Analyzing Fatality and Catastrophic Accidents Investigations Summaries

Team: Text Miners

Institute of Systems Science, NUS

[Abhilasha: A0163276B, Muni: A0163382E, Nandini: A0163272J, Neeraja: A0163327H,

Pradeep: A0163453H, Rahul: A0163314N]

1 CONTENTS

ABSTRACT

This work is part of assignment to apply Text Mining concepts on a Fatality and Catastrophic Investigation summaries from Construction industry. We would like to understand, process and apply some proven text analytics methods to gain insights in to the data as well as recommend some of the changes to the processes to improve the safety measures for workers in construction industry.

2 EXECUTIVE SUMMARY

Our business problem is to analyze the fatality and catastrophic data mainly provided in textual form to gain insights to following main areas:

1. Which type of accidents are more coming in fatal or catastrophic accidents?
2. What are the riskier occupations in such accidents?
3. Which part of human body are more prone to be injured in such accidents?
4. What are the common activities that the victims were engaged in prior to the accident?
5. What are some of the recommendation to improve the safety measures

Our answers to the above questions are:

1. **Caught in/between Objects**, **Falls**, and **Collapse of Object** are the top 3 types of accidents.

2. **Driver**, **Operator** and **Carpenter** came out to be riskier occupations in such accidents.

3. **Finger**, **Arm** and **Hand** is found to be more prone to be injured in such accidents.

4. **Operating the machines** caused caught in/between objects, **Cleaning** resulting in to fall by slipping, exposure to chemical substance. Fires/Explosion and **driving** are the major activities victim was performing at the time of accident.

5. Based on the learning we can think of building a smart accident preventions application. This app will have potential to predict the foreseen risk based on type of work being carried out. It can act as a smart alert system to raise the alert about posed risk with severity before an activity to be started. Apart from this the major accident types are caused by mostly lapses of standard procedures or negligence. Due care and stricter process should be implemented for such accident types as they are fatal.

For the detailed explanation, refer to Section 2.6

We followed the various concepts taught in the class mentioned below to conduct our analysis:

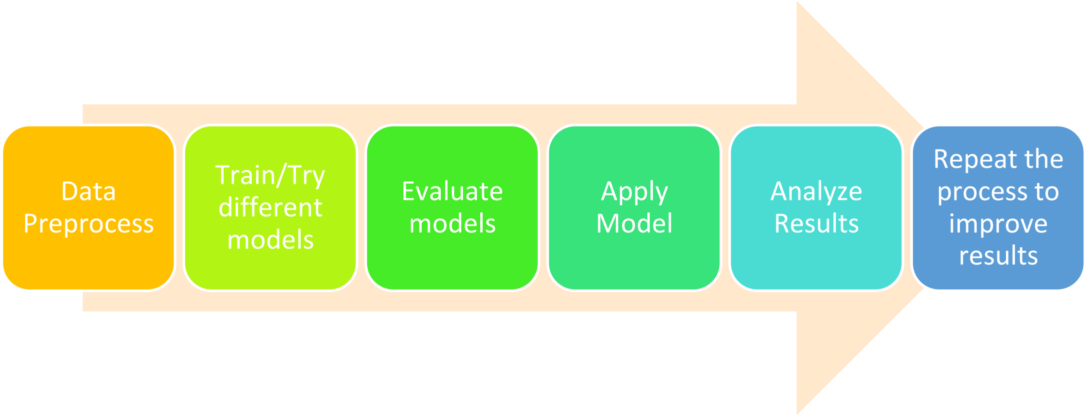
- Text Pre-Processing
- POS tagging
- Classification using SVM
- Random Forest (Ensemble)
- Clustering
- LDA
- Word2vec

2.1 Data Selection and Preprocessing

We have been provided 2 sets of data files:

1. MsiaAccidentCases – This is already a classified/labeled data which could be used to train and test the model
2. osha – This is unlabeled data and something we want to gain insights to.

   On a high system level, we followed the process as mentioned below:


Data Preprocess
Train/Try different models
Evaluate models
Apply Model
Analyze Results
Repeat the process to improve results

The classified data contain 3 variables:

- **Cause:** This is a classified type of the accidents. There are total 12 categories.
- **Title**: This is a title given to the accident case
- **Summary**: This is the detailed report over the accident.

Osha contains only title and summary about the incident. This is our unseen data which will be applied the knowledge learned from the already classified data.

2.2 Data Gathering and Partition

MsiaAccident cases already contains a separation between train and test data. We used the same for our training and evaluation purposes.

However, number of records in train/test is quite small 181 and 51 respectively. We followed 2 steps to enrich the train/test data:

Step 1

As mentioned in Section 2.5, we manually tagged the records from Osha and used them for train/test.

Approximately, 347 and 125 records were used for Train and test respectively. This approach showed the consistent classification between the models.

Step 2

We also made use of last column in Osha.xslx to improve the overall accuracy of the model. As such the data represented in *MsiaAccident* was quite low. There was comment like section for about 3.5k records which mentioned the cause of the accident. We filtered out those records which had cause mentioned and built a dataset which was tagged using the cause mentioned in those records against categories in the *MsiaAccident* cases. We combined all the data from above steps and split this data randomly in to 70/30 ratio.

Table below shows the causes mapped to categories in train data:

Causes In Osha	**Mapped Category Train Data**
Caught in stationary equipment	Caught in/between Objects
Elevator (struck by elevator or counter-weights)	
Crushed/run-over by construction equipment during	
Crushed/run-over/trapped of operator by operating	
Drown non-lethal fall	Drowning
Asphyxiation/inhalation of toxic vapor	Exposure to Chemical Substances
Heat/hypothermia	Exposure to Extreme Temperatures
Fires/Explosion	Fires and Explosion
Fall *….*	Falls
Struck by falling object/projectile	Collapse of Object
Collapse of structure	
Trench collapse	
Wall (earthen) collapse	
Electrocution by equipment contacting wire	Electrocution
Electrocution by touching exposed wire/source	
Electric shock other and unknown cause	
Electrocution from equipment installation/tool use	
Crushed/run-over of non-operator by operating cons	Struck By Moving Objects
Crushed/run-over by highway vehicle	

Overall the data partition looked like below:

Table (1) Data Partition

Type	MsiaAccident	Manual Tagging	Osha Commented Records	Total
Train Set	181	347	2493	3021
Test Set	51	125	1069	1245

2.3 Feature Engineering

We followed the different techniques taught for cleaning and preprocessing the data. This helped us feed the quality data for training our classification models. To be able to use the preprocessing code for different models, we have put it in a common util file named *"tm_assignment_util.py". The source and data is organized as below:*

1. *src -> Data-> Train_MsiaAccidentCases.xlsx*
2. *src->Data-> Test_MsiaAccidentCases.xlsx*
3. *src->Data-> osha.xlsx*
4. *src -> *.ipynb*

Here are the steps followed for the same:

2.3.1 Merge Title and summary columns:

We merged the title and summary columns in to a single Title_Summary_Case column. This helped us utilize the maximum information available in the document. Title column present important keywords which could help train a good model.

2.3.2 Remove 'Others' Column

Others column was added with 1 incomplete record and hence decided to be removed.

2.3.3 Word Clouds by Causes

We generated the various word clouds to understand the train data and words under each category on raw data. In an iterative process, it helped us clean and create meaningful tokens vector space model for the document. Below are the word clouds grouped by Causes. In the final phases after LDA, this helped us choose the right category on different clusters produced by LDA:

Figure (1) *Caught in/between Objects*

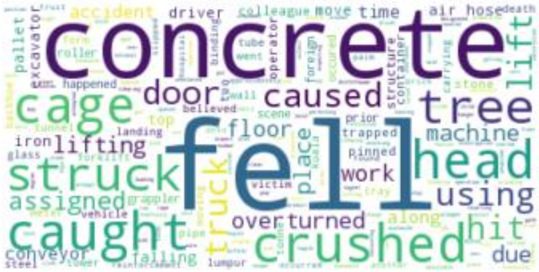


Figure (2) *Collapse of object*

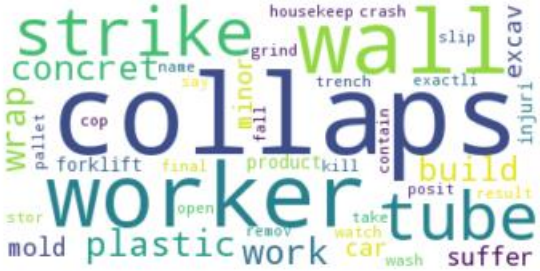


Figure (3) *Drowning*

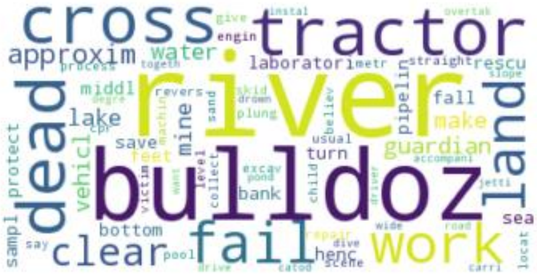


Figure (4) *Electrocution*

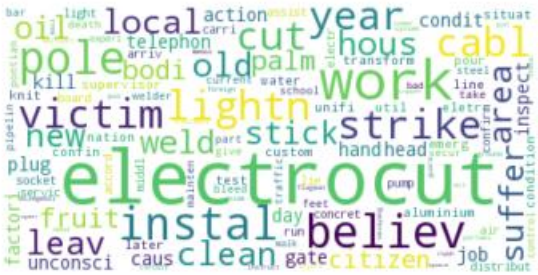


Figure (5) *Exposure to Chemical substances*

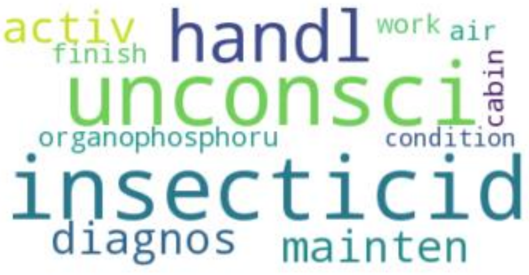


Figure (6) *Falls*

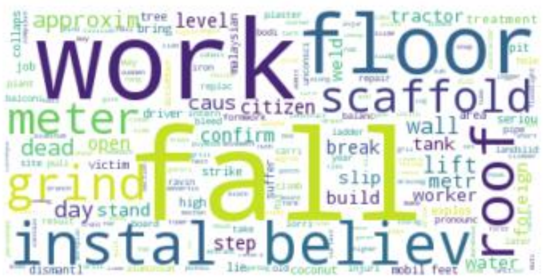


Figure (7) *Exposure to extreme temperatures*

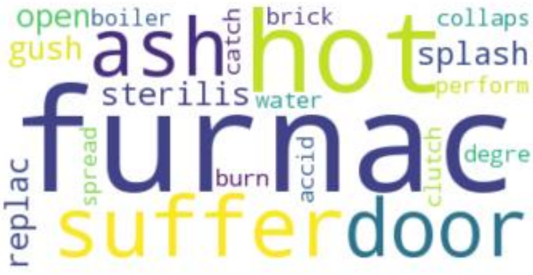


Figure (8) *Others*

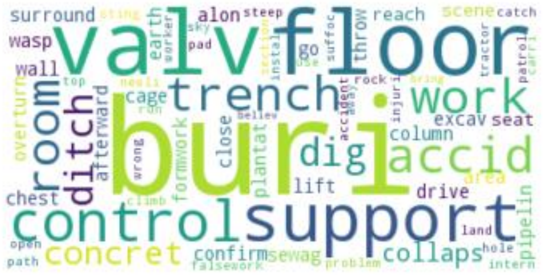


Figure (9) *Fires and Explosion*

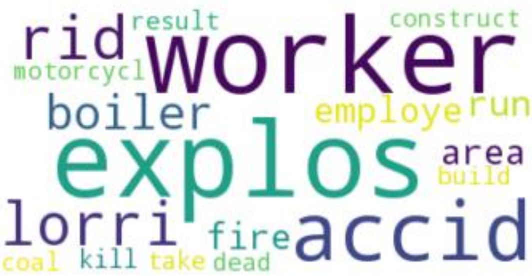


Figure (10) *Struck by Moving Object*

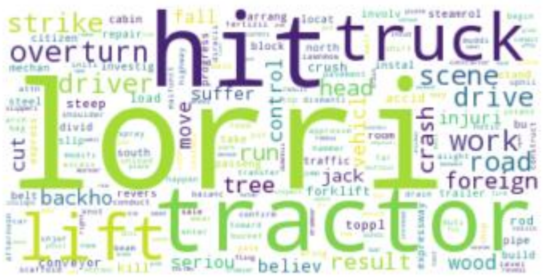
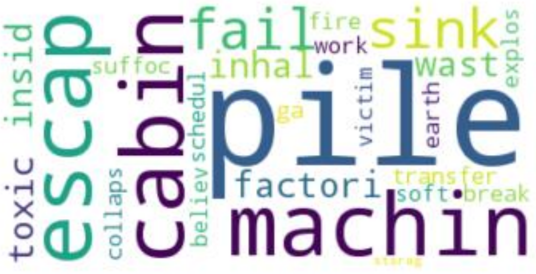


Figure (11) *Suffocation*



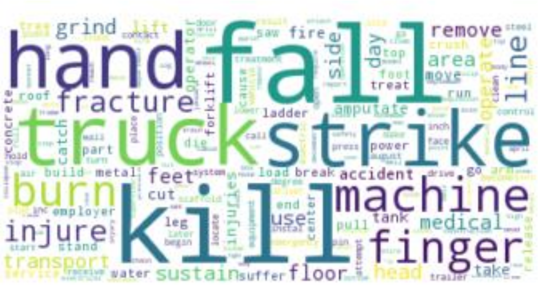
2.3.4 Data Cleanup

Below are the data cleanup steps:

1. Apply POS_TAG and remove function words
2. Remove punctuations
3. Change the words to lowercase
4. Consider words with length > 2
5. Apply stopwords
6. Apply nltk lemmatize to retrieve the base forms of the words

After working in an iterative way to improve upon the tokens, word cloud for *Osha data* was generated as shown below. We could immediately observe that *fall/strike* seems to be the major reasons of incidences and lorry/tractor are the major vehicles involved in the accidents.

Figure (12) *Document word cloud after clean-up Osha.xslx*



Since the above diagram doesn't quantify each cause weightage, we plotted Bar graph and Scree plot on Frequency distribution of top 10 words (Figure 13). We would also want to know as to what proportion of the text is taken up with such words (Figure 14). Here are the results:

Figure (13) Top 10 words Frequency Distribution

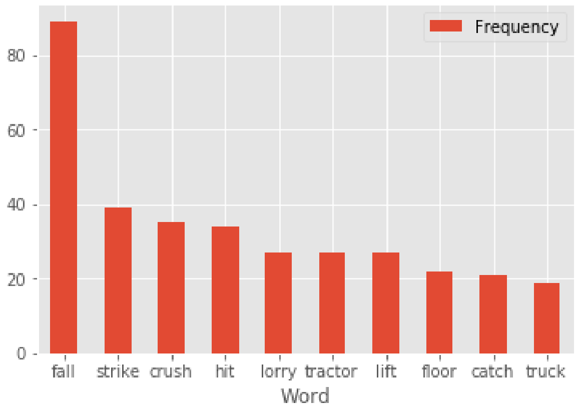


Figure (14) Proportion in whole text

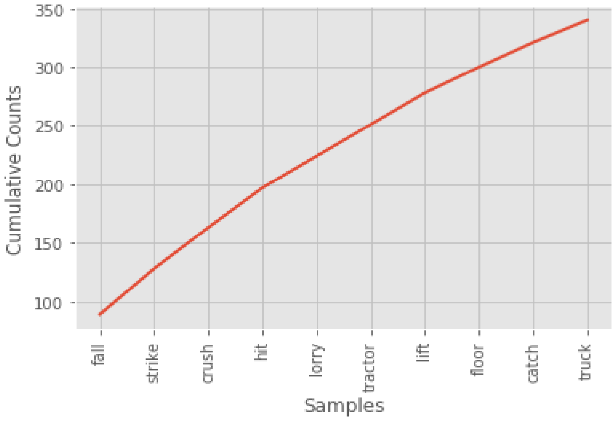


Figure (14) and **(15)** below shows the distribution in **Osha data**.

Figure (14) Top 10 words Frequency Distribution

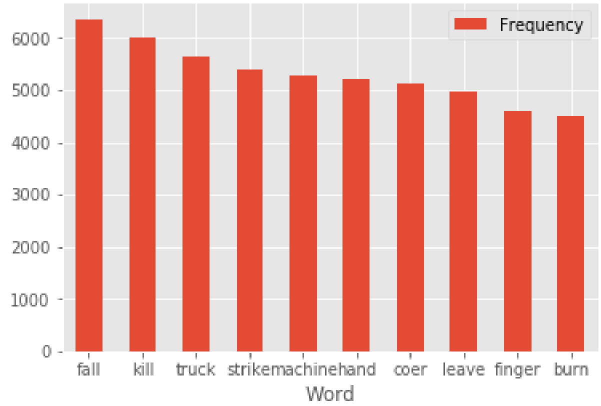
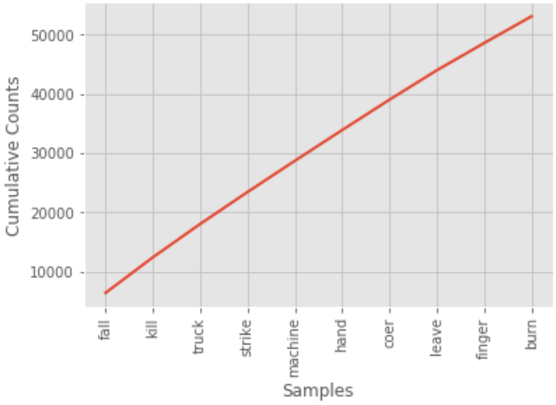


Figure (15) Cumulative Frequency Distribution



We can clearly observe that *"Fall"* remains the highest cause of injury and *"truck"* is the major vehicle involved in accidents. Another observation is *"Hand"* which seems to be top affected part of body in accident. This will further be analyzed as part of details analysis on body parts and occupations.

2.4 Model Building

2.4.1 Naïve Bayes

We started with Naïve Bayes model training and testing since the basic assumptions of Feature independence holds true. We are not required to do much semantic analysis and bag of words approach fits this problem statement. As learned through the course, Naïve Bayes has following features:

1. Very fast, Low Storage Requirement
2. Robust to irrelevant features
3. Very good in domains with many equally important features

The equation for the same is shown as below in Figure (16) below:

Figure (16) Multinomial Naïve Bayes Parameter Estimation

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Based on our final results, we found out that SVM and Random forests performed better than Naïve Bayes. Hence we stopped spending more effort on Naïve Bayes.

2.4.2 SVM

SVM is a well-known and promising algorithm for Text Categorization. It fits the properties of text as mentioned below:

High Dimensional input space: when categorizing text, one has to deal with many features. Since SVMs are overfitting protection which doesn’t necessarily depend on the number of features, they have the potential to handle these large feature spaces.

Few irrelevant features: One way to avoid these high dimensional input spaces is to assume that most of the features are irrelevant. Feature selection tries to determine those. However there are only few irrelevant features in text categorization and hence SVM is helpful without removing relevant features in attempt to reduce dimension

Document Vectors are sparse: For each document $d_{i,}$ the corresponding document vector contains only few entries which are not zero. SVM is proven to be well suited for problems with dense concepts and sparse instances.

Most text categorization problems are linearly separable: All assumed categories are linearly separable and so are many of the text categorization tasks. SVMs could be used effectively to find such linear separators.

Code Snippet for SVM is as shown below:

```
text_clf = Pipeline([('vect', CountVectorizer()),

                ('tfidf', TfidfTransformer(use_idf=True)),

                 ('clf', SGDClassifier()) ])
```

Please note that initially we kept the use_idf = False since the data was skewed however when we attained a good number of records in train/test data. We turned the flag ON.

2.4.3 Random Forest

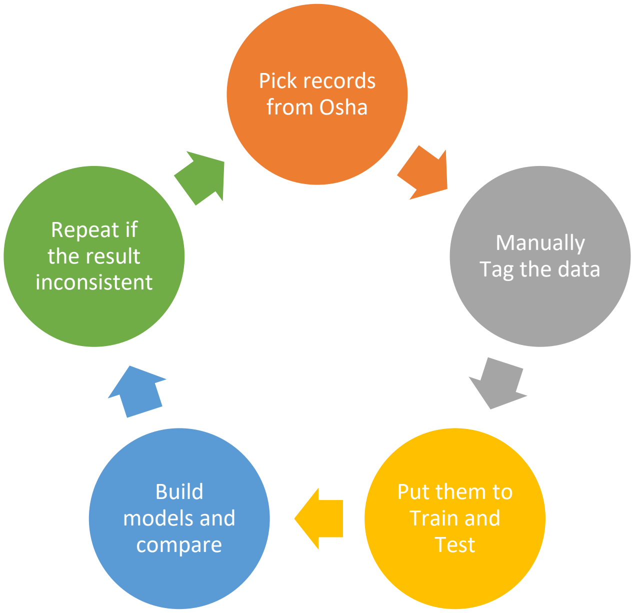
Random forests combine the predictions of multiple decision trees. The dataset is repeatedly divided into subtrees, guided by the best combination of variables. However, finding the right combination of variables can be difficult. For instance, a decision tree constructed based on a small sample might not be generalizable to future, large samples. To overcome this, multiple decision trees could be constructed, by randomizing the combination and order of variables used. The aggregated result from these forests of trees would form an ensemble, known as a random forest. This is an ensemble method and proved to be effective in various classification algorithms.

2.5 MODEL PERFORMANCE AND TUNING

Unlike the standard data, text data poses many issues to baseline the model performance. In our experiment, we noticed that when the Osha data was predicted using the different algorithms like SVM, Naïve Bayes and Random Forest, results were not consistent. There was big difference between the categories reported by different algorithms. Hence, we were not confident about our models stability.

To attain model stability, following approach was followed:

Figure (16) Process to achieve Stable classification between models



Below is the comparison of models shown ***Before and After*** the manually tagging and Predicting on Osha data:

Type of Accident	SVM	Random Forest	SVM	Random Forest	SVM	Random Forest
	Before the manual tagging		**After the manual tagging**		**After adding 3.5k tagged data**	
Caught in/between Objects	3279	4575	5487	6053	4026	4162
Collapse of object	216	73	597	431	1442	1566
Drowning	373	248	188	105	175	122
Electrocution	1551	3127	1205	949	859	964
Exposure to Chemical substance	436	135	577	295	879	579
Exposure to extreme Temperature	1204	102	1224	1413	578	292
Falls	4181	5232	2602	2455	2122	3146
Fires and Explosion	972	89	1177	490	1329	1226
Other	273	96	496	214	223	67
Struck By Moving Objects	3545	2578	2251	3651	1020	635
Suffocation	293	68	519	267	177	71

Below table shows the different metrics comparison:

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	~61%	.64	.61	.58
SVM	~83%	.83	.82	.82
Random Forest	~78%	.77	.76	.76

SVM performed the best for us and the classification report for different categories are shown as below:

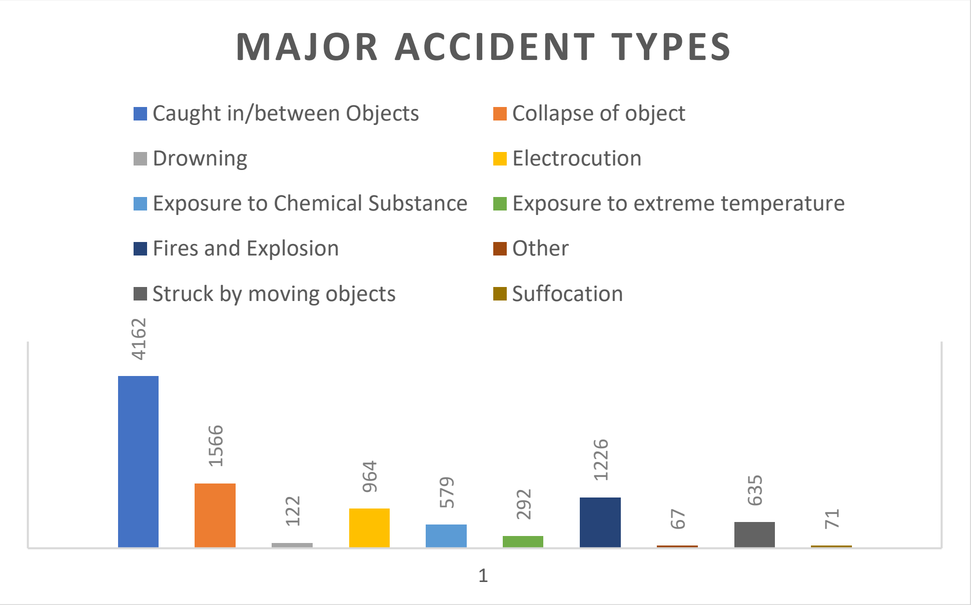
Causes	Precision	Recall	F1-score	support
Caught in/between objects	0.69	0.64	0.67	129
Collapse of object	0.65	0.77	0.70	132
Drowning	0.75	0.88	0.81	17
Electrocution	0.89	0.94	0.92	35
Exposure to Chemical Substances	0.71	0.67	0.69	30
Exposure to extreme temperatures	0.67	0.63	0.53	23
Falls	0.92	0.90	0.91	413
Fires and Explosion	0.64	0.81	0.72	36
Other	0.88	0.78	0.82	9
Struck By Moving Objects	0.51	0.55	0.50	53
Suffocation	0.89	0.71	0.68	29

2.6 Answers To Problem Statements

2.6.1 Major types of accidents

Based on the classification results of our best algorithm (SVM), below **Figure (17)** shows the distribution of accidents accident types:

Figure (17) Major Types of Accidents

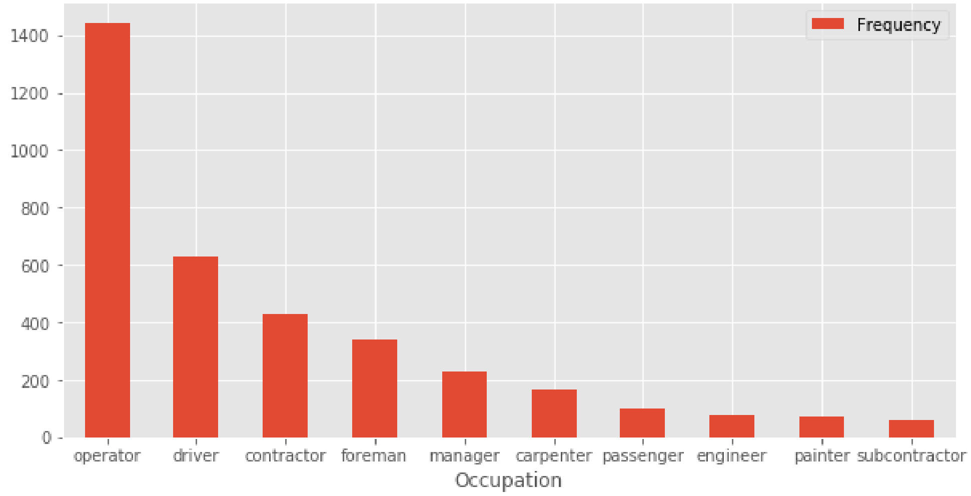


2.6.2 Riskier occupations

We followed the following process to extract the occupation names:

1. Using the Osha data prepared a preliminary occupation list
2. Observed the keyword in each category and put them in to word2Vec vocabulary
    a. Found out top 30 similar words for that keyword
3. List of occupations from word2Vec were added to the list prepared at step (1)
4. Fetch the Noun, Singular Noun and Plural Nouns using Regex parser

Figure (18) Riskier Occupations

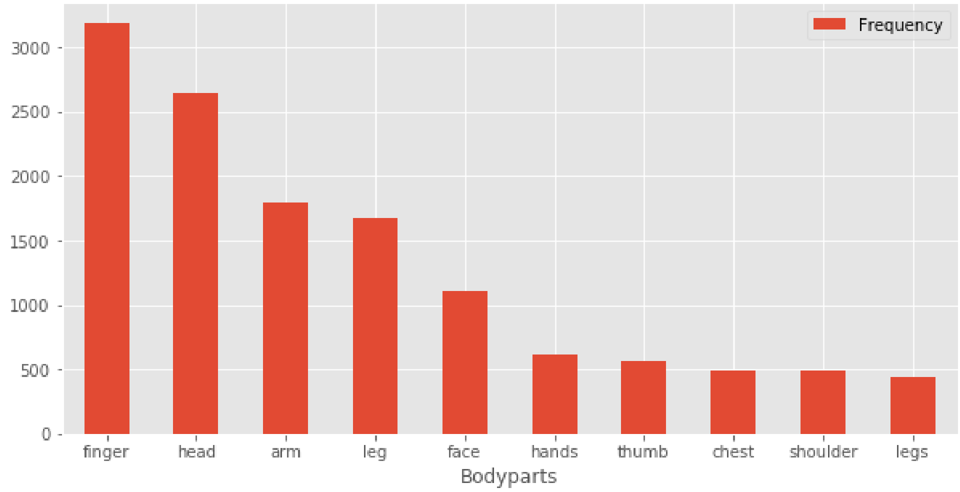


2.6.3 Human body prone to injuries

Similar approach as Riskier occupations was followed to extract the human body frequency distribution:

1. Using the Osha data prepared a preliminary body part list
2. Observed the keyword in each category and put them in to word2Vec vocabulary
    a. Found out top 30 similar words for that keyword
3. List of occupations from word2Vec were added to the list prepared at step (1)
4. Fetch the Noun, Singular Noun and Plural Nouns using Regex parser

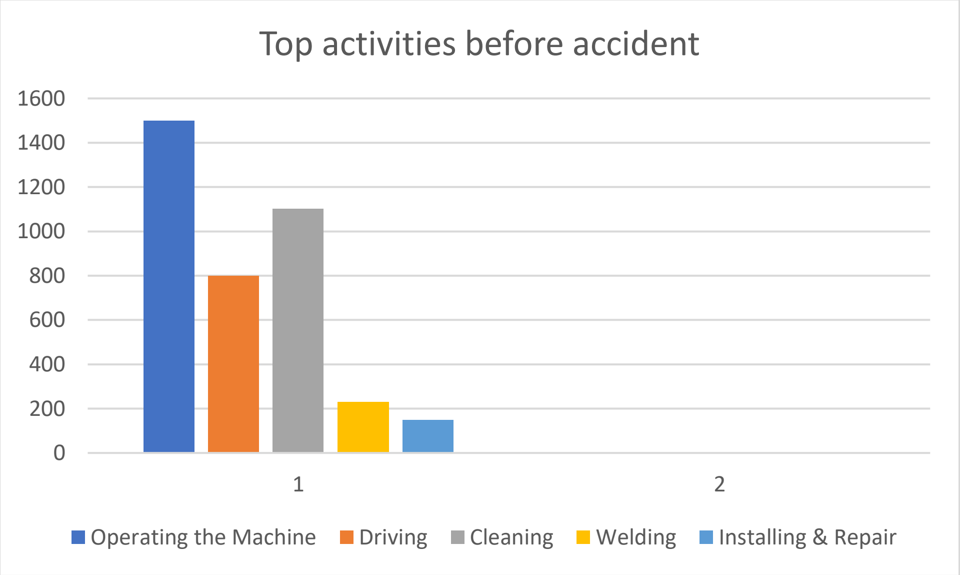
Figure (19) Body Parts prone to Injuries



2.6.4 Common activities that victims were engaged

We followed the following process to extract the common activities that victims were engaged in before accident took place:

1. Using the Osha data prepared a preliminary verb list
2. Using LDA we categorized these verbs into 5 topics
3. List of verbs in each topic were displayed in the order of relevance
4. Removed those verbs which were related to accidents like crushed, fall etc.
5. Finally filtered out the most common topics along with the top verbs associated with that topic.



2.6.5 Recommendation to improve safety measures

Based on the above data findings, following recommendations should be taken for safety measures:

1. Safety gloves and helmets should be enforced while carrying out work
2. Activities like machine handling, Cleaning tasks are activities carried out before accident and

needs extra checklist to carry out the task

3. A smart app could be built for supervisors who could be advised for the critical risks of a scheduled tasks and related suggestions could be followed as part of safety procedures.

2.7 REFERENCES

1. http://www.nltk.org/book/ch01.html
2. https://web.stanford.edu/class/cs124/lec/naivebayes.pdf
3. https://code.google.com/archive/p/word2vec/