KE-5205 Text Mining

Continuous Assignment

by
Group 7

Mikhail Raphael | Rishu Raj | Sandeep Yadav | Sourabh Metha | Tan Thong Loon | Yosin Anggusti

List Of Contents

1. Executive Summary

Despite improvement in recent years, the construction industry remains the top contributor for workplace fatalities in Singapore. Many have proposed the use of leading indicators to help construction-related organisations to improve safety of working facilities in Singapore. In this project, we used rigorous text mining methods with iterative development methodology CRISP-DM to help improve our results and findings. Our scope of work is mainly based on the dataset of Osha.xlsx and MsiaAccidentCases.xlsx and below are our main findings and recommendations.

36.5% of the cause of fatalities in construction workplace turns out to be fire and explosion. As part of text mining finding, we also found that 302 cases are labelled as exposure to extreme temperatures, which can be closely interchangeable terms for fire and explosion. Thus, combining the two categories, we deduced that 38.3% cases are at least temperature related. Therefore, we recommend that there is always a temperature monitoring around the area and also, it will be good to provide workers who work with fire or high temperature with fire-resistant jacket. This will hopefully drive down the main cause of fatalities record in the workplace.

In terms of objects that are involved in the fatalities, ladder, forklift and truck turns out to be the main objects. As such, we recommend to cautious while using these objects. Employers might want to send workers to Ladder Safety Training to educate workers on how to reduce the risk of the fatalities. In terms of vehicle safety, trucks and forklifts driver might want to be more careful while operating the vehicles, especially when there are workers around the area. As part of the effort to reduce the accident rates, construction organisations might want to ensure that there is safety zone within vehicle operating area.

This recommendation is further affirmed by the finding from the risky occupations where job involving operation of heavy machineries like truck, forklift and also carpenter turn out to be the riskiest jobs. These are consistent with the finding of the objects and thus, affirming the recommendation to look into safety zone area.

Common activities in which victim was engaged during accident are naturally using a tool/machine, operating a machine, driving, installing, removing, moving etc. However, one insight is also cleaning activities which turn out to be the top activities involved in the cases. Cases related to this specific activity includes cleaning of machineries which were not shutdown properly, thus again, was affirming our finding in top objects involved.

Lastly, since 92.3% of the cases involved single victims, there might be a higher self awareness in terms of safety.

2. Introduction

2.1 Purpose of Analysis

The purpose of analysis is to investigate the:

- Distribution of different types of accidents in terms of their main causes
- Kind of objects causing the accidents
- Occupations that are at higher risk of accidents
- Common activities that the victims were engaged in prior to the accident
- Victim count of whether the accidents involve single or multiple victims

In terms of Business and Text Mining Goals, we are interested in:

- Extracting the meaning and specific information from unstructured text to answer and examine the kind of objects that are commonly involved in the accidents, the kind of occupations that are generally more disposed towards accidents, the common activities the victims were involved in prior to the accidents and the severity of the accidents in terms of the victim counts
- Automatically putting text into categories. Of the 2 datasets available to us, only 1 dataset is labelled. The other dataset contains cases of accidents that are unlabelled. To investigate the overall distribution of the accidents in both the manually labelled and the unlabelled cases, it will be useful to devise an automatic method to tag the documents into specific categories.

3. Main Body

For this project, we uses CRISP-DM framework to guide us in refining the process and results of the text mining. Based on the outlined business goals above, we start understanding the data, followed by iterative process of data preparation, model building, evaluation before the final conclusion of the insight.

Diagram 1. CRISP-DM Framework

3.1 Data Understanding

Using the CRISP-DM framework, the first step of the problem lies in understanding the dataset. Doing a quick breakdown of the labelled dataset, we can see that majority (31%) of the labelled dataset contains accidents labelled as Falls (Figure 1). At the same time, we also notice that there are causes that are similar to each other like Others and Other. There is also some ambiguity in terms of the causes for example, Struck by Moving Object can be understood as similar to Collapse of object if the cause is struck by falling tree.

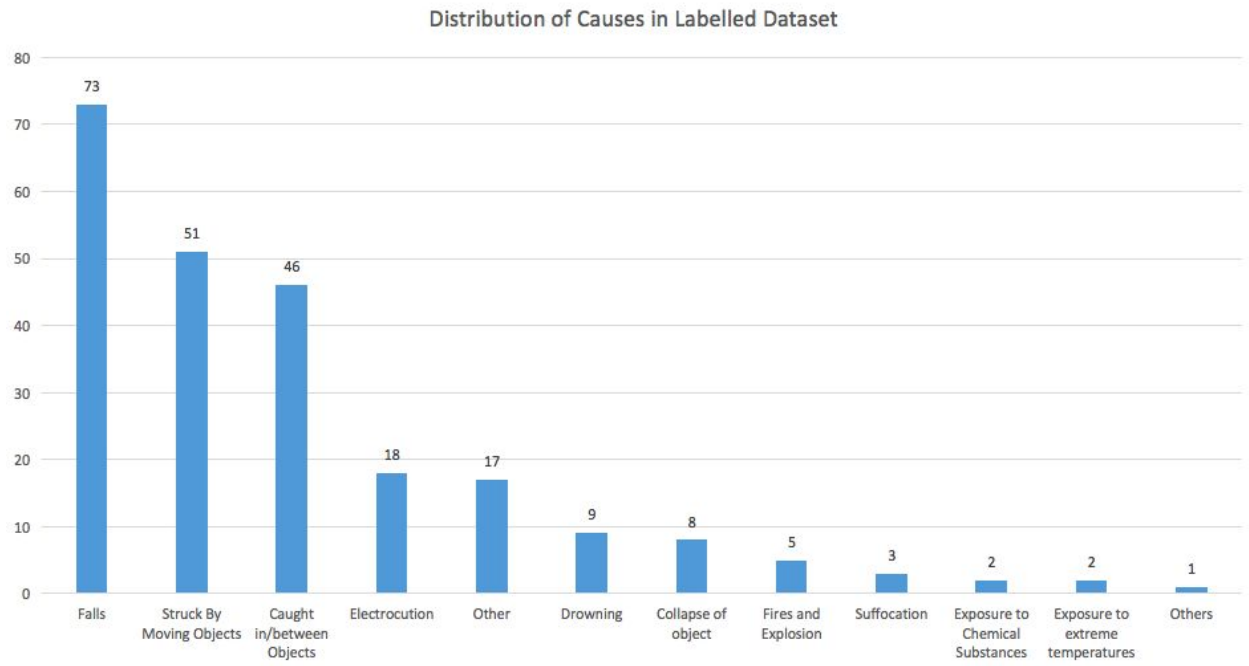


Figure 1. Distribution of Causes in Labelled Dataset

Looking deeper into the causes and the title cases, we noticed that that some of the data are incorrectly labelled for example case 183 in the MsiaAccidentCases where the case was labelled as electrocution even though the victim was crushed by a water tank.

The summary column of dataset contains a lot of potential objects but finding the appropriate object responsible for accident is challenging whereas the same can extracted easily from title column. For most of the accidents that involves someone falling from somewhere, objects are not responsible for accident.

We also observed that the way the accidents and victims described are different for the two given dataset (Osha and MsiaAccidentCases). Thus, to help detect multiple or single victim involved in a case, we chose to focus on the main dataset of Osha as studying the pattern in

MsiaAccidentCases document might not be helpful in extracting the distribution of the number of victims of Osha.

On a random check of the data set to understand what were the occupations of victims, we found that there were cases where the occupation of the victim is not mentioned, and the activity of the victim during the time of accident is not a direct relation to their occupation. For example, employee transferring product from bulk trailer to railcar, got injured when he slipped and fell to the ground. We also noticed that there were reports where victims were referred with their occupation. We also came across reports where the activity of the victim prior to the accident was mentioned as 'walking on a elevated catwalk', with no hints on the occupation of the victim. This encouraged our team to channelize different text mining techniques first, on the title tag of the accident report.

3.2 Data Preparation

In terms of preparing the dataset, we felt that the preparation of the data is slightly different with regards to the text mining goals we are trying to achieve. As such, the data preparation will be divided into data preparation for classification and data preparation for information extraction.

3.2.1 Data Preparation for Text Classification

The first preprocessing step that the team did was to to clean up any observed data quality issues. In this case, we replaced the one and only Others label for case 52 with the Other label. We also recategorised some of the ambiguous cases that we notice like case 183 where the case is labelled as electrocution even though there are no mention of electricity in the dataset.

Following which, the team extracted the text from the summary cases column and performed text preprocessing: removal of unicodes, removal of punctuations, setting the text to lowercase, stemming the text using a Snowball Stemmer and removal of numbers and stopwords. Similarly, this preprocessing was executed on the summary cases column of the unlabelled data before combining the two preprocessed dataset into a combined dataset. This combined text dataset was then tokenized and a term frequency-inverse document frequency matrix (TF-IDF) was built on these tokens to store the distribution of the terms in both the labelled and unlabelled accidents. For the configuration of the TF-IDF, we investigated the distribution of the top 30 words and realise that the word most commonly appearing is the word employee ('employe') which appears in almost all of our documents (Figure 2).

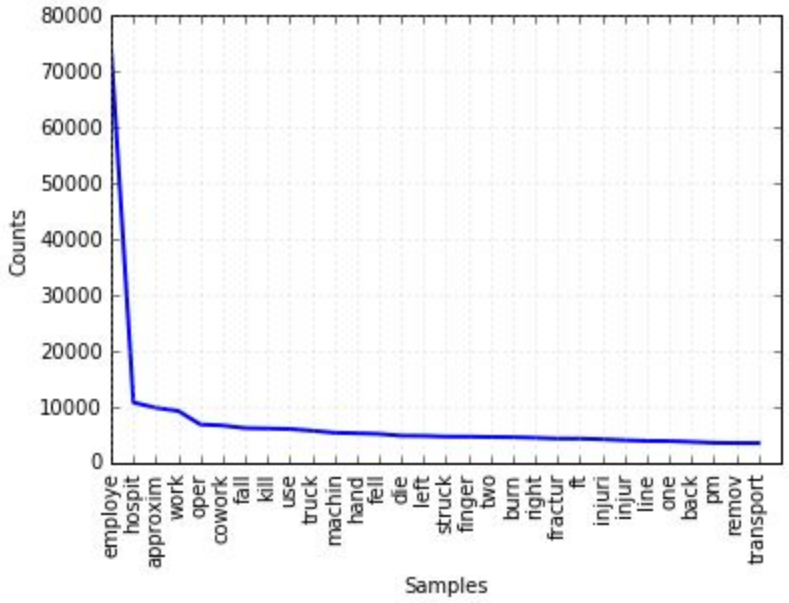


Figure 2. Distribution of Top 30 Words

There is also a large number of terms with only below 10 frequency counts. Hence for the TF-IDF, we set the max_df to 1.0 (i.e. no terms appearing in all documents should be used for the TF-IDF) and min_df to 10 (i.e. no terms with term frequency 10 and below would be used for the TF-IDF). The resulting TF-IDF is a matrix of 16558 x 4123 where 4123 is the number of remaining terms. This TF-IDF will be used for the classification model.

3.2.2 Data Preparation for Information Extraction

The difference between the preprocessing step for information extraction and for classification is the sentence tokenizing of the accident cases, word tokenizing of the sentences followed by Part of Speech (POS) Tagging without any removal of stop words. Using the pos tagger from nltk package, we are able to identify the nouns and verbs and store them separately into two separate lists. The purpose of this extraction is to facilitate the grouping of nouns into possible objects involved in the accidents, possible hints of the occupation the person is involved in as well as the number of victims. On the other hand, the grouping of tokens into verbs helps to facilitate the identification of activities the victims might be engaged in prior to the accident happening.

As part of the process to identify number of victims, we used various technique including POS for specific noun (‘NN’ and ‘NNS’) to help identify singular and plural nouns where we cannot get enough hints from general keywords like ‘employee’, ‘worker’ and ‘victim’. One observation of pos_tag from nltk library in this specific account is that it is not very accurate as it gives tag like ‘datesicestablishment’ and ‘name30843203806/28/20051742mader’. However, we still managed to use the remaining tags to help identify the potential victim and categorise them to either multiple or single victim based on the grammar. For this exercise, we avoid lemmatization,

stop word removal and lowercase update to preserve the accuracy of pos_tag as much as possible. The POS tagging was also only applied after sentence tokenizing.

3.3. Modeling and Solutioning

3.3.1 Classification of Causes

To categorise the unlabelled cases in Osha.xlsx, the team implemented and compared the accuracies of creating the classification model from a few approaches.

1. Implementing the classification model using a Support Vector Machine and doing 10 fold Cross Validation with 10 repetitions to estimate the out of sample error rate as this will minimise the .
    - The SVM was configured with a Cost Factor of 10000, gamma setting of auto and using a radial basis function.
2. Implementing the classification model using an ensemble of K-Nearest Neighbour looking at 3 neighbours, Support Vector Machine, Logistic Regression using Majority Voting as the decision method for the ensemble and doing 10 fold Cross Validation with 10 repetitions to estimate the out of sample error rate.

The results of the out of sample accuracy rate show that a single support vector machine actually performs relatively well compared to the ensemble model with 70.6%% vs 67.5% respectively. This might be due to the ensemble not taking into consideration the different weightage relative to the confidence of prediction of the individual models.

3.3.1.1 Improving Classification by Feature Engineering

Re-examining the dataset to try and improve the modelling results, we realised that using the tf-idf as training data has some pitfalls as there might be some terms that are synonyms of each other for example fall and slip, fire and flame. However, the tf-idf method will recognise them as different terms and hence not learn to use these semantic features for classification. Another downside of using tf-idf is that the tf-idf might get very sparse if the terms used are unique across documents. As such, this might affect the classification accuracy.

The first way to improve the classification accuracy will be to try to reduce the features by doing singular value decomposition (SVD). Iterating the SVD from maximum 4000 number of components to 1000 components and plotting a scree plot (Figure 3), we notice a sharp dip in the variance explained when the SVD moved from 2000 to 1000 components. Therefore, using a svd with 2000 components to extract out 2000 features from the original TF-IDF, we can hopefully both reduce the number of features necessary to explain the variance in the model as well as account for considerable accuracy. The two models are then retrained and their accuracies noted.

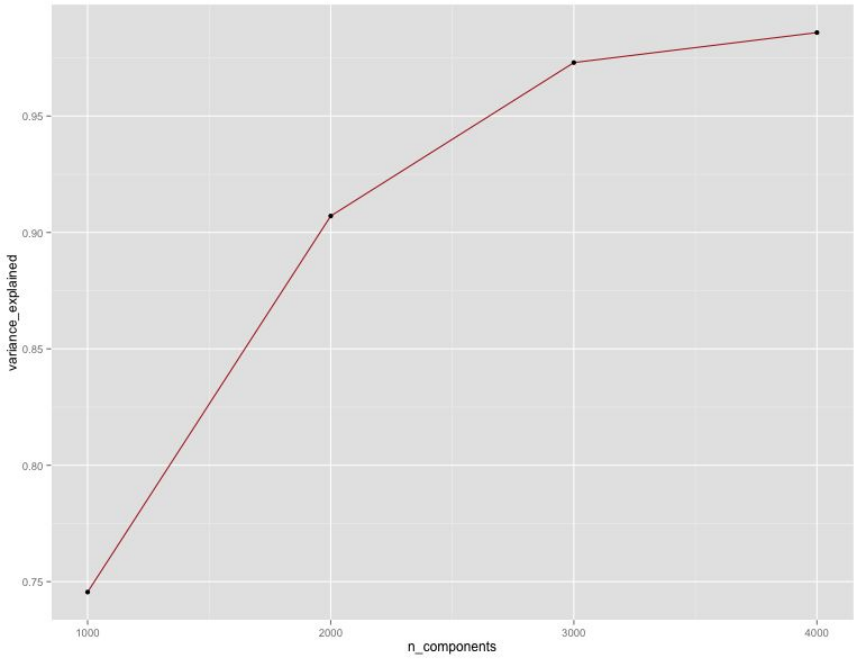


Figure 3. Scree Plot for finding Number of Components for TruncatedSVD

The 2nd way to improve the classification accuracy would be to do feature hashing to allocate the features into buckets of hash locations. For this hashing, instead of using the TF-IDF, the full summary and title text was used to pass to the feature hashing algorithm in NLTK to hash the features into separate buckets. After feature hashing, the models are retrained and their accuracies noted.

Comparing the models built on the engineered features and the models built on the original tf-idf, we notice that there is a significant improvement of 5% in the accuracy for SVM model built on Feature Hashing (Table 1) and also for the ensemble model. For the TruncatedSVD, the accuracy also improved by 0.4% for the SVM model and 0.3% for the ensemble model. This might be due to feature hashing capturing the context of synonyms without discarding any of the features like TF-IDF and TruncatedSVD capturing the variance explained without the use of unimportant features.

Model \ Features	**TF-IDF (min_df=10, max_df=1.0)**	**Feature Hashing**	**Truncated SVD on TF-IDF (n_components = 2000)**
SVM	0.7066	0.7520	0.7106
Ensemble	0.6753	0.7398	0.6788

Table 1. Comparison of Accuracies before and After Feature Engineering

As such, it seemed that feature hashing would be a better choice for the classification model. Using the feature hashing model, the team went on to predict the accident cause labels for the unlabelled cases. With this prediction, the distribution of the causes of the traffic accidents are as Figure 4 with most of the cases being caused by Fires and Explosions.

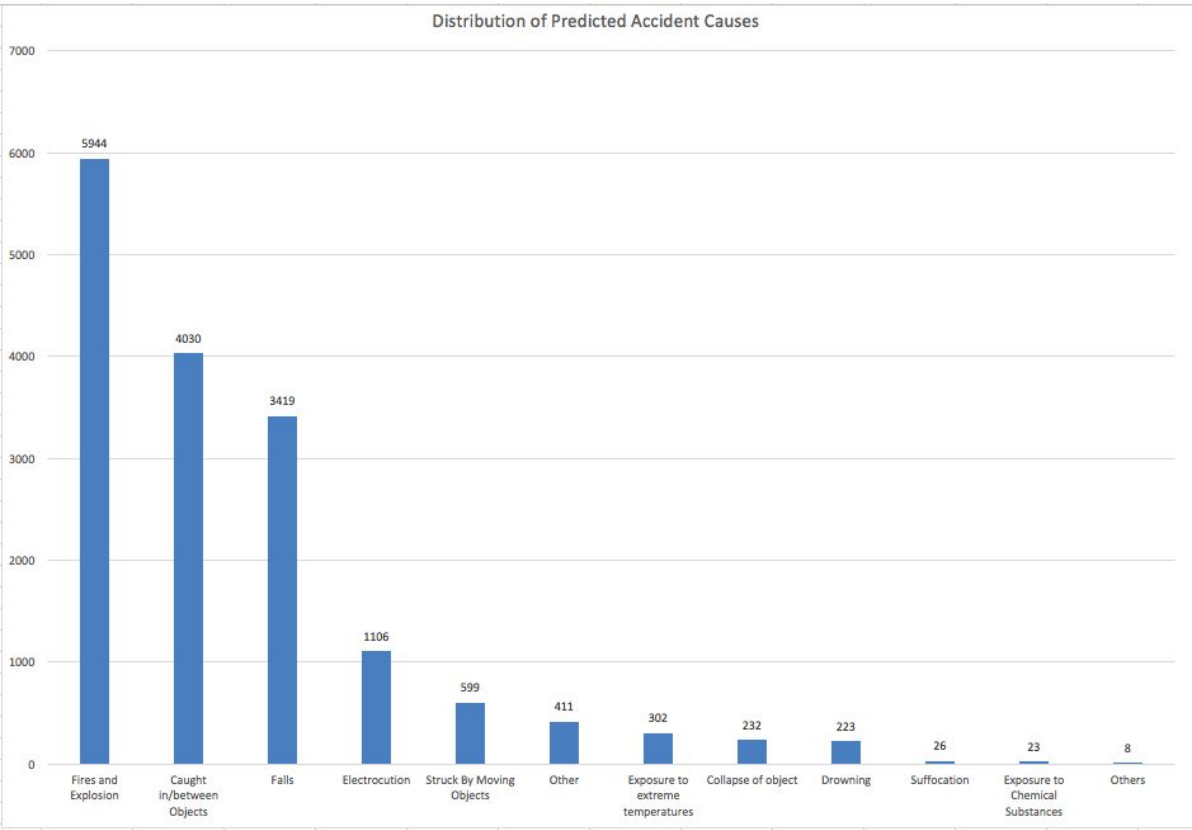


Figure 4. Distribution of Predicted Accident Cases

Randomly examining 10 accidents for the predicted causes in Osha, we noticed that the prediction was only right 40% of the time. The incorrect causes appeared to consist mostly of Fires and Explosions.

3.3.2 Extraction of Kind of Objects

For extracting information about objects responsible for accidents, we used different approaches to get the insights. We tried extracting Proper Nouns from the summary column text especially the nouns that comes after preposition using POS tagging text mining technique. Because of presence of a number of keywords, it was difficult to identify the appropriate object responsible for accidents. The keywords extracted from the summary column gave a very vague result with low precision.

A closer look at the Title column gave a better perspective of extracting objects responsible for accidents. For most title entries, Proper Noun followed by preposition seemed to be the objects

involved in the accidents. Examples of some of them are struck by Forklift, caught in conveyor belt, overrun by truck, injured by grinder etc. We tried extracting all those proper nouns that follows preposition/subordinating conjunction using POS tagging. The results were fairly good. We were able to get objects for more than 65 percentage of accidents. The most frequent objects responsible for accidents were found out to be fall, falling, Ladder, truck, forklift and explosion.

Diagram 2. Word Cloud After Stemming And Punctuation Removal

The word cloud (Diagram 2) was obtained after doing stemming and removing punctuation from list of potential objects responsible for accidents. Common potential objects apart from fall are ladder, forklift, truck, trench, conveyer, chemical, steel and tree.

Diagram 3. Cloudword after Removing ‘Fall’ Keyword

The above word cloud (Diagram 3) excludes fall and falling keywords as in most of the accidents that involve someone falling, objects are generally not responsible. We extracted the

most frequent terms (Diagram 4, Figure 5) with frequency of number of accidents they were involved in.

Diagram 4. Most Frequent Objects with Frequencies

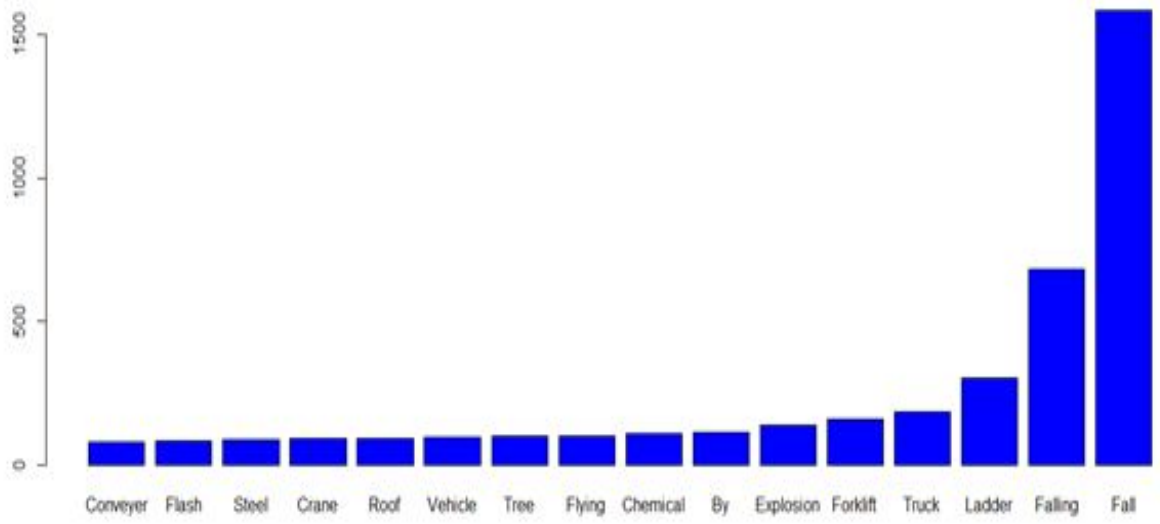


Figure 5. Histogram of Most Frequent Objects

3.3.3 Extraction of Risky Occupations

While trying to determine the different occupations among victims, thus identifying riskier occupations, we noticed that the first noun of the title tag of the report often refers to the victims with their occupations, hence as the first iteration, we extracted the first occurrence of the noun from the report and examined if this was a valid occupation. From this process, we were able to identify distinct occupations like carpenter, foreman worker. For the 2nd iteration, we then used this extracted list and searched for the occupations in the accident descriptions where the title had not revealed the victim occupations.

Concurrently, our team used the “occupations” dictionary extracted from ‘Java Gate Annie Gazetteer’ to pattern match through the accident descriptions to find if there are occupations corresponding that can be extracted from the accident report. For this we use regular expressions

```
-- regex --      <var> = r"""
                  NP: {<NN>+}
                 {<NNS>}
                  """.
```

Iteration 1:

For iteration one, the ‘title’ column from the osha.xlsx was used to fetch occupations. Using regular expressions (as given above), our team extracted the ‘Noun - Noun’ combinations in each of the accident report titles. There were about 982 distinct occupations that came up after the first set of iterations. However close examination of the results revealed that some of the ‘occupations’ in the iteration 1 output list were not ‘valid’ occupation and hence need to be discarded. For this process, our team decided to examine the most repeated ‘occupations’ and then discard the ones that were to be considered ‘invalid’ during manual observation.

Iteration 2:

For iteration two, our team used the ‘occupations’ list obtained from iteration 1 to fetch against possible ‘occupations’ under the ‘description’ tag of those records in the accident report file for which the ‘Noun-Noun’ regex execution on their corresponding titles did not fetch results.

Concurrently our team also used an ‘occupations’ dictionary from the ‘Java Gate Annie Gazetteer’ to match against all accident report ‘descriptions’ in the accident report document (osha.xlsx) .

A few assumption made:

i) Key words found in the ‘occupations’ list such as ‘Employee’, ‘worker’ are not considered as valid occupations.

ii) All ‘occupations’ present in the dictionary are assumed to be valid.

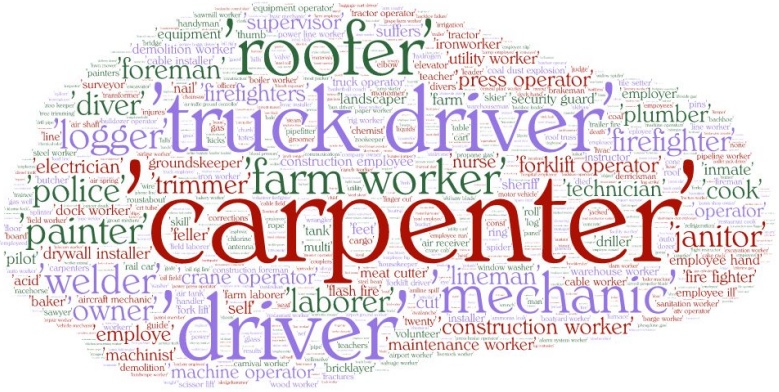


Diagram 5. Word Cloud of Occupations

The word cloud above (Diagram 5) shows the risky occupations obtained from the accident report. We observed that the prominent occupations that are considered as risky are those of carpenter, driver, welder etc.

3.3.4 Extraction of Common Activities

In order to extract information such as activities in which employees were engaged when the accident happen, we tried two approaches -

1. One using Part of Speech tagging in order to extract verbs which may refer to activities. Problem with using only this approach was that all verbs were extracted even if they were not related to employee/worker and filtering terms which were not activity became hard as lots of manual work was required.
2. Another approach we used was to extract all employee based instances (sentences) using regular expressions and then extracting activities based on existing pattern in data or known activities or verbs. We used this approach and divided the information extraction into 2 phases:
    a. Extract all sentences were employee(s) was referred. Below regular expressions were used:

```
([Ee][Mm][Pp][Ll][Oo][Yy][Ee][Ee]\\s+(..\\s+)*was\\s+[a-zA-Z]+ing\\s+([a-zA-Z0-9[:punct:]]+\\s+){0
,5})
(([Ee][Mm][Pp][Ll][Oo][Yy][Ee][Ee][Ss]{0,1}|[Hh][Ee]|[Tt][Hh][Ee][Yy])\\s+([a-zA-Z0-9[:punct:]]+\\s
+)*(was|were)\\s+([a-zA-Z]+\\s+)*([a-zA-Z]+ing\\s+|[a-zA-Z]+ed\\s+)([a-zA-Z0-9[:punct:]]+\\s+){0,
5)
```

    b. Extract activities based on pattern in data or known activities or all verbs. During our analysis we found that activities like clean and repair occurred several times which we used in regex to directly extract. Also, we found that most of the sentences where employee(s) reference was followed by was/were had activities ending with ing and ed (except for working) so we used below regex to extract all such activities:
    ((was|were)\\s+([a-zA-Z]+ing|[a-zA-Z]+[^z]ed)\\s+[a-z]{0,3}\\s{0,3}[a-zA-Z]+\\s+)

    In the end, in order to improve completeness of our regex (where no activity was found using above regex) we extracted all verbs ending with ing.

Further, in order to analyze information, we did pre-processing such as removing punctuations, converting to lowercase, removing stopwords etc and tokenized processed information into unigrams and bigrams. Results are as of Figure 6:

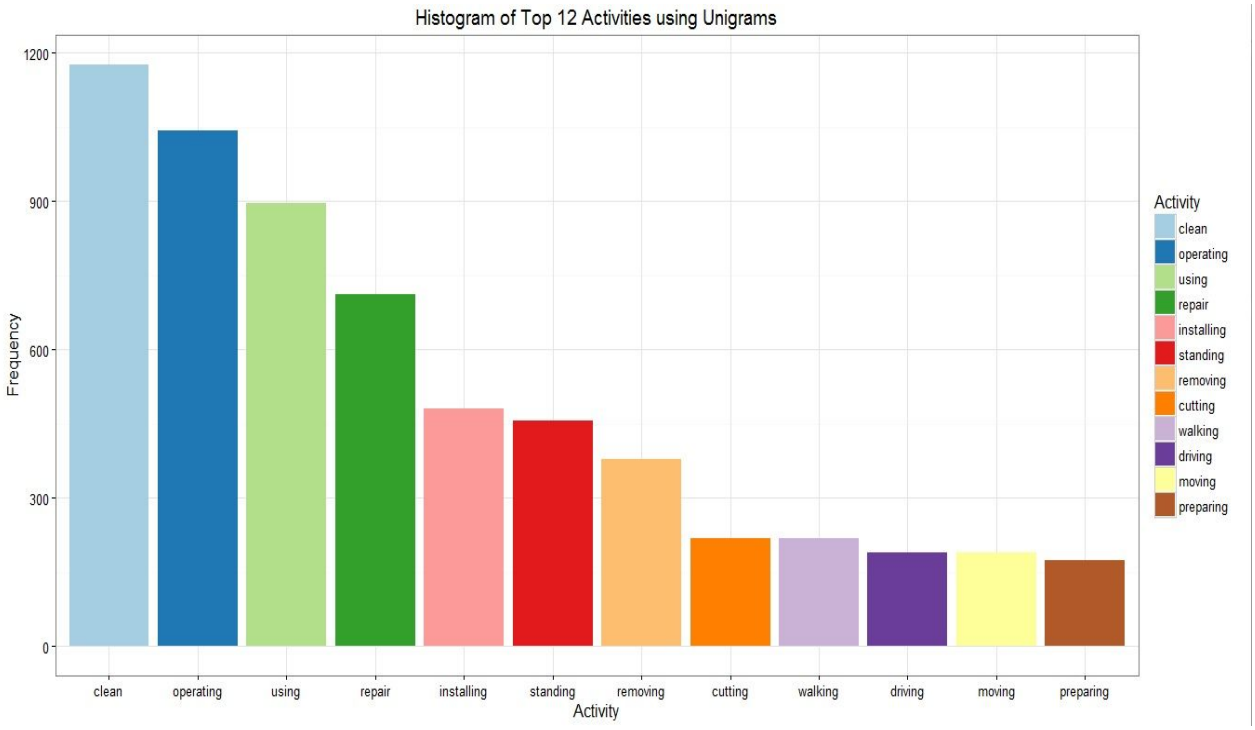


Figure 6. Distribution of Top 12 Activities Using Unigrams

As we can see, unigrams generalise common activities but we could not make out what type of activities are related with using and performing. Below is the list of activities which are related to using. Using below graph, we can conclude that using associates to using a tool/machine.

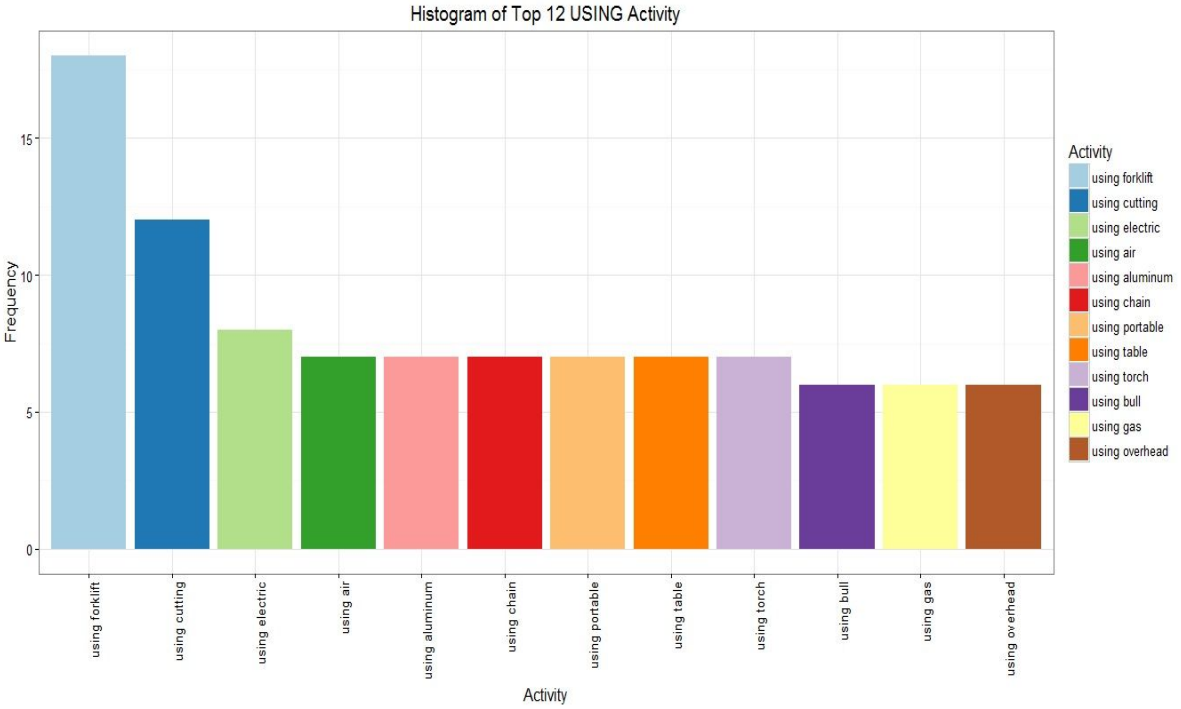


Figure 7. Distributions

Similarly for Moving activity (Figure 7), we can conclude that the term moving is associated to moving an object from a bigram perspective.

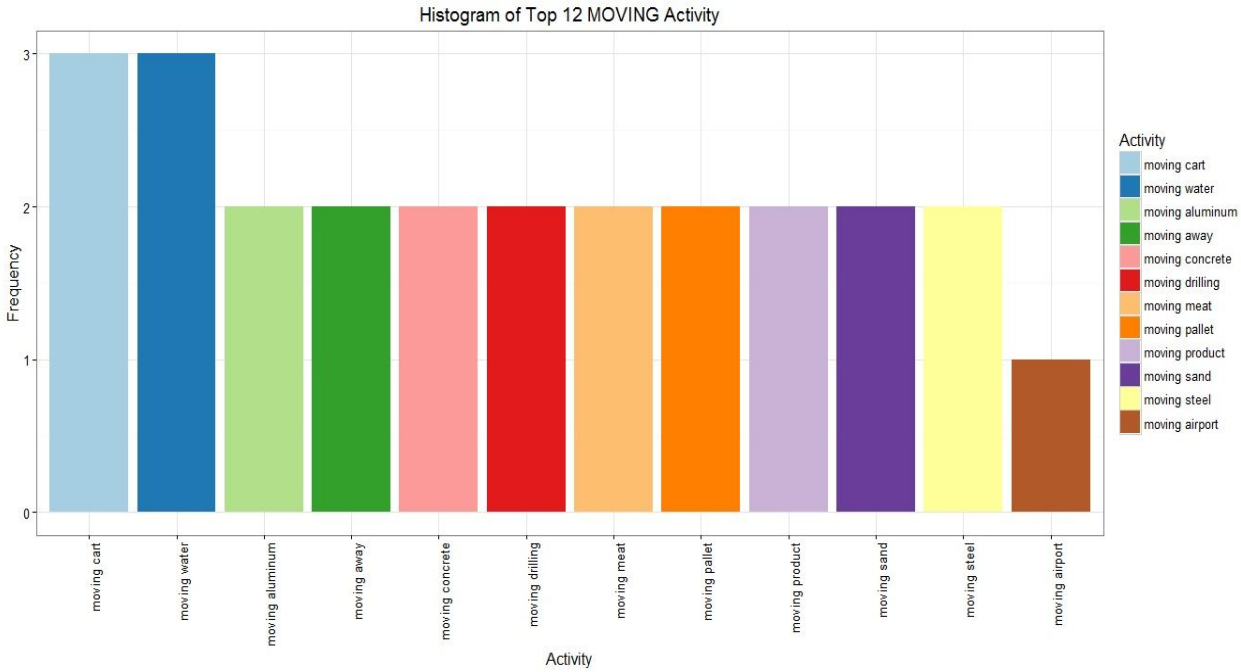


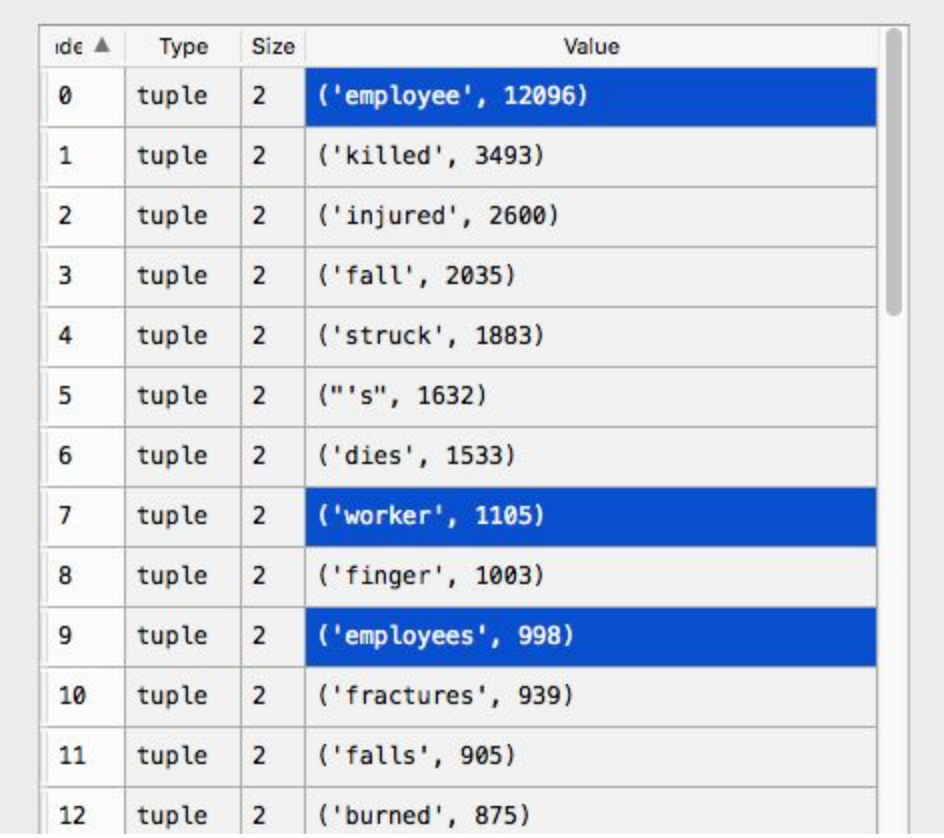
Figure 8. Histogram of Top 12 MOVING Activity

Results could be further improved by building a dictionary of activities using terms obtained above and using it to extract activities directly.

Diagram 6. Word Cloud for Common Activities

3.3.5 Extraction of Number Of Victims

In identifying number of victims involved in each case, we use CRISP-DM methodology to iteratively improve the model we are building. Before we start with the first iteration, we were trying to understand the text and were doing tokenization to identify the most appearing words which could give us a hint on the number of victims involved. It appears that there are common words highlighted in below diagram.



ıdɛ ▲	Type	Size	Value
0	tuple	2	('employee', 12096)
1	tuple	2	('killed', 3493)
2	tuple	2	('injured', 2600)
3	tuple	2	('fall', 2035)
4	tuple	2	('struck', 1883)
5	tuple	2	("'s", 1632)
6	tuple	2	('dies', 1533)
7	tuple	2	('worker', 1105)
8	tuple	2	('finger', 1003)
9	tuple	2	('employees', 998)
10	tuple	2	('fractures', 939)
11	tuple	2	('falls', 905)
12	tuple	2	('burned', 875)

Figure 9. Dictionary of Most Frequent Tokens

The first iteration was reading the title column and grouping the single and multiple based on derived commonly appearing words: 'employee', 'victim' and 'worker'. Just based on whether its plural or not, we get the results of 1,078 multiple victims and 13,194 number of single victim cases, with 2,051 cases (12.5%) unlabelled due to the absence of the three keywords.

To improve this model, we run the second iteration with the observation of the common syntax (‘#1’,’#2’,’#3’) identified from the description column of the cases. With this, check if the description contains #3 or #2 or #1 respectively. With this, we realise almost all the cases can labelled. Below are the result of the case distribution with 501 cases involving 3 victims, 993 involving 2 victims And 14,338 involving single victim with 491 remaining cases unlabelled. In this iteration, we managed to drop the unlabelled data from 12.5% to just 3%.

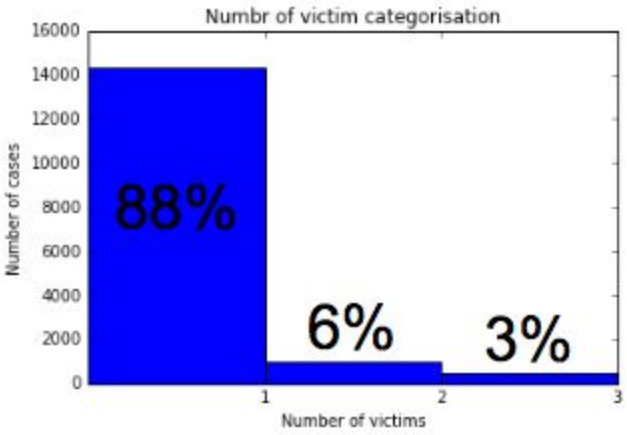


Figure 10. Distribution of Number of Victims

In the third iteration, we are trying to label the remaining 491 unlabelled cases. To do this, we identify the plural and singular noun based on 'NN' and 'NNS' pos_tag from nltk library. We assume that the noun which identifies occupation like firefighters are the victims in the case. This assumption has proven to make sense based on our random noun checks like 'operator' and 'electrician' which are the victims in case number 201621844 and 170158908 respectively. With this, we managed to improve labelling to 98% of the cases. Finally, we observed the remaining 304 cases unlabelled and deduced that mainly these cases fall into single victim category. As such, we presented the case of single to multiple final proportion to be : 90.3% to 9.7%.

In order to evaluate the accuracy of the labeling on number of victims, we randomly picked 10 cases and check if the labeling in third iteration is done correctly. It turns out that out of 10 cases, there are 4 wrong labelling such as case 202651907 (row 40) and case 201035995 (row 50) due to the use of #2 in the descriptions while the actual case is just involving one victim, as mentioned in title column. As such, we continue to iterate to the fourth iteration where we try to interpret the title column first before going to the description column with the result of 1,255 involving single victim and 15,068 involving multiple victims. To check accuracy, we randomly select 10 labelled data and verified by reading the case manually. It shows all 10 to be accurate. Thus, we decided to stop at this iteration for the labelling of the case.

4. Conclusions

Our team explored the use of TF-IDF, Dimension Reduction using TruncatedSVD on the TF-IDF and Feature Hashing on full text to train both a SVM model and the ensemble model for the predictive classification of the accident causes. From our prediction results, it can be seen that majority of the accidents in the unlabelled dataset were labelled as Fires and Explosions, followed by Caught In Between Objects. This distribution was surprisingly different from the original training dataset distribution. Random sampling the predicted results also revealed that the labelling was not entirely accurate for certain labels like Fires and Explosions, explaining the high numbers of Fires and Explosion. On hindsight, we learnt that using accuracy to measure the result of this automated classification might be inappropriate especially in the event of unbalanced class labels as the classifier would classify most of the causes to the highly occurring cases to achieve a high accuracy. Another lesson we learnt was that choosing the right features was important to achieving good results in a classifier. However, this is often a compromise between choosing a fast and good enough classifier versus one that is more complex and requires more storage capacity like feature hashing with the default number of features.

In terms of information extraction, we learnt that with text mining techniques, we were able to extract objects responsible for about 65 percent of accidents records in dataset. The most common objects identified are fall, Ladder, Truck and Forklift. POS tagging is not always accurate. E.g. “falling” word was tagged as proper noun by POS tagger . For some of the accidents such as accidents where victim fell from somewhere, no objects would be responsible for accidents in those cases. Similarly for common activities, just using POS tagging did not provide comprehensive results. We had to use additional processing such as extraction using regular expression to further improve our analysis.
Further iteration of text mining could be applied in the summary column of dataset for the result which did not give any clear objects responsible for accident. For example, results such as falling and flying are vague .Further mining can be done to find out objects which could be falling or flying and which could have resulted in accident .

In terms of business findings, 36.5% of the cause of fatalities in construction workplace turns out to be fire and explosion. Combining this with related category of exposure with extreme temperature, we deduced that 38.3% cases are at least temperature related. Therefore, we recommend that there is always a temperature monitoring around the area and also, it will be good to provide workers who work with fire or high temperature with fire-resistant jacket. This will hopefully drive down the main cause of fatalities record in the workplace.

In terms of objects that are involved in the fatalities, ladder, forklift and truck turns out to be the main objects. As such, we recommend to cautious while using these objects. Employers might want to send workers to Ladder Safety Training to educate workers on how to reduce the risk of

the fatalities. In terms of vehicle safety, trucks and forklifts driver might want to be more careful while operating the vehicles, especially when there are workers around the area. As part of the effort to reduce the accident rates, construction organisations might want to ensure that there is safety zone within vehicle operating area.

This recommendation is further affirmed by the finding from the risky occupations where job involving operation of heavy machineries like truck, forklift and also carpenter turn out to be the riskiest jobs. These are consistent with the finding of the objects and thus, affirming the recommendation to look into safety zone area.

When trying to determine the risky operations our team found that in general, occupations involving handling of heavy machinery such as forklift, cranes, trucks etc were to be classified as 'riskier' than the others occupations. We also noticed that, in most cases victims of accidents were mostly bystanders who were with people having 'risky' operations. Our team had to take a few assumptions when considering operations as 'risky'.

Common activities in which victim was engaged during accident are naturally using a tool/machine, operating a machine, driving, installing, removing, moving etc. However, one insight is also cleaning activities which turn out to be the top activities involved in the cases. Cases related to this specific activity includes cleaning of machineries which were not shutdown properly, thus again, was affirming our finding in top objects involved.
In terms of technical findings, during our analysis, we found that unigrams help in generalising analysis while bigrams help in finding more specific activities/descriptions.

In terms of number of victims distribution in Osha cases, we found that 9 in 10 cases involve single victim. Most of the cases involve single victim.

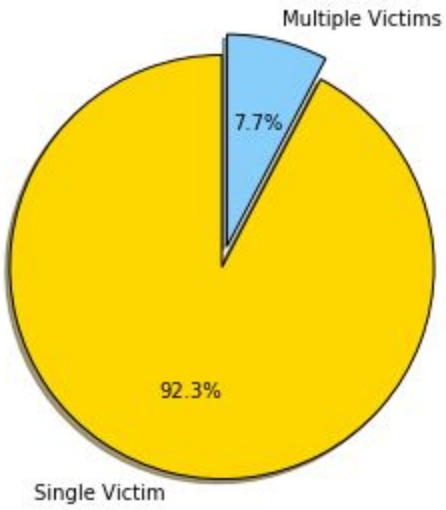


Figure 11. Single VS Multiple Victims

Since most of the cases involve single victim, what construction industry may want to do is to have a buddy system where a construction worker is given a partner to take care of, ensuring buddy's personal protection and safety wears are worn all the time in the construction area. That way, there is always one person who is taking care of another person. With each worker having a partner who help look after him, it will hopefully lower down the number of single victim cases which happen to be the majority of the cases. There needs to be higher self awareness also in order to increase personal safety.

4.1 Further Work

Further work can be done on:

- Generalizing regular expression to improve soundness and completeness
- Building domain specific dictionaries for occupations, objects and activities to supplement information extraction and improve accuracy.
- Implement spell checking & correction and word splitting (where there is no space between words) in order to further clean data.
- Using Clustering for examining the clusters of the accidents to reduce the labels for the causes as some labels are ambiguous. Clustering can also be used to do labeling for the causes based on their membership
- Improving clustering and classification through the use of deep learning techniques like word2vec to learn the word embeddings for each of the word and and multi-word expressions tokenizer for tokenizing words.
- Using F1-measure to measure the performance of the classifiers instead of accuracy.

5. List of References

[1] ANNIE Gazetteer Occupation List from Java GATE