



**ANALYSIS OF CONSTRUCTION SAFETY HAZARDS USING
OPEN DATA AND TEXT MINING**

BY

MS. N.K.A. HESHANI RUPASINGHE

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE
(ENGINEERING AND TECHNOLOGY)**

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2020

COPYRIGHT OF THAMMASAT UNIVERSITY

**ANALYSIS OF CONSTRUCTION SAFETY HAZARDS USING
OPEN DATA AND TEXT MINING**

BY

MS. N.K.A. HESHANI RUPASINGHE

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE
(ENGINEERING AND TECHNOLOGY)**

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2020

COPYRIGHT OF THAMMASAT UNIVERSITY

THAMMASAT UNIVERSITY

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THESIS

BY

MS. N.K.A. HESHANI RUPASINGHE

ENTITLED

ANALYSIS OF CONSTRUCTION SAFETY HAZARDS USING OPEN DATA AND
TEXT MINING

was approved as partial fulfilment of the requirements for
the degree of Master of Science (Engineering and Technology)

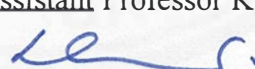
on December 23, 2020

Chairman




(Assistant Professor Kevin Tantisevi, Ph.D.)

Member and Advisor




(Associate Professor Kriengsak Panuwatwanich, Ph.D.)

Member



(Associate Professor Mongkut Piantanakulchai, Ph.D.)

Director



(Professor Pruettha Nanakorn, D.Eng.)

Thesis Title	ANALYSIS OF CONSTRUCTION SAFETY HAZARDS USING OPEN DATA AND TEXT MINING
Author	Ms. N.K.A. Heshani Rupasinghe
Degree	Master of Science (Engineering and Technology)
Faculty/University	Sirindhorn International Institute of Technology/ Thammasat University
Advisor	Assoc. Prof. Dr. Kriengsak Panuwatwanich
Academic Years	2020

ABSTRACT

The construction industry is known to have the highest rate of injuries and accidents around the world. Many research studies are engaged in analysing risks based on this industry using various techniques, and safety is a key consideration. According to the existing literature, it has been found that the contribution of sources of hazards such as worker factor, technological factor, natural factor, surrounding activity and organisational factors are the severe cause of accidents. However, data collection for these studies has been carried out through questionnaires and interviews, which could be susceptible to subjectivity and inaccuracies. There is a likelihood of answers being biased, skipped questions, interpretation issues, lack of nuance and accessibility issues. Besides, interviews rely on the respondent's ability to accurately and honestly answer the questions being asked without involving fear for legal circumstances and being bias. The root cause for the occupational injuries might not be reliably alarmed whenever the organisational factors such as safety and management issue forefront the accident. Thus, the objectives of this research study are to employ Natural Language Processing technique for automatic extraction of sources of hazards and hazards within sources of hazards from open data, to evaluate the performance of various existing statistical classifiers with developed classifiers, to identify the distribution of extracted sources of hazards and the contribution of hazards relevant to each source of hazard through Pareto Analysis.

The existing statistical classifiers required hundred thousand of data for better performance and work poorly whenever the positive training sample lacks data. Therefore, this study presents a comprehensive methodology for rule-based extraction tool development for sources of hazards identification with 95% threshold and classifier training for hazard identification within each source of hazard extracted. Random Forest showed the highest

accuracy in extracting hazards from each source of hazard. Descriptive results showed that the worker factors were the highest root cause of the occupational accidents. Further, Pareto analysis showed that 80% of the contributing hazards for each source of hazard factor. In addition, secondary analysis using one-way ANOVA showed that Summer has the highest potential for causing accidents compared to other seasons. The developed rule-based classifier and trained Random Forest classifier to analyse digital textual documents would be a great help for the present, and future of construction safety management and this study would be a well-defined domain for such analysis.

Keywords: Construction industry, Hazards, Sources of hazards, Text analysis, Frequency analysis, safety science



ACKNOWLEDGEMENT

It is a great pleasure to devote this page in writing an appreciation for each individual to acknowledge the countless support given in the success of this study.

I am highly indebted to my advisor, Assoc. Prof. Dr Kriengsak Panuwatwanich for his guidance, supervision and continuous support which has given throughout the study period.

Also, I would like to thank my committee chairman Asst. Prof. Dr Kevin Tantisevi and member Assoc. Prof. Dr Mongkut Piantanakulchai for their valuable comments and insights given in the success of this study.

I would like to express my gratitude the Excellent Foreign Students (EFS) scholarship programme offered by Sirindhorn International Institute of Technology, Thammasat University, for me to get a higher educational opportunity.

Further, I would make this a chance to pay great regards to my parents for their support and care given to me by always standing beside me during all the hard times I have passed and for motivating me to achieve my goals.

Last but not least, I would like to acknowledge the help given by my friends, all the non-mentioned academic and non-academic staff members who assisted me in countless ways to make this report a success.

Ms. N.K.A. Heshani Rupasinghe

TABLE OF CONTENTS

	Page
ABSTRACT	(1)
ACKNOWLEDGEMENT	(3)
TABLE OF CONTENTS	(4)
LIST OF TABLES	(8)
LIST OF FIGURES	(10)
LIST OF SYMBOLS/ABBREVIATIONS	(11)
CHAPTER 1 INTRODUCTION	1
1.1 Research background	1
1.2 Research problem	2
1.3 Research scope	3
1.4 Research objectives	4
CHAPTER 2 REVIEW OF LITERATURE	5
2.1 Development of occupational health and safety	5
2.2 Hazards in construction	8
2.3 Sources of hazards in construction	10
2.3.1 Definition	10
2.3.2 Sources of hazards	10
2.4 Construction accident predictors	14
2.5 Accident predictive models	14
2.6 Automation in construction	15
CHAPTER 3 RESEARCH METHODOLOGY	18

3.1 Data collection	19
3.2 Categories of sources of hazards	19
3.3 Text analysis	20
3.3.1 Natural Language Processing (NLP)	22
3.3.2 Key phrase extraction for lexicon building	22
3.3.3 Identification of N-grams	25
3.3.4 Composing extraction rules	27
3.3.5 Data pre-processing	30
3.3.6 Validation of the tool	32
3.4 Comparison of the model with existing classifiers	33
3.4.1 Support Vector Machine (SVM)	33
3.4.2 Random Forest (RF)	33
3.4.3 k Nearest Neighbours (kNN)	34
3.4.4 Kernel SVM	34
3.4.5 Naïve Bayesian (NB)	34
3.5 Classifier training for hazard identification	35
3.5.1 Identified hazards in natural factors	36
3.5.2 Identified hazards in organisational factors	37
3.5.3 Identified hazards in surrounding activity	38
3.5.4 Identified hazards in technological factors	39
3.5.5 Identified hazards in worker factors	40
3.5.6 Utilised libraries and modules	42
3.5.7 TFIDF (Term Frequency Inverted Term Frequency)	43
3.5.8 Bag of word model (BoW)	43
3.5.9 Split data into a training set and test set	44
3.5.10 Fit trained data into a classifier	44
3.6 Frequency analysis	44
3.7 One-way ANOVA	44
3.7.1 Validation of the data for normality	45
3.7.2 Validation of the data for homogeneity	45
3.7.3 Non-parametric Kruskal-Wallis H test	45
3.7.4 Post hoc tests	46
3.7.5 Calculation of effect size	46

CHAPTER 4 RESULTS AND DISCUSSION	47
4.1 Performance evaluation of developed rule-based classifier	47
4.1.1 Comparison of rule-based model performance with statistical classifiers	48
4.2 Performance evaluation of the classifiers trained to identify hazards	48
4.3 Distribution of sources of hazards	49
4.3.1 Worker factor hazards	50
4.3.2 Technological factor hazards	51
4.3.3 Surrounding activity related hazards	51
4.3.4 Organisational factor hazards	52
4.3.5 Natural factor hazards	53
4.3.6 Comparison of extracted results with existing literature	53
4.3.7 Monthly distribution of sources of hazards	55
4.3.8 Seasonal distribution of sources of hazards	56
4.4 Normalisation of accident data	57
4.5 Literature study on seasonal variations of occupational accidents	59
4.6 One-way ANOVA	60
4.6.1 Assessing the assumptions	60
4.6.2 Outputs of non-parametric Kruskal-Wallis H test statistics for different seasons, years and months	67
4.6.3 Outputs of non-parametric Kruskal-Wallis H test for seasonal change on natural factor	68
4.6.4 Outputs of non-parametric Kruskal-Wallis H test for sources of hazards depending on the month	68
4.6.5 Welch test on natural factors based on monthly accident Data	69
4.7 Outputs of post hoc test for seasonal behaviour of accidents	69
4.7.1 Dependent variable: Total number of accidents	69
4.7.2 Dependent variable: Total number of accidents per 50 billion dollars	71
4.8 Output of post hoc test for monthly behaviour of accidents	74
4.8.1 Dependent variable: Total number of accidents	74
4.8.2 Dependent variable: Total number of accidents per 50 billion dollars	76
4.9 Output of seasonal distribution of construction spending	78
4.10 Output of post hoc test for seasonal behaviour of accidents based on sources of hazards	80

4.11 Output of post hoc test for monthly behaviour of accidents based on sources of hazards	86
4.12 Discussion on seasonal variation of accidents distribution	89
CHAPTER 5 CONCLUSION	91
5.1 Conclusion	91
5.2 Contribution	92
5.3 Future work	94
REFERENCES	95
APPENDICES	108
APPENDIX A	109
APPENDIX B	110
APPENDIX C	126
APPENDIX D	136
APPENDIX E	138
APPENDIX F	142
APPENDIX G	143
APPENDIX H	148
APPENDIX I	153
BIOGRAPHY	162

LIST OF TABLES

Tables	Page
2.1 Types of errors and violations	12
2.2 Use of text mining in construction industry	15
3.1 Categories of sources of hazards	20
3.2 Sample of extracted key phrases	23
3.3 Sample for lexicon	26
3.4 Summery of number of phrases in N-grams files and number of rules for each factor	26
3.5 Comprehensive list of extraction rules	28
3.6 Examples for hazards in natural factors	36
3.7 Examples for hazards in organisational factors	37
3.8 Examples for hazards in surrounding activity	38
3.9 Examples for hazards in technological factors	39
3.10 Examples for hazards in worker factors	40
3.11 Description of imported libraries	43
4.1 Summary of model performance at each iteration	47
4.2 Performance of statistical classifiers	48
4.3 Performance of trained classifiers	48
4.4 Tabulated data for sources of hazards distribution	49
4.5 Test of normality for seasonal data	60
4.6 Test of normality for yearly data	61
4.7 Test of normality for monthly data	62
4.8 Test of normality for sources of hazards in each season	63
4.9 Test of normality for sources of hazards in each month	64
4.10 Test of homogeneity variances on seasonal data	66
4.11 Test of homogeneity variances on yearly data	66
4.12 Test of homogeneity variances on monthly data	67
4.13 Kruskal Wallis test statistics	67
4.14 Test statistics for seasonal change on natural factor ^{a,b}	68
4.15 Test statistics for monthly accident data on sources of hazard ^{a,b}	69
4.16 Welch test on natural factor	69
4.17 Multiple comparisons of total number of accidents occurred in each season	70
4.18 Descriptive results of total number of accidents occurred in each season	70

4.19 ANOVA of total number of accidents occurred in each season	71
4.20 Test of normality	71
4.21 Test of homogeneity variances	72
4.22 ANOVA of total number of accidents per 50 billion dollars	72
4.23 Multiple comparisons of construction accidents per 50 billion dollars	73
4.24 Descriptive of normalised total of construction accidents	73
4.25 Significant multiple comparisons of total number of accidents occurred in each month	74
4.26 Descriptive results of total number of accidents occurred in each month	75
4.27 ANOVA of total number of accidents occurred in each month	76
4.28 Test of normality	76
4.29 Test of homogeneity variances	77
4.30 Significant multiple comparisons of total number of accidents per 50 billion dollars	77
4.31 ANOVA of construction accidents per 50 billion dollars	77
4.32 Descriptive results of construction accidents per 50 billion dollars	78
4.33 Test of normality for construction spending on seasons	78
4.34 ANOVA for construction spending in million dollars	79
4.35 Multiple comparisons of construction spending in million dollars with seasons	79
4.36 Descriptive of construction spending in million dollars	80
4.37 ANOVA of total number of accidents occurred due to sources of hazard in each season	83
4.38 Multiple comparisons of total number of accidents occurred due to each sources of hazard in each season	83
4.39 Descriptive results of total number of accidents occurred due to each sources of hazard depending on season	85
4.40 Significant multiple comparisons of total number of accidents occurred due to each sources of hazard in each month	87
4.41 Descriptive results of total number of accidents occurred due to each sources of hazard depending on month	87
4.42 ANOVA of total number of accidents occurred due to sources of hazard in each month	89

LIST OF FIGURES

Figures	Page
2.1 Development of occupational health and safety	7
2.2 Hierarchy of influences in construction	9
3.1 Flow of the research methodology	18
3.2 Rule based classifier building and validation	21
3.3 Sample trigram list	31
3.4 Sample of bag of word model	44
4.1 Distribution of sources of hazards	49
4.2 Worker factor hazards	50
4.3 Technological factor hazards	51
4.4 Surrounding activity-related hazards	52
4.5 Organisational factor hazards	52
4.6 Natural factor hazards	53
4.7 Monthly distribution of accidents	56
4.8 Seasonal distribution of sources of hazards	57
4.9 Construction spending data distribution	58
4.10 Monthly accident distribution	59
4.11 Normal Q-Q plot for each season (a) Winter (b) Spring (c) Summer (d) Autumn	61

LIST OF SYMBOLS/ABBREVIATIONS

Symbols/ Abbreviations	Terms
NLP	Natural Language Processing
TM	Text Mining
OHS	Occupational Health and Safety
OIICS	Occupational Injury and Illness Classification Manual
CAD	Computer Aided Design
OHSA	Occupational Health and Safety Administration
NAICS	North American Industry Classification System
US	United States
DT	Decision Tree
RF	Random Forest
NB	Naïve Bayesian
ANN	Artificial Neural Network
PPR	Post-Project Reviews
SVM	Support Vector Machine
k-NN	k- Nearest Neighbours
BoW	Bag of Words
SPSS	Statistical Package for Social Sciences
NF	Natural Factor
OF	Organisational Factor
SA	Surrounding Activity
TF	Technological Factor
WF	Worker Factor
ANOVA	Analysis of Variances

CHAPTER 1

INTRODUCTION

Occupational accidents have been a vital phenomenon in the construction industry (Yılmaz & Kanit, 2018). In order to minimise these construction accidents, many studies have considered further analysis based on different variables. Regardless of the technological improvements to develop the construction health and safety management, the contribution of sources of hazards such as worker factors, natural factors, surrounding activities, organisational factors and technological factors on the health and safety cannot be downplayed. Therefore, construction health and safety-related research studies have a great enthusiasm on investigating effective health and safety management system (Chen & Jin, 2012), constructing frameworks and models on-site safety climate (Choudhry et al., 2009; Q. Li et al., 2017), and forecasting and amplifying the performance of the safety (Fang et al., 2015; Xia et al., 2018). Further, in recent years automated technologies have become more widespread application in construction site health and safety management (De Melo et al., 2017; Dong et al., 2018; P. X. Zou et al., 2017).

This chapter consists of a brief introduction to the research background on construction health and safety-related accidents, research problem, research scope and the objectives. Reviewed literature study including the development of health and safety, hazards in construction, sources of hazards, construction accident predictors, and automation in construction are discussed in Chapter 2. The methodology of data aggregation, sources of hazards identification, text analysis for extraction of sources of hazards and one-way ANOVA is presented in Chapter 3. Chapter 4 discusses the findings on the performance of developed text mining tool, distribution of categorical variables obtained from open data, and text mining, identified hazards related to each source of hazard and the significance of seasonal change on occupational accidents obtained by following the methodology. Chapter 5 includes the conclusion, contribution, and future work.

1.1 Research background

The construction industry is known as an industry which has the highest rate of injuries and accidents around the world. Construction site safety has been discussed since the second half of the 19th century, and safety is key consideration. According to the International Labour Organisation, over 2.78 million fatalities occur due to work-related accidents or ill health

annually. Further, there has been recorded more than 374 million non-fatal workplace injuries every year, which follows four days of continuous absences in work (*Health and safety at work*).

Occupational health and safety are not a mere responsibility of contractors, designers and consultant. It is also a responsibility of the client to advocate site safety (Jitwasinkul & Hadikusumo, 2011). Lack of safety measures goes beyond the health concerns since the cost of construction injuries plays a notable role on the financial stability and increase the overall expenses of construction up to 15% (Hallowell, 2011).

However, the factors which influence the construction safety cannot be eliminated entirely due to both technical and economic reasons (Hoła, 2010). Technological improvement of the era demands more construction designs, and hence workplace safety has begun to play a significant role in the construction industry. Therefore, it has been scrutinised in different views and angles of the researchers in the world for years (Fredericks et al., 2005; Hassanein & Hanna, 2008; Liaudanskiene et al., 2010).

In recent years, some research studies have been even trying to introduce Natural Language Processing (NLP) which is a text mining technique to the construction industry for addressing the management issues of textual documents such as automatic analysis of injury reports (Tixier et al., 2016b), automatic clustering of construction project documents based on textual similarity (Al Qady & Kandil, 2014), and retrieval of CAD drawings (Hsu, 2013). Text mining has immense potential for future project improvement, avoiding mistakes, making aware of previously unknown facts through high accuracy (Choudhary et al., 2009).

1.2 Research problem

Construction site accidents analysis is a prevalent topic among the researchers even though the workplace accident mitigation is yet an inefficient task. Most of the construction industries around the world generate catastrophe investigation report which includes a full description of the event after an accident occurred (Zhang et al., 2019). These data have been publicly available for research use in some countries, and yet, utilisation of these data for safety management is scarce. Further, the research that employs techniques such as text mining and Natural Language Processing to extract data from construction data reports are becoming popular in construction automation (Choudhary et al., 2009; Zhang et al., 2019; Y. Zou et al., 2017). However, none of the research has utilised these reports to identify the root cause of the accident. Moreover, real hazard inside the source of hazard has never been identified through these open data.

Data aggregation for most of the research studies which analyse hazards are mainly accomplished through questionnaire surveys, semi-structured interviews and structured surveys which are not a very reliable source of collecting data (Debios, 2009). The survey questionnaires often filled with biased answers, skipped questions, interpretation issues, lack of nuance and accessibility issues as well as interviews rely on the respondent's ability to accurately and honestly answer the questions being asked without involving fear for legal circumstances and being bias. The root cause for the occupational injuries might not be reliably alarmed whenever the organisational factors such as safety and management issue forefront the accident. Thus, the accuracy of the answers in questionnaires and interviews lacks due to the above reasons.

Therefore, this study adopts the data collected from catastrophe investigation reports where publicly made available in occupational health and safety organisation for the hazard identification (*Severe Injury Reports | Occupational Health and safety Administration*). However, these reports are gathered in Excel files in thousands of rows, and manual extraction of required data for analysis is time-consuming. Therefore, this study attempts to develop a text mining technique to extract sources of hazard from open data.

1.3 Research scope

The scope of this research work is to develop a tool to extract information from 8,940 final narrative data of construction site severe injury data reports which collected through catastrophe investigation reports from January 2015 to September 2019. These are publicly made available at occupational health and safety administration website of the United States. The study develops a rule-based classification for the sources of hazard extraction and assesses the performance of the developed classifier with existing statistical classifiers. Further, the study employs the foremost existing statistical classification tool to identify actual hazard within each source of hazard. The performance of the classifiers is evaluated using the average weighted F1 score and F1 score. Then the obtained categorical data is analysed by using descriptive statistical analysis and Pareto Analysis.

The development of text mining model ensures the reduction of the time taken to extract data manually and utilisation of data reported through catastrophe investigation reports amplifies the accuracy of the data rather than the conventional data collection methods such as questionnaires and interviews. Besides, trained classifier for hazard identification assures the immediate recognition of hazards consists of sources of hazards.

In addition to the principal analysis discussed above, secondary analysis is conducted using one-way ANOVA to discover the seasonal and monthly variation of the extracted sources

of hazards.

1.4 Research objectives

The research problems encountered under Heading 1.2, is addressed by defining three main objectives. They are:

- a) To employ NLP technique for automatic extraction of sources of hazards;
- b) To evaluate the performance of various existing statistical classifiers with developed classifiers; and
- c) To gain practical insights from the extracted sources of hazards and hazards through statistical analyses.



CHAPTER 2

REVIEW OF LITERATURE

2.1 Development of occupational health and safety

The history of public health concerned articles can be found since the late 1750s. According to Century (1800), public health concept was derived from many historical ideas, trial and error methods with the improvement of basic sciences and through technology and epidemiology. Early in the 19th century, construction workplace had no health and safety measures for physical labours. Thus, it seems that construction labour health and safety was not addressed in this era. However, the situation began to change by healthier means with the establishment of insurance plans in the second half of the 19th century. Employees started to purchase insurance plans rather than being acquisitive. Also, some of the employers willingly provided insurance plans for their employees (Durisko, 1997). Moreover, some workers started to leave their career due to the unsafe work environment, and employers had to raise the wages of the employees who were engaged in high-risk activities. As a result of these conflicts between the employee and the employer, industry policy changes were started to develop.

In 20th century, researchers started to pay attention to the health and safety of the construction site occupants. Accidents and ill-health were particularly severe, and according to the statistics of U.S Department of Labour, three hundred out of one hundred thousand (300/100000) employees were killed annually in mining activities which was the most popular industry in the 1900s. As per Durisko (1997) half of such injured parties were given a little compensation which was equal to half of the annual wages of the victim and safety was little concern in the climate. However, in 1910 worker compensation law was introduced in New York to provide worker compensation at a fixed rate. Since then, National Safety Council was founded in 1913 to promote the health and safety of the Americans, U.S. Department of Labour was established in 1913 to focus on occupational health and safety, Federal Compensation Act was declared in 1916 to comfort workers who suffer from injuries or prone to illnesses while working. Eventually, by 1921, forty-four number of states successfully adopted the worker compensation law.

Prevention of injuries and protection of their health were given special attention (Woodbury, 1927) and noise and hearing protection was ranked the first place as mining works became a prevalent industry in the era (Barrett & Calhoun, 1900; Burns et al., 1962).

Nevertheless, the very first use of safety nets and hard hats were started while constructing the Golden Bridge in San Francisco, USA in 1933.

Laufer (1987) stated that accidents prevention measures were introduced from various concerns such as humanitarian, legal, company image and cost. The systematic study of accidents cost was initially documented by Schnee Heinrich in the 1920s by classifying the costs as direct and indirect. In some other studies conducted in the late 1950s, accident cost was classified as insured and uninsured (Simonds & Grimaldi, 1956). The total costs caused only by the occupational hazards in terms of wages, medical expenses, insurance claims, production delays, lost time of co-workers and equipment damage was estimated by national safety council of United States and accounted that 15 billion dollars during 1974. It was approximately 1% of the gross national production at that time (Ashford, 1975). Since then, many safety investigators focused their attention on accident costs (Bird et al., 1974; Everett & Frank Jr, 1996; Gilmore, 1970; Leopold & Leonard, 1987; Rinefort, 1977).

In 1971 Occupational Health and Safety Administration (OSHA) was created by occupational health and safety act which was passed in 1970 to ensure the safety of the workers in the site by setting and introducing standards while providing training, outreach, education and assistance (Durisko, 1997).

After all, the National Occupational Research Agenda was unveiled in 1996 in order to research on reduction of construction site injuries and illnesses precisely. Hence, in the 21st-century construction site, health and safety have become vastly discussed topic among researchers and continuing to improve workplace safety through various assessments to eliminate fatalities.

In recent years, some researchers have been even trying to introduce Natural Language Processing (NLP) which is a text mining technique to the construction industry for addressing the management issues of textual documents such as automatic analysis of injury reports (Tixier et al., 2016b), automatic clustering of construction project documents based on textual similarity (Al Qady & Kandil, 2014) and retrieval of CAD drawings (Hsu, 2013). Text mining has immense potential for future project improvement, avoiding mistakes, making aware of previously unknown facts through high accuracy (Choudhary et al., 2009). Novel research in construction automation further utilises artificial intelligence and integrated hybrid model named Symbiotic Grated Recurrent Unit (SGRU) for the safety assessment of construction projects (Cheng et al., 2020).

Above discussed development of health and safety in the construction industry is summarised and presented below in Figure 2.1.

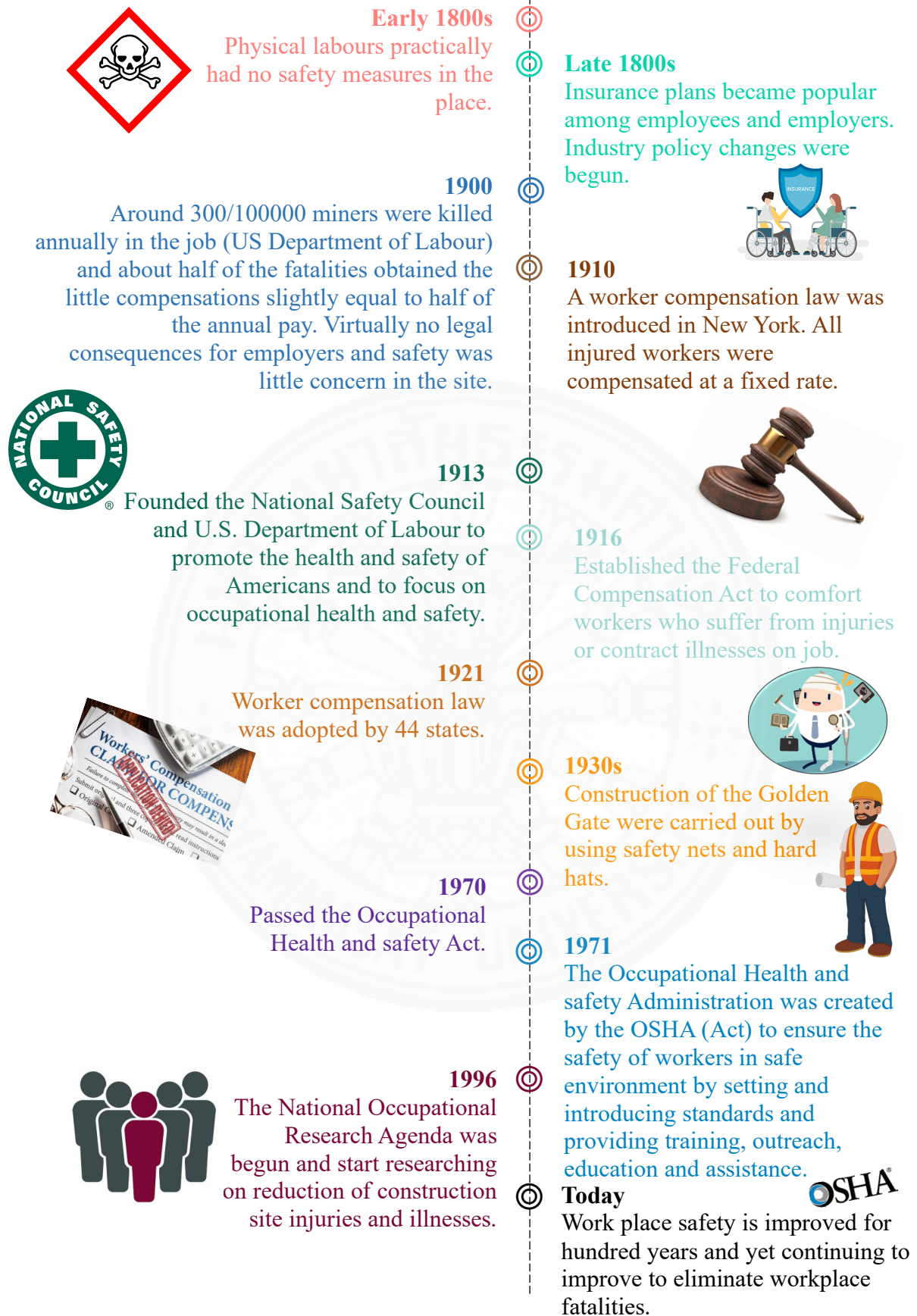


Figure 2.1 Development of occupational health and safety

2.2 Hazards in construction

In occupational health and safety, term hazard has been defined in several ways. It is defined as an agent which has a high probability of causing harm to an endangered target in a particular environment. Simply, the Health and Safety Commission defines a hazard as “the potential cause to harm”. The term hazard defines by Hamid et al. (2003) as anything potential to cause harm such as formwork arrangements, excavation, roof works, working on scaffoldings, etc. Further, it is defined as a situation which is connected to the work process or production process and is denoted by such a layout or state of factors of this operation, which may arise an accident in work or occupational disease (Carter & Smith, 2006). However, the international labour organisation defines a hazard as the inherent potential to cause injury or damage to people’s health.

Researchers have found two main categories of hazards in the construction industry. Firstly, injury hazards which are usually attached to the process of works or equipment used in different climatic conditions (Davies & Tomasin, 1996). Secondly, risks of ill health or health hazard which is typically categorised under chemical, physical and biological hazards (King & Hudson, 1985). However, the injury hazards can be identified at the time of occurrence and health hazards can be noted after symptoms started to appear in long term basis. Thus, the term hazard can be simply encapsulated as any substance or a factor which causes ill-health or loss of life.

The existence of hazards in the construction workplace is an outturn of the actions such as planning and preparation, site environment, project management and safety culture (Hide et al., 2003). Further, underlying the influences for accident causation, Hide et al. (2003) developed the accident causation model by stating a hierarchical influence on construction accidents. As shown in Figure 2.2, all the accidents are a combination of originating influences, shaping factors, worker factors, site factors and material and equipment factors. It demonstrates that the accidents appear from the failures of the worker interactions with their workplace, material and equipment. Doubles arrows in the models illustrate multiple two-way interactions between those primary factors. These immediate accident circumstances are influenced by three main shaping factors; worker factors, site factors and material/equipment factors. These shaping factors are influenced by the originating influences; permanent works design, project management, construction processes safety culture and risk management, as shown in Figure 2.2. However, higher the hierarchy, it shows that the accidents are pioneered from client requirements, economic climate and construction education which frames construction itself a potential cause for harm.

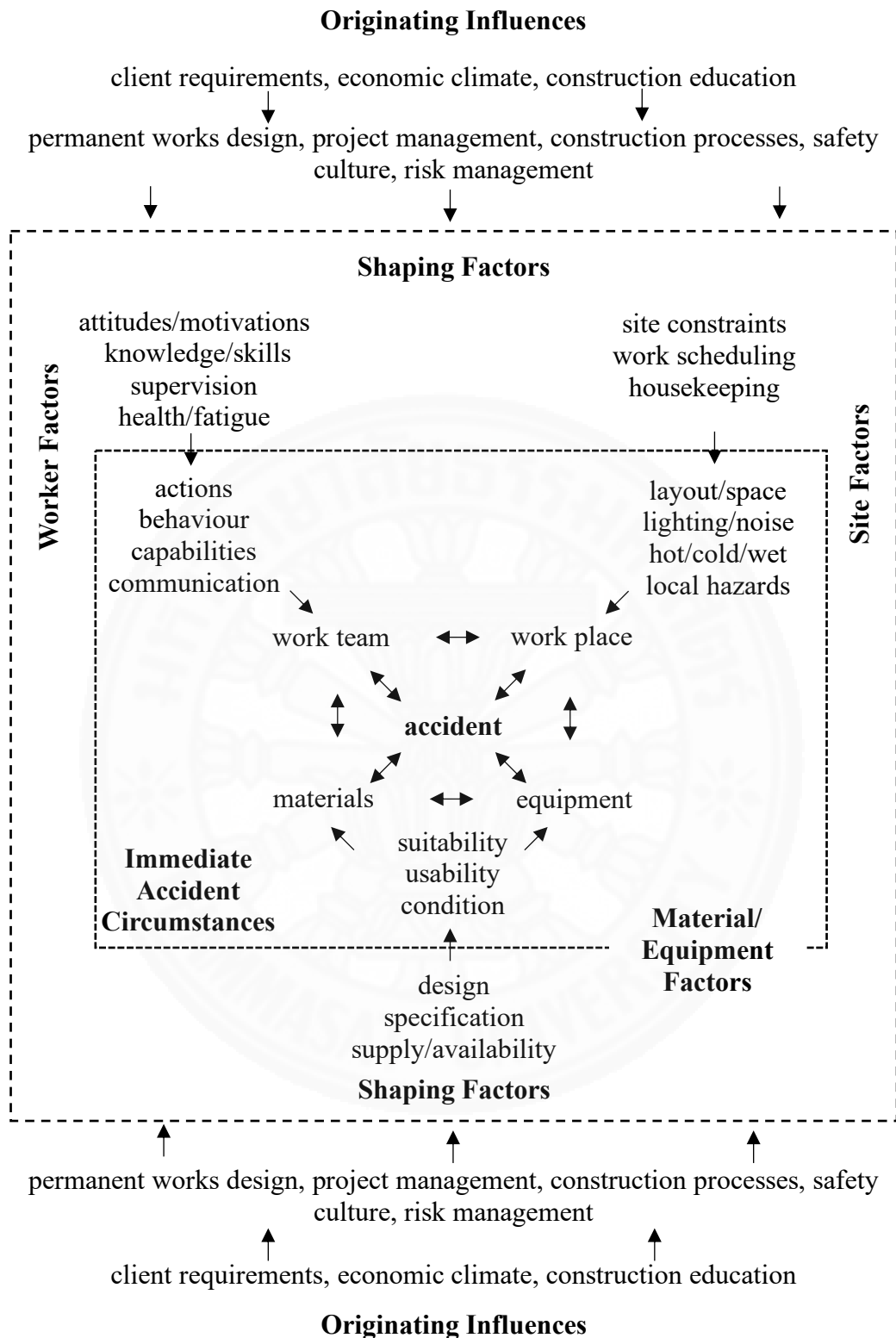


Figure 2.2 Hierarchy of influences in construction

2.3 Sources of hazards in construction

2.3.1 Definition

A hazard source is a location or condition that can give rise to a hazard (*Severe Injury Reports | Occupational Health and safety Administration*). Further, ("Risk management - Principles and guidelines Sydney: Standards Australia,") given that the hazard sources are defined as elements which alone or in combination have the intrinsic potential to give rise to risk. The commonly used classification of hazard includes biological (micro-organism, bacteria, viruses, insects, plants and animals etc.), chemical (toxicants, toxins which affects the body or chemicals that leads to fire and explosion), physical (electricity, pressure, noise, height, vibration), ergonomics (repetitive movement, manual handling, workplace design, job and task design) and psych social (stress, violence, other workplace stressors) (Macdonald, 2012). These hazard categories; physical, ergonomics, chemical and psychosocial are escalated by human factors while psychosocial hazards rise due to organisational factors as well as, biological hazards can be influenced due to natural factors by figuring those as hazard sources. Therefore, discussion on different sources of hazards is provided below.

2.3.2 Sources of hazards

“Universal framework” by McClay (1989) discovered three main components in an accident as risk, human actions and practical limitations. Distraction theory of Hinze (1997) proclaimed that the pressure arises from the workload can cause the workers to distract from the hazards and amplify the risk of accidents. The “Constraints-response” model developed, illustrates that distal factors (project conditions and management decisions) can cause for inappropriate site environment or actions (proximal factors) which leads to an accident (Suraji et al., 2001). Therefore, distal factors and the proximal factors are considerable two accident influencing factors in the construction industry.

The top managements’ viewpoint about safety (Levitt, 1975) or organisational culture (Molenaar et al., 2002), safety climate (Mohamed, 2002), foreman practices (Levitt & Samelson, 1993) and turnover (Hinze, 1980) are highly influenced on safety performance on the site. Furthermore, Mitropoulos et al. (2005) identified another three sources of hazards; work technology, physical conditions and surrounding activities.

However, the source of hazard, human factors occur due to human errors are identified as the central element of the accident (Mitropoulos et al., 2005). In combination to all of the above, Hoła (2010) stated that sources of hazards in the construction industry are technical factors

which cause due to machines, materials and devices (work technology) the organisational factors which occur due to subcontractor systems (management) who carry out the tasks which include safety culture, scheduling, work planning, supervision and procedures. Human factors which occur due to human errors, the external environment which goes along with the physical conditions and other civil structures situated within or outside the building site which can be matched with surrounding activities (Mitropoulos et al., 2005).

2.3.2.1 Work technology

Tools (power tools, scaffolds, cranes, drilling machines, cutting tools etc.), materials (chemicals, gasses, liquids etc.) and actions (loading, unloading, climbing, descending etc.) required to perform a specific task is involved in work technology. Either operating the activity in the pre-planned work conditions and climate or performing the work in different climate under different conditions may involve different hazards (Mitropoulos et al., 2005). However, technological hazards were less defined and some overlaps with the interpretation of planned human activities which causes hazards (Gunn, 1990; Kasperson & Pijawka, 1985). Tool, equipment or machinery breakdown, technical fault and errors are initiated with human involvement even though these are categorised as technological hazard factors (Gunn, 1990).

2.3.2.2 Physical conditions

The physical working environment of the site, such as deep excavations, floor openings, confined spaces, overhead material storages, overhead power lines, trenches, and layout etc. can cause another set of hazards. Also, environmental conditions such as illumination, wind, heat, cold, vapour, noise etc. may multiply or deduct the effect of hazards (Mitropoulos et al., 2005).

2.3.2.3 Surrounding activities

Surrounding activities such as dropping of objects from elevated levels, heavy vehicle movement, vibrations due to sheet piling etc. also generates threats in the site (Mitropoulos et al., 2005).

2.3.2.4 Human factors

The HSE definition for human factor is “Human factors refers to environmental, organisational and job factors and human and individual characteristics which influence behaviour at work in a way which can affect health and safety” (Kerr et al., 2009). According

to Muir and Thomas (2004), human factors are the issues concerning the workers' performance, and the possible contributions cause an accident event. Woods and Dekker (2000) stated that broadly used definition for human factor is "human factor concerns the interaction between people, their characteristic and abilities, organisation and management and technology".

Reason (1990) identified human factors in five different ways; three types of errors and two types of violations, as shown in Table 2.1.

Table 2.1 Types of errors and violations

Type	Hazard	Description
Errors	Slips and Lapse	These are the type of human errors which occur with little or absence of conscious thought. It is an unintended error while executing the correct plan.
	Mistakes	This can also be represented as decisional errors executed in such a way that correct plan in an inappropriate manner. It is purely a planned activity yet, inappropriate for the situation.
	Perceptual errors	These are the errors occurred due to misinterpretation of the actual work.
Violations	Routine violations	Routine violations are habitual deviations from the original plan, which were tolerated by the supervision. Occupants who behave oppose to the specified practise standards are falling into this category. As an example, driving in higher speed than 5mph in a place where the speed limit is limited to a maximum of 5mph is a routine violation.
	Exceptional violations	These violations are usually rare in the workplace and usually occurs in emergencies due to unusual circumstances. It is an intentional or instinctive reaction to the situation.

According to the Ferrell's Human Factor Theory by Russell Ferrell, above mentioned human errors are occurred due to overload, incompatibility and improper activities (Abdelhamid & Everett, 2000). Moreover, human performance in a particular situation is varied by the workers' expectation on what will happen, their past experiences in similar situation and assessment on consequences. Nevertheless, emotional factors such as fatigue, pressure,

teamwork, stress and fear can affect human performance and decision making (Garrett & Teizer, 2009). As specified by the HEAT (Human Error Awareness Taxonomy), organisational influences, supervisory influences, preconditions, and acts/events provoke the probability of triggering human errors.

2.3.2.5 Organisational factors

Organisational factors contribute to the most significant influence of individual behaviour and group behaviour (Kerr et al., 2009). They include poor health and safety culture, poor work planning, leading to high work pressure, scheduling, supervision and procedures, management based on one-way communication, inadequate responses to previous incidents etc. In the hierarchical risk breakdown introduced by (Rezakhani, 2012) shows that project management which includes in organisational factors is a primary level hazard factor and hierarchy is followed by technical and managerial complexity, planning and controlling, project team selection, decision making, communication and unavailability of resources respectively. Since the human involves in design, manufacture, management of complex technological systems, operate and maintain, it is clear that the human actions and decisions compromised in every organisational factor (Reason, 2016). Swiss Cheese model stated that the workers involved in unsafe acts due to latent failures such as poor design, gaps in supervision, undetected manufacturing defects or maintenance failures, unworkable procedures, clumsy automation, shortfalls in training, less than adequate tools and equipment and unseen for years and if aligned, it becomes an active failure in the system (Reason, 1990; Reason, 2017).

2.3.2.6 Natural hazards

United Nations International Strategy for Disaster Reduction (UNISDR) (2009) defines a natural hazard as a natural process or phenomenon which might negatively affect the society and environment. Gill and Malamud (2016) show earthquakes, floods, tsunamis, landslides, the ground collapses, storm, lightning and extreme temperature both hot and cold etc. as examples for natural hazards. However, natural factors also triggered by human activities and cause accidents which damages not only to the construction industry but also to a considerable area. For example, the landslide occurred in the Kashmir earthquake in 2005 was aggravated by the road construction (Owen et al., 2008).

2.4 Construction accident predictors

Construction accidents are caused by many different site circumstances. Many studies have been conducted to investigate those contributing factors using past data. However, researchers claim that main contributors such as safety risk and low hazard recognition have been downplayed while discussing the predictive models (Carter & Smith, 2006; Namian et al., 2016; Tixier et al., 2014). Safety climate has been widely claimed as a predictor of safety outcomes (Glendon & Litherland, 2001; Johnson, 2007; Zohar, 1980). Tam et al. (1998) have identified safety management strategies such as post-accident investigation, safety training, safety awards and portion of subcontracting as safety outcome predictors.

Neeleman et al. (2001) found that personnel characteristics as one of the main contributing factors for accidents. Studies show that old age is more likely to be a contributing factor as the mental and physical ability to adapt to the environment decreases with the increment of age (Garg, 1991). Despite that Jiang et al. (2011) identified controversial result that in general, many young workers tend to have accidents rather than the old workers. Moreover, studies claim that gender is associated with the occurrences of accidents, and male workers have a higher potential in being a victim (Loomis et al., 1997). Chi et al. (2005) showed that inexperience of the workers in small to medium size construction industries is highly prone to accidents. In contrast, Zhang et al. (2016) persist that statistical significance of individual differences of the workers are not considerable, whereas site conditions have a negative impact on accidents.

Further, age, company size, location, mechanism, geography, season, gender, experience, time, week, type of work and weather are considered as casual factors in previous research studies (Arquillos et al., 2012; Chi et al., 2013; Huang & Hinze, 2003; Jeong, 1998; Ling et al., 2009).

2.5 Accident predictive models

In addition to the common accident-causing factors discussed above, recent research studies adopt forecasting technologies using existing data and suggest the need of the more empirical and quantitative research to prevent the accidents (Tixier et al., 2016a). Thus, Yan et al. (2005) utilised the multiple logistic regression model to characterize the rear-end accidents at signalized intersections. Nishimoto et al. (2017) developed a severe injury prediction algorithm based on large scale data and under-triage control. Choi et al. (2020) developed a predictive model based on national data for fatalities of construction workers by applying

machine learning, and Gerassis et al. (2017) developed a Bayesian decision tool for accident analysis in embankment construction.

2.6 Automation in construction

Construction automation has become a widespread topic in recent years. There is an increment in adopting new technologies such as robotics, sensory machines and ICT (Balaguer & Abderrahim, 2008). Research studies have found means to automate construction planning (Faghihi et al., 2015), enhance the productivity and quality (Kamaruddin et al., 2016), material handling (Alumbugu et al., 2019), defect identification in concrete (Liu et al., 2019), and building information modelling (BIM) (Li et al., 2019; Liu et al., 2019; Matarneh et al., 2019; Ozturk, 2020; Tang et al., 2019). Moreover, artificial intelligence (AI) has been employed for planning (Jiao et al., 2019), safety management (Baker et al., 2019; Nozaki et al., 2018), and construction management (Ko & Cheng, 2003).

Text mining is a widespread Artificial Intelligence technology that uses NLP to transform unstructured documents into normalised, structured data for information retrieval, data mining, machine learning, statistics, and computational linguistics (Rai, 2019). Therefore, it has been used recently by construction-related research for construction management processes, and safety analysis. Table 2.2 summarises the findings of research studies which used text mining in the field of construction.

Table 2.2 Use of text mining in construction industry

Paper	Use of text mining	Notes
The needs and benefits of Text Mining applications on Post-Project Reviews (Choudhary et al., 2009)	To identify the need for text mining and its benefits for post-project reviews (PPR).	It uncovers the patterns, associations and trends from PPR reports.
Identifying work-related injuries: comparison of methods for interrogating text fields (McKenzie et al., 2010)	Interrogated the text mining field to identify work-related injuries presented on emergency departments in Queensland to inform the surveillance of work-related injury using narrative text.	Basic keyword search, index search and context analytic text mining were used.

Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims. (Bertke et al., 2012)	Utilised text mining to classify worker's medical compensations into "claim causation" categories.	Naïve Bayesian classifier was used and found that the implementation of lower-level categories could significantly drop the accuracy of the predictions.
Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques (Fan & Li, 2013)	Build a model to retrieve similar cases for alternative dispute resolution in construction accidents using text mining techniques.	Vector spaced model was used for retrieval.
Injury narrative text classification using factorization model (Chen et al., 2015)	Utilised text mining to automatically classify narratives in emergency medical reports into injury codes.	Non-negative matrix factorization (NNMF) based classifier is used for training, and it achieved the best performance among other classifiers.
Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports (Tixier et al., 2016b)	Utilised NLP as a tool for text mining to extract precursors and outcomes from unstructured injury reports.	Rule-based content analysis system was used to analyse the data and manually encoded 2200 injury reports from Desvignes (2014) were utilised for training.
Construction accident narrative classification: An evaluation of text mining techniques (Goh & Ubeynarayana, 2017)	Text mining is used to label the document using the predefined class of labels.	The initial round of classification was done using six classifiers, and SVM has identified as the best classifier.
An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green	Text mining technique is adopted to extract the features of green building-related cases.	

building design (Shen et al., 2017)		
Construction site accident analysis using text mining and Natural Language Processing techniques (Zhang et al., 2019)	Developed and ensembled model using NLP and text mining to identify the most common accident-causing objects and common types of accident events.	Ensembled model and SVM were identified as best models for retrieval.

Accordingly, text mining has a vital contribution to the construction industry to overcome many analyses associated with a large amount of data reports.



CHAPTER 3

RESEARCH METHODOLOGY

This chapter presents the methodology adopted in this study to reach the objectives presented in Heading 1.4. There are five main steps in this methodology, and they are data collection, identification of sources of hazards, extraction tool building, frequency analysis and one-way ANOVA for comparison of the means. The flow of the methodology is shown in Figure 3.1.

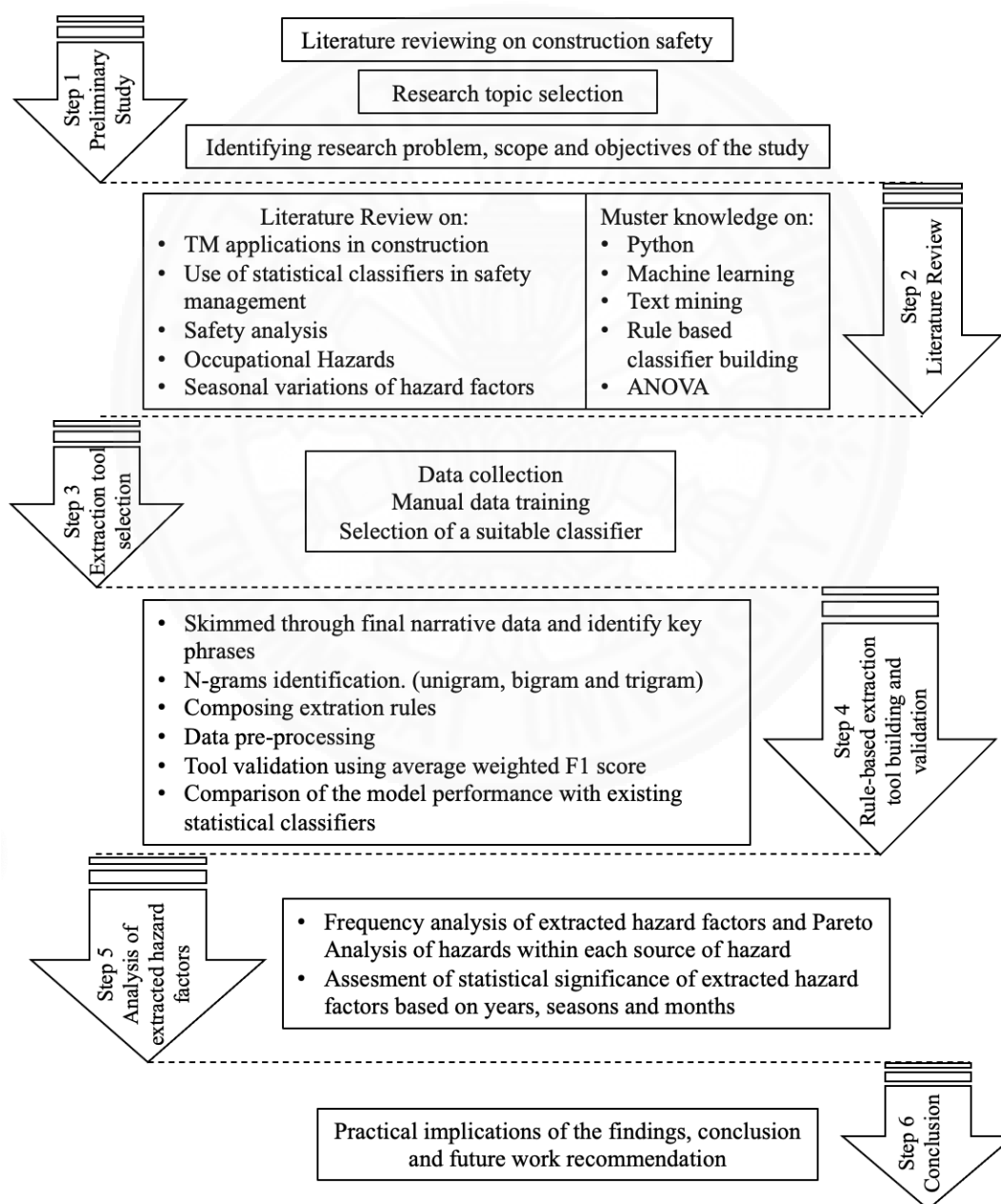


Figure 3.1 Flow of the research methodology

3.1 Data collection

Data for this study was obtained from severe injury open data set available at the Occupational Health and Safety Administration, department of labour, United States (*Severe Injury Reports | Occupational Health and safety Administration*). It had been recorded the accident data from January 02, 2015, to September 30, 2019. Not only the construction site accidents, but also the accidents occur in agriculture, forestry, fishing, hunting, mining, utilities, manufacturing, transportation and warehousing, finance and insurance, educational services, public administration, information, wholesale trade etc. was recorded and available for free download. Downloaded data consisted of event date, company or the employer, address, primary NAICS (North American Industry Classification System), severity, nature of accidents, part of the body damaged, cause of the accident, final narrative and secondary source of the accident. All together 50,032 data were recorded.

However, the present study focused on the construction accident-related data only. It was extracted using the primary NAICS, and Macro was used for the separation of data. (See APPENDIX A). After the extraction, 8,940 number of constructions-related accident data was obtained, and the study utilised the accident description available in ‘final narrative’ column to identify the sources of hazards through text mining.

3.2 Categories of sources of hazards

Sources of hazards were defined as a condition or a factor which would lead to hazard in the construction site environment. Thirteen number of sources of hazards (originating influences, shaping factors, worker factors, site factors, material and equipment factors, distal factors, proximal factors, work technology, physical conditions, surrounding activities, human factors, organisational factors and natural factors) were identified while reviewing the existing literature and they are discussed in Heading 2.2 and Heading 2.3. However, it had been identified that the human factor was the cause for many described hazardous events, and the human was the potential cause for harm. Therefore, the study divided the human factors into worker factors, organisational factors and technological factors. However, technological factors were defined by including the material and equipment factors too. Organisational factors were defined in the study as a combination of the distal factors and originating influences. Further, surrounding activities were defined as a combination of physical conditions and site factors. Finally, five primary sources of hazards were identified and listed in Table 3.1.

Table 3.1 Categories of sources of hazards

Source of Hazard	Description
Worker factor	Every possible error, violation, mental and health issue, lack of skill and behaviour of the workers inside the site environment
Technological factor	Tool, material, machinery or equipment breakdown and technical faults and errors occur while utilizing
Natural factor	Any natural phenomenon which negatively affects the site condition and causes harm
Organisational factor	Safety management, project conditions, management decisions and controlling which are beyond the level of the worker
Surrounding activity	Activities in progress inside or outside the site environment other than the activity that the victim is engaged

3.3 Text analysis

Text analysis is a tool/automated process which is employed to extract and classify information from a document in a textual format such as emails, customer review reports, survey responses, tweets etc. Moreover, it alters the textual data into numerical data which can be further utilised in data mining algorithms (Williams & Gong, 2014) such as k-Nearest Neighbour (kNN), Support Vector Machines (SVM), Naïve Bayesian algorithm, k-means, etc. However, such algorithms perform dreadfully when the positive sample for training is limited for each category (Prabowo & Thelwall, 2009). For such instances, hand-coded rules and keyword dictionaries were used to integrate human judgment and knowledge into the text mining system by increasing the accuracy (Tixier et al., 2016b). Therefore, due to the scarcity of open data reports rectified under Heading 3.1., text mining through NLP was carried out using the rule-based method as shown in Figure 3.2.

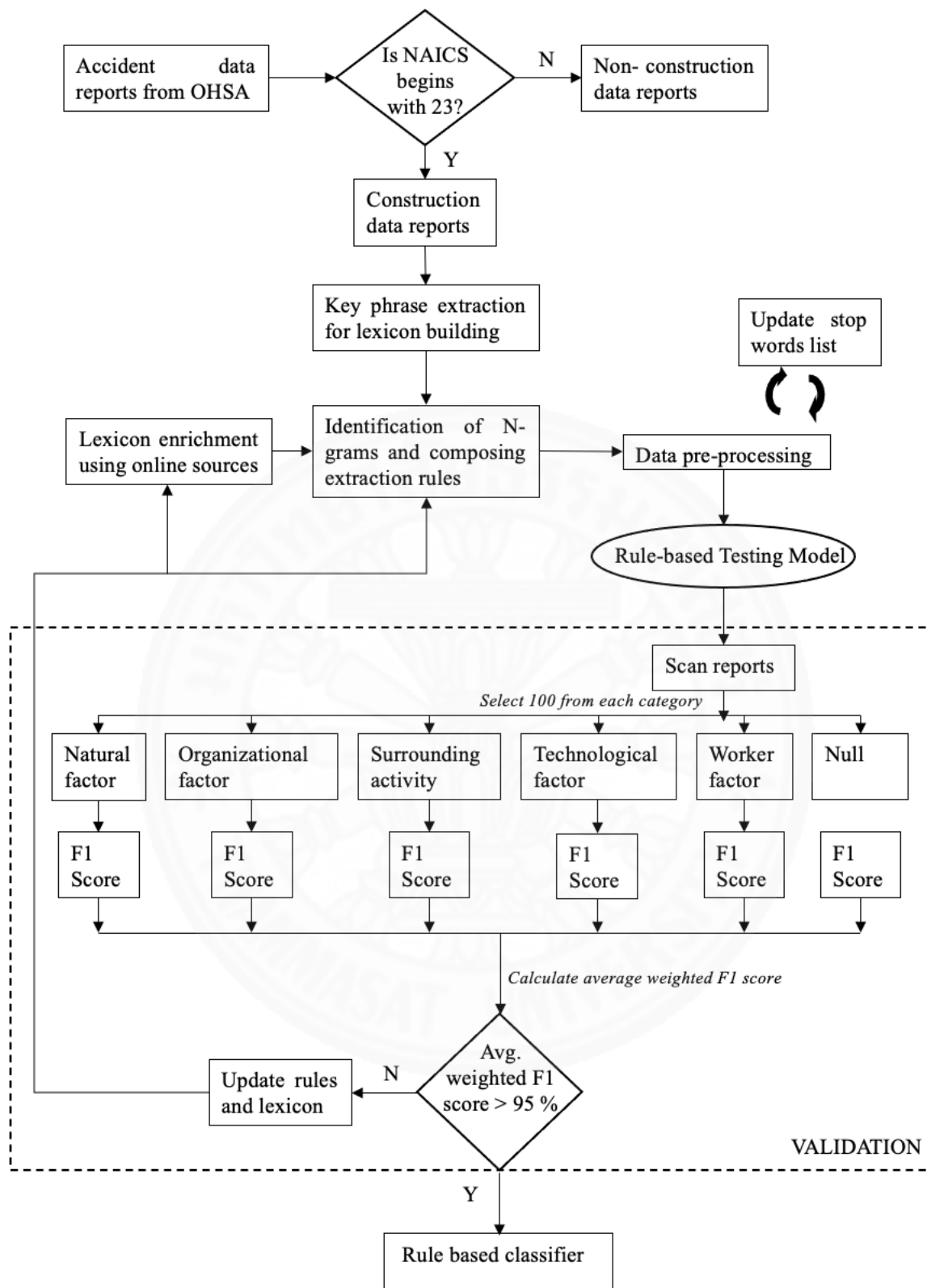


Figure 3.2 Rule based classifier building and validation

3.3.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a technique utilised in multiple areas such as computer science, mathematics and information engineering, artificial intelligence and computer linguistics (*Field of computer science and linguistics*, 2019). Solely, NLP is an interaction between the computer and the human languages. In this study, NLP was used as a tool to extract information from the final narrative data column obtained in the open data set. However, specific platforms, libraries, packages and data pre-processing was required to extract data.

The developing tool utilised was Spyder 3.2.3, which promotes and facilitates the use of Python 3.5 for scientific and engineering software development. Anaconda Navigator 1.3.1 created by Continuum Analytics was used as the desktop graphical user interface which supports to launch Spyder and easily manage Conda packages, environments and channels without using command line prompt.

Before every text mining activity through NLP, manual text mining was required to train the computer to extract required features through NLP. However, in this scenario, manual text mining was represented as lexicon building as the text mining was accomplished through the rule-based method.

3.3.2 Key phrase extraction for lexicon building

The first step in designing the rule-based content analysis system was to extract critical phrases which describe the cause for the accident. These critical phrases were identified by reviewing the initial 1500 data and sample of extracted phrases, and the classification output was shown in Table 3.2.

For the first incident in Table 3.2, the source of the hazard was identified as natural factors as the key phrase which describes the accident was the word phrase '*gust of wind displaced a tree*'. Also, in the second example, the door was slammed due to '*strong gust of wind*'. Therefore, it also categorised under natural factor. In the third example, the accident was caused when employees '*lost their grip*'. Hence, it is a worker factor which occurred due to error made by the workers. In the fourth instance, the accident was occurred when the worker '*slipped and stepped on*' a protruding nail. In the fifth example, the accident was occurred as the worker '*jumped off*' the trailer. In the sixth example, the accident was occurred due to a surrounding activity which was a motor vehicle driven by a member of the general public. In the seventh example, it is mentioned that '*No fall protection was being used at the time*'. This showed that there were no proper safety precautions in the area and showed poor

safety management. Therefore, it was categorised under organisational factor. In the eighth example, there was no precise phrase to identify the key phrase for the cause of the accident. Therefore, it was categorised as null data in the study. In the ninth example, the accident was occurred since the stock of foot panels in an under-construction building *'fell on the employee'*. Thus, this is categorised under surrounding activity. Again, in the tenth example, the accident was occurred as the worker *'did not have a personnel fall arrest system'*. Failing to assign fall arrest systems to the worker or the failing to manage the safety of the workers was a matter of the organisation, and therefore, it was categorised under organisational factors. In the eleventh example, aerial lift act abnormally was a technological factor caused for the accident. Thus, it was categorised as a technological hazard source. In the twelfth example, the accident was occurred as the *'jib attachment slipped off'*. Therefore, this was categorised under technological factor.

Table 3.2 Sample of extracted key phrases

No	Final Narrative	Phrase for lexicon building	Source of Hazard
1	A worker was installing a retaining wall next to Highway 285 when a gust of wind displaced a tree next to the highway. The tree struck the worker, resulting in back, pelvis, and ankle injuries.	gust of wind displaced a tree	Natural factor
2	On or about February 18, 2015, an employee suffered an amputation of part of his left pinkie finger when an entry door slammed on it due to a strong gust of wind.	gust of wind	Natural factor
3	Two employees in a scissor lift were wrecking forms, taking forms down, and removing the forms. They lost their grip on the form, which fell and hit an employee walking below on the shoulder, knocking him to the ground. He hit his head on the concrete, causing a wound to the head, a broken collarbone, and a broken rib.	lost their grip	Worker factors

4	On Friday, January 2, 2015, at 10:30 p.m., a crew of employees were changing out wooden decking boards on a platform offshore. One of the employees was walking the platform, slipped, and stepped on a protruding nail that entered the employee's right foot. First aid was applied to the employee, and the employee was later sent to a clinic for treatment. The employee was admitted to a hospital at 3:00 a.m. on January 5, 2015, from swelling on his foot due to the incident.	slipped, and stepped on	Worker factors
5	Employee was loading plastic pipes onto a utility trailer. He jumped off the trailer and broke his ankle.	jumped off	Worker factors
6	At approximately 4:30 PM on January 28, 2015, an employee removing cones in the "closed lane" was struck by a motor vehicle driven by a member of the general public. The employee's left leg sustained a fractured femur and fibula. The employee was transported and admitted to Geisinger Medical Centre in Danville, PA.	vehicle driven by a member of the general public.	Surrounding activities
7	Employee fell approximately 15 feet. No fall protection was being used at the time.	No fall protection	Organisational factors
8	Finger cut accident.	-	Null
9	A stack of 12 x 4-foot panels of sheet rock in the hallway of an under-construction building fell on the employee and broke his leg.	fell on the employee	Surrounding activities

10	An employee fell from the second floor, at a height of approximately 10 feet, while installing sheetrock. The employee did not have a personal fall arrest system at the time of the accident.	employee did not have a personal fall arrest system	Organisational factors
11	An employee was pinched between an aerial lift and walkway. While attempting to come down, the lift acted abnormally. It instead went up and pinched the employee between the lift and walkway.	lift acted abnormally	Technological factors
12	Employee was transferring a mini dumpster into a larger dumpster when the jib attachment slipped off, hit the ground, bounced, and hit the employee.	jib attachment slipped off	Technological factors
13	Employee broke a leg at a construction yard.	-	Null
14	Employee caught hand in system.	-	Null

3.3.3 Identification of N-grams

N-gram is a series of n number words taken from a given document or speech. These N-grams can be a letter, word, syllables or base pairs according to the administration of the term (*N-gram*). In this study N-grams were identified as unigrams, bigrams and trigrams with the help of identified key phrases and sample lexicon were presented in Table 3.3. As an example, unigrams of the sentence ‘employee fell off’ are (‘employee’), (‘fell’), (‘off’). Bigrams of the same sentence are (‘employee’, ‘fell’), (‘fell’, ‘off’). Trigrams are (‘employee’, ‘fell’, ‘off’). The list of N-grams obtained by manually analysed reports was presented in Table 3.2 and consisted of only contributing words to extract the root cause of the accident. The full version of the Lexicon is available in APPENDIX B. Further, N-grams were enriched by identifying probable common words related to each category for future anticipated cases. For instance, for the source of hazard, natural factor, unigrams such as ‘thunderstorm’, ‘landslide’, ‘tornado’ and ‘tsunami’ was added even though they were not found in firstly observed reports.

These identified N-grams were then written in separate text files under headings such as ‘nfuni.txt’ for unigrams of natural factors, ‘nfbi.txt’ for bigrams of natural factors, ‘nftri.txt’

for trigrams of natural factors etc., for further modification and to import inside the Python code.

Table 3.3 Sample for lexicon

Source of hazard	Unigram	Bigram	Trigram
Natural factor	climate cyclone earthquake	bush fire ground collapse heat wave	gust wind caused ice/slick roads strong gust wind
Organisational factor	bitten unguarded unhooked	fall protection floor opening management issues	guardrail not place high work pressure no fall protection
Surrounding activity	automobile dog motorcycle	dog chased flagging traffic general public	exposed fall hazards foreign body injected unknown object fell
Technological factor	electrocuted entangled malfunction	blade broke electrical shock hook broke	broke from choke cap came off carbon monoxide poisoning hazardous chemical splashed jib attachment slipped
Worker factor	attempted dizziness fainted	accidentally stepped became dizzy experience cramping	employee become overheated experienced severe dehydration lose his balance

Table 3.4 Summary of number of phrases in N-grams files and number of rules for each factor

Source of hazard	Number of Unigrams	Number of Bigrams	Number of Trigrams	Number of rules
Natural factor	10	14	8	3
Organisational factor	5	23	19	3
Surrounding activity	5	25	66	3
Technological factor	20	172	213	3
Worker factor	26	147	171	6

3.3.4 Composing extraction rules

Table 3.5 demonstrates the comprehensive list of extraction rules according to the order of implementation, and every rule was written as a combination of statements using the ‘if-else if’ function in Python.

According to the manually extracted data set, unigram ‘*attempt*’ was a more frequent act while occurring an accident and always lead to worker factor. Therefore, before searching for other unigrams, the word ‘*attempt*’ was searched first in the final narrative unigram set and which has ‘*attempt*’ were categorised under worker factor. However, the word ‘*attempt*’ can be either written in as same as ‘*attempt*’ or can be written as ‘*attempting*’, ‘*attempted*’ etc. These types of suffixes were removed while creating N-grams lists and removal of these terms and conversion of unstructured data into structured data were discussed below in Heading 3.3.5. After categorising all the records which had word ‘*attempt*’ under worker factors, bigrams of natural factors were checked. This was done to remove the conflicts which are likely to arouse by having N-grams, which can fall into two or more categories. As an example, final narrative “*As the employee was holding the window unit, a gust of wind blew and the employee lost his grip on it*” had bigrams of [“*wind*”, “*blew*”] which falls to natural factor and [“*lost*”, “*his*”, “*grip*”] which fall to worker factor. However, it was clear that the worker lost his grip due to the gust of wind and root cause for the accident was the gust of wind. By checking the bigrams of the natural factors, the root cause of the accident was identified and categorised under the most accurate sources of hazard

Then in the third and fourth step [“*he*”, “*slip*”] and [“*employee*”, “*slip*”] was checked with the bigrams of the final narrative. This can be explained using the following example. “*The injured employee slipped and fell in front of the other employee who was operating a Lull at the time. The Lull ran over the injured employee’s foot.*” Here the accident was not occurred as the Lull ran over the employee. This might fall into surrounding activity if trigrams were checked before checking bigram of [“*employee*”, “*slip*”]. The accident occurred as the employee slipped on. Therefore, this way of ordering rules allowed the program to categorised these final narratives under worker factor. After that, trigrams, bigrams and unigrams were checked respectively for each category. However, in each category, worker factors were checked at last to minimise the cooccurrences as discussed above. In the final step, all the final narratives which did not fall into any other category was named as null. This included the final narratives which did not state the cause of the accident directly.

Table 3.5 Comprehensive list of extraction rules

Step No	Statement returns true if... (elif)	Description	Output
1	('attempt') in unigram	Checks whether the word 'attempt' is in the unigram list of final narrative data (unigram)	Worker factor
2	any(check in bigram for check in bigramnf)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of natural factors (bigramnf)	Natural Factor
3	('he', 'slip') in bigram	Checks whether the words 'he', 'slip' are in the bigram list of final narrative data	Worker factor
4	('employe', 'slip') in bigram	Checks whether the words 'employe', 'slip' are in the bigram list of final narrative data	Worker factor
5	any(check in trigram for check in trigramnf)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of natural factors (trigramnf)	Natural Factor
6	any(check in trigram for check in trigramof)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of organisational factors (trigramof)	Organisational factor
7	any(check in trigram for check in trigramsa)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of surrounding activity (trigramsa)	Surrounding activity
8	any(check in trigram for check in trigramtf)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of technological factors (trigramtf)	Technological factor

9	any(check in trigram for check in trigramwf)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of worker factors (trigramwf)	Worker factor
10	any(check in bigram for check in bigramof)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of organisational factors (bigramof)	Organisational factor
11	any(check in bigram for check in bigramsa)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of surrounding activity (bigramsa)	Surrounding activity
12	any(check in bigram for check in bigramtf)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of technological factors (bigramtf)	Technological factor
13	any(check in bigram for check in bigramwf)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of worker factors (bigramwf)	Worker factor
14	any(check in unigram for check in unigramtf)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of technological factors (unigramtf)	Technological factor
15	any(check in unigram for check in unigramof)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of organisational factors (unigramof)	Organisational factor
16	any(check in unigram for check in unigramsa)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of surrounding activity (unigramsa)	Surrounding activity

17	any(check in unigram for check in unigramwf)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of worker factors (unigramwf)	Worker factor
18	any(check in unigram for check in unigramnf)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of natural factors (unigramnf)	Natural Factor
19	else	Final narratives which do not fall into any of above	Null

3.3.5 Data pre-processing

Data pre-processing was required before the creation of N-grams using Python as the final narrative data was unstructured. Pre-processing includes punctuation removal, uppercase to lowercase, tokenisation, stop word removal, stemming and lemmatisation and append.

a) Punctuation removal: This step includes removing all the characters except alphabetic characters. Punctuations are not treated as a significant character in NLP, and it usually increases the size of the training dataset. Therefore, in this step, word complexity due to “Employee1” and “Employee2” was eliminated and treated as one word “Employee”. However, the white spaces among between words are kept as it is for uncomplicated handling of the document.

b) Uppercase to lower case: All the uppercase letters were converted into lower case letters and thus, eliminated the complications aroused due to the same word treated as different words. After the transformation, as an example “Machine” and “machine” is treated as one word “machine”.

c) Tokenisation: The document is chopped into words and created a token for each word. After the tokenisation sentence “employee was fallen down” will turn in to an array of words [“employee”, “was”, “fallen”, “down”].

d) Stop word removal: These are the most common words that exist in a sentence and which adds low value to the meaning of the sentence in text mining (*Dropping common terms: stop words*). Generally, stop words are determined by the frequency of word appear in the document and then filtered the most frequent terms such as “a”, “an”, “the”, “is”, “was”, “were”, “the”, ‘not’, ‘didn’t”, “and”, “be”, “by”, “for”, “from”, “to”, “I”, “my”, “he”, “myself” etc. However, to maintain the sense of the N-grams, some stop words were removed from the

stop word list, and some words were added to the stop word list. For instance, if an accident occurred due to “not having fall protection system” is an organisational factor. This will not be accurately identified if the stop word list consists of “not”. Therefore, the word “not” was removed. Likewise, words such as “not”, “didn’t”, “weren’t”, ‘wasn’t”, “off”, “over”, “on”, “out”, “by”, “with”, “her”, “his”, “it”, “through”, “lower”, “himself”, “herself”, “themselves” etc. were removed from the original stop word list. Moreover, to reduce the complexity of the meaning of the sentence words such as “ft”, “feet”, “inches”, “degree” etc. were added to the stop words list.

e) Stemming and lemmatisation: The same word can be express in different formats due to their tense (eg. “call”, “called”, “calling”), the singularity (eg. “calls”) and plurality (eg. “call”). Moreover, the same word can be expressed in word family that have the same base format (“collect”, “collectively”, “collection”). Scenarios, as mentioned above, were standard grammatical rules and eliminated and convert into its base form in this process.

f) Append: In the final data pre-processing step, usually, all the documents are appended into a corpus. However, in this study, each sentence was appended to lists of N-grams, and separate lists of unigrams, bigrams and trigrams were created. Created trigram lists can be shown as in Figure 3.3.

ide ▲	Type	Size	Value
0	tuple	3	('fell', 'upper', 'upper')
1	tuple	3	('upper', 'upper', 'elev')
2	tuple	3	('upper', 'elev', 'foreign')
3	tuple	3	('elev', 'foreign', 'bodi')
4	tuple	3	('foreign', 'bodi', 'inject')
5	tuple	3	('bodi', 'inject', 'backho')
6	tuple	3	('inject', 'backho', 'ran')
7	tuple	3	('backho', 'ran', 'over')
8	tuple	3	('ran', 'over', 'piec')
9	tuple	3	('over', 'piec', 'lumber')

Figure 3.3 Sample trigram list

3.3.6 Validation of the tool

The validation process was carried out in each time after extraction of sources of hazards by using the output obtained in an Excel file. A random number was assigned to each injury report and separated to different Excel sheets according to the source of hazard using the Macro. (See APPENDIX D). However, to eliminate the display of manually trained data first, these outputs were then sorted according to the random number. Finally, the first hundred (100) data reports were scrutinised in each category and accuracy obtained through F1 score, which described below in Heading 3.3.6.1 was recorded at each repetition. This validation process played a crucial role in establishing a decidedly accurate system.

After a careful examination over six repetitions, the model performance was evaluated through counting all attributes for average weighted F1 score. However, it should be noted that imperfections are possible as the manual text analysis was done by the author and model can be further investigated and fixed by refining the n-gram files accordingly.

3.3.6.1 F1 score

Model performance is measured through F1 score proposed by Buckland and Gey (1994) as it has been widely used in reviewed literature (Choi et al., 2020; Fan & Li, 2013; Gerassis et al., 2017). It is a measure of test accuracy and provides a more realistic measure of model performance using precision and recall. A threshold of 95% was selected to be obtained from the implemented rule-based classifier. It is calculated by using Equation (3.1).

$$\text{F1 Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3.1)$$

Where;

TP = True positives

FP = False positives

FN = False negatives

Precision = $TP / (TP + FP)$

Recall = $TP / (TP + FN)$

However, conventional F1 score does not calculate the value for each label in the dependent variable. Therefore, the weighted average F1 score was obtained using Equation (3.2), and it was used to evaluate the tool performance in this study.

$$\text{Average Weighted F1 Score} = \sum_{i=1}^N \left(\frac{S_i}{T} \times F1_i \right) \quad (3.2)$$

Where;

N is the total number of labels.

S_i is the number of actual instances in the i^{th} label.

T is the actual predictions of all labels.

$F1_i$ is the F1 score of the i^{th} label.

N in this study was six (6), which was the number of categories. True positives (TP) in this study were the ones which were extracted correctly from the injury data report through the rule-based text mining model. False positives (FP) also called as ‘error’ which was detected falsely and wrongly categorised. Finally, false negatives (FN) were the ones which were falsely detected and should fall into the rest of the categories.

3.4 Comparison of the model with existing classifiers

Comparison of the model was made using the existing classifiers on the same data set, and performance was evaluated using the average weighted F1 score. For this, five main existing classifiers were discovered from existing literature. They were SVM, RF, k-NN, Kernel SVM and NB.

3.4.1 Support Vector Machine (SVM)

Support Vector Machine is a discriminative classifier which is associate with supervised learning algorithms. This classifier looks at the objects very close to the boundary of maximum margin hyperplane and tries to train the model. As an example, in a classification of apples and oranges, SVM looks at apples which are more like an orange and oranges which are more like an apple. However, Non-Support Vector Machines try to look at objects which are far from the boundary and tries to train the model.

3.4.2 Random Forest (RF)

Random Forest classifier is an advanced version of Decision Tree classifier which consists of several individual decision trees that operate as an ensemble model. However, the accuracy increases when the number of individual trees involved increases. This adopts a general technique of bootstrap aggregating to train trees. Predictions for new data point x' is

given by taking the average of all predictions from individual decision trees on x' or by taking the highest vote. This denotes by the following Equation (3.3).

$$f^{\wedge} = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (3.3)$$

Where;

In a given training set bagging repeatedly B times, $X = x_1, x_2, \dots, x_n$ called X_b and responses $Y = y_1, y_2, \dots, y_n$ called Y_b . For $b = 1, 2, \dots, B$. Decision Tree classifier f_b is trained on X_b and Y_b .

3.4.3 k Nearest Neighbours (kNN)

This is a non-linear non-parametric method use for both classification and regression. Input consists of 'k' nearest samples, and output can be differ depending on classification or regression. This model performs well when the sample has more neighbours, and classification accuracy will increase. However, having too many neighbours would cause overfitting and predicted results would be low for a different sample.

3.4.4 Kernel SVM

This is as same as SVM. The only difference is Kernel is used when the data distribution is non-linear.

3.4.5 Naïve Bayesian (NB)

Naïve Bayesian classifier non-linear model derived from Bayes theorem, as shown in Equation (3.4), which is in the family of simple probabilistic classifiers.

$$p(A/B) = \frac{p(B/A) \times p(A)}{P(B)} \quad (3.4)$$

Where;

A and B are events and $p(B) \neq 0$.

$p(A/B)$ is a conditional probability which is the likelihood of occurring event A while given that the B is true.

$p(B/A)$ is also a conditional probability which is the likelihood of event B occurring whenever the A is true.

$p(A)$ and $P(B)$ are probabilities of event A and event B independently.

In conditional probability of Naïve Bayesian given problem features are vectorized into a vector X where X denotes by $X_1, X_2 \dots$ and X_n . n represents independent variables and assigns to instant probabilities for each k possible outcomes of class C_k as $p(C_k/X_1, X_2, \dots X_n)$, and can be demonstrated by using Bayes theorem as in Equation (3.5).

$$p(C_k/X) = \frac{p(X/C_k) \times p(C_k)}{P(X)} \quad (3.5)$$

Bayesian probability Equation (3.5) can also be represented as in Equation (3.6).

$$\text{Posterior probability} = \frac{\text{likelihood} \times \text{prior probability}}{\text{marginal likelihood}} \quad (3.6)$$

Therefore, Naïve Bayesian classifier predicts the probabilities for each member class depending on the fact given that the new record belongs to a particular class. Achieved highest probability will be then defined as the most likely class where the new record belongs.

3.5 Classifier training for hazard identification

Extracted sources of hazards were separated into different Excel sheets using Macro (APPENDIX D). These each source of hazards factors were then examined for possible actual hazard. These hazard factors were identified by using the N-grams taken to identify the sources of hazards and 3the existing literature as a reference.

For instance, Peterson et al. (2013) stated that both extreme heat and cold waves effect the USA's both construction and non-construction industry. Walsh et al. (2016) stated that wind and hail hazards are significant cause of accidents in Australia. Bobick (2004) stated that fall through open areas including roof and floor areas are severe cause of accidents. Robbins et al. (2008) illustrated the importance of managing insects due to risk associated with mosquito borne diseases. Limited site visibility hazard is a key consideration in excavation works and in underground construction work (Zhou & Ding, 2017). Sorock et al. (1996) showed that motor vehicle crashes into road construction work zones are also a severe cause of accidents due to

general public activities using narrative text from insurance claims. Casteel and Peek-Asa (2000) showed the importance of managing workplace environment to manage possible robberies. Hinze et al. (2006) showed that the laceration hazards accounted for most frequency accidents and injure to fingers and hands. Templer (1995) claimed that falls from staircases, ladders and scaffolding are due to mis stepping, traffic, loose balance, tripping etc. Lortie and Rizzo (1999) developed a classification system to identify slip-trip-fall hazard events and loose balance hazards events separately. Lortie and Rizzo (1999) further claimed that loose balance hazards were underestimated and falling accidents were falsely identified. Thus, these studies showed the importance of actual hazards categorisation into their sources of hazards and the following sections discuss the identified hazards related to each source of hazards in this study.

3.5.1 Identified hazards in natural factors

It had been identified three main hazards related to natural factors based on phrases extracted while extracting the sources of hazards. Examples for each hazard are shown in Table 3.6 with the related final narrative.

Table 3.6 Examples for hazards in natural factors

Hazard	Example form Final Narrative	N-grams
Gust of wind (Peterson et al., 2013; Walsh et al., 2016)	On or about February 18, 2015, an employee suffered an amputation of part of his left pinky finger when an entry door slammed on it due to a strong <i>gust of wind</i> .	[('gust', 'wind')]
Wet weather (Walsh et al., 2016)	An employee installing siding fell from a ladder breaking his left leg below the knee. The ladder was on the deck and slipped due to misty, <i>wet weather</i> conditions.	[('wet', 'weather')]
Sudden heat wave (Peterson et al., 2013)	An employee showed signs of heat stress while operating a paver during a <i>heat wave</i> . The employee was hospitalized.	[('heat', 'wave')]

3.5.2 Identified hazards in organisational factors

Six main hazards related to organisational factors were identified based on phrases extracted while extracting the sources of hazards. Examples for each hazard is shown in Table 3.7 with the related final narrative.

Table 3.7 Examples for hazards in organisational factors

Hazard	Example from Final Narrative	N-grams
Unprotected equipment or areas (Bobick, 2004)	An employee fell 14 <i>feet through a hole</i> in the floor. The hole was unguarded at the time of the incident.	[('fell', 'through', 'hole')]
Work without proper inspection over PPE (Zhang et al., 2020)	An employee was helping to run wire cable onto the boom of a crawler crane when the loose part of his <i>safety harness</i> lanyard got caught in the rotating drum.	[('safeti', 'har')]
Venomous species inside the working environment (Robbins et al., 2008)	An employee was <i>bitten</i> by a rattlesnake.	[('bitten')]
Use of unskilled labour (Vitharana et al., 2015)	Employees were dismantling/removing heavy equipment at the Amazon facility. One of the employees got on a stand-up forklift to move some things while cleaning the area and got his leg caught between a concrete wall and the forklift. The employee was <i>not trained</i> to operate the forklift.	[('not', 'train')]
Contaminated site environment (X. Li et al., 2017)	An employee was hospitalized on 7/27/2015 for a <i>bacterial infection</i> in his calf that developed after contacting contaminated water that was in a pipe he was cutting.	[('bacteri', 'infect')]
Limited site visibility (Zhou & Ding, 2017)	An employee was looking at excess equipment in an area with <i>limited visibility</i> . He fell approximately 4 feet from the floor area	[('limit', 'visibl')]

3.5.3 Identified hazards in surrounding activity

Eight main hazards related to organisational factors were identified based on phrases extracted while extracting the sources of hazards. Examples for each hazard is shown in Table 3.8 with the related final narrative.

Table 3.8 Examples for hazards in surrounding activity

Hazard	Example from Final Narrative	Phrase
Struck by foreign object or person (Cheng et al., 2020)	An employee was <i>struck by</i> a portable aluminium extension ladder that fell from the side of the east wall.	[('struck', 'by')]
Falling of objects (Zhang et al., 2019)	An employee was walking when an <i>unknown object fell</i> from above and hit his head. He was hospitalized with head and neck pain and swelling.	('unknown', 'object', 'fell')
Fault of surrounding employee (Lee et al., 2020)	An employee was working on a closed I-64 exit ramp at IL Route 177. A vehicle being <i>driven by a site employee struck a nearby roller</i> , which then struck the employee, breaking both legs.	[('driv', 'by', 'site'),('struck', 'by', 'nearby')]
General public activities (Sorock et al., 1996)	A <i>motorcycle</i> entered the work zone and struck an employee who was cleaning up a pothole. The work zone had TCD.	['motorcycl']
Remote control work (Schiffbauer & Ganoë, 1999)	An employee was admitted to the hospital after being pinned between a <i>remote-control</i> dirt compactor and pylon.	('remot', 'control')
Robbery (Mullgn, 1997)	On 8/4/15, at approximately 5:20 a.m., two employees were <i>robbed</i> while stopped at a Fuel Depot gas station to refuel their work vehicle. The first employee was shot three times in the stomach and the second employee was beaten up, suffering a fractured jaw, busted lip, and bruises all over his body. Both employees were hospitalized.	['rob']

Blasting activity (Kecojevic & Radomsky, 2005)	An employee was hit by the <i>blast of a sand blaster</i> , which injected media into his leg.	[('blast', 'sand', 'blaster')]
Surrounding animal attack (Treves et al., 2011)	An employee was running a phase line for an electric customer when the resident's <i>dog chased</i> him. He tried to climb over a 4-foot fence to escape and his left pant leg and foot got caught on a post. He fell over the fence, breaking the tibia and fibula in his left leg.	[('dog', 'chase')]

3.5.4 Identified hazards in technological factors

Ten main hazards related to technological factors were identified based on phrases extracted while extracting the sources of hazards. Examples for each hazard is shown in Table 3.9 with the related final narrative.

Table 3.9 Examples for hazards in technological factors

Hazard	Example from Final Narrative	N-grams
Part of machinery or structures failed (Chinniah, 2015)	An employee was <i>struck in the head</i> by a metal frame when a <i>crane hook failed</i> .	[('struck', 'in', 'head'), ('cran', 'hook', 'fail')]
Unstable work area (Kim et al., 2016)	An employee sustained a fractured right leg when a <i>trench collapsed</i> .	[('trench', 'collaps')]
Electrical work (Dalziel, 1972)	An employee suffered <i>arc flash burns</i> while performing electrical work on an energized bus.	[('arc', 'flash'), ('flash', 'burn')]
Malfunctioning of machinery (Teizer & Cheng, 2015)	Two workers at the National Counter Terrorism Center were on a suspended scaffold when the <i>scaffold malfunctioned</i> causing the employees to fall.	[('scaffold', 'malfunc')]
Release of energy in extreme manner (Eckhoff, 2016)	An employee was cutting a barrel. The barrel <i>exploded</i> and the employee suffered several burns and broke his left leg.	[('explod')]

Fluid leak in machinery or pipe (Sweeney, 1988)	A propane-powered space heater was used inside the cab of an elevated crane during workplace operations. The space heater <i>leaked</i> , resulting in an <i>explosion</i> and fire.	[‘explos’, ‘leak’]
Entanglement (Townsend & Barker, 2014)	An employee was lowering a fire hose off of the side of a ship using a rope when the <i>rope became tangled</i> around his index finger, amputating it at the first knuckle.	[(‘rope’, ‘becam’, ‘tangl’)]
Hazardous chemical usage (Deacon & Smallwood, 2001; Helander, 1991)	An employee was hospitalized for <i>carbon monoxide poisoning</i> while working at a construction site.	[(‘carbon’, ‘monoxid’, ‘poison’)]
Lacer work (Hinze et al., 2006; Shendell et al., 2012)	An <i>employee lacerated</i> his right hand and was hospitalized.	[(‘employe’, ‘lacer’)]
Pressure release of machinery (Albert et al., 2014)	An employee was filling a tank with water and was struck by a <i>pressurized water blast</i> that knocked him to the ground. The employee hit his head on the pavement and sustained a concussion.	[(‘pressur’, ‘water’, ‘blast’)]

3.5.5 Identified hazards in worker factors

Fourteen main hazards related to worker factors were identified based on phrases extracted while extracting the sources of hazards. Examples for each hazard is shown in Table 3.10 with the related final narrative.

Table 3.10 Examples for hazards in worker factors

Hazard	Final Narrative	N-grams
Inaccurate foot placement (Templer, 1995)	An employee was stepping across a pile of dirt in the way of his work when he <i>misstepped</i> and fell, shattering his left ankle.	[‘misstep’]

Lack of skill (Notelaers et al., 2007)	An employee sustained battery acid burns to the face while <i>attempting</i> to start an air compressor on the back of a service vehicle by boosting the battery with jumper cables.	['attempt']
Mishandling (Khan et al., 2019)	An <i>employee dropped</i> a concrete form on his hand, causing an injury that required surgery.	[('employee', 'drop')]
Misbehave (Eriksson & Lind, 2016)	An <i>employee used the running</i> line that was attached to the hook of a crane to <i>pull himself up</i> . The left tip of his middle finger and half of his ring finger got caught between pinch points (the rope and pulley), resulting in an amputation.	[('employee', 'use', 'run'), ('pull', 'himself', 'up')]
Instability of the worker (Lortie & Rizzo, 1999)	An employee was on stairs applying insulation above his head when he <i>lost his balance</i> and fell, hitting a protruding pipe and breaking his left hip.	[('lost', 'hi', 'balanc')]
Health issue (Vitharana et al., 2015)	An employee was picking up trash at the site. At the end of the day, he started having cramps and <i>not feeling well</i> . He was hospitalized.	[('not', 'feel', 'well')]
Irresponsible work (Teizer et al., 2013)	An employee was dismantling a scaffold while <i>unaware</i> that he was tied off onto it. He fell from the scaffold, approximately 7 feet to the ground below.	['unawar']
Loss of attention (Hasanzadeh et al., 2017)	An employee was dispatched to a parts distributor to pick up parts. The employee was <i>standing by</i> the loading dock. While doing so, a truck backed up to the dock and pinned the employee against the dock.	[('stand', 'by')]
Irresponsible work of another employee	An employee was trying to show a coworker the controls on a skid steer to deactivate the interlock on the fork attachment. While the employee was pointing out the disengagement	[('cowork', 'immedi', 'press')]

(Stoilkovska et al., 2015)	button, the <i>coworker immediately pressed</i> it, causing the forks to lift and pin the injured employee between the fork mast and the cab. The employee suffered a fractured hip.	
Lost control over the task performing (Sacks et al., 2009)	An employee was dumping a portion of a load from a buggy when he <i>lost control</i> of the buggy and descended down a ramp. His right leg was pinned and broken between the equipment's gas tank and a wall.	[('lost', 'control')]
Inspecting work (Dzeng et al., 2016; Poh et al., 2018)	An employee fell through a skylight while <i>checking</i> for ice on the roof of the building.	['check']
Assisting an employee (Poh et al., 2018)	An employee was <i>assisting</i> a supervisor to clear jammed material from a machine. A 40-pound knife blade moved and amputated the employee's fingertip.	['assist']
Unseen hazards (Gheisari & Esmaili, 2016)	An employee was de-rigging the suspended scaffolding on the bridge. He did <i>not see</i> a cable and ran into the cable clips resulting in an abdominal hematoma.	[('not', 'se')]
Work place violence (Husk, 1992)	An employee was being <i>harassed by</i> a coworker. The coworker picked up a product and threw it at the employee, striking his right cheek. He required stitches.	[('harass', 'by')]

3.5.6 Utilised libraries and modules

Libraries are a definitive collection of scripts associated with python programming in order to simplify the process of programming while allowing the programmer to utilise the script without rewriting. At the same time, modules provide only a single functionality. Libraries imported for classifier training was presented in Table 3.11.

Table 3.11 Description of imported libraries

Libraries	Description
os	This is a module which used to work with operating system dependent functions.
numpy	numpy is imported to support for large multi-dimensional arrays and matrices along with a collection of mathematical functions to operate with these arrays.
pandas	Pandas is a software library used for data manipulation, data importing, and analysis.
re	re is a regular expression library which helps to identify string characters and to match strings with other strings or set of strings.
nltk	A natural language tool kit is open-source library denoted by nltk to work with human language processing steps.
sklearn	This is an open-source library which provides many supervised and unsupervised learning algorithms.

3.5.7 TFIDF (Term Frequency Inverted Term Frequency)

TFIDF is a statistical measure to illustrate the importance of a word in a document in a collection of documents. TF denotes the Term Frequency, and it defines the word is more important if it appears several times in the target document. IDF denotes the Inverted Term Frequency, and it defines the word is more vital if it appears a smaller number of times in the target document. TFIDF is utilised in the Bag of Word model to train the classifiers.

3.5.8 Bag of word model (BoW)

Bag of words is a text representation model while engaging with machine learning algorithms and it is easier to understand and implement for machine learning problems such as language modelling and document classification (Juanals & Minel, 2018). This model is presented in a table format, and columns represent the different words in all the documents while rows represent each document. Each of the column and row headings is assigned a numerical value and count the number of times each word appears in the document and place the count in the respective cell. Created sample BoW is shown in Figure 3.4.

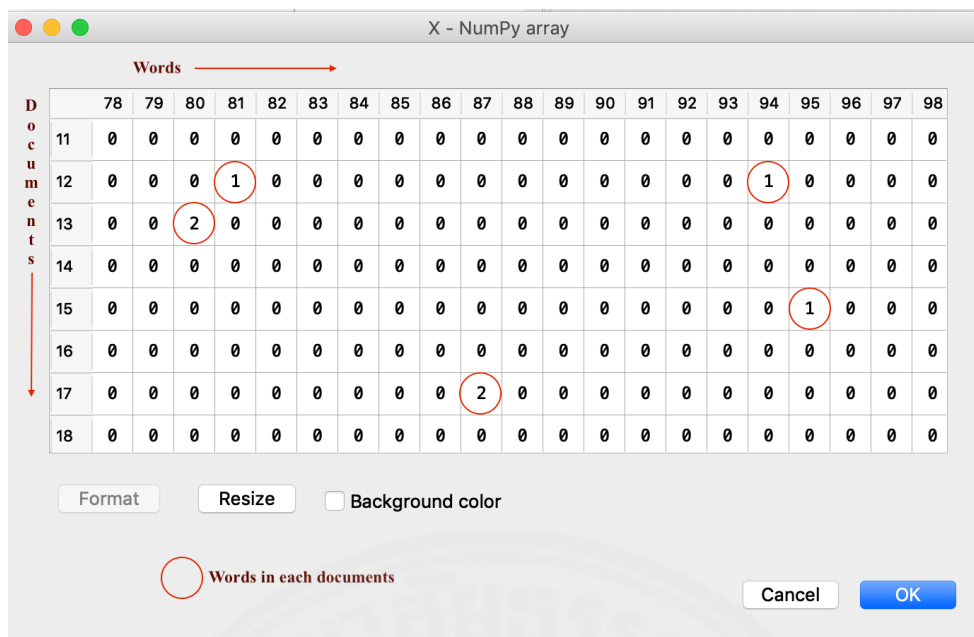


Figure 3.4 Sample of bag of word model

3.5.9 Split data into a training set and test set

Data set was split into training and test set to examine the accuracy of the results. In the data set, 70% of the data were used to train the classifiers with predefined outputs. Rest was used to predict the results and compared with the predefined outputs.

3.5.10 Fit trained data into a classifier

Classifiers discussed in Heading 3.4 were used for evaluating the best performing classifier. F1 score was used to measure the performance.

3.6 Frequency analysis

After all the sources of hazard were extracted from 8940 number of data records, descriptive analysis was carried out for extracted sources of hazards by using IBM SPSS statistical 23 software. Pareto charts were drawn for identified hazards to identify the hazards which have the highest contribution to cause an accident using 80-20 rule.

3.7 One-way ANOVA

The one-way analysis of variance (ANOVA) was used to determine the statistically significant relationships between sources of hazards and the number of accidents occurred in each season, year and month. First, statistical significance was evaluated for each season, year and month with the respective total number of accidents. Also, seasonal significance was

assessed with the respective number of the normalised total number of accidents by using construction spending. Then statistical significance was checked for each source of hazards to identify the seasonal, yearly and monthly behaviours of accidents. For all the cases, the null hypothesis was assumed as there is no statistically significant relationship and means are not significantly different from each other. The null hypothesis can be presented in Equation (3.7).

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k \quad (3.7)$$

Where μ = group mean and k = number of groups.

However, the one-way ANOVA is an omnibus test statistic which cannot specify the statistically significant groups. Thus, post hoc test was needed to be conducted. In addition to that, there are three main assumptions which are required for one-way ANOVA to be valid.

- a) The dependent variable which is being compared should be normally distributed.
- b) The data should be homogeneous, and populations variances should be equal.
- c) Data should be independent.

3.7.1 Validation of the data for normality

Normal distribution of the data was checked using the “explore” function in SPSS. Depending on whether the data is normally distributed or not, selection of ANOVA test method varied. If the data is typically distributed, post hoc test was conducted directly. Otherwise, the non-parametric Kruskal-Wallis H test was performed.

3.7.2 Validation of the data for homogeneity

Homogeneity of the variances was checked using homogeneity variances test in SPSS. If the data does not meet the homogeneity variances assumption, the Welch test was performed to identify the significance.

3.7.3 Non-parametric Kruskal-Wallis H test

Non-parametric Kruskal-Wallis test is a rank-based non-parametric test which is used to determine the statistical significance of two or more groups of independent variables. There are five assumptions to be satisfied with the data for this test to be valid. They are;

- a) The dependent variable should be measured in an interval or ratio level.
- b) Independent variable should consist of two or more categorical, independent groups.
- c) Data should be independent observations.

- d) Data distribution has the same shape.
- e) There needs to be homogeneity of variances.

However, this test also an omnibus test which cannot identify the statistically significant groups. Thus, this will also require a post hoc test to identify the statistically significant groups after determining the significance. Besides, the Welch test needed to be conducted if the homogeneity of variances assumption was violated.

3.7.4 Post hoc tests

Post hoc tests were conducted to identify statistically significant groups. According to the data, the sample sizes for every season, year and month differs from each other as the data consists of data from January 2015 to September 2019. Therefore, variances tend to differ from one another. Also, variances for sources of hazards factors tend to differ highly as the sample sizes are different. Thus, Games Howell procedure was selected as the post hoc test, which is generally the best performing test when similar variances are not assumed.

3.7.5 Calculation of effect size

SPSS does not routinely provide an effect size for one-way ANOVA. Thus, the effect size was calculated using the between-group effect (SS_M) and the total amount of variance in the data (SS_T). However, the measure of this effect size can slightly be biased as it is purely based on sums of squares from the sample. Therefore, this study calculates the omega squared (ω^2) using Equation (3.8). The df_M in the equation is degrees of freedom for the effect, and MS_R is the mean squared error.

$$\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} \quad (3.8)$$

CHAPTER 4

RESULTS AND DISCUSSION

This chapter conveys the results and discussions acquired through the adopted methodology in Chapter 3. The results obtained after implementing the rule-based classifier and its performance results on five sources of hazard were discussed in detail in Heading 4.1. Further, the performance of hazards identification classifiers was discussed in Heading 4.2. Starting from Heading 4.3, the chapter discusses the distribution of sources of hazards and hazards related to each factor with Pareto analysis.

In the latter part of the chapter, a secondary analysis was carried out to identify the seasonal and monthly behaviours of the accidents while normalizing the data using construction spending. Lastly, the results of one-way ANOVA and post hoc test were discussed.

4.1 Performance evaluation of developed rule-based classifier

The summary of the model performance was presented in Table 4.1. Seven iterations were required to achieve a threshold of 95%. However, the fourth iteration shows that it has achieved the threshold. Nevertheless, the average weighted F1 score testing for the fifth iteration was reduced as the F1 score of surrounding activity of the fifth iteration went significantly below the threshold of 95%. Hence, the sixth and seventh iterations were performed.

Table 4.1 Summary of model performance at each iteration

Iteration	Avg. Weighted F1 Score	F1 Score of NF	F1 Score of OF	F1 Score of SA	F1 Score of TF	F1 Score of WF	F1 Score of Null
1	0.85	0.59	0.92	0.88	0.89	0.85	0.84
2	0.93	0.75	0.97	0.96	0.95	0.94	0.91
3	0.94	0.86	0.98	0.95	0.96	0.96	0.92
4	0.97	0.97	0.99	0.96	0.99	0.96	0.94
5	0.94	0.97	0.98	0.89	0.97	0.93	0.92
6	0.98	0.97	0.99	0.96	0.99	0.99	0.98
7	0.98	1	0.98	0.95	1	0.99	0.99

4.1.1 Comparison of rule-based model performance with statistical classifiers

The accuracy obtained through the rule-based model was compared with the existing statistical classifiers found in the reviewed literature. Table 4.2 shows the performance of each tested statistical classifiers and their performance. SVM performs better than any other classifier and yet only achieves 81% accuracy. The least accuracy of 28% was achieved by Naïve Bayesian classifier. Thus, this proves the requirement of the rule-based classifier for extraction of sources of hazards from the final narratives.

Table 4.2 Performance of statistical classifiers

Classifier	F1 score
Rule- based classifier	0.95
Support Vector Machine	0.81
Random Forest classifier	0.71
K-nearest Neighbours	0.53
Kernel SVM	0.47
Naïve Bayesian classifier	0.28

4.2 Performance evaluation of the classifiers trained to identify hazards

According to Table 4.3, Random Forest classifier has been identified as the best classifier for extraction of hazard factors from each source of hazard factor. Kernel SVM records the least performance as the data was not nonlinear. Natural factor hazard achieved the same accuracy due to lack of data for training and testing the classifier.

Table 4.3 Performance of trained classifiers

Classifier	F1 score				
	Worker factor hazards	Technological factor hazards	Surrounding activity hazards	Organisational factor hazards	Natural factor hazards
SVM	0.942	0.940	0.989	0.982	0.80
NB	0.622	0.737	0.975	0.982	0.80
<i>RF</i>	<i>0.946</i>	<i>0.942</i>	<i>0.984</i>	<i>0.982</i>	<i>0.80</i>
kNN	0.908	0.900	0.918	0.965	0.80
Kernel SVM	0.658	0.475	0.888	0.859	0.80

4.3 Distribution of sources of hazards

Figure 4.1 below illustrates proportions the distribution of five types of sources of hazard extracted through developed rule-based classification tool and Table 4.4 showed the tabulised data. Overall, in 8940 data, worker factor was the most significant, which accounted for more than one-third of the sources of hazard. Least contribution to the accident cause becomes natural factors by having 0.6% contribution. Technological factors compromised by 22.4%. Surrounding activity and organisational factor represent an accident contribution of 7.3% and 2.1% respectively. However, 29.8% of the data is categorised under null by the tool as they do not mention the cause for the accident.

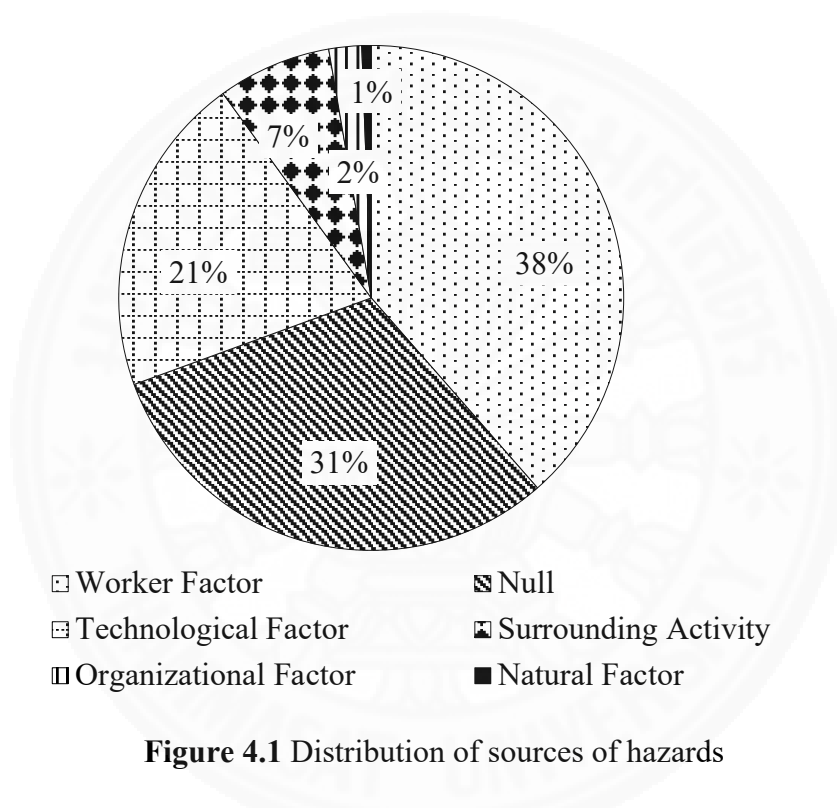


Figure 4.1 Distribution of sources of hazards

Table 4.4 Tabulated data for sources of hazards distribution

Source of Hazard	Frequency	Percent (%)
Worker Factor	3448	38.6
Null	2761	30.9
Technological Factor	1838	20.6
Surrounding Activity	655	7.3
Organizational Factor	188	2.1
Natural Factor	50	0.6
Total	8940	100

4.3.1 Worker factor hazards

The bar chart in Figure 4.2 the percentile representation of worker factor-related hazards. The study findings show that 29.9% of the occupational accidents due to worker factors were due to inaccurate foot placement, including mis stepping, slipping and tripping. Lack of skill was accounted for 21.6% of worker factor-related hazards. Data obtained through the final narrative shows that workers lack skill in operating machinery, performing electrical work, performing underground work etc. The mishandling of the equipment, objects or part of the structures and contributed for 13.7% of the worker factor accidents. The fourth place was taken by health issues including dehydration, fainting, dizziness, cramping, loss consciousness, feeling sick or light-headed, experiencing pain, pass out etc. and accounted for 8.2%. Misbehaviour of the workers caused for 7.4% of the worker factor, and it includes attempting to do untrained work, fails to remember, walking in hazardous areas without permission and violations such as drinking, smoking and not following safety rules. Instability of the worker caused 7% of the accidents. The irresponsible work, loss of attention, irresponsible work of another employee, lost control over the task performing, inspecting work, assisting an employee, unseen by the worker and workplace violence contributed for 4.6%, 4.1%, 1.4%, 0.9%, 0.6%, 0.5%, 0.1% and 0.1% respectively. The Pareto analysis showed that 80% of the worker factor-related accidents were caused by inaccurate foot placement, lack of skill, mishandling, misbehaving and instability of the worker. Therefore, industry should mainly focus on addressing these hazards to mitigate accidents which highly likely to occur due to worker factors.

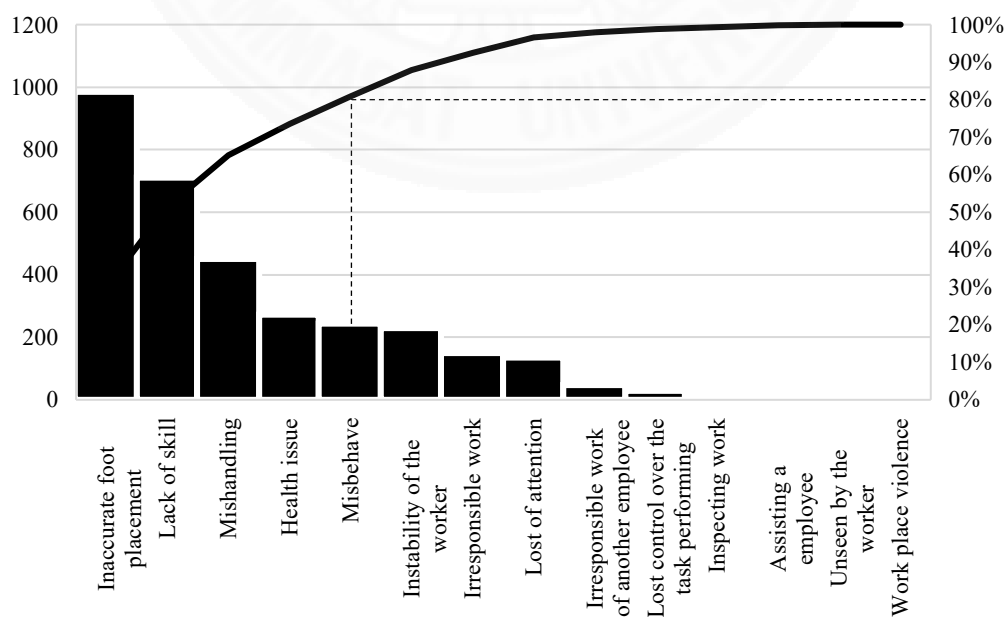


Figure 4.2 Worker factor hazards

4.3.2 Technological factor hazards

Figure 4.3 shows the technological factors related to everyday hazards. The Pareto analysis showed that 80% of the technological factor-based accidents were due to hazards including parts of machinery or equipment failure (28.7%), unstable work area (21.8%) and electrical work (19.8%). Malfunctioning of machinery accounted for 19.2% of the accidents while the release of energy in an extreme manner, the fluid leak in machinery or pipe, entanglement, hazardous chemical usage, lacer work and pressure release of machinery caused for 1.7%, 1.6%, 1.5%, 1.2%, 0.6% and 0.3% respectively.

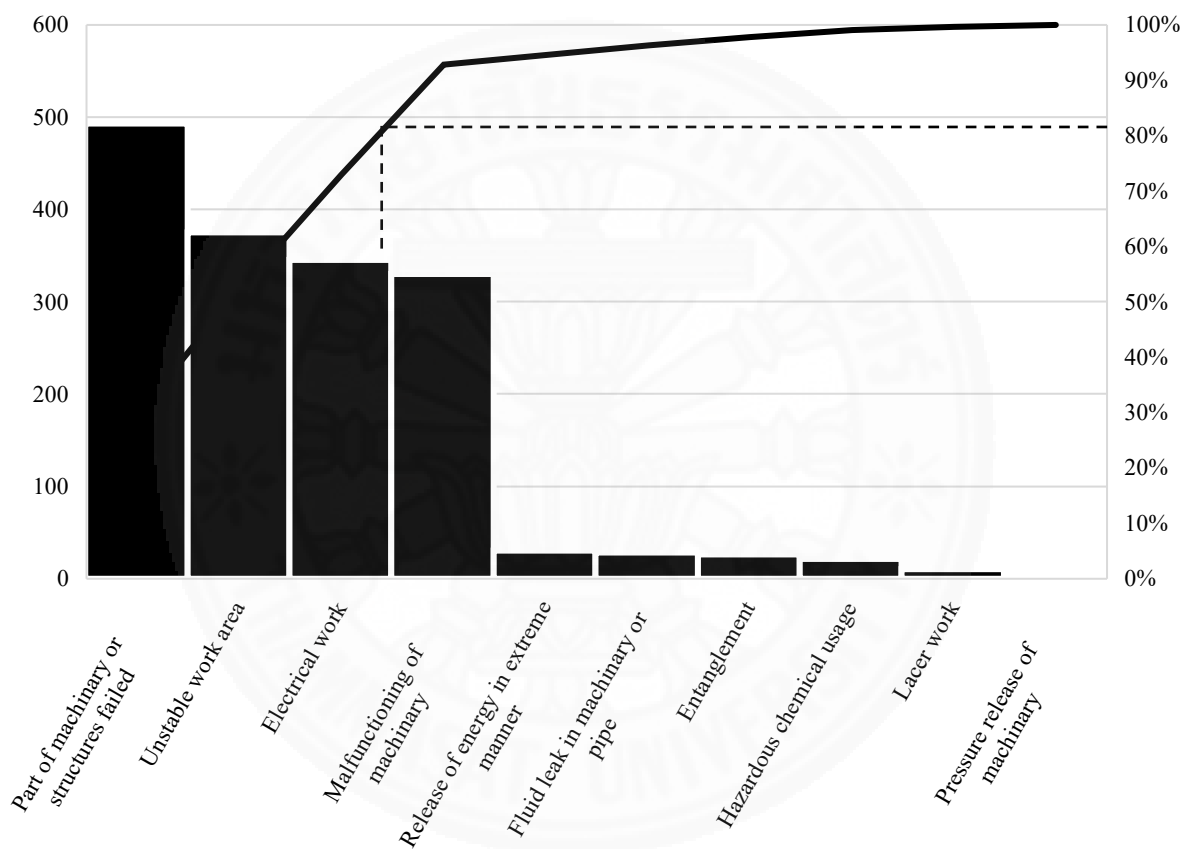


Figure 4.3 Technological factor hazards

4.3.3 Surrounding activity related hazards

The chart in Figure 4.4 showed the identified surrounding activity based on occupational hazards. According to the Pareto analysis, 80% of the surrounding activity-based accidents were due to struck-by hazards (47.8%) and falling of objects (35.5%). Other hazards such as fault of surrounding employee, general public activities, remote control work, robbery, blasting activity and the surrounding animal attack were taken 8.3%, 5.7%, 0.8%, 0.3%, 0.3% and 0.5% respectively.

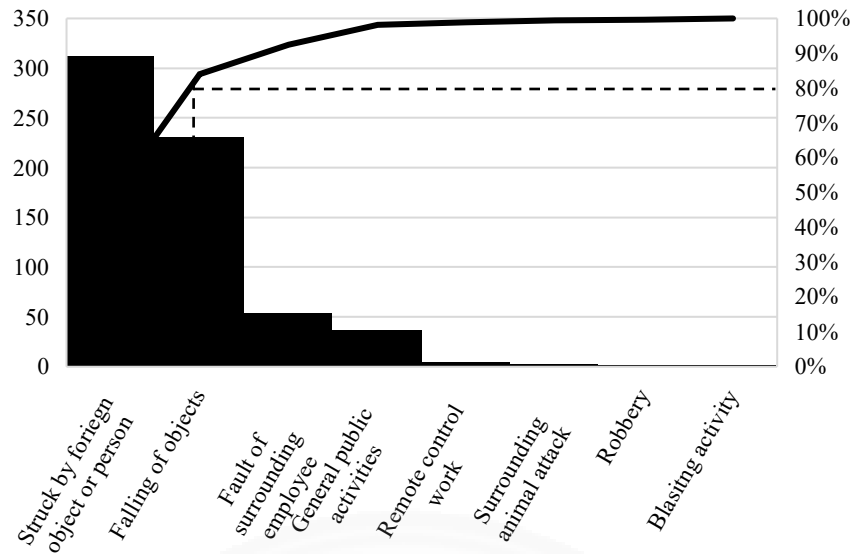


Figure 4.4 Surrounding activity-related hazards

4.3.4 Organisational factor hazards

The bar chart in Figure 4.5 showed the hazards related to organisational factors. Safety management issues such as unprotected equipment or areas, work without proper inspection over PPE, venomous species inside the working environment, use of unskilled labour, contaminated site environment and limited site visibility caused for 55.3%, 32.4%, 8.5%, 2.1%, 1.1% and 0.5% respectively. 80% of these accidents were due to unprotected equipment or areas and works without proper inspection over PPE. Thus, it is clear that the safety management issues play a significant role in occupational accidents other than the project conditions, management decisions and controlling which defines an organisational factor.

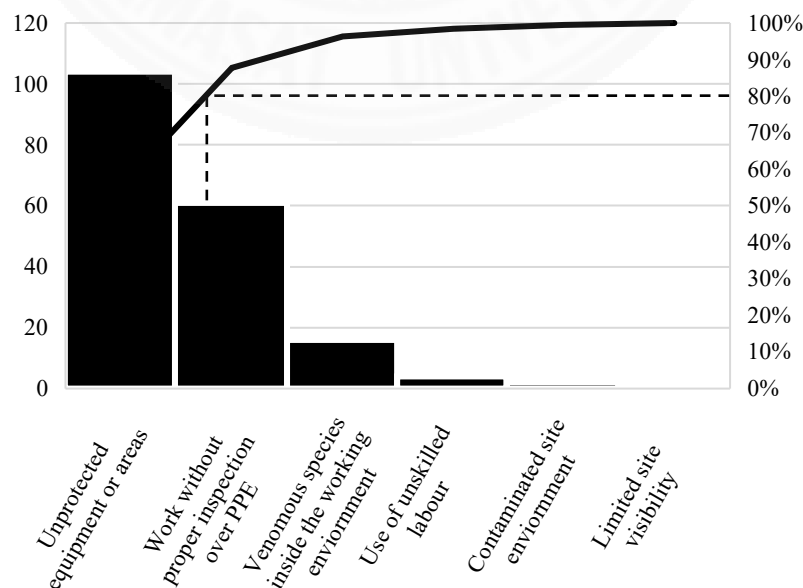


Figure 4.5 Organisational factor hazards

4.3.5 Natural factor hazards

The bar chart in Figure 4.6 illustrates three distinct hazards which caused the natural factor accidents. A gust of wind is the most influencing attribute in the open data set and shows 94% contribution. Sudden heatwaves show the least contribution of 2% while wet weather influences the natural factor by 4%. Less number of hazards were exit in natural factors as it is the least source of hazard.

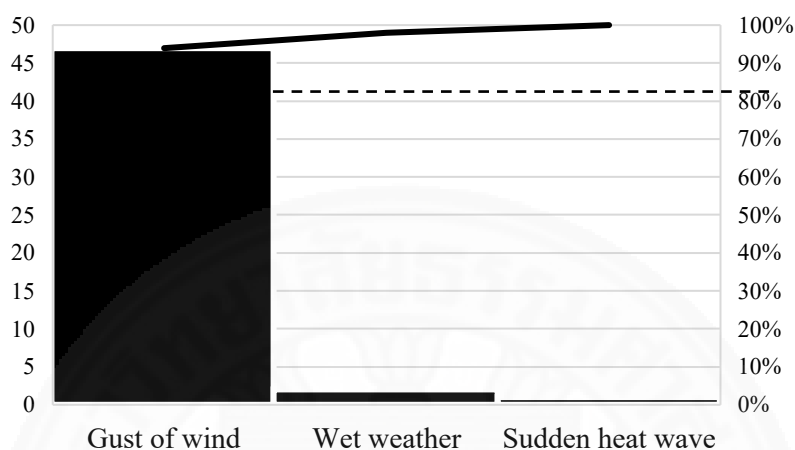


Figure 4.6 Natural factor hazards

4.3.6 Comparison of extracted results with existing literature

There had been found many studies which illustrate the construction hazards. These are combined results of various acts of the human. Therefore, the existing literature discussed more on the human factor involvement towards occupational accidents. The Health and Safety Executive defined the human factor as “Human factors refers to environmental, organisational and job factors and human and individual characteristics which influence behaviour at work in a way which can affect health and safety” (Kerr et al., 2009). The human factors also concern the interaction between people, their characteristics, abilities, organisation, management and technology (Woods & Dekker, 2000). Therefore, in this study, the human factors were divided into worker factors, technological factors, organisational factors and surrounding activity.

Further, existing studies focused on identifying hazards contributing to occupational accidents. For instance, Cheng et al. (2020) found that traffic, the collapse of an object, falls, caught in/between objects, struck by moving objects, others exposure to a chemical substance, fires and explosion, electrocution, struck by a falling object, and exposure to extreme temperatures are the severe cause of accidents. Zhang et al. (2019) found exposure to extreme temperature, exposure to chemical substances, struck by falling objects, fires and explosion, traffic, caught in between objects, electrocution, struck by moving objects, collapse of objects

and falls are the more significant cause of accidents. Zhang et al. (2020) categorised hazards to three primary levels: critical causes including delayed hazards elimination, inadequate safety inspection, important causes including inadequate execution of construction plan, inadequate subcontractor management, and general causes including inadequate use of PPE, incomplete construction plan.

However, in this study, root causes for the hazards were initially identified and then related actual hazards were identified for each source of hazard factor. This study showed that worker factor hazards are a combination of inaccurate foot placement, lack of skill, mishandling, misbehaviours, instability of the workers, health issues, heat stress, irresponsible work, workplace violence, and loss of attention. Accidents due to inaccurate foot place are inclusive of slipping, tripping and mis stepping. However, some studies claimed that slipping, tripping, falling, loose balance and lost attention are monocausal classification (Axelsson & Carter, 1995; Manning, 1988). However, Lortie and Rizzo (1999) further claimed that loose balance hazards were underestimated and falling accidents were falsely identified. Thus, this study classified them separately and hazards contributions were identified. Therefore, a possible management of these accidents are recommended since inaccurate foot placement contribute most to the worker factor-based accidents.

The technological factor hazards are a combination of structures/machinery failures, hazardous work area, electrical work, malfunctioning of machinery, the release of energy, fluid leaks, entanglement hazards, hazardous chemical usage, and lacer work. The strict adherences to the safety measures are recommended to prevent accidents due to hazardous work area, hazardous chemical usage, electrical work and lacer work.

Also, the organisational factor accidents are related to unprotected equipment/areas, lack of inspection over PPE, venomous species inside the working environment, unskilled labour, contaminated site environment, and limited site visibility. These accidents are directly the outcomes of safety management. Thus, the management crew should implement proper guidelines to alert the workers on unprotected areas and eliminate accidents due to these hazards. The species inside the working environment are mostly the mosquitoes and other bugs. Hence, proper water resources and site environment management must be a key consideration.

The surrounding activity-based hazards are a combination of struck by foreign object/person, falling of objects from upper elevations, faults of surrounding co-workers, general public activities, remote control work, surrounding animal attack, robberies, and blasting activities. The struck by hazards and falling objects hazards are very common types of hazards which has been discussed in several existing literature. However, proper prevention

methods are required for these hazards. Also, traffic hazards due to general public should be address through proper implementation of traffic control measures. Casteel and Peek-Asa (2000) illustrate the importance of implementing crime prevention through environmental design to reduce the robberies inside workplace. Thus, such methods should be utilized in the workplace to minimise the robbery hazards.

Lastly, the natural factors accidents are due to gust of wind, wet weather, and sudden heat waves. These hazards show the need of investigating severe weather hazards in construction industry and the requirement for the implementation of various weather condition management strategies for safety management.

Thus, this study categorised the hazards into their leading causes for proper visualization rather than expressing hazard in general. The above-discussed hazards categorised by Zhang et al. (2020) showed that those hazards are the outcomes of safety management issues, and in this study, those hazards have been categorised under organizational factor. Further, hazards identified by Zhang et al. (2019) and Cheng et al. (2020) are a combination of worker factors, technological factors and surrounding activity. Identifying the root cause of the accidents helps the industry to address the source group directly for safety management. Therefore, this study initially attempts to highlight the source of the hazard. Hence, the industry could address the root cause, and industry could gain insights for better safety management.

4.3.7 Monthly distribution of sources of hazards

It had been found that there is a significant variation in accident causation depending on the month. Thus, a secondary analysis was performed to identify the monthly variations of accident behaviour. According to Figure 4.7, it was clear that in each year, June and July had caused a severe rate of accidents. December and January caused a lesser number of accidents compared to other months. Thus, it seems that seasonal behaviour had been influenced by the rate of accidents throughout the year. Hence, seasonal behaviour of the accidents was analysed in the following sections.

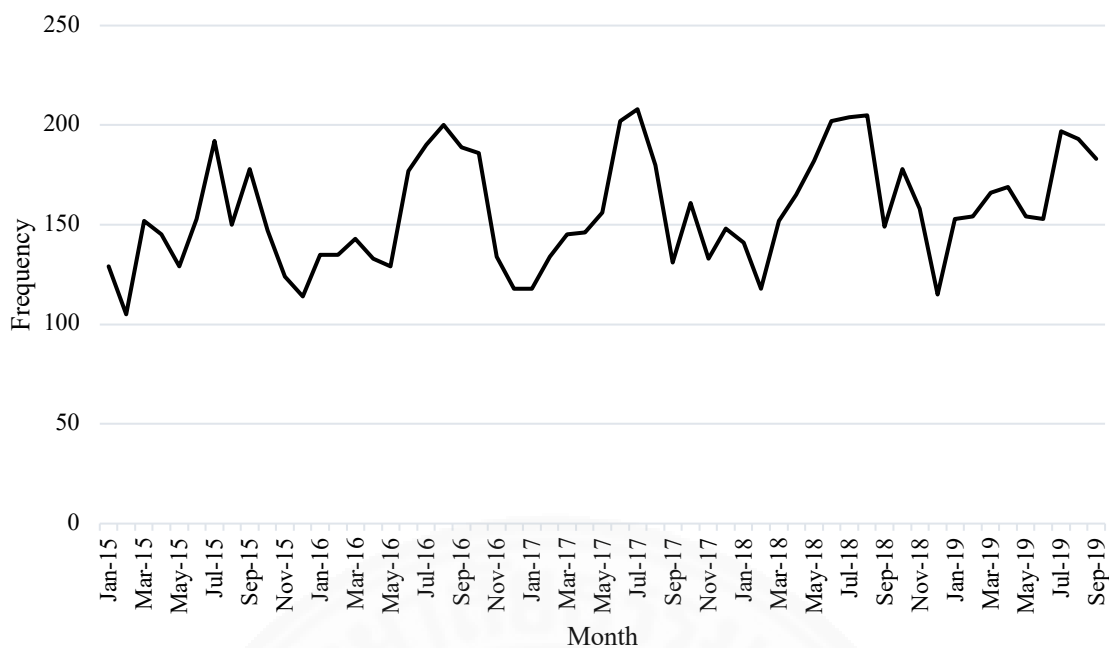


Figure 4.7 Monthly distribution of accidents

4.3.8 Seasonal distribution of sources of hazards

According to Heigl (2018b), Autumn and Winter were considered as dangerous seasons for construction activities due to the cold and slippery and wet ground. Spring and Summer also subjected to the issues since moisture and rain continue in Spring, and heat affects severely in Summer (Heigl, 2018b). Besides, Tschetter and Lukasiewicz (1983) state that outdoor construction activities are prone to decline in Winter compared to other seasons. Also, there is 80% of the constructional labour force decline in Winter (Tschetter & Lukasiewicz, 1983). Since the USA is subjected to higher seasonal variations as above, seasonal change is highly likely to affect the accident rate and sources of hazard distribution.

Seasonal variation of each source of hazard factor is shown in Figure 4.8. It clearly illustrates that worker factor accidents are more severe in mid of Summer to the beginning of Autumn compared to Winter and Spring. Winter and Spring have more severe accidents due to natural factors. Technological factors also tend to have a rise in the Summer season. Since the technological factors were defined, including machinery and equipment breakdowns, technical faults and errors while utilizing them, the rise in temperature in Summer can affect to increase the technological factor accidents in Summer.

However, a higher number of construction accidents does not clearly illustrate that the particular month or season is highly prone to accidents. Also, it is unlikely to have equivalent functioning of construction activities in every state throughout the year. Public sectors tend to

decline investments in the Winter months (Industry, 1979). Thus, the behaviour of accidents shown in the Figure 4.8 implies that: (1) Construction activeness might cause for a higher number of accidents in Summer; and (2) Temperature rise could cause for higher accident rate in Summer. Therefore, these observations were further investigated and their findings presented in the subsequent sections.

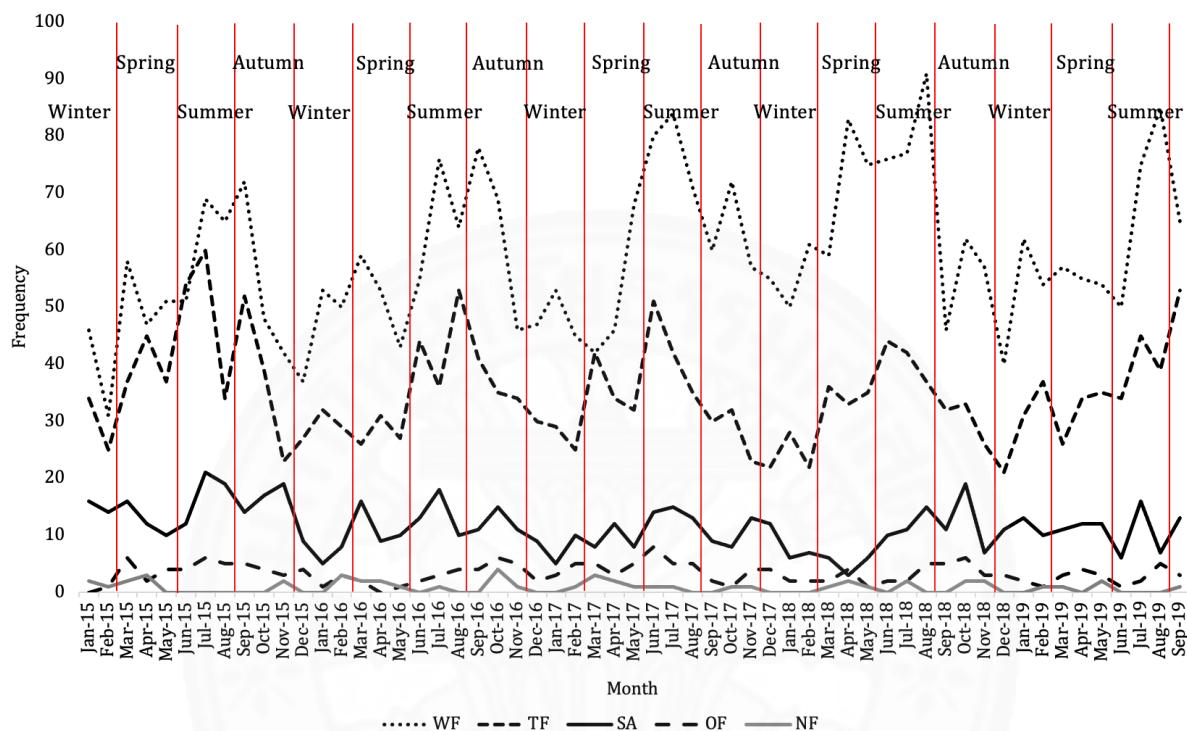


Figure 4.8 Seasonal distribution of sources of hazards

4.4 Normalisation of accident data

Seasonal distribution of the construction accident data showed that there is a higher possibility of accidents in Summer months. Since, one of the possibilities for this behaviour was construction activeness, its' effect needed to be removed to illustrate the temperature effects on accidents clearly. Thus, the effect of construction activeness was removed using construction spending data.

Construction spending data available on Census Bureau (*Construction Spending Survey*, 2020), the United States was utilised to normalised the total accidents in each state. Data was downloaded from January 2015 to September 2019 since the accidents data records were extracted for this period. Data were recorded in millions of dollars and normalizing the total number of accidents data using the original values (January 2015: 73,805 million dollars) were not reflective of the actual accident data. For instance, in January 2015, there had been recorded

129 number of accidents for 73,805 million dollars in construction spending. The normalisation using the 73,805 million dollars as the denominator would result in 0.00175 construction accidents which is not reflective of the accidents. Therefore, the denominator was selected as 50,000, which is also representative of the least spending. Then the construction spending of January 2015 was 1.48 in 50 billion dollars, and there were 87.4 accidents per 50 billion dollars.

Afterwards, accidents were presented as the number of accidents per 50 billion dollars. Total construction spending data were available depending on the state, month and year for state and local construction and private non-residential construction excluding power, communication and railroad sectors. This includes safety, non-railroad transportation, highway and street construction, waste disposal and sewerage construction, water supply, conservation and development.

Distribution of construction spending in 50 billion dollars over the above period is presented in Figure 4.9, and it clearly shows that Summer months tend to have higher spending on construction. Winter months tend to have lower construction spending.

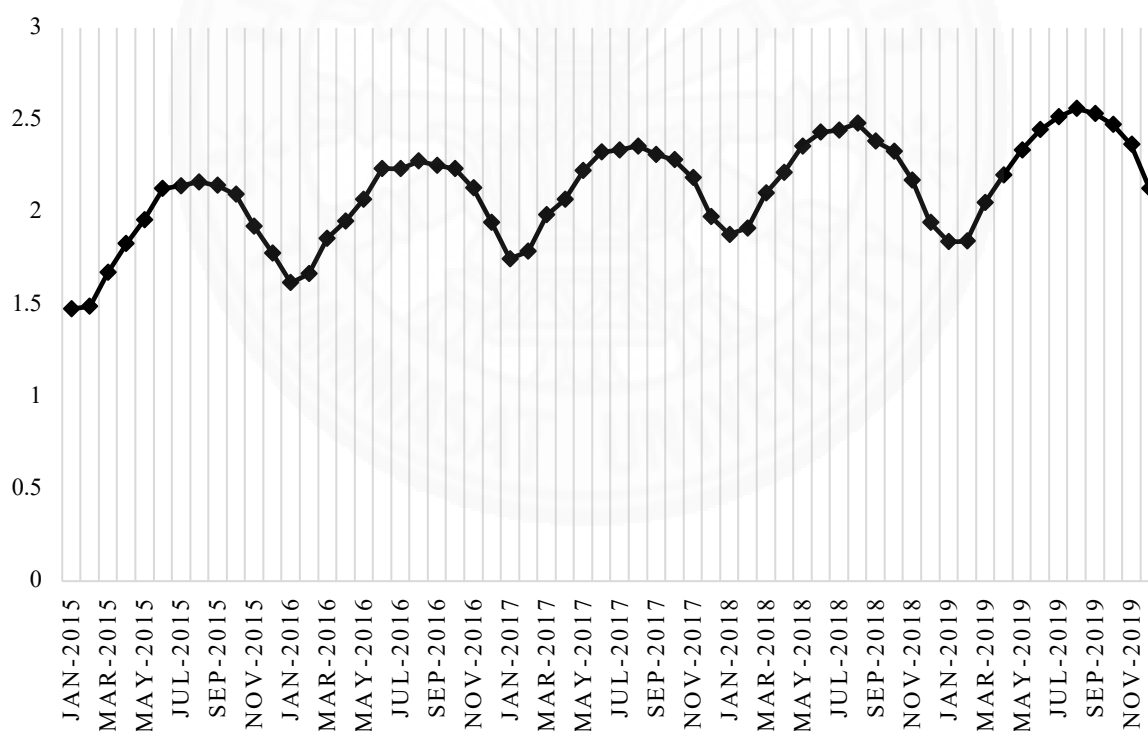


Figure 4.9 Construction spending data distribution

Figure 4.10 shows the accident distribution before and after normalisation. Accidents per 50 billion dollars show a notable trend as in total accidents.

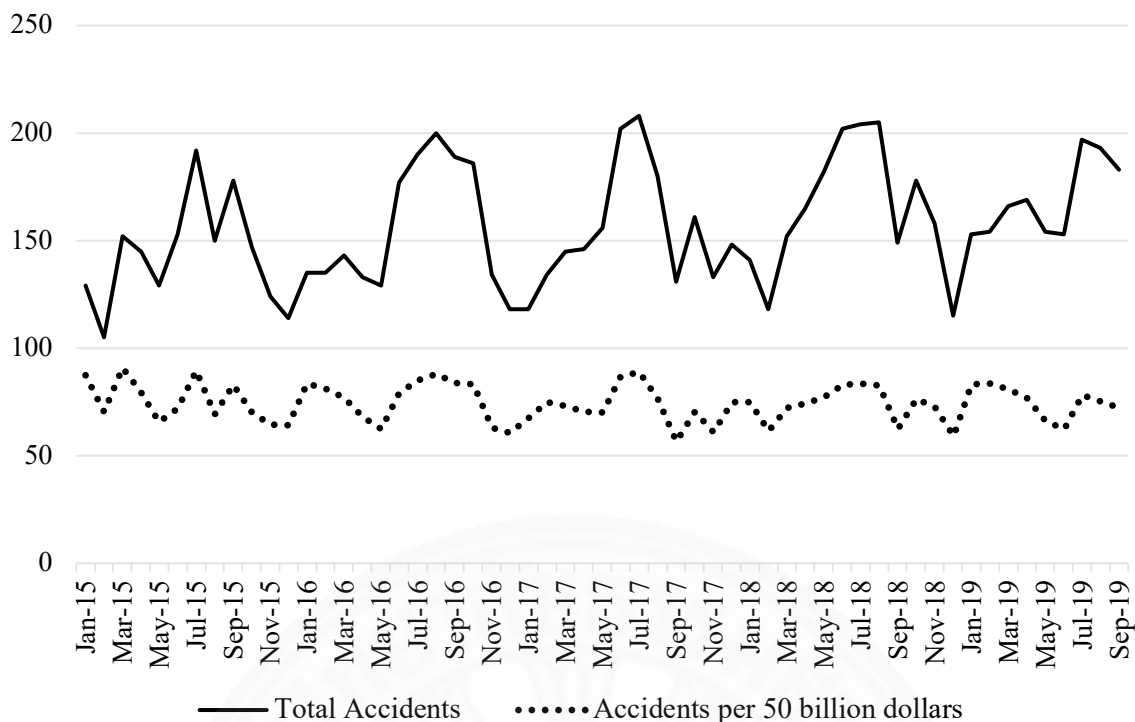


Figure 4.10 Monthly accident distribution

4.5 Literature study on seasonal variations of occupational accidents

Since the construction worker has been exposed to many kinds of occupational accidents, many studies have discussed seasonal variations on occupational accidents based on different countries. Hinze et al. (2005) stated by using OSHA data from 1990 to 2000 that struck by accident occurrences were higher in Summer months and October. Ling et al. (2009) claimed that the beginning of the rainy season caused the rise of fatalities in Singapore due to the rush to complete work. López et al. (2011) analysed the occupational accidents during lunchtime in Spain, and the results of Chi-square analysis showed that seasons were having a significant impact on fatal accidents around lunch break. Also, Arquillos et al. (2012) differentiate the climatic zones in Spain using seasons and found that accident severity is highly influenced by the seasonal effects. Liao (2012) illustrate using Bayesian classification that in Taiwan fall accidents of workers in age 21-40 and 41-60 are higher in Summer and fall-related accidents are higher in Winter due to wet weather.

Although that many research studies found a higher accident rate in Summer, research study conducted in Brazil by Brito et al. (2019) and in New Zealand by Robb and Barnes (2018) found that beginning and the end of the year would cause for higher accident rates. In contrast, Dumrak et al. (2013) discovered using Chi-square analysis that seasonal changes were not

strongly associated ($\chi^2 = 23.386$; $df = 15$; $p = 0.076$) with occupational accidents in South Australia.

However, it seems that recent studies which focus on seasonal behaviour of occupational accidents are rare and very few studies focused on the seasonal change of accidents in the USA. Thus, this study adopts one-way ANOVA to examine the significance of seasonal and monthly behaviour of accidents based on both normalised and non-normalised total accidents data. Moreover, this study attempts to observe the variations of each extracted hazard factor depending on the seasons.

4.6 One-way ANOVA

One-way ANOVA was performed for each season, year, month and for all sources of hazards to identify the significance of the hazard factors depending on the seasonal and monthly variations. Each factor was tested for assumptions, and required omnibus test was conducted. Results were shown in below sections.

4.6.1 Assessing the assumptions

4.6.1.1 Test of normality distribution

Normality distribution of data was checked for each season, year, month and each source of the hazard before performing one-way ANOVA and results are presented in below sections.

a) Assessing the normality distribution of seasonal accident data

According to Table 4.5, it can be seen that the significance value of the Shapiro-Wilk test for Summer is less than 0.05. Thus, Summer data were deviated from the normal distribution and Non-parametric Kruskal Wallis Test needed to be conducted.

Table 4.5 Test of normality for seasonal data

	Season	Shapiro-Wilk		
		Statistic	df	Sig.
Total	Winter	.938	14	.394
	Spring	.961	15	.718
	Summer	.828	15	.009
	Autumn	.911	13	.189

Normal Q-Q plot is shown in Figure 4.11 for each season. According to Figure 4.11 (a), (b) and (d), the data are distributed closely to the diagonal line. Thus, the data are normally distributed. Nevertheless in Figure 4.11 (c), data points are distributed away from the diagonal line. Thus, data of the Summer is not normally distributed.

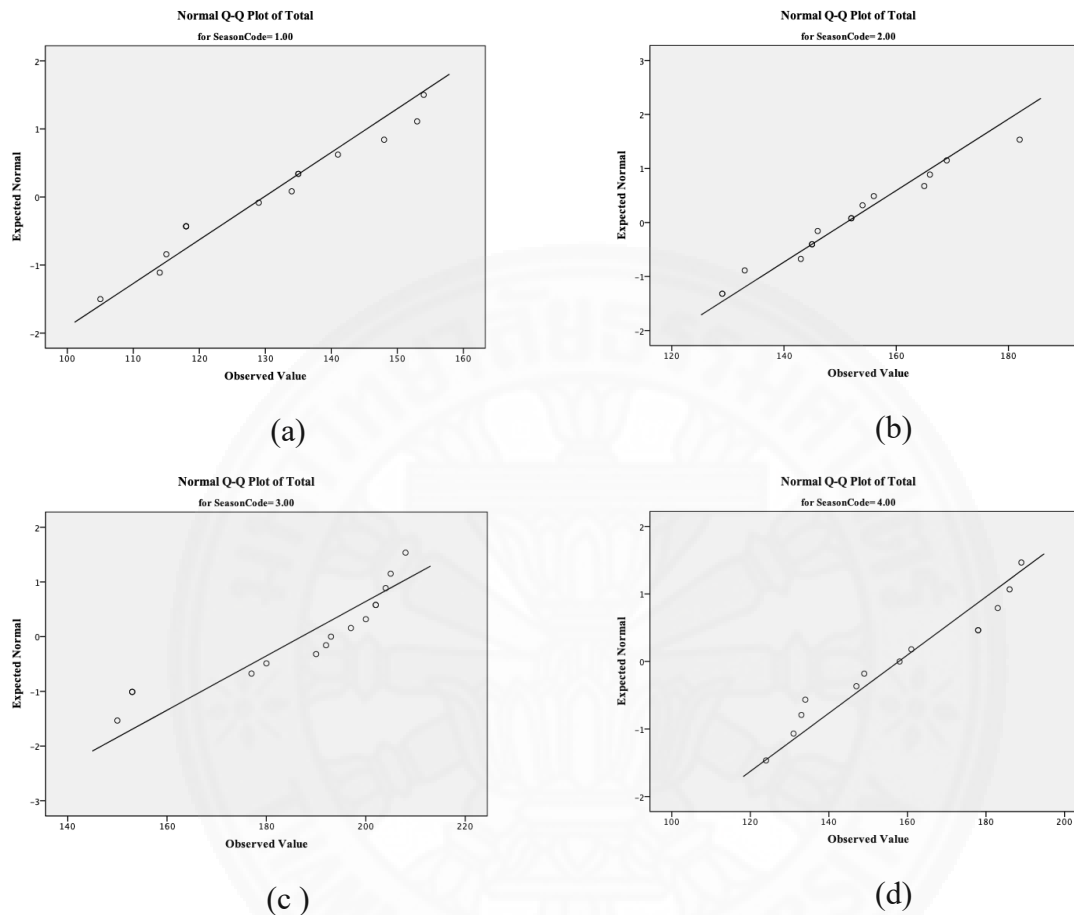


Figure 4.11 Normal Q-Q plot for each season (a) Winter (b) Spring (c) Summer (d) Autumn

b) Assessing the normality distribution of yearly accident data

According to Table 4.6, the significance value of the Shapiro-Wilk test for each year is higher than 0.05, except for 2016. Thus, 2016 data were deviated from the normal distribution and Non-parametric Kruskal Wallis Test needed to be conducted.

Table 4.6 Test of normality for yearly data

	Year	Shapiro-Wilk		
		Statistic	df	Sig.
Total	2015	.958	12	.750
	2016	.843	12	.030

	2017	.908	12	.199
	2018	.936	12	.442
	2019	.839	9	.056

c) Assessing the normality distribution of monthly accident data

According to Table 4.7, it can be seen that the significance value of the Shapiro-Wilk test for each month is higher than 0.05, except for December. Thus, December data were deviated from the normal distribution and Non-parametric Kruskal Wallis Test needed to be conducted.

Table 4.7 Test of normality for monthly data

	Month	Shapiro-Wilk		
		Statistic	df	Sig.
Total	January	.999	5	1.000
	February	.972	5	.891
	March	.886	5	.338
	April	.915	5	.501
	May	.890	5	.357
	June	.821	5	.118
	July	.932	5	.613
	August	.884	5	.328
	September	.883	5	.323
	October	.961	4	.783
	November	.872	4	.308
	December	.718	4	.019

d) Assessing the normality distribution of sources of hazard in each season

According to Table 4.8, it can be seen that the significance value of the Shapiro-Wilk test for each source of hazard in each season is higher than 0.05 except for natural factors. Therefore, Games Howell post hoc test was conducted for all sources of hazards directly except for natural factors. Natural factor data were deviated from the normal distribution and Non-parametric Kruskal Wallis Test needed to be conducted.

Table 4.8 Test of normality for sources of hazards in each season

Sources of Hazard	Season	Shapiro - Wilk		
		Statistic	df	Sig.
Worker Factor	Winter	.966	14	.818
	Spring	.920	15	.192
	Summer	.952	15	.549
	Autumn	.949	13	.588
Technological Factor	Winter	.970	14	.874
	Spring	.947	15	.471
	Summer	.922	15	.204
	Autumn	.906	13	.163
Surrounding Activity	Winter	.968	14	.842
	Spring	.959	15	.679
	Summer	.984	15	.990
	Autumn	.950	13	.597
Organisational Factor	Winter	.958	14	.694
	Spring	.971	15	.874
	Summer	.925	15	.232
	Autumn	.948	13	.536
Natural Factor	Winter	.681	14	.000
	Spring	.896	15	.082
	Summer	.606	15	.000
	Autumn	.872	13	.015

e) Assessing the normality distribution of sources of hazard in each month

According to Table 4.9, It can be identified that the normal distribution is violated in March and April for worker factors, April and August for technological factors, surrounding activities in April, organisational factors in August and natural factors in January, June, September and November. Further, natural factors are constant in August and December. Thus, they have been omitted when conducting the normality test. Non-parametric Kruskal Wallis Test was then conducted to identify the significance in sources of hazards depending on the month of the accident.

Table 4.9 Test of normality for sources of hazards in each month

Source of Hazard	Month	Shapiro-Wilk		
		Statistic	df	Sig.
Worker Factor	January	.929	5	.593
	February	.965	5	.840
	March	.649	5	.003
	April	.767	5	.043
	May	.948	5	.725
	June	.810	5	.097
	July	.952	5	.749
	August	.875	5	.286
	September	.972	5	.887
	October	.910	4	.481
	November	.832	4	.174
	December	.950	4	.719
Technological Factor	January	.974	5	.899
	February	.884	5	.330
	March	.869	5	.263
	April	.752	5	.031
	May	.893	5	.375
	June	.938	5	.655
	July	.858	5	.222
	August	.773	5	.047
	September	.870	5	.265
	October	.920	4	.538
	November	.802	4	.105
	December	.912	4	.492
Surrounding Activity	January	.806	5	.090
	February	.916	5	.502
	March	.884	5	.329
	April	.735	5	.021
	May	.961	5	.814
	June	.912	5	.482

	July	.991	5	.984
	August	.994	5	.992
	September	.953	5	.758
	October	.906	4	.462
	November	.982	4	.911
	December	.849	4	.224
Organisational Factor	January	.961	5	.814
	February	.881	5	.314
	March	.867	5	.254
	April	.881	5	.314
	May	.894	5	.377
	June	.845	5	.180
	July	.867	5	.254
	August	.552	5	.000
	September	.902	5	.421
	October	.848	4	.220
	November	.863	4	.272
	December	.863	4	.272
Natural Factor	January	.552	5	.000
	February	.828	5	.135
	March	.881	5	.314
	April	.828	5	.135
	May	.883	5	.325
	June	.552	5	.000
	July	.881	5	.314
	September	.552	5	.000
	October	.971	4	.850
	November	.729	4	.024

4.6.1.2 Test of homogeneity variances

Test of homogeneity variances of data was checked for each season, year, month and each sources of hazard before performing one-way ANOVA and results are presented in below sections.

a) Assessment of homogeneity variances assumption on seasonal accident data

In this step, homogeneity variances were assessed for total accidents occurred in each season and all the accidents occurred due to each source of the hazard. According to Table 4.10 seasonal distribution of accident data achieved the homogeneity variances assumption. Thus, Kruskal Wallis Test can be performed to these data.

Table 4.10 Test of homogeneity variances on seasonal data

	Levene Statistic	df1	df2	Sig.
Total	1.926	3	53	.137
Worker Factor	.638	3	53	.594
Technological Factor	2.130	3	53	.107
Surrounding Activity	.271	3	53	.846
Organisational Factor	.622	3	53	.604
Natural Factor	1.300	3	53	.284

b) Assessment of homogeneity variances assumption on yearly accident data

According to Table 4.11, the yearly distribution of data achieved the homogeneity variances assumption. Thus, Kruskal Wallis test can be performed to these data.

Table 4.11 Test of homogeneity variances on yearly data

	Levene Statistic	df1	df2	Sig.
Total	1.345	4	52	.266
Worker Factor	.701	4	52	.594
Technological Factor	.984	4	52	.425
Surrounding Activity	.547	4	52	.702
Organisational Factor	.272	4	52	.895
Natural Factor	1.572	4	52	.196

c) Assessment of homogeneity variances assumption on yearly accident data

According to Table 4.12, a monthly distribution of data achieved the homogeneity variances assumption except for the natural factors. Kruskal Wallis Test can be performed to accident data except for the natural factors. Welch test needed to be conducted for natural factors.

Table 4.12 Test of homogeneity variances on monthly data

	Levene Statistic	df1	df2	Sig.
Total	1.486	11	45	.170
Worker Factor	1.379	11	45	.216
Technological Factor	1.385	11	45	.213
Surrounding Activity	1.202	11	45	.313
Organisational Factor	1.722	11	45	.099
Natural Factor	2.036	11	45	.047

4.6.2 Outputs of non-parametric Kruskal-Wallis H test statistics for different seasons, years and months

Data for each season, year and month were violated the normal distribution assumption for one-way ANOVA and hence, Non-parametric Kruskal Wallis H test was conducted. According to Table 4.13, Chi-square value for the season is 29.726 for 3 degrees of freedom which is higher than the critical Chi-square value of 7.81. Thus, it can be concluded that there is a significant statistical relationship between the mean number of accidents and accident occurred season. However, Chi-square value for the year is 7.283 for 4 degrees of freedom which is less than the critical Chi-square value of 9.49. Thus, there is no significant statistical relationship between the mean number of accidents and accident occurred year. Nevertheless, Chi-square value for the month is 34.483 for 11 degrees of freedom which is higher than the critical Chi-square value of 19.68. Thus, it can be concluded that there is a significant statistical relationship between the mean number of accidents and accident occurred month.

Table 4.13 Kruskal Wallis test statistics

Grouping Variable	Chi-Square Value	Degrees of Freedom	Asymptotic Sig.
Season	29.726	3	0.000
Year	7.283	4	0.122
Month	34.483	11	0.000

4.6.3 Outputs of non-parametric Kruskal-Wallis H test for seasonal change on natural factor

Data for sources of hazards: worker factor, technological factor, surrounding activity and organisational factor was satisfied with the normal distribution assumption. Thus, statistical significance was directly obtained through ANOVA, and significant groups were identified using Games Howell post hoc test. However, natural factors violated the normal distribution assumption, and thus, the non-parametric Kruskal Wallis test was conducted. According to Table 4.14, Chi-square value is 14.163 for 3 degrees of freedom which is higher than the critical Chi-square value of 7.81. Thus, it can be concluded that there is a significant statistical relationship between seasonal change and the mean number of accidents occurred due to the natural factor.

Table 4.14 Test statistics for seasonal change on natural factor ^{a,b}

	Natural Factor
Chi-Square	14.163
df	3
Asymp. Sig.	.003
a. Kruskal Wallis Test	
b. Grouping Variable: Season	

4.6.4 Outputs of non-parametric Kruskal-Wallis H test for sources of hazards depending on the month

According to Table 4.15, it can be seen that the organisational factors and surrounding activity have no statistically significant relationship. Critical Chi-square value for 11 degrees of freedom is 19.68, which is higher than the Chi-square values shown in Table 4.15. Further, the technological factor and worker factor shows Chi-square values higher than the critical chi-square value. Thus, there is a significant statistical relationship between accident occurred month, and the mean number of accidents occurred due to technological factor and worker factor. However, natural factors were eliminated in this process as it violated the homogeneity variances assumption.

Table 4.15 Test statistics for monthly accident data on sources of hazard ^{a,b}

	Organisational Factor	Surrounding Activity	Technological Factor	Worker Factor
Chi-Square	14.551	14.025	32.141	25.949
df	11	11	11	11
Asymp. Sig.	.204	.232	.001	.007
a. Kruskal Wallis Test				
b. Grouping Variable: Month				

4.6.5 Welch test on natural factors based on monthly accident Data

According to Table 4.16, the significance of accidents based on natural factors depending on months cannot be identified due to the lack of data.

Table 4.16 Welch test on natural factor

	Statistic	df1	df2	Sig.
Welch
Robust tests of equality of means cannot be performed for Natural Factor because at least one group has 0 variance.				

4.7 Outputs of post hoc test for seasonal behaviour of accidents

4.7.1 Dependent variable: Total number of accidents

Results of the Games Howell's post hoc test analysis for different seasons are shown in Table 4.17. It can be seen that each season was compared with the remaining seasons to identify possible statistically significant groups. According to the Table 4.17, there is a significant mean difference in the number of accidents occurred in Winter and Spring ($p=0.005$), Winter and Summer ($p=0.000$), Winter and Autumn ($p=0.008$), Spring and Summer ($p=0.000$), Summer and Autumn ($p=0.008$) at the significance of 0.05 level.

Moreover, descriptive results in Table 4.18 and multiple comparisons in Table 4.17 shows that the mean number of accidents is higher in Spring (151 ± 15 , $p=0.005$), Summer (187 ± 20 , $p=0.000$) and Autumn (158 ± 23 , $p=0.008$) compared to that of Winter (128 ± 16). Further, the mean number of accidents occurred in Spring (151 ± 15 , $p=0.005$) and Autumn (158 ± 23 , $p=0.008$) are comparatively lower than the number of accidents occurred in Summer ($187 \pm$

20). Thus, it can be concluded that there is a high possibility of accidents in Summer compared to other seasons and low possibility of accidents in Winter. However, there was no statistically significant difference between the number of accidents that occurred in Spring (151 ± 15) and Autumn (158 ± 23 , $p=0.810$).

Table 4.17 Multiple comparisons of total number of accidents occurred in each season

Test Procedure: Games-Howell				
(I) Season	(J) Season	Mean Difference (I-J)	Std. Error	Sig.
Winter	Spring	-21.2810*	5.7049	.005
	Summer	-57.2810*	6.6592	.000
	Autumn	-27.9835*	7.6670	.008
Spring	Winter	21.2810*	5.7049	.005
	Summer	-36.0000*	6.4983	.000
	Autumn	-6.7026	7.5276	.810
Summer	Winter	57.2810*	6.6592	.000
	Spring	36.0000*	6.4983	.000
	Autumn	29.2974*	8.2743	.008
Autumn	Winter	27.9835*	7.6670	.008
	Spring	6.7026	7.5276	.810
	Summer	-29.2974*	8.2743	.008

*. The mean difference is significant at the 0.05 level.

Table 4.18 Descriptive results of total number of accidents occurred in each season

	N	Mean	Std. Deviation	Std. Error
Winter	14	129.786	15.5770	4.1631
Spring	15	151.067	15.1066	3.9005
Summer	15	187.067	20.1298	5.1975
Autumn	13	157.769	23.2133	6.4382
Total	57	156.842	27.6946	3.6682

The effect size for the seasonal data was calculated using Equation (3.8) and values in Table 4.19. The parameters are given as follows.

$$\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{24463.047 - (3)348.840}{42951.579 + 348.840} = 0.541 \quad (4.1)$$

$$\omega = \sqrt{0.541} = 0.735 \quad (4.2)$$

The effect size for seasonal behaviour of accident is 0.735. Since the benchmark for larger effect size was taken as 0.5, it can be concluded that the effect of seasons on accidents is a sustentative finding.

Table 4.19 ANOVA of total number of accidents occurred in each season

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	24463.047	3	8154.349	23.376	.000
Within Groups	18488.532	53	348.840		
Total	42951.579	56			

4.7.2 Dependent variable: Total number of accidents per 50 billion dollars

According to Table 4.20, Shapiro-Wilk test shows a normal distribution of construction accidents per 50 billion dollars in different seasons. Also, according to Table 4.21, the homogeneity of the variances is assumed. Thus, ANOVA was used to identify the significance of the seasonal change of accidents in normalised total accidents.

Table 4.20 Test of normality

	Season	Shapiro-Wilk		
		Statistic	df	Sig.
Construction accident rate per 50 billion dollars	Winter	.924	14	.255
	Spring	.961	15	.712
	Summer	.939	15	.369
	Autumn	.934	13	.388

Table 4.21 Test of homogeneity variances

Construction accident rate per 50 billion			
Levene Statistic	df1	df2	Sig.
.900	3	53	.447

According to Table 4.22, ANOVA shows that there is a significant statistical relationship between normalised total accidents with seasons. Thus, the Games Howell post hoc test was conducted to identify significant groups.

Table 4.22 ANOVA of total number of accidents per 50 billion dollars

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	672.074	3	224.025	3.162	.032
Within Groups	3754.617	53	70.842		
Total	4426.691	56			

According to the multiple comparisons in Table 4.23, It can be seen that each season was compared with the remaining seasons to identify possible statistically significant groups. According to Table 4.23, there is a significant mean difference in accidents per 50 billion dollars in Summer and Autumn ($p=0.038$) at the significance of 0.05 level.

Moreover, descriptive results in Table 4.24 and multiple comparisons in Table 4.23 shows that mean number accidents per 50 billion dollars are higher in Summer (80 ± 8 , $p=0.038$), compared to that of Winter (73 ± 9), Spring (73 ± 7), and Autumn (70 ± 9). Thus, it can be concluded that there is a high possibility of accidents in Summer compared to other seasons. However, there was no statistically significant difference between the number of accidents occurred per 50 billion dollars in Winter (73 ± 9), Spring (74 ± 7) and Autumn (70 ± 9).

Table 4.23 Multiple comparisons of construction accidents per 50 billion dollars

Games-Howell				
(I) Season	(J) Season	Mean Difference (I-J)	Std. Error	Sig.
Winter	Spring	-.256	3.160	1.000
	Summer	-6.638	3.264	0.203
	Autumn	2.672	3.588	0.878
Spring	Winter	.256	3.160	1.000
	Summer	-6.381	2.738	0.115
	Autumn	2.929	3.117	0.784
Summer	Winter	6.638	3.264	0.203
	Spring	6.381	2.738	0.115
	Autumn	9.310*	3.222	0.038
Autumn	Winter	-2.672	3.588	0.878
	Spring	-2.929	3.117	0.784
	Summer	-9.310*	3.222	0.038

*. The mean difference is significant at the 0.05 level.

Table 4.24 Descriptive of normalised total of construction accidents

	N	Mean	Std. Deviation	Std. Error
Winter	14	73.378	9.592	2.563
Spring	15	73.634	7.158	1.848
Summer	15	80.016	7.827	2.021
Autumn	13	70.705	9.050	2.510
Total	57	74.583	8.890	1.1776

Depending on the dependent variable, total and total per 50 billion dollars, statistically significant groups have changed. However, both dependent variables show that there is a more significant number of accidents in Summer compared to the other seasons.

4.8 Output of post hoc test for monthly behaviour of accidents

4.8.1 Dependent variable: Total number of accidents

According to the multiple comparisons in Table 4.25, each month's accident rate was compared with the remaining months to identify possible statistically significant groups. (The complete table is included in APPENDIX G) Results showed that there is a significant mean difference between January and July ($p=0.001$), February and July ($p=0.008$), March and July ($p=0.001$), April and July ($p=0.015$), July and November ($p=0.016$), July and December ($p=0.012$), August and December ($p=0.037$) at the significance level of 0.05. Further, descriptive results in Table 4.26 and multiple comparisons in Table 4.25 showed that mean number of accidents occurred in January (135 ± 13 , $p=0.001$), February (129 ± 18 , $p=0.008$), March (152 ± 9 , $p=0.001$), April (152 ± 15 , $p=0.015$), November (137 ± 14 , $p=0.016$) and December (124 ± 16 , $p=0.012$) are lower compared to that of July (198 ± 8). Also, the mean number of accidents occurred in August (186 ± 22 , $p=0.037$) is comparatively higher than that of December (124 ± 16). However, there is no statistically significant mean difference in between May, June, September and October with other remaining months.

Table 4.25 Significant multiple comparisons of total number of accidents occurred in each month

(I) Month	(J) Month	Mean Difference (I-J)	Std. Error	Sig.
January	July	-63.0000*	6.7882	.001
February	July	-69.0000*	8.9989	.008
March	July	-46.6000*	5.3009	.001
April	July	-46.6000*	7.5498	.015
July	January	63.0000*	6.7882	.001
	February	69.0000*	8.9989	.008
	March	46.6000*	5.3009	.001
	April	46.6000*	7.5498	.015
	November	60.9500*	8.0459	.016
	December	74.4500*	8.8262	.012
August	December	61.8500*	12.7641	.037

November	July	-60.9500*	8.0459	.016
December	July	-74.4500*	8.8262	.012
	August	-61.8500*	12.7641	.037

Table 4.26 Descriptive results of total number of accidents occurred in each month

	N	Mean	Std. Deviation	Std. Error
January	5	135.200	13.0843	5.8515
February	5	129.200	18.5930	8.3150
March	5	151.600	9.0167	4.0324
April	5	151.600	15.0266	6.7201
May	5	150.000	22.1246	9.8944
June	5	177.400	24.5010	10.9572
July	5	198.200	7.6942	3.4409
August	5	185.600	22.0068	9.8417
September	5	166.000	24.8797	11.1265
October	4	168.000	17.4547	8.7274
November	4	137.250	14.5459	7.2730
December	4	123.750	16.2558	8.1279
Total	57	156.842	27.6946	3.6682

The effect size was calculated using Equation (3.8) and data in Table 4.27 for the monthly data to discover the significance of the finding. The parameters are as follows.

$$\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{28304.879 - (11)325.482}{42451.579 + 325.482} = 0.58 \quad (4.3)$$

$$\omega = \sqrt{0.58} = 0.76 \quad (4.4)$$

The effect size for monthly behaviour of accident is 0.76. Since the benchmark for larger effect size was taken as 0.5, it can be concluded that the effect of months on the accident is a sustentative finding.

Table 4.27 ANOVA of total number of accidents occurred in each month

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	28304.879	11	2573.171	7.906	.000
Within Groups	14646.700	45	325.482		
Total	42951.579	56			

4.8.2 Dependent variable: Total number of accidents per 50 billion dollars

According to Table 4.28, Shapiro-Wilk test shows a normal distribution of construction accidents per 50 billion dollars in different months. Also, according to Table 4.29, the homogeneity of the variances is assumed. Thus, ANOVA was used to identify the significance of the seasonal change of accidents in normalised total accidents.

Table 4.28 Test of normality

	Month	Shapiro-Wilk		
		Statistic	df	Sig.
Construction accident rate per 50 billion dollars	January	.911	5	.472
	February	.956	5	.782
	March	.886	5	.335
	April	.970	5	.874
	May	.917	5	.510
	June	.954	5	.764
	July	.929	5	.587
	August	.974	5	.903
	September	.904	5	.433
	October	.877	4	.325
	November	.875	4	.316
	December	.857	4	.249

Table 4.29 Test of homogeneity variances

Construction accident rate per 50 billion dollars			
Levene Statistic	df1	df2	Sig.
1.167	11	45	.336

According to the multiple comparisons Table 4.30, results showed that there is a significant mean difference between May and July ($p=0.02$), July and November ($p=0.02$) at the significance level of 0.05. (The complete table is available in APPENDIX H) Further, descriptive results in Table 4.32 and multiple comparisons in Table 4.30 showed that mean number of accidents occurred in May (68 ± 5 , $p=0.02$) and November (65 ± 5 , $p=0.02$) are lower compared to that of July (85 ± 5).

Table 4.30 Significant multiple comparisons of total number of accidents per 50 billion dollars

(I) Month	(J) Month	Mean Difference (I-J)	Std. Error	Sig.
May	July	-16.79*	3.27	0.02
July	May	16.79*	3.27	0.02
	November	19.85*	3.31	0.02
November	July	-19.85*	3.31	0.02

Table 4.31 ANOVA of construction accidents per 50 billion dollars

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1830.364	11	166.397	2.884	.006
Within Groups	2596.328	45	57.696		
Total	4426.691	56			

Table 4.32 Descriptive results of construction accidents per 50 billion dollars

Month	N	Mean	Std. Deviation	Std. Error
January	5	79.34	7.99	3.57
February	5	74.35	8.76	3.92
March	5	78.79	7.58	3.39
April	5	73.85	4.50	2.01
May	5	68.26	5.69	2.54
June	5	76.69	9.65	4.32
July	5	85.06	4.60	2.06
August	5	78.29	7.13	3.19
September	5	71.62	12.12	5.42
October	4	75.06	6.16	3.08
November	4	65.21	5.18	2.59
December	4	64.71	7.11	3.55
Total	57	74.58	8.89	1.18

4.9 Output of seasonal distribution of construction spending

According to Table 4.33, Shapiro-Wilk test shows that construction spending is normally distributed among seasons. Thus, ANOVA was conducted and found that construction spending is significant between season. These results are presented in Table 4.34.

Table 4.33 Test of normality for construction spending on seasons

	Season	Shapiro-Wilk		
		Statistic	df	Sig.
Total State and Local Construction Spending (million dollars)	Winter	.911	14	.163
	Spring	.976	15	.933
	Summer	.952	15	.562
	Autumn	.982	13	.987

Table 4.34 ANOVA for construction spending in million dollars

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6375399938.509	3	2125133312.836	32.162	.000
Within Groups	3502008727.070	53	66075636.360		
Total	9877408665.579	56			

Results of the Games Howell's post hoc test analysis for different seasons based on construction spending are shown in Table 4.35. There is a significant mean difference in construction spending in Winter and Spring ($p=0.001$), Winter and Summer ($p=0.000$), Winter and Autumn ($p=0.000$), Spring and Summer ($p=0.001$) at the significance of 0.05 level. Moreover, descriptive results in Table 4.36 and multiple comparisons in Table 4.35 shows that mean construction spending is lower in Winter compared to that of Summer, Spring and Autumn. Further, the mean of construction spending in Spring and Autumn are lower than the construction spending of Summer. Thus, it can also be a reason for accidents to become higher in Summer than in Winter.

Table 4.35 Multiple comparisons of construction spending in million dollars with seasons

(I) Season	(J) Season	Mean Difference (I-J)	Std. Error	Sig.
Winter	Spring	-14040.895*	3290.921	.001
	Summer	-28034.029*	2825.626	.000
	Autumn	-22589.198*	3014.247	.000
Spring	Winter	14040.895*	3290.921	.001
	Summer	-13993.133*	3056.204	.001
	Autumn	-8548.303	3231.394	.062
Summer	Winter	28034.029*	2825.626	.000
	Spring	13993.133*	3056.204	.001
	Autumn	5444.831	2756.068	.224
Autumn	Winter	22589.198*	3014.247	.000
	Spring	8548.303	3231.394	.062
	Summer	-5444.831	2756.068	.224

*. The mean difference is significant at the 0.05 level.

Table 4.36 Descriptive of construction spending in million dollars

	N	Mean	Std. Deviation	Std. Error
Winter	14	88925.57	8143.55	2176.45
Spring	15	102966.47	9560.22	2468.43
Summer	15	116959.60	6979.10	1801.99
Autumn	13	111514.77	7518.86	2085.35
Total	57	105149.84	13280.9	1759.09

4.10 Output of post hoc test for seasonal behaviour of accidents based on sources of hazards

According to Table 4.37, all the sources of hazards show a significant statistical relationship based on the season of the accident occurred. However, ANOVA value of the natural factors was taken by the Kruskal Wallis test as it violated the normal distribution assumption. Then the statistically significant seasonal groups were identified using a post hoc test.

Table 4.38 presents the multiple comparisons between seasons depending on each source of hazard factor. Accidents occurred due to worker factors has a statistically significant mean difference in Winter and Spring ($p=0.000$) and Spring and Summer ($p=0.012$). Accidents occurred due to technological factor shows a significant mean difference in Winter and Spring ($p=0.017$), Winter and Summer ($p=0.000$) and Spring and Summer ($p=0.005$). Accidents occurred due to organisational factors were significant between Winter and Autumn ($p=0.045$). Natural factors related accidents were significant between Winter and Spring ($p=0.044$) and Spring and Summer ($p=0.002$).

Moreover, multiple comparisons in Table 4.38 and descriptive results in Table 4.39 shows that mean number of accidents occurred in Summer (71 ± 12) due to worker factor is higher compared to that of Winter (49 ± 9) and Spring (57 ± 11). The mean number of accidents occurred in Winter (28 ± 5) due to technological factor is less than that of Spring (34 ± 5), and the mean number of accidents occurred in Summer (43 ± 8) due to technological factor is higher compared to that of Winter (28 ± 5) and Spring (34 ± 5). Further, the mean number of accidents occurred in Winter (2 ± 1) due to organisational factor is less compared to that of Autumn (4 ± 1). Besides, multiple comparisons convey the message that mean number of accidents occurred in

Winter is less than that of the Spring and Summer is higher than the Spring. However, due to the lower sample size obtained through text mining with open data, this conclusion can differ when performing with a higher sample size.

Moreover, Table 4.38 shows that there is no significant mean difference in multiple comparisons in the surrounding activity even though the global effect in Table 4.37 showed a significance. According to Addinsoft (2019), there can be several reasons for contradictory results in ANOVA versus multiple comparisons. They are

- a. Lack of statistical power,
- b. A high number of factor levels,
- c. The weak significance of the global effect,
- d. A conservative multiple comparison test.

However, these reasons cannot be applied to surrounding activity as the sample sizes are the same as others and global significance also at 0.020 level. Thus, this was thoroughly examined and found that Tian et al. (2018) illustrates, it is always essential to perform pairwise comparisons, regardless of the significant status of the ANOVA and findings should be reported. Thus, it can be concluded that the surrounding activity has no statistically significant mean differences with the seasonal change.

The effect size was calculated using Equation (3.8) and data in Table 4.37 for the seasonal effect on each source of hazard to discover the significance of the finding.

a) Worker factor

$$\omega_{WF}^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{3779.35 - (3)123.608}{10330.561 + 123.608} = 0.326 \quad (4.5)$$

$$\omega_{WF} = \sqrt{0.326} = 0.57 \quad (4.6)$$

The effect size for seasonal behaviour of accidents occurred due to worker factor is 0.57. Since the benchmark for larger effect size was taken as 0.5, it can be concluded that the effect of seasonal behaviour on accidents due to worker factor is a sustentative finding.

b) Technological factor

$$\omega_{TF}^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{1741.220 - (3)50.51}{4418.246 + 50.51} = 0.355 \quad (4.7)$$

$$\omega_{TF} = \sqrt{0.326} = 0.59 \quad (4.8)$$

The effect size for seasonal behaviour of accidents occurred due to technological factor is 0.59. Since the benchmark for larger effect size was taken as 0.5, it can be concluded that the effect of seasonal behaviour on accidents due to technological factor is a sustentative finding.

c) Organisational factor

$$\omega_{OF}^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{24.859 - (3)2.699}{167.930 + 2.699} = 0.098 \quad (4.9)$$

$$\omega_{OF} = \sqrt{0.098} = 0.31 \quad (4.10)$$

The effect size for seasonal behaviour of accidents occurred due to organisational factor is 0.31. Since the benchmark for medium effect size was taken as 0.3, it can be concluded that the effect of seasonal behaviour on accidents due to organisational factor has a medium effect.

d) Surrounding activity

$$\omega_{SA}^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{152.967 - (3)14.249}{908.140 + 14.249} = 0.119 \quad (4.11)$$

$$\omega_{SA} = \sqrt{0.119} = 0.35 \quad (4.12)$$

The effect size for seasonal behaviour of accidents occurred due to surrounding activity is 0.35. Since the benchmark for medium effect size was taken as 0.3, it can be concluded that the effect of seasonal behaviour on accidents due to surrounding activity has a medium effect.

e) Natural factor

$$\omega_{NF}^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{12.722 - (3)0.857}{58.140 + 0.857} = 0.172 \quad (4.13)$$

$$\omega_{NF} = \sqrt{0.172} = 0.41 \quad (4.14)$$

The effect size for seasonal behaviour of accidents occurred due to natural factor is 0.41. Since the benchmark for medium effect size was taken as 0.3, it can be concluded that the effect of seasonal behaviour on accidents due to natural factor has a medium effect.

Table 4.37 ANOVA of total number of accidents occurred due to sources of hazard in each season

		Sum of Squares	df	Mean Square	F	Sig.
Worker Factor	Between Groups	3779.350	3	1259.783	10.192	.000
	Within Groups	6551.212	53	123.608		
	Total	10330.561	56			
Technological Factor	Between Groups	1741.220	3	580.407	11.491	.000
	Within Groups	2677.026	53	50.510		
	Total	4418.246	56			
Surrounding Activity	Between Groups	152.967	3	50.989	3.579	.020
	Within Groups	755.173	53	14.249		
	Total	908.140	56			
Organisational Factor	Between Groups	24.859	3	8.286	3.070	.036
	Within Groups	143.071	53	2.699		
	Total	167.930	56			
Natural Factor	Between Groups	12.722	3	4.241	4.949	.004
	Within Groups	45.418	53	.857		
	Total	58.140	56			

Table 4.38 Multiple comparisons of total number of accidents occurred due to each sources of hazard in each season

Dependent Variable	(I) Season	(J) Season	Mean Difference (I-J)	Std. Error	Sig.
Worker Factor	Winter	Spring	-7.8095	3.7624	.188
		Summer	-22.4095*	3.9462	.000
		Autumn	-10.6813	3.9566	.058
	Spring	Winter	7.8095	3.7624	.188
		Summer	-14.6000*	4.3470	.012
		Autumn	-2.8718	4.3564	.911
	Summer	Winter	22.4095*	3.9462	.000
		Spring	14.6000*	4.3470	.012
		Autumn	11.7282	4.5161	.068

	Autumn	Winter	10.6813	3.9566	.058
		Spring	2.8718	4.3564	.911
		Summer	-11.7282	4.5161	.068
Technological Factor	Winter	Spring	-6.0000*	1.8674	.017
		Summer	-15.3333*	2.4296	.000
		Autumn	-6.8462	2.9183	.126
	Spring	Winter	6.0000*	1.8674	.017
		Summer	-9.3333*	2.4951	.005
		Autumn	-.8462	2.9730	.992
	Summer	Winter	15.3333*	2.4296	.000
		Spring	9.3333*	2.4951	.005
		Autumn	8.4872	3.3548	.081
	Autumn	1.00	6.8462	2.9183	.126
		2.00	.8462	2.9730	.992
		3.00	-8.4872	3.3548	.081
Surrounding Activity	Winter	Spring	-.4238	1.2845	.987
		Summer	-3.6905	1.4090	.065
		Autumn	-3.2033	1.4009	.130
	Spring	Winter	.4238	1.2845	.987
		Summer	-3.2667	1.4281	.126
		Autumn	-2.7795	1.4201	.231
	Summer	Winter	3.6905	1.4090	.065
		Spring	3.2667	1.4281	.126
		Autumn	.4872	1.5336	.989
	Autumn	Winter	3.2033	1.4009	.130
		Spring	2.7795	1.4201	.231
		Summer	-.4872	1.5336	.989
Organisational Factor	Winter	Spring	-.6429	.5736	.680
		Summer	-1.5762	.6173	.075
		Autumn	-1.5659*	.5578	.045
	Spring	Winter	.6429	.5736	.680
		Summer	-.9333	.6580	.499
		Autumn	-.9231	.6025	.434

	Summer	Winter	1.5762	.6173	.075
		Spring	.9333	.6580	.499
		Autumn	.0103	.6443	1.00
	Autumn	Winter	1.5659*	.5578	.045
		Spring	.9231	.6025	.434
		Summer	-.0103	.6443	1.00
Natural Factor	Winter	Spring	-.9619*	.3445	.044
		Summer	.2381	.2970	.853
		Autumn	-.5055	.4139	.620
	Spring	Winter	.9619*	.3445	.044
		Summer	1.2000*	.2851	.002
		Autumn	.4564	.4054	.678
	Summer	Winter	-.2381	.2970	.853
		Spring	-1.2000*	.2851	.002
		Autumn	-.7436	.3659	.214
	Autumn	Winter	.5055	.4139	.620
		Spring	-.4564	.4054	.678
		Summer	.7436	.3659	.214

Table 4.39 Descriptive results of total number of accidents occurred due to each sources of hazard depending on season

		N	Mean	Std. Deviation	Std. Error
Worker Factor	Winter	14	48.857	8.7077	2.3272
	Spring	15	56.667	11.4497	2.9563
	Summer	15	71.267	12.3431	3.1870
	Autumn	13	59.538	11.5370	3.1998
	Total	57	59.246	13.5821	1.7990
Technological Factor	Winter	14	28.000	4.7068	1.2579
	Spring	15	34.000	5.3452	1.3801
	Summer	15	43.333	8.0504	2.0786
	Autumn	13	34.846	9.4943	2.6332
	Total	57	35.175	8.8824	1.1765

Surrounding Activity	Winter	14	9.643	3.3422	.8932
	Spring	15	10.067	3.5750	.9231
	Summer	15	13.333	4.2201	1.0896
	Autumn	13	12.846	3.8911	1.0792
	Total	57	11.456	4.0270	.5334
Organisational Factor	Winter	14	2.357	1.3927	.3722
	Spring	15	3.000	1.6903	.4364
	Summer	15	3.933	1.9074	.4925
	Autumn	13	3.923	1.4979	.4154
	Total	57	3.298	1.7317	.2294
Natural Factor	Winter	14	.571	.9376	.2506
	Spring	15	1.533	.9155	.2364
	Summer	15	.333	.6172	.1594
	Autumn	13	1.077	1.1875	.3294
	Total	57	.877	1.0189	.1350

4.11 Output of post hoc test for monthly behaviour of accidents based on sources of hazards

The Games Howell test was conducted to identify significant months for the sources of hazards: worker factor and technological factor. Surrounding activity and organisational factor was not considered as Kruskal Wallis H test given that those are not significant depending on the month. Also, natural factor was not considered as the Welch test unable to find any significance due to lack of data.

According to the Table 4.40, worker factor-related accidents were significant in January and July ($p=0.005$), February and July ($p=0.044$), March and July ($p=0.027$), July and November ($p=0.032$) and July and December ($p=0.016$). Technological factor-related accidents were significant only between June and December ($p=0.036$). Table 4.41, showed that July (76 ± 5) has a higher number of accidents caused by worker factors compared to that of January (53 ± 6), February (48 ± 11), March (55 ± 7), November (50 ± 8) and December (45 ± 8). Moreover, accidents occurred due to technological factor is higher in June (45 ± 8) compared to December (25 ± 4).

These results can be verified using web page data by Kennedy and Lindsey (2020). It states that the USA experience the hottest days of the year at the end of July (15 to 31). Even

though the amount of solar radiation reached a peak in mid of June, Kennedy and Lindsey (2020) state that temperature tends to increase into July.

Table 4.40 Significant multiple comparisons of total number of accidents occurred due to each sources of hazard in each month

Dependent Variable	(I) Month	(J) Month	Mean Difference (I-J)	Std. Error	Sig.
Worker Factor	January	July	-23.4000*	3.5609	.005
	February	July	-28.0000*	5.5749	.044
	March	July	-21.2000*	4.0546	.022
	April	January	4.0000	7.2650	1.000
	July	January	23.4000*	3.5609	.005
		February	28.0000*	5.5749	.044
		March	21.2000*	4.0546	.022
		November	25.7000*	4.5266	.032
	December	31.4500*	4.6693	.016	
	November	July	-25.7000*	4.5266	.032
December	July	-31.4500*	4.6693	.016	
Technological Factor	June	December	20.4000*	4.0571	.036
	December	June	-20.4000*	4.0571	.036

Table 4.41 Descriptive results of total number of accidents occurred due to each sources of hazard depending on month

		N	Mean	Std. Deviation	Std. Error
Worker Factor	January	5	52.800	5.8907	2.6344
	February	5	48.200	11.2561	5.0339
	March	5	55.000	7.3144	3.2711
	May	5	56.800	15.1394	6.7705
	April	5	58.200	13.0269	5.8258
	June	5	62.400	14.4326	6.4545
	July	5	76.200	5.3572	2.3958

	August	5	75.200	12.1737	5.4443
	September	5	64.200	12.2556	5.4809
	October	4	62.750	10.6888	5.3444
	November	4	50.500	7.6811	3.8406
	December	4	44.750	8.0156	4.0078
	Total	57	59.246	13.5821	1.7990
Technological Factor	January	5	30.800	2.3875	1.0677
	February	5	27.600	5.8138	2.6000
	March	5	33.400	7.1274	3.1875
	May	5	35.400	5.5045	2.4617
	April	5	33.200	3.8987	1.7436
	June	5	45.400	7.7330	3.4583
	July	5	45.000	9.0000	4.0249
	August	5	39.600	7.7330	3.4583
	September	5	41.600	10.7842	4.8229
	October	4	34.750	3.0957	1.5478
	November	4	26.500	5.1962	2.5981
	December	4	25.000	4.2426	2.1213
	Total	57	35.175	8.8824	1.1765

The effect size was calculated using Equation 3.8 and data in Table 4.42 for the monthly effect on worker factor and technological factor to discover the significance of the finding.

a) Worker factor

$$\omega_{WF}^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{5021.261 - (11)117.984}{10330.561 + 117.984} = 0.356 \quad (4.15)$$

$$\omega_{WF} = \sqrt{0.356} = 0.59 \quad (4.16)$$

The effect size for monthly behaviour of accidents occurred due to worker factor is 0.59. Since the benchmark for larger effect size was taken as 0.5, it can be concluded that the effect of monthly behaviour on accidents due to worker factor is a sustentative finding.

b) Technological factor

$$\omega_{WF}^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} = \frac{2443.696 - (11)43.879}{4418.246 + 43.879} = 0.356 \quad (4.17)$$

$$\omega_{TF} = \sqrt{0.439} = 0.66 \quad (4.18)$$

The effect size for monthly behaviour of accidents occurred due to technological factor is 0.66. Since the benchmark for larger effect size was taken as 0.5, it can be concluded that the effect of monthly behaviour on accidents due to technological factor is a sustentative finding.

Table 4.42 ANOVA of total number of accidents occurred due to sources of hazard in each month

		Sum of Squares	df	Mean Square	F	Sig.
Worker Factor	Between Groups	5021.261	11	456.478	3.869	.001
	Within Groups	5309.300	45	117.984		
	Total	10330.561	56			
Technological Factor	Between Groups	2443.696	11	222.154	5.063	.000
	Within Groups	1974.550	45	43.879		
	Total	4418.246	56			

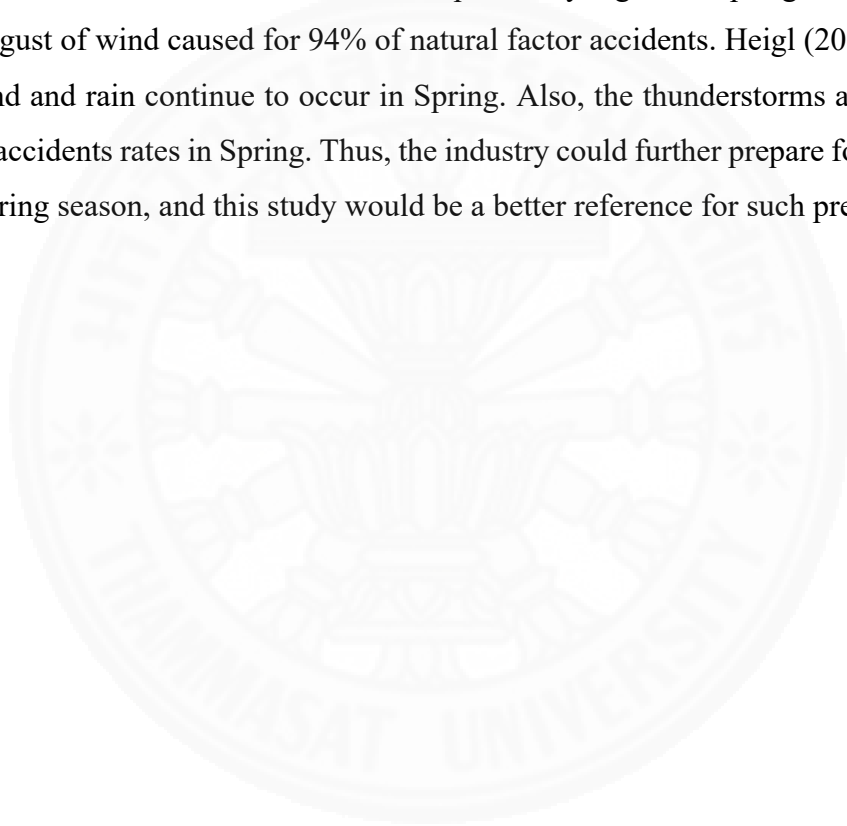
4.12 Discussion on seasonal variation of accidents distribution

In the secondary analysis of this study, it has been found that seasonal variations have a significant ($p=0.001$) impact on occupational accidents. One-way ANOVA results showed that Summer (mostly July) has the highest contribution to the accidents with both normalised and non-normalised data. Since the construction activeness effects were eliminated in normalised data, results illustrate that temperature variation has a higher contribution to the USA's occupational accidents. The average annual temperature throughout the states ranges from 21.5 degrees of Celsius to -3.0 degrees of Celsius. However, this combines the highest temperature values of 56.7 degrees of Celsius and the lowest of -45.0 degrees of Celsius (Osborn, 2020). Thus, this study shows that these temperature variations affect the different source of hazards in various ways. For instance, worker factors show a higher possibility of accidents in Summer. Also, identified worker factor-based hazards showed that health issues accounted for 8.2% of worker factor accidents and contributed as one of the hazard factors for 80% of the accidents. The lack of skill is also contributed to 80% of the accidents, and another fact is that Summer is

when new workers enter the workplace (Liao, 2012). Thus, the industry should focus on worker factor-based hazards causing accidents by implementing proper radiation depletion methods, organising better awareness programs and providing adequate knowledge on the work for unskilled labourers before the commencement of work.

The technological factor accidents are also higher in Summer compared to the other seasons. Since the technological factors involve the accidents occur while utilising the technology, Summer can cause heat exhaustion machinery and would results malfunction of the machinery. Therefore, the industry should pay significant attention to worm down machinery and better maintenance of machinery throughout Summer.

The natural factor-based accidents are comparatively higher in Spring compared to other seasons. The gust of wind caused for 94% of natural factor accidents. Heigl (2018a) states that moisture, wind and rain continue to occur in Spring. Also, the thunderstorms and heavy rains cause higher accidents rates in Spring. Thus, the industry could further prepare for such hazards before the Spring season, and this study would be a better reference for such preparations.



CHAPTER 5

CONCLUSION

5.1 Conclusion

The construction industry is known as one of challenging industry in terms of safety. Safety has been a key consideration over the years and many research studies were engaged in analysing risks based on the construction industry using various technique. According to the existing literature, it had been found that the contribution of sources of hazards such as worker factor, technological factor, natural factor, surrounding activities and organisational factors were the severe cause of accidents. Therefore, this study aimed to identify the contribution of each source of hazard on occupational accidents. Further, the study identified the hazards within each source of hazard and the leading hazards using Pareto analysis.

The study adopts NLP and TM techniques for extraction of required data from open data downloaded from OSHA. Rule-based extraction tool was developed since the existing statistical classifiers required a large amount of data to achieve higher accuracy. The developed classifier was evaluated using the average F1 score and achieved 95% accuracy by training 1500 data while SVM achieved 81% accuracy for 8490 data. Then the developed rule-based classifier was utilised to extract sources of hazards, and frequency analysis was conducted using SPSS 23. Moreover, hazards inside the sources of hazards were also identified and classifiers; SVM, RF, kNN, Kernel SVM, and NB were trained to identify the best extraction tool for hazards. The classifiers trained were evaluated using the F1 score, and RF was identified as the best classifier for extraction of the hazards.

The frequency analysis showed that worker factors were the highest root cause of the occupational accidents followed by the technological factor, surrounding activity, organisational factor and natural factor. Further, Pareto analysis showed that 80% of the worker factor-related hazards were inaccurate foot placement, lack of skill, mishandling, misbehaving and instability of the worker. Also, 80% of the technological factor-based accidents were due to hazards including parts of machinery or equipment failure (28.7%), unstable work area (21.8%) and electrical work (19.8%). The 80% of the surrounding activity-based accidents were due to struck-by hazards (47.8%) and falling of objects (35.5%). Moreover, 80% of organisational factor-based accidents were due to unprotected equipment or areas and works without proper inspection over PPE. Lastly, the gust of wind is the most influencing hazard in natural factor and accounts for 94% contribution.

In addition, research had been identified that accidents were influenced by both construction activeness and seasonal variation. Thus, a secondary analysis was conducted by normalizing the accident data to remove the effect of construction activeness. The seasonal variations were then examined using one-way ANOVA, and significant seasonal and monthly groups had been identified using Games Howell post hoc test. Results showed that the mean number of accidents per 50 billion dollars were higher in Summer (80 ± 8 , $p=0.038$) compared to that of Winter (73 ± 9), Spring (73 ± 7), and Autumn (70 ± 9). Further, the mean number of accidents occurred in May (135 ± 13 , $p=0.001$), February (129 ± 18 , $p=0.008$), March (152 ± 9 , $p=0.001$), April (152 ± 15 , $p=0.015$), November (137 ± 14 , $p=0.016$) and December (124 ± 16 , $p=0.012$) are lower compared to that of July (198 ± 8). Hence, it was concluded that accidents were more severe in Summer, mostly in July compared to other months.

Moreover, seasonal based analysis on the sources of hazards showed that the worker factor and technological factor-based accidents inclined in Summer, while organisational factor-based accidents tend to rise in Autumn. Also, natural factor-based accidents have a rise in Spring.

5.2 Contribution

Occupational health and safety are one of the main considerations and openly available digital accidents records have taken greater attention in novel research studies in safety management. Various kinds of AI models have been developed and paid significant attention to improve the accuracy of the prediction models. Also, studies have captured the accidents causing events such as fall, struck by objects, caught in between, exposure to environmental heat, traffic etc.

However, the main contribution of the rule-based extraction tool formulated in this study is to instantaneously extract the sources of hazards in construction accident data reports with 95% accuracy. The tool utilised N-gram files and set of rules which provide the opportunity to employ human knowledge and judgment to enhance the accuracy. The combination of the developed rule-based classifier and trained RF classifier provided the path for effortless identification of occupational accident-causing hazards relevant to each source of hazard. Despite the fact that existing models utilised the OSHA accident data records, the hazards identified in this study is detailed, and the approach is straightforward. From the practitioners' point of view,

- Rule-based tool along with the RF classifier provides a lower degree of complexity to employ on safety management in any construction industry.
- The tool is robust enough to identify sources of hazard and hazards within each source of hazard in any other domain through a proper adjustment of N-gram files.
- The developed rule-based classifier and RF classifier to analyse digital textual documents would be much helpful for present, and future of construction safety management and this study would be a well-defined domain for such analysis.
- The findings of the study showed that 38% of the construction accidents were generated due to worker factors. Thus, the industry should minimise the errors and violations caused by a worker by implementing proper guidelines.
- Also, identified worker factor-based hazards showed that health issues accounted for 8.2% of worker factor accidents and contributed as one of the hazard factors for 80% of the accidents. This was the leading cause for more severe accidents in the Summer season.
- The lack of skill is also contributed to 80% of the accidents, and another fact is that Summer is when new workers enter the workplace. The hazards which cause 80% of the accidents within each source of hazards are the ones which required greater attention in minimising construction accidents.
- Thus, the industry should focus on worker factor-based hazards causing accidents by implementing proper radiation depletion methods, organising better awareness programs and providing adequate knowledge on the work for unskilled labourers before the commencement of work.
- The industry should also pay significant attention to worm down machinery and better machinery maintenance throughout the Summer season since technological factor-based hazard is higher in Summer.
- The study also showed that safety issues such as unprotected areas and poor use of personnel protective equipment cause 80% of organisational factor-based accidents. These are direct safety management issues, and developing proper methods would eliminate the accident from its root.
- Further, a gust of wind caused 94% of the accidents can be prevented by using proper wind barriers over the working environment. Also, the thunderstorms and heavy rains cause higher accidents rates in Spring. Thus, the industry could further prepare for such hazards before the Spring season, and this study would be a better reference for such preparations.

Therefore, this study industry could gain insights to reduce workplace accidents by developing better safety management strategies.

5.3 Future work

Despite the satisfactory results shown by the developed tool, several future improvements are feasible. Since the model requires a basic literacy in Python, the development of a graphical user interface can be considered for secure industrial usage. The vocabulary in N-gram files related to construction accidents can be further improved to increase the accuracy of the tool. The existence of a larger number of null data in the open data reduces the viability of real-life industry usage of the tool. Thus, it is recommended to develop more feasible rules to avoid this major limitation in the study. The hazards categorised can be subjective to the author. Thus, it is also recommended to insert expertise knowledge during categorisation phases to avoid such subjectiveness. The rule-based classifier developed focused on extracting the most prominent source of hazard factor for the accident causation. However, an accident can be caused due to multiple factors. For instance, an accident can be a combination of technological factors and worker factors. Therefore, it is recommended to implement possible new rules or upgrade the rules by combining several existing rules using AND/OR operator.

In terms of safety management, the National Institute of Occupational Health and Safety (NIOHS) of the USA provides fatality assessment and control evaluation reports (FACE) to make fatality prevention through design. The primary intention of establishing the FACE program was to provide access to the full report of hundreds of accident fatality reports. These reports can also be further analysed using the developed methodology to recognise the hazards and effects can be determined. OSHA also provides severe injury accident records since 1983. However, data from 2015 were recorded in Excel. Rest of these data can be manually arranged into an Excel and can be utilised to investigate the hazard factors further, and this would also lead to higher accuracy in natural factor hazard identification. Since there is a fewer number of natural factor accidents in this study, hazard identification showed a lower accuracy. Thus, the use of more data from FACE reports and OSHA data reports would increase accuracy.

REFERENCES

- Abdelhamid, T. S., & Everett, J. G. (2000). Identifying root causes of construction accidents. *Journal of construction engineering and management*, 126(1), 52-60.
- Addinsoft. (2019). *How to interpret contradictory results between anova and multiple pairwise comparisons?* Retrieved September 6 from https://help.xlstat.com/s/article/how-to-interpret-contradictory-results-between-anova-and-multiple-pairwise-comparisons?language=en_US
- Al Qady, M., & Kandil, A. (2014). Automatic clustering of construction project documents based on textual similarity. *Automation in construction*, 42, 36-49.
- Albert, A., Hallowell, M. R., Kleiner, B., Chen, A., & Golparvar-Fard, M. (2014). Enhancing construction hazard recognition with high-fidelity augmented virtuality. *Journal of construction engineering and management*, 140(7), 04014024.
- Alumbugu, P. O., Shakantu, W. W., John, T. A., & Ola-Awo, A. W. (2019). Automation in construction materials handling: the case study in north central nigeria.
- Arquillos, A. L., Romero, J. C. R., & Gibb, A. (2012). Analysis of construction accidents in Spain, 2003-2008. *Journal of safety research*, 43(5-6), 381-388.
- Ashford, N. A. (1975). Worker health and safety: an area of conflicts. *Monthly Labor Review*, 98(9), 3-11.
- Axelsson, P.-O., & Carter, N. (1995). Measures to prevent portable ladder accidents in the construction industry. *Ergonomics*, 38(2), 250-259.
- Baker, H., Hallowell, M. R., & Tixier, A. J.-P. (2019). AI Predicts Independent Construction Safety Outcomes from Universal Attributes. *arXiv preprint arXiv:1908.05972*.
- Balaguer, C., & Abderrahim, M. (2008). *Robotics and automation in construction*. BoD—Books on Demand.
- Barrett, E. A., & Calhoun, R. A. (1900). Noise & Hearing Protection-Development Of Two Training Exercises For Drillers.
- Bertke, S., Meyers, A., Wurzelbacher, S., Bell, J., Lampl, M., & Robins, D. (2012). Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims. *Journal of safety research*, 43(5-6), 327-332.
- Bird, F. E., Cecchi, F., Tilche, A., & Mata-Alvarez, J. (1974). *Management guide to loss control*. Institute Press.

- Bobick, T. G. (2004). Falls through roof and floor openings and surfaces, including skylights: 1992–2000. *Journal of construction engineering and management*, 130(6), 895-907.
- Brito, L., Rodrigues, M., & Paiva, J. G. S. (2019). A computational system for temporal visual analysis of labour accident data. 2019 23rd International Conference Information Visualisation (IV),
- Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, 45(1), 12-19.
- Burns, C., Ottoboni, F., & Mitchell, H. W. (1962). Health hazards and heavy construction. *American Industrial Hygiene Association Journal*, 23(4), 273-281.
- Carter, G., & Smith, S. D. (2006). Safety hazard identification on construction projects. *Journal of construction engineering and management*, 132(2), 197-205.
- Casteel, C., & Peek-Asa, C. (2000). Effectiveness of crime prevention through environmental design (CPTED) in reducing robberies. *American journal of preventive medicine*, 18(4), 99-115.
- Century, T. (1800). A History of Public Health. *Eighteenth-Century*, 1600(8).
- Chen, L., Vallmuur, K., & Nayak, R. (2015). Injury narrative text classification using factorization model. *BMC medical informatics and decision making*, 15(S1), S5.
- Chen, Q., & Jin, R. (2012). Safety4Site commitment to enhance jobsite safety management and performance. *Journal of construction engineering and management*, 138(4), 509-519.
- Cheng, M.-Y., Kusoemo, D., & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in construction*, 118, 103265.
- Chi, C.-F., Chang, T.-C., & Ting, H.-I. (2005). Accident patterns and prevention measures for fatal occupational falls in the construction industry. *Applied ergonomics*, 36(4), 391-400.
- Chi, S., Han, S., & Kim, D. Y. (2013). Relationship between unsafe working conditions and workers' behavior and impact of working conditions on injury severity in US construction industry. *Journal of construction engineering and management*, 139(7), 826-838.
- Chinniah, Y. (2015). Analysis and prevention of serious and fatal accidents related to moving parts of machinery. *Safety science*, 75, 163-173.
- Choi, J., Gu, B., Chin, S., & Lee, J.-S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. *Automation in construction*, 110, 102974.

- Choudhary, A. K., Oluikpe, P., Harding, J. A., & Carrillo, P. M. (2009). The needs and benefits of Text Mining applications on Post-Project Reviews. *Computers in Industry*, 60(9), 728-740.
- Choudhry, R. M., Fang, D., & Lingard, H. (2009). Measuring safety climate of a construction company. *Journal of construction engineering and management*, 135(9), 890-899.
- Construction Spending Survey*. (2020, October 06, 2020). U.S. Census Bureau. Retrieved September, 10 from <https://www.census.gov/construction/c30/c30index.html>
- Dalziel, C. F. (1972). Electric shock hazard. *IEEE spectrum*, 9(2), 41-50.
- Davies, V. J., & Tomasin, K. (1996). *Construction safety handbook*. Thomas Telford.
- De Melo, R. R. S., Costa, D. B., Álvares, J. S., & Irizarry, J. (2017). Applicability of unmanned aerial system (UAS) for safety inspection on construction sites. *Safety science*, 98, 174-185.
- Deacon, C., & Smallwood, J. (2001). Hazardous chemical substances. *African Newsletter*, 11(1), 17.
- Debios, S. (2009). *Advantages and disadvantages of questionnaires*. Retrieved June, 20 from <https://surveyanyplace.com/questionnaire-pros-and-cons>
- Desvignes, M. (2014). Requisite empirical risk data for integration of safety with advanced technologies and intelligent systems. *University of Colorado*.
- Dong, S., Li, H., & Yin, Q. (2018). Building information modeling in combination with real time location systems and sensors for safety performance enhancement. *Safety science*, 102, 226-237.
- Dropping common terms: stop words*. Retrieved 2019 september 22 from <https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html>
- Dumrak, J., Mostafa, S., Kamardeen, I., & Rameezdeen, R. (2013). Factors associated with the severity of construction accidents: The case of South Australia. *Construction Economics and Building*, 13(4), 32-49.
- Durisko, J. (1997). *The History of Safety in a Construction Environment*. Retrieved 2020 February 20 from <https://blog.vingapp.com/corporate/the-history-of-safety-in-a-construction-environment>
- Dzeng, R.-J., Lin, C.-T., & Fang, Y.-C. (2016). Using eye-tracker to compare search patterns between experienced and novice workers for site hazard identification. *Safety science*, 82, 56-67.

- Eckhoff, R. K. (2016). *Explosion hazards in the process industries*. Gulf Professional Publishing.
- Eriksson, P. E., & Lind, H. (2016). Strategies for reducing moral hazard in construction procurement: a conceptual framework. *Journal of Self-Governance and Management Economics*, 4(1), 7.
- Everett, J. G., & Frank Jr, P. B. (1996). Costs of accidents and injuries to the construction industry. *Journal of construction engineering and management*, 122(2), 158-164.
- Faghihi, V., Nejat, A., Reinschmidt, K. F., & Kang, J. H. (2015). Automation in construction scheduling: a review of the literature. *The International Journal of Advanced Manufacturing Technology*, 81(9-12), 1845-1856.
- Fan, H., & Li, H. (2013). Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques. *Automation in construction*, 34, 85-91.
- Fang, D., Jiang, Z., Zhang, M., & Wang, H. (2015). An experimental method to study the effect of fatigue on construction workers' safety performance. *Safety science*, 73, 80-91.
- Field of computer science and linguistics*. (2019). Retrieved September 22 from https://en.wikipedia.org/wiki/Natural_language_processing
- Fredericks, T. K., Abudayyeh, O., Choi, S. D., Wiersma, M., & Charles, M. (2005). Occupational injuries and fatalities in the roofing contracting industry. *Journal of construction engineering and management*, 131(11), 1233-1240.
- Garg, A. (1991). Ergonomics and the older worker: An overview. *Experimental Aging Research*, 17(3), 143-155.
- Garrett, J., & Teizer, J. (2009). Human factors analysis classification system relating to human error awareness taxonomy in construction safety. *Journal of construction engineering and management*, 135(8), 754-763.
- Gerassis, S., Martín, J., García, J. T., Saavedra, A., & Taboada, J. (2017). Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. *Journal of construction engineering and management*, 143(2), 04016093.
- Gheisari, M., & Esmaeili, B. (2016). Unmanned aerial systems (UAS) for construction safety applications. Construction Research Congress 2016,
- Gill, J. C., & Malamud, B. D. (2016). Hazard interactions and interaction networks (cascades) within multi-hazard methodologies. *Earth System Dynamics*, 7(3), 659.
- Gilmore, C. L. (1970). Accident prevention and loss control.

- Glendon, A. I., & Litherland, D. K. (2001). Safety climate factors, group differences and safety behaviour in road construction. *Safety science*, 39(3), 157-188.
- Goh, Y. M., & Ubeynarayana, C. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention*, 108, 122-130.
- Gunn, S. W. A. (1990). The language of disasters. *Prehospital and Disaster Medicine*, 5(4), 373-376.
- Hallowell, M. R. (2011). Risk-based framework for safety investment in construction organizations. *Journal of construction engineering and management*, 137(8), 592-599.
- Hamid, A. R. A., Yusof, W. Z. W., & Singh, B. S. B. J. (2003). Hazards at construction sites.
- Hasanzadeh, S., Esmaili, B., & Dodd, M. D. (2017). Impact of construction workers' hazard identification skills on their visual attention. *Journal of construction engineering and management*, 143(10), 04017070.
- Hassanein, A. A., & Hanna, R. S. (2008). Safety performance in the Egyptian construction industry. *Journal of construction engineering and management*, 134(6), 451-455.
- Health and safety at work*. Retrieved October 9 from <https://www.ilo.org/global/topics/safety-and-health-at-work/lang--en/index.htm>
- Heigl, C. (2018a). How Seasonal Temperature Changes Affect the Construction Industry. *Operating insights*.
- Heigl, C. (2018b). How Seasonal Temperature Changes Affect the Construction Industry. <https://www.constructconnect.com/blog/seasonal-temperature-changes-affect-construction-industry>
- Helander, M. G. (1991). Safety hazards and motivation for safe work in the construction industry. *International Journal of Industrial Ergonomics*, 8(3), 205-223.
- Hide, S., Atkinson, S., Pavitt, T. C., Haslam, R., Gibb, A. G., & Gyi, D. E. (2003). Causal factors in construction accidents.
- Hinze, J. (1980). Turnover, New workers and safety-closure. *Journal of the construction division-ASCE*, 106(3), 432-432.
- Hinze, J. (1997). *The distractions theory of accident causation* (CIB REPORT), Issue.
- Hinze, J., Devenport, J. N., & Giang, G. (2006). Analysis of construction worker injuries that do not result in lost time. *Journal of construction engineering and management*, 132(3), 321-326.
- Hinze, J., Huang, X., & Terry, L. (2005). The nature of struck-by accidents. *Journal of construction engineering and management*, 131(2), 262-268.

- Hoła, B. (2010). Methodology of hazards identification in construction work course. *Journal of Civil Engineering and Management*, 16(4), 577-585.
- Hsu, J.-y. (2013). Content-based text mining technique for retrieval of CAD documents. *Automation in construction*, 31, 65-74.
- Huang, X., & Hinze, J. (2003). Analysis of construction worker fall accidents. *Journal of construction engineering and management*, 129(3), 262-271.
- Industry, U. S. O. o. C. (1979). *Annual Report - Office of Construction Industry Services*. U.S. Department of Labor, Labor-Management Services Administration, Office of Construction Industry Services.
- Jeong, B. Y. (1998). Occupational deaths and injuries in the construction industry. *Applied ergonomics*, 29(5), 355-360.
- Jiang, G., Choi, B. C., Wang, D., Zhang, H., Zheng, W., Wu, T., & Chang, G. (2011). Leading causes of death from injury and poisoning by age, sex and urban/rural areas in Tianjin, China 1999–2006. *Injury*, 42(5), 501-506.
- Jiao, Z., Yao, P., Zhang, J., Wan, L., & Wang, X. (2019). Capability construction of C4ISR based on AI planning. *IEEE Access*, 7, 31997-32008.
- Jitwasinkul, B., & Hadikusumo, B. H. (2011). Identification of important organisational factors influencing safety work behaviours in construction projects. *Journal of Civil Engineering and Management*, 17(4), 520-528.
- Johnson, S. E. (2007). The predictive validity of safety climate. *Journal of safety research*, 38(5), 511-521.
- Juanals, B., & Minel, J.-L. (2018). An Instrumented Methodology to Analyze and Categorize Information Flows on Twitter Using NLP and Deep Learning: A Use Case on Air Quality. International Symposium on Methodologies for Intelligent Systems,
- Kamaruddin, S. S., Mohammad, M. F., & Mahbub, R. (2016). Barriers and impact of mechanisation and automation in construction to achieve better quality products. *Procedia-Social and Behavioral Sciences*, 222, 111-120.
- Kasperson, R. E., & Pijawka, K. D. (1985). Societal response to hazards and major hazard events: Comparing natural and technological hazards. *Public Administration Review*, 45, 7-18.
- Kecojevic, V., & Radomsky, M. (2005). Flyrock phenomena and area security in blasting-related accidents. *Safety science*, 43(9), 739-750.
- Kennedy, C., & Lindsey, R. (2020, July 7, 2020). *If things go as “normal,” most U.S. locations will have their hottest day of the year by the end of July*. Retrieved September 7 from

<https://www.climate.gov/news-features/featured-images/if-things-go-%E2%80%9Cnormal%E2%80%9D-most-us-locations-will-have-their-hottest-day#:~:text=For%20most%20of%20the%20country,to%20keep%20increasing%20into%20July.>

- Kerr, R., McHugh, M., & McCrory, M. (2009). HSE Management Standards and stress-related work outcomes. *Occupational medicine*, 59(8), 574-579.
- Khan, M. W., Ali, Y., De Felice, F., & Petrillo, A. (2019). Occupational health and safety in construction industry in Pakistan using modified-SIRA method. *Safety science*, 118, 109-118.
- Kim, H., Lee, H.-S., Park, M., Chung, B., & Hwang, S. (2016). Automated hazardous area identification using laborers' actual and optimal routes. *Automation in construction*, 65, 21-32.
- King, R. W., & Hudson, R. (1985). *Construction hazard and safety handbook*. Butterworth-Heinemann.
- Ko, C.-H., & Cheng, M.-Y. (2003). Hybrid use of AI techniques in developing construction management tools. *Automation in construction*, 12(3), 271-281.
- Laufer, A. (1987). Construction accident cost and management safety motivation. *Journal of Occupational Accidents*, 8(4), 295-315.
- Lee, Y.-C., Shariatfar, M., Rashidi, A., & Lee, H. W. (2020). Evidence-driven sound detection for prenotification and identification of construction safety hazards and accidents. *Automation in construction*, 113, 103127.
- Leopold, E., & Leonard, S. (1987). Costs of construction accidents to employers. *Journal of Occupational Accidents*, 8(4), 273-294.
- Levitt, R. E., & Samelson, N. M. (1993). *Construction safety management*. John Wiley & Sons.
- Li, Q., Ji, C., Yuan, J., & Han, R. (2017). Developing dimensions and key indicators for the safety climate within China's construction teams: A questionnaire survey on construction sites in Nanjing. *Safety science*, 93, 266-276.
- Li, X., Jiao, W., Xiao, R., Chen, W., & Liu, W. (2017). Contaminated sites in China: countermeasures of provincial governments. *Journal of Cleaner Production*, 147, 485-496.
- Li, X., Shen, G. Q., Wu, P., & Yue, T. (2019). Integrating building information modeling and prefabrication housing production. *Automation in construction*, 100, 46-60.
- Liao, C.-W. (2012). Pattern analysis of seasonal variation in occupational accidents in the construction industry. *Procedia Engineering*, 29, 3240-3244.

- Liaudanskiene, R., Varnas, N., & Ustinovichius, L. (2010). Modelling the application of workplace safety and health act in Lithuanian construction sector. *Technological and Economic Development of Economy*, 16(2), 233-253.
- Ling, F. Y. Y., Liu, M., & Woo, Y. C. (2009). Construction fatalities in Singapore. *International Journal of Project Management*, 27(7), 717-726.
- Liu, Z., Cao, Y., Wang, Y., & Wang, W. (2019). Computer vision-based concrete crack detection using U-net fully convolutional networks. *Automation in construction*, 104, 129-139.
- Loomis, D. P., Richardson, D. B., Wolf, S. H., Runyan, C. W., & Butts, J. D. (1997). Fatal occupational injuries in a southern state. *American journal of epidemiology*, 145(12), 1089-1099.
- López, M. A. C., Fontaneda, I., Alcántara, O. J. G., & Ritzel, D. O. (2011). The special severity of occupational accidents in the afternoon: "The lunch effect". *Accident Analysis & Prevention*, 43(3), 1104-1116.
- Lortie, M., & Rizzo, P. (1999). Reporting and classification of loss of balance accidents. *Safety science*, 33(1-2), 69-85.
- Macdonald, W. (2012). Models of causation: Health determinants in HaSPA (Health and Safety Professionals Alliance), the core body of knowledge for generalist OHS professionals. *Tullamarine, VIC: Safety Institute of Australia*.
- Manning, R. B. (1988). *Village revolts: social protest and popular disturbances in England, 1509-1640*. Oxford University Press.
- Matarneh, S. T., Danso-Amoako, M., Al-Bizri, S., Gaterell, M., & Matarneh, R. (2019). Building information modeling for facilities management: A literature review and future research directions. *Journal of Building Engineering*, 24, 100755.
- McClay, R. E. (1989). Toward a more universal model of loss incident causation. *Professional Safety*, 34(1), 15.
- McKenzie, K., Campbell, M. A., Scott, D. A., Discoll, T. R., Harrison, J. E., & McClure, R. J. (2010). Identifying work related injuries: comparison of methods for interrogating text fields. *BMC medical informatics and decision making*, 10(1), 1-10.
- Mitropoulos, P., Abdelhamid, T. S., & Howell, G. A. (2005). Systems model of construction accident causation. *Journal of construction engineering and management*, 131(7), 816-825.
- Mohamed, S. (2002). Safety climate in construction site environments. *Journal of construction engineering and management*, 128(5), 375-384.

- Molenaar, K., Brown, H., Caile, S., & Smith, R. (2002). Corporate culture. *Professional Safety*, 47(7), 18-27.
- Muir, H., & Thomas, L. (2004). Passenger safety and very large transportation aircraft. *Measurement and Control*, 37(2), 53-58.
- Mullgn, E. A. (1997). Workplace Violence: cause for concern or the construction of a new category of fear? *Journal of Industrial Relations*, 39(1), 21-32.
- N-gram*. Retrieved March 20 from <https://en.wikipedia.org/wiki/N-gram>
- Namian, M., Albert, A., Zuluaga, C. M., & Behm, M. (2016). Role of safety training: Impact on hazard recognition and safety risk perception. *Journal of construction engineering and management*, 142(12), 04016073.
- Neeleman, J., Ormel, J., & Bijl, R. (2001). The distribution of psychiatric and somatic ill health: Associations with personality and socioeconomic status. *Psychosomatic medicine*, 63(2), 239-247.
- Nishimoto, T., Mukaigawa, K., Tominaga, S., Lubbe, N., Kiuchi, T., Motomura, T., & Matsumoto, H. (2017). Serious injury prediction algorithm based on large-scale data and under-triage control. *Accident Analysis & Prevention*, 98, 266-276.
- Notelaers, G., De Witte, H., Van Veldhoven, M., & Vermunt, J. K. (2007). Construction and validation of the short inventory to monitor psychosocial hazards. *Médecine du Travail et Ergonomie*, 44(1/4), 11.
- Nozaki, D., Okamoto, K., Mochida, T., Qi, X., Wen, Z., Tokuda, K., Sato, T., & Tamesue, K. (2018). AI management system to prevent accidents in construction zones using 4K cameras based on 5G network. 2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC),
- Osborn, L. (2020). *Average Annual Temperature for Each US State*. Current Results. Retrieved September 20 from <https://www.currentresults.com/Weather/US/average-annual-state-temperatures.php>
- Owen, L. A., Kamp, U., Khattak, G. A., Harp, E. L., Keefer, D. K., & Bauer, M. A. (2008). Landslides triggered by the 8 October 2005 Kashmir earthquake. *Geomorphology*, 94(1-2), 1-9.
- Ozturk, G. B. (2020). Interoperability in building information modeling for AECO/FM industry. *Automation in construction*, 113, 103122.
- Peterson, T. C., Heim Jr, R. R., Hirsch, R., Kaiser, D. P., Brooks, H., Diffenbaugh, N. S., Dole, R. M., Giovannetone, J. P., Guirguis, K., & Karl, T. R. (2013). Monitoring and understanding changes in heat waves, cold waves, floods, and droughts in the United

- States: state of knowledge. *Bulletin of the American Meteorological Society*, 94(6), 821-834.
- Poh, C. Q., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. *Automation in construction*, 93, 375-386.
- Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2), 143-157.
- Rai, A. (2019). *What is Text Mining: Techniques and Applications?* .
<https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/>
- Reason, J. (1990). The contribution of latent human failures to the breakdown of complex systems. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 327(1241), 475-484.
- Reason, J. (2016). *Managing the risks of organizational accidents*. Routledge.
- Reason, J. (2017). The contribution of latent human failures to the breakdown of complex systems. In *Human error in aviation* (pp. 5-14). Routledge.
- Rezakhani, P. (2012). Classifying key risk factors in construction projects. *Buletinul Institutului Politehnic din Iasi. Sectia Constructii, Arhitectura*, 58(2), 27.
- Rinefort, F. C. (1977). A new look at occupational safety... a cost-benefit analysis of selected Texas industries. *Professional Safety*, 22(9), 8-13.
- Risk management - Principles and guidelines Sydney: Standards Australia, (2009).
- Robb, D., & Barnes, T. (2018). Accident rates and the impact of daylight saving time transitions. *Accident Analysis & Prevention*, 111, 193-201.
- Robbins, P., Farnsworth, R., & Paul Jones III, J. (2008). Insects and institutions: Managing emergent hazards in the US Southwest. *Journal of Environmental Policy and Planning*, 10(1), 95-112.
- Sacks, R., Rozenfeld, O., & Rosenfeld, Y. (2009). Spatial and temporal exposure to safety hazards in construction. *Journal of construction engineering and management*, 135(8), 726-736.
- Severe Injury Reports | Occupational Health and safety Administration*. Retrieved September 21 from <https://www.osha.gov/severeinjury/>
- Shen, L., Yan, H., Fan, H., Wu, Y., & Zhang, Y. (2017). An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design. *Building and Environment*, 124, 388-401.

- Shendell, D. G., Mizan, S. S., Marshall, E. G., Kelly, S. W., Therkorn, J. H., Campbell, J. K., & Miller, A. E. (2012). Cut-laceration injuries and related career groups in New Jersey career, vocational, and technical education courses and programs. *Workplace health & safety*, 60(9), 401-409.
- Simonds, R. H., & Grimaldi, J. V. (1956). *Safety management: Accident cost and control*. RD Irwin.
- Sorock, G. S., Ranney, T. A., & Lehto, M. R. (1996). Motor vehicle crashes in roadway construction workzones: an analysis using narrative text from insurance claims. *Accident Analysis & Prevention*, 28(1), 131-138.
- Stoilkovska, B. B., Žileska Pančovska, V., & Mijoski, G. (2015). Relationship of safety climate perceptions and job satisfaction among employees in the construction industry: the moderating role of age. *International journal of occupational safety and ergonomics*, 21(4), 440-447.
- Suraji, A., Duff, A. R., & Peckitt, S. J. (2001). Development of causal model of construction accident causation. *Journal of construction engineering and management*, 127(4), 337-344.
- Tam, C., Fung, I., & Ivan, W. (1998). Effectiveness of safety management strategies on safety performance in Hong Kong. *Construction Management & Economics*, 16(1), 49-55.
- Tang, S., Shelden, D. R., Eastman, C. M., Pishdad-Bozorgi, P., & Gao, X. (2019). A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. *Automation in construction*, 101, 127-139.
- Teizer, J., & Cheng, T. (2015). Proximity hazard indicator for workers-on-foot near miss interactions with construction equipment and geo-referenced hazard areas. *Automation in construction*, 60, 58-73.
- Teizer, J., Cheng, T., & Fang, Y. (2013). Location tracking and data visualization technology to advance construction ironworkers' education and training in safety and productivity. *Automation in construction*, 35, 53-68.
- Templer, J. (1995). *The staircase: studies of hazards, falls, and safer design* (Vol. 2). MIT press.
- Tian, C., Manfei, X., Justin, T., Hongyue, W., & Xiaohui, N. (2018). Relationship between Omnibus and Post-hoc Tests: An Investigation of performance of the F test in ANOVA. *Shanghai archives of psychiatry*, 30(1), 60.

- Tixier, A. J.-P., Hallowell, M. R., Albert, A., van Boven, L., & Kleiner, B. M. (2014). Psychological antecedents of risk-taking behavior in construction. *Journal of construction engineering and management*, *140*(11), 04014052.
- Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016a). Application of machine learning to construction injury prediction. *Automation in construction*, *69*, 102-114.
- Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016b). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in construction*, *62*, 45-56.
- Townsend, A. K., & Barker, C. M. (2014). Plastic and the nest entanglement of urban and agricultural crows. *PLoS One*, *9*(1), e88006.
- Treves, A., Martin, K. A., Wydeven, A. P., & Wiedenhoeft, J. E. (2011). Forecasting environmental hazards and the application of risk maps to predator attacks on livestock. *BioScience*, *61*(6), 451-458.
- Tschetter, J., & Lukaszewicz, J. (1983). Employment changes in construction: secular, cyclical, and seasonal. *Monthly Lab. Rev.*, *106*, 11.
- Vitharana, V., De Silva, G., & De Silva, S. (2015). Health hazards, risk and safety practices in construction sites—a review study. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, *48*(3).
- Walsh, K., White, C. J., McInnes, K., Holmes, J., Schuster, S., Richter, H., Evans, J. P., Di Luca, A., & Warren, R. A. (2016). Natural hazards in Australia: storms, wind and hail. *Climatic Change*, *139*(1), 55-67.
- Williams, T. P., & Gong, J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in construction*, *43*, 23-29.
- Woodbury, R. M. (1927). *Workers health and safety: A statistical program*. Macmillan Company, New York.
- Woods, D., & Dekker, S. (2000). Anticipating the effects of technological change: a new era of dynamics for human factors. *Theoretical issues in ergonomics science*, *1*(3), 272-282.
- Xia, N., Zou, P. X., Liu, X., Wang, X., & Zhu, R. (2018). A hybrid BN-HFACS model for predicting safety performance in construction projects. *Safety science*, *101*, 332-343.
- Yan, X., Radwan, E., & Abdel-Aty, M. (2005). Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. *Accident Analysis & Prevention*, *37*(6), 983-995.

- Yilmaz, M., & Kanit, R. (2018). A practical tool for estimating compulsory OHS costs of residential building construction projects in Turkey. *Safety science, 101*, 326-331.
- Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in construction, 99*, 238-248.
- Zhang, L., Liu, Q., Wu, X., & Skibniewski, M. J. (2016). Perceiving interactions on construction safety behaviors: Workers' perspective. *Journal of Management in Engineering, 32*(5), 04016012.
- Zhang, W., Zhu, S., Zhang, X., & Zhao, T. (2020). Identification of critical causes of construction accidents in China using a model based on system thinking and case analysis. *Safety science, 121*, 606-618.
- Zhou, C., & Ding, L. (2017). Safety barrier warning system for underground construction sites using Internet-of-Things technologies. *Automation in construction, 83*, 372-389.
- Zohar, D. (1980). Safety climate in industrial organizations: theoretical and applied implications. *Journal of applied psychology, 65*(1), 96.
- Zou, P. X., Lun, P., Cipolla, D., & Mohamed, S. (2017). Cloud-based safety information and communication system in infrastructure construction. *Safety science, 98*, 50-69.
- Zou, Y., Kiviniemi, A., & Jones, S. W. (2017). Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Automation in construction, 80*, 66-76.



APPENDIX A**MACRO FOR CONSTRUCTION RELATED DATA SEPARATION**

```
Private Sub CommandButton1_Click()  
    'All is the worksheet with all of the data  
    'Sheet1 is the construction related data set  
    '23 is the code for construction data according to primary NAICS  
    a = Worksheets("All").Cells(Rows.Count, 1).End(xlUp).Row  
    For i = 2 To a  
        x = Left(Cells(i, 12).Value, 2)  
        If x = 23 Then  
            Worksheets("All").Rows(i).Copy  
            Worksheets("Sheet1").Activate  
            b = Worksheets("Sheet1").Cells(Rows.Count, 1).End(xlUp).Row  
            Worksheets("Sheet1").Cells(b + 1, 1).Select  
            ActiveSheet.Paste  
            Worksheets("All").Activate  
        End If  
    Next  
    Application.CutCopyMode = False  
    ThisWorkbook.Worksheets("All").Cells(1, 1).Select  
End Sub
```

APPENDIX B

LEXICON

Source of hazard	Unigram	Bigram	Trigram
Natural factor	gust thunderstorm climate topography landslide earthquake flood tsunami cyclone tornado	high wind bush fire wild fire wind caused volcanic eruption land slide heat wave during thunderstorm ground collapse wet weather strong wind wind blew Ice/slick roads Gust wind	wind displaced tree gust wind caused violent thunderstorm occurred wind caught them wind caused plate Strong gust wind Ice/slick roads lightning struck the crane boom
Organisational factor	unguarded unhooked bitten stung wasp	not trained clumsy automation less tools unworkable procedures inadequate responses safety culture safety harnessing fall protection roof opening floor opening bacterial infection open hole not worn	no fall protection high work pressure fell through hole guardrail not placed not wearing seatbelts copper head snake performing work 80-degree personal fall arrest personal fall protection ppe not connected rattle snake bit stairs not guardrail temporary employee using guardrail not place

		<p>poor design</p> <p>complex technology</p> <p>managerial complexity</p> <p>limited visibility</p> <p>no railing</p> <p>guard rail</p> <p>Unprotected stairwell</p> <p>management issues</p> <p>venomous snake</p>	<p>Employee exposed steam</p> <p>fell through a hatch</p> <p>Not personal fall</p> <p>Slipped muddy conditions</p> <p>fell through an opening</p>
Surrounding activity	<p>someone</p> <p>motorcycle</p> <p>automobile</p> <p>robbed</p> <p>dog</p>	<p>fell on</p> <p>struck by</p> <p>fell from</p> <p>motor vehicle</p> <p>vehicle struck</p> <p>public vehicle</p> <p>approaching truck</p> <p>driven by</p> <p>general public</p> <p>struck another</p> <p>traffic vehicle</p> <p>passenger vehicle</p> <p>beam rolled</p> <p>remote control</p> <p>Adjacent employee</p> <p>hit by</p> <p>flagging traffic</p> <p>privately driven</p> <p>dog chased</p> <p>coworker struck</p> <p>beam fell</p> <p>object swung</p> <p>unit hit</p> <p>nearby grinder</p>	<p>fell from upper</p> <p>from upper elevation</p> <p>foreign body injected</p> <p>backhoe ran over</p> <p>piece lumber fell</p> <p>excavator made contact</p> <p>swiped by trailer</p> <p>stack sheetrock fell</p> <p>fell during removal</p> <p>struck by motor</p> <p>falling piece metal</p> <p>removed from upper</p> <p>falling piece metal</p> <p>by backhoe bucket</p> <p>struck by car</p> <p>kelly bar dropped</p> <p>stung numerous times</p> <p>plywood fell over</p> <p>Excavator struck employee</p> <p>piece plywood fell</p> <p>after hitting hydro-mobile</p> <p>blast sand blaster</p> <p>piece of metal fell</p> <p>trash cart fell</p>

		another employee	<p>boom section fell</p> <p>piece conduit fell</p> <p>piece concrete fell</p> <p>SUV turned onto the street at a high rate of speed</p> <p>block of wood fell off</p> <p>steel post then fell</p> <p>car struck him</p> <p>city inspector drove up past</p> <p>piece of PVC pipe fell from above</p> <p>made an abrupt backward turn</p> <p>piece of granite</p> <p>board from the fifth floor fell off</p> <p>Tree stump came</p> <p>rock knocked over</p> <p>roller struck him</p> <p>Struck in head</p> <p>delivery truck struck</p> <p>Struck head by</p> <p>Struck chest by</p> <p>plug from the block valve blew out</p> <p>piece slag fell</p> <p>reversing bulldozer struck</p> <p>roofing paper fell off</p> <p>Another vehicle collided</p> <p>Car drove through</p> <p>passing driver struck</p> <p>Allow car pass</p> <p>Car hit wood</p> <p>truck hit bump</p>
--	--	------------------	---

			<p>exposed fall hazards piece steel fell company pickup truck Fell upper elevation rock rolled down vehicle ran through bobcat operator struck Unknown object fell semi truck struck Concrete supports struck debris fell from fell from backhoe falling piece material</p>
Technological factor	<p>collapsed combustion dislodged electrocuted entangled explosion exploded failed kicked leak loose lose malfunction shocked sparked tilted unbalanced unexpectedly unstable vibration</p>	<p>one snapped spud shifted arc flash topsail separated Rolled back broke away excavator shifted flash fire cap slipped struts shifted beam broke ATV flipped burst free pipe moved machine activated pipe shattered wrong direction Turnstile moved ladder fell weight shifted</p>	<p>access plate slipped anti-slip feet amputated open pin end frame detached automatic valve closed beams fell on tanker valve was not closed received electrical burns hose started shake being suspended by unclogging pump truck fan blade struck rocker beam slid off garage door opened chain whipped around telehandler rolled over hot water condensate released out of a steam line directly above chain hoist snapped</p>

		snapped forward rolled off tool discharged ladder snapped gate slipped Fire started Rope broke cart tipped buggy slid barrier rolled became entangled became imbalanced became tangled became unstable bounced off become overloaded began roll began tilt Tractor rolled belt broke between mechanical blade broke blew apart Pump clogged saw shattered bounced up braces broke broke free broke off buggy tipped cage broke it ignited carbon monoxide	natural gas flared and caught fire scaffolding plank broke tailgate closed contact with pesticides boom came down quickly fire blew out temporary suspender spun sharp metal scratched band saw activated trailer rolled over ladder's feet were not properly positioned table was not stabilized boom struck a slab truss rolled over rebar slid off excavator bucket moved instead boom went sideways lift went off frame flipped over backhoe forks disengaged scaffold rolled over panel fell over scaffold plank broke machine rolled over pipe came free board fell striking concrete chipped off lever got stuck rear outrigger penetrated attachment rocked back bucket of a trackhoe fell
--	--	---	--

		chain snapped conveyor grabbed cord snapped crane broke drill cycled drywall fell down reverse ejected from electric shock energized by electrical shock energized cable energized wiring explosion occurred flammable gasses flange slipped flash burn flew off fork slipped form fell gave way gave out generator fell grinder jumped guardrail flung gun punctured hit rut hole scaffold hook broke hook slipped hose broke hose disconnected it rotated	pressurized pipe burst broken grinding wheel presurized pipe burst block fell over concrete shifted struck metal splintered off track-hoe struck push pole fell nail gun discharged shard flew up forklift moved forward boom fell on brick slid down blade became stuck blown with air bit got stuck angle iron slipped bridge piling fell Fell through gap Tie wire broke Air got into rollback bed struck got into line Flipped it over Pump shot concrete Pipe cap exploded Scaffold tipped over Ladder slid out Pry bar slipped Backhoe struck stem Chainsaw kicked back Car created loop Forklift rolled over
--	--	---	---

	<p>it slipped</p> <p>it swung</p> <p>it tipped</p> <p>it turned</p> <p>jack slipped</p> <p>jib fell</p> <p>kick out</p> <p>kicked back</p> <p>kicked out</p> <p>knife slipped</p> <p>ladder acted</p> <p>ladder buckled</p> <p>ladder slipped</p> <p>ladder tipped</p> <p>ladder, broke</p> <p>lurched forward</p> <p>load shifted</p> <p>load struck</p> <p>lost power</p> <p>magnet released;</p> <p>methane gas</p> <p>mount moved</p> <p>moved suddenly</p> <p>one snapped</p> <p>outrigger retracting</p> <p>panel energized</p> <p>panel slipped</p> <p>pile rolled</p> <p>pin broke</p> <p>pipe shifted</p> <p>pipe slipped</p> <p>pipe struck</p> <p>pipes slipped</p>	<p>Crane traveled backward</p> <p>Pressurized water blast</p> <p>Truck rolled over</p> <p>broke from choke</p> <p>bundles fell off</p> <p>hammer tipped over</p> <p>cap came off</p> <p>cable rose upward</p> <p>Ball dropped suddenly</p> <p>came off roller</p> <p>carbon monoxide intoxication.</p> <p>carbon monoxide poisoning</p> <p>carriage came down</p> <p>cart went around</p> <p>caught hand system</p> <p>caused by sledgehammer</p> <p>aerial lift moved</p> <p>chain came off</p> <p>flash fire occurred</p> <p>tailgate came down</p> <p>truck moved forward</p> <p>ladder slid sideways</p> <p>chain fall jumped</p> <p>vapor fire occurred</p> <p>clogged drill head</p> <p>contact with live</p> <p>contact with machine's</p> <p>contacted energized wire</p> <p>contacted overhead electric</p> <p>contacted overhead power</p> <p>contacting live electrical</p> <p>conveyor came over</p> <p>cross arm slipped</p>
--	--	---

	plate slipped plate shifted point broke post slipped pouring broke pressure injected pressure released prolonged exposure pulled through pump caught pump shifted railing broke plate bent received shock vise triggered receiving laceration rod ejected whip around rolled backwards rolled forward rope broke run over ran over rolled out saw jumped strap broke saw exploded separated from shot out slid backwards slipped off Skytrak fell spark fell	cup tool disintegrated cylinder fell machine caught on dismounting delivery truck dolly tipped over drum pushed chute driving turned over. equipment turned on excavator backed over exposed hydraulic fluid exposed to steam fitting not fit fusion pipe fell fan turned on form came free gas saw ignited gun bounced off gun went off guard rail fell hammer slipped hatch cover fell hazardous chemical splashed hit scissor lift hook slipped off hooks not latched hydraulic pump busted impaled by power inadvertently went off jib attachment slipped jib fell of knife went sideways ladder became displaced ladder slid away
--	--	--

	<p>sprung off</p> <p>strap broke</p> <p>struck live</p> <p>steel fell</p> <p>splashed by</p> <p>sulfuric acid</p> <p>swung down</p> <p>sheave up</p> <p>table moved</p> <p>truss fell</p> <p>trailer shifted</p> <p>tip over</p> <p>tipped over</p> <p>truss shifted</p> <p>unexpectedly released</p> <p>weak area</p> <p>weld broke</p> <p>twisted fell</p> <p>wrench broke</p> <p>wrench slipped</p>	<p>ladder slid down</p> <p>lift acted abnormally</p> <p>lifting mechanism snapped</p> <p>live electrical line</p> <p>load slipped off</p> <p>loading chute fell</p> <p>lost control rigging</p> <p>knocking it over</p> <p>machine blade caught</p> <p>machine grabbed piece</p> <p>machine stuck reverse</p> <p>made contact power</p> <p>manhole cover flipped</p> <p>material shifted</p> <p>metal rebar broke</p> <p>millwright tripped over</p> <p>nail gun bumped</p> <p>nipped by concrete</p> <p>not secured properly</p> <p>not securely rigged</p> <p>not set up</p> <p>not shut off</p> <p>Contact with live</p> <p>panel shorted out</p> <p>pallet came down</p> <p>part forks fell</p> <p>pinned by elevator</p> <p>pipe slid off</p> <p>platform came down</p> <p>press door closed</p> <p>rail rolled off</p> <p>rebar sticking out</p> <p>rigging material slipped</p>
--	--	--

			rigging fell out rope got caught repeatedly strike safety gate released saw kick back sheets slid off spud shifted supportive strap broke SkyTrak tip over slid out place slipped off trailer strap came off standing on broke started spin around started moving forward stone weighing tipped struck by hazards steel louvers slipped tape popped off truck lurched backwards truck tipped over wall form fell wood caught blade wire rope pulled wall fell outward ladder knocked off wall panel fell ladder knocked over wood got caught roller rolled over leaned it against
Worker factor	assisting attempt	accidentally stepped accidentally shot	bypassed safety trigger carrying large fan

attempted	Employee dropped	catch his balance
attempting	Stepped in	clamps got caught
checking	adjusted scaffold	employee put his hand
felt	became dizzy	conducting maintenance
forgot	became ill	activities
fainted	became infected	connecting pin removed
grabbed	became lightheaded	contact live wire
helping	began cramping	contact station transformer
missed	began taking	contacted live conduit
misstepped	blacked out	employee was refueling a water
slip	employee activated	pump when the gasoline began
slipped	came contact	to overflow out of the fuel tanks
spraying	inhaled vapors	dismantling sections of
thinking	slumped over	scaffolding
touched	caught it	contacted energized circuit
tried	changed length	contacted energized electric
tripped	chipping concrete	coworker's machete struck
trying	car entered the	Began to feel ill
unaware	construction area	forklift drove over
walked	climbed over	trailer operator moved
walking	walked up	Employee used running
dizziness	climbing out	He stepped into
maintainance	cut hand	threw trash out
Repair	Began arguing	coworker ran over
	driver took	pulling ladder towards
	employee attempting	overcome by heat
	employee pulled	dropped by another
	employee came	shaking out steel
	employee contacted	by another subcontractor
	employee de-	lost sight of the employee
	energized	working on the ground close to
	He jumped	an excavator
	employee dismounted	coworker swung a hammer

		<p>employee dismounted</p> <p>employee dropped</p> <p>employee dropped</p> <p>employee exited</p> <p>employee hit</p> <p>employee hooked</p> <p>employee jumped</p> <p>employee lacerated</p> <p>employee operating</p> <p>employee reached</p> <p>employee removing</p> <p>employee slid</p> <p>started vomiting</p> <p>employee stumbled</p> <p>employee touched</p> <p>employee transitioning</p> <p>employee tripped</p> <p>employee twisted</p> <p>employee unhooked</p> <p>employee's powered</p> <p>employees detached</p> <p>entered excavation</p> <p>everyone let go</p> <p>exited cab</p> <p>experience cramping</p> <p>experienced</p> <p>dehydration</p> <p>experienced headache</p> <p>experienced pain</p> <p>experiencing cramps</p> <p>feel ill</p> <p>feel lightheaded</p> <p>feeling dehydrated</p>	<p>Employee moved close</p> <p>he swung the hammer</p> <p>Coworker immediately pressed</p> <p>Driver cab accidentally</p> <p>operator turned vehicle</p> <p>standing near coworker</p> <p>placed against the wall</p> <p>cleaning head came out of the</p> <p>tank through the manway</p> <p>Breaking right hip</p> <p>operator's knee struck</p> <p>Trying reattach another</p> <p>walking between barrier</p> <p>employee shot himself</p> <p>Fist fight another</p> <p>Suffered cardiac event</p> <p>Collided another employee</p> <p>Another employee moved</p> <p>Assisting another employee</p> <p>Suffered abdominal hernia</p> <p>injured employee working</p> <p>Training another employee</p> <p>Employee removed gravity</p> <p>turned speak another</p> <p>lost their balance</p> <p>Employee suffered shock</p> <p>Employee hand resting</p> <p>Pulling himself off</p> <p>he became overheated</p> <p>employee stepped backward</p> <p>coworker started machine</p> <p>crane began lift</p> <p>crossing back over</p>
--	--	--	---

	<p>felt dehydrated. felt dizzy felt lightheaded felt sick felt pain heat issues heat exhaustion heat stress stepped backwards he experienced operator inadvertently employee experienced felt strange fingers pulled fingers slipped poor health heat stroke foot slipped foot slipped foreign object hand slipped harassed by he exited he hit he missed he moved he taking jumped off jumped on kicked away leaned against leaned back leg cramps</p>	<p>did not deploy employee become overheated employee began sweating employee bumped hatch employee came employee developed infection employee inadvertently moved employee jersey barrier employee lifted up employee passed out employee removed screws employee run over employee started cramping employee stumbled over employee suffered dehydration. employee tied off employee used foot employee's finger contacted employee's finger got employee's foot went employee's work belt excavator operator backed experience heart-attack experience heat symptoms experienced back pain experienced severe cramping experienced severe dehydration experienced stomach pain experiencing elbow soreness finger came contact foot became entangled forklift operator pushed hand got pulled</p>
--	---	--

	lose balance	he came contact
	lose control	he ended up
	lost balance	he inadvertently actuated
	lost consciousness	he inadvertently bumped
	lost control	he let off
	lost sight	he lowered it
	manually positioning	he made sharp
	missed step	he moved board
	operator raised	he reached left
	passed out	he turned fell
	pulled off	he went back
	severe dehydration	he went off
	shocked fell	he went put
	slipped fell	heat stress symptoms.
	slipped off	heat stress symptoms.
	slipped on	helping another employee
	standing above	his scissor lift
	standing besides	hole his safety
	standing by	injured employee on
	standing close	leaned against it
	standing next	legs became wobbly
	standing on	loose his balance
	standing underneath	Lose his balance
	started getting	lost her balance
	started having	lost her footing
	stepped away	lost her grip
	stepped back	lost his balance
	stepped back	lost his footing
	stepped between	lost his grip
	stepped down	lost their footing
	stepped in	lost their grip
	stepped into	lost his vision
	stepped off	moving rolling scaffold

		stepped on stepped onto stepped over stepping off stepping on stepping out stepping over suffered cramps tripped dropped tripped fell tripped on tripped over truck reversing employee lifted	not feeling well not tied off not tied off operator not see operator extended outrigger operator lowered bucket operator moved trench operator not see pinned employee's thumb possible heat stress. pull himself up pulling de-energized conductor reached his hand removing floor hole removing flange off rested his hand rolled his ankle servicing head pulley forklift operator hit shoulder accidently struck slipped on ice slipped on mat started having cramps straining his groin suffered renal failure took phone call, trailer stairs removed truck operator latched turned his back turned over landed unclipped his belt unclipped his belt unseen by employees
--	--	--	---

			went grab it while doing hit work place violence. worker struck worker employee moved close stepped on loose employee stepped off
--	--	--	---



APPENDIX C

SPYDER PYTHON CODE FOR RULE-BASED TEXT MINING

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Sat Feb 29 15:21:31 2020
@author: heshanirupasinghe
"""

import os
os.chdir('/Users/heshanirupasinghe/Desktop/Machine Learning /Machine Learning A-Z
New/Part 7 - Natural Language Processing/Section 36 - Natural Language Processing')

import xlwt
from xlwt import Workbook
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv('nds.csv')
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
corpus = []

wb = Workbook(encoding = 'utf-8')
sheet1 = wb.add_sheet('Sheet 1')

text_file = open('nfuni.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
```

```

ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
unigrammf = set(review)

```

```

text_file = open('ofuni.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
unigramof = set(review)

```

```

text_file = open('sauni.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
unigramsa = set(review)

```

```

text_file = open('tfuni.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
unigramtf = set(review)

```

```

text_file = open('wfuni.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
unigramwf = set(review)

```

```

text_file = open('nfbi.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
bigramnf = list(nltk.bigrams(review))
bigramnf_set=set(bigramnf)

```

```

text_file = open('ofbi.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
bigramof = list(nltk.bigrams(review))
bigramof_set=set(bigramof)

```

```

text_file = open('sabi.txt', 'r')

```

```

read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
bigramsa = list(nltk.bigrams(review))
bigramsa_set=set(bigramsa)

```

```

text_file = open('tfbi.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
bigramtf = list(nltk.bigrams(review))
bigramtf_set=set(bigramtf)

```

```

text_file = open('wfbi.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
bigramwf = list(nltk.bigrams(review))
bigramwf_set=set(bigramwf)

```

```

text_file = open('nftri.txt', 'r')
read_file = text_file.read()

```

```

word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
trigramnf = list(nltk.trigrams(review))
trigramnf_set=set(trigramnf)

```

```

text_file = open('oftri.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
trigramof = list(nltk.trigrams(review))
trigramof_set=set(trigramof)

```

```

text_file = open('satri.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
trigramsa = list(nltk.trigrams(review))
trigramsa_set=set(trigramsa)

```

```

text_file = open('tftri.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)

```



```

review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
trigramtf = list(nltk.trigrams(review))
trigramtf_set=set(trigramtf)

```

```

text_file = open('wftri.txt', 'r')
read_file = text_file.read()
word_list = re.sub('[^a-zA-Z]', ' ', read_file)
review = word_list.lower()
word_list = review.split()
ps = PorterStemmer()
review = [ps.stem(word) for word in word_list if not word in
set(stopwords.words('english'))]
trigramwf = list(nltk.trigrams(review))
trigramwf_set=set(trigramwf)

```

```

for i in range(0, 7417):
    review = re.sub('[^a-zA-Z]', ' ', dataset['Final Narrative'][i])
    review = review.lower()
    review = review.split()
    ps = PorterStemmer()
    review = [ps.stem(word) for word in review if not word in
set(stopwords.words('english'))]
    trigram=list(nltk.trigrams(review))
    trigram_set=set(trigram)
    bigram= list(nltk.bigrams(review))
    bigram_set= set(bigram)
    unigram=set(review)
    if (('attempt') in unigram):
        sheet1.write(i, 1, 'attempt')
        sheet1.write(i, 0, dataset['Final Narrative'][i])

```

```

sheet1.write(i, 2, 'Worker Factor')
elif any(check in bigram for check in bigramnf) :
    c = list(bigram_set & bigramnf_set)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Natural Factor')
elif (('he', 'slip') in bigram):
    sheet1.write(i, 1, ('he', 'slip'))
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Worker Factor')
elif (('employe', 'slip') in bigram):
    sheet1.write(i, 1, ('employe', 'slip'))
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Worker Factor')
elif any(check in trigram for check in trigramnf) :
    c = list(trigram_set & trigramnf_set)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Natural Factor')
elif any(check in trigram for check in trigramof):
    c = list(trigram_set & trigramof_set)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Organisational Factor')
elif any(check in trigram for check in trigramsa):
    c = list(trigram_set & trigramsa_set)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Surrounding Activity')
elif any(check in trigram for check in trigramtf):

```

```

c = list(trigram_set & trigramtf_set)
d= ‘.join(str(c))
sheet1.write(i, 1, d)
sheet1.write(i, 0, dataset[‘Final Narrative’][i])
sheet1.write(i, 2, ‘Technological Factor’)
elif any(check in trigram for check in trigramwf):
c = list(trigram_set & trigramwf_set)
d= ‘.join(str(c))
sheet1.write(i, 1, d)
sheet1.write(i, 0, dataset[‘Final Narrative’][i])
sheet1.write(i, 2, ‘Worker Factor’)
elif any(check in bigram for check in bigramof):
c = list(bigram_set & bigramof_set)
d= ‘.join(str(c))
sheet1.write(i, 1, d)
sheet1.write(i, 0, dataset[‘Final Narrative’][i])
sheet1.write(i, 2, ‘Organisational Factor’)
elif any(check in bigram for check in bigramsa):
c = list(bigram_set & bigramsa_set)
d= ‘.join(str(c))
sheet1.write(i, 1, d)
sheet1.write(i, 0, dataset[‘Final Narrative’][i])
sheet1.write(i, 2, ‘Surrounding Activity’)
elif any(check in bigram for check in bigramtf):
c = list(bigram_set & bigramtf_set)
d= ‘.join(str(c))
sheet1.write(i, 1, d)
sheet1.write(i, 0, dataset[‘Final Narrative’][i])
sheet1.write(i, 2, ‘Technological Factor’)
elif any(check in bigram for check in bigramwf):
c = list(bigram_set & bigramwf_set)
d= ‘.join(str(c))
sheet1.write(i, 1, d)
sheet1.write(i, 0, dataset[‘Final Narrative’][i])

```

```

sheet1.write(i, 2, 'Worker Factor')
elif any(check in unigram for check in unigramtf):
    c = list(unigram & unigramtf)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Technological Factor')
elif any(check in unigram for check in unigramof):
    c = list(unigram & unigramof)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Organisational Factor')
elif any(check in unigram for check in unigramsa):
    c = list(unigram & unigramsa)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Surrounding Activity')
elif any(check in unigram for check in unigramwf):
    c = list(unigram & unigramwf)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Worker Factor')
elif any(check in unigram for check in unigramnf):
    c = list(unigram & unigramnf)
    d=' '.join(str(c))
    sheet1.write(i, 1, d)
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Natural Factor')
else:
    sheet1.write(i, 0, dataset['Final Narrative'][i])
    sheet1.write(i, 2, 'Null')

```

```
wb.save('classification outputs.xls')
```



APPENDIX D

SOURCES OF HAZARD SEPARATION FOR VALIDATION

```

Sub Button4_Click()
For i = 2 To 7418
word = Cells(i, 18).Value
  If word = "Natural Factor" Then
    Worksheets("DATA").Rows(i).Copy
    Worksheets("NF").Activate
    b = Worksheets("NF").Cells(Rows.Count, 1).End(xlUp).Row
    Worksheets("NF").Cells(b + 1, 1).Select
    ActiveSheet.Paste
    Worksheets("DATA").Activate
  Else
  If word = "Null" Then
    Worksheets("DATA").Rows(i).Copy
    Worksheets("NULL").Activate
    b = Worksheets("NULL").Cells(Rows.Count, 1).End(xlUp).Row
    Worksheets("NULL").Cells(b + 1, 1).Select
    ActiveSheet.Paste
    Worksheets("DATA").Activate
  Else
  If word = "Organisational Factor" Then
    Worksheets("DATA").Rows(i).Copy
    Worksheets("OF").Activate
    b = Worksheets("OF").Cells(Rows.Count, 1).End(xlUp).Row
    Worksheets("OF").Cells(b + 1, 1).Select
    ActiveSheet.Paste
    Worksheets("DATA").Activate
  Else
  If word = "Surrounding Activity" Then
    Worksheets("DATA").Rows(i).Copy
    Worksheets("SA").Activate
    b = Worksheets("SA").Cells(Rows.Count, 1).End(xlUp).Row

```

```
Worksheets("SA").Cells(b + 1, 1).Select
ActiveSheet.Paste
Worksheets("DATA").Activate
Else
If word = "Technological Factor" Then
Worksheets("DATA").Rows(i).Copy
Worksheets("TF").Activate
b = Worksheets("TF").Cells(Rows.Count, 1).End(xlUp).Row
Worksheets("TF").Cells(b + 1, 1).Select
ActiveSheet.Paste
Worksheets("DATA").Activate
Else
If word = "Worker Factor" Then

Worksheets("DATA").Rows(i).Copy
Worksheets("WF").Activate
b = Worksheets("WF").Cells(Rows.Count, 1).End(xlUp).Row
Worksheets("WF").Cells(b + 1, 1).Select
ActiveSheet.Paste
Worksheets("DATA").Activate
End If
End If
End If
End If
End If
End If
Next
End Sub
```

APPENDIX E

CLASSIFIER TRAINING FOR ORGANISATIONAL FATORS

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Tue Aug 27 10:35:26 2019

@author: heshanirupasinghe
"""

# Natural Language Processing
import os
os.chdir('/Users/heshanirupasinghe/Desktop/Machine Learning /Machine Learning A-Z
New/Part 7 - Natural Language Processing/Section 36 - Natural Language Processing/Test
files')

# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
# Importing the dataset
#dataset = pd.read_csv('nd.csv', delimiter = '\t', quoting = 3)
dataset = pd.read_csv('hof.csv')
#dataset = pd.read_csv('nd.csv')
#dataset = dataset0.iloc[:, :-1].values
#y = dataset.iloc[:, 3].values
# Cleaning the texts
"Cleaned version of the review is called as 'review' here."
import re
import nltk # Library for removing irrelevant words
#nltk.download('stopwords') #downlord the stop word list
from nltk.corpus import stopwords
```



```

from nltk.stem.porter import PorterStemmer # Class for stemming
corpus = [] #Corpus is a collection of text.

# Delete row at index 0
#dataset = np.delete( dataset[21:22], dataset[21:22])
#text = dataset['Final Narrative'][2]
#pm.get_phrases(text) phrase machine only takes the noun phrases.

for i in range(0, 188):
    review = re.sub('[^a-zA-Z]', ' ', dataset['Phrase'][i])
    review = review.lower()
    review = review.split()
    ps = PorterStemmer() #Object creation of the stemming class
    review = [ps.stem(word) for word in review if not word in set(stopwords.words('english'))]
    review = ' '.join(review)
    corpus.append(review)# appending all the reviews to corpus

# Creating the Bag of Words model
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
X = cv.fit_transform(corpus).toarray()# These are like independent ariables
y = dataset.iloc[:, 3].values # These are like dependent variables

#Feature scaling is not needed.

#Label encoding/ Encoding independent variable
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
#labelencoder_X = LabelEncoder()
"""X[:, 0] = labelencoder_X.fit_transform(X[:, 0])
onehotencoder = OneHotEncoder(categorical_features = [0])
X = onehotencoder.fit_transform(X).toarray()"""
# Encoding the Dependent Variable
labelencoder_y = LabelEncoder()

```

```
y = labelencoder_y.fit_transform(y)

#a= [y,y1]

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)

# Fitting Naive Bayes to the Training set
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, y_train)

# Fitting SVM to the Training set
from sklearn.svm import SVC
classifier= SVC(kernel='linear', random_state=0)
classifier.fit(X_train, y_train)

# Fitting Random forest to the Training set
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=1000,criterion='entropy', random_state=0)
classifier.fit(X_train, y_train)

# Fitting K nearest neighbours to the Training set
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors =5,p=2, metric ='minkowski')
classifier.fit(X_train, y_train)

# Fitting Kernel SVM to the Training set
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0)
classifier.fit(X_train, y_train)
```

```
# Fitting Decision Tree Classification to the Training set
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)

# Predicting the Test set results
y_pred = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)

from sklearn.metrics import f1_score
f1_score(y_test,y_pred, average='micro', sample_weight=None)

#plt.legend(loc='upper left', bbox_to_anchor=(1,1))

''' (TP = # True Positives, TN = # True Negatives, FP = # False Positives, FN = # False
Negatives):

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 Score = 2 * Precision * Recall / (Precision + Recall)'''
```

APPENDIX F
CRITICAL VALUES OF CHI-SQUARE DISTRIBUTION

df	p	
	0.05	0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.97	40.29
23	35.17	41.64
24	36.42	42.98

df	p	
	0.05	0.01
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89
35	49.80	57.34
40	55.76	63.69
45	61.66	69.96
50	67.50	76.15
60	79.08	88.38
70	90.53	100.43
80	101.58	112.33
90	113.15	124.12
100	124.34	135.81
200	233.99	249.45
300	341.40	359.91
400	447.63	468.72
500	553.13	576.49
600	658.09	683.52
700	762.66	789.97
800	866.91	895.98
900	970.90	1001.63
1000	1074.68	1106.97

APPENDIX G

MULTIPLE COMPARISONS OF TOTAL NUMBER OF ACCIDENTS

OCCURRED IN EACH MONTH

Test Procedure: Games-Howell						
(I) Month	(J) Month	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
January	February	6.0000	10.1676	1.000	-39.861	51.861
	March	-16.4000	7.1063	.537	-48.570	15.770
	April	-16.4000	8.9107	.766	-55.520	22.720
	May	-14.8000	11.4952	.955	-68.419	38.819
	June	-42.2000	12.4218	.193	-101.426	17.026
	July	-63.0000*	6.7882	.001	-94.701	-31.299
	August	-50.4000	11.4499	.065	-103.748	2.948
	September	-30.8000	12.5714	.483	-90.941	29.341
	October	-32.8000	10.5075	.269	-85.111	19.511
	November	-2.0500	9.3347	1.000	-46.332	42.232
	December	11.4500	10.0151	.976	-37.406	60.306
February	January	-6.0000	10.1676	1.000	-51.861	39.861
	March	-22.4000	9.2412	.497	-67.398	22.598
	April	-22.4000	10.6911	.639	-69.672	24.872
	May	-20.8000	12.9244	.866	-77.714	36.114
	June	-48.2000	13.7550	.149	-109.513	13.113
	July	-69.0000*	8.9989	.008	-114.303	-23.697
	August	-56.4000	12.8841	.051	-113.108	.308
	September	-36.8000	13.8903	.385	-98.854	25.254
	October	-38.8000	12.0543	.219	-94.288	16.688
	November	-8.0500	11.0470	.999	-58.290	42.190
	December	5.4500	11.6277	1.000	-47.682	58.582
March	January	16.4000	7.1063	.537	-15.770	48.570
	February	22.4000	9.2412	.497	-22.598	67.398
	April	.0000	7.8371	1.000	-36.445	36.445

	May	1.6000	10.6846	1.000	-52.353	55.553
	June	-25.8000	11.6756	.596	-85.915	34.315
	July	-46.6000*	5.3009	.001	-69.911	-23.289
	August	-34.0000	10.6358	.255	-87.650	19.650
	September	-14.4000	11.8347	.963	-75.503	46.703
	October	-16.4000	9.6139	.812	-70.209	37.409
	November	14.3500	8.3160	.807	-29.589	58.289
	November	27.8500	9.0732	.309	-21.822	77.522
April	January	16.4000	8.9107	.766	-22.720	55.520
	February	22.4000	10.6911	.639	-24.872	69.672
	March	.0000	7.8371	1.000	-36.445	36.445
	May	1.6000	11.9608	1.000	-52.686	55.886
	June	-25.8000	12.8538	.684	-85.305	33.705
	July	-46.6000*	7.5498	.015	-82.929	-10.271
	August	-34.0000	11.9172	.316	-88.037	20.037
	September	-14.4000	12.9985	.982	-74.765	45.965
	October	-16.4000	11.0148	.900	-69.198	36.398
	November	14.3500	9.9023	.916	-31.422	60.122
	December	27.8500	10.5462	.403	-21.882	77.582
	May	January	14.8000	11.4952	.955	-38.819
February		20.8000	12.9244	.866	-36.114	77.714
March		-1.6000	10.6846	1.000	-55.553	52.353
April		-1.6000	11.9608	1.000	-55.886	52.686
June		-27.4000	14.7635	.759	-92.056	37.256
July		-48.2000	10.4757	.080	-102.680	6.280
August		-35.6000	13.9556	.420	-96.540	25.340
September		-16.0000	14.8896	.987	-81.270	49.270
October		-18.0000	13.1934	.940	-78.006	42.006
November		12.7500	12.2799	.989	-43.545	69.045
December		26.2500	12.8048	.663	-32.033	84.533
June		January	42.2000	12.4218	.193	-17.026
	February	48.2000	13.7550	.149	-13.113	109.513
	March	25.8000	11.6756	.596	-34.315	85.915

	April	25.8000	12.8538	.684	-33.705	85.305
	May	27.4000	14.7635	.759	-37.256	92.056
	July	-20.8000	11.4848	.771	-81.525	39.925
	August	-8.2000	14.7282	1.000	-72.722	56.322
	September	11.4000	15.6160	.999	-56.794	79.594
	October	9.4000	14.0081	1.000	-54.442	73.242
	November	40.1500	13.1513	.264	-20.864	101.164
	December	53.6500	13.6427	.100	-8.840	116.140
July	January	63.0000*	6.7882	.001	31.299	94.701
	February	69.0000*	8.9989	.008	23.697	114.303
	March	46.6000*	5.3009	.001	23.289	69.911
	April	46.6000*	7.5498	.015	10.271	82.929
	May	48.2000	10.4757	.080	-6.280	102.680
	June	20.8000	11.4848	.771	-39.925	81.525
	August	12.6000	10.4259	.964	-41.571	66.771
	September	32.2000	11.6465	.388	-29.522	93.922
	October	30.2000	9.3812	.292	-24.776	85.176
	November	60.9500*	8.0459	.016	16.244	105.656
	December	74.4500*	8.8262	.012	23.748	125.152
August	January	50.4000	11.4499	.065	-2.948	103.748
	February	56.4000	12.8841	.051	-.308	113.108
	March	34.0000	10.6358	.255	-19.650	87.650
	April	34.0000	11.9172	.316	-20.037	88.037
	May	35.6000	13.9556	.420	-25.340	96.540
	June	8.2000	14.7282	1.000	-56.322	72.722
	July	-12.6000	10.4259	.964	-66.771	41.571
	September	19.6000	14.8546	.952	-45.540	84.740
	October	17.6000	13.1540	.946	-42.231	77.431
	November	48.3500	12.2375	.098	-7.723	104.423
	December	61.8500*	12.7641	.037	3.762	119.938
September	January	30.8000	12.5714	.483	-29.341	90.941
	February	36.8000	13.8903	.385	-25.254	98.854
	March	14.4000	11.8347	.963	-46.703	75.503

	April	14.4000	12.9985	.982	-45.965	74.765
	May	16.0000	14.8896	.987	-49.270	81.270
	June	-11.4000	15.6160	.999	-79.594	56.794
	July	-32.2000	11.6465	.388	-93.922	29.522
	August	-19.6000	14.8546	.952	-84.740	45.540
	October	-2.0000	14.1410	1.000	-66.502	62.502
	November	28.7500	13.2927	.609	-33.054	90.554
	December	42.2500	13.7791	.256	-20.955	105.455
October	January	32.8000	10.5075	.269	-19.511	85.111
	February	38.8000	12.0543	.219	-16.688	94.288
	March	16.4000	9.6139	.812	-37.409	70.209
	April	16.4000	11.0148	.900	-36.398	69.198
	May	18.0000	13.1934	.940	-42.006	78.006
	June	-9.4000	14.0081	1.000	-73.242	54.442
	July	-30.2000	9.3812	.292	-85.176	24.776
	August	-17.6000	13.1540	.946	-77.431	42.231
	September	2.0000	14.1410	1.000	-62.502	66.502
	November	30.7500	11.3606	.386	-24.462	85.962
	December	44.2500	11.9260	.144	-13.110	101.610
November	January	2.0500	9.3347	1.000	-42.232	46.332
	February	8.0500	11.0470	.999	-42.190	58.290
	March	-14.3500	8.3160	.807	-58.289	29.589
	April	-14.3500	9.9023	.916	-60.122	31.422
	May	-12.7500	12.2799	.989	-69.045	43.545
	June	-40.1500	13.1513	.264	-101.164	20.864
	July	-60.9500*	8.0459	.016	-105.656	-16.244
	August	-48.3500	12.2375	.098	-104.423	7.723
	September	-28.7500	13.2927	.609	-90.554	33.054
	October	-30.7500	11.3606	.386	-85.962	24.462
	December	13.5000	10.9068	.962	-39.102	66.102
December	January	-11.4500	10.0151	.976	-60.306	37.406
	February	-5.4500	11.6277	1.000	-58.582	47.682
	March	-27.8500	9.0732	.309	-77.522	21.822

	April	-27.8500	10.5462	.403	-77.582	21.882
	May	-26.2500	12.8048	.663	-84.533	32.033
	June	-53.6500	13.6427	.100	-116.140	8.840
	July	-74.4500*	8.8262	.012	-125.152	-23.748
	August	-61.8500*	12.7641	.037	-119.938	-3.762
	September	-42.2500	13.7791	.256	-105.455	20.955
	October	-44.2500	11.9260	.144	-101.610	13.110
	November	-13.5000	10.9068	.962	-66.102	39.102

*. The mean difference is significant at the 0.05 level.



APPENDIX H
MULTIPLE COMPARISONS OF TOTAL NUMBER OF ACCIDENTS
PER 50 BILLION DOLLARS

(I) Month	(J) Month	Mean Difference (I-J)	Std. Error	Sig.
January	February	4.98	5.30	1.00
	March	0.54	4.93	1.00
	April	5.49	4.10	0.94
	May	11.07	4.39	0.44
	June	2.64	5.60	1.00
	July	-5.73	4.12	0.93
	August	1.04	4.79	1.00
	September	7.71	6.49	0.97
	October	4.28	4.72	1.00
	November	14.13	4.42	0.22
	December	14.62	5.04	0.30
February	January	-4.98	5.30	1.00
	March	-4.44	5.18	1.00
	April	0.50	4.40	1.00
	May	6.09	4.67	0.95
	June	-2.34	5.83	1.00
	July	-10.71	4.42	0.50
	August	-3.94	5.05	1.00
	September	2.73	6.69	1.00
	October	-0.71	4.98	1.00
	November	9.14	4.70	0.71
	December	9.64	5.29	0.77
March	January	-0.54	4.93	1.00
	February	4.44	5.18	1.00
	April	4.94	3.94	0.96
	May	10.53	4.24	0.45

	June	2.10	5.49	1.00
	July	-6.27	3.96	0.87
	August	0.50	4.65	1.00
	September	7.17	6.39	0.98
	October	3.73	4.58	1.00
	November	13.59	4.27	0.22
	December	14.08	4.91	0.32
April	January	-5.49	4.10	0.94
	February	-0.50	4.40	1.00
	March	-4.94	3.94	0.96
	May	5.58	3.24	0.82
	June	-2.84	4.76	1.00
	July	-11.22	2.88	0.09
	August	-4.45	3.77	0.97
	September	2.23	5.78	1.00
	October	-1.21	3.68	1.00
	November	8.64	3.28	0.41
	December	9.13	4.08	0.59
	May	January	-11.07	4.39
February		-6.09	4.67	0.95
March		-10.53	4.24	0.45
April		-5.58	3.24	0.82
June		-8.43	5.01	0.83
July		-16.79*	3.27	0.02
August		-10.03	4.08	0.46
September		-3.36	5.99	1.00
October		-6.79	4.00	0.82
November		3.06	3.63	1.00
December		3.55	4.37	1.00
June		January	-2.64	5.60
	February	2.34	5.83	1.00
	March	-2.10	5.49	1.00

	April	2.84	4.76	1.00
	May	8.43	5.01	0.83
	July	-8.37	4.78	0.80
	August	-1.60	5.37	1.00
	September	5.07	6.93	1.00
	October	1.63	5.30	1.00
	November	11.48	5.03	0.56
	December	11.98	5.59	0.62
July	January	5.73	4.12	0.93
	February	10.71	4.42	0.50
	March	6.27	3.96	0.87
	April	11.22	2.88	0.09
	May	16.79*	3.27	0.02
	June	8.37	4.78	0.80
	August	6.77	3.79	0.79
	September	13.44	5.80	0.55
	October	10.01	3.70	0.40
	November	19.85*	3.31	0.02
	December	20.35	4.11	0.06
	August	January	-1.04	4.79
February		3.94	5.05	1.00
March		-0.50	4.65	1.00
April		4.45	3.77	0.97
May		10.03	4.08	0.46
June		1.60	5.37	1.00
July		-6.77	3.79	0.79
September		6.67	6.29	0.99
October		3.24	4.43	1.00
November		13.09	4.11	0.22
December		13.58	4.77	0.33
September	January	-7.71	6.49	0.97
	February	-2.73	6.69	1.00

	March	-7.17	6.39	0.98
	April	-2.23	5.78	1.00
	May	3.36	5.99	1.00
	June	-5.07	6.93	1.00
	July	-13.44	5.80	0.55
	August	-6.67	6.29	0.99
	October	-3.44	6.23	1.00
	November	6.41	6.01	0.98
	December	6.91	6.48	0.99
October	January	-4.28	4.72	1.00
	February	0.71	4.98	1.00
	March	-3.73	4.58	1.00
	April	1.21	3.68	1.00
	May	6.79	4.00	0.82
	June	-1.63	5.30	1.00
	July	-10.01	3.70	0.40
	August	-3.24	4.43	1.00
	September	3.44	6.23	1.00
	November	9.85	4.03	0.49
	December	10.34	4.70	0.60
	November	January	-14.13	4.42
February		-9.14	4.70	0.71
March		-13.59	4.27	0.22
April		-8.64	3.28	0.41
May		-3.06	3.63	1.00
June		-11.48	5.03	0.56
July		-19.85*	3.31	0.02
August		-13.09	4.11	0.22
September		-6.41	6.01	0.98
October		-9.85	4.03	0.49
December		0.49	4.40	1.00
December	January	-14.62	5.04	0.30

February	-9.64	5.29	0.77
March	-14.08	4.91	0.32
April	-9.13	4.08	0.59
May	-3.55	4.37	1.00
June	-11.98	5.59	0.62
July	-20.35	4.11	0.06
August	-13.58	4.77	0.33
September	-6.91	6.48	0.99
October	-10.34	4.70	0.60
November	-0.49	4.40	1.00



APPENDIX I

**MULTIPLE COMPARISONS OF TOTAL NUMBER OF ACCIDENTS
OCCURRED DUE TO EACH SOURCES OF HAZARD IN EACH
MONTH**

Dependent Variable	(I) Month	(J) Month	Mean Difference (I-J)	Std. Error	Sig.
Worker Factor	January	February	4.6000	5.6815	.998
		March	-2.2000	4.2000	1.000
		April	-4.0000	7.2650	1.000
		May	-5.4000	6.3937	.997
		June	-9.6000	6.9714	.929
		July	-23.4000*	3.5609	.005
		August	-22.4000	6.0481	.149
		September	-11.4000	6.0811	.746
		October	-9.9500	5.9584	.827
		November	2.3000	4.6573	1.000
		December	8.0500	4.7961	.829
	February	January	-4.6000	5.6815	.998
		March	-6.8000	6.0033	.980
		April	-8.6000	8.4368	.991
		May	-10.0000	7.6994	.957
		June	-14.2000	8.1854	.813
		July	-28.0000*	5.5749	.044
		August	-27.0000	7.4148	.121
		September	-16.0000	7.4418	.611
		October	-14.5500	7.3418	.696
		November	-2.3000	6.3317	1.000
		December	3.4500	6.4345	1.000
	March	January	2.2000	4.2000	1.000
		February	6.8000	6.0033	.980

		April	-1.8000	7.5193	1.000
		May	-3.2000	6.6813	1.000
		June	-7.4000	7.2360	.989
		July	-21.2000*	4.0546	.022
		August	-20.2000	6.3514	.232
		September	-9.2000	6.3828	.917
		October	-7.7500	6.2660	.959
		November	4.5000	5.0448	.996
		December	10.2500	5.1732	.697
	April	January	4.0000	7.2650	1.000
		February	8.6000	8.4368	.991
		March	1.8000	7.5193	1.000
		May	-1.4000	8.9320	1.000
		June	-5.6000	9.3541	1.000
		July	-19.4000	7.1819	.404
		August	-18.4000	8.6879	.628
		September	-7.4000	8.7109	.998
		October	-5.9500	8.6257	1.000
		November	6.3000	7.7840	.998
		December	12.0500	7.8678	.888
	May	January	5.4000	6.3937	.997
		February	10.0000	7.6994	.957
		March	3.2000	6.6813	1.000
		April	1.4000	8.9320	1.000
		June	-4.2000	8.6948	1.000
		July	-18.0000	6.2992	.346
		August	-17.0000	7.9737	.620
		September	-6.0000	7.9987	.999
		October	-4.5500	7.9059	1.000
		November	7.7000	6.9778	.983
		December	13.4500	7.0712	.735
	June	January	9.6000	6.9714	.929
		February	14.2000	8.1854	.813

		March	7.4000	7.2360	.989
		May	5.6000	9.3541	1.000
		April	4.2000	8.6948	1.000
		July	-13.8000	6.8848	.687
		August	-12.8000	8.4439	.899
		September	-1.8000	8.4676	1.000
		October	-.3500	8.3799	1.000
		November	11.9000	7.5107	.869
		December	17.6500	7.5975	.535
	July	January	23.4000*	3.5609	.005
		February	28.0000*	5.5749	.044
		March	21.2000*	4.0546	.022
		May	19.4000	7.1819	.404
		April	18.0000	6.2992	.346
		June	13.8000	6.8848	.687
		August	1.0000	5.9481	1.000
		September	12.0000	5.9816	.686
		October	13.4500	5.8568	.568
		November	25.7000*	4.5266	.032
		December	31.4500*	4.6693	.016
	August	January	22.4000	6.0481	.149
		February	27.0000	7.4148	.121
		March	20.2000	6.3514	.232
		May	18.4000	8.6879	.628
		April	17.0000	7.9737	.620
		June	12.8000	8.4439	.899
		July	-1.0000	5.9481	1.000
		September	11.0000	7.7253	.928
		October	12.4500	7.6291	.854
		November	24.7000	6.6626	.129
		December	30.4500	6.7604	.054
	September	January	11.4000	6.0811	.746
		February	16.0000	7.4418	.611

		March	9.2000	6.3828	.917
		May	7.4000	8.7109	.998
		April	6.0000	7.9987	.999
		June	1.8000	8.4676	1.000
		July	-12.0000	5.9816	.686
		August	-11.0000	7.7253	.928
		October	1.4500	7.6552	1.000
		November	13.7000	6.6925	.665
		December	19.4500	6.7899	.316
	October	January	9.9500	5.9584	.827
		February	14.5500	7.3418	.696
		March	7.7500	6.2660	.959
		May	5.9500	8.6257	1.000
		April	4.5500	7.9059	1.000
		June	.3500	8.3799	1.000
		July	-13.4500	5.8568	.568
		August	-12.4500	7.6291	.854
		September	-1.4500	7.6552	1.000
		November	12.2500	6.5812	.751
		December	18.0000	6.6802	.395
	November	January	-2.3000	4.6573	1.000
		February	2.3000	6.3317	1.000
		March	-4.5000	5.0448	.996
		May	-6.3000	7.7840	.998
		April	-7.7000	6.9778	.983
		June	-11.9000	7.5107	.869
		July	-25.7000*	4.5266	.032
		August	-24.7000	6.6626	.129
		September	-13.7000	6.6925	.665
		October	-12.2500	6.5812	.751
		December	5.7500	5.5509	.988
	December	January	-8.0500	4.7961	.829
		February	-3.4500	6.4345	1.000

		March	-10.2500	5.1732	.697
		May	-12.0500	7.8678	.888
		April	-13.4500	7.0712	.735
		June	-17.6500	7.5975	.535
		July	-31.4500*	4.6693	.016
		August	-30.4500	6.7604	.054
		September	-19.4500	6.7899	.316
		October	-18.0000	6.6802	.395
		November	-5.7500	5.5509	.988
Technological Factor	January	February	3.2000	2.8107	.976
		March	-2.6000	3.3615	.998
		May	-4.6000	2.6833	.814
		April	-2.4000	2.0445	.974
		June	-14.6000	3.6194	.133
		July	-14.2000	4.1641	.231
		August	-8.8000	3.6194	.507
		September	-10.8000	4.9396	.611
		October	-3.9500	1.8804	.642
		November	4.3000	2.8089	.874
		December	5.8000	2.3749	.507
		February	January	-3.2000	2.8107
	March		-5.8000	4.1134	.930
	April		-7.8000	3.5805	.597
	May		-5.6000	3.1305	.788
	June		-17.8000	4.3267	.074
	July		-17.4000	4.7917	.139
	August		-12.0000	4.3267	.339
	September		-14.0000	5.4791	.438
	October		-7.1500	3.0259	.518
	November		1.1000	3.6756	1.000
	December		2.6000	3.3556	.999
	March	January	2.6000	3.3615	.998
February		5.8000	4.1134	.930	

		April	-2.0000	4.0274	1.000
		May	.2000	3.6332	1.000
		June	-12.0000	4.7032	.420
		July	-11.6000	5.1342	.558
		August	-6.2000	4.7032	.953
		September	-8.2000	5.7810	.925
		October	-1.3500	3.5434	1.000
		November	6.9000	4.1122	.836
		December	8.4000	3.8288	.594
	April	January	4.6000	2.6833	.814
		February	7.8000	3.5805	.597
		March	2.0000	4.0274	1.000
		May	2.2000	3.0166	.999
		June	-10.0000	4.2450	.514
		July	-9.6000	4.7181	.671
		August	-4.2000	4.2450	.993
		September	-6.2000	5.4148	.977
		October	.6500	2.9079	1.000
		November	8.9000	3.5791	.460
		December	10.4000	3.2496	.219
	May	January	2.4000	2.0445	.974
		February	5.6000	3.1305	.788
		March	-.2000	3.6332	1.000
		April	-2.2000	3.0166	.999
		June	-12.2000	3.8730	.252
		July	-11.8000	4.3863	.399
		August	-6.4000	3.8730	.841
		September	-8.4000	5.1284	.843
		October	-1.5500	2.3315	1.000
		November	6.7000	3.1289	.624
		December	8.2000	2.7459	.288
	June	January	14.6000	3.6194	.133
		February	17.8000	4.3267	.074

		March	12.0000	4.7032	.420
		May	10.0000	4.2450	.514
		April	12.2000	3.8730	.252
		July	.4000	5.3066	1.000
		August	5.8000	4.8908	.976
		September	3.8000	5.9346	1.000
		October	10.6500	3.7889	.358
		November	18.9000	4.3255	.062
		December	20.4000*	4.0571	.036
	July	January	14.2000	4.1641	.231
		February	17.4000	4.7917	.139
		March	11.6000	5.1342	.558
		May	9.6000	4.7181	.671
		April	11.8000	4.3863	.399
		June	-.4000	5.3066	1.000
		August	5.4000	5.3066	.991
		September	3.4000	6.2817	1.000
		October	10.2500	4.3123	.524
		November	18.5000	4.7906	.113
		December	20.0000	4.5497	.075
	August	January	8.8000	3.6194	.507
		February	12.0000	4.3267	.339
		March	6.2000	4.7032	.953
		May	4.2000	4.2450	.993
		April	6.4000	3.8730	.841
		June	-5.8000	4.8908	.976
		July	-5.4000	5.3066	.991
		September	-2.0000	5.9346	1.000
		October	4.8500	3.7889	.952
		November	13.1000	4.3255	.266
		December	14.6000	4.0571	.152
	September	January	10.8000	4.9396	.611
		February	14.0000	5.4791	.438

		March	8.2000	5.7810	.925
		May	6.2000	5.4148	.977
		April	8.4000	5.1284	.843
		June	-3.8000	5.9346	1.000
		July	-3.4000	6.2817	1.000
		August	2.0000	5.9346	1.000
		October	6.8500	5.0652	.932
		November	15.1000	5.4781	.366
		December	16.6000	5.2688	.263
	October	January	3.9500	1.8804	.642
		February	7.1500	3.0259	.518
		March	1.3500	3.5434	1.000
		May	-.6500	2.9079	1.000
		April	1.5500	2.3315	1.000
		June	-10.6500	3.7889	.358
		July	-10.2500	4.3123	.524
		August	-4.8500	3.7889	.952
		September	-6.8500	5.0652	.932
		November	8.2500	3.0242	.397
		December	9.7500	2.6260	.154
	November	January	-4.3000	2.8089	.874
		February	-1.1000	3.6756	1.000
		March	-6.9000	4.1122	.836
		May	-8.9000	3.5791	.460
		April	-6.7000	3.1289	.624
		June	-18.9000	4.3255	.062
		July	-18.5000	4.7906	.113
		August	-13.1000	4.3255	.266
		September	-15.1000	5.4781	.366
		October	-8.2500	3.0242	.397
		December	1.5000	3.3541	1.000
	December	January	-5.8000	2.3749	.507
		February	-2.6000	3.3556	.999

		March	-8.4000	3.8288	.594
		May	-10.4000	3.2496	.219
		April	-8.2000	2.7459	.288
		June	-20.4000*	4.0571	.036
		July	-20.0000	4.5497	.075
		August	-14.6000	4.0571	.152
		September	-16.6000	5.2688	.263
		October	-9.7500	2.6260	.154
		November	-1.5000	3.3541	1.000



BIOGRAPHY

Name Ms. Neththi Kumara Appuhamilage Heshani Rupasinghe
Date of Birth October 19, 1992
Education 2017: Bachelor of Science in Engineering, Specialized in Civil Engineering, Faculty of Engineering, University of Ruhuna, Hapugala, Galle, Sri Lanka, 80000.
2020: Master of Science in Engineering and Technology (Civil Engineering), School of Civil Engineering and Technology, Sirindhorn International Institute of Technology, Thammasat University, Pthumthani, Thailand, 12000.

Publications

Rupasinghe, H., & Rengarasu, T. M. (2018, May). Development of Driving Cycles for Galle. In *2018 Moratuwa Engineering Research Conference (MERCOn)* (pp. 108-113). IEEE.
Rupasinghe, N.K.A.H., & Panuwatwanich, K., (2020). Extraction and analysis of construction safety hazard factors using open data. *IOP Conf. Ser.: Mater. Sci. Eng.* 849 012008