Machine learning for text mining based on prediction of occupational accidents and safety risk calculation

Chang Li, Bing Pan, Zheng Xiang, Lixuan Zhang, Lee Chen, Don Chen
Faculty of Computer Science and Information System, Universiti Teknologi MARA (UiTM), Malaysia

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

1.0 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. 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 [1-8]. Thus, these 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 [9-14].

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) [15-21].

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 [22-28].

2.0 LITERATURE REVIEW

// 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.

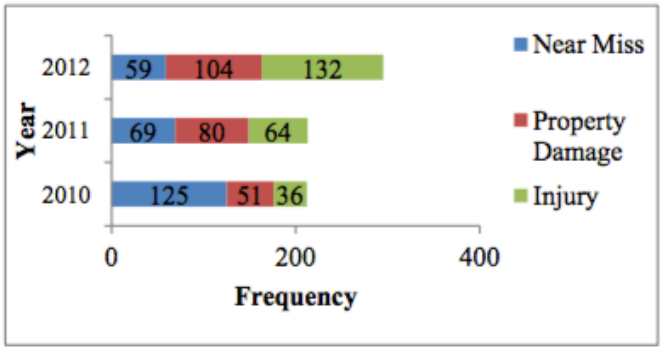

Fig. 1. Incident Categories in 2010-2012.

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 [1-11]. 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 [12-19]. BN has been implemented to outline the causes behind the falls from the height. Moreover, some researchers used BN for prediction of accidental incidents. 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. It is accompanied by many studies that addressed Bayesian theorem, data fusion, ensemble, clustering, decision tree and so on [29-34].

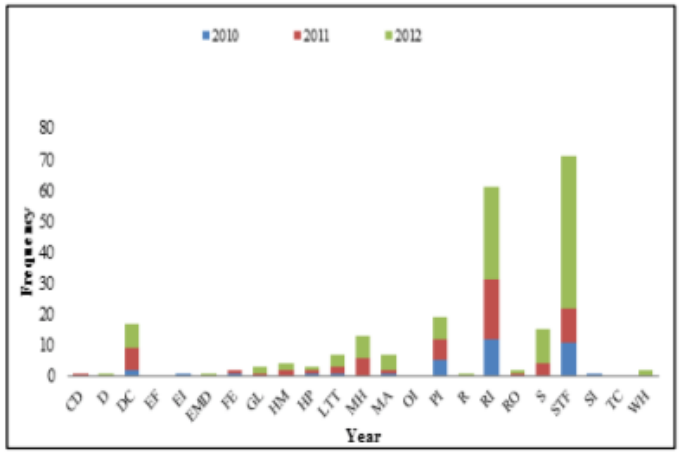

Fig. 2. Occurrence of primary causes in the years 2010-12.

Some important works related to the application of text mining in the field of accident analysis are road

crash analysis in Australia, aviation accident analysis, and so on. Some of the researchers use Leximancer concept mapping, text retrieval method, and link analysis. 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 [35-40].

3.0 RESEARCH METHODOLOGY

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. 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. 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. FTA is a top-down method designed to analyze the effects of initiating faults and events on any complex system. 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. 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.

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. 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.

4.0 RESULT

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: 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: 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).

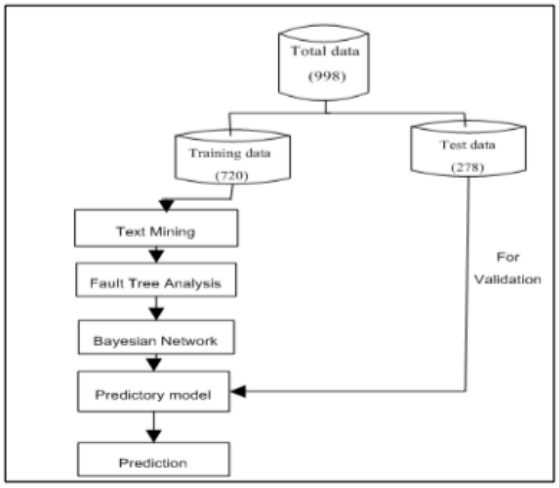

Fig. 3. Flowchart of the proposed methodology.

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%.

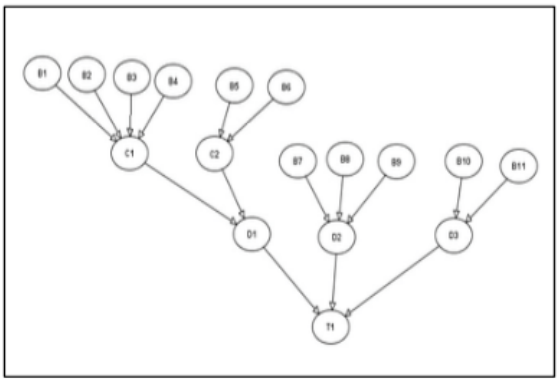

Fig.4. BN of STF at department A2.

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. 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.

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		

5.0 CONCLUSION

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. 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. 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.

REFERENCES

[1] Dimitrijevic, Branislav, et al. Segment-Level Crash Risk Analysis for New Jersey Highways Using Advanced Data Modeling. No. CAIT-UTC-NC62. Rutgers University. Center for Advanced Infrastructure and Transportation, 2020.

[2] Li, Chang, et al. "Machine learning for text mining based on prediction of occupational accidents and safety risk calculation." Australian Journal of Engineering and Applied Science 13.6 (2020): 11-17.

[3] Bahrami, Javad, Viet B. Dang, Abubakr Abdulgadir, Khaled N. Khasawneh, Jens-Peter Kaps, and Kris Gaj. "Lightweight implementation of the lowmc block cipher protected against side-channel attacks." In *Proceedings of the 4th ACM Workshop on Attacks and Solutions in Hardware Security*, pp. 45-56. 2020.

[4] Chen, Lee, et al. "Machine learning established by using crowdsourced investigation vehicle data for forecast of expressway crash risk ." International Journal of Applied Science and Information Science 11.8 (2020): 356-363.

[5] Ahmadinejad, Farzad, Javad Bahrami, Mohammd Bagher Menhaj, and Saeed Shiry Ghidary. "Autonomous Flight of Quadcopters in the Presence of Ground Effect." In *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pp. 217-223. IEEE, 2018.

[6] Zhang, Lixuan, et al. "Machine Learning Models established toward the Car Smash Injury Difficulty." European Journal of Applied Engineering and Basic Sciences 19.17 (2020): 4678-4685.

[7] Bozorgasl, Zavareh, and Mohammad J. Dehghani. "2-D DOA estimation in wireless location system via sparse representation." In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 86-89. IEEE, 2014.

[8] Xiang, Zheng, et al. "Machine Learning process for injury severity prediction and Traffic accidents classification." International Journal of Management System and Applied Science 23.12 (2020): 997-1003.

[9] Amini, Mahyar, and Aryati Bakri. "Cloud computing adoption by SMEs in the Malaysia: A multi-perspective framework based on DOI theory and TOE framework." Journal of Information Technology & Information Systems Research (JITISR) 9.2 (2015): 121-135.

[10] Silva, Philippe Barbosa, Michelle Andrade, and Sara Ferreira. "Machine learning applied to road safety modeling: A systematic literature review." Journal of traffic and transportation engineering (English edition) 7.6 (2020): 775-790.

[11] Amini, Mahyar. "The factors that influence on adoption of cloud computing for small and medium enterprises." (2014).

[12] AlMamlook, Rabia Emhamed, et al. "Comparison of machine learning algorithms for predicting traffic accident severity." 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). IEEE, 2019.

[13] Amini, Mahyar, et al. "Development of an instrument for assessing the impact of environmental context on adoption of cloud computing for small and medium enterprises." Australian Journal of Basic and Applied Sciences (AJBAS) 8.10 (2014): 129-135.

[14] Rezapour, Mahdi, Amirarsalan Mehrara Molan, and Khaled Ksaibati. "Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models." International journal of transportation science and technology 9.2 (2020): 89-99.

[15] Amini, Mahyar, et al. "The role of top manager behaviours on adoption of cloud computing for small and medium enterprises." Australian Journal of Basic and Applied Sciences (AJBAS) 8.1 (2014): 490-498.

[16] Rahim, Md Adilur, and Hany M. Hassan. "A deep learning based traffic crash severity prediction framework." Accident Analysis & Prevention 154 (2021): 106090.

[17] Amini, Mahyar, and Nazli Sadat Safavi. "Critical success factors for ERP implementation." International Journal of Information Technology & Information Systems 5.15 (2013): 1-23.

[18] Siebert, Felix Wilhelm, and Hanhe Lin. "Detecting motorcycle helmet use with deep learning." Accident Analysis & Prevention 134 (2020): 105319.

[19] Amini, Mahyar, et al. "Agricultural development in IRAN base on cloud computing theory." International Journal of Engineering Research & Technology (IJERT) 2.6 (2013): 796-801.

[20] Yang, Yang, et al. "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods." Physica A: Statistical Mechanics and Its Applications 595 (2022): 127083.

[21] Amini, Mahyar, et al. "Types of cloud computing (public and private) that transform the organization more effectively." International Journal of Engineering Research & Technology (IJERT) 2.5 (2013): 1263-1269.

[22] Fu, Yuchuan, et al. "A decision-making strategy for vehicle autonomous braking in emergency via deep reinforcement learning." IEEE transactions on vehicular technology 69.6 (2020): 5876-5888.

[23] Amini, Mahyar, and Nazli Sadat Safavi. "Cloud Computing Transform the Way of IT Delivers Services to the Organizations." International Journal of Innovation & Management Science Research 1.61 (2013): 1-5.
[24] Alkinani, Monagi H., Wazir Zada Khan, and Quratulain Arshad. "Detecting human driver inattentive and aggressive driving behavior using deep learning: Recent advances, requirements and open challenges." Ieee Access 8 (2020): 105008-105030.
[25] Amini, Mahyar, and Nazli Sadat Safavi. "A Dynamic SLA Aware Heuristic Solution For IaaS Cloud Placement Problem Without Migration." International Journal of Computer Science and Information Technologies 6.11 (2014): 25-30.
[26] Wahab, Lukuman, and Haobin Jiang. "Severity prediction of motorcycle crashes with machine learning methods." International journal of crashworthiness 25.5 (2020): 485-492.
[27] Amini, Mahyar, and Nazli Sadat Safavi. "A Dynamic SLA Aware Solution For IaaS Cloud Placement Problem Using Simulated Annealing." International Journal of Computer Science and Information Technologies 6.11 (2014): 52-57.
[28] Muhammad, Khan, et al. "Deep learning for safe autonomous driving: Current challenges and future directions." IEEE Transactions on Intelligent Transportation Systems 22.7 (2020): 4316-4336.
[29] Sadat Safavi, Nazli, et al. "An effective model for evaluating organizational risk and cost in ERP implementation by SME." IOSR Journal of Business and Management (IOSR-JBM) 10.6 (2013): 70-75.
[30] Cai, Qing, et al. "Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data." Transportation research part A: policy and practice 127 (2019): 71-85.
[31] Sadat Safavi, Nazli, Nor Hidayati Zakaria, and Mahyar Amini. "The risk analysis of system selection and business process re-engineering towards the success of enterprise resource planning project for small and medium enterprise." World Applied Sciences Journal (WASJ) 31.9 (2014): 1669-1676.
[32] Wahab, Lukuman, and Haobin Jiang. "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity." PLoS one 14.4 (2019): e0214966.
[33] Sadat Safavi, Nazli, Mahyar Amini, and Seyyed AmirAli Javadinia. "The determinant of adoption of enterprise resource planning for small and medium enterprises in Iran." International Journal of Advanced Research in IT and Engineering (IJARIE) 3.1 (2014): 1-8.
[34] Ziakopoulos, Apostolos, and George Yannis. "A review of spatial approaches in road safety." Accident Analysis & Prevention 135 (2020): 105323.
[35] Safavi, Nazli Sadat, et al. "An effective model for evaluating organizational risk and cost in ERP implementation by SME." IOSR Journal of Business and Management (IOSR-JBM) 10.6 (2013): 61-66.
[36] Mannering, Fred, et al. "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis." Analytic methods in accident research 25 (2020): 100113.
[37] Khoshraftar, Alireza, et al. "Improving The CRM System In Healthcare Organization." International Journal of Computer Engineering & Sciences (IJCES) 1.2 (2011): 28-35.
[38] Mokhtarimousavi, Seyedmirsajad. "A time of day analysis of pedestrian-involved crashes in California: Investigation of injury severity, a logistic regression and machine learning approach using HSIS data." Institute of Transportation Engineers. ITE Journal 89.10 (2019): 25-33.
[39] Abdollahzadegan, A., Che Hussin, A. R., Moshfegh Gohary, M., & Amini, M. (2013). The organizational critical success factors for adopting cloud computing in SMEs. Journal of Information Systems Research and Innovation (JISRI), 4(1), 67-74.
[40] Silva, Philippe Barbosa, Michelle Andrade, and Sara Ferreira. "Machine learning applied to road safety modeling: A systematic literature review." Journal of traffic and transportation engineering (English edition) 7.6 (2020): 775-790.