TEXT MINING REPORT

A novel approach for accident risk mitigation in the construction industry

TEAM : FT12

SHIVARAM ANDIYAPPAN SELVARAJ - A0163267B
JOSEPH PABLO VARUN -A0163323N
ARUN SUGUMAR GURUMOORTHY - A0163347A
PRAVEEN KUMAR GUDDALA KUMARAN - A0163242N
SOWNDARYA SATYAMURTHI -A0163229E
HAO SUYA -A0163339B

LIST OF CONTENTS

EXECUTIVE SUMMARY

Text mining can be used to make the large quantities of unstructured data accessible and useful. Machine learning algorithms can be applied to large document databases. Data in the documents are unstructured,thus making it inappropriate for applying conventional data mining techniques. Whereas, Text mining approach consists of techniques that can be applied over the text documents to derive useful insights .

Text mining has a wide range of applications that include:

1. Finding meaning from unstructured text data
2. Text classification and categorisation
3. Information extraction
4. Identification of relevant text documents
5. As an aid to improve accuracy in predictive modelling or unsupervised learning.

This report provides an analysis and evaluation of application of text mining techniques on the fatality and catastrophe investigation summary reports obtained from the workplace. The summaries provide a complete description of the incident,further analysis can be made on the data to mitigate the identified risks and prevent the occurrence of similar accidents. The methods of analysis include tokenization,stemming,lemmatization,stop word removal and model building.

The business objective for this scenario is to mine the accident summary data to extract answers for following objective questions:

1.Which type of accidents (in terms of main causes) are more common in fatal or catastrophic accidents?

2. What are the more risky occupations in such accidents?

3.Which parts 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?

To answer the above questions, various text mining techniques are studied and compared against one another to check the suitability for getting an answer for the above business questions. This includes identifying the part of speech and subject-object dependencies.Since text mining is an iterative process, the process is repeated for different iterations and the best results out of these iterations are produced.

The answers to the business questions are the findings obtained from the results of analysis. Observations show that caught between objects or struck between moving objects are the most common types of accidents in construction industry. Some other incidents are falling, fire and explosion and electrocutions. The analysis for finding most common objects associated with the accidents are forklifts, machines, roof and ladder.Apart from the incidents mentioned,there are numerous other incidents. More risky occupations associated with accidents are identified as machine operator, carpenter, and laborer. Common activities during which accidents occur are operating heavy vehicles forklift and cranes.

The findings could be used to focus on the most common factors that influence accidents in the construction industry and the people associated with the accidents. The personnel can be provided special trainings and can be provided safety apparels during the work sessions to mitigate the risk of accidents. The infrastructure and the equipments causing accidents can be altered to mitigate accidents.

INTRODUCTION:

The aim of this assignment is to analyse the two given datasets *MsiaAccidentCases.xlsx* ,*osha.xlsx* and *MsiaAccidentCases.xlsx* is a small sample of accident summaries collected from Malaysia, which have been manually labelled with 9 main causes of accidents. The other document, *osha.xlsx*, is much larger and it's not labeled with the main causes.

In this report,we are using text mining techniques learned in class to analyze the data, extract useful information, build necessary models and presented our findings on the four questions. We have used python language and the relevant packages for our work and have presented our findings in the form of graphs ,charts and word clouds.

The business goal is to provide a detailed analysis of various accidents take place in the workplace to the construction industry and thereby enhancing the safety measures. The aim of text mining in this context is to extract useful information from the unstructured data from the *Osha.xlsx* and to provide insights to discover the more frequent hazards occurred over the certain period .

QUESTION 1:

Which type of accidents (in terms of main causes) are more common in fatal or catastrophic accidents?

Techniques used: supervised learning and Frequency.

The causes are recategorized into 7 categories as follows:

1.Caught in/between Objects/Collapse of object

2.Electrocution

3.Falls

4.Fires and Explosion / Exposure to extreme temperatures / Chemical substance

5.No air -- Drowning/Suffocation

6.Struck By Moving Objects

7.Stung by Wasps

Data Cleaning:

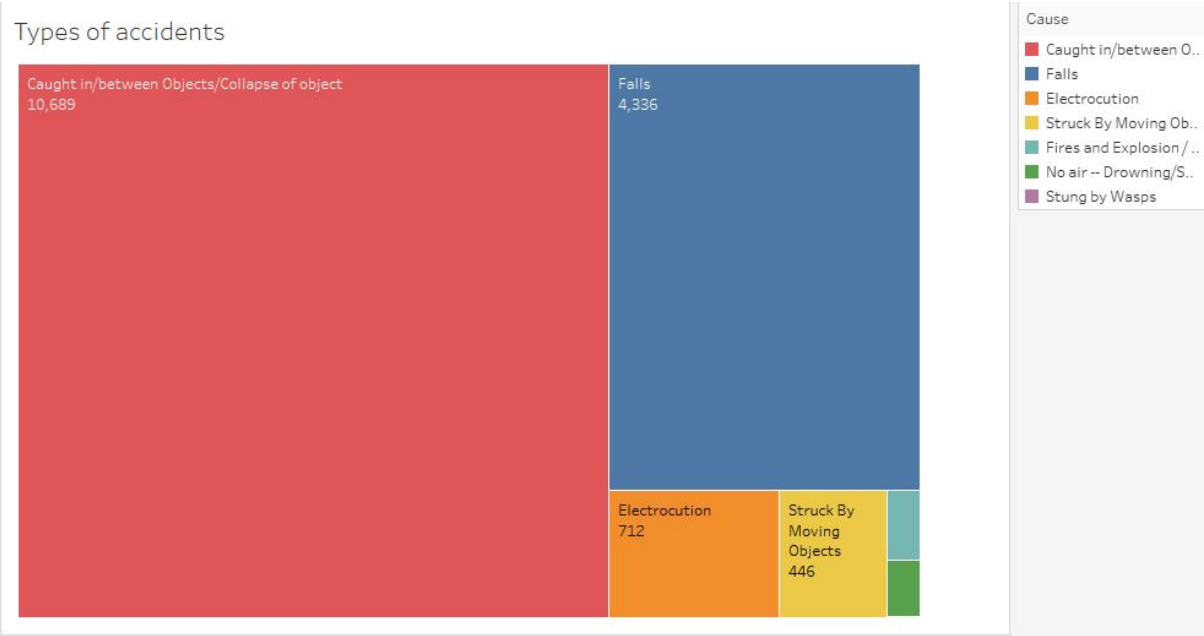
Stop word removal , lemmatization were performed.

1. Based on the reclassified cases, Naïve Bayes, SVM, SVM with stochastic gradient descent classifier and Decision Tree models were built, which was also checked with test set and compared the accuracy.
2. SVM with stochastic gradient descent classifier model gives the best accuracy. This model was applied to the large dataset to predict the causes of the cases. Predicted results of larger data set were checked at random manually.
3. Then frequency function was applied on the predicted causes and got the top 2 common accidents are "Caught in/between Objects/Collapse of object" and "Falls". The Frequency table as follow:

Cause	
Caught in/between Objects/Collapse of object	10,966
Falls	3,300
Electrocution	942
Fires and Explosion / Exposure to extreme temperatures / Chemical subst..	520
Struck By Moving Objects	418
No air -- Drowning/Suffocation	174
Stung by Wasps	3

MODELS	ACCURACY
NAIVE BAYES	51%
DECISION TREES	60%
SVM	72%
SVM WITH STOCHASTIC GRADIENT DESCENT	77%

DASHBOARD :



WORD CLOUD REPRESENTATION :

Word Cloud 1

Using SVM with stochastic gradient descent classifier model

Word Cloud 2
Using POS tagging

QUESTION 2:

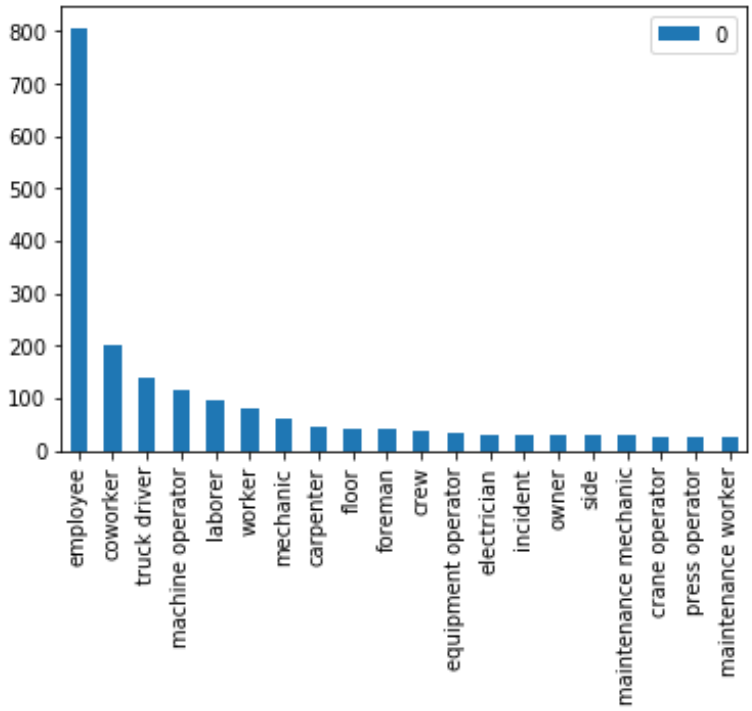
1. What are the more risky occupations in such accidents?

Method:

A new header is added to the Osha dataset. Among the columns ,the 'title' column is taken and tokenized. After the chunking process,the grammar rules are checked .Finally the nouns are extracted and written to the csv file 'occupation.csv'. The output csv file 'occupation.csv' contains all the occupations involving accidents .

RESULTS:

we get this bubble chart(after removing employee and coworker) showing the numbers with the frequency of occupation.



QUESTION 3:

Which parts of human body are more prone to be injured in such accidents?

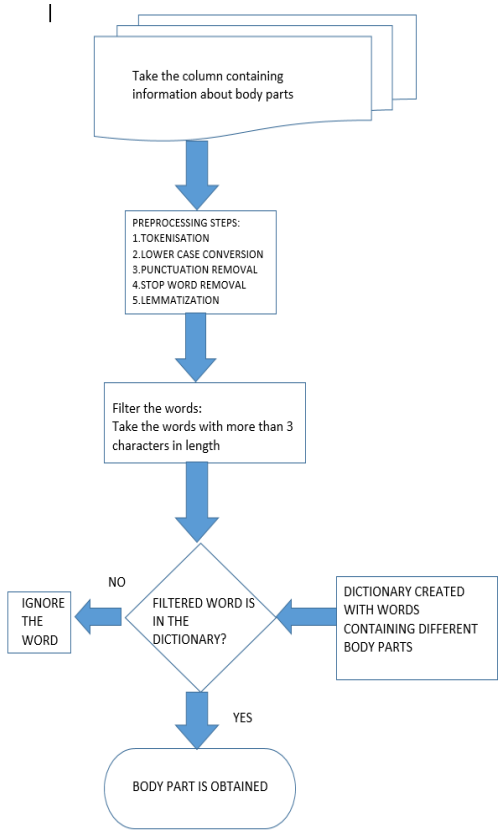
Method:

For finding the most affected body parts , there are some steps to be followed as described in the flowchart below:

The data file Osha.xlsx is taken and given a header for its columns for our reference as shown in 'copy of Osha.xlsx' data. The content in the column 'type' has information about the body parts injured during the accident.

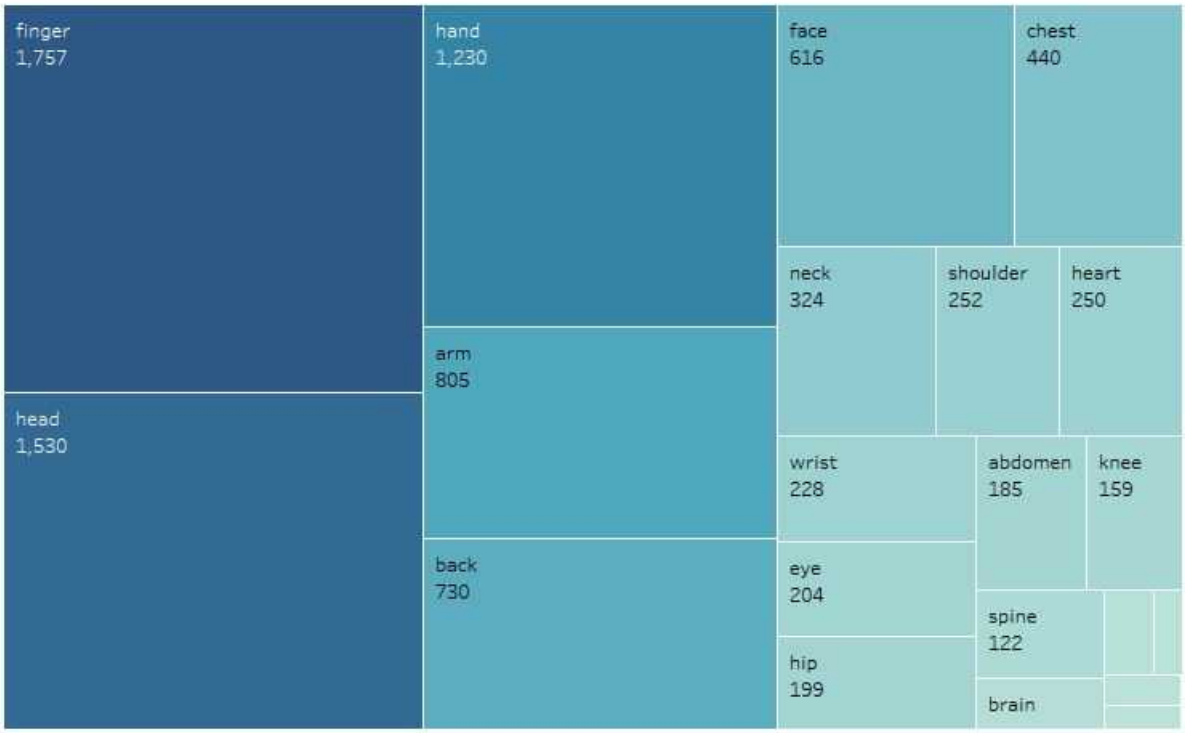
A dictionary is created which contains all the body parts that are affected in an accident.This dictionary is built with thorough study on the effect of accidents on human morphology and the references are mentioned in the annexure.

FLOWCHART FOR THE PROCESS



RESULTS

DASHBOARD



Word Cloud for affected body parts:

FREQUENCY CHART FOR AFFECTED BODY PARTS

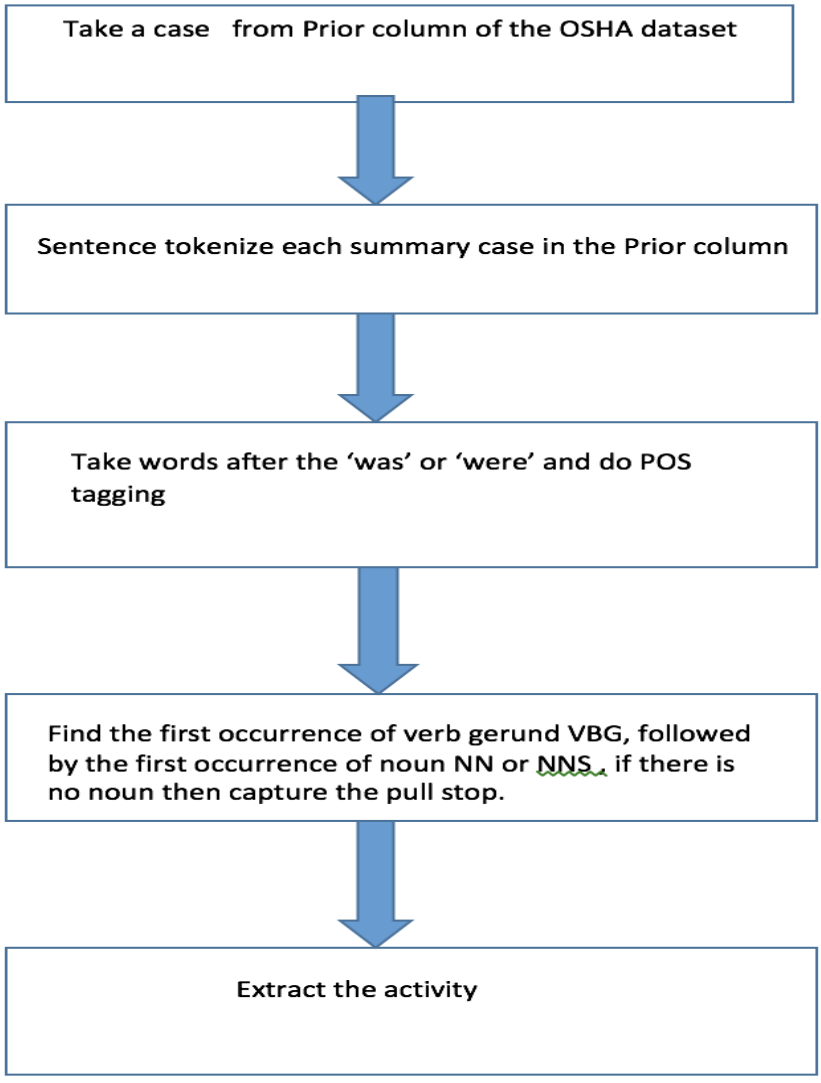
Col	
finger	1,757
head	1,530
hand	1,230
arm	805
back	730
face	616
chest	440
neck	324
shoulder	252
heart	250
wrist	228
eye	204
hip	199
abdomen	185
knee	159
spine	122
brain	70
stomach	45
buttock	27
kidney	26
palm	20

QUESTION 4

What are the common activities that the victims were engaged in prior to the accident?

The summary case in the column named 'Prior' contains the summary case of activities before the accident. The first sentence contains the activity done before the accident.The first sentence is taken and processed as per the flowchart below and the activities are extracted.

PROCESS FLOW:



RESULTS:

Following the above process, the activities prior to the accident are filtered and written to a csv file. This file can be found in /Final_deliverables/Question_4/activity.csv.

CONCLUSION

FINDINGS:
1.We have labelled the OSHA dataset based on the model built from Msia dataset.We have grouped the initial nine categories in the Msia dataset into 7 categories.We can find that the Msia data is insufficient and Osha has more information. The training data is inadequate . Combining the categories helped us in reducing the redundancy of the accident causes.SVM with stochastic gradient descent gives more accuracy compared to other models.
2.Occupations such as truck driving,machine operating and labour possess more serious threats.
3.More injuries happen to the finger,head and hands .The workplace should be provided with safety apparels such as helmets and gloves to prevent these injuries.

4.Activities such as driving heavy vehicles , working with metals such as foundry ,ceiling and roof work can lead to accidents ,hence such activities should be done with more precautions.

Challenges Faced:
1.As there are many different cases with different words,for example driving a truck riding a car ,information extraction is difficult for these types.
2. We faced difficulty in writing grammar rules due to the limitations in POS tagging such as picking ‘noun’ as ‘noun’ and ‘verb’ as ‘verb’ .
3. While extracting ‘activities’ ,the rule did not work if there is no ‘was’ or ‘were’

FURTHER STUDY

The synonyms of words and the different forms of verbs should be studied in depth to improve the classification accuracy.

HOW THE RESULTS CAN BE USED?

1. The results of questions 2 and 3 can be looked upon and mismatched cases can be further researched to identify better grammars to extract the answers with higher accuracy.

2. The result of the classifier on Osha can be used to identify misclassified results so that we can train more on those cases by collecting more training data for such misclassified data.

FILE NOMENCLATURE

Please refer the table for the location of various files containing the programs for various questions:

QUESTIONS	PATH
QUESTION 1	/Final_deliverables/Question_1
QUESTION 2	/Final_deliverables/Question_2
QUESTION 3	/Final_deliverables/Question_3
QUESTION 4	/Final_deliverables/Question_4

REFERENCES

1.nltk.tag.stanford — NLTK 3.2.5 documentation
2.Preprocessing Techniques for Text Mining - An Overview
Dr. S. Vijayarani1, Ms. J. Ilamathi2, Ms. Nithya3
Assistant Professor1, M. Phil Research Scholar2, 3
Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India1, 2, 3
3. Sentex - youtube channel
4.Class notes and sample programs.
4.https://www.safeopedia.com/definition/687/human-factors-causing-accidents
5.https://www.bls.gov/opub/ted/2015/type-of-injury-or-illness-and-body-parts-affected-by-nonfatal-injuries-and-illnesses-in-2014.htm
6.http://www.hopesandfears.com/hopes/now/question/213759-which-body-part-gets-injured-the-most
7.https://www.britannica.com/science/human-disease/Physical-injury