KE 5205
TEXT MINING
PROJECT REPORT

TEXT ANALYTICS ON CONSTRUCTION FATAL ACCIDENTS SUMMARY REPORT

TEAM MEMBERS

Huang FuXing	A0163461J
Huang YuBo	A0163463E
Low Kang Jiang	A0074752B
Ma Ben	A0078067W
Tey Peng Mok	A0163350N

Contents

1.0 EXECUTIVE SUMMARY

In Singapore, the top contributor for workplace fatalities remain as construction industry despite there are improvement in recent years. There are a lot of information generated by Fatality and Catastrophe Investigation Summary. The summaries can be used as a useful resource to identify the occupations, body parts and workplace activities that have a higher risk associated to fatal accidents.

By applying proper text mining techniques on summary report. A project managers and safety professionals can reduce the occurrence of similar accidents by taking appropriate measures to mitigate the identified risks.

Given two datasets of accidents cases with recorded summary reports. We come up with a solution to build a classification model using the *Malaysia Accident Cases* dataset with labelled causes to classify the causes of accident in *Osha Accident Cases* dataset with unlabelled causes. Then, we extracted the key information such as the risky occupations, body parts more prone to injure and activities conducted when a fatal accident happened from *Osha* dataset with classified causes.

This report aims to provide a detailed explanation and analysis on the text mining and information extraction process leading to answer 4 business problems as required by the project. The project starts with a cleaning stage in which the column name for two datasets were standardized to avoid ambiguity. Then we proceed to build a classification model using the labelled *Malaysia* dataset to classify the accident causes for *Osha* dataset. The classification model is built using the voting ensemble algorithm instead of a single classification model to improve the overall robustness of the model. The performance of the classification model built then be validated using the labelled test data.

We proceed to information extraction using the cleaned and labelled *Osha* dataset to extract the occupations, body parts and working activities associated to the fatal accidents. Several visualisation tools have been used in the project such as bar chart and word cloud to visualise the most common occupations, body parts and working activities when the fatal accidents happened.

Generally, we attempted each question with different information extraction approaches using different text mining techniques such as key words approach and natural language processing approach as the nature of the information extracted are different. We have tried on several customization methods to improve the overall accuracy of information extracted for each problem and will explain them in detailed in the reports.

In conclusion, by using proper classification, text mining techniques and information extraction approaches, project manager and safety professionals can extract insightful information from the accident summary for further decision-making process.

2.0 BUSINESS OBJECTIVE

2.1 CURRENT CHALLENGES

Construction accidents caused by poor construction safety have caused significant human suffering, affect project progress and costs. Eventually, a poor safety record will damage the reputation of the industry and companies involved hence it is always a top interest for a construction company to determine the factors that causing the happening of an accident.

In fact, the inspection finding, audit score and safety climate surveys proposed to forecast the safety performance and improve the safety risk controls were not justifiable and reliable as they are not created using rigorous methods.

Generally, many safety professionals do agree that there is a lot of insight can be extracted from the Fatality and Catastrophe Investigation Summary that was generated after an accident happened. These summaries can be used as a good resource to identify and analysed the factors such as occupations, body parts and activities conducted associated to the happening of an accident. However, given that there are thousands of summary reports generated for all the accidents that happened throughout the years, it is very challenging for experts to read through the summary reports one by one to extract the useful information.

Hence, the business objective of this project is to use text mining and analytics to enhance the overall process of extracting information from the accident summary report. The 5 problems that required to be answered for the project for *Osha* dataset are listed below:

1. What are the causes for each of the accidents of *Osha* dataset?
2. Which type of accidents in terms of causes are more common in fatal or catastrophic accidents?
3. Which type of occupations are riskier in such fatal or catastrophic accidents?
4. Which parts of human body are more prone to be injured in such fatal or catastrophic accidents?
5. What are the most common activities that the victims were engaged in prior to the accident?

2.2 PROPOSED SOLUTION

The text mining and information extraction of accidents summary reports has been carried out using the Natural Language Toolkit(NLKT) package available in Python. Other than that, several data analysis and data visualization package have also been used to conduct the necessary data preprocessing and data visualization work such as *pandas, numpy, matplotlib, seaborn*. etc. We also using *sklearn* library available in Python to conduct machine learning to build the classification model to classify the unlabeled accident causes in Osha dataset. One of the highlight for this project is we did using different text mining techniques to approach each problem.

3.0 TEXT ANALYTICS

3.1 DATA UNDERSTANDING AND PRE-PROCESSING

Dataset used:

'MsiaAccidentCases.csv' & 'osha.csv'

Dataset stored:

'MsiaAccidentCases_clean.csv' & 'osha_clean.csv'

Two datasets have been provided in this project, the *Malaysia Accident Cases(Malaysia)* and *Osha Accident Cases(Osha)* dataset. Before we start to solve the problem, we have conducted preliminary data analysis to understand the data structure and datatype of both datasets.

The main difference between the two datasets is the accidents in *Malaysia* dataset have been labelled with causes whereas the accidents in *Osha* dataset have not been labelled with causes. Hence, one of the business problem is to label the *Osha* dataset with accident causes.

Malaysia dataset contains 235 recorded cases with 3 attributes which are the case title, case description and accident cause. All the cases have been informed to be labelled with 11 accident causes. However, from the analysis we noticed there are 12 unique causes as there are two class of causes which are labelled with "*Other*" and "*Others*" but belongs to the same class. Hence, we modified the label for these two classes to standardize them to become a single unique class type. There are no missing data in *Malaysia* dataset.

Osha dataset contains 16323 recorded with 5 attributes which are accident ID, case title, case description, key terms 1 and key terms 2. The attributes of *Osha* dataset have not been labelled with column header. We rename the two important columns of both *Malaysia* and *Osha* dataset with "*title*" and "*description*" to avoid ambiguity in the later analysis. The cleaned and well labelled header *Malaysia* and *Osha* dataset has been stored and exported to be used in problem 0.

3.2 PROBLEM 0: WHAT ARE THE CAUSES FOR EACH OF THE ACCIDENTS OF OSHA DATASET?

Dataset used:

'MsiaAccidentCases_clean.csv' & 'osha_clean.csv'

Dataset stored:

'osha_clean_cause_labelled.csv'

As mentioned, the *Osha* dataset have not been labelled with accident causes. This mean that the information provided by original dataset are not sufficient to identify the most common cause of fatal accident in the later stage. As a result, we have built a classification model using the 235 accident cases in *Malaysia* dataset to classify the accident cause for 16323 accidents cases in *Osha* dataset for next stage analysis.

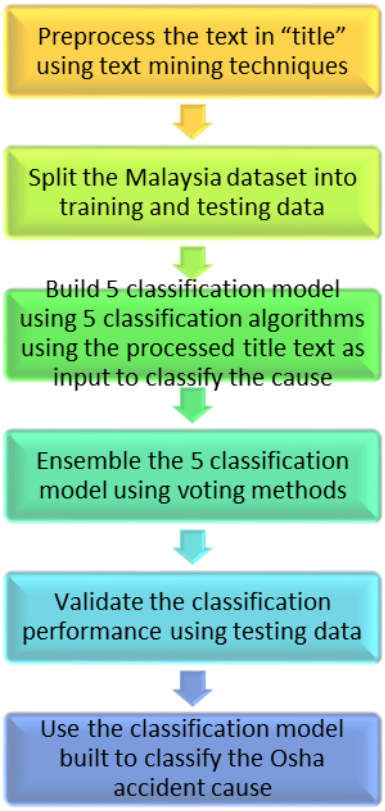

Figure 1: Overall classification model building process to classify accident cause in *Osha* dataset

Figure 1 show the overall classification model building process. We start the model building by processed the text in "*title*" and "*description*" using a steps by steps text mining process before feed into 5 classification algorithms. Then a voting ensemble model has been built using all the previous classification models built to improve the overall robustness of the model. The model then has been stored and have been be used to classify the cause of *Osha* dataset.

As mentioned, we conduct preliminary data explanatory analysis on *Malaysia* dataset to better understand the distribution of the accident cause in *Malaysia* dataset. Figure 2 shows the count plot of the number of accidents case by cause.

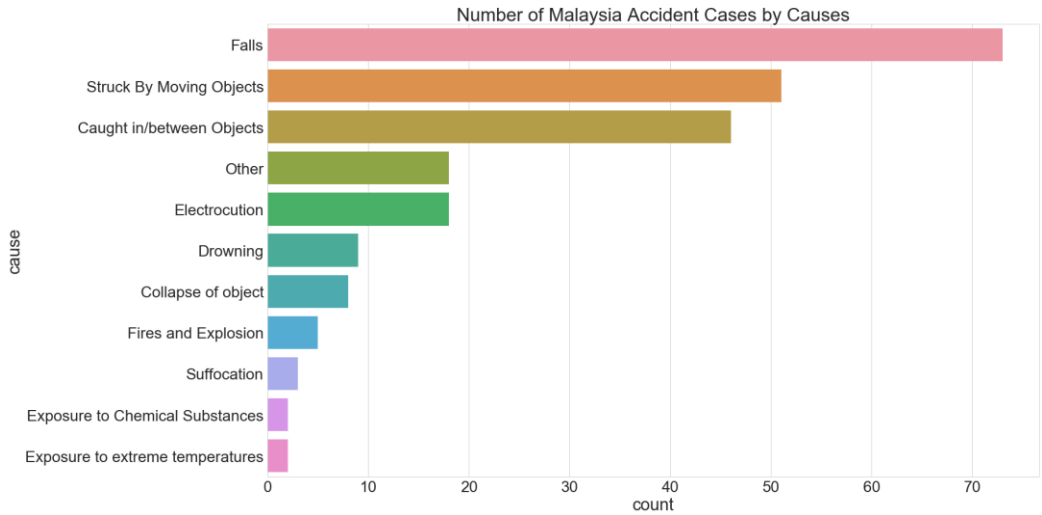

Figure 2: Count plot of the number of *Malaysia* accidents case by causes

As shows in Figure 2, there are 11 cause categories and each row in *Malaysia* dataset which represent a case has been labelled with a cause. Before we feed the text into the machine learning algorithms, we

conduct necessary text mining techniques to extract the bag of words to be fed into the classification algorithm.

```
# define a function for text mining with the following steps:
# 1. remove the non English words
# 2. tokenize the string for each row
# 3. remove punctiation
# 4. convert each of the token to lower case
# 5. remove stopwords
# 6. lemmatize each of the token
# 7. join the tokens back into string

mystopwords = stopwords.words("English")
wnlemma = nltk.WordNetLemmatizer()
def pre_process(text):
    text = re.sub(r'\d+', '', text)
    tokens = nltk.word_tokenize(text)
    tokens_nop = [word for word in tokens if word not in string.punctuation]
    tokens_lower = [ word.lower() for word in tokens_nop ]
    tokens_nostop = [word for word in tokens_lower if word not in mystopwords]
    tokens_lemma = [wnlemma.lemmatize(word) for word in tokens_nostop]
    text_after_process = " ".join(tokens_lemma)
    return(text_after_process)
```

Figure 3: Steps by steps text preprocessing of text in "*title*" and "*description*" columns

Figure 3 shows the steps by steps text mining process of how the text has been processed. Figure 4 shows the snapshot of the data-frame after text mining process. Noticed that after the process of text mining, the length of processed "*title*" and "*description*" text has been reduced significantly to only contain the important keywords that are useful to build the classification model.

cause	title	description	length_title	length_description	title_processed	description_processed	length_title_processed	length_description_processed
Caught in/between Objects	Died being caught in between machines	The accident occurred as victim was assigned t...	37	288	died caught machine	accident occurred victim assigned inspect main...	19	214
Other	Died been buried	The accident occurred during the floor concret...	16	173	died buried	accident occurred floor concreting work falsew...	11	117
Struck By Moving Objects	Died crushed by entrance arch	Victim with four co-workers were installing wo...	29	251	died crushed entrance arch	victim four co-worker installing wood plate in...	26	178

Figure 4: Snapshot of preprocessed dataframe of *Malaysia* dataset

To build the classification model, we split the *Malaysia* dataset into training and testing set with 7 to 3 ratios. The training set are used to build the classification models whereas the testing data are used to validate the accuracy of the classification models built. There are 5 classification algorithms used which are Naïve Bayes, Decision Tree, Support Vector Classifier, Random Forest and Neural Network.

Originally, we use both bag of words in "*title_preprocessed*" and "*description_preprocessed*" as an input to classify the causes. Table 1 and table 2 shows the classification performance by using the bag of words in '*description_preprocessed*' and '*title_preprocessed*' respectively. As shows, the accuracy of the models built using bag of words of "*title_preprocessed*" is significantly higher than the models built using bag of words of '*description_preprocessed*", this is due to the facts that "*title_preprocessed*" have captured the main cause keywords concisely. As expected, the overall accuracy of ensemble model is slightly better than each the individual classification model as shown in confusion matrix in Figure 5. As a result, we use the Voting Ensemble model built using processed '*title_preprocessed*' text as our finalized classification model.

Classification Algorithms	Precision	Recall	F1-Score
Naïve Bayes	0.49	0.52	0.45
Decision Tree	0.52	0.51	0.49
Support Vector Classifier	0.58	0.61	0.59
Random Forest	0.57	0.61	0.56
Neural Network	0.62	0.63	0.62
Voting Ensemble	0.57	0.62	0.57

Table 1: Classification Performance of models built by using processed '*description_preprocessed*' text

Classification Algorithms	Precision	Recall	F1-Score
Naïve Bayes	0.70	0.76	0.71
Decision Tree	0.72	0.75	0.73
Support Vector Classifier	0.80	0.77	0.77
Random Forest	0.74	0.77	0.75
Neural Network	0.80	0.79	0.78
Voting Ensemble	0.77	0.82	0.79

Table 2: Classification Performance of models built by using '*title_preprocessed*' text

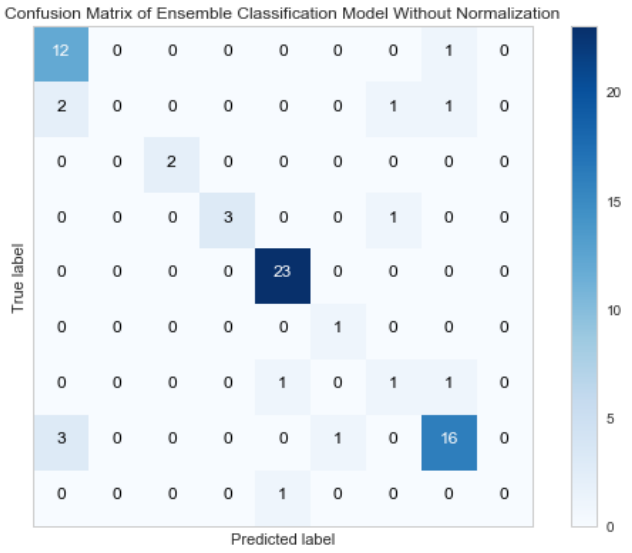

Figure 5: Confusion matrix of Voting Ensemble Classification Model

Figure 6 shows the snapshot of the data frame of *Osha* dataset with labelled cause classified using the classification model built using *Malaysia* dataset. As this is an unsupervised learning, we are unable to validate the accuracy using the testing data. However, an inspection of the first few cases indicated that the caused labelled are reliable. Figure 7 shows the count-plot of the labelled causes for *Osha* accident cases. The top 3 common causes of accidents in Osha dataset are "*Struck by Moving Object", "Caught in between Objects"* and *"Falls"*. The labelled *Osha* dataset has been stored and exported to be used in Problem 1.

.

title	description	length_title	length_description	title_processed	description_processed	length_title_processed	length_description_processed	cause
Two Workers Are Struck By Motor Vehicle And O...	On August 27 2013 Employees #1 and #2 of T...	59	777	two worker struck motor vehicle one killed	august employee templar inc. construction comp...	42	500	Struck By Moving Objects
Employee Is Struck By Bales Of Wire And Killed	On August 26 2013 Employee #1 with Lee Iro...	48	1552	employee struck bale wire killed	august employee lee iron metal company inc. us...	32	1040	Struck By Moving Objects
Employee Is Splashed With Hot Water And Is Bu...	On July 14 2013 Employee #1 vacuum pump tr...	51	1539	employee splashed hot water burned	july employee vacuum pump truck driver operato...	34	1007	Exposure to extreme temperatures

Figure 6: Snapshot of the data frame of *Osha* dataset with labelled cause classified by using the classification model built using *Malaysia* dataset

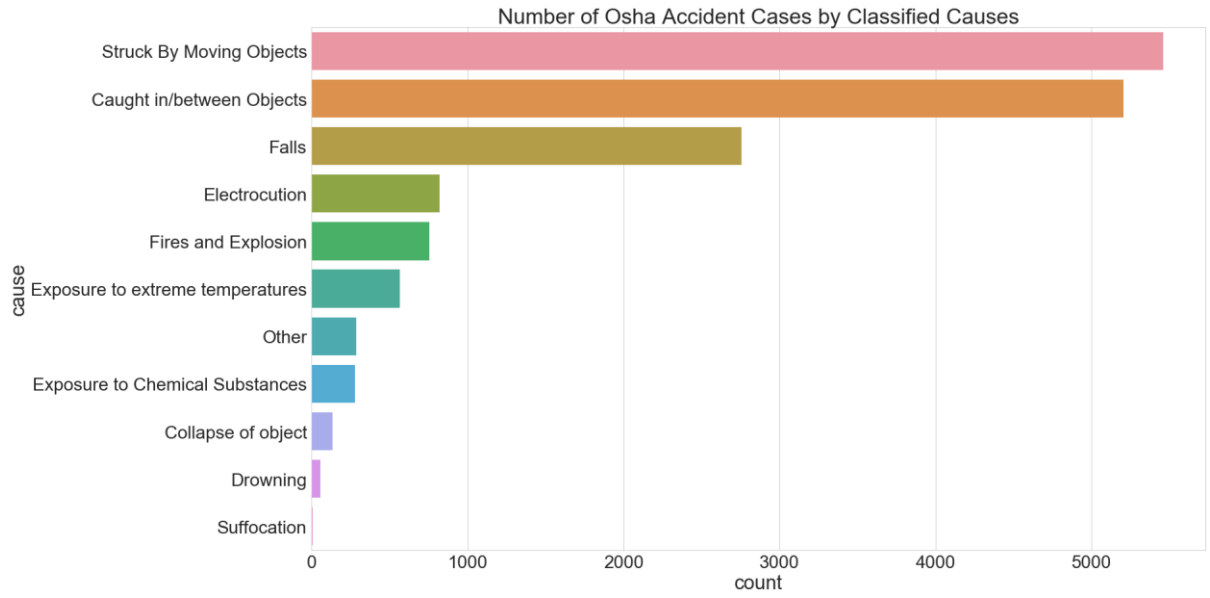

Figure 7: Count-plot of the labelled causes for *Osha* accident cases

3.3 PROBLEM 1: WHICH TYPE OF ACCIDENTS IN TERMS OF CAUSES ARE MORE COMMON IN FATAL OR CATASTROPHIC ACCIDENTS?

Dataset used:

'osha_clean_cause_labelled.csv'

Dataset stored:

'osha_fatal_clean_cause_labelled.csv'

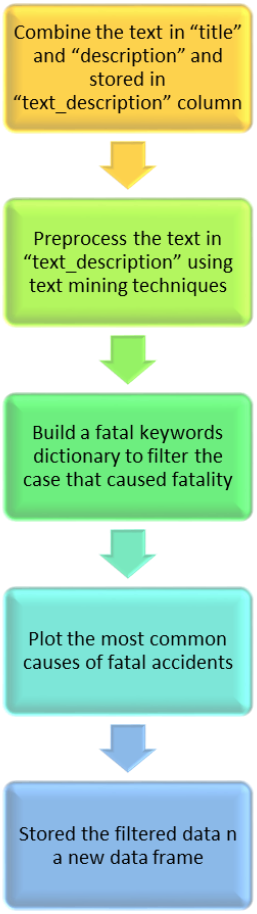

Figure 8: Steps by steps approaches to determine the most common causes that cause fatal accidents.

Figure 8 shows the steps how we approach to solve problem 1. We have built a fatal keywords dictionary to filter the *Osha* dataset with labelled cause that caused fatality. Each of the words in "*title*" and "*description*" are then be compared with each of the keywords in the dictionary to determine whether the case had caused fatality or not.

```
# define a dictionary with fatality keywords
fatal_keywords = ['kill', 'kills', 'killed', 'killing', \
                  'murder', 'murders', 'murdered', 'murdering', \
                  'die', 'dies', 'died', 'dying', 'dead', 'death', 'deadly', \
                  'fatal', 'fatally', 'fatality', 'lethal' \
                  'decease', 'deceases', 'deceased', 'deceasing' \
                  'lifeless', 'breathless', 'catastrophic', 'catastrophal', 'pass away', 'passed away']
```

Figure 9: Dictionary with fatality keywords that has been used to filter the fatal cases

Figure 9 shows the dictionary that we use to filter the fatal cases. We use various lemmatization form of keywords that is a synonym to the "*fatal*" keywords such as "*kill*", "*murder*", "*die*", and "*decease*".

As a result, 6568 cases out of 16323 *Osha* accidents case have been filtered out which contains the fatality keywords as shown in snapshot of data frame in Figure 10.

	incident_ID	title	description	cause	title_description	title_description_processed
0	202561825	Employee Falls From Flatbed Trailer And Later...	On August 30 2013 Employee #1 was working f...	Struck By Moving Objects	Employee Falls From Flatbed Trailer And Later...	employee fall flatbed trailer later dy august ...
1	200361855	Two Workers Are Struck By Motor Vehicle And O...	On August 27 2013 Employees #1 and #2 of T...	Struck By Moving Objects	Two Workers Are Struck By Motor Vehicle And O...	two worker struck motor vehicle one killed aug...
2	200361863	Employee Is Struck By Bales Of Wire And Killed	On August 26 2013 Employee #1 with Lee Iro...	Struck By Moving Objects	Employee Is Struck By Bales Of Wire And Kille...	employee struck bale wire killed august employ...
3	202673471	Foreman Is Fatally Crushed When Forklift Tips...	At approximately 6:30 a.m. on May 13 2013 E...	Caught in/between Objects	Foreman Is Fatally Crushed When Forklift Tips...	foreman fatally crushed forklift tip approxima...

Figure 10: Snapshot of data frame that contains only fatal accidents of *Osha* dataset

As shows in Figure 11, most of the cases that involved fatality are caused by the "*struck by moving objects*". This is the most common causes of fatality in *Osha* accident cases and there is a need for safety professionals to review the safety procedure on the task that involve heavy moving objects. The labelled fatal *Osha* dataset has been stored and exported to be used in Problem 2

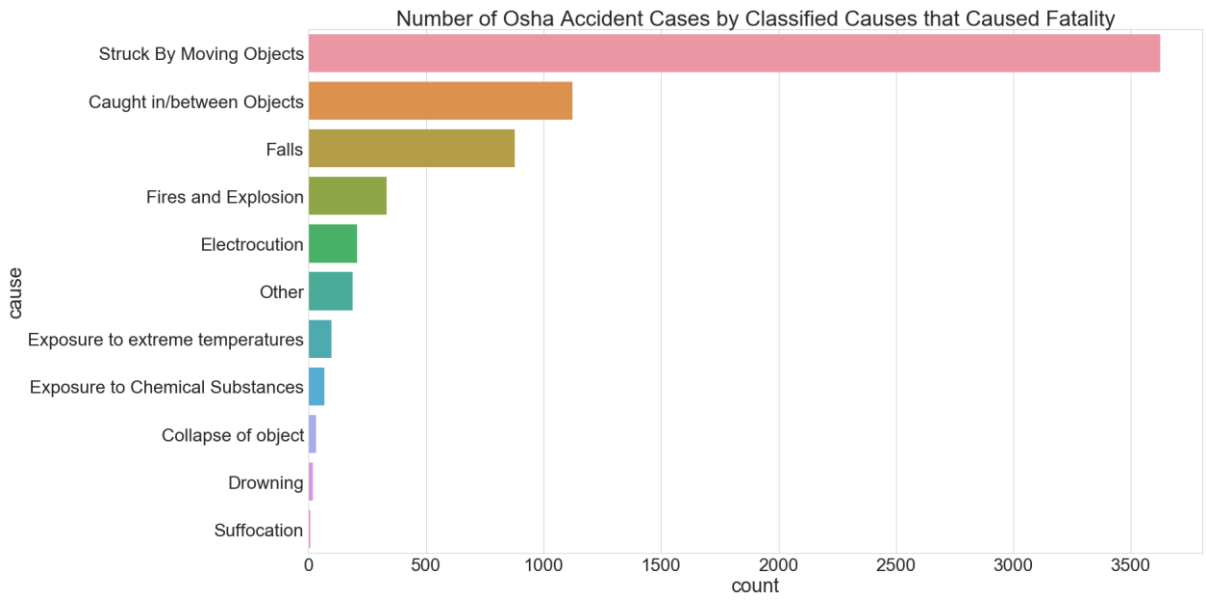

Figure 11: Number of *Osha* Accident cases by Classified Causes that Caused Fatality

3.4 PROBLEM 2: WHICH TYPE OF OCCUPATIONS ARE RISKIER IN SUCH FATAL OR CATASTROPHIC ACCIDENTS?

Dataset used:

'osha_fatal_clean_cause_labelled.csv' & "construction occupation name.txt"

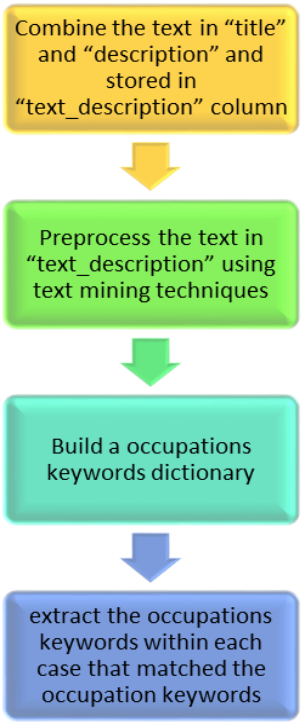

Figure 12: Steps to extract the occupations for each of fatal case in *Osha* dataset

Figure 12 shows the steps to extract the occupations keywords for each of fatal case in *Osha* dataset. We have built an occupations dictionary to extract the occupations keyword from stored *Osha* dataset with labelled causes that caused fatality.

The whole process starts with combine the text in "*title*" and "*description*" and store it as a new "*title_description*" column. Then, we define a *pre_process* function to conduct necessary text processing on the "*title_description*" column. The 7 preprocessing steps has been summarized as shown in Figure 3.

```
# define a function for occupations word extraction with the following steps:
# 1. remove the non English words
# 2. tokenize the string for each row
# 3. extract the unique occupation keyword
# 4. join the tokens back into string

occupations_dict_token = nltk.word_tokenize(occupations_dict.lower())

def extraction(text):
    text = re.sub(r'\d+', '', text)
    tokens = nltk.word_tokenize(text)
    tokens_occupation = [word for word in tokens if word in occupations_dict_token]
    text_after_process = " ".join(set(tokens_occupation))
    return(text_after_process)
```

Figure 13: Extraction function to extract the unique occupations within the text of each accident case

Other than the *pre_process* function, we also built an *extraction* function to extract the unique occupation that available in the predefined occupation dictionary as shown in Figure 13. The function comparing the processed "*title_description*" word of bags in each row of the data frame and extracting the unique occupations keywords that matched the dictionary to be stored in a new column. Figure 14 shows the resulted data frame with a new target column to store the extracted occupations name for each case.

incident_ID	title	description	cause	title_description	title_description_processed	occupation
202561825	Employee Falls From Flatbed Trailer And Later...	On August 30 2013 Employee #1 was working f...	Struck By Moving Objects	Employee Falls From Flatbed Trailer And Later...	employee fall flatbed trailer later dy august ...	
200361855	Two Workers Are Struck By Motor Vehicle And O...	On August 27 2013 Employees #1 and #2 of T...	Struck By Moving Objects	Two Workers Are Struck By Motor Vehicle And O...	two worker struck motor vehicle one killed aug...	worker
200361863	Employee Is Struck By Bales Of Wire And Killed	On August 26 2013 Employee #1 with Lee Iro...	Struck By Moving Objects	Employee Is Struck By Bales Of Wire And Kille...	employee struck bale wire killed august employ...	
202673471	Foreman Is Fatally Crushed When Forklift Tips...	At approximately 6:30 a.m. on May 13 2013 E...	Caught in/between Objects	Foreman Is Fatally Crushed When Forklift Tips...	foreman fatally crushed forklift tip approxima...	foreman

Figure 14: Resulted data frame with a new target column to store the extracted occupations name

The main challenge of this approach is some of the row do not contain any occupation key words as shown in Figure 14. The approach can be further enhanced by include the synonym of the occupations key words in the dictionary.

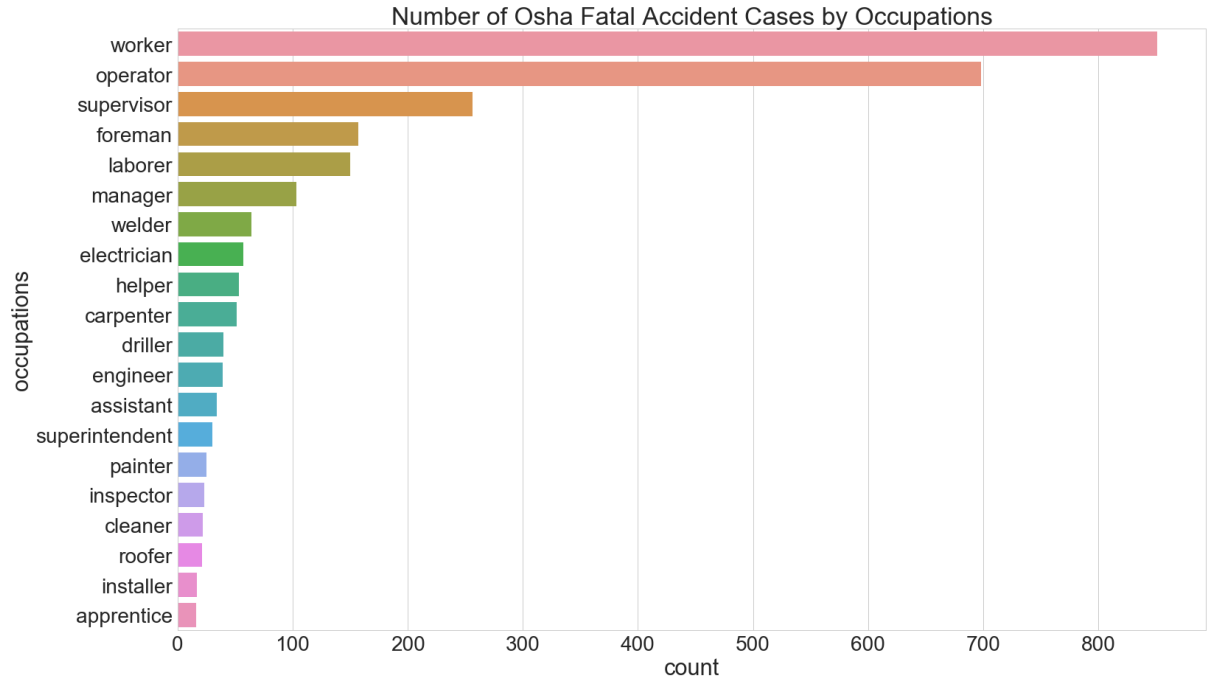

Figure 15: Top 10 risky occupations that are associated with fatal accidents

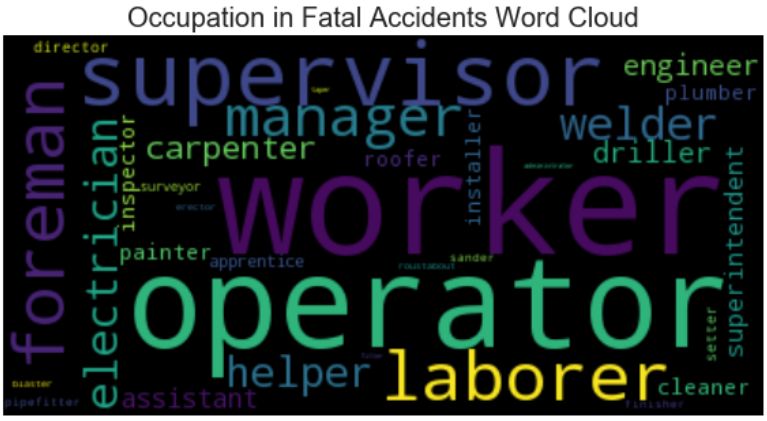

Figure 16: Word Cloud of Occupation Distribution of *Osha* fatal accidents

Figure 15 and Figure 16 shows the distribution of the top 10 risky occupations that are associated with fatal accidents. We have noticed that worker and operator are the most common occupation that associated with fatal accidents. Safety officer can use this info to review the safety workflow for operator to reduce the fatal cases.

3.5 PROBLEM 3: WHICH PARTS OF HUMAN BODY ARE MORE PRONE TO BE INJURED IN SUCH FATAL OR CATASTROPHIC ACCIDENTS?

Dataset used: 'osha_fatal_clean_cause_labelled.csv' & "body parts name.txt"

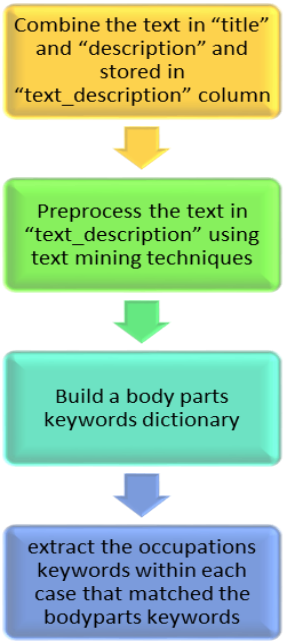

Figure 17: Steps to extract the associated body parts for each of fatal case in *Osha* dataset.

Figure 17 shows the steps to extract the associated body parts for each of fatal case in *Osha* dataset. We have built a body parts dictionary to extract the occupations keyword from stored *Osha* dataset with labelled causes that caused fatality.

The whole process starts with combine the text in "*title*" and "*description*" and store it as a new "*title_description*" column. Then, we define a *pre_process* function to conduct necessary text processing on the "*title_description*" column. The 7 preprocessing steps has been summarized as shown in Figure 3.

Other than the *pre_process* function, we also built an *extraction* function to extract the unique body parts that available in the predefined body parts dictionary. The function comparing the processed "*title_description*" word of bags in each row of the data frame and extracting the unique body parts keywords that matched the dictionary to be stored in a new column. Figure 18 shows the resulted data frame with a new target column to store the extracted body parts name for each case.

incident_ID	title	description	cause	title_description	title_description_processed	title_description_process	body_parts
202561825	Employee Falls From Flatbed Trailer And Later...	On August 30 2013 Employee #1 was working f...	Struck By Moving Objects	Employee Falls From Flatbed Trailer And Later...	employee fall flatbed trailer later dy august ...	employee fall flatbed trailer later dy august ...	abdomen
200361855	Two Workers Are Struck By Motor Vehicle And O...	On August 27 2013 Employees #1 and #2 of T...	Struck By Moving Objects	Two Workers Are Struck By Motor Vehicle And O...	two worker struck motor vehicle one killed aug...	two worker struck motor vehicle one killed aug...	
200361863	Employee Is Struck By Bales Of Wire And Killed	On August 26 2013 Employee #1 with Lee Iro...	Struck By Moving Objects	Employee Is Struck By Bales Of Wire And Kille...	employee struck bale wire killed august employ...	employee struck bale wire killed august employ...	leg back blood abdomen torso head face

Figure 18: Resulted data frame with a new target column to store the extracted body parts name

Figure 19 and 20 shows the distribution of the top 10 common body parts that associated with fatal accidents. As expected, head are the most common body parts that is prone to be injured in such fatal accidents. Safety officer can use this info to review the safety policy that better protect the head of construction worker

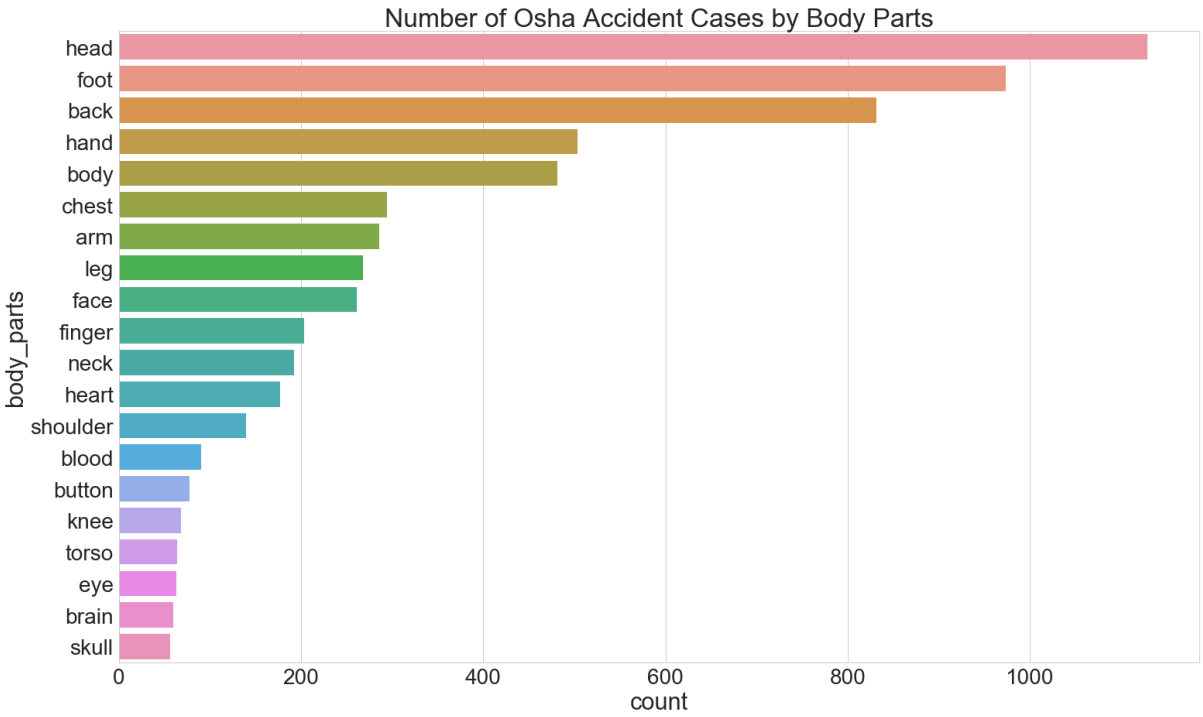

Figure 19: Distribution of the top 10 common body parts that associated with fatal accidents

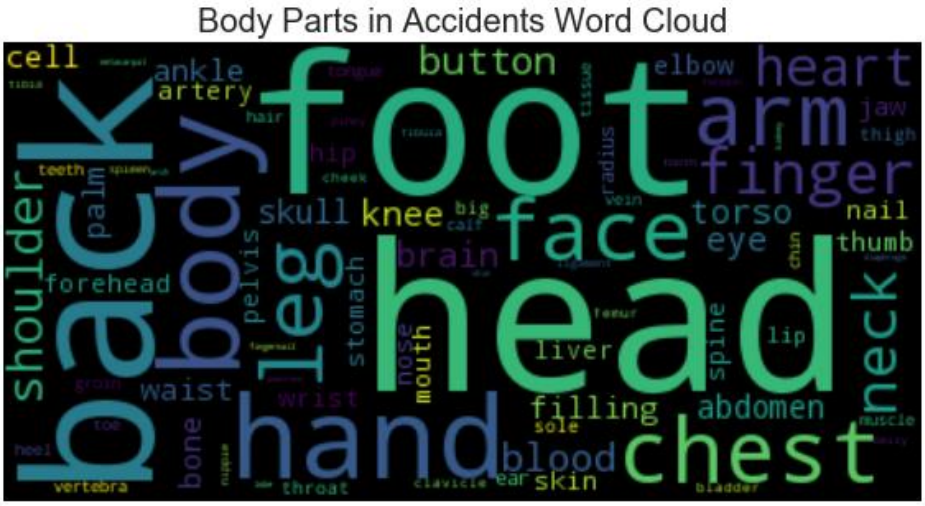

Figure 20: Word Cloud of Body Parts Distribution of *Osha* fatal accidents

3.6 PROBLEM 4: WHAT ARE THE MOST COMMON ACTIVITIES THAT THE VICTIMS WERE ENGAGED IN PRIOR TO THE ACCIDENT?

Dataset used:

'osha_fatal_clean_cause_labelled.csv'

We approach the activity information extraction problem different from the occupation and body parts keywords extraction. Instead of building our own dictionary, we use the post-tag and grammar pattern parsing to extract the activity information for each accident case. Figure 21 shows the steps on how we approach this problem.

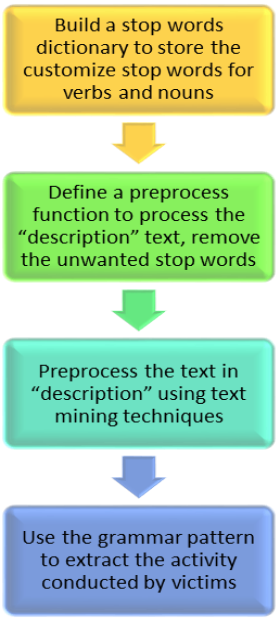

Figure 21: Steps on extracting activities conducted by victims during the fatal accidents

The whole process starts with defining a *pre_process* function to apply necessary text processing on the "*description*" column which contains activity info. The 6 preprocessing steps has been summarized as shown in Figure 22. Notice that this function is different as compare to *pre_process* function for previous problems as it does not include removing of punctuation and English stop words. This is because we conclude that the punctuation and English stop word are useful token for information extraction using pos-tag and grammar pattern parsing.

```
# build a stopword dictionary with verb and noun that are not related to activity conducted during accident
verb_stopwords = set(['killing', 'lying', 'causing'])
noun_stopwords = set(['hospital', 'ambulance', 'treatment', 'amputation', 'explosion', 'boom', 'accident', \
                      'laborer', 'employee', 'coworker', 'crew', 'carpenter', 'painter', 'employer', 'operator', \
                      'firm', 'company', 'facility'])
```

```
# define a function for text mining with the following steps:
# 1. tokenize the string for each row
# 2. convert each of the token to lower case
# 3. remove verb stopwords
# 4. remove noun stopwords
# 5. pos-tag each of the token
# 6. return the pos-tagged tokens
def pre_process(text):
    tokens = nltk.word_tokenize(text)
    tokens_lower = [ word.lower() for word in tokens ]
    tokens_nostop_verb = [word for word in tokens_lower if word not in verb_stopwords]
    tokens_nostop_noun = [word for word in tokens_nostop_verb if word not in noun_stopwords]
    tokens_pos = pos_tag(tokens_nostop_noun)
    return(tokens_pos)
```

Figure 22: Steps by steps *pre_process* function to process the text file in "*description*" column

Figure 23 shows the grammar pattern use to extract activities information conducted by victims. Basically, we extract all the present continuous activity following with an object. The main challenge for this problem is we had accidentally extracted some of the activities that are not related to work activities. To solve the issue, we have built our own verb and noun stop words to exclude those words that is not related to work activities as show in Figure 22.

```
# define the grammar to capture activity conducted during the accident happened
grammar = r'''
ACT: {<VBG><RB|IN|DT>+<JJ.>?<N.*>+}
'''
chunker = nltk.RegexpParser(grammar)
```

Figure 23: Grammar pattern to extract activities information

Figure 24 shows the bar chart of the top 20 most common activities that the victims were engaged in prior to the fatal accident. We have noticed victims are usually stand in front or on top of an object before the accidents happened. Also, there is a high number of victims are operating a forklift when the fatal accidents happen. Hence, it is necessary for the project manager and safety officer to review the operating and safety procedure of using a forklift as well as increase the safety awareness of worker on surrounding environment when standing.

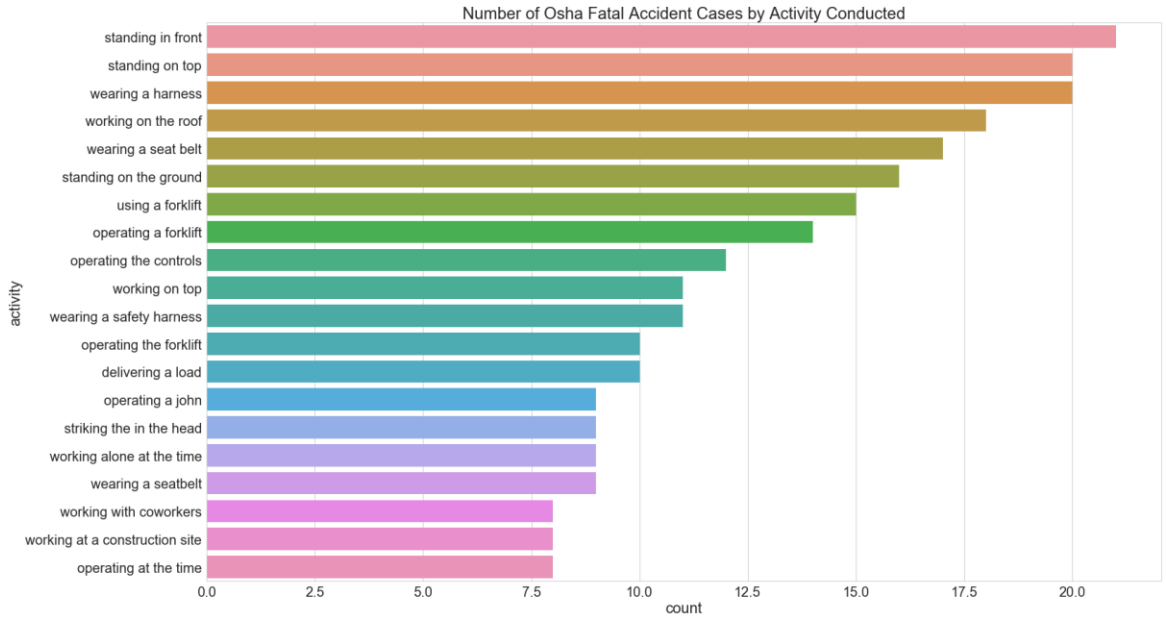

Figure 24: bar chart of the top 20 most common activities that the victims were engaged in prior to the fatal accident.

We also conduct further analysis on the activities key terms to analyze the verbs and nouns of the activities keywords list separately. We use the pos-tag as a condition check to separate the verbs and nouns.

As shown in Figure 25, victims are usually working on truck, machine, forklift and crane when the accidents happen. Perhaps it shows that it is catastrophic when they are struck by this heavyweight object.

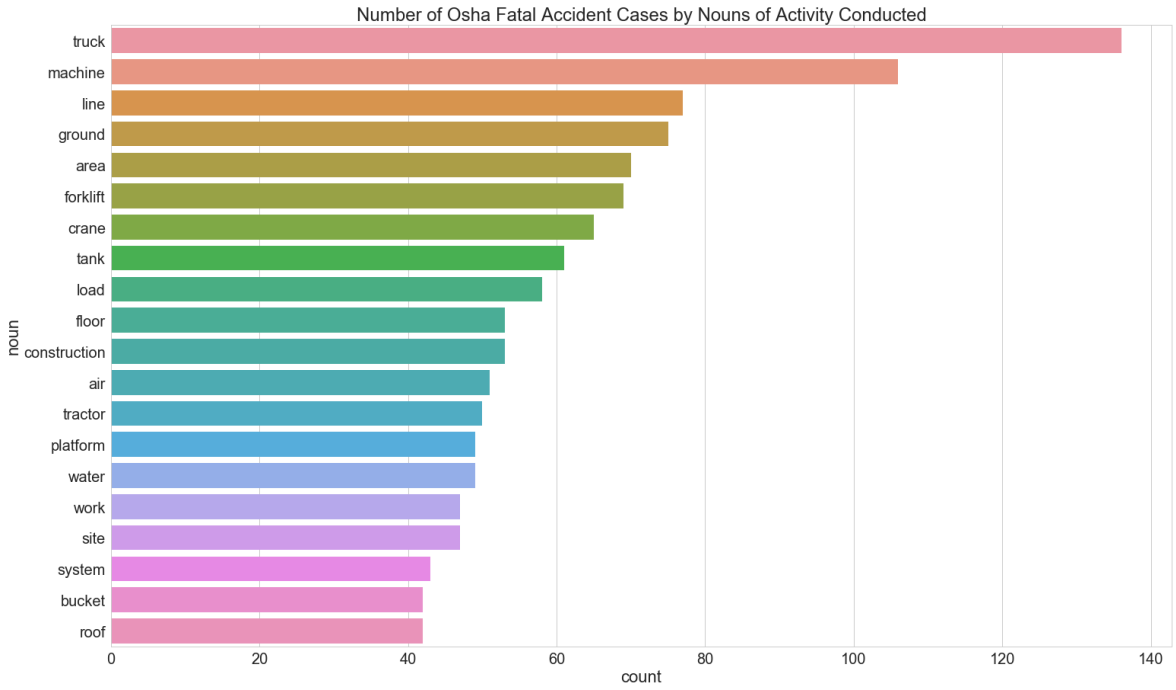

Figure 25: Number of Osha Fatal Accident cases by Nouns of Activity Conducted.

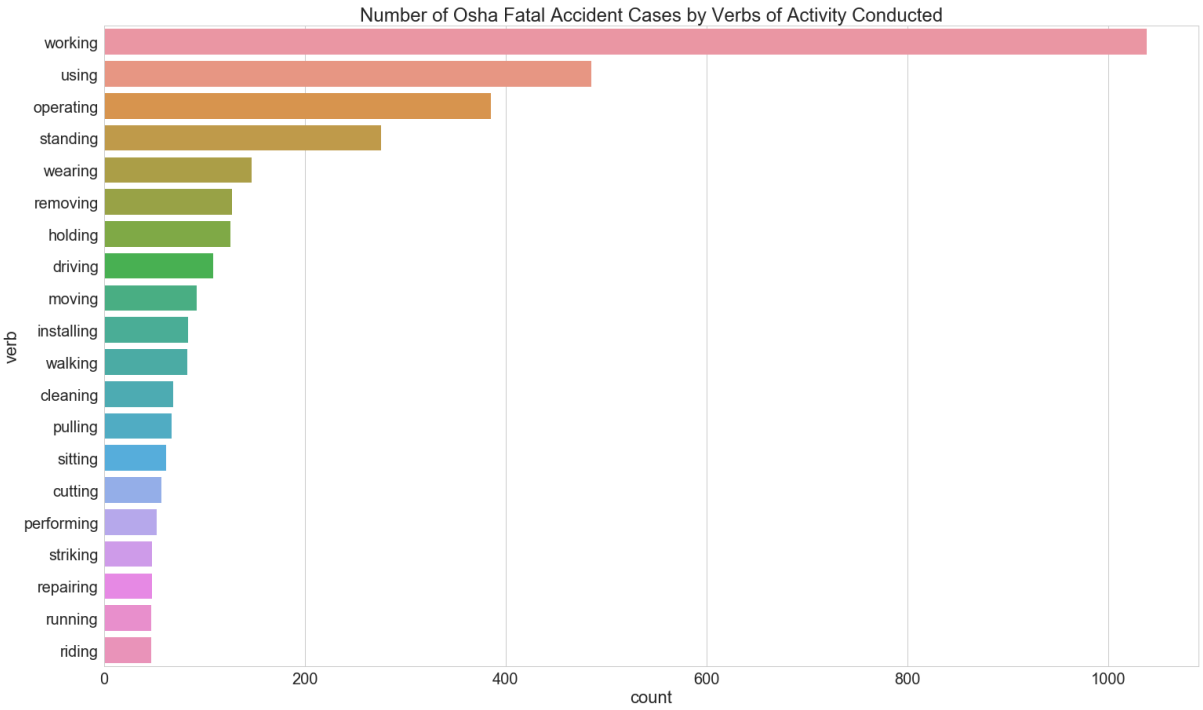

Figure 26: Number of Osha Fatal Accident cases by Verbs of Activity Conducted

Figure 26 shows the count plot of the verb of fatal activities. It shows that victims are usually standing or removing an object when the accident happened. These might be a useful info for the safety officer.

Figure 27 shows the word cloud of the activity keywords when the fatal accidents. It can provide insight for the project manager on which are the common activity and objects that usually associated with the happening of a fatal accident. Noticed that instead of using default collocation analysis, we constructed the word cloud by define each word as single and not collocated to each other as the activity list are preprocessed filtered text with no proper grammar structure.

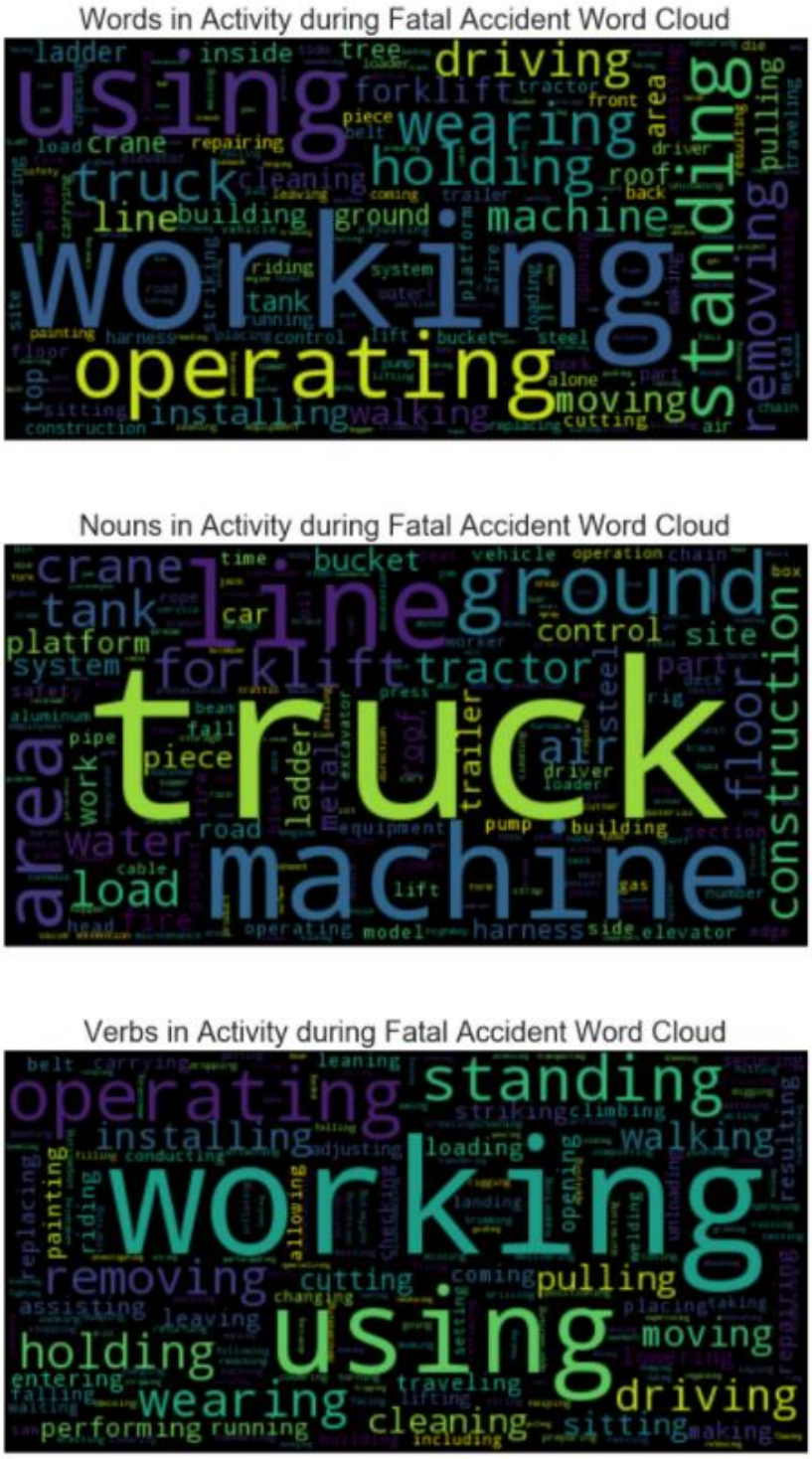

Figure 27: Word cloud of the activity keywords when the fatal accidents

4.0 CONCLUSIONS

In conclusion, by using proper text mining techniques and information extraction approaches, we can build a classification model to classify the unlabelled cause of Fatality and Catastrophe Investigation Summary of Accident Cases in a company. Various useful information such as the risky occupations, body parts that more prone to be injured and activities conducted prior to the happening of a fatal accident can also be extracted. This information can be extracted by key terms matching either by build a dictionary or using pos-tag and grammar pattern parsing techniques.

One of the challenges that we faced in this project is we are unable to build a dictionary to match the activities key terms in the text file as activities is a very generic and it would be difficult to define a dictionary that include all the available activities conducted in the construction site. Hence, by using a proper pos-tag and grammar pattern parsing technique, we can extract the activity information more efficiently and accurately. The main fall back is we need to define a list of stop words ourselves to remove those words that capture by the activities grammar pattern but does not related to working activities.

By using the information extracted, a project manager and safety professionals can plan necessary measure to mitigate the risks of happening of a fatal accident. For instance, our results show that fatality accidents are usually caused by victims struck by moving objects. Project manager and safety professionals might need to review the safety protocol of operating a moving object and raise the awareness of worker who standing near the moving objects. One of the riskier occupation that we discovered in the project is operator, that is necessary for project manager to review the operating work procedure of machine and vehicles operator to determine whether they are well trained before allowing to operate the machines or vehicles.

Besides, there is a need for safety professionals to review the protection of head of the workers as we found out that the fatality always caused by the severe injured of victim's head. It might necessary to make the wearing of certified helmet compulsory or increase the toughness of the helmet used to better protect the head of worker. Further text analysis on the activities string also revealed that victims are usually working on heavy moving objects such as truck, machine, forklift and crane when a fatal accident happen, there is a need to avoid worker to stand closely with operating heavy objects.

Nevertheless, one of the challenges of this project is the *Malaysia* dataset with labelled cause are very small with only 235 cases. The cases are also imbalanced causing the inadequate training data for the minority cause. Hence, it might be fruitful to label a portion of accident data in *Osha* dataset manually to be use as another source of training data. To overcome the issue of imbalanced data, analyst might want to oversample and under sample the dataset accordingly.

As regarding to information extraction, we have demonstrated in the problems on how we adopt different information extraction methodology to extract required information from the text. For occupation and body parts information extraction, we have built a valid terms dictionary to extract the occupations and body parts terms that we are interested. We also built a synonym dictionary in problem to extract the accident cases that caused fatality. For the activities information extraction, filter dictionaries have been created to filter the unwanted stop words that captured by the grammar pattern.

The area that we can explore in the future for this project is to apply ontology-based approach and deep learning to extract the information required. For instance, we can have used *Word2Vec* package for word embedding to extract the activities that is highly associated with a certain accident cause.

5.0 REFERENCES

S. Bird, E. Klein, E. Loper. *Natural Language Processing with Python --- Analyzing Text with the Natural Language Toolkit.*