Text Mining the Contributors to Rail Accidents

Donald E. Brown, *Fellow, IEEE*

***Abstract*—Rail accidents represent an important safety concern for the transportation industry in many countries. In the 11 years from 2001 to 2012, the U.S. had more than 40 000 rail accidents that cost more than $45 million. While most of the accidents during this period had very little cost, about 5200 had damages in excess of $141 500. To better understand the contributors to these extreme accidents, 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. While a number of studies have looked at the fixed fields, none have done an extensive analysis of the narratives. 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 results show that predictive accuracy for accident costs significantly improves through the use of features found by text mining and predictive accuracy further improves through the use of modern ensemble methods. Importantly, this study 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 only fixed field analysis of the accident reports.**

***Index Terms*—Rail safety, safety engineering, latent Dirichlet allocation, partial least squares, random forests.**

I. 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 make them difficult to read by personnel from outside the industry.

The FRA makes the data from these accident reports available on-line at [2]. Over the last 12 years the number of fields have 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 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.

Manuscript received February 28, 2015; revised June 14, 2015 and August 3, 2015; accepted August 18, 2015. Date of publication September 25, 2015; date of current version January 29, 2016. The Associate Editor for this paper was F.-Y. Wang.

The author is with the Data Science Institute, University of Virginia, Charlottesville, VA 22904 USA (e-mail: brown@virginia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2015.2472580

II. Related Work

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.

One of the most well-studied areas of rail safety concerns rail crossings by roadways. A recent application of fuzzy sets and clustering to guide the selection of rail crossings for active safety systems (e.g., bells, lights, and barriers) is in [3]. Tey *et al.* [4] describe the use of logistic regression and mixed regression to model the behavior of drivers at railway crossings. The paper by Akin and Akbas [5] describes the use of neural networks to model intersection crashes and intersection characteristics, such as, lighting, surface materials, etc. Taken together these papers show the use data mining to better understand the factors that can influence and improve safety at rail crossings.

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 provide methods to detect anomalies in massive amounts of traffic data and then cluster these detections according to different attributes. Similarly D'Andrea *et al.* mined Twitter and used support vector machines to detect traffic events [7]. Another recent application of text mining is to license plate recognition [8]. These authors use Levenshtein text-mining 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.

Other researchers have used text mining of reports. In this category Nayak *et al.* [13] used text mining to analyze road crash data in Australia. For text mining they employed Leximancer concept mapping as implemented in a commercial product available through Leximancer [14]. This technique uses naive Bayes classifiers to identify concepts from co-occurring words. Smith and Humphreys [15] provide an overview description of Leximancer concept mapping, as well as, an assessment of whether the approach is "grounded in practice" using The Personal Memoirs of U.S. Grant [16]. The paper by Nayak *et al.* [13] uses the Leximancer product to produce concepts from road crash accident reports in Australia. They conclude that the concepts found by this approach do improve understanding of the common causes of road accidents.

Other researchers [17] combined text retrieval methods and link analysis to study U.S. National Transportation Safety Board (NTSB) aviation accident reports from 2001 to 2003 (about 3k reports). They used concept chain querying to detect links between topics across text documents. They focused on retrieval not prediction. These same data were the focus of a chapter in a recent book [18]. The chapter gives a tutorial on using the commercial statistical package STATISTICA to improve predictive performance of models for accident outcomes. They use a four element categorical variable that classifies the accidents as substantial, destroyed, minor, and none as the response. They combine structured text with key words found from text mining with boosted trees (a method discussed below) and show some improvement in an ROC curve for classification of substantial damage over models with just structured text inputs.

The work we present in this paper differs from and extends previous work in the transportation safety literature in at least four ways. First this paper describes a broader comparison of techniques than previous studies. Specifically Section V gives results for comparisons between no text mining and two contemporary approaches to text mining in combination with three approaches to supervised learning. This three by three design provides a broader range of evaluation than any previous study. Second, this paper focuses on rail accident reports over a longer time span than other studies; namely, 11 years. Third none of the text mining analytics described here have previously been applied to rail accident damage assessment. Finally, the methods in this paper are all available through open source software (R) and the code used in the analysis is also freely and openly available.

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. Boosting provides an iterative approach to combining the outputs from a sequence of simple or *weak* learners to generate a more accurate combined estimate. In principle, the *weak* learners can be any supervised learning procedure, although for the results here we use classification trees (see [19]).

Fundamentally the approach can be considered as an additive expansion of simple basis functions (see [20]). For a dependent or response variable, f, (e.g., cost of an accident) which is a function of a vector of predictor or independent variables, x, this expansion has the form

$$f(x) = \sum_{m=1}^{M} \beta_m b(x, \gamma_m). \tag{1}$$

The $\beta_m, m = 1, \ldots, M$, are basis function coefficients and the $b(x, \gamma_m) \in \Re$ are functions of the vector argument, x, with parameters γ.

Bagging or bootstrap aggregation builds multiple models by resampling subsets of the original data. The subsets are created by randomly removing some of the original predictor variables. Again the estimates from the resulting models are combined to produce one overall estimate.

The bagging method we use in this work is random forests [21] and it again combines results from multiple classification trees. Specifically, the random forest algorithm builds a group or forest of tree-based models where each tree uses a subset

of the predictor variables with a resampling of the training data. The predictions from each tree are combined using simple averaging or medians. We can measure variable importance in random forests by analyzing tree performance (e.g., RMSE) for trees without each of the variables. Large increases in RMSE for trees without a variable suggest that the variable not included is important in predicting accident costs and, hence, it is a contributor to understanding accident damage.

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 semistructured 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 semistructured.

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).

In this paper we use a current extension on the basic term-document matrix. We employ probabilistic indexing and topic models. The indexing models assume a document d contains a topic z with probability, $\Pr(z|d)$. Each topic consists of certain words, $w_i, i = 1, \ldots, N$, where N is the number of possible words. A document is formed by choosing the words for the topics according to probabilities, $P(w_i|z)$. Typically in text mining the latent indices are found using singular value decomposition (SVD) [12]. Rather than use SVD we employ partial least squares (PLS) [22]. PLS has been used in information retrieval [23] and text analysis [24] and these preliminary results have been promising.

PLS is similar to principal components in that it constructs latent variables that are linear combinations of predictors. Unlike principal components, which use only the predictor variables, PLS linear combinations are formed to maximize the covariance between the predictor and the response variables. So, for a matrix of predictor variables, X, and response variables, Y, PLS extracts factors and loadings from $Y^T X X^T Y$. This approach works well in supervised learning problems with large numbers of predictor variables, such as text analysis, where the number of words is very large.

The probabilistic topic model we use is known as Latent Dirichlet Allocation (LDA) [25]. LDA performs well for different tasks like document organization and image labeling [26], [27]. As described in [25] we can represent the standard LDA model as a generative process for documents composed of topics that are themselves composed of words. The generative process for each document d in a collection D can be described as follows. First, draw T topics from a Dirichlet distribution $\beta_t \sim \mathrm{Dir}_V(\eta)$, where a topic is a distribution over V words. Second, for each document d, draw topic proportions from another Dirichlet distribution: $\theta_d \sim \mathrm{Dir}_K(\alpha)$. Third, to obtain each word $w_{d,n}$ in the document d, draw a topic $z_{d,n}|\theta_d \sim \mathrm{Multinomial}(\theta_d)$ and then draw a word $w_{d,n}|z_{d,n}, \beta_{1:K} \sim \mathrm{Multinomial}(\beta_{z_{d,n}})$.

LDA is a "bag of words" approach that uses no semantic content in the documents. Although we will not use it for the results in this paper, we have extended the basic LDA approach to include simple semantics [28]. Of greater applicability to our work here we have combined LDA and generalized additive models to understand and more accurately predict incidents, such as, hit-and-run accidents [29] and [30]. These results provided the foundation for the work on analyzing other critical incidents, such as the train accidents described in this paper.

Also, of relevance to the work in this paper is the research and development on Positive Train Control (PTC). The National Transportation Safety Board (NTSB) has named PTC as one of its "most-wanted" initiatives for national transportation safety [31]. Beginning in 2001 the railroads deployed components of PTC on small sections of track to test and validate its usefulness. A complete list of these deployments is in [31]. PTC requires a number of technologies, some of which have not been deployed. Research and development results are beginning to produce these needed technologies. Henzel [32] describes the use of eddy current sensors to provide more precise location of trains for positive control. Parallel control for emergency response is presented in [33]. Meyers *et al.* [34] describe risk assessment methods for evaluating the safety of PTC. They also discuss the many challenges in performing this risk assessment. The work we describe in the subsequent sections of this paper can better inform these risk assessments. In particular, the text mining approach we describe can enable a better understanding of the characteristics of accidents that PTC may prevent and those that it cannot.

III. 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.

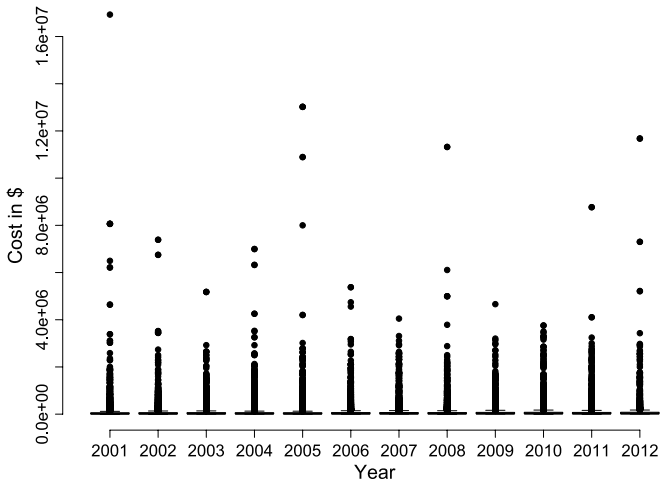


Fig. 1. Box plots of total accident damage from 2001–2011.

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.

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.

IV. Data Structuring and Cleaning

Before discussing the analytics used in this study we need to further describe how we structured and cleaned the data. As noted in Section III there are 5471 unduplicated, extreme damage accidents after removing the one that occurred due to the attacks on 9/11. Further data cleaning described in this section reduced the data set by 2 additional points to 5469.

We randomly divided the reports into training and test sets. The training set contains 3667 accidents and the test set has 1802. The total accident damage for observations in the test set ranges from $143.2k to $13M with a median of $342.2k. The total accident damage for accidents in the test set ranges from $143.4k to $13M with a median of $342.4k. As noted below we made several small changes from the random draw to better balance the test set.

From the base FRA structured data we formed 4 numeric predictor variables: 1) Number of cars; 2) Number of operators (crew size); 3) Speed at the time of the accident; and 4) Weight. We also formed 4 categorical predictors: 5) Cause (as shown in Table I); 6) Train type (as shown in Table II); 7) Accident Type (as shown in Table III); and Track type (as shown in Table IV). Since categorical variables require special handling for modeling it is important to understand the structuring and cleaning of each of them.

TABLE I
First Character Cause Codes & Extreme Accident Frequency

Code	Cause	Frequency (%)
T	Rack, Roadbed and Structures	2,180 (40)
S	Signal and Communications	50 (1)
M	Miscellaneous	905 (17)
H	Train operation - Human Factors	1,389 (25)
E	Mechanical and Electrical	945 (17)

TABLE II
Train Types

Code	Type	Frequency (%)
1	Freight	4,067 (74)
2	Passenger	195 (4)
3	Commuter	40 (1)
4	Work	28 (0)
5	Single car	39 (1)
6	Cut of cars	185 (3)
7	Yard/Switching	762 (14)
8	Light locomotive	76 (1)
9	Maintenance/Inspection	33 (0)
A	Maintenance of way	39 (1)
B	Other B	3 (0)
C	Other C	0 (0)
D	Other D	2 (0)

TABLE III
Accident Types

Code	Type	Frequency (%)
1	Derailment	4,102 (75)
2	Head-on collision	93 (2)
3	Rear-end collision	160 (3)
4	Side collision	316 (6)
5	Raking collision	56 (1)
6	Broken train collision	21 (0)
7	Highway-rail crossing	220 (4)
8	Railroad grade crossing	2 (0)
9	Obstruction	78 (1)
10	Explosive detonation	0 (0)
11	Fire/violent rupture	70 (1)
12	Other impacts	263 (5)
13	Other as in narrative	88 (2)

TABLE IV
Track Types

Code	Cause	Frequency (%)
1	Main	3,858 (71)
2	Yard	1,254 (23)
3	Siding	190 (3)
4	Industry	167 (3)

As noted in Section I Cause is actually a four character code which indicates a hierarchical decomposition of causal factors. For instance, E0 indicates a brake failure and E02L indicates a broken brake pipe or connection. The first letter of this code takes one of the values T, S, M, H, and E with the meanings shown in Table I. For this study we used only the coarse categorization given the by the first character.

Table I also shows the frequency of occurrence of cause types in the extreme damage data set. The physical infrastructure of the network, which includes the tracks, roadbed, bridges and other structures, accounts for about 40% of the extreme damage accidents. Human factors is the second most common cause and is cited in 25% of the extreme accidents.

Another categorical variable in this study is train type. The extreme event accidents involve 13 different train types with the labels and frequencies shown in Table II. Notice that the last 3 labels represent types not defined by the previous categories. Since there was only one accident with other type C, this was removed from the data (so is it shown as 0 in Table II). Additionally, since other type D occurred twice, we randomly assigned one of them to the test set and the other to the training set. The random draw of the test set placed one of the accidents involving other type B in the test set and left the remaining two in the training set. Notice that almost three quarters of the extreme damage accidents involve freight trains.

The third categorical variable used in this study is accident type. There are 13 different types of accidents in the reports. Table III shows the codes, type of accident and frequency of occurrence in the extreme damage data. The most common extreme accident with over 4000 incidents is derailment (coded 1). The remaining 12 types together account for roughly 24% of the extreme damage accidents. The accident reports contained only one instance of an explosion. So, this accident was removed from the study (and is shown as 0 in Table III). There were two instances of railroad grade crossing accidents and one was randomly put in the test set and the other in the training set.

The final categorical variable in this study is type of track which is coded 1–4 as shown in Table IV. 70 of the accident reports contained no entry for type of track. We used k-nearest neighbor with $k = 50$ to impute the missing values. Table IV shows that more than 70% of the extreme damage accidents occurred on main track.

The categorical variables can have interaction effects on the cost of an extreme accident. To illustrate this, Fig. 2 shows the accident damage in the extreme accidents with the recorded cause and accident type. For this graphic we show two levels for accident type, derailment and non-derailment or other. As this figure shows there are more costly derailment accidents caused by track and roadbed problems and mechanical and electrical failures. On the other hand, human factors seem to cause more costly non-derailment accidents. In general accidents caused by signaling and communications have lower costs, while accidents with human factors and miscellaneous causes account for the 14 most costly accidents in the period studied.

The narrative data in the 15 free text fields were combined into one document per accident. Before performing text mining we removed numbers and the stop words from each document.

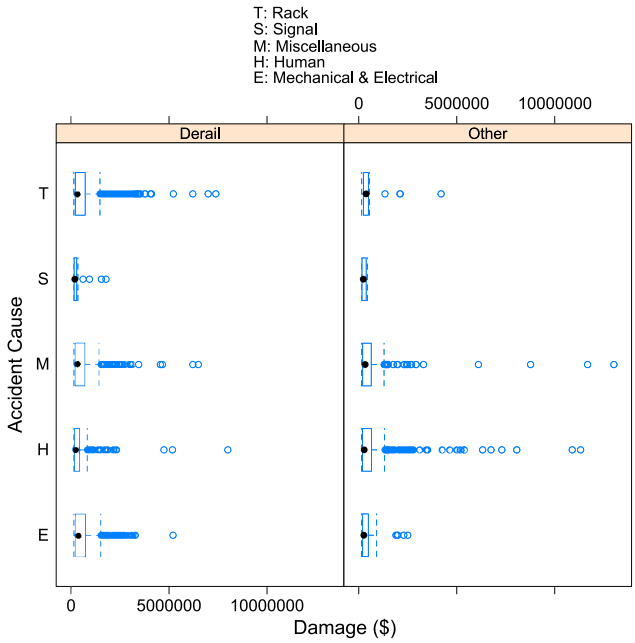


Fig. 2. Box plots of accident damage by cause 2001–2012.

We also stemmed the words in the documents. We then created a term-document matrix with cell values that give the frequency of the terms in each accident document weighted by the frequency of the terms in all documents. For this study we used an inverse frequency weighting or $\log(N/n_i)$ where N is the total number of documents (accidents) and n_i is the number of occurrences of the i^{textth} term. This matrix is the input to the analysis of text as described in the next section.

V. 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 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 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.

TABLE V
UNIQUE WORDS IN THE 10 TOPICS IN THE ACCIDENT REPORTS

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

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 affirmative then which text and non-text features best predict accident severity. Answering this last questions will enable preliminary understanding of contributors to rail accidents.

Once the data were structured and cleaned (Section IV) we proceeded to address the first study question: Do the narratives 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 V 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 material 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 common (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 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,

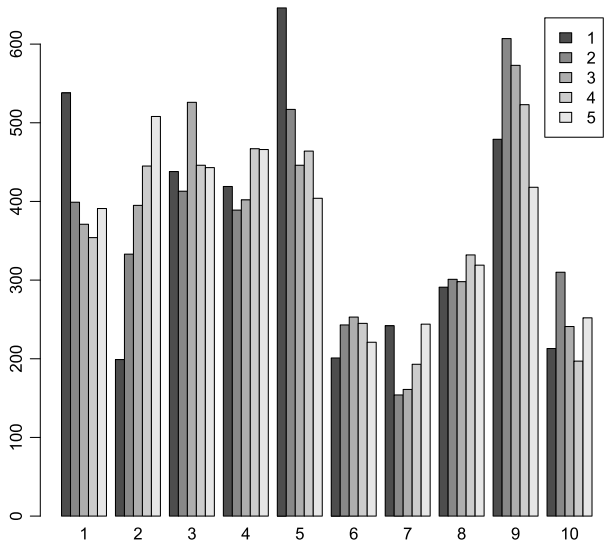


Fig. 3. Frequency of 10 topics in the accident reports.

TABLE VI
SIGNIFICANT TOPICS IN THE OLS MODEL

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

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 VI 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 in topic 4 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.

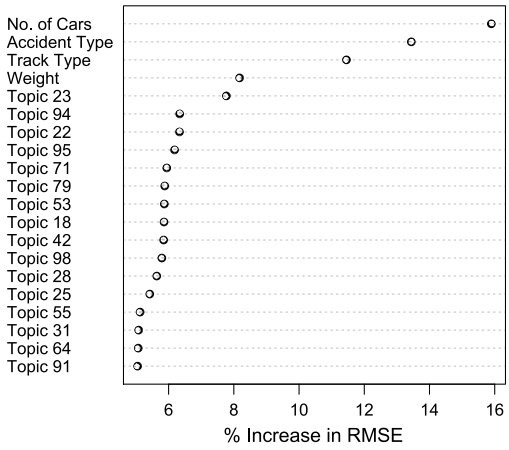


Fig. 4. Variable importance for the random forest model with 100 LDA topics.

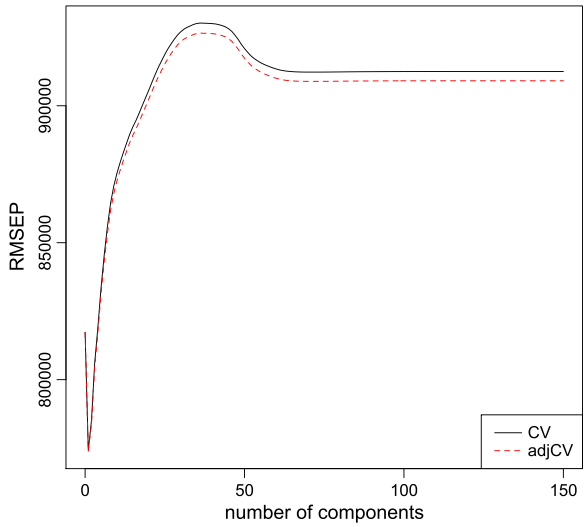


Fig. 5. RMSE from cross-validation with different numbers of components.

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 removed. 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

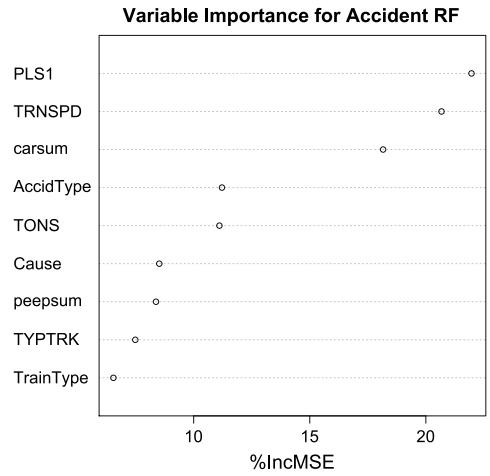


Fig. 6. Variable importance for the random forest model with the PLS predictor.

TABLE VII
TEXT MINING RMSE

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

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.

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. 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 models (see Fig. 6).

Table VII 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 deicide if their use is appropriate

TABLE VIII
IMPORTANT LDA TOPICS IN THE RANDOM FOREST MODEL

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

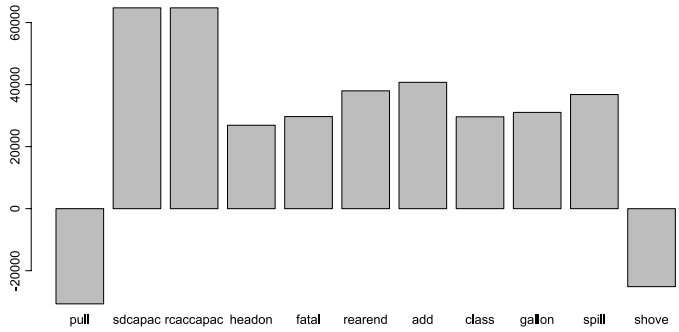


Fig. 7. Example large coefficients for narrative words obtained with PLS.

for transportation safety modeling as described in this paper. The results show that it is.

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 provides 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 VII 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, Table VIII shows words in five of the more important topics identified in Fig. 4. 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 HORIZONTAL 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," in Topic 22 in Table VIII. Shove occurs in 265 extreme accidents over the 11 year period with a total cost of $67 307 107.

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. PRIMARY CAUSE: *DEBRIS* IN FLANGEWAY CHOPPER DOOR OPERATING ROD FROM 29TH HEAD CAR LANX 8124.

All three of the words discussed here illustrate another important 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 and contributing cause fields in the accident report. Table IX shows the relevant codes. 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."

However, the coding of the cause of the accidents does not necessarily match the narrative. Fig. 8 shows the high level coding of the accidents where shove was contained in the narrative. While most of these accidents were coded as human factors caused accidents, only 83 of them had one of the codes associated with shove as the primary cause and 7 had one of

TABLE IX
CODES FOR CURVE AND SHOVE

Code	Definition
H301:	Car(s) shoved out and left out of clear
H306:	Shoving movement, absence of man on or at leading end of movement
H307:	Shoving movement, man on or at leading end of movement, failure to control
H505:	Lateral drawbar force on curve excessive, train handling
H506:	Lateral drawbar force on curve excessive, train makeup
H507:	Lateral drawbar force on curve excessive, car geometry (short car/long car combination)

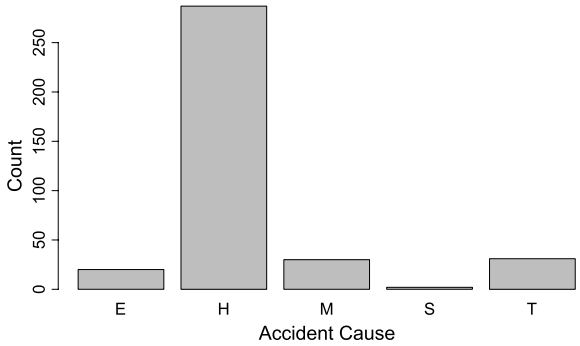


Fig. 8. Coded causes for accidents with shove in the narratives.

Fig. 9. Coded causes for accidents with curve in the narratives.

these codes as a contributing cause. So in only about $1/3$ of the cases would a safety engineer looking only at the fixed fields know that shoving was a factor in the accident.

The situation is even more opaque for curve. As Fig. 9 shows most of the accidents with curve in the narrative are not even coded as human factors accidents. Only 7 accidents have one of the curve codes as a primary cause and only 1 accident has a curve code as a contributing cause. The means that in only about 3% of the cases would a safety engineering using the fixed fields for analysis be aware of curvature as relevant to rail accidents.

VI. CONCLUSIONS AND FUTURE RESEARCH

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.

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 VII.

The work described in this paper only focused on incidents with extreme accident damage. As noted in Section III the cost of accidents is not highly correlated with death and injury. 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, ideally, described with a probabilistic model that further enhances understanding of the contributors to accidents.

Finally, as described in Section V the work here used standard methods to clean the narratives. However, train accident narratives use jargon common to the rail transport industry and classical stemming and stop word removal do not necessarily 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 from text that takes advantage of language characteristics particular to the rail transport industry.

REFERENCES

[1] "Railroad safety statistics—2009 Annual report—Final," Federal Railroad 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 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, 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]. 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-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 iterative 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.

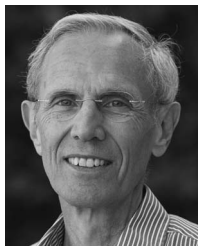
[30] X. Wang and D. E. Brown, "The spatio-temporal modeling for criminal incidents," *Security Inf.*, vol. 1, no. 2, pp. 1–17, Feb. 2012.

[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 localization 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.



Donald E. Brown (F'01) received the B.S. degree from the U.S. Military Academy, West Point, NY, USA, the M.S. and M.E. degrees from the University of California, Berkeley, CA, USA, and the Ph.D. degree from the University of Michigan, Ann Arbor, MI, USA. He is currently the Director of the Data Science Institute, University of Virginia, Charlottesville, VA, USA, and the William Stansfield Calcott Professor of Systems and Information Engineering. His research focuses on techniques that enable the combination of different types of data for prediction and engineering design. Dr. Brown was a recipient of the IEEE Joseph Wohl Career Achievement Award and the IEEE Norbert Wiener Award for Outstanding Research.