DETECTING POTENTIAL VEHICLE CONCERNS USING NATURAL LANGUAGE PROCESSING APPLIED TO AUTOMOTIVE BIG DATA

Monica G. Eboli
Catherine M. Maberry
Ian A. Gibbs
Ramsi Haddad
Emerging Issues Analytics, Global Vehicle Safety, General Motors
United States of America

Paper Number 19-000000

ABSTRACT

A large volume of unstructured data exists in the automotive industry and needs to be analyzed to detect potential vehicle concerns. Much of this data is textual in nature since customer complaints are made through call center interactions and warranty repairs. Current approaches to detect potential vehicle concerns in text data include various keyword search methods. In this paper, we apply Natural Language Processing (NLP) and shallow machine learning methods on text data to create classifiers to detect the potential vehicle concern of airbag non-deployment. For this potential vehicle concern, we show the performance of multinomial Naïve Bayes (NB), Support Vector Machine (SVM) and Gradient Boosted Trees (GBT) classifiers against keyword search methods. We present challenges of classification model development related to the nature of automotive data and limited training data. Our findings provide insights on robust text classification approaches that can improve identification of potential vehicle concerns.

INTRODUCTION

Automotive corporations and the U.S. federal government [1] are driving improvements in product safety through the collection and analysis of both structured and unstructured (text) data. Despite their efforts, a common problem facing large corporations today is how to extract meaningful insights about product safety from large volumes of unstructured, noisy data that they have accumulated in many disparate systems. These data systems present clear opportunities for analyzing actionable information regarding product complaints and potential defects, but are commonly referred to as "dark data" because they are not easily analyzed due to their unstructured nature [2]. Consider the text data of vehicle warranty claims, call center transactions, and product complaints on social media; these sources all contain valuable information that may describe potential vehicle concerns, but are not represented in a relational structure that can be easily queried. In addition, large corporations, such as General Motors (GM), spend resources maintaining these data systems and encounter challenges efficiently extracting actionable information from them because these systems were not originally created for safety event detection.

At the same time, the U.S. federal government has created several incident reporting and complaint collection systems for a variety of industries: the Food and Drug Administration's (FDA) Adverse Event Reporting System (AERS) [3] for the pharmaceutical industry, the Federal Aviation Administration's (FAA) Aviation Safety Reporting System (ASRS) [4] for the aviation industry, and the National Highway Traffic Safety Administration's (NHTSA) Vehicle Owner Questionnaire (VOQ) [5] and Transportation Recall Enhancement, Accountability and Documentation (TREAD) [6] for the automotive industry. The effort by the U.S. federal government in creating these systems is due to the public interest in ensuring that products created by these industries are safe for consumers. Yet, the fundamental problem still exists; all of these systems contain large volumes of dark data because they all have varying degrees of unstructured data in the form of text.

Ultimately, private industry and the U.S. federal government have a vested interest in developing techniques for the transformation of unstructured data into structured data to facilitate detection and monitoring of potential vehicle concerns within the automotive industry. For both private industry and the government there is a need to produce statistics that provide an overview of how certain types of product failures are reduced in response to their actions

(product recalls, bulletins, etc.), but also to identify trends that should be addressed by those actions in the first place [7].

In this paper we will describe GM's efforts in utilizing Natural Language Processing (NLP) and shallow machine learning methods to transform unstructured text data into structured data that describes potential vehicle concerns [8]. The specific focus will be on the issue of airbag non-deployment, but we have expanded our approach to many other significant potential vehicle concerns. To the best of our knowledge, we believe this publication to be the first instance of NLP and shallow machine learning to be presented in the context of safety monitoring within the automotive industry.

Data for this effort originates from a variety of sources ranging from GM internal data (warranty claims, customer call center transactions) to public data managed by the U.S. government (NHTSA VOQ). For the scope of detecting narratives that involve airbag non-deployment, results presented in this paper will be constrained to NHTSA VOQ and TREAD data. Given their utility for the task of text classification, we present results for multinomial naïve Bayes, support vector machine and gradient boosted trees classifiers compared to traditional keyword-based pattern matching methods. We also discuss fundamental components of developing these classifiers, such as training set development and NLP pipeline development. Through the work described in this paper, it is our hope that we significantly advance the concept of detection and monitoring of potential vehicle concerns within the automotive industry.

BACKGROUND

In order to improve collision outcomes for occupants, front airbags work in concert with seat belts to restrain driver and front passenger seat occupants by inflating when vehicle sensors, measuring acceleration at various vehicle locations, indicate a moderate to severe frontal impact [9]. Airbag deployment is controlled during collision by a complex algorithm that assesses data from multiple vehicle sensors, such as occupant presence, change in velocity (delta V), time to max delta V, principal direction of force (PDOF) and others, to determine whether frontal airbags should deploy [9]. The complexity of the algorithm may contradict the assumption by vehicle occupants that the airbags were faulty in not deploying during a collision.

The value in detecting potential airbag non-deployment events is to enable investigation into these potential vehicle safety concerns by vehicle safety engineers. In the effort to decrease fatalities in frontal collisions related to potential system failure of frontal airbags, reliable, accurate, and robust detection methods in unstructured data are a critical first step.

Data Sources for Classification

Human-labeled datasets required for supervised methods were sourced from two corpora – NHTSA VOQ and TREAD data.

NHTSA VOQ is a publicly available dataset and consists of customer safety complaints about automotive products.

> JULY 20, 2016, WE WERE REAR ENDED BY A HONDA ACCORD DOING 40 MPH WHILE WE WERE STOPPED AT A LIGHT. THE IMPACT SLAMMED US INTO THE CHEVY SILVERADO IN FRONT OF US. NEITHER AIRBAG DEPLOYED.

The customer complaints dataset "contains all safety-related defect complaints received by NHTSA since January 1, 1995" [5]. NHTSA receives complaint documents from various sources including: (1) online submissions from by the general public, (2) vehicle owner questionnaire submitted by the general public, (3) the auto safety hotline submitted by the hotline operator and (4) consumer letters.

Figure 1. Sample NHTSA VOQ document describing a potential airbag non-deployment event.

TREAD is a GM-internal data source that is of interest for potential vehicle concern monitoring because it consolidates data from many disparate systems. The TREAD Act, which describes the requirements for GM's TREAD data system, was created in response to the Ford/Firestone issue and significantly changed the information automotive OEMs must report to the U.S. federal government [6]. The TREAD Act requires manufacturers to submit information related to substantially similar vehicles that may have different names, foreign fatalities, notices of foreign safety recalls and other safety campaign information and Early Warning Reporting (EWR). The EWR component of the TREAD Act results in GM compiling data from many disparate systems. As such, GM's TREAD data provides a broad cross-section of data from many business areas and large volumes of unstructured text data. The centralized nature of this data source is the primary motivating factor for its use in analyzing potential vehicle concerns.

Prior Detection Methods

Prior to the work described in this paper, potential vehicle concerns were monitored in both GM internal and public data sources using IBM Watson Explorer (previously known as IBM Content Analytics). Watson Explorer provides a proprietary version of Apache Lucene that employs an Unstructured Information Management Architecture (UIMA) pipeline for indexing, searching and analyzing text data. Watson Explorer annotators and dictionaries were used as the primary method for transforming unstructured data into structured data [10].

Annotators are compound rule sets for labeling text documents for a specific potential vehicle concern. Each rule within an annotator is designed to match a specific pattern of text. The pattern of text defined by an annotator may be a specific sequence or utilize Boolean logic to detect the presence of one or more words in a sentence, paragraph or document. Dictionaries are used to define the terms used in pattern matching.

An annotator defining airbag non-deployment could be applied to the document in Figure 1. Such an annotator would include a sequence rule matching a pattern of negation ("NEITHER"), followed by the airbag system ("AIRBAG"), followed by a mention of deployment ("DEPLOYED"). This complex rule would require a negation dictionary, an airbag system dictionary, and a deployment dictionary. All dictionaries would be required to include synonyms, misspellings and alternative forms of the base terms. For example, one would need to account for representations of the airbag system as "AIRBAG", "AIRBAGS", "AIR BAG", "SUPPLEMENTAL RESTRAINT SYSTEM", etc. A document stating, "AIRBAGS NEVER DEPLOYED" would not be flagged by the sequence rule because the airbag system is now in the first position of the pattern and negation is in the second position. To mitigate this issue, a Boolean rule would need to be developed that looks for airbag system, negation and deployment in the same sentence while ignoring sequencing. Utilizing Boolean logic loosens the pattern and can lead to tradeoffs between false positives and false negatives.

An IBM Watson Explorer annotator was designed to detect airbag non-deployment in NHTSA VOQ and TREAD. This annotator was developed using subject matter expertise and the same training data was used to develop machine learning methods described in later sections. The airbag non-deployment annotator serves as the baseline for comparing performance of machine learning methods.

METHODS

All document classification models combine supervised machine learning classification with the addition of standard Natural Language Processing (NLP) techniques to effectively transform unstructured text data into structured data. The general process used in this exercise consisted of: (1) training set development, (2) text preprocessing, (3) model development, and (4) and model assessment.

Training Data
Training sets were first developed to facilitate machine learning model development. Vehicle safety experts at GM collaborated to define airbag non-deployment events and related document characteristics. Potentially related document characteristics vary by dataset and consist of inclusion and exclusion criteria focused on terminology, key words, and circumstances described. Resulting definitions were documented and used as the basis for training data sampling and labeling.

The training set was developed for airbag non-deployment using NHTSA VOQ and TREAD data. Training samples from NHTSA VOQ were not restricted to GM manufactured vehicles since airbag non-deployment allegations are found among other automotive manufacturers. Qualitatively, the customer complaints describing airbag non-deployment within NHTSA VOQ were notably homogeneous across automotive manufacturers. The resultant labeled training set contained 2003 documents of which 1000 were sourced from NHTSA VOQ and 1003 were sourced from TREAD data. Within the overall labeled training set, there were 916 positive examples of airbag non-deployment and 1087 negative examples.

Text Preprocessing
Text preprocessing is commonly implemented in text analytics solutions. The goal of text preprocessing is to increase the homogeneity of the corpus through data standardization, aggregation of semantically similar terms and removal of words that contribute little to analysis.

Multiple text preprocessing techniques were used in this exercise including: (1) case standardization, (2) stop word removal, (3) contraction expansion, (4) lemmatization, (5) standardization of dollar values, units of speed and numbers and (6) removal of non-alpha-numeric characters. These techniques were applied to the labeled datasets prior to analysis using a custom developed Python program.

Model Development
NLP and machine learning pipelines were developed to evaluate the use of different machine learning methods to detect airbag non-deployment narratives. In text analytics, it is common that most effort in model development is spent on feature extraction. Features in this context are individual measurable properties extracted from the text that will be used to predict labels on documents (classification) [11]. Furthermore, features extracted in text analytics may include n-grams which are sequences of two or more words (i.e. “red wine” is a bi-gram, “engine control module” is a tri-gram). Pipelines are a collection of processes that can be used to transform data and fit classifiers in a defined sequence. Figure 2 depicts the steps included in the model fitting pipelines. Steps include: (1) feature extraction, (2) feature encoding, (3) feature selection and (4) classifier fitting. Pipeline parameters were tuned using grid search, evaluated using industry standard metrics, tested for generalizability using cross validation, and developed using open source data science technologies.

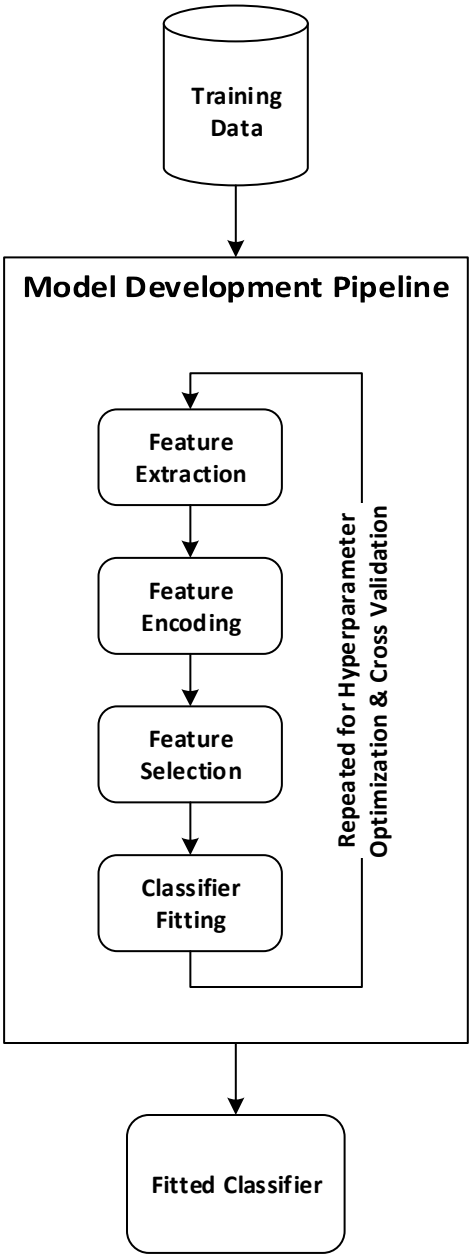

Figure 2. Model development pipeline.

Feature Extraction & Encoding: Features were extracted from the corpus as individual words and also included n-grams, which are a sequence of n adjoining words in a document. For example, if a document read "NEITHER AIRBAG DEPLOYED" bi-gram extraction, where n equals 2, would yield "NEITHER AIRBAG" and "AIRBAG DEPLOYED." Text features were restricted to bag of words (BOW) representations [12].

All features extracted from the labeled data were encoded using Term Frequency-Inverse Document Frequency (TF-IDF, Equation 1). TF-IDF is a feature encoding technique that weights how important a word is in a document within a corpus. Term Frequency ($tf_{t,d}$) describes the frequency of a term or token (t) within a particular document (d) while Inverse Document Frequency (idf_t) describes the inverse of the number of documents in the corpus that contain the term (t)

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \quad \text{(Equation 1)}$$

Since both TF and IDF are raw frequency measures, it is common to utilize re-scaled and smoothed variants. Sublinear TF ($wf_{t,d}$) is described in Equation 2. A smoothed version of IDF is described in Equation 3 where n_d is the number of documents in the corpus.

$$wf_{t,d} = \begin{cases} 1 + logtf_{t,d} & if\ tf_{t,d} > 0 \\ 0 & otherwise \end{cases} \quad \text{(Equation 2)}$$

$$idf_t = log\,\frac{1+n_d}{1+df_t} + 1 \quad \text{(Equation 3)}$$

In general, frequently occurring terms have low TF-IDF weight and rare terms have a high TF-IDF weight. TF-IDF can be used to devalue words such as "VEHICLE" and "DRIVE" which frequently appear in automotive data but were not filtered out during text preprocessing. Furthermore, "DEPLOY" occurs in a considerable number of documents in the training data and would have a lower TF-IDF weight compared to other words.

Feature Selection: Feature selection is applied prior to classifier training to select the most relevant features for classification. The chi-squared test measures the dependence of the features on the classes being modeled. Features identified as being independent of a class will have a low chi-square test statistic and are not considered useful for classification. Each extracted feature is ranked by the chi-square test statistic from largest to smallest and the top q% of features are used for the model algorithm. The proportion of features selected, dictated by q, is one of the parameters varied in the model fitting pipelines.

Binary Classifiers: Given that detection of airbag non-deployment is a signal detection problem with a binary outcome/class (presence or absence, 1 or 0, yes or no), several binary classification machine learning algorithms were utilized in the model fitting pipelines. The objective in utilizing these algorithms is to fit a model on the extracted and selected features such that the model generates accurate predictions about the binary classes in the training data. These classifiers are discussed below.

Naïve Bayes: Multinomial Naïve Bayes (NB) is a widely used generative classifier in which the conditional probability is used to determine whether a document belongs to a class [12]. The most well-known use of NB in NLP is in sentiment analysis [12].

NB assumes independence for all features and can work well depending on the validity of this assumption. In NLP, NB feature independence would require a word in a document to occur independently of every other word. Since word independence is a false assumption regarding text, we included n-gram word sequences to partially capture word dependence.

Linear Support Vector Machines: Support vector machines (SVM) is a classifier that divides data into two classes using a hyperplane that maximizes the separation of data in each class [13, 14]. SVM has been used in text classification successfully and tests have shown it to be better than naïve Bayes in document classification [15]. SVM is well-suited for high-dimensional data and text feature extraction commonly results in hundreds of thousands of features.

Gradient Boosted Trees: Gradient Boosted Trees (GBT) is an ensemble model that uses several weak learners together to minimize the loss of the model [16]. The composition of the results from the weak learners is performed by gradient descent. New trees are iteratively added to the ensemble to reduce a loss function. Generating a GBT model is computationally expensive because each tree is a sub-classifier that is individually developed and the ensemble classifier is refitted at each iteration.

GBT have been used in sentiment analysis in situations in which there are insufficient data to train other classifiers successfully. For example, it has been used in sentiment analysis with the Greek language where data is not plentiful [17]. Since we are working with a relatively small amount of training data, this method could be optimal for our situation. The largest limitation of this method is the computational cost created by the high-dimensional feature space of NLP problems.

Cross Validation & Hyperparameter Optimization: A critical component of the model development process included the application of a rigorous and systematic method to find the optimal models and their respective parameters. Grid search is an exhaustive hyperparameter optimization technique where model pipelines are fitted using all possible combination of supplied parameters. To find the optimal parameters for the models described in this paper we applied a grid search over relevant parameters in the model pipeline. For example, we varied feature extraction and selection parameters, such as the chi-square proportional cutoff. Each of the models also had specific parameters relevant to those models. For example, the prior probability parameter was varied for the NB model and the slack parameter was varied for the SVM model.

Evaluation and testing is performed using k-fold cross-validation. In k-fold cross validation the data is divided into k groups. Five groups were used in this exercise. Four of the five groups are used to fit the model pipelines and the remaining group is used to evaluate the trained classifier. This process is repeated until each group is used for evaluation of the classifier. The pipeline with the highest average validation F1 score is determined by the grid search.

Model Assessment

The intersections of labeled and classifier predicted classes are depicted in Figure 3. In this case, true positives (TP) are occurrences where the classifier correctly indicated an airbag non-deployment event. True negatives (TN) are occurrences where the classifier correctly did not indicate an airbag non-deployment event. False positives (FP) are occurrences where the classifier incorrectly indicated an airbag non-deployment event. False negatives (FN) are occurrences where the classifier incorrectly did not indicate an airbag non-deployment event. False positives and false negatives are analogous to Type I and Type II errors in statistical hypothesis testing respectively.

		Classifier Predicted Classes	
		1	**0**
Labeled Classes	**1**	**True Positive (TP)**	**False Negative (FN)**
	0	**False Positive (FP)**	**True Negative (TN)**

Figure 3. Confusion matrix used to measure binary classification processes.

Precision, recall, and F1 are commonly used metrics to assess binary classification methods [15]. These metrics build upon test results described within the confusion matrix (Figure 3). Precision is a measure of model performance and is expressed in Equation 4 where precision is calculated by dividing true positive occurrences (TP) by the sum of TP and false positives (FP) occurrences.

$$precision = \frac{TP}{TP+FP} \text{ (Equation 4)}$$

Recall is a measure of completeness and is expressed in Equation 5 where recall is calculated by dividing true positive occurrences (TP) by the sum of TP and false negative (FN) occurrences.

$$recall = \frac{TP}{TP+FN} \text{ (Equation 5)}$$

F1 is the harmonic mean of recall and precision and was used to assess overall model performance in this exercise. Equation 6 states that F1 is two times the product of precision and recall divided by the sum of precision and recall.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision+recall} \text{ (Equation 6)}$$

RESULTS

Three machine learning classifiers, Multinomial Naïve Bayes (NB), Support Vector Machines (SVM), and Gradient Boosted Trees (GBT) were compared to text annotators to understand which method performed better in identifying potential airbag non-deployment events in text. The primary method of comparison was by F1 score.

Across all data, SVM and GBT showed a similar performance with identical F1 scores of 91.3% (Table 1, graphed in Figure 4). Consistent with identical F1 scores, both SVM and GBT had very similar recall (92.5% and 92.0%) and precision (90.2% and 90.7%) as shown (Table 1). The NB classifier had the poorest performance of the three tested machine learning models (F1 NB 87.8% compared to 93.1% for SVM and GBT, Table 1). Despite its poor performance, the NB classifier performed far better than the annotator across all the data (F1 62.4%, Table 1).

When analyzing by data source (Figure 5), the annotator shows a competitive performance for TREAD (F1 71.2%), which out-performs Naïve-Bayes (F1 64.2%). For VOQ, however, the annotator showed inferior performance (F1 59.1%, Table 2) versus NB (F1 92.3%, Table 2). The profile of TREAD, which is a collection of disparate data sources, is likely to be the reason for the reduced F1 classifier scores relative to the more consistent data exhibited by VOQ. In all cases, however, neither the annotator nor the NB model outperformed the SVM or GBT models (Table 2).

Table 1.
Comparative results of machine learning classifiers across all input data using F1, Recall and Precision.

	F1	Recall	Precision
Annotator	**62.4%**	**54.3%**	**73.5%**
NB	**87.8%**	**89.1%**	**86.6%**
SVM	**91.3%**	**92.5%**	**90.2%**
GBT	**91.3%**	**92.0%**	**90.7%**

Table 2.
Comparative results of machine learning classifiers by data source using F1.

F1	VOQ	TREAD
Annotator	**59.1%**	**71.2%**
NB	**93.4%**	**64.2%**
SVM	**95.0%**	**76.3%**
GBT	**94.6%**	**78.4%**

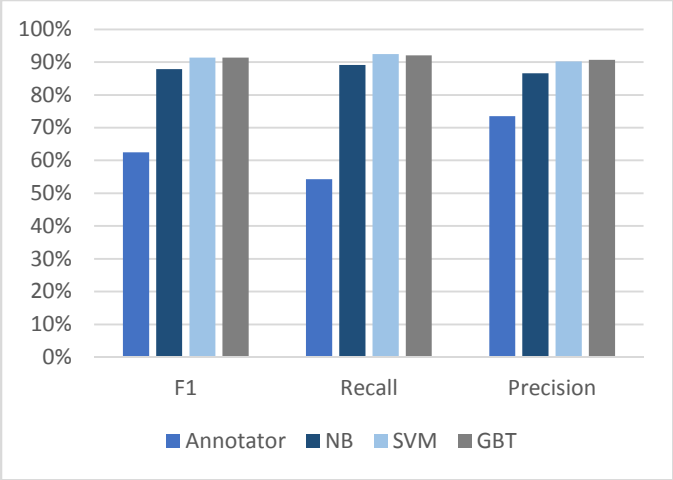

Figure 4. Graph of comparative results of machine learning classifiers across all input data using F1, Recall and Precision. Machine learning is represented by Naïve-Bayes (NB), Support Vector Machine (SVM) and Gradient Boosted Trees (GBT).

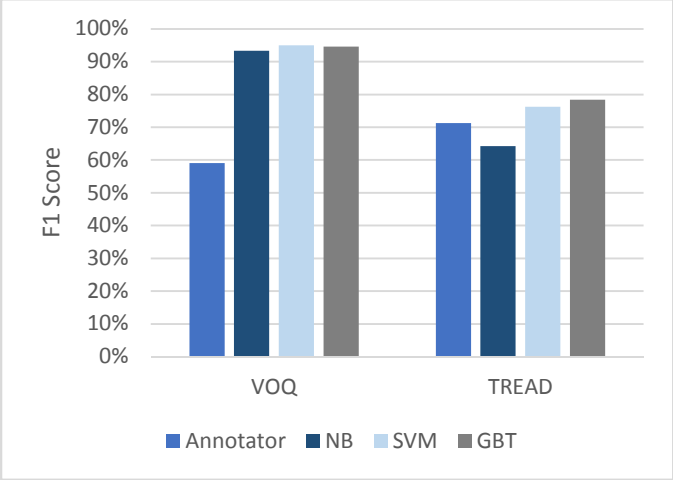

Figure 5. Graph of comparative results of machine learning classifiers by data source using F1. Machine learning is represented by Naïve-Bayes (NB), Support Vector Machine (SVM) and Gradient Boosted Trees (GBT).

CONCLUSIONS

The results shown in this paper illustrate the potential power for machine learning in transforming unstructured "dark data" into meaningful safety event detection. Machine learning methods demonstrated greatly improved classification performance (F1 score, precision, recall) in NHTSA VOQ and TREAD data than IBM Watson Explorer annotators for classification of airbag non-deployment narratives. This was true even in our scope, where training data was scarce and from a variety of data sources. Machine learning models also exhibited better balanced classification solutions compared to annotators which would tend towards having high recall at the cost of precision or vice versa.

The machine learning models yielded the worst classification performance on TREAD data. Indeed, the F1 scores for the three machine learning models tested were 16.2% to 29.2% worse for TREAD data relative to VOQ data. TREAD is the compilation of many disparate data sources at GM. We believe the reduced classification performance of TREAD is consistent with the heterogeneous nature of the data. It is likely that more training data

will be required to increase TREAD classification performance since more TREAD data will include more of each of its constituent data sources.

Despite the improvement in detecting potential airbag non-deployment events, these methods have a number of limitations related to machine learning approaches in general. First, these supervised machine learning approaches require human-labeled data in large quantities to use for training data. Second, machine learning models can only account for features (words) that have been observed in training data. If GM, the U.S. federal government and/or the U.S. public at large were to develop a new term for an airbag, then that term would be unknown to the model described in this paper unless new training data with the new term in it were used to re-fit a new version of the model. Last, our methods ignore the structure of documents and additional information, such as parts of speech (POS), for words. Such grammatical information may improve the robustness and increase the performance of our predictive models.

The application of machine learning methods for detection of potential vehicle concerns presents a robust, reliable and accurate solution. The transformation of unstructured text data into structured data enables subsequent time series analysis of potential airbag non-deployment signals, including comparative trend analysis, anomaly detection, and control charting. Future work will also focus on methods to improve model performance and reduce potential training data bias. Extracting additional features from the text, such as word POS tags, Named Entity Recognition (NER) tagging or tagging text with an ontology, may provide significant performance gains. Additionally, word embeddings could be used as an alternative feature encoding scheme which would capture the semantic meaning of the words being modeled [18]. Utilizing the concept of "data programming" to create large training sets quickly may also enable the transition from shallow learning methods (NB, SVM and GBT) to deep learning methods, such as recurrent neural networks (RNN) utilizing long short-term memory (LSTM) [19].

The data sources used in this paper represent one public and one internal GM data source. Given the robustness of machine learning text classification methods, we intend to expand the application of these models to other publicly available and GM internal data sources. Social media data, such as Twitter, Facebook and automotive forums, contain similarly unstructured data that may describe airbag non-deployment events that are valuable to detect.

We have applied our NLP and machine learning methods to other areas of potential vehicle concern and have been able to increase safety event detection F1 scores by 8 – 24% (Figure 6). In addition, these increases in F1 score for safety event detection have occurred rapidly. While annotator development in IBM Watson Explorer required detailed development of a deterministic ruleset by a human over months, a machine learning algorithm arrives at an optimal solution in minutes. Given that training data is required for both approaches, the transition from annotators to machine learning methods was a natural one.

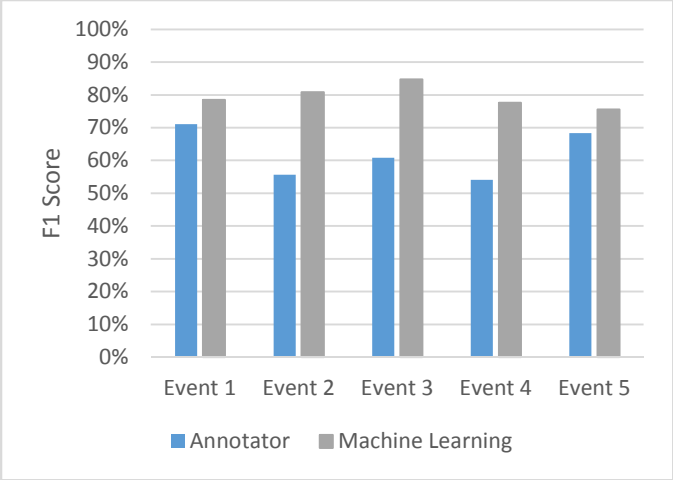

Figure 6. Classification performance improvements for other vehicle safety events. For each event, a specific machine learning model was developed. Each machine learning model is compared to an existing annotator based on F1 score.

REFERENCES

[1] US Department of Transportation, National Highway Traffic Safety Administration. 2015. "NHTSA's Path Forward." June. https://www.nhtsa.gov/staticfiles/communications/pdf/nhtsa-path-forward.pdf.

[2] Zhang, Ce, Jaeho Shin, Christopher Re, Michael Cafarella, and et al. 2016. "Extracting Databases from Dark Data with DeepDive." *Proceedings of the ACM-Sigmod International Conference on Management of Data.* 847-859.

[3] US Food & Drug Administration. 2017. *Questions and Answers on FDA's Adverse Event Reporting System (FAERS).* https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugeffects/.

[4] US Federal Aviation Administration. 2017. *Aviation Safety Reporting System.* https://asrs.arc.nasa.gov/.

[5] US Department of Transportation, National Highway Traffic Safety Administration. 2017. *SaferCar.gov.* https://www.safercar.gov/.

[6] US Department of Transportation, National Highway Traffic Safety Administration. 2017. *TREAD Act.* https://one.nhtsa.gov/cars/rules/rulings/index_treadact.html.

[7] Johnson, Chris. 2002. "Software tools to support incident reporting in safety-critical systems." *Safety Science* 40: 765-780.

[8] Tanguy, Lodovic, Nikola Tulechki, Assaf Urieli, Eric Hermann, and et al. 2016. "Natural Language Processing for aviation safety reports: from classification to interactive analysis." *Computers in Industry* (Elsevier) 78: 80-95.

[9] Gabler, Hampton C., and John Hinch. 2008. "Evaluation of Advanced Air Bag Deployment Algorithm Performance using Event Data Recorders." *Annals of Advances in Automotive Medicine / Annual Scientific Conference* 52: 175-184.

[10] Zhu, Wei-Don, Bob Foyle, Daniel Gagne, Vijay Gupta, and et al. 2014. *IBM Watson Content Analytics: Discovering Actionable Insight from Your Content.* 3rd ed. Poughkeepsie, NY: IBM. https://www.redbooks.ibm.com/redbooks/pdfs/sg247877.pdf.

[11] Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning.* New York, NY: Springer.

[12] Jurafsky, Daniel, and H. James Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd ed. Upper Saddle River, NJ: Prentice-Hall.

[13] Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. "LIBLINEAR: A Library for Large Linear Classification." *Journal of Machine Learning Research* 9: 1871-1874.

[14] Ho, Chia-Hua, and Chih-Jen Lin. 2012. "Large-scale Linear Support Vector Regression." *Journal of Machine Learning Research* 13: 3323-3348.

[15] Joachims, Thorsten. 2002. Learning to Classify Text Using Support Vector Machines. New York, NY: Springer.

[16] Friedman, H. Jerome. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189-1232.

[17] Athanasiou, Vasileios, and Manolis Maragoudakis. 2017. "A Novel, Gradient Boosting Framework for Sentiment Analysis in Languages where NLP Resources Are Not Plentiful: A Case Study for Modern Greek." *Algorithms* 10 (1): 34.

[18] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and et al. 2013. "Distributed Representations of Words and Phrases and their Compositionality." *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems.* 3111-3119.

[19] Ratner, Alexander, Christopher De Sa, Sen Wu, Daniel Selsam, and et al. 2017. "Data Programming: Creating Large Training Sets, Quickly." *arXiv:1605.07723 [stat.ML].*