International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 14th Aug 2019. Link :http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-08

Title TO PREVENT RAIL ACCIDENTS BY THE CONTRIBUTORS USING TEXT DATA MINING

Volume 08, Issue 08, Pages: 342–349.

Paper Authors

ROOPA TIRUMALASETTI, A.MALATHI

Guntur Engineering College

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

TO PREVENT RAIL ACCIDENTS BY THE CONTRIBUTORS USING TEXT DATA MINING

[1]ROOPA TIRUMALASETTI, [2]A.MALATHI$_{M.TECH}$
[1]M.Tech (Computer Science Engineering),Guntur Engineering College
[2]Assistant professor, Guntur Engineering College
[1]sindhujaakhila@gmail.com,[2]sireeraj19@gmail.com

Abstract: Security worries for the transportation business in numerous nations. In the 11 years Rail mischance's represent a critical from 2001 to 2012, the U.S. had more than 40000 rail mischance's that cost more than $45 million .While a large portion of the mishaps amid this period had next to no cost, around 5200 had harms in abundance of $141500. To better comprehend the supporters of these extraordinary mishaps, the Federal Railroad Administration has required the railways required in mischance to submit reports that contain both fixed field sections and stories that portray the qualities of the mischance. While various reviews have taken a gander at the fixed fields, none have done a broad examination of the accounts. This paper portrays the utilization of content mining with a mix of methods to naturally find mischance attributes that can advise a superior comprehension of the supporters of the mishaps. The review assesses the efficacy of content mining of mishap accounts by evaluating prescient execution for the expenses of extraordinary mischances. The outcomes how that prescient exactness for mischance costs significantly enhances using highlights found by content mining and prescient precision additionally enhances using present day group techniques. Imperatively, this review likewise appears through case cases how the findings from content mining of the stories can enhance comprehension of the supporters of rail mishaps in ways unrealistic through just fixed field examination of the mischance reports.

Keywords—Rail safety, safety engineering, latent Dirichletallocation, partial least squares, random forests.

INTRODUCTION

IN the 11 years from 2001 to 2012 the U.S. had more than 40000 rail accidents with a total cost of $45.9 M. These accidents resulted in 671 deaths and 7061 injuries. Since 1975 the Federal Railroad Administration (FRA) has collected data to understand and find ways to reduce the numbers and severity of these accidents. The FRA has set ―an ultimate goal of zero tolerance for rail-related accidents, injuries, and fatalities‖ [1]. A review of the data collected by the FRA shows a variety of accident types from derailments to truncheon bar entanglements. Most of the accidents are not serious; since, they cause little damage and no injuries. However, there are some that cause over $1 Mindamages, deaths of crew and passengers, and many injuries. The problem is to understand the

characteristics of these accidents that may inform both system design and policies to improvesafety. After every mischance a report is finished and submitted to the FRA by the railroad organizations included. This report has various fields that incorporate qualities of the prepare or prepares, the work force on the trains, the natural conditions (e.g., temperature and precipitation), operational conditions (e.g., speed at the season of mischance, most elevated speed before the mishap, number of autos, and weight), and the essential driver of the mishap. Cause is a four character, coded passage in view of in light of 5 general classes (examined in Section IV). The FRA additionally gathers information on the expenses of every mischance decomposed into damages to track and equipment to incorporate the quantity of unsafe material autos harmed. Also, they report the quantity of wounds and passings from every mischance.The FRA makes the information from these mischance reports accessible on-line at [2]. In the course of the most recent 12 years the quantity of fields have changed just somewhat, in spite of the fact that there are some missing qualities. For example,the track thickness field is missing over 90% of its qualities. After every mischance a report is finished organizations included. This report has various fields that incorporate qualities of the prepare or prepares, the work force on the trains, the natural conditions (e.g., temperature and precipitation), operational conditions (e.g., speed at the season of mischance, most elevated speed before the mishap, number of autos, and weight), and the essential driver of the mishap. Cause is a four character, coded passage in view of in light of 5 general classes (examined in Section IV). The FRA additionally gathers information on the expenses of every mischance decomposedintodamagesto track and equipmentto incorporate the quantity of unsafe material autos harmed. Also, they report the quantity of wounds and passings from every mischance.

The paper by Akin and Akbas [5] depicts the utilization of neural systems to model convergence accidents and crossing point attributes, for example, lighting, surface materials, and so on. Taken together these papers demonstrate the utilization information mining to better comprehend the variables that can impact and enhance wellbeing at rail intersections . Recent work has shown the applicability of data and text mining to broader classes of safety and security problems relevant to transportation. For example, the use of data mining techniques for anomaly detection in road networks is illustrated by the work of [6]. They providemethods to detect anomalies in massive amounts of traffic data and then cluster these detections according to different attributes. Similarly D'Andreaet al. mined Twitter and used supportvectormachines to detect traffic events [7]. Another recent application of text mining is to license plate recognition [8]. These authors use Levenshteintextmining in combination with a Bayesian approach to increase the accuracy of automated license plate matching. Cao et al., use data mining in combination with rule-based and machine learning approaches to perform traffic sentiment analysis [9]. Speech processing and message feature extraction have been used for detection of intent in traveler screening [10].

Recently results by [11] show the use of text mining for fault diagnosis in high-speed rail systems. The authors of this work use probabilistic latent semantic analysis [12] in combination with Bayesian networks for diagnosis of faults in vehicle onboard equipment. They assessed their method through two experiments that obtained real fault detection data on the Wuhan-Guangzhou high speed rail signaling system.

Data from the rail accidents in the US To comprehend the qualities of rail mishaps in the U.S. we utilize the information accessible on mishaps for a long time (2001–2012) [2]. The information comprise of yearly reports of mishaps and every yearly set has 141 factors. The announcing factors really changed over this period yet we utilize the subset of 141 that were reliable all through the 11 years. The factors are a blend of numeric, e.g., mishap speed, clear cut, e.g., hardware sort, and free content. The free content is contained in 15 story handle that depict the mishap. Each field is restricted to 100 bytes and that gives a sum of 1500 bytes to depict the mischance. Under 0.5% of the mishap reports have any content in the fifteenth field. The normal number of words in an account is 22.8 and the middle is 19. The biggest story has a 173 words and the littlest has 1.

Data structuring and cleaning Before examining the examination utilized as a part of this review we have to additionally depict how we organized and cleaned the information. As noted in Section III there are 5471 unduplicated, outrageous harm mishaps in the wake of expelling the one that happened because of the assaults on 9/11. Assist information cleaning portrayed in this area decreased the informational index by 2 extra indicates 5469. We arbitrarily partitioned the reports into preparing and test sets. The preparation set contains 3667 mishaps and the test set has 1802. The aggregate mishap harm for perceptions in the test set reaches from $143.2k to $13M with a middle of $342.2k add up to mischance harm for mishaps in the test set reaches from $143.4k to $13M with a middle of $342.4k. As noted underneath we rolled out a few little improvements from the arbitrary attract to better adjust the test set.

III. ANALYSIS OF THECONTRIBUTERS TO RAIL ACCIDENTS

The study in this paper looked at different analytical approaches to understand contributors to rail accidents, and specifically, to rail accident damage. To achieve this goal, this study sought to answer three major questions:

1) Do the narratives in accident reports contain features that can improve the predictive accuracy of accident severity?

2) Do ensemble methods provide significant performance lift in the prediction of accident severity?

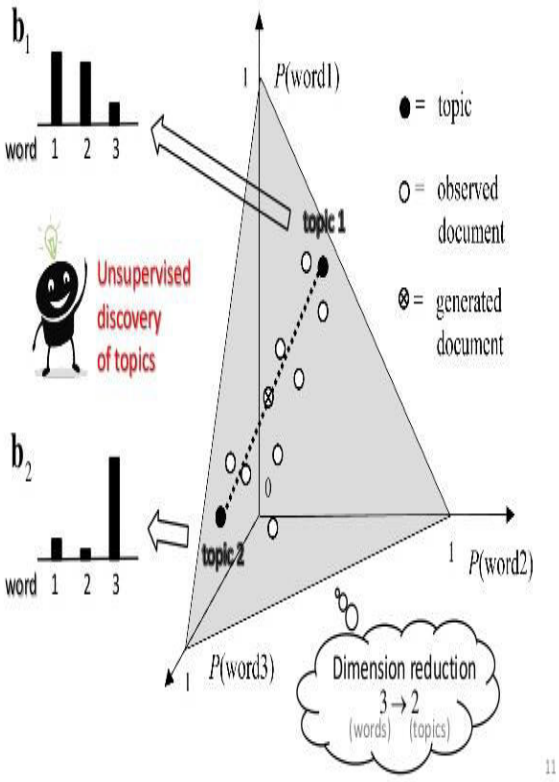
3) Can text mining of accident narratives improve our understanding of rail accidents?

The first question is important because there is no existing study of the automated use of narrative text for understanding accidents. If text can more accurately predict outcomes then its analysis has the potential to improve our understanding of the accidents. Notice that we do not deceive ourselves in thinking we can accurately predict accident damage using the small set of variables provided by the

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

accident reports. Our goal is to use predictive accuracy as a metric in assessing the efficacy of using text and data mining to understand contributors to accident damage.

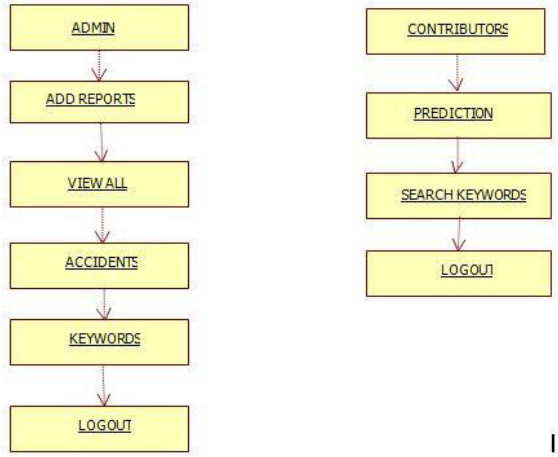
SYSTEM ARCHITECTURE:



DATA FLOW DIAGRAM:

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.



Module Descriptions:

Accident Report Generation:

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.

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.

Text Mining Techniques:

Latent Dirichlet Allocation (LDA): LDA provides a method to identify topics in text. We applied LDA to the accident narratives to obtain 10 and 100 topics. To incorporate LDA topics into these ensemble models we again score each topic in each narrative by the proportion of

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

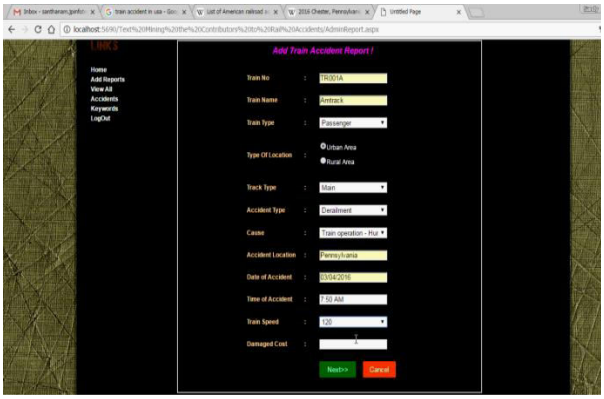
topic words in the narrative. In order to compare the importance of topics, we also used the ensemble models with the top ten most important words in each topic.

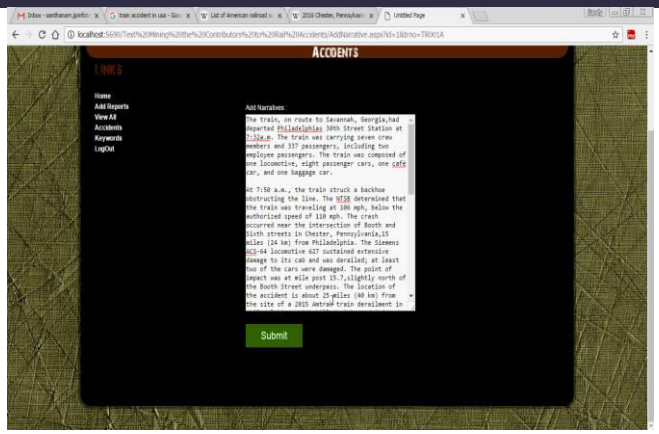
Partial Least Squares (PLS):we measure importance as the percent change in root mean square error (RMSE) in the out-of-bag sample when that variable is removed. The results indicate that of the 20 most important variables 16 are LDA topics. For PLS we first obtained 1000 words from the LDA topics. We then found the estimated number of PLS components using cross-validation. We incorporate the PLS component into the accident damage models using two approaches. In the first approach we use a two step process. We first predict damage with only the PLS component. In the second approach we us the PLS component to estimate the coefficients for each word and directly use the results as another predictor variable, the PLS predictor, in the random forest model.

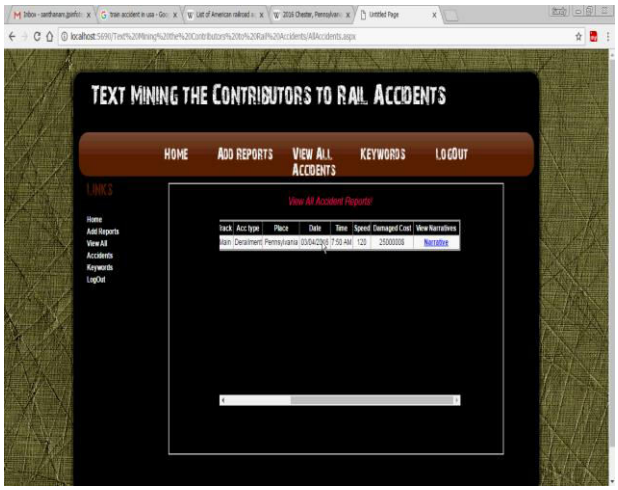
Stored In databases:

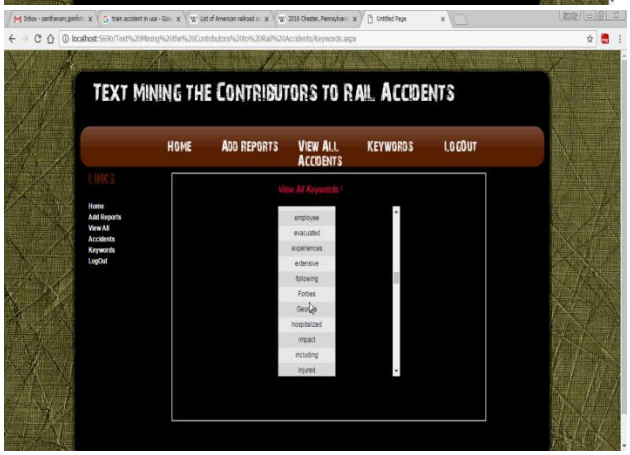
Text databases are semi structured because in addition to the free text they also contain structured fields that have the titles, authors, dates, and other Meta data. The accident reports used in this paper are semi structured.

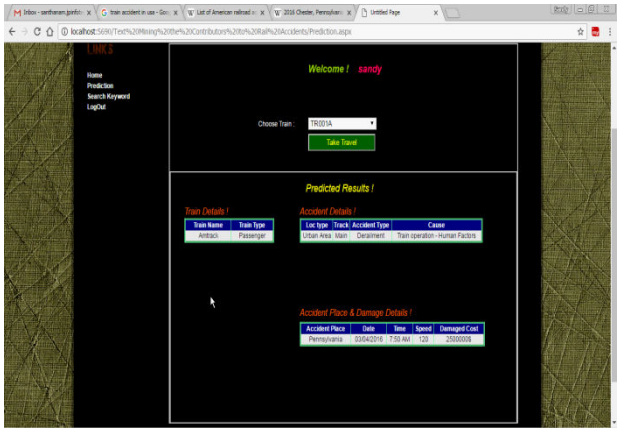
RESULTS



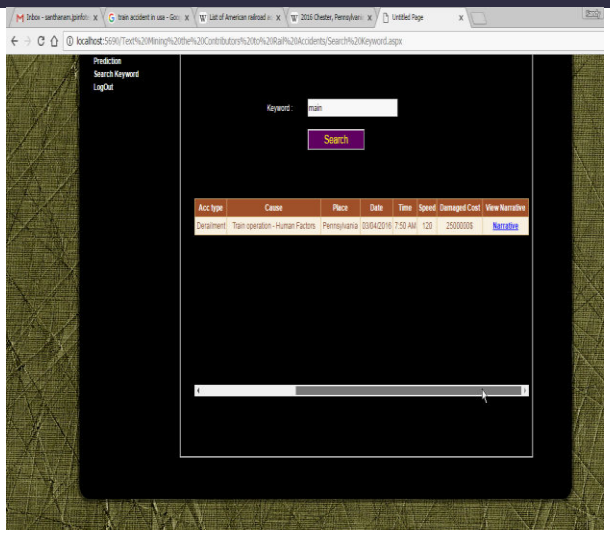








International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org



FUTURE ENHANCEMENT

There are also several areas of future work that will provide more fundamental advances in the use of text mining for train safety engineering. The first is to exploit the ability of narratives to represent the current state of safety while the fixed fields are locked into the understanding available at the time of the database design. Hence, research is needed to provide a temporal representation of the evolution of narratives, since this temporal review will possibly expose areas where safety has improved, as well as, the current and evolving challenges. A second of fundamental research need is to characterize the variation and uncertainty inherent in text mining techniques. In this study the use of both LDA and PLS did not give consistent results with different training and test set selections. These differences need to be formally characterized and, ideally, described with a probabilistic model that further enhances understanding of the contributors to accidents.

CONCLUSION

The results presented in Section V show that the combination of text analysis with ensemble methods can improve the accuracy of models for predicting accident severity and that text analysis can provide insights into accident characteristics not available from only the fixed field entries. As shown in Table VII the improvements provided by text and ensemble modeling are dramatic even without working to optimize the performance of the ensemble methods for these data. This suggests that these techniques should be added to the toolkit and training of train safety engineers. Additionally as discussed in Section V and made evident in Figs. 8 and 9 the use of text analysis can enhance the safety engineers overall understanding of the contributors to accidents in ways not possible with only analysis of the fixed fields. Modern text analysis methods make the narratives in the accident reports almost as accessible for detailed analysis as the fixed fields in the reports. More importantly as the examples illustrated, text mining of the narratives can provide a much richer amount of information than is possible in the fixed fields. This makes sense since the narratives can describe the characteristics of the accident in more detail, while the fixed fields are limited to the structure and schema of the original database designers.

REFERENCES

[1] "Railroad safety statistics—2009 Annual report—Final," Federal Railroad Admin., Washington, DC, USA, Apr. 2011.

[2] "Office of safety analysis," Federal Railroad Administration, Washington, DC, USA, Oct. 2009. [Online]. Available: http://safetydata.fra.dot.gov/officeofsafety/

[3] G. Cirovic and D. Pamucar, "Decision support model for prioritizing railway level crossings for safety improvements: Application of the adaptive neuro-fuzzy system," *Expert Syst. Appl.*, vol. 40, pp. 2208–2223, 2013.

[4] L.-S. Tey, G. Wallis, S. Cloete, and L. Ferreira, "Modelling driver behaviour towards innovative warning devices at railway level crossings," *Neural Comput. Appl.*, vol. 51, pp. 104–111, Mar. 2013.

[5] D. Akin and B. Akbas, "A neural network (NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics," *Sci. Res. Essays*, vol. 5, pp. 2837–2847, 2010.

[6] H. Gonzalez, J. Han, Y. Ouyang, and S. Seith, "Multidimensional data mining of traffic anomalies on large-scale road networks," *Transp. Res. Rec.*, vol. 2215, pp. 75–84, 2011.

[7] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from Twitter stream analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2269–2283, Mar. 2015.

[8] F. Oliveira-Neto, L. Han, and M. K. Jeong, "An online self-learning algorithm for license plate matching," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1806–1816, Dec. 2013.

[9] J. Cao *et al.*, "Web-based traffic sentiment analysis: Methods and applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 844–853, Apr. 2014.

[10] J. Burgoon *et al.*, "Detecting concealment of intent in transportation screening: A proof of concept," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 103–112, Mar. 2009.

[11] Y. Zhao, T. H. Xu, andW. Hai-feng, "Text mining based fault diagnosis of vehicle on-board equipment for high speed railway," in *Proc. IEEE 17th Int. Conf. ITSC*, Oct. 2014, pp. 900–905.

[12] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 50–57.

[13] R. Nayak, N. Piyatrapoomi, J. W. R. Nayak, N. Piyatrapoomi, and J. Weligamage, "Application of text mining in analysing road crashes for road asset management," in *Proc. 4th World Congr. Eng. Asset Manage.*, Athens, Greece, Sep. 2009, pp. 49–58.

[14] "Leximancer Pty Ltd." [Online]. Available: http://info.leximancer.com/ academic

[15] A. E. Smith and M. S. Humphreys, "Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping," *Behav. Res. Methods*, vol. 38, no. 2, pp. 262–279, 2006.

[16] U.S. Grant, The Personal Memoirs of U.S. Grant., 1885. [Online]. Available: http://www.gutenberg.org/files/4367/4367-pdf/4367-pdf.pdf

[17] W. Jin, R. K. Srihari, H. H. Ho, and X. Wu, "Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques," in *Proc. 7th IEEE Int. Conf. Data Mining*, Omaha, NE, USA, Oct. 2007, pp. 193–202.

[18] D. Delen*et al.*, *Practical Text Mining and Statistical Analysis for Non-StructuredTextDataApplications*. Waltham,MA, USA:Academic, 2012.

[19] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

[20] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.