Volume 6, Issue 2, April 2019 ISSN: 2348-8565 (Online)

IJMERT

International Journal of Modern Engineering and Research Technology

Website: http://www.ijmert.org Email: editor.ijmert@gmail.com

# A Sophisticated Approach of Data Mining for Road Accidents Investigation

**Roopali Sahu**
*Research Scholar*
*Shri Ram Group of Institutions*
*Jabalpur, (M. P.) [INDIA]*
*Email: rupasahui29@gmail.com*

**Jayesh Jain**
*Assistant Professor*
*Department of Computer Science & Engineering*
*Shri Ram Group of Institutions*
*Jabalpur, (M. P.) [INDIA]*
*Email: jayeshjain.it@gmail.com*

**Anupam Choudhary**
*Lecturer*
*Department of Computer Science*
*Kalaniketan Polytechnic College*
*Jabalpur, (M. P.) [INDIA]*
*Email: choudharyanupam7@yahoo.com*

## ABSTRACT

*There has been massive increase in road accidents over the past years. Various research and experts have tried to come up with a solution to solve road accidents issue so that people don't lose their loved ones and their valuable life. Lakhs of accidents are committed in a year all over the world. Even the developed countries have not been able to stop this 7. Developing countries such as ours, India has also tried to stop road accidents happening in the country but still it is growing at a rapid rate. Data mining is an extraction method of useful information where we go through a large amount of data and extract useful and hidden information from vert large datasets. In this paper, we have performed one such branch of data mining called as text mining where we uncover useful information related to road accidents and to ultimately identify the real causes of road accidents so that the same can be avoided and steps can be taken to make sure the accidents don't occur. We have performed a research analysis on the huge road accidents dataset and came up with a conclusion of different causes of road accidents which are found to be the most common one. RapidMiner is a data mining software which we have used to perform implementation of our dissertation work. The research paper contains the introduction following with the methodology where we have used Naïve Bayes classifier to classify different causes of accidents into different classes and understand patterns of road accidents causes. Different data mining algorithms are applied to understand real causes of road accidents. The road accident data has been taken from various sources such as Newspapers, e-papers, Government portals, forums, etc.*

***Keywords:—*** *Road accidents, Data Mining, Text Mining, Classification, Naive Bayes Classifier.*

## I. INTRODUCTION

The most valuable thing in the world is human life. Every year millions of people lose their life because of road accidents and this has been increasing very rapidly over the past few years. Data mining is a process where we extract and discover hidden information and knowledge which can be very useful and can not be seen easily in the dataset. In this paper, we have taken road accidents data from various sources such as newspapers, e-papers, forums, offline connect people to people, etc. The data collected from different sources is merged into a single dataset. Once the data is integrated into a single unit, this huge dataset is given input to pre-processing where all the inconsistencies are removed. After removing inconsistencies. Different data mining algorithms are applied to uncover different causes of road accidents to understand why these accidents occur and how much and which causes are the most common and largely responsible for causing road accidents and ultimately taking life of millions of people. We have also performed a research survey on different causes with its solution and how those solution can be implemented in the real world. Data mining has various branches from Text mining to image mining, video mining etc[1]. In this paper, we have performed implementation on Text mining where we have uncovered useful information from text which is stored in word files and excel files[2].

### *a. Data Mining*

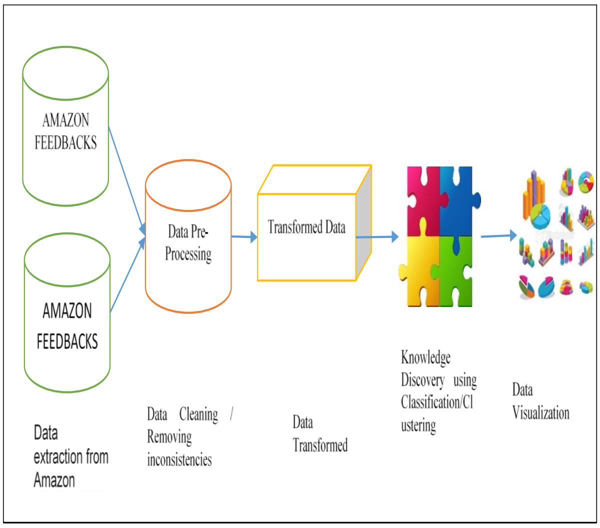

AMAZON FEEDBACKS
AMAZON FEEDBACKS
Data Pre-Processing
Transformed Data
Data extraction from Amazon
Data Cleaning / Removing inconsistencies
Data Transformed
Knowledge Discovery using Classification/Cl ustering
Data Visualization

*Figure 1: Data Mining Process*

The above diagram depicts the data mining process. Firstly, the dataset is collected from different sources. In this case, the data of road accidents is taken from various different sources such as forums, portals, Internet data, data.gov.in, etc. The data of around previous 5 years is taken to understand patterns of road accidents. The collected data is stored and integrated into a single unit. Next step is to take this data and apply pre-processing to clean the data. As the collected data is coming from various sources, there are inconsistencies and noise in the data which must be removed to perform mining in the data and extract knowledge from it. This process is called as pre-processing where all the irrelevant data is removed, noise is removed and missing values are recovered. Next step involves transforming the data into appropriate form for mining. Once the transformation is done, the dataset is applied into classification where the data mining classification techniques are applied to understand patterns and different causes of road accidents. There are other methods such as clustering and rule association algorithms which are also applied to understand knowledge from the huge datasets[3]. Once the knowledge is discovered, the final step is Data visualization where we visualise and analyse data from different angles and different views using Pie charts, Bar charts, etc. In this paper, data mining classification method is applied to uncover different causes of road accidents and are classified into different classes. Naïve Bayes classifier is one of the most popular classifiers is used to perform this classification of road accidents datasets into different classes and causes[4]. RapidMiner platform is used for training and testing (implementation) of the work.

## II. LITERATURE REVIEW

Dr. E. Suganyaet. al proposed a research paper titled as, “Analysis of road accidents in India using data mining classification algorithms” [5]. In this paper, the researchers applied different data mining techniques of classification to make classes of different causes of accidents. The source data is taken from open gov website of Indian Government. The limitation in this work is that the learning and Training is not provided to the system where it can make predictions from past learning and for the new dataset. The performance parameters to check how correct the predictions are used only three in number. Also, no real time data has been used for performing implementation.

Dr. Zhang et. al proposed a research work titled as, “Method of road traffic accidents causes analysis based on data mining[6].” The researchers in this paper have worked on different road accidents causes. Naïve Bayes, K-NN classifiers are applied on datasets to understand patterns of road accidents. The limitation in this work is that no performance parameters are used to check the accuracy of the system.

E.H. Kaur et. al have proposed a research paper titled as, “Prediction of the cause of accident- and accident-prone location on roads using data mining techniques.” This paper contains the work of researchers on causes of road accidents which are uncovered using different correlation analysis techniques and visualization techniques [7]. The dataset is taken from Rajasthan State. Clustering has been applied to make different clusters of road highways in State and districts. Performance Parameters are not applied and no real causes are listed as the condition of roads are blamed for road accident causes.

P. S. Kasbeet. al proposed a research papertitled as, “A Review on Road Accident Data Analysis Using Data Mining Techniques.” In this research paper, researchers have performed analysis on road accidents causes[8]. This paper contains only the survey on different causes and no implementation work has been performed. Accuracy which is achieved is also low. Real time data is also not used or applied.

## III. METHODOLOGY

### *Naive Bayes classifier:*

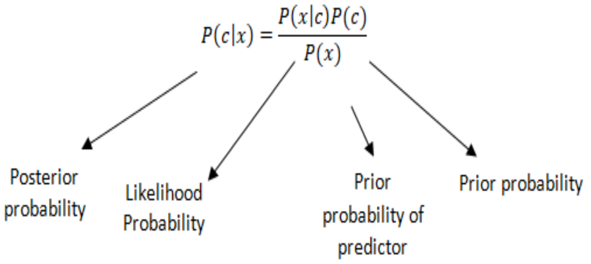
Data mining classification is a popular method for classifying data items and uncover hidden and useful knowledge and information from the datasets. There are several classification algorithms which are used to implement classification [9]. In this paper, we discuss the classifier used in our dissertation work. Naïve Bayes classifier is used in our implementation [10].

Naive Bayes classifier which is one of the most popular and largely used classifier. The research work is implemented using this classifier. The basis for Naïve Bayes classifier is probability theorem where the classification and prediction for target class is done on the basis of probabilities. The probability for different classes is determined and the class with highest probability becomes the target class for the object [11]. Naïve Bayes classifier is first trained against the known output where the objects are given to the classifier as input and the predicted class is the known output to the classifier [12] Once the training is done, the testing is performed where the Naïve Bayes Classifier predicts the class for unknown object. The name is given as “Naïve” to the classifier as the classifier is totally naïve where the features of object in one class are totally independent of features of another class. That is, the features do not dependent on each other and are totally unaware of each other's presence [13]. The class is predicted for the unknown object. The Naïve Bayes classifier is not limited in two class binary classification, it can be used for more than two classes also as used in this dissertation work where we have different road accidents causes as different classes [14]. It has been widely accepted classifier for making accurate and correct predictions.

### *b. Equation:*

The Naïve Bayes classifier equation is[15]: -

These are the following parameters in the equation:


$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$
Posterior probability
Likelihood Probability
Prior probability of predictor
Prior probability

- P(c|x) is a value which represents the posterior probability that represents the output of the equation.
- P(c) is the already known probability called as apriori probability.
- P(x|c) represents the likelihood probability where the class is denoted by "C" and features are denoted by "X".
- P(x) represents feature x, apriori probability.

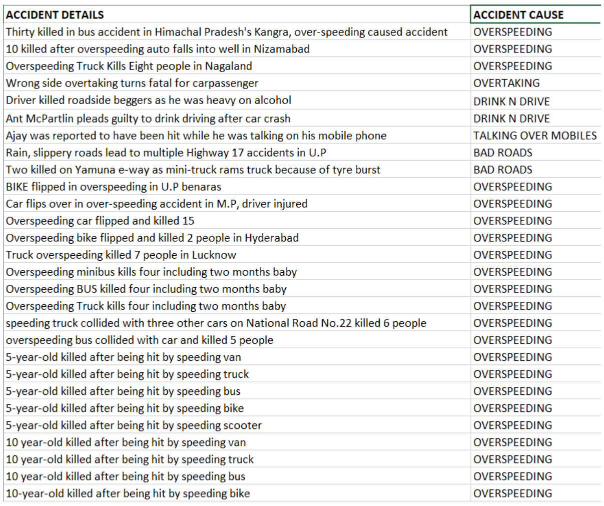
The above equation is applied and the calculation is performed for all the target classes as there are several classes for road accidents representing different causes of road accidents [16]. The Posterior probability is determined for each and every target class and the class having highest posterior probability is the target class for the object whose class is to be determined [17].

Data mining implementation can be done using different software platforms. There are various software platforms such as R, RapidMiner, etc. In this work, we have used RapidMiner software platform for implementation of the research work. RapidMiner is an open source software tool which is widely used for implementing data mining work. It is used by researchers and students for performing data mining work. It can be downloaded for free from the Internet and can be used for performing data mining work implementation. It uses a drag and drop operation for implementing different classification algorithms which can be performed by not only programmers but also non-programmers. It is been widely known for its simple usage, its fast results. Accuracy and open source software. It can be used for variety of data mining tasks such as Classification, Clustering, and Rule Association Mining implementation.

As we can see from the above diagram,a new process is created in RapidMiner which is our main process. This process is created for training the system. There are different operators used from the operator window in the process window for implementation. Read Excel is the operator for using reading excel files. The whole dataset is stored in this excel file. Process documents operator removes inconsistencies from the data. Validation operator checks the performance of the system.

### *C. Training Dataset*

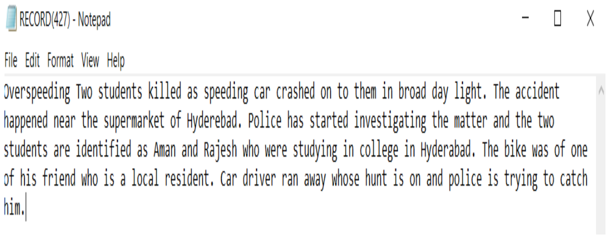
To make system capable for making predictions, the system is trained against the known dataset. The output for the objects is already known and the system gets learning from those known examples. We have taken more than 3000+ records of road accident data to train the system. Two labels are defined where first column defines the accident details and second column specifies the cause of the accident.



| 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*

### *d. Testing Dataset*



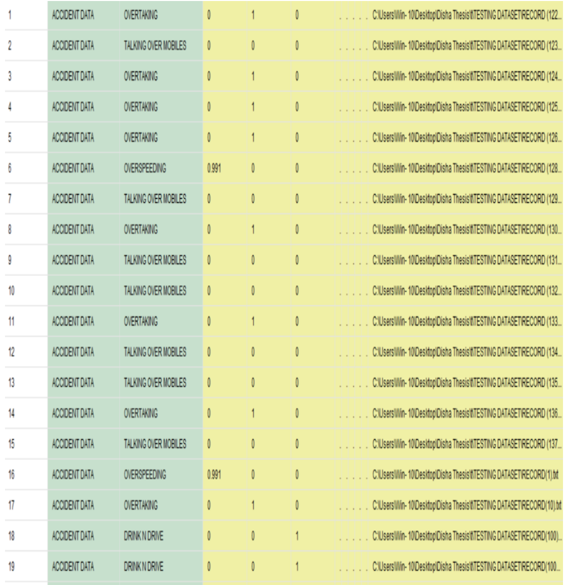
Overspeeding Two students killed as speeding car crashed on to them in broad day light. The accident happened near the supermarket of Hyderebad. Police has started investigating the matter and the two students are identified as Aman and Rajesh who were studying in college in Hyderabad. The bike was of one of his friend who is a local resident. Car driver ran away whose hunt is on and police is trying to catch him.

*Figure 4: Testing dataset*

After training, testing is performed using testing dataset where 2000+ road accidents records are taken to check and verify how correct the predictions are made by the system. Testing dataset is stored in notepad text files. The above figure shows the testing dataset file.

## IV. Result & Analysis

This section contains the result obtained after applying different methodologies of data classification. The data has been collected from various sources of E-Newspapers, Forums, Portals, etc. The data for road accidents is stored in an excel file. There are some common causes for road accidents which are uncovered through data mining such as OVERSPEEDING, TALKING OVER PHONES, BAD ROADS, OVERTAKING, etc. The results are obtained in Rapidminer by creating a new process and running it. The result obtained is as follows:



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

*Figure 5: RapidMiner results for Road Accidents.*

The results can be seen in the above figure where we can see the system has correctly predicted class for the unknown objects. The class of road accident cause is correctly predicted from our system by providing the road accident data. The developed system automatically makes the predictions and no manual efforts is needed.

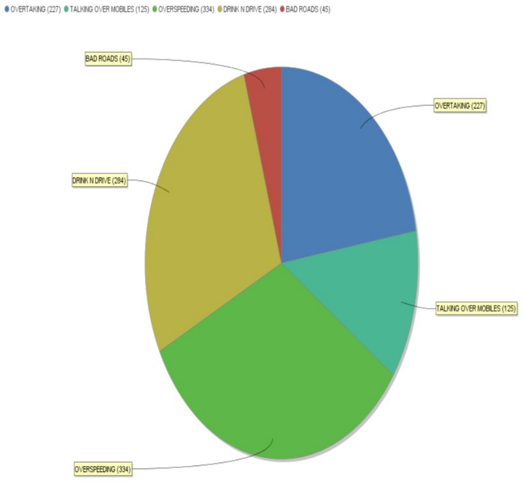


*Figure 6: Result Pie Chart*

Total 1000+ records are taken to obtain the results. The above pie chat shows the different causes of road accidents. As we can see, over speeding cause accounts to 334 accidents 33.4%, Overtaking caused 227 accidents i.e. 22.7%, Bad roads caused 45 number of accidents which is 4.5%. Talking over mobiles accounts to 125 which is 12.5%. In our research we found these causes to be the most common one.

## V. Conclusion and Future Work

We conclude the research paper by specifying the domain i.e. Data mining which we used to extract hidden knowledge from huge datasets of road accidents. We applied Naïve Bayes classifier to classify road accidents into different causes of road accidents. In future we aim to further increase the dataset and cluster the road accident data for different states and aim to use different classifier such as SVM, ID3 and Neural Network.

## 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. Suganyaet. al "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] M. Zhang et.al, "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] E. H. Kaur et al, "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] P. S. Kasbe et al "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 AryehKontorovich. "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.

* * * * *