Proceedings of the 2016 Joint Rail Conference
JRC2016
April 12-15, 2016, Columbia, SC, USA

JRC2016-5757

TEXT MINING ANALYSIS OF RAILROAD ACCIDENT INVESTIGATION REPORTS

Trefor Williams
Rutgers University
Piscataway, NJ, USA

John Betak
Collaborative Solutions, LLC
Albuquerque, NM, USA

Bridgette Findley[1]
University of Texas-San Antonio
San Antonio, TX, USA

ABSTRACT

The National Transportation Safety Board in the United States and the Transportation Safety Board of Canada publish reports about major railroad accidents. The text from these accident reports were analyzed using the text mining techniques of probabilistic topic modeling and k-means clustering to identify the recurring themes in major railroad accidents. The output from these analyses indicates that the railroad accidents can be successfully grouped into different topics. The output also suggests that recurring accident types are track defects, wheel defects, grade crossing accidents, and switching accidents. A major difference between the Canadian and U.S. reports is the finding that accidents related to bridges are found to be more prominent in the Canadian reports.

INTRODUCTION

For major railroad accidents the U.S. National Transportation Safety Board (NTSB), and the Transportation Safety Board of Canada (TSBC), produce reports that study accident causes and give recommendations for corrective actions. The advent of various text mining methods now allow the text of these reports to be studied in a systematic way to identify the major types of railroad accidents that occur. The potential exists to identify the major causes of railroad accidents to determine areas where changes in operating procedures and geometric design practices may be warranted.

Existing research in railroad accidents has focused on the analysis of numerical data using advanced statistical techniques to obtain probabilities of accident occurrence [1,2]. Text mining can now be used to derive additional information about railroad accidents to augment existing work. Data mining techniques including text mining, decision trees, support vector machines and classification have been successfully applied to other areas of transportation including the analysis of traffic accident data [3] and the prediction of highway project construction costs [4].

In this paper two methods of grouping accident reports into themes are demonstrated. This paper describes the use of probabilistic topic modeling and k-means clustering of accident report text to identify the major types of recurring railroad accidents and to indicate that text mining can be a useful tool in analyzing railroad data bases containing text.

ACCIDENT REPORTS

We obtained accident reports from the web sites of the NTSB and the TSBC. In total, reports for 312 Canadian accidents and 167 U.S. accidents were used. The Canadian reports included the period from 1991 to 2014. The U.S. reports included the period from 1993 to 2014. It was not possible to use all of the data available from the NTSB website because the PDF accident investigation files older then 1993 were not of sufficient quality to allow for optical character recognition.

Even though entire reports were input, data preparation was not an onerous task. The data mining software we employed, Rapid Miner, is able to accept files in text format for processing. All of the reports we used were PDF files downloaded from the government websites that had been originally generated using a word processor. Files of this type can be easily converted to text files using a program like Adobe

[1] Formerly Research Assistant at Rutgers University

Acrobat. Each individual report was then submitted as a file to the data mining software.

The majority of railroad investigations involve freight train accidents, such as collisions and derailments, but the NTSB places special emphasis on train accidents that involve the traveling public, such as passenger train and rail transit accidents. The criteria for investigating a railroad accident include fatalities or substantial damage. NTSB issues safety recommendations in its reports [5].

For Canadian railroad accidents, the primary criterion for determining if an occurrence in any transportation mode will be investigated is whether or not such analysis is likely to lead to a reduction of risk to persons, property, or the environment. The TSBC defines 5 classes of accidents that may merit public inquiry and investigation and these can be found at http://www.tsb.gc.ca/eng/lois-acts/evenements-occurrences.asp.

PROBABILISTIC TOPIC MODELING

Text mining algorithms now provide the capability to examine large text documents to extract new information. Text mining is often defined in the context of discovering previously unknown information that is implicit in the text but not immediately obvious [6]. In the context of the railroad accident reports produced by NTSB and TSBC there is the potential to systematically examine the text and identify major recurring accident themes and to learn more about the factors involved in major accidents. Topic modeling algorithms are statistical methods that analyze the words of unstructured original texts to automatically discover the themes that run through them. Topic models automatically organize a text collection into its major themes.

A frequently used topic-modeling algorithm is Latent Dirichlet Allocation (LDA) Details of the LDA Algorithm are given by Blei [7]. LDA is a generative probabilistic model for collections of discrete data such as text corpora. The underlying assumption of LDA is that a text document will consist of multiple themes. LDA is a three-level hierarchical Bayesian model where each item of a collection of text is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. For text modeling, the topic probabilities provide an explicit representation of a document [8]. Additionally, a topic model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. In other words, the LDA algorithm automatically identifies words that occur in the accident reports and forms then into ranked topics. In this case, we are using the LDA algorithm's ability to find themes in the NTSB and TSBC accident reports. The groups of words automatically generated (the topics) may represent the most frequent types of railroad accidents that occur. In addition, the LDA ranks the topics by their probability of occurrence and also ranks the words in each topic by their probability of occurrence within the topic.

LATENT DIRICHLET ALLOCATION RESULTS

Table 1 shows the words included in each LDA topic for the NTSB accident investigations. The table shows major words included in each topic and the number of accident reports that had the highest probability of belonging to the topic. Because accident reports contain multiple themes, an accident report may be included in several topics with varying probabilities. Therefore, the number of reports column shows the number of reports for which it was the most highly ranked topic. The words for each topic are shown in order of the ranking of their importance in the topic. For example, for Topic 15 in Table 1, the most important word is "fatigue", and "accident" is the second most important. Examination of Table 1 indicates that major topics in the NTSB reports included switching accidents, derailments involving wheels and track defects, accidents involving signals, accidents involving rail joints and grade crossing accidents. Topics 1, 3, 5, 8, and 10 include track related terms, implying that track and switch defects are major accident factors. Topic 7 is clearly a topic related to hazardous materials, yet only four accident reports are grouped in this topic.

Table 1. LDA Topics from NTSB Reports

Topic ID	Words	Number of Reports
1	train track bnsf dispatcher crew safety engineer north south railroad	11
2	accident information report board safety railroad cars car inspection factual	3
3	rail track derailment inspection car cars head wheel area defects	20
4	crossing grade accident railroad safety highway steel truck crossings traffic	10
5	track rail railroad accident joint report train derailment area ballast	9
6	engineer train signal safety passenger amtrak locomotive railroad emergency csxt	9
7	emergency fire materials hazardous train response cars car responders incident	4
8	track switch accident work yard employees safety main position cars	36
9	tank cars car cn train accident atsf rtc water pipeline	5
10	train track safety system wmata transit circuit station operator metrorail	9
11	brake train brakes air braking safety cars dynamic car locomotive	10
12	safety accident bridge training railroad report accidents management southern ntsb	3
13	train ntsb locomotive accident railroad engineer sleep collision crew conductor	12
14	train signal accident operating speed mph collision time stop red	20
15	fatigue accident safety medical tank work report employee board employees	3

The Topics 1, 6, 8, 10 and 14 include words that can be linked to training issues. For example, Topic 11 is a category related to accidents involving train braking. This implies that some accidents occur due to lack of experience in handling air brakes. Topic 1 includes the words dispatcher, crew, safety and engineer. This implies that accidents grouped in this topic occurred because of safety errors involving railroad employees.

Table 2 shows the results from the TSBC reports. Examining the results from the TSBC LDA analysis indicates some topics that differ from the LDA analysis of the NTSB reports. Topics identified include derailments due to track defects, grade crossing accidents, accidents involving signals, accidents involving wheels and wheel bearings, and switching accidents. Not surprisingly, track related defects are again found in multiple topics. However, accidents involving bridges seem more prominent in the Canadian reports with Topic 10 showing bridge as the most prominent word.

Table 2. LDA Topics from TSBC Reports

Topic ID	Words..	Number of Reports
1	wheel bearing axle car roller wheels train bearings service cp	28
2	car tank cars fracture inspection aar dangerous product goods impact	19
3	brake brakes cars air hand locomotive car applied application pipe	18
4	safety information board cn rail transportation equipment action control operation	12
5	work training locomotive train operating crew time employees railway rules	16
6	rail derailment track defects cars car derailed gauge curve head	43
7	track switch yard main movement cars cn position switches assignment	24
8	train locomotive cars locomotives end cn forces trains car emergency	19
9	crossing crossings driver railway road vehicle warning train truck safety	37
10	bridge track water inspection railway river inspections cn derailment subgrade	17
11	emergency passenger locomotive passengers safety train fuel car coach rail	11
12	train approximately mile board report occurrence time mph speed north	14
13	track rail safety inspection maintenance mile cpr cn traffic ties	23
14	train signal crew track rtc locomotive trains system stop signals	29
15	tc safety railway cars response risk rail fire tank oil	3

CLUSTERING MODELS

An alternative method of studying the grouping of railroad accident text is Clustering. Clustering techniques can be used to group the words that occur in the accident investigation reports. K-means clustering is a technique that has been widely applied in other industries but it has not previously been applied to railroad accident data. For example, clustering can be applied to automated image processing [6]. K-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [8]. These clusters reflect some mechanism that is at work in the domain from which instances are drawn. This mechanism causes some instances to bear a stronger resemblance to each other than they do to the remaining instances [9].

Applied to the railroad accident report text, the clusters produced are groups of words from the accident text that frequently occur together. Groups of words are produced that are similar to LDA topics. We have found that the text of both the Canadian and U.S. accident investigations can successfully be clustered indicating that there are recurring groupings of words that can be successfully clustered. Using the Rapid Miner open-source data mining software, the accident investigation texts were clustered using the k-means clustering process. K-means clustering works by partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Rapid Miner is a well-known toolbox for implementing various types of data mining techniques without the need to write computer code.

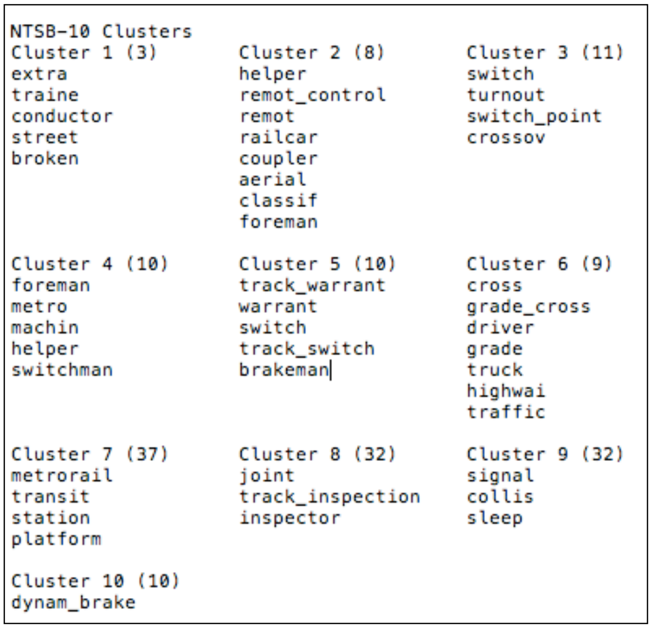
Using the text processing capabilities of Rapid Miner the text is preprocessed to translate it into a numerical format that is useable by the k-means algorithm. Six steps were used to put the text into a format useable by the clustering algorithm [9]:

1. Transform Cases. Uppercase letters were removed from the text.
2. Tokenize. The unstructured text is transformed into a sequence of tokens. Single words were defined as tokens. The tf-idf formulation was used to scale a tokens value.
3. Stemming. In this data transformation, related word tokens are normalized into a single form. For example "walking" would be transformed to "walk" [10].
4. Filter Stop words. Common words like "and" and "but" are removed by removing words on a predefined list.
5. Filter Token Length. This filter removes words that are less than three characters long.
6. Generate n-gram terms. In this process, Rapid Miner has been set to allow two word tokens to also be generated. This differs from the LDA analysis where only single words were considered.

After these transformations are completed, a matrix with each token (Words and word pairs from the accident reports.) as a column is generated. The rows are the individual accident reports and each cell represents the number of times a word occurred in the report. For this analysis, the matrix columns are the words from the accident reports. The matrix is used as the input to the k-means clustering process. Different numbers of clusters were studied and best results were found with k set to 10 for both the Canadian data and the NTSB reports.

CLUSTERING RESULTS

Figure 1 shows the results of the k-means clustering analysis of the NTSB reports. The figure also shows the number of reports grouped in a cluster. The figure shows that the clustering algorithm is able to discern groupings in the text of the accident investigations. Clusters identified from the NTSB accident investigation reports include track warrants, switches, signals, track inspection, and grade crossings. Several of the clusters possibly identify accident groupings that may warrant additional investigation. Examination of the words in each cluster indicates that the clustering algorithm automatically produced meaningful clusters of accident types and that applying the clustering technique shows that there are underlying mechanisms in the railroad accident domain. In



```
NTSB-10 Clusters
Cluster 1 (3)       Cluster 2 (8)       Cluster 3 (11)
extra               helper              switch
traine              remot_control       turnout
conductor           remot               switch_point
street              railcar             crossov
broken              coupler
                    aerial
                    classif
                    foreman

Cluster 4 (10)      Cluster 5 (10)      Cluster 6 (9)
foreman             track_warrant       cross
metro               warrant             grade_cross
machin              switch              driver
helper              track_switch        grade
switchman           brakeman            truck
                                        highwai
                                        traffic

Cluster 7 (37)      Cluster 8 (32)      Cluster 9 (32)
metrorail           joint               signal
transit             track_inspection    collis
station             inspector           sleep
platform

Cluster 10 (10)
dynam_brake
```

Figure 1. Clustering of NTSB Accident Reports

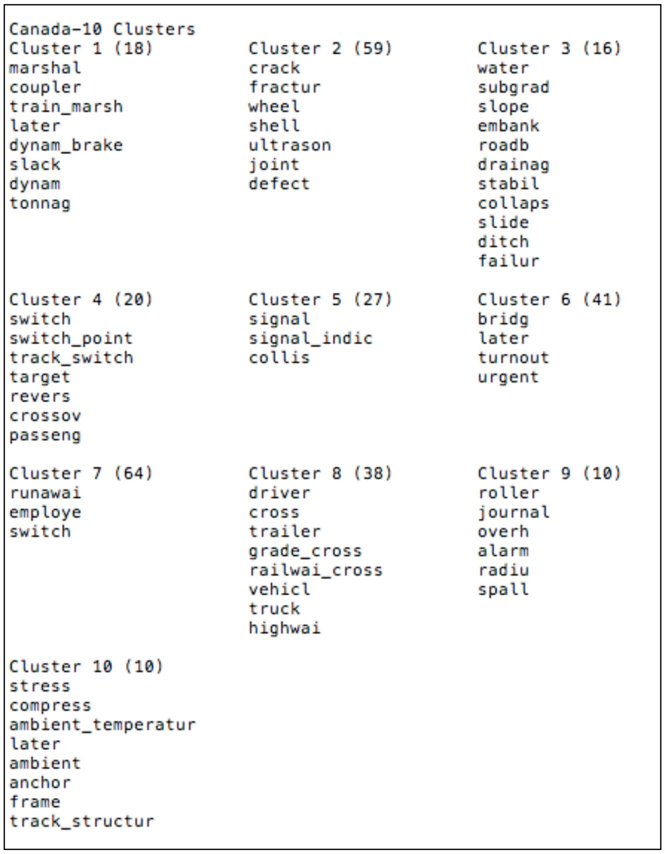
particular, Cluster 5 finds a grouping of accidents related to track warrants, Cluster 6 identifies grade crossing accidents and trucks, and Cluster 9 identifies accidents involving signals, collisions and sleep.

Figure 2 shows the words included in each of the 10 clusters found from the TSBC accident investigation. The results of the clustering from the Canadian accident investigations had some significant differences with the results obtained from the analysis of the NTSB reports. Major clusters in the Canadian data include marshaling, wheel cracks and fractures, water and drainage issues, bridges, switching, grade crossings and signals.

Cluster 1 has “dynamic braking”, “slack”, and “coupler” (which suggests inappropriate conductor or train driver operation). Possibly, these accidents indicate a need for improved train crew training. Clusters 3 and 5 deal with switches, turnouts, crossovers. While Cluster 8 includes the word “joints.” Therefore, these three clusters are finding switch-related problems for a total of 53 accident reports. The cluster containing the largest number of Canadian investigations is Cluster 7. It includes the words “runaway”, “employee” and “switch.” This cluster also suggests training issues are involved and that training to prevent runaways is needed. Another major difference found between the Canadian and American clusters is the prominence of the word “bridge” in cluster 6 of the TSBC results, while the word does not appear in any of the clusters generated from the NTSB reports. Forty-one TSBC reports were grouped in Cluster 6.

CONCLUSIONS

The analysis has shown that text mining can be a useful tool to better understand the types of railroad accidents that occur.



```
Canada-10 Clusters
Cluster 1 (18)      Cluster 2 (59)      Cluster 3 (16)
marshal             crack               water
coupler             fractur             subgrad
train_marsh         wheel               slope
later               shell               embank
dynam_brake         ultrason            roadb
slack               joint               drainag
dynam               defect              stabil
tonnag                                  collaps
                                        slide
                                        ditch
                                        failur

Cluster 4 (20)      Cluster 5 (27)      Cluster 6 (41)
switch              signal              bridg
switch_point        signal_indic        later
track_switch        collis              turnout
target                                  urgent
revers
crossov
passeng

Cluster 7 (64)      Cluster 8 (38)      Cluster 9 (10)
runawai             driver              roller
employe             cross               journal
switch              trailer             overh
                    grade_cross         alarm
                    railwai_cross       radiu
                    vehicl              spall
                    truck
                    highwai

Cluster 10 (10)
stress
compress
ambient_temperatur
later
ambient
anchor
frame
track_structur
```

Figure 2. Clusters from Canadian Accident Reports

The LDA topics generated and the words grouped in each k-means cluster have identified accident themes from the accident reports. In addition, the different methods of analyzing the text, produce similar results further confirming the existence of recurring railroad accident types. Main accident themes of rail and track defects, wheel defects, grade crossing accidents and switching accidents were identified. In addition, it was found that many of the accident categories indicate areas where additional training can reduce accidents. There are some differences found in the U.S. and Canadian reports. In particular, both the clustering and LDA analyses found that accidents involving bridges were more prominent in Canada. Additionally, accidents involving runaway cars are prominent in the Canadian clustering analysis.

These findings provide useable information that identifies major recurring types of railroad accidents. The results suggest accident types where more detailed investigations are warranted. Future research suggested by this work is to perform the text mining and clustering for different time periods to see if major accident groupings change over time. Additionally, the probabilities of word occurrence produced by the LDA algorithm can be used in predictive models to predict the number of expected accident types.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2009-ST-061-CCI002-07 and the Command, Control and Interoperability Center for Advanced Data Analysis (CCICADA) at Rutgers, the State University of New Jersey.

REFERENCES

[1] Barkan, C. P. L., Anderson, C. T., and Source, R., 2003, "Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk," Transportation Research Record, (1825) pp. 66-8.

[2] Liu, Z., Saat, M. R., Qin, X., 2013, "Analysis of U.S. Freight-Train Derailment Severity using Zero-Truncated Negative Binomial Regression and Quantile Regression," Accident Analysis and Prevention, 59 pp. 87-6.

[3] Chong, M., Abraham, A., and Paprzycki, M., 2005, "Traffic Accident Analysis using Machine Learning Paradigms," Informatica, 29 pp. 89-9.

[4] Williams, T., and Gong, J., 2014, "Predicting Construction Cost Overruns using Text Mining, Numerical Data and Ensemble Classifiers," Automation in Construction, **43**pp.23-6.

[5] National Transportation Safety Board, 2014, "2014 Annual Report to Congress," National Transportation Safety Board, Washington DC

[6] Blei, D. M., 2012, "Probabilistic Topic Models," Communications of the ACM, **55**(4) pp. 77-7.

[7] Blei, D. M., Ng, A. Y., Jordan, M. I., 2003, "Latent Dirichlet Allocation," Journal of Machine Learning Research, **3**(4) pp. 993-1022.

[8] Kanungo, T., Mount, D. M., Netanyahu, N. S., 2000, "The analysis of a simple k-means clustering algorithm," Proceedings of the sixteenth annual symposium on Computational geometry, Anonymous ACM, pp. 100-109.

[9] Witten, I.H., and Eibe, F., 2005, "Data mining: practical machine learning tools and techniques," Morgan Kaufmann Publishers, San Francisco.

[10] Weiss, S., Indurkhya, N., and Zhang, T., 2010, "Predictive text mining: a practical guide," Springer-Verlag, London.