Abstract- Train accidents have a great impact on society because of the trauma, injuries, fatalities and the associated costs. This project presents an analysis of significant train accidents occurring in world from 2000 to 2015. Collision accidents generally lead to more casualties than derailment accidents, and the most frequent cause of accidents is human error. Additionally, most train accidents occur during summer. Accident investigation and analysis are key to reinforcing and improving railway safety. Many railway accidents have been caused by degraded human performance and human error, but the tasks of train drivers and signalers have remained essentially the same. These findings can provide railway leaders with lessons and rules learned from past accidents thus facilitating the establishment of a safer railway operation environment around world.

New technologies and equipment have gradually reduced railway accidents, little or no investigation has been conducted to check whether railway performance shaping factors (R-PSFs). Attributed to degraded human performance, have changed or remained constant. Decision tree algorithm was applied in analyzing the collected data. The results show that although most derived rules are unique, some rules are worth noting. The results show that predictive accuracy for accident costs significantly improves through the use of features found by text mining. Accuracy can be further improved through the use of modern ensemble methods. Importantly, this project also shows through case examples, how the findings from text mining of the narratives can improve understanding of the contributors to rail accidents in ways not possible through fixed field analysis of the accident reports only.

Keywords: Rail accidents, railway safety, railway performance shaping factors, text mining, J48, decision tree, C4.5, narrative accident reports

1. INTRODUCTION

Indian Railways has one of the largest networks in the world criss-crossing the nation from North to South and East to West. Railway transporting system is experimenting a deep transformation now a day. The continuous technological evolution and rising the standard of living of the population has led today to an increase in the volume of rail traffic both referred to the passenger and freight transport. Despite railways being a statistically safe form of transportation rail accidents may still happen. Safety in railway transport must be understood as the safe performance of their functions by the main components. From this perspective it denotes a growing interest in the issue of safety in railway. Infrastructure must be understood as the safe performance of their functions by the main components. From this perspective it denotes a growing interest in the issues of safety in railway transport. Therefore many data mining techniques are come in to existence, like text mining.

For a long time, ministry of railways in world is a relative monopoly agency. When an accident happened, the ministry of railways would form an investigation team to investigate; the team would release the investigation report on the accident ultimately. But the investigation report only could only be seen by the railway staff, there was no way of knowing for the outside world basically. After the reform of the ministry of railways in 2013, this situation has changed. In this year, world’s railway system released annual railway traffic fatality statistics publicly at first time, 1336 people lost their life due to train accidents. In 2014, this figure reached 1232. On the whole, the number of deaths declined in 2014, but the outlook for railway safety in world is still grim. The railway system is a major component of the economy of most countries, daily transporting millions of passengers as well as millions of dollars’ worth of goods from origin to destination (1).
Therefore, the relevant operational, regulatory, and governmental bodies of every country with a rail network aim for a safe, highly reliable, and excellent quality railway system (2). In the United Kingdom in particular, railways have played a substantial role in society’s daily life and economy since the 1820s. Their crucial role is apparent in a recent Rail Safety and Standards Board Ltd. (RSSB) annual safety performance report (3), which states that for 2013 to 2014 about 1.59 billion passenger journeys, 60.1 billion passenger kilometers, and 48.5 million freight train kilometers were recorded. This paper describes an investigation to understand the possible predictors or contributors to accidents obtained from “mining” the narrative text in rail accident reports. To do this the approach integrates a combination of analytical methods to first identify the accidents of interest and then look for relationships in the structured and unstructured data that may suggest contributors to accidents. This study evaluates the efficacy of the features found from text mining using models containing these features to predict the costs of extreme accidents. In performing this evaluation the study also considers the usefulness of modern ensemble approaches incorporating these text-mined features to predict accident costs. Finally, the study teases apart the text-mined features, whose importance is confirmed by predictive accuracy, for their insights into the contributors to rail accidents. The purpose of this final analysis is to understand the insights for rail safety that text mining can provide to the exclusion of fixed field reports.

These studies revealed some interesting results, however, they are unable to properly analyse the cognitive aspects of the causes of the crashes. They often opt to leave out significant qualitative and textual information from data sets as it is difficult to create meaningful observations. The consequence of textual ignorance results in a limited analysis whereby less substantial conclusions are made. Text mining methods attempt to bridge this gap. Text Mining is discovery of new, previously unknown information, by automatically extracting it from different written (text) resources. Text mining methods are able to extract important concepts and emerging themes from the collection of text sources. Used in a practical situation, the possibilities for knowledge discovery through the use of text mining is immense. To our knowledge, there is limited or no reputable studies that have utilized text mining in this data domain, however, earlier studies in the field indicate a real need for textual mining in order to better understand the contextual relationships of road crash data.

2. LITERATURE REVIEW

Railway accidents cause numerous casualties every year, thereby attracting the interest of many researchers and analysts. Existing research on train accidents is empirically and methodologically diverse. From an empirical standpoint, most studies have attempted to identify the risk factors that influence the severity of train accident casualties. Many of these studies report that train accidents often result from a chain or sequence of events as opposed to a single cause.

Reinach and Viale[3] analyzed six train accidents and developed a human error analysis and classification model with 36 probable, contributing factors. The results demonstrate that each accident was associated with multiple contributing factors. Related studies in other countries have reported similar results. Although these models can be used to assign blame for accidents, they are usually ineffective in preventing future ones. Additionally, numerous studies have focused on the factors influencing the severity of accidents.

Ilkjær and Lind[4] analyzed a railway accident that occurred in Denmark in 1994 and concluded that carriage interior (the layout of the seat, is there an armrest on the seat, do the luggage place safe or not and so on) has a major influence on personal injuries.

Mirabadi and Sharifian[1] investigated Iranian Railways accident data from 1996 to 2005 and concluded that human error, wagon and track contribute most often to increased accident severity.

Evans[2] analyzed fatal train accidents on Europe’s main line railways from 1980 to 2009. The results show that the most common immediate causes of serious accidents at railroad crossings are errors or violations by railroad users. Other studies have concentrated on developing computer-controlled systems to prevent the train accidents.1,10,11 From a methodological standpoint, researchers have applied a variety of statistical approaches to analyze the train accident rates and trends in various countries. Evans9 data on almost all fatal railway accidents in the United Kingdom from 1967 to 2003, found downward trends in all the main classes of accidents per train kilometre in the 27 years leading up to 1993. In addition, nonparametric statistical methods have been widely used in analyzing train accidents.

Chong et al [17] applied artificial neural networks and decision trees to model the severity of injury resulting from train accidents. In all cases, the decision trees outperformed the neural networks. Yan et al.13 applied hierarchical tree-based regression to explore the train-vehicle crash prediction and analysis at passive highway-rail grade crossings and concluded that installing stop signs can effectively improve safety at these crossings estimated the overall trend in the number of fatal train
collisions and derailments per train-kilometer to be 6.3% per year from 1990 to 2009 in Europe, with a 95% confidence interval. However, there are statistically significant differences among different European countries in the rate of fatal train accidents. Evans10 analyzed Liu et al[14] developed two regression models provided a better understanding of train derailment severity distribution. Knowledge discovery and data mining techniques have also been applied in many recent explorations of this topic.

Wong and Chung[28] explored accident occurrence in Taiwan using rough set theory, a technique that has been used to make decisions in the presence of uncertainty and vagueness.16,17Mirabadi and Sharifian3 applied associated rules techniques to analyze the data from accidents on Iranian railways in order to reveal previously unknown relationships among the data. The patterns extracted through these methods can be utilized to develop regulations and rules to help prevent similar accidents from occurring in the future.

Hwangeletal [22] Performed supervised co-clustering of phenotypes and genes simultaneously by integrating various sources of phenotypic and genomic data as well as prior knowledge. Their approach enabled discovery of disease classes based on the molecular underpinnings of the phenotypes and the molecular interactions in a network. The authors proposed a phenomenological co morbidity network of diseases that is based on medical claims data. The network was made up of two layers. The first layer contains links representing the conditional probability for co morbidity while links that contain respective statistical significance are in the second layer.

They showed that the network undergoes dramatic structural changes across the lifetime of patients.

2.1 CURRENT LIMITATIONS

The current design The analysis in this study is limited to mining the causal text relating to 'Groundings', 'Collisions', 'Machinery Failures' and 'Fire' related accidents. The scope of the study has also been limited by focusing only on pattern classification and connectives methods for extracting the causal relations to keep the study to a reasonable size. There are quite a few challenges when dealing with accident investigation reports. The reports are written in the natural language with no standard template. Misspellings and abbreviations are often found. Detection of compound words such as "safety culture", "spirit status", etc are difficult as order of importance is unknown. The contextual meaning of the words "safety" and "culture" differs significantly but the word "safety culture" has a different meaning altogether. Therefore, context and semantics play an important role in text mining. To date, they have not reported large scale analysis of the narratives for information that could inform safety policies and design. They focused on retrieval not prediction.

2.2 PROPOSED WORK

Proper study of investigation reports plays an important role while analyzing the railway accidents for taking safety measures. These investigation reports consist of both structured and unstructured data. Even though analysis of structured data is an easy task compared with unstructured data, but analyzing the unstructured data gives more accurate details about the accident than structured data because it is written in natural language with no standard templates.

In the proposed system, analysis of unstructured (textual) data reports was achieved by using text mining. Text mining provides a means for efficient and informative scanning of accident cases of interest without reading the actual entire report. Therefore, text mining in this context is seen as a useful tool in understanding accidents and their influencing factors. Here in the proposed system classification algorithm J48 along C4.5 was used. New variables from the unstructured text were extracted which were later used for predicting the likelihood of attorney involvement and the severity of claims. Interesting themes were identified in the responses to the survey data.

• This paper describes an investigation to understand the possible predictors or contributors to accidents obtained from “mining” the narrative text in rail accident reports. To do this the approach integrates a combination of analytical methods to first identify the accidents of interest and then look for relationships in the structured and unstructured data that may suggest contributors to accidents.

• Finally, the paper highlights apart the text-mined features, whose importance is confirmed by predictive accuracy, for their insights into the contributors to rail accidents. The purpose of this final analysis is to understand the insights for rail safety that text mining can provide to the exclusion of fixed field reports.

3. ALGORITHMS

This paper integrates methods for safety analysis with accident report data and text mining to uncover contributors to rail accidents. This section describes related work in rail and, more generally, transportation safety and also introduces the relevant data and text mining techniques. This paper integrates methods for safety analysis with accident report data and text mining to uncover contributors to rail accidents. This section describes related work in rail and, more generally, transportation safety and also introduces the relevant data
and text mining techniques. Based on the collected data, this paper proceeds to establish a rail accident decision table. To the author’s knowledge, this paper is the first such attempt to analyze Chinese train accidents. Our approach consists of two stages. In the first stage, we use rough set theory as a tool for data pre-processing, to remove redundant knowledge from established information systems and to provide the means to deal with missing data. In the second stage, we adopt associated rules analysis to propose rules based on the past accidents’ data. These rules can uncover unknown relationships that can be the basis of forecast and decision. The number of published reports was not constant over time. With continuous technological improvements such as installation of the Train Protection and Warning System, the number of accidents and significant accidents in recent years has dropped significantly and so fewer cases were extracted for the period of 1997 to 2012. To achieve a more accurate distribution of the number of the R-PSFs for the whole period of interest, the proportion of operator errors in the 467 reports was intended to be similar to the number of errors of the final set of 237 reports. Text mining is concerned with finding patterns in unstructured text. This field has become increasingly important because of the large amounts of data available in documents, news articles, research papers, and accident reports. In many cases text databases are semi structured because in addition to the free text they also contain structured fields that have the titles, authors, dates, and other meta data. The accident reports used in this paper are semi structured. One of the key goals of text mining is to characterize the contents of the documents through pattern discovery. These patterns may then be used for improved information retrieval or, as in this paper, for input into predictive models. Regardless of the ultimate goal, most text mining begins with vector space models where documents are represented by term-document matrices. These matrices have terms as headers for the rows and documents as headers for the columns. The values in the cells give the count or frequencies of a term (row) in a document (column).

3.1 Generate Accident Report

This paper integrates methods for safety analysis with accident report data and text mining to uncover contributors to rail accidents. This section describes related work in rail and, more generally, transportation safety and also introduces the relevant data and text mining techniques.

3.2 Characteristics of Accident Report

This report has a number of fields that include characteristics of the train or trains, the personnel on the trains operational conditions (e.g., speed at the time of accident, highest speed before the accident, number of cars, and weight), and the primary cause of the accident. This field has become increasingly important because of the large amounts of data available in documents, news articles, research papers, and accident reports.

3.3 Stored In Databases

Text databases are semi structured because in addition to the free text they also contain structured fields that have the titles, authors, dates, and other Meta data. The accident reports used in this paper are semi structured.

3.4 J48 DECISION TREE

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable.
Basic Steps in the Algorithm:

Step 1: In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class.

Step 2: The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.

Step 3: Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

Step 4: Counting Gain
This process uses the “Entropy” which is a measure of the data disorder.

\[ \text{Entropy} (y) = \sum_{i=1}^{n} \frac{|y_i|}{|y|} \log \left( \frac{|y_i|}{|y|} \right) \]

\[ \text{Entropy} (\bar{y}|j) = \sum_{i=1}^{n} \frac{|y_i|}{|\bar{y}|} \log \left( \frac{|y_i|}{|\bar{y}|} \right) \]

Step 5: Gain:

\[ \text{Gain}(\bar{y}, j) = \text{Entropy}(\bar{y}) - \text{Entropy}(\bar{y}|j) \]

The objective is to maximize the Gain, dividing by overall entropy due to split argument by value. Because of the outliers this is a significant step to the result. Some instances are present in all data sets which are not well-defined and differ from the other instances on its neighborhood. The classification is performed on the instances of the training set and tree is formed. The pruning is performed for decreasing classification errors which are being produced by specialization in the training set. Pruning is performed for the generalization of the tree.

3.5 C4.5 ALGORITHM

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. Authors of the Weka machine learning software described the C4.5 algorithm as a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date.

Basic steps in algorithm

Step 1: Select one attribute from a set of training instances.

Step 2: Select an initial subset of the training instances.

Step 3: Use the attribute and the subset of instances to build a decision tree.

Step 4: Use the rest of the training instances (those not in the subset used for construction) to test the accuracy of the constructed tree.

Step 5: If all instances are correctly classified – stop.

Step 6: If instances are incorrectly classified, add it to the initial subset and construct a new tree.

Step 7: Iterate until
(i) A tree is built that classifies all instance correctly
OR
(ii) A tree is built from the entire training set

4. RESULTS

In these scenario rail accidents has been predicted by applying different statistical methods. The text data set is preprocessed for output evaluation. Once our text dataset is preprocessed then we need to generate the report for corresponding test data.

In fig 2 accident information is given as input dataset.

Fig 2: Textual data is given as test set

In fig 3 and 4 the textual data is evaluated using our proposed model and which is visualized in graph.
5. Conclusion

In this Paper, show cases the combination of methods to improve the accuracy of models which predict accident severity. Text analysis with ensemble methods can provide insights into accident characteristics. Modern text analysis methods make the narratives in the accident reports almost as accessible for detailed analysis as the fixed fields in the reports. Text mining of the narratives provides a much richer amount of information than is possible in the fixed fields. Standard methods to clean the narratives have been employed. Train accident narratives use jargon common to the rail transport industry, since classical stemming and stop word removal do not do a good job of characterizing the words used in this industry. For train safety analysis, text mining could benefit from a careful look at ways to extract features and advantage of language characteristics particular to the rail transport industry. Herein this project, railway accident reports are investigated.

This approach can be extended to investigate different types of narrative reports pertaining to other transport domains like airways and roadways. Also, it can be implemented for ensuring safety in industries where safety is major concerned.

References

[2]”Fatal train accidents on Europe railways” by Evans http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2861873

Authors Profile

M Naveen Kumar, pursuing full time Ph.D from Centurion University of Technology & Management. He has 10 Years of extensive teaching experience in various domains such as Data mining, machine learning and Big Data analytics including various cloud models.
G Kumari, pursuing full time Ph.D from Andhra University. She has 10 years of teaching experience in various areas such as Machine Learning, Big Data Analytics, Data Mining and various cloud computing models.

Dr. A. M. Sowjanya, has done B.Tech and M.Tech in Computer Science. Her Ph.D in Incremental Clustering. She is currently working as Assistant Professor in College of Engineering, Andhra University. Her research interests includes Machine Learning, Data Mining and Big Data Analytics.