INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1407 - 1411
RECEIVED : 16/11/2016. PUBLISHED: 28/11/2016
November 28, 2016

A CONTEXT AWARE CLOUD BASED TEXT MINING RECOMMENDATION FRAMEWORK VEHICLE ACCIDENTS USING BIME

[1]L.Mahalakshmi, [2]S. Karthik
[1]PG Scholar, Dept Of CSE, Ganadipathy Tulsi's Jain Engineering College, Vellore
[2]Asst Prof,Dept Of CSE,Ganadipathy Tulsi's Jain Engineering College, Vellore

Abstract:

The accidents on the day to day incidents play a major role in terms of safety concern for the transportation industry. The statistics of road, rail, and air accidents incur a huge loss to the country. Though most of the minor accidents costs very little, it impacts the life of several humans during the time. In order to better understand the accident reports involved, this application is developed on the backend process of text mining where the Apriori is applied using the BIME tool. The Extensive analysis is collecting the data sets of the incidents happened on various factors. It predicts the accuracy on contribution of those accidents and the steps for avoidance. This application 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 under- standing of the contributors to accidents in ways not possible through only fixed field analysis of the accident reports. The BIME tool is used on top of Big Query Table to discover the data analytics of the accidents and its loses. This application overcomes the existing methodologies of manual prediction about the causes and losses incurred.

Keywords : BIME tool, Big Query, Extensive analysis.

1. INTRODUCTION

There is a growing concern in the maritime industry regarding human and organizational factors that affect sailing performance and the overall safety of ship operations in and onboard [6]. This concern stems from a recent rise in commercial maritime accidents caused by ill-fated decisions taken by higher level management. This is further highlighted by academic research showing direct ties between organizational factors and safe performance of maritime crew of the ship. However, effective tools or methodologies for identifying and mitigating potentially harmful human and organizational factors before they cause an accident are yet to be developed. The purpose of the present research is to extract the causal patterns from accident investigation reports. These patterns study human and organizational factors affecting safety culture and discuss models of safety culture used to design assessment techniques. A careful investigation of these patterns provides an opportunity to improve and manage safety in the future [53]. This study aspires to model causal parameters relating accidents During the last century, sea trade has been increased due to technological advancements. Hence, increasing number of ships are sailing on the world seas. Modern ships are getting faster, bigger and highly automated. Though these technological advancements are beneficial, they still pose a challenge in themselves. Accidents at sea still occur and the consequences to people, ship or environment, are often greater than before [26]. These accidents are investigated by a maritime accident investigation

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1407 - 1411
RECEIVED : 16/11/2016. PUBLISHED: 28/11/2016
November 28, 2016

board. The board reports how the accident occurred, the circumstances, causes, consequences and rescue operations. These reports also provide recommendations for preventing similar accidents. The reports are long, detailed and systematic examinations of marin. According to , causal patterns from the accident investigation reports provide information on various mechanisms behind accidents. Unfortunately, in the maritime field, no standard reporting formats exist and data collection from the textual reports is a laborious task . Text mining provides a means for efficient and informative scanning of accident cases of interest without reading the actual report. Therefore, text mining in this context is seen as a useful tool in understanding accidents and their influencing factors.

2. RELATED WORK

Pattern recognition is a subfield in machine learning with a purpose of developing methods that recognize meaningful patterns from the data. Pattern recognition has seen applications in the fields of 1) computational fluid dynamics . In other words, pattern classification observes the environment to learn and distinguish patterns of interests and make reasonable decisions about the pattern (or finding the correct class represented by the pattern).

Fig.1. Basic representation of pattern classifier

The decision of the pattern classifiers depend on the prior available patterns. The more relevant patterns are available for the pattern classifier, the better the decision will be. Pattern classification methods are of two types, supervised methods and unsupervised methods. The major difference between supervised and unsupervised methods is the process of learning, during which the characteristics of the data are learned by the classifier. In supervised classification methods, the pattern x = (x1, x2,..., xn) along with its associated label or class yi , yi∈ {1,2,. . . ,P} &i∈ {1,2,..., k}, form the training dataset S,{(xi , yi),i = {1,2,..., k}}. During the training phase, the classifier learns from the existing patterns with their corresponding labels. The trained classifier can then be used to predict the labels for the new unseen data or test data. On the contrary, unsupervised methods do not use labels yi along with the patterns xi during training. The unsupervised methods estimate the hidden patterns in the data to group the given data into several groups or clusters. Hence, unsupervised methods are also referred to as Clustering Methods. This study used two supervised methods, Support Vector Machines (SVM) and na¨ıve Bayes classifiers to classify causal and non-causal patterns.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1407 - 1411
RECEIVED : 16/11/2016. PUBLISHED: 28/11/2016
November 28, 2016

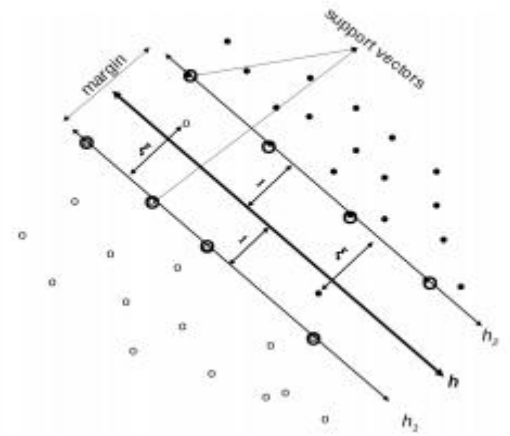

Fig.2. The optimal separating hyperplane

3. DATA REPRESENTATION

The data used in the study is 'MAIB accident investigation reports'. Marine Accident Investigation Branch (MAIB 2) is a branch of the Department for Transport located in Southampton, UK. MAIB has four teams of experienced accident investigators, each comprising a principal inspector and three inspectors drawn from the nautical, engineering, naval architecture or fishing disciplines. The role of the MAIB is to contribute to safety at sea by determining the causes and circumstances of marine accidents and working with others to reduce the likelihood of such accidents recurring in the future. There are 11 categories of accident investigation reports which are Machinery, Fire/Explosion, Injury/Fatality, Grounding, Collision/Contact, Flooding/Foundering, Listing/Capsize, Cargo Handling Failure, Weather Damage, Hull Defects and Hazardous Incidents. But this study concentrates only on 4 types of accident types with a total of 135 investigation reports as shown in the Table VI. Each report, on an average contains 60 pages which are divided into 3 sections viz: 1) narrative 2) analysis and 3) conclusions. Narrative section describes the summary of the accident, while every possible detail regarding the accident is analyzed in the analysis section. A maritime accident investigation report is written in a natural language, by different investigating officers and hence does not follow a standard reporting format. This makes the investigation reports inconsistent and noisy. If data is inconsistent, the text mining algorithms under-perform. The text data also contains some special formats like number formats, date formats and the most common words that are unlikely to help text mining such as prepositions, articles, and pronouns that are to be eliminated. In order to extract data which is consistent and accurate, data preprocessing methods are crucial. This section of the study reviews some simple NLP processing tasks that are used in the experiments, such as, tokenization and stemming using Natural Language Toolkit (NLTK). The NLTK, is a suite of Python libraries and programs for symbolic and statistical natural language processing . NLTK includes graphical demonstrations and sample data. It is accompanied by extensive documentation. This may sound trivial as the text is already stored in machine-readable formats. Nevertheless, some problems are still left, like the removal of punctuation marks. Other characters like brackets, hyphens, etc. require

processing as well. Furthermore, the text should be lower cased to cater consistency in the documents. The main use of tokenization is identifying the meaningful significant words.

4. RESULT ANALYSIS

The dataset is the collection of causal relations marked by three domain experts. The experts marked a total of 151 causal sentences in four accident investigation reports. These 151 causal sentences with an addition of 151 non-causal sentences from the same accident investigation reports are combined to form a complete dataset containing 302 sentences. Out of them, 70% i.e 212 sentences(106 causal and 106 non-causal) are considered as the training set and remaining 30% i.e 90 sentences (45 causal and 45 non-causal) are considered to be test set. Inconsistency can arise from different number formats or time formats. Another

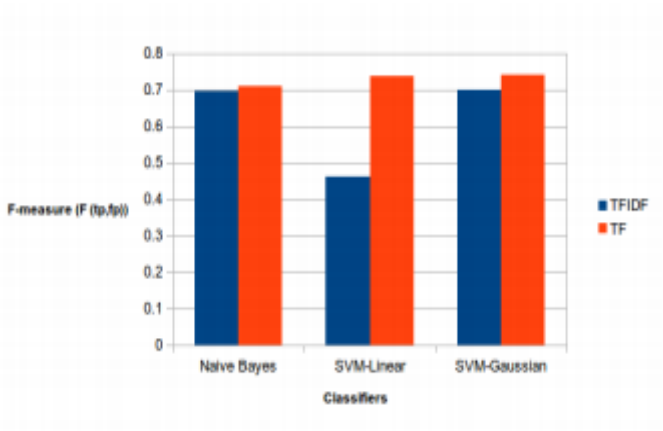

Fig.3. Comparison of F-measure (Ft p, f p) on validation sets for TF & TFIDF weights.

problem is abbreviations and acronyms which have to be transformed into a standard form. If we compute the frequencies of the words in a corpus, and arrange them in decreasing order of frequency, then the product of the frequency of a word and its rank (its position in the list) is more or less equal to the product of the frequency and rank of another word. So frequency of a word is inversely proportional to its rank. That is, the frequency of words multiplied by their ranks in a large corpus is approximately constant.

CONCLUSION

To conclude, it is possible to say that experts' marked causal relations from four different accident investigation reports were fairly sufficient to classify and extract causal patterns from other accident investigation reports. The results also suggest that usage of connecting words were influential on classification results. It was evident from this analysis, that pattern classification method outweighs the connectives method. It is still unclear which of the approaches are most suitable for exacting causal relations from maritime accident reports. When there are many similar methods available it is difficult to choose which one to use. In such a case simplicity and reputation of the method and experience of its usage can influence the decision. This research might embark on developing effective tools and methodologies in future for identifying human and organizational factors present in the accident investigation reports.

REFERENCES

[1] K. Artana, D. Putranta, I. Nurkhalis, and Y. Kuntjoro. Development of simulation and data mining concept for marine hazard and risk management. In Proceedings of the 7th International Symposium on Marine Engineering (24-28 October 2005), 2005.

[2] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 16(5):412–424, 2000.

[3] R. Basili, A. Moschitti, and M. T. Pazienza. A hybrid approach to optimize feature selection process in text classification. In AI* IA 2001: Advances in Artificial Intelligence, pages 320–326. Springer, 2001.

[4] C. Bishop. Pattern recognition and machine learning, volume 4. Springer Verlag, 2006.

[5] D. C. Blair. Language and representation in information retrieval. Elsevier North-Holland, Inc., 1990.

[6] J. F. Bradford. The growing prospects for maritime security cooperation in southeast asia. Technical report, DTIC Document, 2005.

[7] M. A. I. Branch, F. Floor, C. House, and C. Place. Bridge watchkeeping safety study. Department for Transportation, Marine Accident Investigation Branch, Southampton, 2004.

[8] N. Cristianini and J. Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[9] J. A. Dowell. Transition words. https://www.msu.edu/ ˜jdowell/135/transw.html, cited 10 March 1997.

[10] L. Egghe. The exact place of zipf's and pareto's law amongst the classical informetric laws. Scientometrics, 20(1):93–106, 1991.