ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

An Advanced Approach of Text Mining for Analysis of Road Accidents

Disha Jain[1], Sitanshu Jain[2]

M.Tech Student, Dept. of Computer Science and Engineering, Gyan Ganga College of Technology, Jabalpur, India[1]

Assistant Professor, Dept. of Computer Science and Engineering, Gyan Ganga Institute of Technology & Sciences, Jabalpur, India[2]

ABSTRACT: Road accidents have been increasing day by day and there seems to be no solution for it. Millions of people lose their valuable life through road accidents. Data and statistics are of immense importance as it can give insights into what are the real factors causing these dreadful accidents and how we can avoid the same. In this research paper, we have worked upon a domain called as Data mining in which we uncover hidden patterns and useful information which can be used to understand what are the causes of these road accidents and what could be the solutions for the same. In this paper, we have applied different classification algorithms to understand the causes of road accidents happening around the world. As this seems to be a very serious issue, we have collected data from wide range of sources ranging from newspapers to articles, forums to open source data. After applying data mining algorithms, we uncovered some patterns and tried to understand the real causes of road accidents and some solutions for the same. The insight we have gained could be immensely useful for government, NGO's and general public for understanding the issue of road accidents. For implementation of our research work, we have used a very popular data mining tool called as RapidMiner. The tool can be downloaded from the Internet for free of cost and is an open source program. There are several branches of Data mining, here we have applied Data mining algorithms in Text, which is called as Text Mining. After going through thousands of records of data we have concluded our research work with the results of causes and solutions for road accidents.

KEYWORDS: Road accidents, Data Mining, RapidMiner, Naive Bayes Classifier.

I. INTRODUCTION

There is nothing more valuable than a human life. Road accidents has killed more people than anything else in the world. There are millions of people being killed in road accidents around the world every year. In this dissertation work, the focus is on the causes of road accidents as to what are the reasons that cause these dreadful accidents and how can we lead to a solution which can limit or minimize these accidents. In this work, we have taken thousands of road accidents records which will be useful in understanding the major causes of road accidents and how those can be resolved. There are various causes because of which the accidents are caused. From Over speeding to Drink and Drive, we have tried to cover all the causes which are somehow involved in road accidents. To understand the patterns of road accidents we need Data; and to gain some useful information from Dataset, there is a concept called as Data Mining[1].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

a. Data Mining

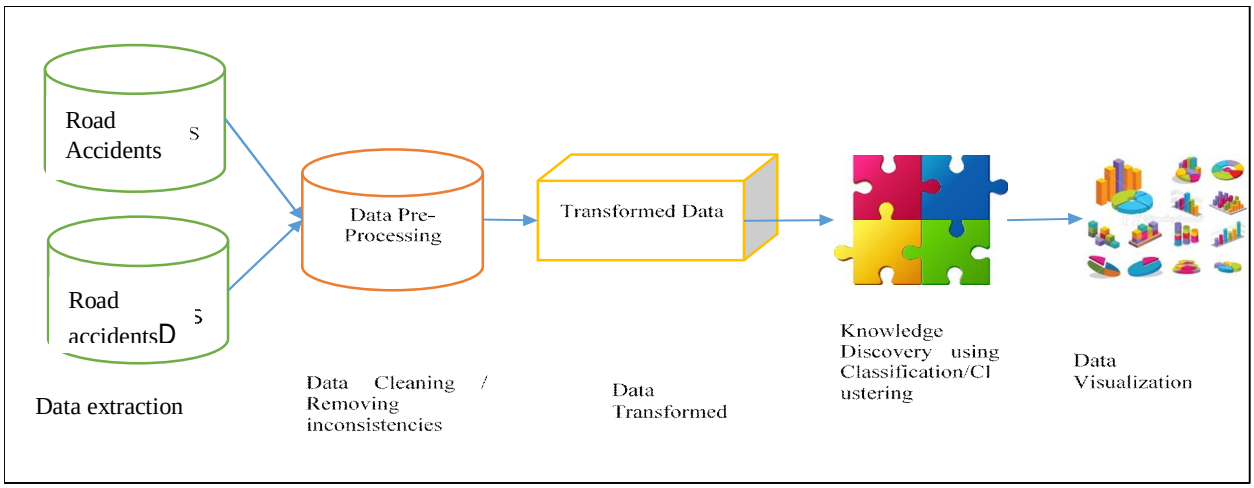


Figure 1: Data Mining Process

The process of Data mining can be seen in the above Diagram. This dissertation work is based on Data Mining domain. In Data mining, the large amount of dataset is taken (more than 5 years, 10 years of data) and we try to uncover some useful and hidden information from the dataset that will be useful in various ways. Firstly, the data is taken in raw form which is huge and usually taken from different sources. This causes noise, inconsistencies and missing values in the data. It is very important to remove these inconsistencies from the data. This is the first step in Data mining process [2]. The inconsistencies are removed and data is made free from noise. This process is called as pre-processing of the data. Now the data is free from all the inconsistencies. Next step is to transform data into a common format which makes it an appropriate form for applying data analytics and data mining. After this step, we apply different data mining techniques such as Classification, Clustering and Rule Association Mining to extract hidden and useful information from the huge dataset. Once we have the hidden information, we can visualise the same using Pie-chart, Bar-chart, etc. The whole process is also termed as Knowledge Discovery.

There are several branches of Data Mining such as Text Mining, Image Mining, Video Mining, etc. In this work, the focus is on Text mining where we have tried to extract hidden patterns from text [3]. The text is the road accident description and details which is taken from various different sources.

There are several sources for road accidents but we must choose the authentic one. We have taken data from different sources such as News websites which are verified and backed by sources, verified forums, Data.gov.in which is the open data government website and offline directly from the people. We accumulated data from these authentic sources and stores the records in text files and excel sheets. In this work, we have applied Data Mining Classification technique where the different causes of road accidents are classified into different classes[4]. The simplest example to understand Classification technique is the E-mail service we use. The mails we receive either go to “inbox “or “spam” based on certain keywords. These are the two classes and the system are trained to decide which class a new mail belongs to on the learning it has received. In the same way, we have classified the causes of road accidents into different classes which we found in the research were the most common and major causes of road accidents happening around the world. To implement Classification, we have used Naive Bayes Classifier which is one of the most popular and efficient classification algorithms. Data Mining task is implemented in a Data mining software. We have implemented this dissertation work in Rapidminer, which is an open source software platform which can be downloaded and used for free from the Internet. It has some very powerful tools which makes the task easy and also provides fast results. The research work is implemented using RapidMiner where the system makes predictions as to which class a particular road

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

accident belongs to. There are various causes of road accidents which we have uncovered in this research work. We have taken 3000+ records for performing Data analytics and predictive analysis.

II. LITERATURE REVIEW

Dr. S. Vijayarani et. al proposed a research work, "ANALYSIS OF ROAD ACCIDENTS ININDIA USING DATA MININGCLASSIFICATION ALGORITHMS" in which the authors have performed an analysis on road accidents in India[5]. Different classification algorithms have been applied on Open Government dataset to classify accident causes into different classes. The limitation in this work is that no Training has been provided to the system where it can make predictions with any new dataset. Also, in this work only three performance parameters have been used, where we have used 5 performance parameters in our research paper.
No implementation details have been mentioned in the research paper with any real time data.

Maolei Zhang et. al proposed a research paper on, "METHOD OF ROAD TRAFFIC ACCIDENTS CAUSES ANALYSIS BASED ON DATA MINING." In this paper, the authors have performed research work on causes of road accidents[6]. No training or testing has been done on real time data. Causes Support Model have been applied to understand causes of road accidents. There are different classification algorithms such as K-NN, Naive Bayes, which are not been applied to understand the real causes of accidents. No performance parameters have been applied to check how well and correct the system is performing.

Ms. Kaur et. al proposed a research paper on, "PREDICTION OF THE CAUSE OF ACCIDENT- AND ACCIDENT-PRONE LOCATION ON ROADS USING DATA MINING TECHNIQUES." In this paper, the authors have applied correlation analysis and exploratory visualization techniques to understand the causes of road accidents in Rajasthan State of India [7]. In this paper, only clustering has been applied and no real causes has been uncovered for road accidents. The authors have focused on the type of road in State highways and districts which have been concluded to be in bad condition and being the causes for road accidents in state. There have been no performance measures taken in this research work.

V. Sakhare et.al proposed a research work on, "A Review on Road Accident Data Analysis Using Data Mining Techniques." In this work, the authors have done a research work on road accidents using different data mining techniques [8]. Only survey has been done and no implementation has been performed using any data mining tool. K-means Clustering has been applied to cluster different causes of accidents. No real time data has been used. Also, the accuracy achieved is low [9].

III. METHODOLOGY

a. Naive Bayes classifier:
Naive Bayes classifier, one of the most popular and largely used classification algorithm is used in our research work for implementation[10]. This classifier is based on Bayes theorem which is a probability theoremi.e. the classification prediction is done on the basis of probabilities. There are different classes defined in the dataset and the classifier predicts the class for objects. This prediction is made using probability. The class having higher probability is predicted as the target class for a particular object. The objects are given as the inputs to the Naive Bayes classifier and it gives the output as the predicted class for the unknown object[11][12]. The name is given as "Naive" because this classifier assumes that the features of objects in one class are independent of any other features in the dataset. That is, one particular feature is totally unrelated to other features. Naive Bayes classifier is not only limited for binary classification i.e. only two classes prediction[13][14]. It is widely accepted in making predictions/ classifications for multiple classes.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

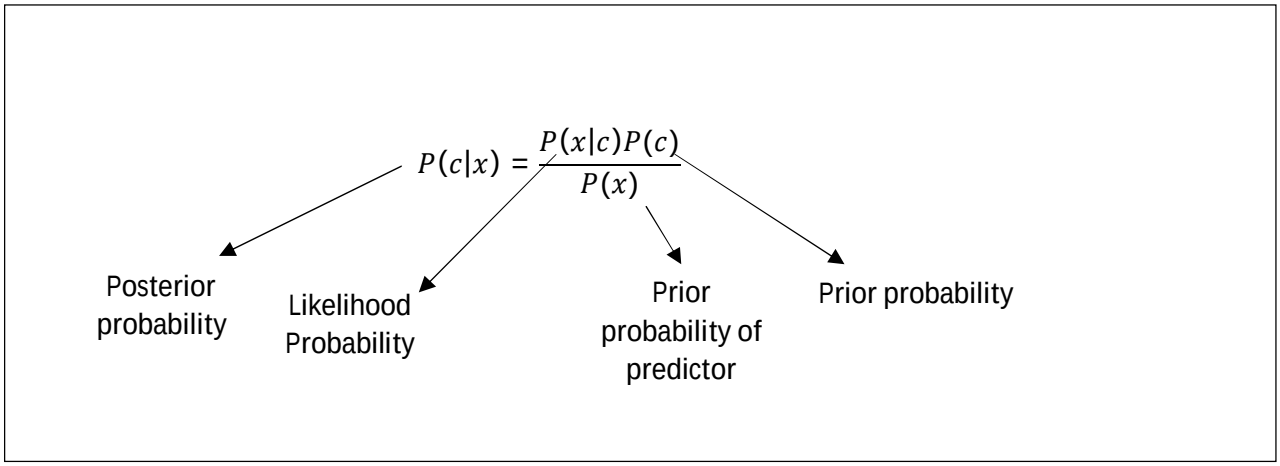
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

b. Equation:
The equation for Naive Bayes is as follows [15]: -



The above equation has following parameters:
P(c|x) denotes the posterior probability which is the result and output of the equation.
P(c) is the aprior probability which we already know. This is also called as known probability for class.
P(x|c) denotes the likelihood probability where class is denoted as "C" and features are denoted as "X".
P(x) denotes the aprior probability of feature 'x'.
We apply the above equation and perform calculation for each target class [16]. There are several classes defined in this dissertation work of road accident causes. Posterior probability is calculated for all the classes and the one with highest value is predicted as the target class for a particular object [17].
There are several software tools for Data mining work. In this research work, **Rapidminer**has been used which is an open source software for performing data analytics and data mining[18]. It needs no purchase for research work and is free for non-commercial use. It is one of the widely used software in the world because of its performance, accuracy and providing fast results. It uses drag and drop operations for data mining work which gives simplicity in usage for non-professionals and non-programmers.
There are various operators which can be used for different data mining tasks. From Market-basket analysis to Clustering, Classification, Association rule, this tool can implement all the techniques.

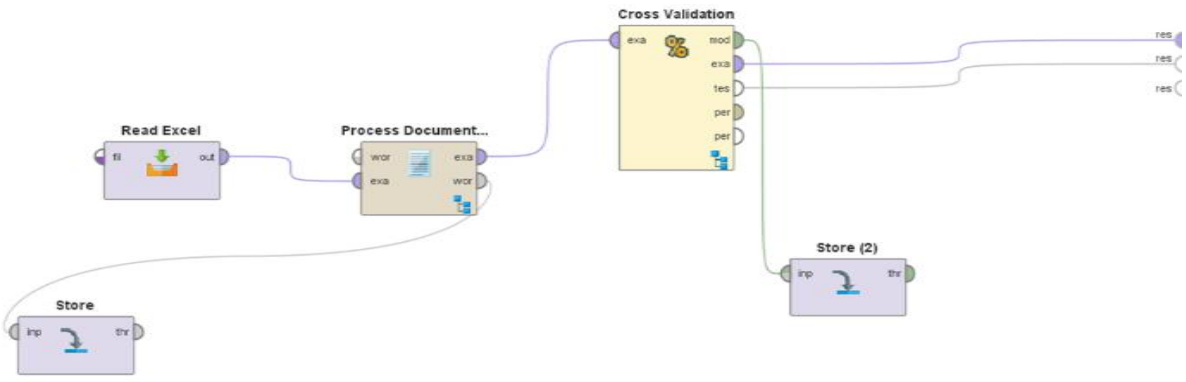


Figure 2 : Rapid Miner main process

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

The above diagram shows the main process in RapidMiner which is created for implementation of the research work. Process is created to give input, add all the operators and produce output. There are different operators used such as Read Excel is used for reading excel files where the dataset is stored. Process documents from files is used to remove inconsistencies from the data. Store operator is used to store the worldlist and model which will be further used in testing process. Validation operator is used to check the performance of the system i.e. how correct the predictions are being made by the system.

a. TRAINING DATASET

For making predictions, the system must be trained. With 'training' the meaning is to give some example records to the system in RapidMiner and let the system train/learn from the dataset. Thousands are records are given with the known output. The system already knows what will be the output for the record. This is termed as training. There are different causes for road accidents which we encountered; the system is trained for all the road accidents causes so that the system can be trained and can make correct predictions. In this research work, we have taken 3000+ records of road accident data in an excel sheet. There are two columns defined: First is the text of road accident and second one is the label class which defines the cause for road accident. The following figure shows how the system is trained:

ACCIDENT DETAILS	ACCIDENT CAUSE
Thirty killed in bus accident in Himachal Pradesh's Kangra, over-speeding caused accident	OVERSPEEDING
10 killed after overspeeding auto falls into well in Nizamabad	OVERSPEEDING
Overspeeding Truck Kills Eight people in Nagaland	OVERSPEEDING
Wrong side overtaking turns fatal for carpassenger	OVERTAKING
Driver killed roadside beggers as he was heavy on alcohol	DRINK N DRIVE
Ant McPartlin pleads guilty to drink driving after car crash	DRINK N DRIVE
Ajay was reported to have been hit while he was talking on his mobile phone	TALKING OVER MOBILES
Rain, slippery roads lead to multiple Highway 17 accidents in U.P	BAD ROADS
Two killed on Yamuna e-way as mini-truck rams truck because of tyre burst	BAD ROADS
BIKE flipped in overspeeding in U.P benaras	OVERSPEEDING
Car flips over in over-speeding accident in M.P, driver injured	OVERSPEEDING
Overspeeding car flipped and killed 15	OVERSPEEDING
Overspeeding bike flipped and killed 2 people in Hyderabad	OVERSPEEDING
Truck overspeeding killed 7 people in Lucknow	OVERSPEEDING
Overspeeding minibus kills four including two months baby	OVERSPEEDING
Overspeeding BUS killed four including two months baby	OVERSPEEDING
Overspeeding Truck kills four including two months baby	OVERSPEEDING
speeding truck collided with three other cars on National Road No.22 killed 6 people	OVERSPEEDING
overspeeding bus collided with car and killed 5 people	OVERSPEEDING
5-year-old killed after being hit by speeding van	OVERSPEEDING
5-year-old killed after being hit by speeding truck	OVERSPEEDING
5-year-old killed after being hit by speeding bus	OVERSPEEDING
5-year-old killed after being hit by speeding bike	OVERSPEEDING
5-year-old killed after being hit by speeding scooter	OVERSPEEDING
10 year-old killed after being hit by speeding van	OVERSPEEDING
10 year-old killed after being hit by speeding truck	OVERSPEEDING
10 year-old killed after being hit by speeding bus	OVERSPEEDING
10-year-old killed after being hit by speeding bike	OVERSPEEDING

Figure 3: Training Dataset

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

a. TESTING DATASET

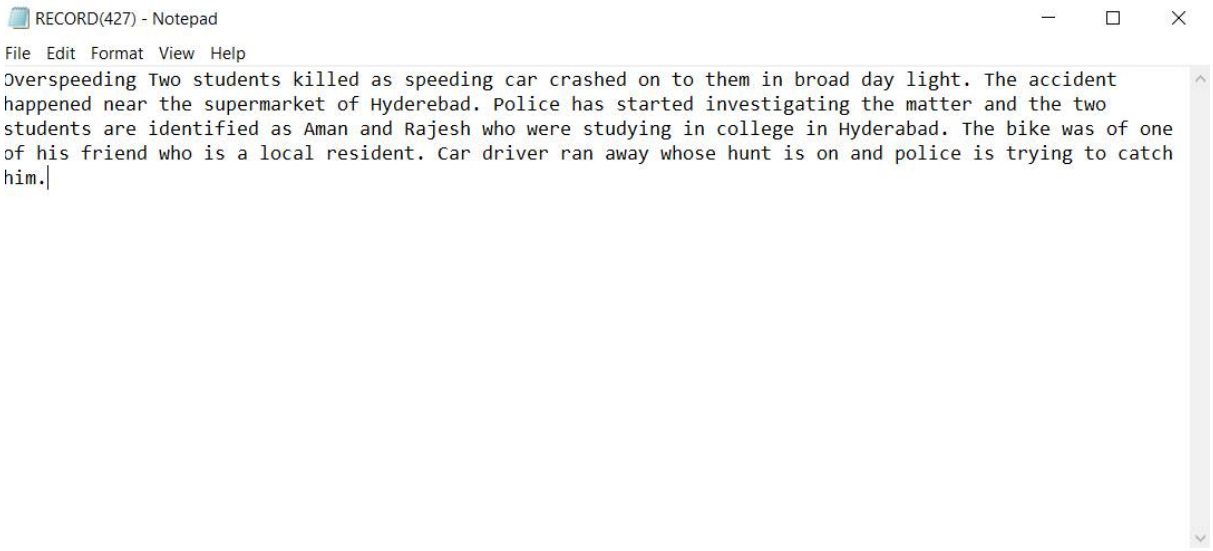


Figure 4 Testing dataset

Once the training is completed, we use Testing dataset. In this dataset, we have taken 2000+ records of road accidents from online news websites. Through this dataset we check how correct the system is making predictions and how correctly the system is trained. Testing is performed on totally new and unknown objects. The road accident data for testing dataset is stored in different notepad text files as shown above in the snapshot. This is very helpful in performing data analysis and gaining hidden information as to what are the real causes which are causing so many road accidents every day around the world.

a. DATA SOURCE: NEWSPAPERS, FORUMS, GOVERNMENT DATA.

The most important thing in Data mining is Data. We have taken data from different sources for performing data mining. These sources are newspapers where there is news about road accidents happening around the world. We have also taken data offline directly from the people who faced road accidents. Some other sources are Online news, forums, government data from data.gov.in website. The data is collected from different sources and merged into one for performing data analytics and understand the real cause of road accident.

IV. RESULT & ANALYSIS

The methodologies are applied to get the desired results. In this research work, the data of Road accidents is taken from different sources such as News, Forums, Portals, etc. 3000+ records are taken to check how correct and accurate the system is performing the predictions. The system predicts the class for different causes of Road accidents. These are some common causes for road accidents which are the major reasons for causing road accidents across the globe. The road accident data is stored in the excel file. The system is trained for predicting the causes which are OVERSPEEDING, OVERTAKING, DRUNKEN N DRIVE, TALKING OVER MOBILE, BAD ROADS. The implementation is done using RapidMiner. We start a new process in RapidMiner and connect all the input ports to output ports. After connecting all the ports, the process is run to get the desired result:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

1	ACCIDENT DATA	OVERTAKING	0	1	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (122...
2	ACCIDENT DATA	TALKING OVER MOBILES	0	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (123...
3	ACCIDENT DATA	OVERTAKING	0	1	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (124...
4	ACCIDENT DATA	OVERTAKING	0	1	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (125...
5	ACCIDENT DATA	OVERTAKING	0	1	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (126...
6	ACCIDENT DATA	OVERSPEEDING	0.991	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (128...
7	ACCIDENT DATA	TALKING OVER MOBILES	0	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (129...
8	ACCIDENT DATA	OVERTAKING	0	1	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (130...
9	ACCIDENT DATA	TALKING OVER MOBILES	0	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (131...
10	ACCIDENT DATA	TALKING OVER MOBILES	0	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (132...
11	ACCIDENT DATA	OVERTAKING	0	1	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (133...
12	ACCIDENT DATA	TALKING OVER MOBILES	0	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (134...
13	ACCIDENT DATA	TALKING OVER MOBILES	0	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (135...
14	ACCIDENT DATA	OVERTAKING	0	1	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (136...
15	ACCIDENT DATA	TALKING OVER MOBILES	0	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD (137...
16	ACCIDENT DATA	OVERSPEEDING	0.991	0	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD(1).txt
17	ACCIDENT DATA	OVERTAKING	0	1	0	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD(10).txt
18	ACCIDENT DATA	DRINK N DRIVE	0	0	1	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD(100)...
19	ACCIDENT DATA	DRINK N DRIVE	0	0	1	. . . . .	C:\Users\Win- 10\Desktop\Disha Thesis\t\TESTING DATASET\RECORD(100...

figure5: RapidMiner results for Road Accidents**.**

The above figure shows the result once we run the RapidMiner process. The system has correctly predicted the class for different accident details. When there was news for road accident as “Speeding vehicle killed 5 people”, the system correctly predicted the class as “OVERSPEEDING”. As the system is trained correctly, this prediction is automatically done by the system and no manual effort is needed.

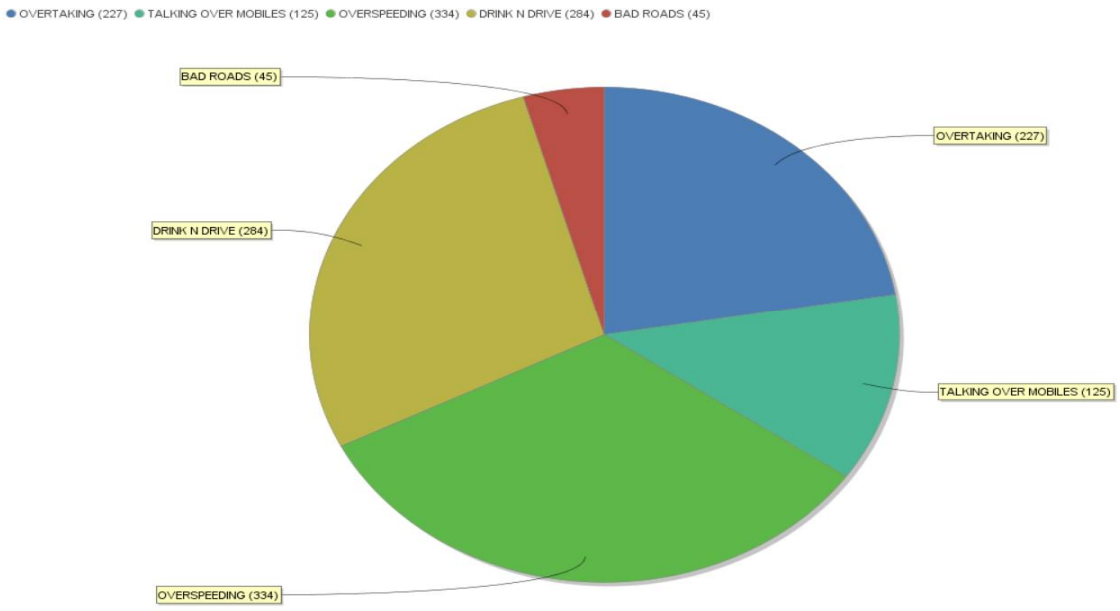


Figure 6: Result Pie Chart

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

From the records of 1000+ records we get following as results: OVERSPEEDING caused 334 no. of accidents which is 33.4%. OVERTAKING has caused 227 number of accidents that is 22.7%.
BAD ROADs accounts to 45 no. of accidents which is 4.5%. DRINK AND DRIVE has 284 of accidents which is 28.4%. TALKING OVER MOBILES has caused 125 of accidents with 12.5%. Here we have developed a system which can work with any dataset and perform predictions for different causes of road accidents. We performed a research considering all the major causes of road accidents and found above causes to be the most common one. This is the result of the dissertation work.

V. CONCLUSION AND FUTURE WORK

In this work, we have worked on Road accident data of past 5 years collected from different sources. The data was huge which we pre-processed and applied Data mining techniques to understand and uncover the real causes of road accidents happening around the world. We used RapidMiner in our work which is a Data Mining tool for implementation of this thesis work. We conclude this work with our findings which are the major causes of road accidents. In future, we aim to apply Clustering on our dataset where we could make different clusters on the basis of population of different states i.e. accidents on the basis of population of different states. In future, we also aim to use different and complex data mining algorithms such as IDE, Neural Network and Multiple Support Vector Machine. We also aim to increase our dataset further from 5k records to more than 20k records for performing data analytics.

REFERENCES

1. "Third IEEE International Conference on Data Mining," *Third IEEE International Conference on Data Mining*, Melbourne, FL, USA, 2003, pp.
2. J Larose, Daniel T., “Discovering knowledge in data: An Introduction to data mining”. John Wiley & Sons, 2014.
3. *Xindong Wu, "Data mining: artificial intelligence in data analysis,"* Proceedings. IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2004. (IAT 2004)., *Beijing, China, 2004, pp. 7-*.
4. Liao, Shu-Hsien, Pei-Hui Chu, and Pei-Yuan Hsiao. "Data mining techniques and applications–A decade review from 2000 to 2011." Expert systems with applications 39.12 (2012): 11303-11311.
5. E. Suganya and S. Vijayarani, "Analysis of road accidents in India using data mining classification algorithms," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 1122-1126.
6. R. Tian, Z. Yang and M. Zhang, "Method of Road Traffic Accidents Causes Analysis Based on Data Mining," 2010 International Conference on Computational Intelligence and Software Engineering, Wuhan, 2010, pp. 1-4.
7. G. Kaur and E. H. 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), Delhi, 2017, pp. 1-7.
8. V. Sakhare and P. S. Kasbe, "A review on road accident data analysis using data mining techniques," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2017, pp. 1-5.
9. A. Poliaková, "What is the impact of road tax collection on the accident status due to the fault of the road?," 2018 XI International Science-Technical Conference Automotive Safety, Casta, 2018, pp. 1-5.
10. H. R. Seth and H. Banka, "Hardware implementation of Naïve Bayes classifier: A cost effective technique," 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, 2016, pp. 264-267.
11. Voznika, Fabricio, and Leonardo Viana. "Data Mining Classification." (2007).
12. Rish, Irina, "An empirical study of the naive Bayes classifier." IJCAI 2001, workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. IBM, 2001.
13. Keogh, Eamonn. "Naive bayes classifier." UCR, and Christopher Bishop “Pattern Recognition Machine Learning”, Springer-Verlag (2006).
14. Berend, Daniel, and Aryeh Kontorovich. "A finite sample analysis of the Naive Bayes classifier." Journal of Machine Learning Research 16 (2015): 1519-1545.
15. Dey, Lopamudra, et al. "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier." arXiv preprint arXiv:1610.09982 (2016).
16. Lior, Rokach, “Data mining with decision trees: theory and applications”. Vol. 81. World scientific, 2014.
17. Yong, Zhou, Li Youwen, and Xia Shixiong. "An improved KNN text classification algorithm based on clustering." Journal of computers 4.3 (2009): 230-237.
18. Rapidminer Studio Documentation, http://docs.rapidminer.com/studio.