**Interim Report|NLP - Capstone Project Group 13 | 2022**

Contents

[1. Team Members 1](#_Toc100417781)

[2. Summary of problem statement, data and findings 1](#_Toc100417782)

[1. Problem Statement 1](#_Toc100417783)

[3. Summary of the approach to EDA and Pre-processing 2](#_Toc100417784)

[1. EDA 2](#_Toc100417785)

[2. Data Pre-processing 11](#_Toc100417786)

[4. Deciding Models and Model Building 11](#_Toc100417787)

[5. How to improve model performance 11](#_Toc100417788)

# Team Members

* Kuldeep
* Shushil
* Raghunathan
* Divya
* Aditya

# Summary of problem statement, data and findings

## Problem Statement

Given a dataset of injuries and accidents in industrial plants we need to build a model (and later chatbot) which will help professionals to obtain information about safety risk given the incident description. This project involves building a chatbot (ties to a model) wherein users would go in type in their incident description and get an output of safety risk. Thus, it is similar to classification analysis model and from technical side building a UI for this classification-based model.

Dataset provided has the below columns:

* Data: timestamp or time/date information
* Countries: which country the accident occurred (anonymized)
* Local: the city where the manufacturing plant is located (anonymized)
* Industry sector: which sector the plant belongs to
* Accident level: from I to VI, it registers how severe was the accident (I means not severe but VI means very severe)
* Potential Accident Level: Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors
* involved in the accident)
* Genre: if the person is male of female
* Employee or Third Party: if the injured person is an employee or a third party
* Critical Risk: some description of the risk involved in the accident
* Description: Detailed description of how the accident happened

# Summary of the approach to EDA and Pre-processing

### EDA

For performing EDA, we converted the Data column to datetime and created additional 2 columns for Month and Year of accident. As per the attached EDA report we carried out the below analysis using the data provided. With the plots we can see the relations between the categorical datapoints and the number of accidents by levels and using Chi-Square test we can analyze the impact of factors like Country, Local, Industry Sector and Employee or Third Party.

1. **Plot of Accident Level by Country**

From below Country\_01 has the highest level of accidents and level 1 accidents too

Chart, bar chart, waterfall chart

Description automatically generated

1. **Plot of Accident Level by Location**

From below Local\_03 has the highest level of accidents and level 1 accidents too

Chart, bar chart, histogram

Description automatically generated

1. **Plot of Accident Level by Industry Sector**

From below Mining has the highest level of accidents and level 1 accidents too

Chart, bar chart, waterfall chart

Description automatically generated

1. **Count plot of Accident Level**

From below level 1 accidents are most frequent

Chart, bar chart

Description automatically generated

1. **Count plot of Potential Accident Level**

From below level 4 accidents are most Potential Accident Level

Chart, bar chart, histogram

Description automatically generated

1. **Count plot of Potential Accident Level w.r.t Accident Level**

From below level 4 accidents are most Potential Accident Level but level 1 accidents are high in occurrence

Chart, bar chart

Description automatically generated

1. **Count plot of Genre w.r.t Accident Level**

From below Male Genre has the most accidents

Chart, waterfall chart

Description automatically generated

1. **Count plot of Employee or Third-Party w.r.t Accident Level**

From below Third Party and Employee both have almost same accident occurrences with high occurrence in leve1 1 accidents

Chart, bar chart

Description automatically generated

1. **Count plot of Critical Risk**

From below Others have the most Critical Risk

Chart

Description automatically generated

Text

Description automatically generated

1. **Count plot of Accident Level by Month**

From below Accidents mostly occur during the first 6 months of the year

Chart, bar chart, histogram

Description automatically generated

1. **Count plot of Accident Level by Year**

2016 has had more accidents as compared to 2017

Chart, bar chart

Description automatically generated

Since the data is not in numeric format, we could not use correlation maps directly and had to perform Chi-Square tests. Below Chi-Square tests were performed.

1. **Impact of Country on Accident Level**

Graphical user interface, text, application

Description automatically generated

1. **Impact of Local on Accident Level**

Graphical user interface, text, application

Description automatically generated

1. **Impact of Industry Sector on Accident Level**

Graphical user interface, text, application

Description automatically generated

1. **Impact of Employee or Third Party on Accident Level**

Graphical user interface, text, application

Description automatically generated

**Word Cloud:**

After removal of stop words from the Description column and lower casing the words in the Description column, below is the Word Cloud.

Text

Description automatically generated

### Data Pre-processing

Below are the data preprocessing steps carried out:

* Lower casing words in the Description column
* Removal of stop words from the Description column
* <Further preprocessing to be carried out during model building>

<Rouge as high as possible>

# Deciding Models and Model Building

For deciding the model for the application, we used the below approaches:

**Embedding: TFID**

Model Performance using TFID: From below we can see that Logistic Regression provides the best accuracy and AUC hence used the same as the best TFID model for the classification task.

Table

Description automatically generated

# How to improve model performance