains a set of highly important attributes which are very powerful in Traffic Management Model: Quarter of the year, Light Condition, Road Junction, Collision Type, Road Surface Condition, Road structure Condition, Weather Condition, Vehicle Characteristics, Crash severity, Human Class and Characteristics, Speed Limit on the related road, Detailed Date and Time, Geographical Location and other attributes which may enhance the accuracy in this model.

The real-time activity patterns are linked with the historical data to predict how the patterns may evolve soon, which is indeed valuable for transportation management [5]. The crash severity model is concerned with predicting the distribution of crashes or injuries by severity, given that a crash has already occurred: the model does not predict crash probability itself.

VI. Experimental Study and Results Discussions

In this section, the detailed results presented regards the experimental studies and make a comparison between the different classifiers and each accuracy.

A. Leeds UK Dataset (Casualty Class and Casualty Severity)

As stated in Prayag Tiwari's studies the researcher applied three classifiers algorithms (decision tree, Naïve Bayes, and SVM), and according to that practical studies they achieved better accuracy by using clustering techniques such as (Self Organizing Map (SOM) and k-modes) based on casualty class to determine clearly that what circumstances affect and who is involved more in an accident between the driver, passenger, or pedestrian, and the better classifier in Prayag Tiwari's work-study is Decision Tree which achieved accuracy 81% better than the (Naïve Bayes, and SVM) and lowest classifier accuracy is SVM which achieved accuracy 75.58% as shown in Fig. 4 [25].

In this practical study, more classifiers applied to classify this dataset based on casualty class and severity class, these classifiers classified data into 3 classes for both. The output of this classifier is determined based on the precision, recall, error rate, and other various factors that play an important role in accuracy measurement. And based on casualty class and severity class, so the Specialists can see clearly that what circumstances affect and who is involved more in an accident between the driver, passenger, or pedestrian. In addition to Prayag Tiwari's Research Work elements, the severity class which is considered an added contribution selected for high traffic management perspective is added. The other contribution which considers an added value in this model is utilizing more classifiers algorithms which are (Random Forest and Logistic Regression). Already The better classifier in this work-study regards Casualty Class is Random Forest classifier which achieved accuracy 87.79% better than others and lowest classifier accuracy is Logistic Regression which achieved accuracy 75% and the Decision Tree is better classifier in this work-study regards Casualty Severity which achieved accuracy 88.02%, and lowest classifier accuracy is Naïve Bayes which achieved accuracy 86.35% as shown in Table II and Fig. 4 and 5.

TABLE II. TRAFFIC MANAGEMENT MODEL LEEDS UK DATASET RESULTS

Leeds UK (Casualty Severity)						Leeds UK (Casualty Class)				
Algorithm	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**
Random Forest	87.12%	0.823	0.871	0.871	0.786	87.79%	0.88	0.875	0.875	0.126
Decision Tree	88.02%	0.88	0.88	0.88	0.22	87.75%	0.883	0.877	0.877	0.127
Naïve Bayes	86.35%	0.816	0.863	0.863	0.776	79.45%	0.792	0.795	0.795	0.189
SVM	87.81%	0.878	0.878	0.878	0.122	77.38%	0.7738	0.7738	0.774	0.226
Logistic Regression	86.85%	0.86846	0.86846	0.868	0.132	75%	0.75	0.75	0.75	0.25

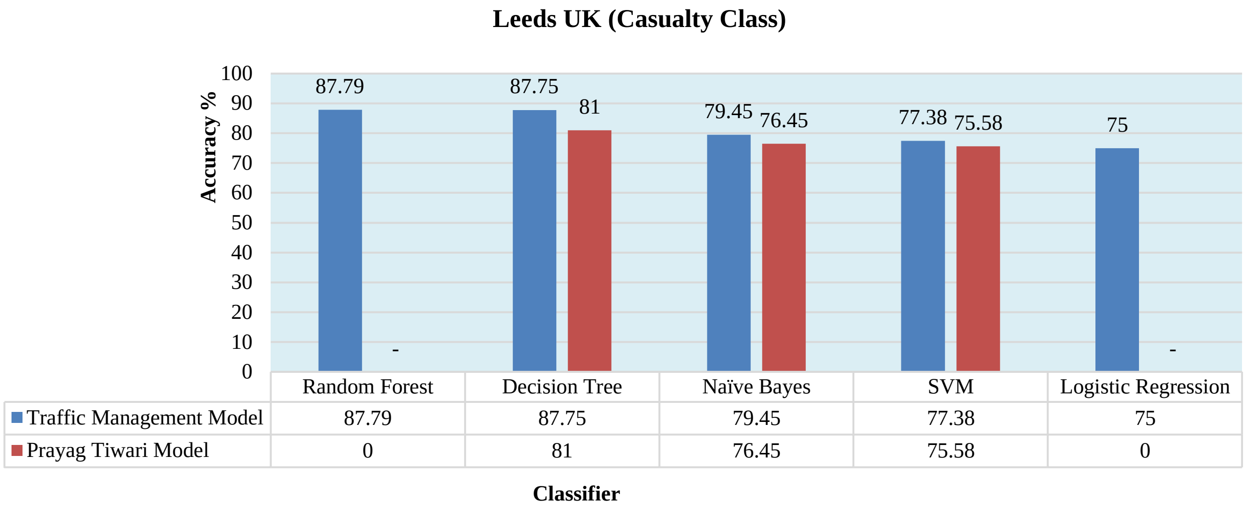

Fig. 4. The Accuracy Chart results between Traffic Management Model and Prayag Tiwari's Research Work.

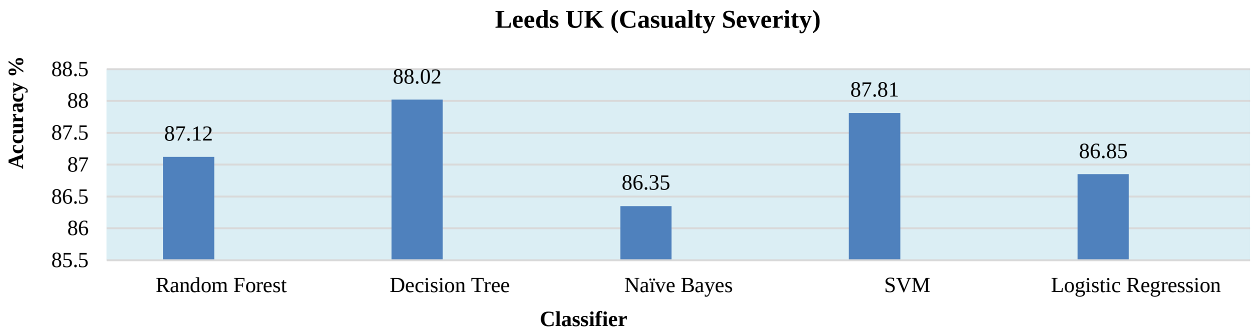

Fig. 5. The Accuracy Chart results in Leeds UK (Casualty Severity Classification).

B. Maryland U.S. Dataset (Speed Limitation)

Speed limit information is found to be very valuable in predicting crash occurrence. However, speed limits may have biased coefficients, most likely attributable to unobserved safety-related effects. So, it is highly significant to define the attributes which affect the speed limitation such as (Light Condition, Road Junction, Road Surface Condition, Road structure Condition, Weather Condition, Vehicle Characteristics, Geographical Location, and other attributes which may enhance the accuracy in this model). In these studies, using Maryland U.S. Dataset 5 classifier algorithms applied to define the most powerful algorithm in Speed Limitation and the Decision Tree Regression is the best algorithm to define the speed limit which achieved 98.93% accuracy and the lowest accuracy achieved by Logistic Regression with accuracy 63.91% as shown in Table III and Fig. 6.

TABLE III. MARYLAND U.S. DATASET (SPEED LIMITATION)

Maryland U.S. (Speed Limitation)					
Algorithm	Accurac y	Precisio n	Reca ll	TP Rate	FP Rate
Random Forest	88.24%	0.88235	0.882	0.882	0.118
Decision Tree Classification	90.59%	0.90588	0.905	0.906	0.094
SVM	82.35%	0.82353	0.823	0.824	0.176
Decision Tree Regression	98.93%	0.9893	0.989	0.989	0.011
Logistic Regression	63.91%	0.63905	0.639	0.639	0.361

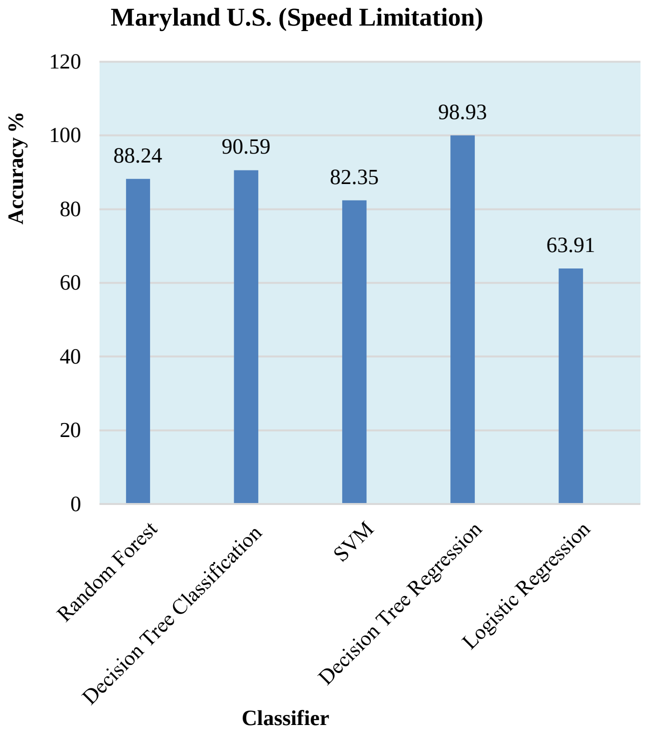

Fig. 6. The Accuracy Chart results in Maryland U.S. (Speed Limitation).

C. Maryland U.S. Dataset (Casualty Class)

According to these practical studies in Maryland U.S. Datasets using more attributes than defined in Leeds UK Datasets to achieve more accuracy and already better accuracy achieved based on Casualty Class so, the Specialists can see clearly that what circumstances affect and who is involved more in an accident between the driver, passenger or pedestrian, and the better classifier in this work-study is Decision Tree Classification which achieved accuracy 89.38% better than the others classifiers used, and lowest classifier accuracy is Logistic Regression which achieved accuracy 83.08% as shown in Table IV and Fig. 7.

TABLE IV. MARYLAND U.S. DATASET RESULT (CASUALTY CLASS)

Maryland U.S. (Casualty Class)					
Algorithm	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**
Random Forest	88.54%	0.8854	0.8854	0.885	0.115
Decision Tree Classification	89.38%	0.89377	0.89377	0.894	0.106
SVM	86.65%	0.86646	0.86646	0.866	0.134
K-Nearest Neighbor	84.90%	0.849	0.849	0.849	0.151
Logistic Regression	83.08%	0.8308	0.8308	0.831	0.169

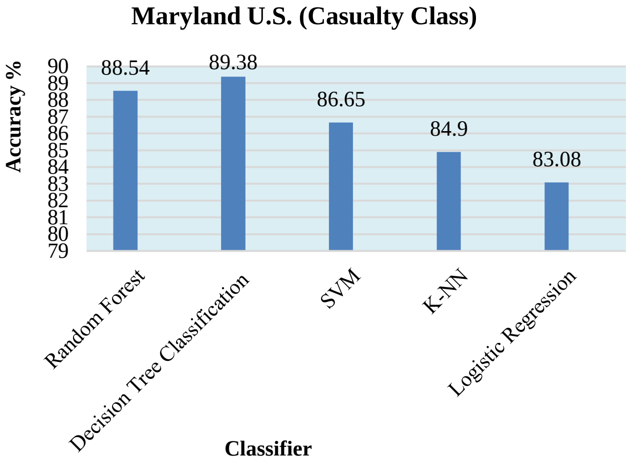

Fig. 7. The Accuracy Chart results in Maryland U.S. (Casualty Class).

D. Maryland U.S. Dataset (Destruction Severity Class)

The accident severity classification is a significant factor in traffic management especially in Emergency Medical Service to define the type of rescue tools and the fast response for support to the Casualty victims. On the other side the accident severity classification has an important impact on road traffic congestion and the more serious the accident, the greater the traffic congestion and it supports the traffic management authorizers for urgent decision making such as preparing alternative routes or preparing the emergency roads. In these studies, using Maryland U.S. Dataset five classifier algorithms applied to define the most powerful algorithm in

Crash Severity Classification and the Random Forest is the best algorithm to define the Destruction severity Class which achieved 98.63% accuracy and the lowest accuracy achieved by Logistic Regression with accuracy 80.8 % as shown in Table V and Fig. 8.

E. *Maryland U.S. Dataset (Accident Occurrence)*

Predicting the occurrence of crashes helps study safety traffic planning and make improvements in the traffic elements. Real-time crash prediction helps identify and prevent crashes before they happen. Therefore, it has become a hot topic in the ITS industry, both crash and non-crash events are needed in crash prediction [15]. So, the proposed traffic management model applied in Maryland U.S. datasets to define the factors which contributed to the accident in this study four classifier algorithms applied to define the most powerful algorithm in Accident Occurrence prediction and the better classifier in this work-study is SVM which achieved accuracy 98.84% better than the other classifier used, and lowest classifier accuracy is K-Nearest Neighbor which achieved accuracy 98.26% as shown in Table VI and Fig. 9.

TABLE V. MARYLAND U.S. DATASET RESULT (DESTRUCTION SEVERITY CLASS)

Maryland U.S. (Destruction Severity Class)					
Algorithm	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**
Random Forest	98.63%	0.98634	0.98634	0.986	0.014
Decision Tree Classification	98.48%	0.9848	0.9848	0.985	0.015
SVM	89.61%	0.89605	0.89605	0.896	0.104
K-Nearest Neighbor	94.76%	0.94764	0.94764	0.948	0.052
Logistic Regression	80.80%	0.80804	0.80804	0.808	0.192

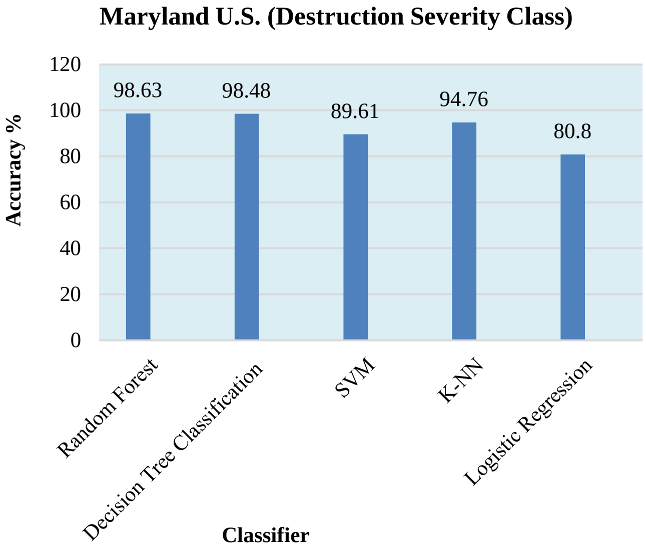

Fig. 8. The Accuracy Chart results in Maryland U.S. (Destruction Class).

TABLE VI. MARYLAND U.S, DATASET RESULT (ACCIDENT OCCURRENCE)

Maryland U.S. (Accident Occurrence)					
Algorithm	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**
Random Forest	98.53%	0.9853	0.9853	0.985	0.015
Decision Tree Classification	98.71%	0.9871	0.9871	0.987	0.013
SVM	98.84%	0.9884	0.9884	0.988	0.012
K-Nearest Neighbor	98.26%	0.9826	0.9826	0.983	0.017

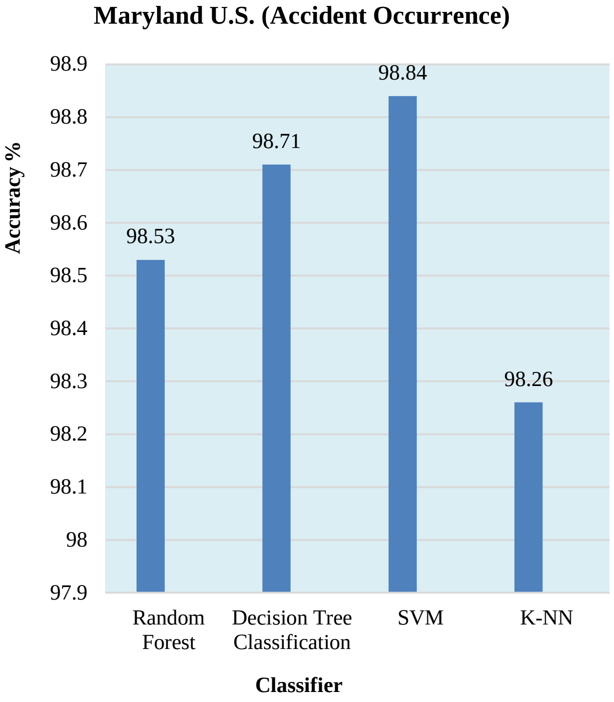

Fig. 9. The Accuracy Chart results in Maryland U.S. (Accident Occurrence).

VII. CONCLUSION AND FUTURE WORK

In these practical studies, there are several Text Mining approaches has been performed to analyze different real datasets which supported by governmental agencies to be trusted data source such as Leeds UK, using different classification and machine learning techniques. One of the contributions is this work-study outperforms Prayag Tiwari's Work-study and better results achieved using more classifiers with higher accuracy from this way based on casualty class and casualty severity so the Specialists can see clearly that what circumstances affect and who is involved more in an accident between the driver, passenger, or pedestrian. And to which degree the victim was affected by the accident. On the other side using Maryland U.S. datasets the previous model applied to classify these datasets into traffic management characteristics needed to define the aspects which have an important role in traffic management processes such as (Accident Casualty Class, Accident Severity Class, Speed Limitation across Geographical Locations, Accident Destruction Class, and Accident Occurrence prediction). Multi algorithms applied to get the optimum classifiers used in this model such as (Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Logistic Regression, and Naïve Bayes). To improve the chances of survival for victims involved in road accidents, it is significant to reduce the response time of rescue teams and to optimize the medical and rescue resources needed. So, the proposed Emergency Medical Service (EMS) once the accident has occurred, it is crucial to managing emergency rescue and resources efficiently and quickly. In future work, the researcher recommends establishing a framework that saves all data to study the usual behavior for each user to further analysis to be used in IoT road traffic applications.

ACKNOWLEDGMENT

I sincerely thank all those who helped me in completing this task, especially the participant supervisors.

REFERENCES

[1] 1Suraj Kumar G Shukla, 2Aadithya Kandeth, 3D. Sai Santhiya, 4Kayalvizhi Jayavel SRM Institute of Science & Technology, Kattankulathur, Chennai “Efficient Traffic Management System” International Journal of Engineering &Technology, 7 (3.12) (2018) 926 -932.

[2] Luci Sumi, and Virender Ranga. National Institute of Technology, Kurukshetra “Sensor enabled Internet of Things for Smart Cities” 2016 Fourth International Conference on parallel, Distriputed, and Grid Computing (PDGC). Conference Paper · December 2016. DOI: 10.1109/PDGC.2016.7913163.

[3] Deepti Goel, Santanu Chaudhury, and Hiranmay Ghosh. 2017. “An IoT Approach for Context-aware Smart Traffic Management Using Ontology”. In Proceedings of WI ’17, Leipzig, Germany, August 23-26, 2017, 8 pages. DOI: 10.1145/3106426.3106499.

[4] Dovydas Skrodenis, Vilnius Gediminas Technical University, Road Research Institute Saulėtekio al. 11, 10223 Vilnius, Lithuania “Road Traffic Management During Special Events” Conference paper, First Online: 18 June 2019 pp 104-109.

[5] Sasan Amini, Eftychios Papapanagiotou, and Fritz Busch, Chair of Traffic Engineering and control, Technical University of Munich “Traffic Management for Major Events” Digital Mobility Platforms and Ecosystems. DOI: 10.14459/2016md1324021.

[6] Sonali Vijay Gaikwad, Archana Chaugule, and Pramod Patil, “Text Mining Methods and Techniques”. International Journal of Computer Applications (0975 – 8887), Volume 85 – No 17, January 2014.

[7] Karlaftis M, Tarko A (1998) “Heterogeneity considerations in accident modeling”. Accid Anal Prev 30(4):425–433.

[8] Lin Xu*, Yang Yue, Qingquan Li, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, “Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data”. 13th COTA International Conference of Transportation Professionals (CICTP 2013), Procedia - Social and Behavioral Sciences 96 (2013) 2084 – 2095.

[9] Guo, Y., Graber, A., McBurney, R.N., Balasubramanian, R., 2010. “Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms”. BMC Bioinformatics 11 (447), 19p.

[10] Theoflatos, A., Chen, C., Antoniou, C., 2019. “Comparing machine learning and deep learning methods for real-time crash prediction”. Transportation Res. Record 1–10.

[11] Poch, Mark, and Fred Mannering. "Negative binomial analysis of intersection-accident frequencies." Journal of transportation engineering 122.2 (1996): 105-113.

[12] Karlaftis M, Tarko A (1998) “Heterogeneity considerations in accident modeling”. Accid Anal Prev 30(4):425–433.

[13] Oh HoonKwon, WonjongRhee·, YoonjinYoon, “Application of classification algorithms for analysis of road safety risk factor dependencies”. Accident Analysis & Prevention, Publisher: Elsevier, Date: February 2015. DOI: 10.1016/j.aap.2014.11.005.

[14] So Young Sohn , Sung Ho Lee “Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea”. February 2003 Safety Science 41(1):1-14, DOI: 10.1016/S0925-7535(01)00032-7.

[15] Junhua Wanga, Tianyang Luoa, Ting Fua,b,∗ “Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach”, ON, N2L 3G1, Canada, ELSEVIER.

[16] Ali S. Al-Ghamdi, “Using logistic regression to estimate the influence of accident factors on accident severity” [J]. Accident Analysis and Prevention, 2002: 729-741.

[17] Sunanda Dissanayake, Jian John Lu. “Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes” [J]. Accident Analysis and Prevention, 2002: 609-618.

[18] Yu, R., Abdel-Aty, M., 2014. “Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data”. Saf. Sci. 63, 50–56.

[19] Oana Victoria Oțăt University of Craiova, Faculty of Mechanical Engineering, 107 Calea Bucuresti Str., Craiova, Romania, otatoana@yahoo.com “traffic management training via dedcated software“ The European Proceedings of Social & Behavioural Sciences EpSBS, ISSN: 2357-1330 , https://doi.org/10.15405/epsbs.2019.08.03.184.

[20] 1Syed Misbahuddin, 2Junaid Ahmed Zubairi, 3Abdulrahman Saggaf, 4Jihad Basuni, 5Sulaiman A-Wadany and 6Ahmed Al-Sofi, “IoT Based Dynamic Road Traffic Management for Smart Cities”, Fredonia NY 14063 Conference Paper · December 2015.

[21] Sagarika Verma, Sayali Badade Computer Science and Engineering, MIT-ADT University, Pune, India. “Traffic Prediction Using Machine Learning” Proceedings of National Conference on Machine Learning, 26th March 2019. ISBN: 978-93-5351-521-8.

[22] Ramzan Talib∗, Muhammad Kashif Hanify, Shaeela Ayeshaz and, Fakeeha Fatimax “Text Mining: Techniques, Applications and Issues” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.

[23] Vinayak Hegdea, Shruthi G. Kinia, and Sahana A, “A Comparative Study of Principal Component Analysis vs Wrapper Method, an Overview of Dimensionality Reduction Techniques Applied in Developing an Undergraduate Student Dropout Model”, International Journal of Control Theory and Applications, Volume 9 • Number 42 • 2016, ISSN: 0974–5572.

[24] Prayag Tiwari University of Padova” Accident Analysis by using Data Mining Techniques” Thesis ·June 2017 DOI: 10.13140/RG.2.2.20091.41766/.

[25] Prayag Tiwari University of Padova, Sachin Agnihotri South Ural State University, Denis Kalitin National University of Science and Technology MISIS, “Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques”, Conference Paper · September 2017, DOI: 10.1007/978-981-10-6430-2_31.