INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 1 (2017) PAGES 1627 - 1639
RECEIVED : 15.02.2017 PUBLISHED : 26.02.2017
February 26, 2017

TEXT MINING HYPOTHESIS, DEALINGS AND NOTION IN PREDICTING REPORTS OF RAILROAD ACCIDENTS.

Vishnupriya. M[1], Vanathi. A[2]
[1]PG Student, Dept. of Computer science and Engineering, Priyadarshini Engineering College, Vaniyambadi,Tamilnadu, India,
[2]Associate Professor, Dept. of Computer science and Engineering, Priyadarshini Engineering College, Vaniyambadi, Tamilnadu, India.

Abstract

Transportation plays an important role in humans life and now in major cities rail transportation is widely being used by various people most of the studies are not been focused on predicting accidents in rail transportation . A survey shows that since 11 years from 2001 to 2012, the U.S. had more than 40,000 rail accidents which cost more than $45 million. Most of the accidents are very expensive during this period up to $141 500. The Federal Railroad Administration has required the railroads involved in accidents to submit reports that contain both fixed field entries and narratives that describe the characteristics of the accident .So this paper proposes a clear idea regarding safety design and policies in railroad transportation and based on earlier estimation on accident reports the cost of repair can be made accurately by using the text mining with combination techniques to automatically discover accident characteristics that can inform better understanding of the contributors to the accidents. The findings from text mining of the narratives can recover understanding of the contributors to railroad accidents in ways not convenient through detached fixed field analysis of the accident reports.

Keywords : Railroad prediction, Design, Dirichlet allocation, Text Mining, Accident Reports, OSHA.

1. INTRODUCTION

In the 11 years from 2001 to 2012 the U.S. had more than 40 000 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 $1M in damages, 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 improve safety. After each accident a report is completed and submitted to the FRA by the railroad companies involved. This report has a number of fields that include characteristics of the train or trains, the personnel on the trains, the environmental conditions (e.g., temperature and precipitation), 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. Cause is a four character, coded entry based on based

on 5 overall categories (discussed in Section IV). The FRA also collects data on the costs of each accident decomposed into damages to track and equipment to include the number of hazardous material cars damaged. Additionally, they report the number of injuries and deaths from each accident.Finally, the accident reports contain narratives which provide a free text description of the accident. These narratives contain more description about the causes and contributors to the accidents and their circumstances. However, for brevity these narratives use railroad specific jargon that makes them difficult to read by personnel from outside the industry.

The FRA makes the data from these accident reports avail-able on-line at [2]. Over the last 12 years the number of fields has changed only slightly, although there are some missing values. For example, the track density field is missing more than 90% of its values. The FRA uses all of these data much as the Federal Aviation Administration uses reports on aviation accidents, namely, to "develop hazard elimination and risk reduction programs that focus on preventing railroad injuries and accidents" [1]. However, as with many safety and regulatory agencies, they can effectively perform analyses on aggregate trends and conditions as shown by the major elements in their report fields. To date, they have not reported large scale analysis of the narratives for information that could inform safety policies and design. This paper describes the use of text mining with a combination of techniques to automatically discover accident characteristics that can inform a better understanding of the contributors to the accidents. The study evaluates the efficacy of text mining of accident narratives by assessing predictive performance for the costs of extreme accidents. The techniques we use from data mining derive from ensemble methods that combine the results from many models or learners to produce a consensus prediction. We apply two types of ensembles: boosting and bootstrap aggregation or bagging. Text mining is concerned with finding patterns in unstructured text. In performing this evaluation the study also considers the usefulness of OSHA 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 railroad 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.

2. DATA FROM RAIL ACCIDENTS IN THE U.S.

To understand the characteristics of rail accidents in the U.S. we use the data available on accidents for 11 years (2001–2012) [2]. The data consist of yearly reports of accidents and each yearly set has 141 variables. The reporting variables actually changed over this period but we use the subset of 141 that were consistent throughout the 11 years. The variables are a combination of numeric, e.g., accident speed, categorical, e.g., equipment type, and free text. The free text is contained in 15 narrative fields that describe the accident. Each field is limited to 100 bytes and that gives a total of 1500 bytes to describe the accident. Less than 0.5% of the accident reports have any text in the 15th field. The average number of words in a narrative is 22.8 and the median is 19. The largest narrative has a 173 words and the smallest has 1. Over the 11 years from 2001 to 2012 there were 42 033 reported accidents. If an accident involves more than one train it generates multiple reports. For this study we condensed these multiple reports into a single report and that gives 36 608 unduplicated accident reports. We also combined fields, such as the numbers of different types of cars (e.g., cabooses) into one field that represented the number of cars. Fig. 1 shows the data are skewed with many low values as indicted by the fact that the boxes in the box plots are lines. The extreme values shown in the figure indicate accidents with higher costs. In year 2001 the 9/11 attacks on the World Trade Center produced an accident that cost almost $17M. Years 2005, 2008, and 2012 also

had costly accidents while 2007, 2009, and 2010 did not have as many extreme accidents.

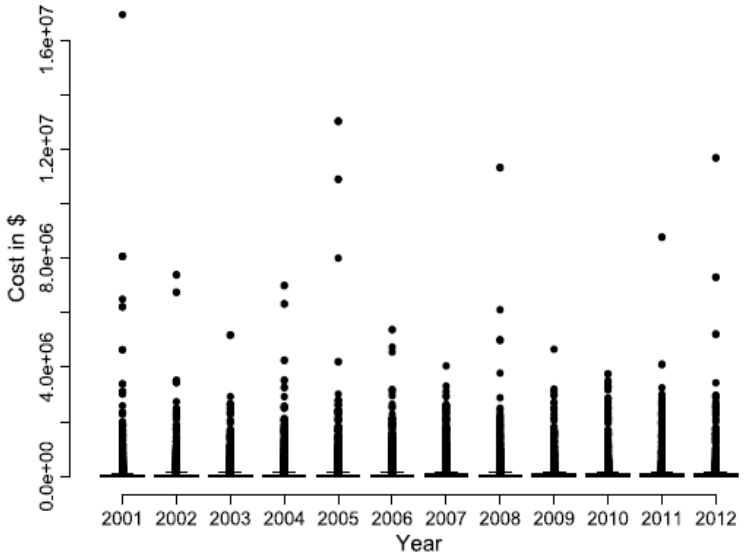

Fig. 1. Box plots of total accident damage from 2001–2011

Given the skewness in these data we focused only on the extreme accidents. To find these we used the box plot extremity point. This point is the location of the upper whisker which is the Upper Fourth plus 1.5× the Fourth Spread. The Upper Fourth is roughly the 75th quantile and the Fourth Spread approximately equals the interquartile range. For these data and this rule, accidents are considered extreme if they have a total cost of more than $141 500. Only 5472 or about 15% of the accidents have damage costs above this value. We also removed the single data point associated with the damage from the 9/11 attacks. This damage, almost $ 17M, was about $4M more than the next most expensive train accident in this 11 year period. Perhaps curiously, accidents with extreme damage do not correlate well with accidents with injuries or loss of life. The correlation between casualties (the sum of total killed and injured) and accident damage is 0.01. This suggests that costly accidents occur to freight trains and that passenger trains have lower equipment and track damage costs. This paper focuses on accidents with extreme cost as measured in dollars and not on injured or killed.

3. OSHA REPORT

A conceptual modelling approach is adopted for this study as shown in Figure 2. Causal factors are identified using database of construction accidents and injuries maintained by OSHA. Interactions among the causal factors are obtained from data mining to form a network of relationship diagrams known as Influence Diagrams. When conditional probability tables for these influence diagrams are obtained through either empirical data or through subject matter experts, Bayesian Belief Network (BBN) is constructed. Conditional probability represents the chance that one event will occur given that a second event has already occurred. For example, referring to the previous example, chance of an accident occurrence given that the site conditions are hazardous is represented by a conditional probability. With the aid of such a BBN, construction safety risk is quantified in a probabilistic form.

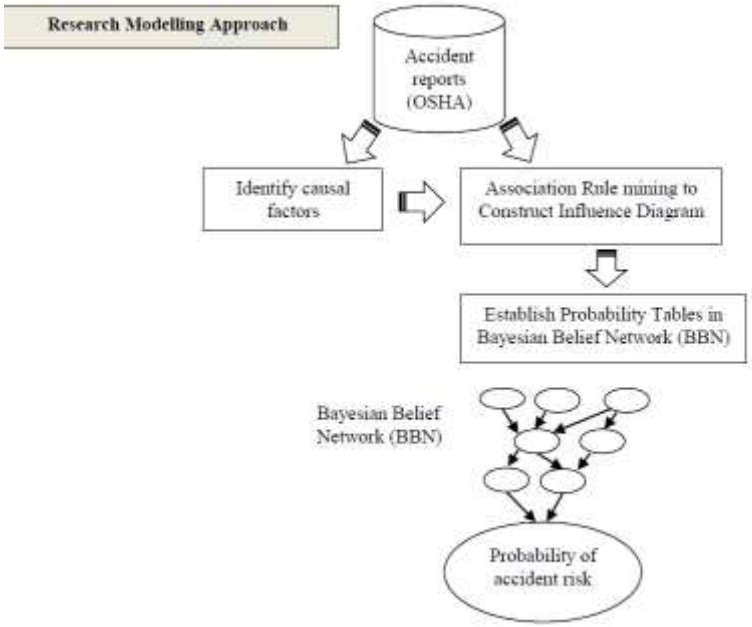

Fig.2. Conceptual modeling approach

The Occupational Safety and Health Administration, a division of the U.S. Departmental of Labor, is the federal agency charged with developing standards for workplace safety. When the agency, known as OSHA, conducts workplace inspections to determine compliance with its standards, it compiles reports of the results of these inspections. These reports are available to the public under the provisions of the Freedom of Information Act. Many of these reports are available online at the OSHA website. The following paragraph outlines how such data may be accessed for the purpose of the safety research such as this. Visit the OSHA website home page and click on the tab "Data and Statistics" located at the top of the page. The data and Statistics' tab will take you to the page where you can access the individual establishment inspection data. Once the "Data and Statistics" page loads, it offers you a series of choices. You will see links to various OSHA databases, such as commonly used statistics, workplace injury, illness and fatality statistics and establishment specific injury and illness data. Each presents OSHA statistics and information from a different point of view. The "Accident Investigation" search allows you to access the accident investigation summaries completed by OSHA investigators following an accident that resulted in a "fatality or catastrophe" reportable on OSHA Form 170. This search can be based on a keyword in the report, a phrase within the report, the date of the event or the industry code. The earliest summaries are those from 1984 and the newest are dated exactly one year earlier than the date you search the database.

4. ANALYSIS OF THE CONTRIBUTORS 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

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 1 (2017) PAGES 1627 - 1639
RECEIVED : 15.02.2017 PUBLISHED : 26.02.2017
February 26, 2017

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 than 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 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.

The second question asks can ensemble methods with text provide additional lift in the prediction of accident severity? Ensemble methods have shown better performance on a variety data mining problems, and if that is also true for train accidents then we can apply these techniques to this important area. Finally, if the answers to both preceding questions are affirma-tive then which text and non-text features best predict accident severity. Answering this last questions will enable preliminary understanding of contributors to rail accidents.

1	2	3	4	5
shove yard pull cut	unit	curv	conductor walk	broken inspect
6	**7**	**8**	**9**	**10**
bridg fire equip oper contain	gallon fuel ton spill approxim capac gatx	truck cross struck stop signal fail	main line travel east side load	hazard materi leak

Table 1. UNIQUE WORDS IN THE 10 TOPICS IN THE ACCIDENT REPORTS

Once the data were structured and cleaned (Section IV) we proceeded to address the first study question: Do the nar-ratives in accident reports contain features that can improve the predictive accuracy of accident severity? To answer this question we used ordinary least squares regression with and without topics found by Latent Dirichlet Allocation (LDA). As noted in Section II. LDA provides a method to identify topics in text. We applied LDA to the accident narratives to obtain 10 and 100 topics. Table 1 shows the unique words in each of the topics for the 10 topic results. These words give insight into the topics. For instance, topic 10 involves hazardous ma-terial leaks and spills; topic 8 concerns crossing accidents; and topic 1 concerns yard accidents. Fig. 3 shows the frequencies of the ten topics in the accident reports. For each topic, this figure shows the number of reports in which it was the most common (labeled 1), next most com-mon (labeled 2), and so forth. For instance, topic 5 is the most common topic in the most accident reports. In contrast topic 2 is the fifth most common topic in most accident narratives. We incorporated the LDA topics into OLS using a score function for each topic. The topic's score was computed as the proportion of topic words contained in the

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 1 (2017) PAGES 1627 - 1639
RECEIVED : 15.02.2017 PUBLISHED : 26.02.2017
February 26, 2017

narrative. So if all the words in topic *j* appear at least once in the narrative for accident *i* then the score, S_{ij} for that topic and accident is 1.0. If only 50% of the topic *j* words appear in narrative for accident *i* then the score is 0.5. If a topic word appears more than once in a narrative the additional appearances do not change the score. For *k* topics, this means that *k* topic variables are included in the OLS where the value of each variable is in the bounded interval [0,1].

Ordinary Least Squares (OLS) predicted accident damage on the test set with a root mean square error (RMSE) of 9.4e5. Including 10 and 100 topics as given by LDA in the OLS produced RMSE results on the test set of 9.3e5 and 9.1e5 respectively. Nested model F-tests showed that both differences had $p < 0.001$. So, clearly incorporating text into the analysis of accidents can improve predicting the costs of these extreme events. Table 2 shows the top 10 words in the five most significant topics in the OLS model. While many of the words in these topics are of obvious importance in the analysis of accidents (e.g., derail), some are not so obvious to those less familiar with accident narratives. For instance, stcc is the standard transportation commodity code. Examples of its use in the narratives are: "ALCOHOLIC BEVERAGE STCC 4910103" and "FOUR OF THESE TANK CARS RELEASED PRODUCT STCC 4914168."We turn now to the second study question: Do ensemble methods provide significant performance lift in the prediction of accident severity? If so, these methods can enable additional insights into the contributors to rail accidents. To answer this question we use the ensemble methods of boosting and bagging as described in Section II with the text mining techniques of LDA and partial least squares (PLS). For boosting we use gradient boosting which treats the approximating functions (see equation (1)) as parameters in a functional gradient descent optimization. Essentially, this algorithm fits a weak learner (e.g., a tree) to approximate the direction of the gradient. For bagging we used random forests.

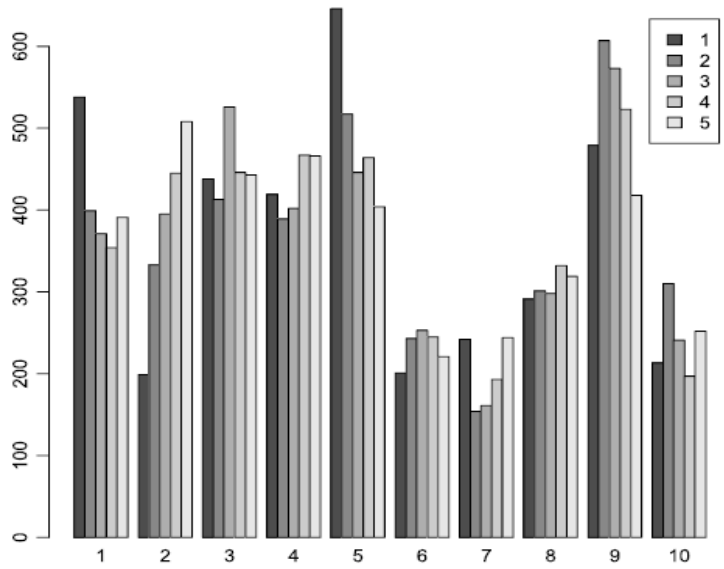

Fig. 3. Frequency of 10 topics in the accident reports

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 1 (2017) PAGES 1627 - 1639
RECEIVED : 15.02.2017 PUBLISHED : 26.02.2017
February 26, 2017

1	2	3	4	5
car	due	releas	stcc	investig
hit	destroy	unknown	gatx	start
derail	factor	amt	utlx	throttl
interlock	causal	admx	gallon	derail
bolt	fifteen	lbs	alcohol	flood
deex	fourteengallon	ethyl	pound	car
green	jecx	nitrat	liquid	gtw
flag	bro	rip	lost	train
train	cebjk	ypa	acid	washout
happen	ken	ammonium	nos	final

Table 2. Significant Topics in the OLS Model

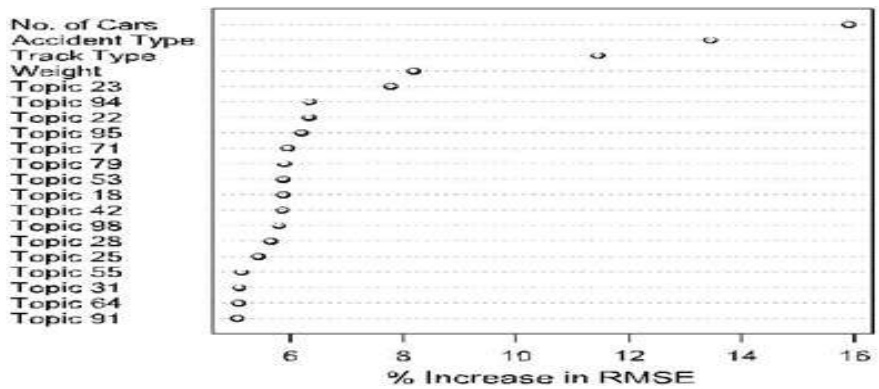

Fig. 4. Variable importance for the random forest model with 100 LDA topics.

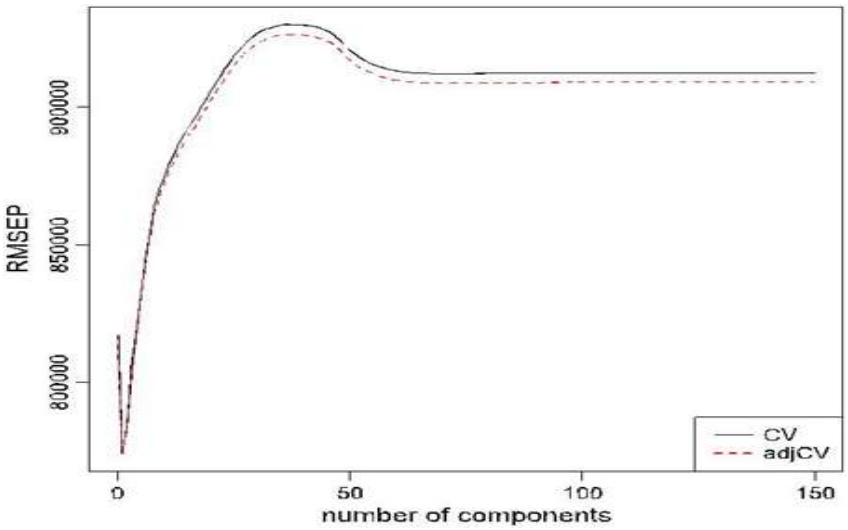

Fig. 5. RMSE from cross-validation with different numbers of components.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 1 (2017) PAGES 1627 - 1639
RECEIVED : 15.02.2017 PUBLISHED : 26.02.2017
February 26, 2017

To incorporate LDA topics into these ensemble models we again score each topic in each narrative by the proportion of 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.Fig. 4 shows the 20 most important variables in the most predictive random forest model. As noted above, we measure importance as the percent change in root mean square error (RMSE) in the out-of-bag sample when that variable is re-moved. The results in Fig. 4 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. Fig. 5 shows the RMSE obtained from cross-validation (CV) for different numbers of components. The minimum is at 1 component and so the models described here only use a single component.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 other words, this prediction was made with only the text as input. We then estimate the residuals from this "text only" prediction using random forest models with the remaining predictor variables. We obtain total accident damage cost estimates by first predicting the residuals and then adding them to the prediction for accident damage from the PLS text model.

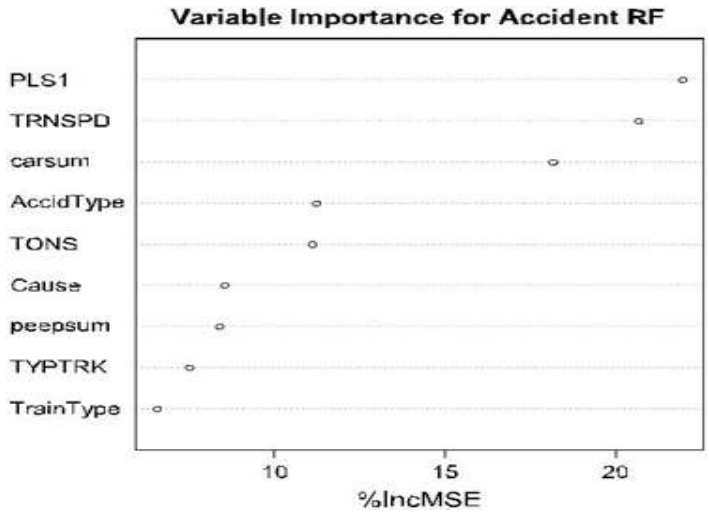

Fig. 6. Variable importance for the random forest model with the PLS predictor.

Text mining	OLS	Random Forests	Gradient Boosting
No text	9.4e5	8.8e5	9.0e5
10 LDA topics	9.3e5	8.4e5	8.9e5
100 LDA topics	9.1e5	8.3e5	8.8e5
PLS Residuals	8.5e5	8.2e5	8.4e5
PLS Variable	8.6e5	8.0e5	8.4e5

Table 3. Text Mining Rmse

In the second approach we us the PLS component to estimate the coefficients for each word and directly use the results as an-other predictor variable, the PLS predictor, in the random forest model. The PLS predictor is then simply a linear combination of the words in the accident narratives. In our tests this PLS predictor was consistently the most important variable used by the random forest

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 1 (2017) PAGES 1627 - 1639
RECEIVED : 15.02.2017 PUBLISHED : 26.02.2017
February 26, 2017

models (see Fig. 6).Table 3 shows the RMSE for the different combinations of supervised learning methods with text mining techniques. These results answer the second question and show that ensemble methods do provide lift in predicting accident severity. As with the OLS results, the values in this table also show that the ensemble methods improve in predictive accuracy with the inclusion of text mining results. As to the type of text mining, PLS shows better performance than LDA.In these tests random forests did better than the methods without text mining and across all text mining techniques. Both random forests and gradient boosting have a number of parameters that an analyst can adjust. For this work we did not attempt to optimize performance of either method but we did vary the number of trees used in the random forests from 100 to 500 (300 did best). We varied the number of trees in gradient boosting from 1000 to 50 000 (50k did best). Our goal in answering the question regarding ensemble methods was not to choose among them, but to decide if their use is appropriate for transportation safety modeling as described in this paper. The results show that it is.

Topic 22	Topic 23	Topic 71	Topic 94	Topic 95
damag	curv	track	crew	car
est	forc	joint	test	leak
bnsfs	degre	pod	ihb	impact
hove	worn	milepost	san	gtw
throught	later	measur	gsabcc	unattend
valx	low	take	pressur	poor
cprs	combin	ment	rogsm	solut
equipm	creat	crosslevel	devic	gear
kmnoa	makeup	trackag	eot	assembl
cmprhj	rail	soo	tox	corros

Table 4. Important LDA Topics In The Random Forest Model

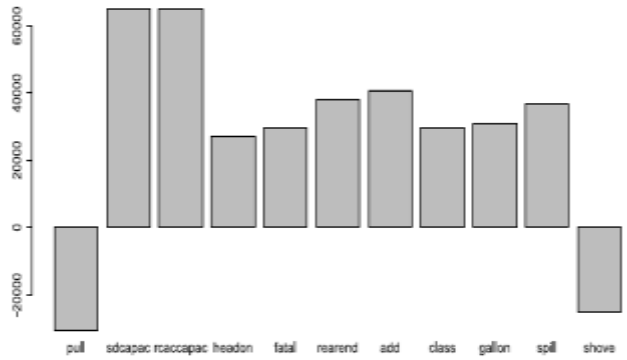

Fig. 7. Example large coefficients for narrative words obtained with PLS.

Since we answered the first two questions affirmatively, we proceed to the final question: Can text mining improve our understanding of rail accidents? Answering this question pro-vides insights into the contributors to rail accidents. Our goal in answering this question is more qualitative than quantitative. So in answering this question we are not seeking to demonstrate predictive accuracy for accident costs but rather to investigate how text mining might support the discovery of important contributors to accidents. The hope is that by understanding these contributors we can improve

safety.The results in Table 4 indicate that we should look to random forest models combined with 100 LDA topics and PLS to provide the foundation for improved understanding. Starting with LDA topics,. Some of these words are easy to interpret and associate with accident damage (e.g., track, joint, leak, gear). Others are not so obvious (e.g., est, hove, curv, rogsm).To get a feel for how these words might inform safety engineering consider "curv." This is a stem of the words curve, curves, curved, etc. There are 225 extreme accidents in the 11 years in the study period that contain "curv." These accidents had a total cost of $11 8179 658. Two example narratives (with emphasis added) that contain curv are below. The discovery of the word "curv" by LDA shows how text analysis can inform safety engineering.

Narrative 1: PULLING SOUTH, DERAILED 2 EMPTY CARS DUE TO ORIZONTAL SPLIT HEAD ON EAST RAIL ON *CURVE*.

Narrative 2: YTY60-07 SHOVING ROCK CARS INTO TYLER ASPHALT WHEN CARS DERAILED IN THE 17' *CURVE* UNDER LOAD OF 40-1 00 TON GONDOLAS. RAIL ROLLED ON THE OUTSIDE, BKTY121466, CNW350575 AND CNW350708 WERE DESTRYED. UP MAINTAINS TRACK.

To further show the importance of text analysis consider the results from PLS. Fig. 7 shows the PLS component 1 coefficients with the 10 largest absolute values for the words in the narratives. In general these words are easier to interpret than the ones in the LDA topics. One of these words, "shove," also appears as an LDA topic word, "hove," Another interesting word in the PLS component is "debri," which is the stem for "debris." This word appears in narratives for only 17 extreme accidents in the 11 year period but the total cost of these 11 accidents is $29 458 075 or almost half as much as the 265 extreme accidents with "shove." Two examples of narratives (emphasis added) with "debris" are given below.

Narrative 1: CNRBW REARENDED 2CNAAW STOPPED ON #2 ML. CPAWE ON #1 ML HIT *DEBRIS* AND DERAILED INTO CARS. UP8088/EM RSD9043/ CAPACITY 5801/SPILL 1205 GALLONS; UP6646/GE RC44AC/CAPACITY 4901/SPILL 3662 GALLONS; UP80 25/EM SD9043/CAPACITY 5801/SPILL 3122 GALLONS FUEL.

Narrative 2: 38JB605 OPERATING NORTH WITH 4 UNITS 63 LOADS AND 49 EMPTIES WHEN 28TH THROUGH 42ND CARS DERAILED. 38. PRI-MARY CAUSE: *DEBRIS* IN FLANGEWAY CHOPPER DOOR OPERATING ROD FROM 29TH HEAD CAR LANX 8124.

All three of the words discussed here illustrate another im-portant benefit of text mining: the insights provided by narrative text are not easily found through analysis of just the structured fields. Consider "debri." As noted a small number of accidents containing this word resulted in significant costs over 11 year. The second narrative example specifically calls out debris as a primary cause of the accident. Yet, debris is not listed among the 389 coded entries for accident cause. So, without careful reading of every narrative or, more practically, without text analysis the safety engineer would be unaware of this import contributor to accidents.

The other two words, “curv” and “shove” have accident cause codes that can be entered in the primary As shown both words are found in three accident cause codes. Each of these codes is a subcategory of human factors accidents, specifically, “Train operation—Human Factors.”

CONCLUSIONS AND FUTURE RESEARCH

The results presented in Section IV show that the combina-tion 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. However, there is much additional work that needs to be done to make these results of even greater use to train safety engineers. As noted several times, the performance of a chosen ensemble method can be improved with optimization. The same is true for the text mining techniques. Experiments with these techniques should yield even greater improvements in performance than those shown in Table 3.The work described in this paper only focused on incidents with extreme accident damage. As noted in Section III the OSHA report is given with BBN. Study is needed of accidents with extreme numbers of casualties to determine their contributors and the similarities and differences of these contributors to those of accidents with extreme costs.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, ide-ally, described with a probabilistic model that further enhances understanding of the contributors to accidents.

REFERENCES

[1] “Railroad safety statistics—2009 Annual report—Final,” Federal Rail-road Admin., Washington, DC, USA, Apr. 2011. [Online]. Available: http://safetydata.fra.dot.gov/OfficeofSafety/publicsite/Publications.aspx

[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 adap-tive 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 be-haviour 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 intersec-tion crashes based upon driver, vehicle and roadway surface characteris-tics,” *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 appli-cations," *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, and W. 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].Avail-able: 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-Structured Text Data Applications*. 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.

[21] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[22] H. Wold, "Estimation of principal components and related models by iter-ative least squares," in *Multivariate Anal.*, P. Krishnaiaah, Ed. New York, NY, USA: Academic, 1966, pp. 391–420.

[23] L. Li, R. D. Cook, and C. Tsai, "Partial inverse regression," *Biometrika*, vol. 94, no. 3, pp. 615–625, Aug. 2007.

[24] M. Taddy, "Multinomial inverse regression for text analysis," *J. Amer. Statist. Assoc.*, vol. 108, no. 503, 2012. [Online]. Available: http://dx.doi.org/10.1080/01621459.2012.734168

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[26] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Handbook of Latent*

Semantic Analysis, vol. 427. Hillsdale, NJ, USA: Erlbaum, 2007.

[27] D. Blei, L. Carin, and D. Dunson, "Probabilistic Topic Models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, Nov. 2010.

[28] X. Wang, M. Gerber, and D. Brown, "Automatic crime prediction using events extracted from Twitter posts," in *Proc. Int. Conf. Social Comput., Behav.-Cultural Model., Prediction*, College Park, MD, USA, Apr. 2012, pp. 231–238.

[29] X. Wang, D. E. Brown, and M. S. Gerber, "Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information," in *Proc. IEEE Intell. Security Inf.*. Washington, DC, USA, Jun. 2012, pp. 36–41.

[30] X. Wang and D. E. Brown, "The spatio-temporal modeling for criminal incidents," *Security Inf.*, vol. 1, no. 2, pp. 1–17, Feb. 2012.

[31] "Positive train control (PTC)," Federal Railroad Admin., Washington, DC, USA, 2012. [Online]. Available: http://www.fra.dot.gov/us/content/784

[32] S. Hensel, C. Hasberg, and C. Stiller, "Probabilistic rail vehicle localiza-tion with eddy current sensors in topological maps," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1525–1536, Dec. 2011.

[33] H. Dong *et al.*, "Emergency management of urban rail transportation based on parallel systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 627–636, Jun. 2012.

[34] T. Meyers, A. Stambouli, K. McClure, and D. Brod, "Risk assessment of positive train control by using simulation of rare events," *Transp. Res. Rec.*, vol. 2289, pp. 34–41, 2012.