Text Mining based Safety Risk Assessment and Prediction of Occupational Accidents in a Steel Plant

Conference Paper · March 2016

DOI: 10.1109/ICCTICT.2016.7514621

Sobhan Sarkar
Indian Institute of Technology Kharagpur

Sammangi Vinay
Indian Institute of Technology Kharagpur

2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)

Text Mining based Safety Risk Assessment and Prediction of Occupational Accidents in a Steel Plant

Sobhan Sarkar
Department of Industrial & Systems Engineering,
Indian Institute of Technology,
Kharagpur, India.
sobhan.sarkar@gmail.com

Sammangi Vinay
Department of Mechanical Engineering
Indian Institute of Technology,
Kharagpur, India.
sammangi.vinay@gmail.com

Jhareswar Maiti
Department of Industrial & Systems Engineering,
Indian Institute of Technology,
Kharagpur, India.
jhareswar.maiti@gmail.com

Abstract - **Occupational accidents are a serious threat to any organization. Occupational accidents in steel industry sector remain a threat as workforce is exposed to different kinds of hazards due to the workplace characteristics. In this study, a unique method is proposed by developing a text mining based prediction model using fault tree analysis (FTA), and Bayesian Network (BN). Free unstructured accident dataset for a period of four years has been used in this study. Text mining approach results in finding the basic events concerning each of primary causes. The basic events, in turn, are utilized in building FT and BN diagram that could predict the occurrence of accidents attributable to different primary causes. The model, so developed, can be considered adequate with 87.5% accuracy. Furthermore, sensitivity analysis is performed for the validation of the model.**

Keywords – Occupational safety; Text Mining; Fault Tree Analysis; Bayesian Network; Sensitivity analysis.

I. INTRODUCTION

Steel industry is one of the most hazardous industries due to its intricate socio-technical structure i.e., it demands massive human labor, high technology which, in turn, makes the safety management system (SMS) of any industry a tough and challenging task. In India, the accident statistics provides evidence of 1433, 1383, 1417 occupational injuries due to fatal accident occurred in 2011, 2012, and 2013, respectively throughout states [1]. Of this, a total of 29, 52, and 31 are the numbers of fatal accidents that took place in those years respectively in steel plant only [1]. Thus, this statistics has encouraged many researchers to probe into the accident events, and to provide some industrial reliable safety solution either by proactive measure through prediction of the occurrence of accidents, or by reactive measure through cause-effect analysis study. In past, safety performance is largely measured by lagging indicators like number of injuries, fatalities, etc. But, with today's advancement, industries are prone to adopt leading indicators like safety observation data etc. to take better decision in SMS beforehand the occurrence of accident. Therefore, early realization of events of accidents mostly has drawn the researchers to investigate properly so as to minimize the risk of accident in occupational domain. In line with this research, our study proposes a proactive safety measure through prediction of accidents in an integrated steel plant.

In this paper, a new safety risk assessment model has been proposed through text mining (TM) concept which actually utilizes the accident database of an integrated steel plant, and tries to find some potential basic events (BE) behind each of the accident primary causes. Consequently, the basic events, thus figured out from TM, are rechecked manually in order to check whether they really act as BE or not for any particular top event/ primary cause like slip-trip-fall (STF). Then, fault tree (FT) diagram, formed for any primary cause, is then transformed into Bayesian Network (BN) in order to investigate the dependencies between parent and child nodes as well as to predict the future probability of occurrence of any primary cause in workplace. Here, BN is evaluated in AgenaRisk software in order to obtain the safety risk potential score (SRPS).

The FT transformed BN-based safety risk assessment model was validated against eight different departments (represented as A1, A2, …, A8) where specific primary events occurred more frequently. This study shows that the ranks of the SRPS are almost steady with the primary event at each department. Finally, sensitivity analysis is done with the help of AgenaRisk software that generates tornado diagrams and sensitivity tables (not mentioned in this paper). From the tornado diagram, we can provide the necessary safety measures required in order to prevent the injuries in particular department. Our main aim is to predict the future occurrence of primary causes in particular departments such that advanced safety proactive measures could be initiated from management to reduce the number of occupational accidents.

The rest of the paper is organized as follows: Section II describes briefly the problem statement of the industry.

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

Section III illustrates some important related works in past. In Section IV, the proposed methodology has been briefly discussed. Results and discussion are given in Section V. Finally, conclusion with future scope has been highlighted in Section VI.

II. PROBLEM STATEMENT

This study primarily focuses on providing a solution to a problem related to an occupational accident in an integrated steel plant in India. The company has been trying to figure out basic causes occurring accidents, and simultaneously to implement some safety measures for their workers. In this industry, after each incident occurrence, data logging describing the short description of event in free text format is maintained that, in turn, results in a huge database which is very hard for human analysis. Therefore, to alleviate the problem occurred in industry, proper evaluation of such huge database including free text needs further development.

A. Dataset:

The dataset consists of details of brief description of events for the last four years (April, 2010 – Dec, 2013). There are 23 primary causes initially pointed out from their database. They are crane dashing (CD), derailment (D), dashing/collision (DC), electric flash (EF), energy isolation (EI), equipment machinery damage (EMD), fire/explosion (FE), gas leakage (GL), hot metals (HM), hydraulic/pneumatic (HP), lifting tools tackles (LTT), material handling (MH), medical ailments (MA), occupational illness (OI), process incidents (PI), rail (R), road incident (RI), run over (RO), skidding (S), slip/trip/fall (STF), structural integrity (SI), toxic chemicals (TC), working at height (WH). As our study is aimed primarily at building a prediction model, total data set of 45 months has been partitioned in the ratio of nearly 70:30 as training and test set, respectively. As a result, out of 998 dataset, 720 observations were used for building the prediction model, and rest is used for validation of the model.

In our study, frequencies of incidents like near miss, property damage and injury from training dataset (33 months) is shown in Fig. 1. It is evident that the frequency of injury is increasing by every year. Fig. 2 shows an in-depth stacked plot of frequency of each primary cause with each year. STF (30.6%) occurs more frequently when compared to other primary events causing the injury. Besides STF, RI (26.29%), PI (8.19%) amount to high percentage of injury cases. Therefore, prevention of these primary causes from occurrence is one of the key challenging tasks in this steel plant.

III. RELATED WORKS

In this section, there is short discussion presented highlighting some of the key papers on accident analysis field. Here, this discussion is two-fold. First, it describes some traditional methods to analyze accidental occurrence, and then some advanced techniques are outlined in line to accidental analysis.

In literature, there are many risk assessment techniques described by the researchers. Some of them, which are relevant and important in this context, are described in this section. Some systematic risk assessment models like fault tree analysis (FTA), Petri nets, decision tree (DT), failure mode and effect criticality analysis (FMECA) are used by many studies [2, 3, 4, 5, 6]. But, there are some disadvantages in implementation of these traditional approaches whenever researchers are more interested in addressing dependencies among various levels of factors that could result in any primary cause of accident. In order to mitigate those limitations, many models like structural equation model (SEM), Bayesian Network (BN) etc. are developed to define the interrelationship between factors [7, 8]. BN has been implemented to outline the causes behind the falls from the height [9]. Moreover, some researchers used BN for prediction of accidental incidents [10].

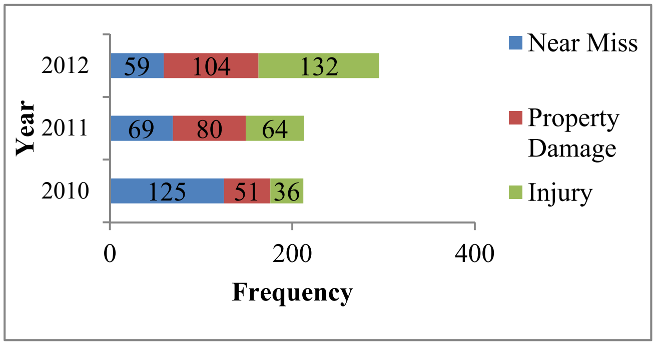

Fig. 1. Incident Categories in 2010-2012.

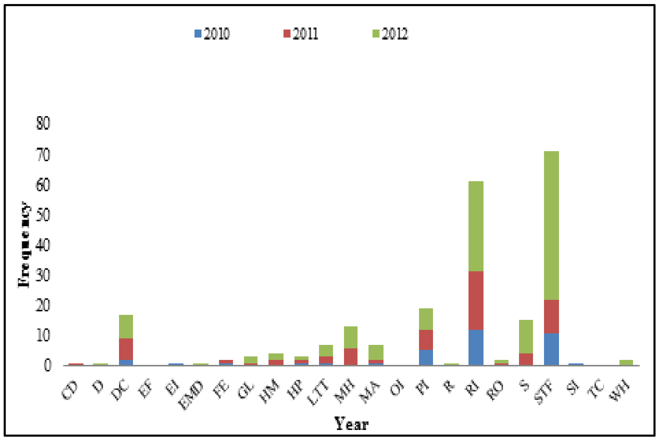

Fig. 2. Occurrence of primary causes in the years 2010-12.

Some modern techniques like data mining have been addressed in accident analysis domain. Chang et al. used Classification and Regression Tree (CART) model and negative binomial regression model for analyzing the traffic accident behavior [11]. It is accompanied by many studies that addressed Bayesian theorem [12], data fusion, ensemble, clustering [13, 14], decision tree [15] and so on.

Some important works related to the application of text mining in the field of accident analysis are road crash analysis in Australia [16], aviation accident analysis [17, 18, 19], and so on. Some of the researchers use Leximancer concept mapping [16], text retrieval method [20], and link analysis [17].

After doing an extensive literature study (though not mentioned in this paper), there is hardly any work done on text mining based risk assessment in steel industrial domain for occupational accident analysis. Thus, this study proposes a potential unique method for accident analysis by incorporating FTA and BN through text mining concept from industrial accidental data base.

IV. METHODS

The methodology is briefly explained in this section below.

A. *Text Mining method*

Text mining is a process of retrieving underlying themes or concepts contained in a large collection of documents. In our study, text mining is used to explore the useful information from a huge accidental unstructured database of an industry [16-17]. To retrieve the information, frequency of each useful word is computed after performing two stage operations i.e., (i) term creation, and (ii) term filtering. In term creation stage, all the terms in character string in a document are tokenized, and in the second stage, all the pre-processing tasks like white space, punctuation, number removal, lower case conversion, stemming, lemmatization, stop words and common words removal are performed. Then, document term matrix (DTM) is created. For each of the primary or top events, this approach is used in order to investigate the existence of basic events and their probability of occurrence in dataset. In order to perform this task, R statistical package, and SAS software have been used.

B. *Creating basic event and finding their probability from Text Mining*

The Document Term Matrix (DTM), thus created, gives us only the important words with corresponding frequencies for the analysis. By manual interpretation, we check whether an appropriate basic event from a combination of words could be generated properly or not. For example, if we filter the ‘slip’ and ‘stair case’ words columns occurring simultaneously, we get a frequency of 13, from which we can infer the basic event as “slipping on stair case”. In this way, we create different possible basic events manually, and find their frequencies. Now, we consider a specific department and a primary event. For instance, there are four observations where hot metal logistics (HML) department and STF occur simultaneously. We derive the basic events of them manually from which we get the basic event probability of those observations.

C. *Fault tree analysis (FTA)*

FTA is a top-down method designed to analyze the effects of initiating faults and events on any complex system [21]. The construction of FTA involves five main steps. The first step is to define the undesired event to study as top event. Once the top event is fixed, all causes affecting the top event are studied and analyzed. After that, FT is constructed based on AND and OR logical gates. Now, the FT is evaluated and analyzed for finding any possible chance of improvement and identify all hazards that are possible in affecting the system. Finally, all possible proactive measures could be undertaken to decrease the occurrence probability. Since, FTA has limited ability to establish complex causal relationships among factors; Bayesian Network (BN) is one of the alternate options for finding out the solution.

D. *Bayesian Network (BN)*

BN, also called belief network, is a statistical model that represents a set of random variables and their conditional dependencies using a directed acyclic graph (DAG). BN consist of nodes, joint nodes, and conditional probability tables (CPTs). When it comes to uncertain inferences, especially while linking various forms of information such as output models, empirical data, expert opinions, BN has a higher efficiency compared to other models [21].

There are two Bayesian network approaches which are mostly used. The first one is to learn from a large amount of training data. And, second is based on experts’ opinion. The transformation of logic gates from FT to BN is often one-to-one i.e., a logic gate in FT is converted into physical node in BN [21]. Nonetheless, the meanings of event node and logic gates differ. An event node represents a variable in the given problem domain, while a logic gate describes logical relationship between the nodes. In the transformation of logic gates, probability values should be mentioned in the CPT in BN that corresponds to logic gates in FT.

The detailed flowchart of the proposed methodology is shown in Fig. 3.

V. RESULTS AND DISCUSSION

In this section, some key findings from our study are discussed. Basic events, obtained from text mining, and thus verified manually, are listed in TABLE I for STF case in A2 department (as an example). Then, the risk prediction score of STF in A2 department is computed and discussed as follows:

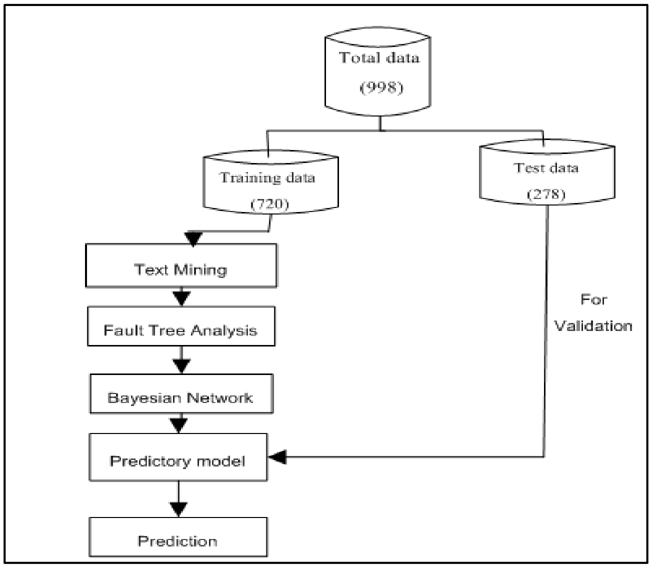

Fig. 3. Flowchart of the proposed methodology.

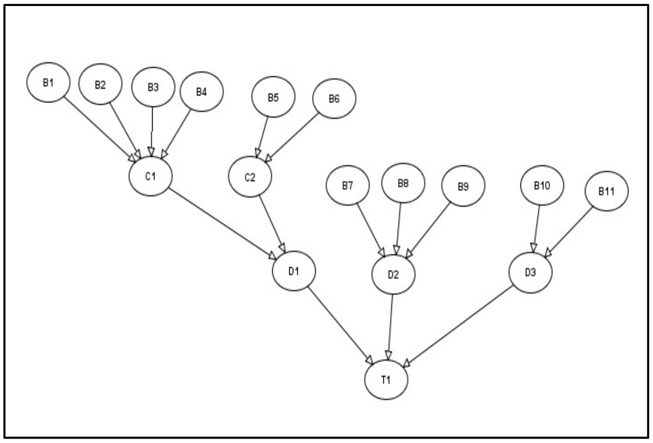

Fig.4. BN of STF at department A2.

The BN-based safety risk assessment model of STF at A2 department is shown in Fig. 4. B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, B11 are the basic events causing the top event (i.e. STF) in A2 department. Basic events are also called the leaf nodes in BN because those are the nodes which do not have children nodes. T1 is called the root node because it has no parent nodes. The remaining nodes D1, D2, D3, C1, and C2 are called intermediate nodes (neither a leaf node nor a root node).

TABLE I: BASIC EVENTS CAUSING STF AT DEPARTMENT A2

Notation	**Meaning**	**Notation**	**Meaning**	**Notation**	**Meaning**
T1	STF	B1	Cleaning the discharge	B7	Foreign particle hits the person
D1	T1 caused by working	B2	Miscommunication between project people	B8	Drain covers are not covered properly
D2	T1 caused by walking	B3	Failure of valve	B9	Slippery Road
D3	T1 caused by travelling	B4	Exposed to Abnormal atmosphere	B10	Slipping on stair-case
C1	Non-Slippage phenomenon while working	B5	Instrument slipped	B11	Just while walking, slipped
C2	Slippage phenomenon while working	B6	Slipped while firing the Oven		

TABLE II: EVALUATION COMPARISON BETWEEN AGENARISK AND TEST DATASET

Department	**STF**	**RI**	**PI**	**Accident from Test data**
A1	4.95	**46.61***	1.60	RI
A2	**67.49***	32.40	58.61	STF
A3	0.00	**18.20***	12.10	RI
A4	**78.37***	0.00	15.27	STF
A5	**28.70***	15.02	10.98	STF
A6	**78.29***	37.21	13.19	STF
A7	31.40	**33.66****	0.00	MH
A8	**46.31***	4.47	9.10	STF

*Note: Bold- highest risk, *-matched, **- not matched*

The model in our study is verified against eight departments (listed in TABLE II). All the safety risk prediction scores obtained from AgenaRisk are listed in this table for top three primary causes (i.e. STF, RI, and PI) against each of the eight departments listed. The last column in this table represents the primary cause with maximum number of occurrence in a particular department, and it is computed from the test data. The safety risk potential score i.e., SRPS of RI in A1 department is 46.61 which is the highest among all other primary causes like STF and PI. So, proposed model predicts RI to be probable primary cause for A1 department in test data which in turn is verified from the statistics from test data. Similarly, for rest of the primary causes, posterior probabilities are computed against each department, and are shown in TABLE II. There is only one misclassification out of eight has been found out that implies the fact that our proposed model has the classification accuracy 87.5%.

A. Sensitivity analysis

Sensitivity analysis is performed in order to further examine the causes that affect the occurrences of three accident primary causes. While using AgenaRisk for sensitivity analysis in BN, a single target node and some sensitivity nodes must be selected. Sensitivity reports such as tornado graphs are generated by AgenaRisk. From the Tornado graph, top sensitivity nodes were preferred based on rank of sensitivity nodes. In this study, Boolean nodes with two values (either true or false) are used. Tornado graphs related to PI, RI, and STF in A5 department are shown in Figs. (5-7).

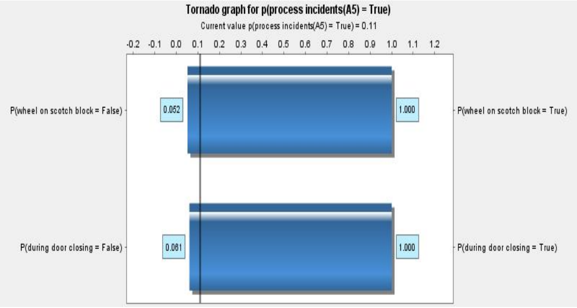

Fig. 5. Tornado diagram for PI in department A5.

In the case of PI, the most significant direct cause is wheel getting on scotch block (Fig. 5). So, proper safety training is necessary to the workers to prevent PI in A5 department. Similarly, the most significant direct cause of RI is the injury occurring during raining (Fig. 6). So, proper care should be adopted while travelling during rain to reduce the RI in A5 department. So, from tornedo diagram (Fig. 7) the most significant direct cause of STF is found out to be the person hitting by instrument. So, proper safety training should be provided to the workers in order to reduce the STF in A5 department. Likewise, tornado diagrams for primary causes in different departments can be generated and proper safety measures can be implemented in order to reduce the occurrence of that primary event. Another important key finding includes that as per SRPS (TABLE II), A4, A1, and A2 departments are more exposed to STF, RI, and PI related problem, respectively, than any other departments. In summary, this model not only estimates the safety risk potential score at various departments in steel plant, but also identifies the sensitive causes at each department.

Like other research, the study has some limitations. First, the dataset consists of only four years incident reports, whereas if it could have included more years, then more number of useful insights regarding root cause analysis would have been explored. Second, text mining is limited to generate the important words and their frequencies, but to figure out the relevant root causes, human effort must be made. That's why, much more time was devoted in analyzing the dataset from preprocessing steps to root cause findings. Finally, the FT diagram has nodes, all connected by OR gates. This study could not address any AND logical explanation from the data base. Furthermore, FTA has limited application in prediction of accidents due to interdependencies among nodes.

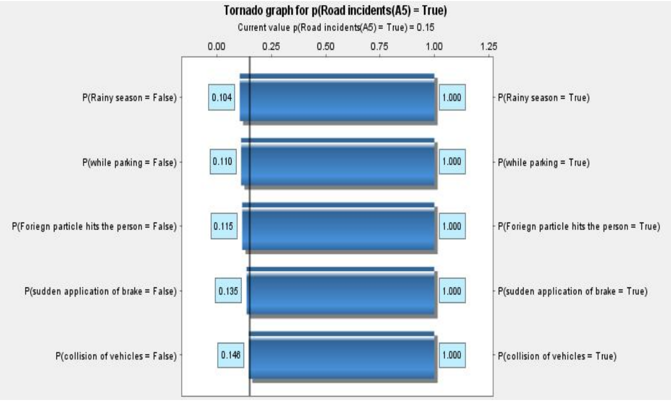

Fig. 6. Tornado diagram for RI in department A5.

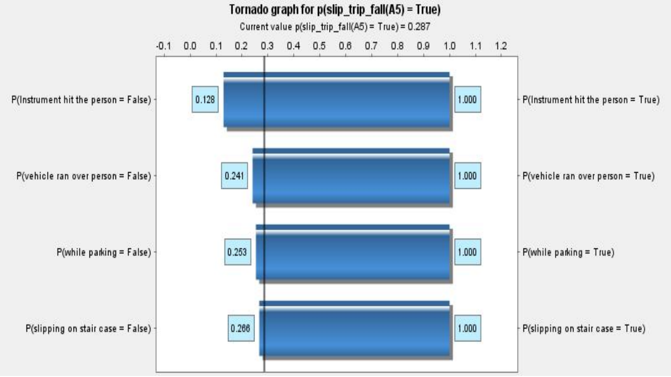

Fig. 7. Tornado diagram for STF in department A5.

VI. CONCLUSION

This study discussed an efficient method in constructing a text mining and FT transformed BN based safety risk assessment model for the steel plant. The results of BN are validated against eight departments in which specific site accidents occurred. The output of BN is highly steady with the accidents events at the departments in test dataset. This implies that the transformation process from FT to BN for all the three primary causes could create a realistic and accurate model. Therefore, based on the assessment of model, and its sensitivity analysis, corresponding department managers could provide proactive preventive safety measures, and ensure better resource utilization to reduce risks of occupational accidents in steel plant.

As a future work, survey can also be done for further verification of our proposed model by expert opinions. Some detailed in-depth analysis could be performed engaging more number of departments as well as more number of primary causes in order to identify the root causes responsible for occurrence of accidents. Furthermore, text mining could be done more accurately and could be accompanied by rule induction mining in order to find out some latent/ hidden rules to avoid incidents to happen. Some unsupervised algorithms like clustering, or association rule mining could be implemented onto this problem with an aim to figure out some inherent grouping of similar events.

ACKNOWLEDGMENT

The authors are grateful to the management of the steel industry for their support from the data collection to the final evaluation of the work.

REFERENCES

[1] Retrieved from: www.indiastat.com/table/crimeandlaw/6/industrialaccidents.

[2] Chi, Chia-Fen, Syuan-Zih Lin, and Ratna Sari Dewi, "Graphical fault tree analysis for fatal falls in the construction industry," *Accident Analysis & Prevention,* vol. 72, pp. 359-369, 2014.

[3] Hauptmanns, U., M. Marx, and T. Knetsch, "GAP—a fault-tree based methodology for analyzing occupational hazards," *Journal of loss prevention in the process industries,* vol. 18, no. 2, pp. 107-113, 2005.

[4] Mistikoglu, Gulgun, Ibrahim Halil Gerek, Ercan Erdis, PE Mumtaz Usmen, Hulya Cakan, and Emrah Esref Kazan, "Decision tree analysis of construction fall accidents involving roofers," *Expert Systems with Applications,* vol. 42, no. 4, pp. 2256-2263, 2015.

[5] Zeng, Sai X., Chun M. Tam, and Vivian WY Tam, "Integrating safety, environmental and quality risks for project management using a FMEA method," *Engineering Economics,* vol. 66, no.1, 2015.

[6] Pinto, Abel, Isabel L. Nunes, and Rita A. Ribeiro, "Occupational risk assessment in construction industry–Overview and reflection," *Safety Science,* vol. 49, no. 5, pp. 616-624, 2011.

[7] Seo, Hee-Chang, Yoon-Sun Lee, Jae-Jun Kim, and Nam-Yong Jee, "Analyzing safety behaviors of temporary construction workers using structural equation modeling," *Safety Science,* vol. 77, pp. 160-168, 2015.

[8] Baksh, Al-Amin, Faisal Khan, Veeresh Gadag, and Refaul Ferdous, "Network based approach for predictive accident modelling," *Safety science,* vol. 80, pp. 274-287, 2015.

[9] Matías, J. M., T. Rivas, J. E. Martín, and J. Taboada, "A machine learning methodology for the analysis of workplace accidents," *International Journal of Computer Mathematics,* vol. 85, no. 3-4, pp. 559-578, 2008.

[10] Deublein, Markus, Matthias Schubert, Bryan T. Adey, Jochen Köhler, and Michael H. Faber, "Prediction of road accidents: A Bayesian hierarchical approach," *Accident Analysis & Prevention,* vol. 51, pp. 274-291, 2013.

[11] Chang, Li-Yen, and Wen-Chieh Chen, "Data mining of tree-based models to analyze freeway accident frequency," *Journal of Safety Research,* vol. 36, no. 4, pp. 365-375, 2005.

[12] Alizadeh, Seyed Shamseddin, Seyed Bagher Mortazavi, and Mohammad Mehdi Sepehri, "Assessment of accident severity in the construction industry using the Bayesian theorem," *International Journal of Occupational Safety and Ergonomics,* vol. 21, no. 4, pp. 551-557, 2015.

[13] Sohn, So Young, and Sung Ho Lee, "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea," *Safety Science,* vol. 41, no. 1, pp. 1-14, 2003.

[14] Chung, Yi-Shih, "Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees," *Accident Analysis & Prevention,* vol. 61, pp. 107-118, 2013.

[15] Kumar, Sachin, and Durga Toshniwal, "A data mining framework to analyze road accident data," *Journal of Big Data,* vol. 2, no. 1, pp. 1-18, 2015.

[16] 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, , pp. 49–58, Sep 2009.

[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] Sjöblom, Olli. "Data Mining in Promoting Aviation Safety Management." In *Safe and Secure Cities*, pp. 186-193. Springer International Publishing, 2014.

[19] Feng, Xia, and Juanjuan Li. "Analyzing pilot-related accidents and incidents by data mining." In *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, vol. 14, pp. V14-325. IEEE, 2010.

[20] Gregor, D., S. Toral, T. Ariza, F. Barrero, R. Gregor, J. Rodas, and M. Arzamendia, "A methodology for structured ontology construction applied to intelligent transportation systems," *Computer Standards & Interfaces,* 2015.

[21] S. S. Leu and C. M. Chang, "Bayesian-network-based safety risk assessment for steel construction projects," *Accident Analysis & Prevention*, vol. 54, pp. 122-133, 2013.