greatlearning
Power Ahead

AIML

CAPSTONE PROJECT

CAPSTONE PROJECT - NLP

FINAL REPORT

MENTOR

MOUSHMI DAS GUPTA

GROUP MEMBERS

NISHTHA ARORA

BARNALI CHATTERJEE

SURAJ P SULLADMATH

JANUARY 2022

TABLE OF CONTENTS

INTRODUCTION

The International Labour Organization (ILO) estimates that some 2.3 million women and men around the world succumb to work-related accidents or diseases every year; this corresponds to over 6000 deaths every single day. Worldwide, there are around 340 million occupational accidents and 160 million victims of work-related illnesses annually. [1].

Industrial accidents cause huge damage to human lives, families of these victims, the industry as well as the environment. To predict the accident, it is significant to investigate past accidental reports. Based on the acquired knowledge safety experts can make the right move to evacuate or decrease the cause of an accident. Abused protective equipment, unmaintained safety articles, and catalogues increase the occurrence of an accident. Performing obligatory safety checks before operating a machine, bringing issues to light, frequent auditing and inspection of the machines would reduce the cause of accidents.[2].

With the advancement in technology, the Natural Language Processing (NLP) technique could also help in-detailed analysis of past accidents with the help of previously collected data and warn humans about the factors that may lead to a severe accident. The current project focuses on a similar problem and would create a chatbot for helping employees, by analyzing the industrial accidents from one of the biggest industries in Brazil and the world.

PROBLEM STATEMENT

Domain: Industrial safety. NLP based Chatbot

Context: The database comes from one of the biggest industries in Brazil and in the world. It is an urgent need for industries/companies around the globe to understand why employees still suffer some injuries/accidents in plants. Sometimes they also die in such an environment.

Data Description: This database basically records accidents from 12 different plants in 03 different countries which every line in the data is an occurrence of an accident. Columns description: ‣ Data: timestamp or time/date information ‣

Countries: which country the accident occurred (anonymised) ‣ Local: the city where the manufacturing plant is located (anonymised)

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)

Gender: 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

Description: Detailed description of how the accident happened

TECHNIQUES USED

Data Visualization and Analysis:

Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends and correlations that might not otherwise be detected can be exposed.

The benefits of data visualization:

When considering business strategies and goals, data visualization benefits decision-makers in several ways to improve data insights. Let’s explore seven major benefits in detail:

- Better analysis
- Quick action
- Identifying patterns
- Finding errors
- Understanding the story
- Exploring business insights
- Grasping the Latest Trends

Natural Language Processing

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand the text and spoken words in much the same way human beings can.

NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to ‘understand’ its full meaning, complete with the speaker or writer’s intent and sentiment.

PROCESS FLOW ANALYSIS

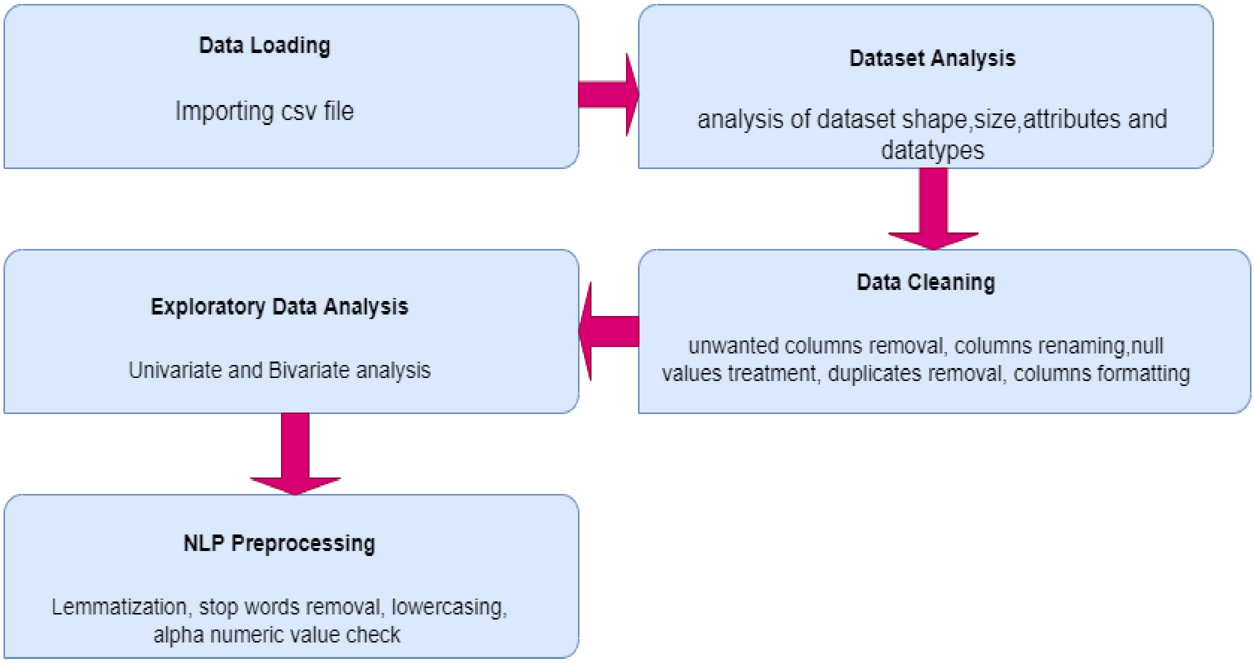

Data Loading
Importing csv file
Dataset Analysis
analysis of dataset shape,size,attributes and datatypes
Data Cleaning
unwanted columns removal, columns renaming,null values treatment, duplicates removal, columns formatting
Exploratory Data Analysis
Univariate and Bivariate analysis
NLP Preprocessing
Lemmatization, stop words removal, lowercasing, alpha numeric value check

FIG 1: DATA CLEANING AND PREPROCESSING FLOW

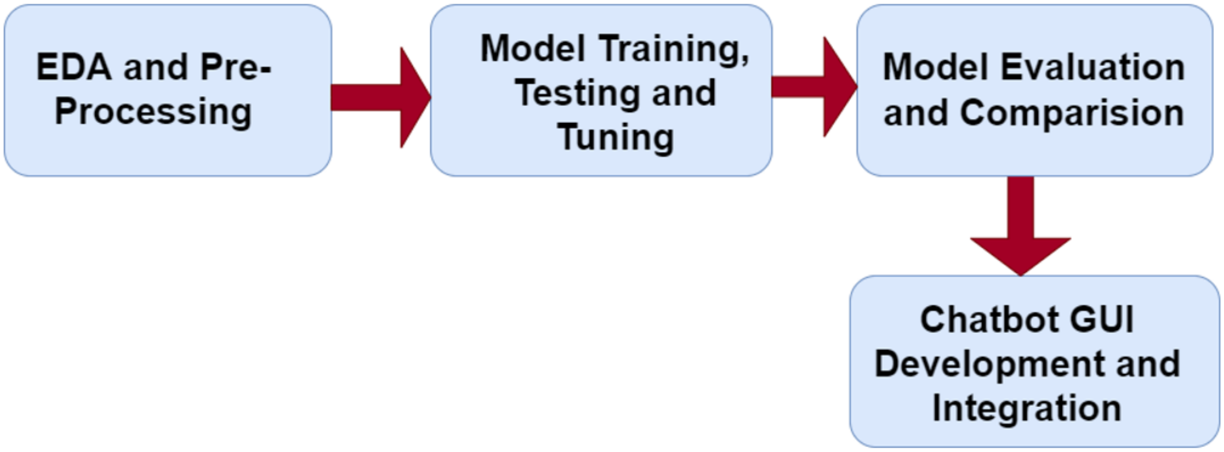
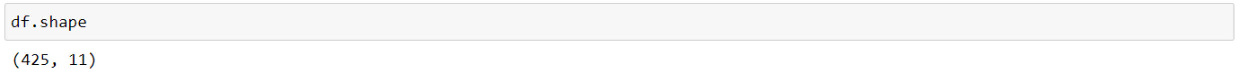

EDA and Pre-Processing
Model Training, Testing and Tuning
Model Evaluation and Comparision
Chatbot GUI Development and Integration

FIG: COMPLETE PROJECT FLOW

PROBLEM INTERPRETATION

Following are the details of the raw dataset :

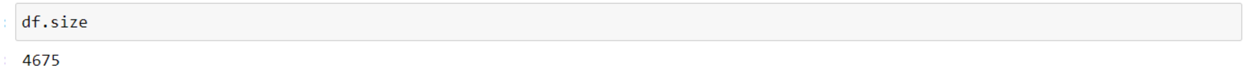
Shape:



```
df.shape
```

```
(425, 11)
```

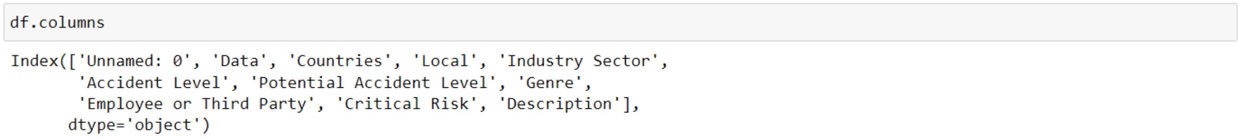
Size:



```
df.size
```

```
4675
```

Columns:



```
df.columns
```

```
Index(['Unnamed: 0', 'Data', 'Countries', 'Local', 'Industry Sector',
       'Accident Level', 'Potential Accident Level', 'Genre',
       'Employee or Third Party', 'Critical Risk', 'Description'],
      dtype='object')
```

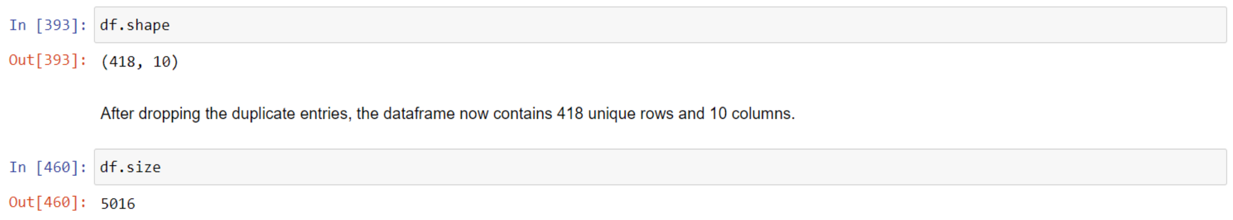
The dataset contains 425 rows and 11 columns

DATA CLEANING:

- The **Data** column contains information about the date and time of the accident. So, the column heading **Data** has been replaced with **Date** to match its description.
- The Genre column contains information about the gender of the victim. So, **Genre** has been replaced with **Gender**.
- The first column **Unnamed: 0** does not contain valuable information related to the accidents. So, this column has been dropped from the dataset.
- No Null/missing values are present.

- Total of 7 duplicate rows are present in the dataset in which all column values are same. This redundant information has been removed.

After performing the steps mentioned above, the transformed dataset shape and size are as follows:



```
In [393]: df.shape
Out[393]: (418, 10)
```

After dropping the duplicate entries, the dataframe now contains 418 unique rows and 10 columns.

```
In [460]: df.size
Out[460]: 5016
```

DATA VISUALIZATION AND ANALYSIS

The univariate,bivariate and multivariate analysis of the dataset helps in understanding the data and finding insights on specific patterns followed in the dataset. The below section performs the analysis on each of the column values provided.

Univariate Analysis

DATE:

Description:

Provides timestamp or time/date information of the accident

Format:

The Date column consists of information in the following format:

YYYY-MM-DD HH:MM:SS

Preprocessing:

The HH,MM,SS section in the Date does not contain any information as these values are marked as zero for all the rows. Hence, they have been removed from the Date column and only the year,month and day have been kept in YYYY-MM-DD format.

Visualization and Analysis:

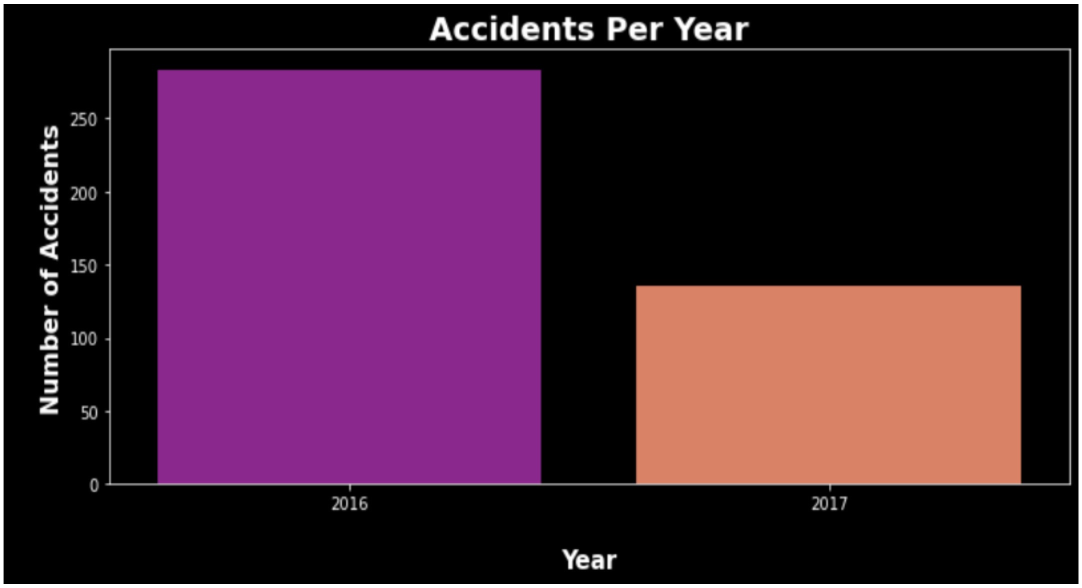


FIG 2: Accidents per Year

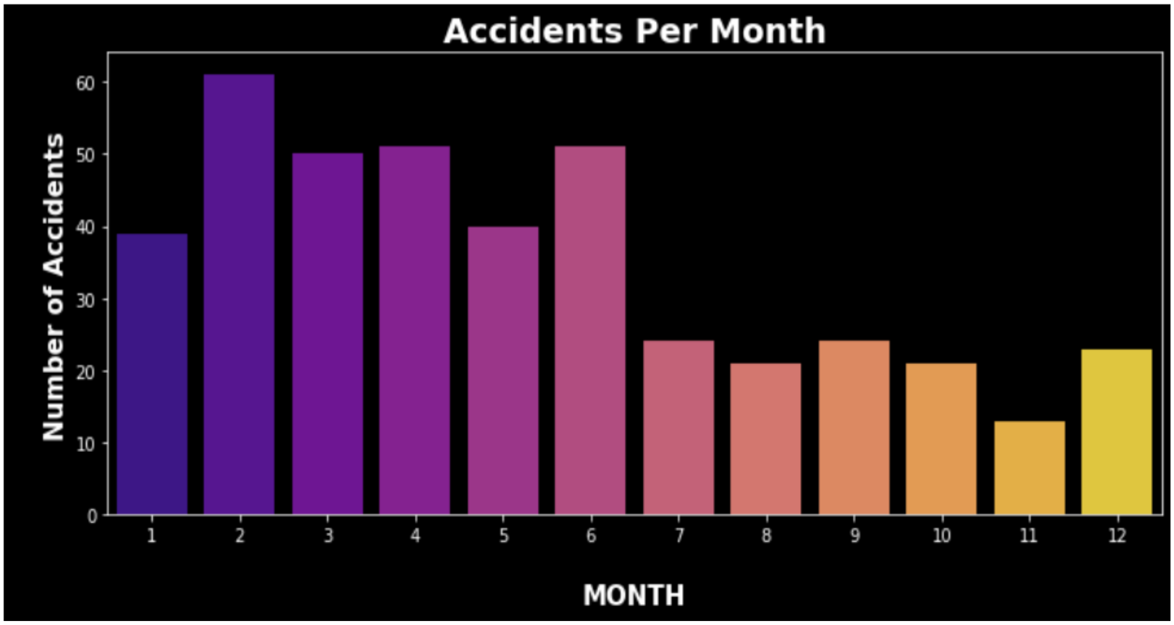


FIG 3: Accidents per Month

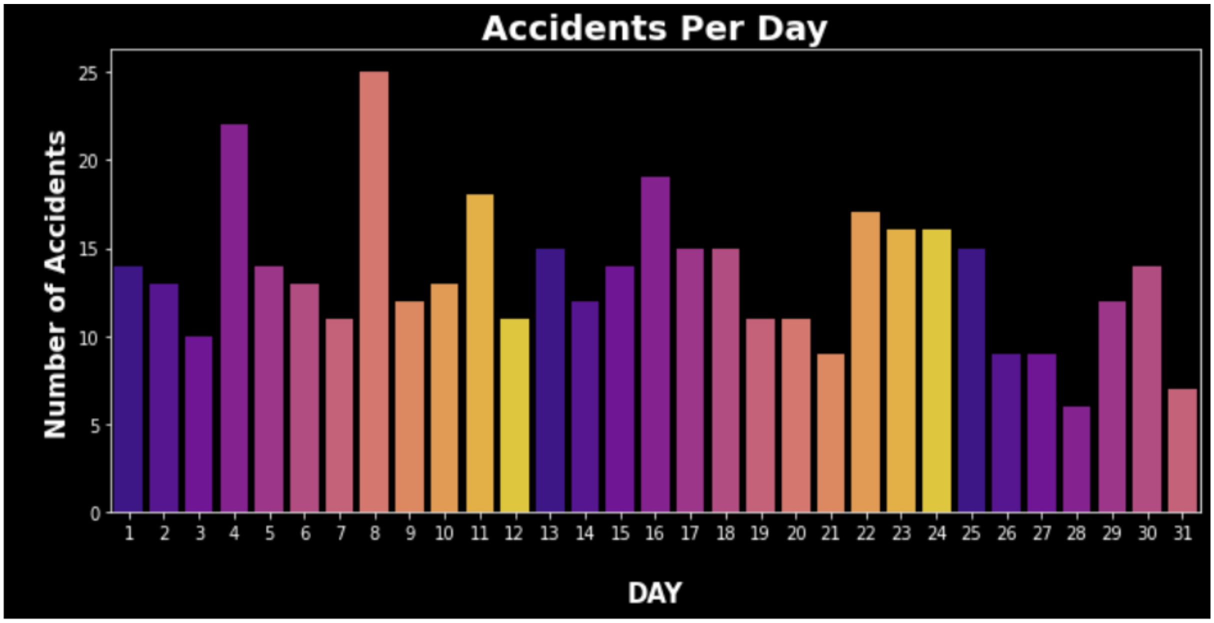


FIG 4: Accidents per Day

Observation:

- 1. The given dataset covers 2 years i.e 2016 and 2017. Maximum accidents have happened in the Year 2017.
- 2. Most of the accidents have happened in first six months of the year.
- 3. The eight day of the month has recorded the maximum no. of accidents.

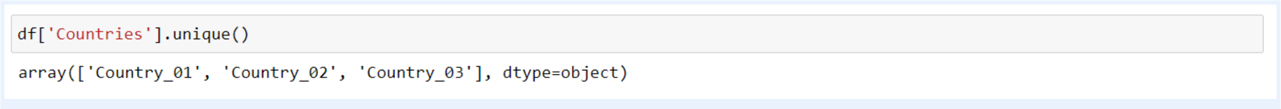
COUNTRIES:

Description:

This column describes the country in which the accident has occurred and country values have been anonymized.

Format:

This column has three unique values specifying 3 different countries.



```
df['Countries'].unique()
```

```
array(['Country_01', 'Country_02', 'Country_03'], dtype=object)
```

Visualization and Analysis:

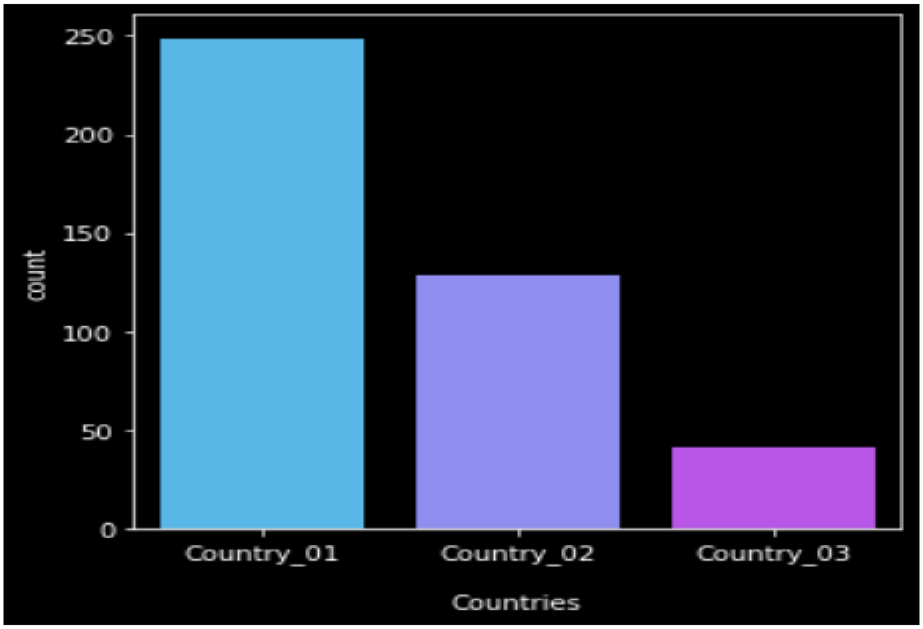


FIG 5: Count of Accidents per Country

Observations:

- There were a total of 251 accidents in country 1 which accounts for 59.06 % of total accidents in all three countries.
- There were a total of 130 accidents in country 2 which accounts for 30.59 % of total accidents in all three countries.
- There were a total of 44 accidents in country 3 which accounts for 10.35 % of total accidents in all three countries.

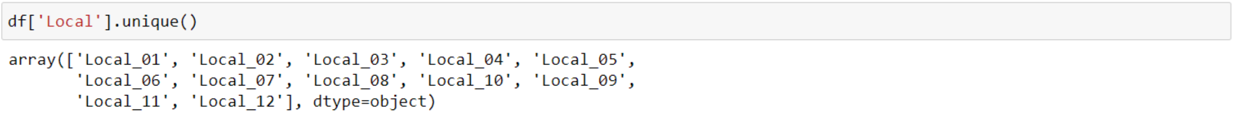
LOCAL

Description:

The Local column contains information about the city where the manufacturing plant is located. This column is anonymised.

Format:

It contains 12 unique values indicating 12 different cities.



```
df['Local'].unique()
```

```
array(['Local_01', 'Local_02', 'Local_03', 'Local_04', 'Local_05',
       'Local_06', 'Local_07', 'Local_08', 'Local_10', 'Local_09',
       'Local_11', 'Local_12'], dtype=object)
```

Visualization and Analysis:

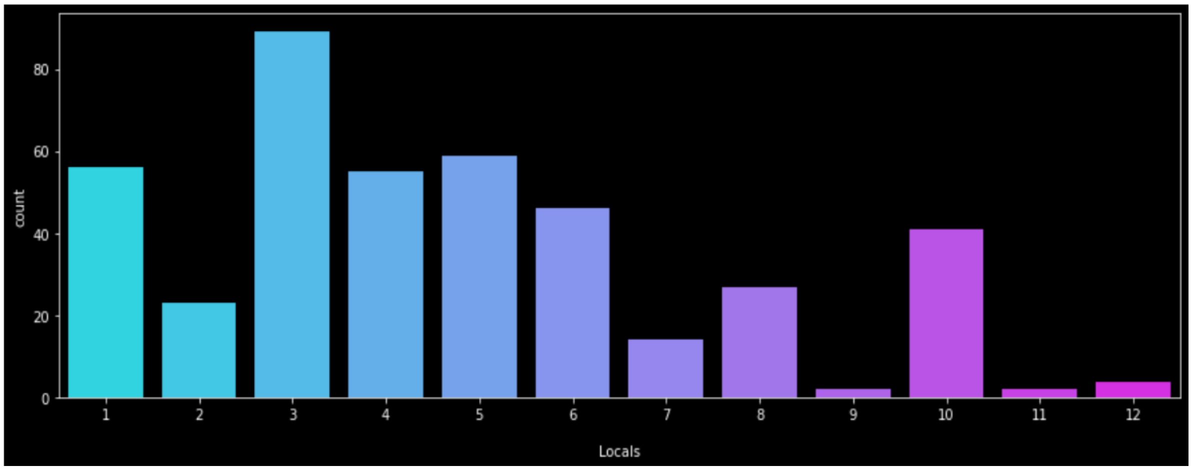


FIG 6: Count of Accidents per City

Observations:

- Maximum accidents have been recorded for Local_03 which accounts for 21.29% of total accidents, followed by Local_05 and Local_01 accounting for 14.11% and 13.40% of the total accidents**.**

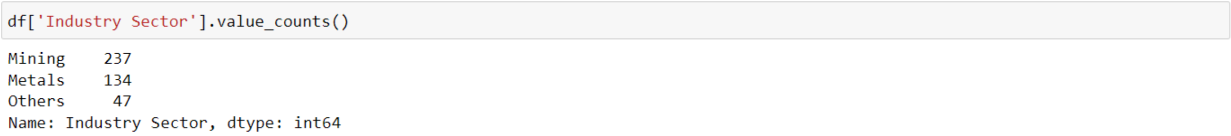
INDUSTRY SECTOR:

Description:

This column provides information about the industry sector to which the plant belongs.

Format:

This column contains three unique values i.e it focuses on 3 unique industry sectors.



```
df['Industry Sector'].value_counts()
```

```
Mining    237
Metals    134
Others     47
Name: Industry Sector, dtype: int64
```

Visualization and Analysis:

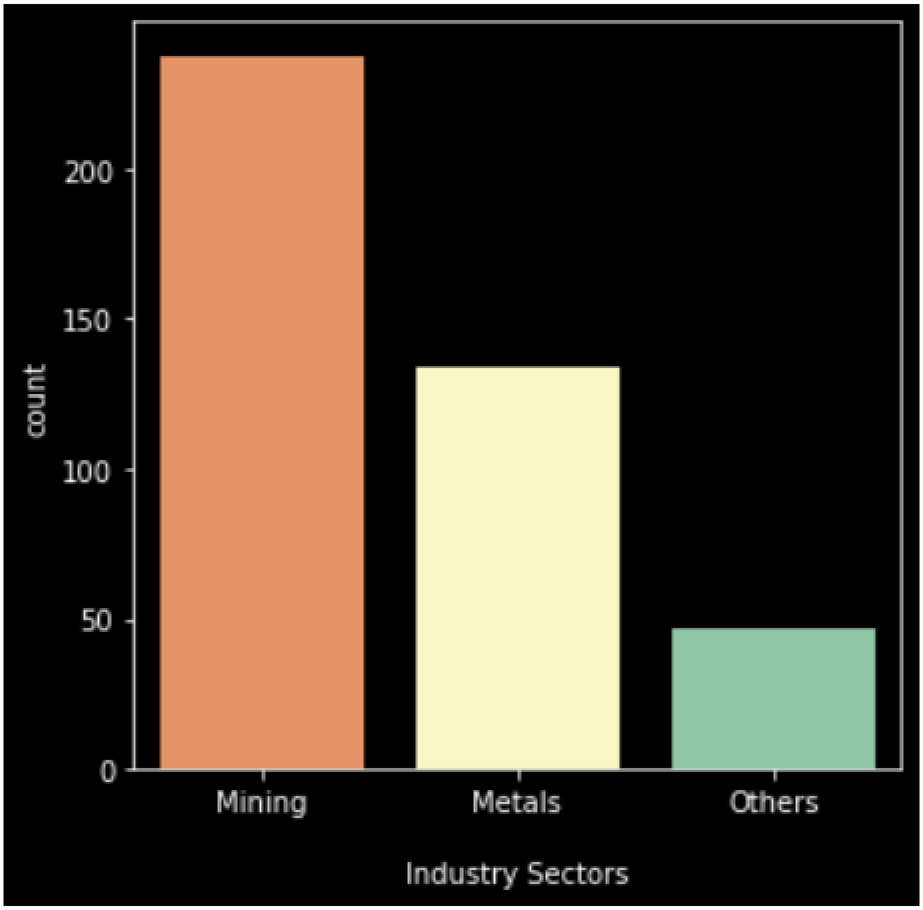


Fig 7: Count of Accidents per Industry Sector

Observations:

- There are 3 industry sectors covered in the dataset - Mining,Metals,others.
- Mining sector covers the majority of the dataset, followed by metals and others.
- 56.71% of accidents occur in Mining sector.
- 32.06% of accidents occured in Metals sector.
- 11.24% of accidents occurred in other sectors.

ACCIDENT LEVEL

Description:

Accident level registers the severity of the accident.

Format:

The values of the accident level ranges from the scale of I to VI. The roman numerals have been changed to arabic numeral format.

Visualization:

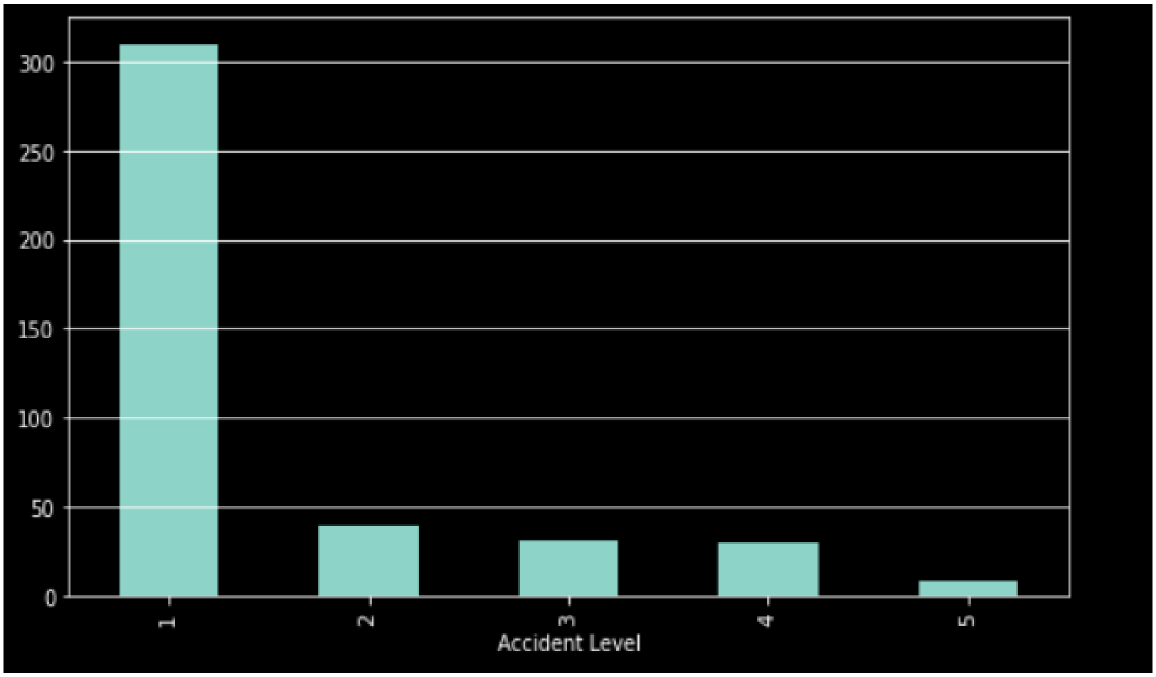


FIG 8: Count of Accident Levels

Observations:

- Accident Level 1 has occurred most number of times as compared to other levels of accidents.

Potential Accident Levels

Description:

Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors involved in the accident

Format:

The values of the potential accident level ranges from the scale of I to VI. The roman numerals have been changed to arabic numeral format.

Visualization and Analysis:

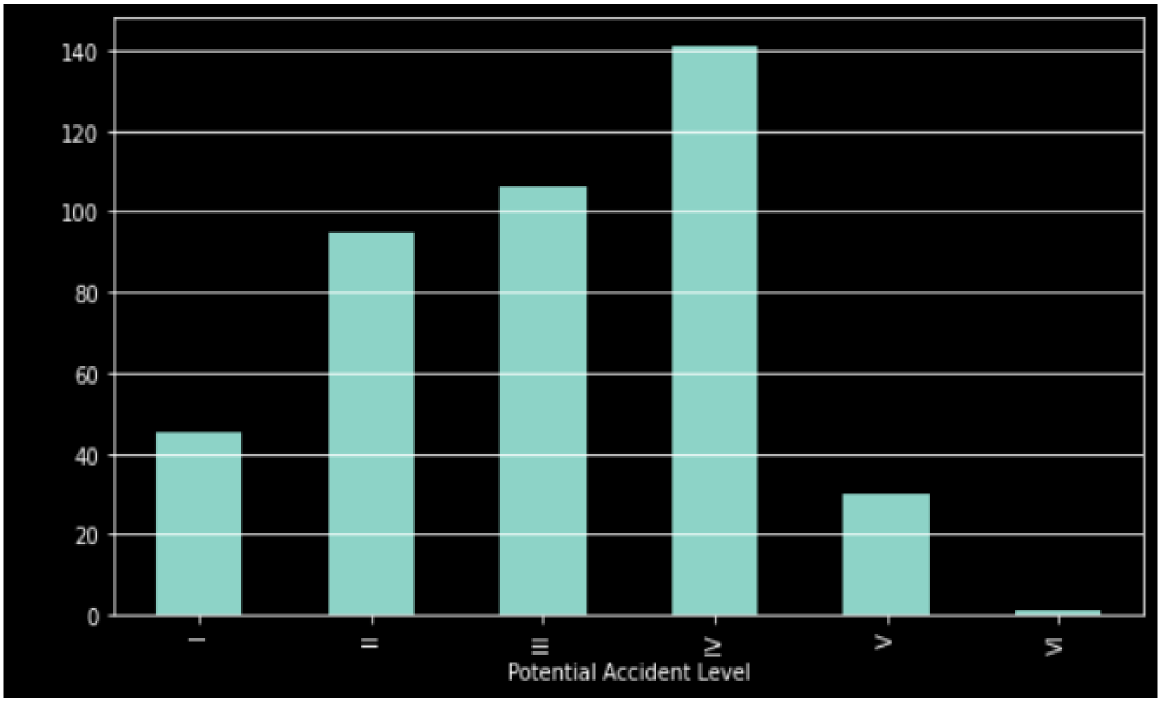


Fig 9: Count of Potential Accident Level

Observations:

- The potential accident level, i.e the severity of accident that could have happened is maximum for accident level 4.

Gender

Description:

The gender of the victim who suffered from the accident.

Format:

It consists of two values i.e Male or Female.

Visualization and Analysis:

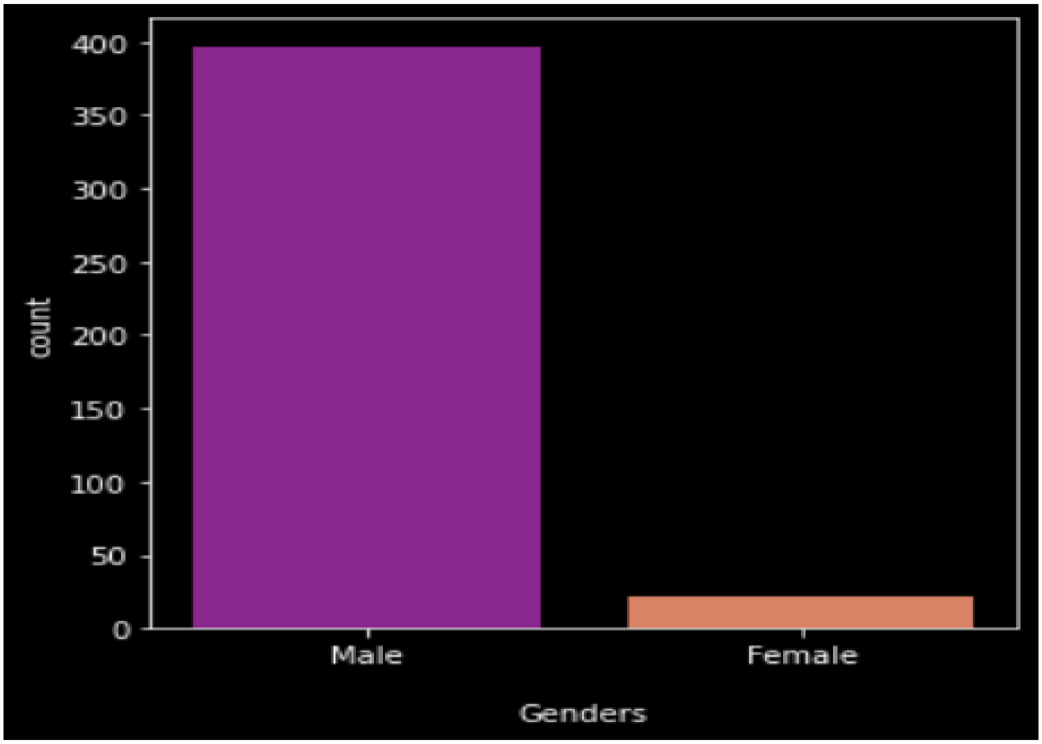


Fig 10: Count of Gender of Victim

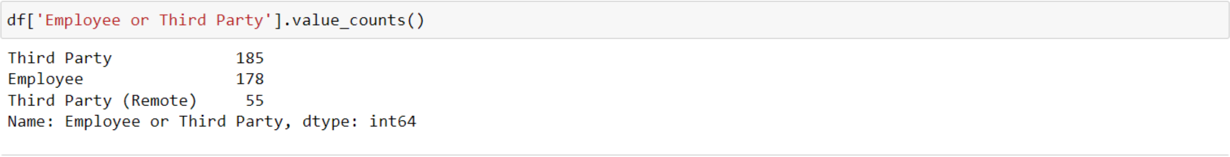
EMPLOYEE OR THIRD PARTY

Description:

If the injured person is an employee or a third party

Format:

It consists of three unique values



```
df['Employee or Third Party'].value_counts()
```

```
Third Party             185
Employee                178
Third Party (Remote)     55
Name: Employee or Third Party, dtype: int64
```

Visualization and Analysis:

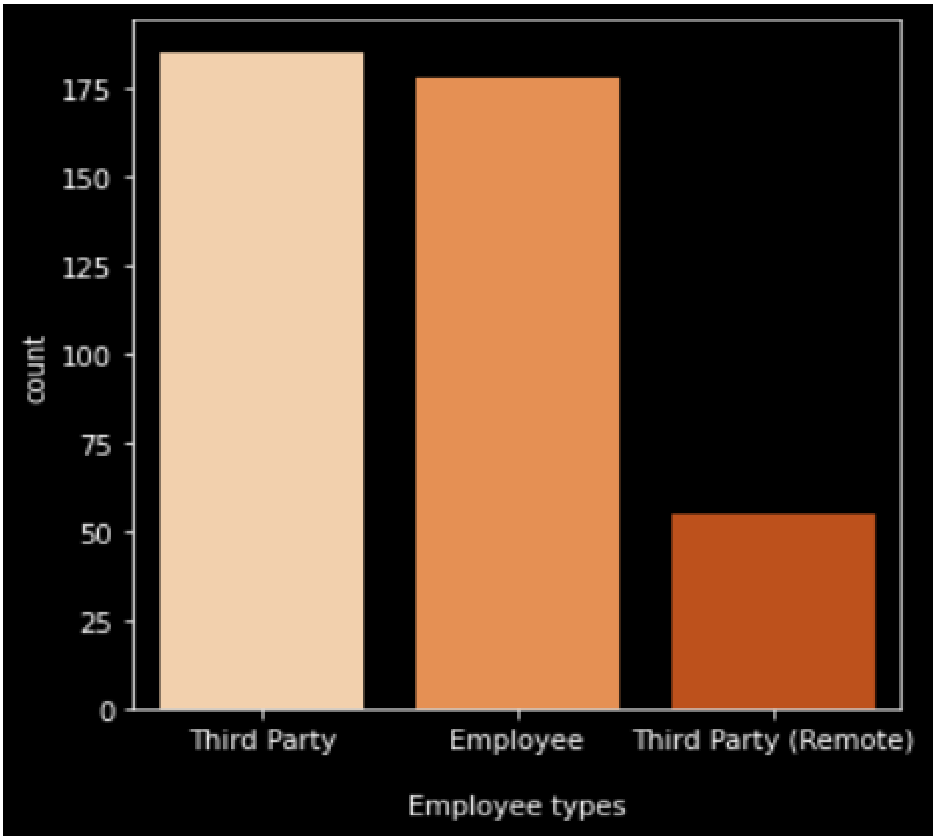


Fig 11: Employee or Third Party Analysis

Observations:

- Employee type of Third party is most prone to Accident risk which accounts for 44.26% of the total accidents.

CRITICAL RISK ANALYSIS

Description:

It provides some description of the risk involved in the accident

Format:

It contains 33 unique values.

Visualization and Analysis:

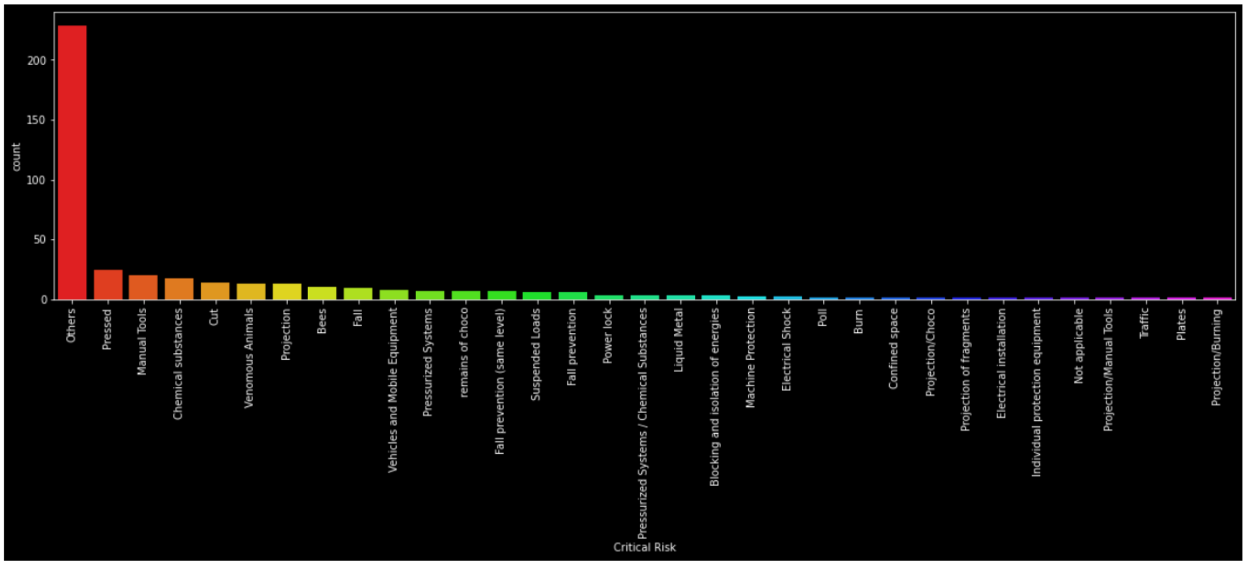


Fig 12: Critical Risk Analysis

Observations:

- There are a total of 33 critical risks covered in the dataset, out of which maximum risks are present under "others" category and post that comes "pressed" category of risk.

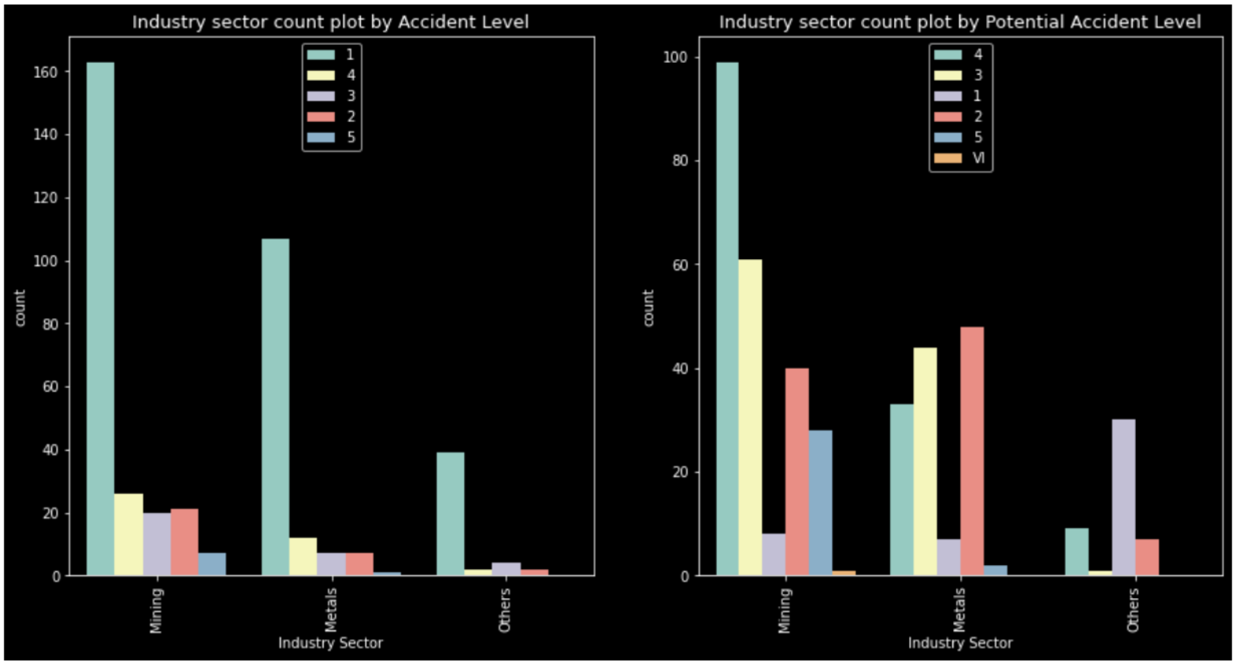


Fig 13:Industry Sector Analysis

Observations:

Industry Sector Vs Accident Level

- Accident level I is highest in all industry sectors (Mining, Metals and Other).
- Most accidents happened in the Mining industry sector.
- Other industry sectors have less accidents in comparison to other industries.
- There are very few cases for Accident level 5.Others sector didn't reported any case for this level

Industry Sector Vs Potential Accident Level

- Potential Accident level 4 is highest in all industry sectors (Mining, Metals and Others).
- Most accidents happened in the Mining industry sector.
- Other industry sectors have less accidents in comparison to other industries.
- These are very few cases for Accident level 6.Metals and Others sector didn;t reported case for this accident level

EDA Summary:

Local

- Highest manufacturing plants are located in Local_03 city and lowest in Local_09 city.

Country

- Percentage(%) of accidents occurred in respective countries: 59% in Country_01, 31% in Country_02 and 10% in Country_03.

Industry Sector

- Percentage(%) of manufacturing plants belong to respective sectors: 57% to Mining sector, 32% to Metals sector and 11% to Others sector.

Country + Industry Sector

- Metals and Mining industry sector plants are not available in Country_03.
- Distribution of industry sectors differ significantly in each country.

Accident Levels

- The number of accidents decreases as the Accident Level increases and increases as the Potential Accident Level increases.

Gender

- There are more men working in this industry as compared to women.

Employee type

- 44% Third party employees, 43% own employees and 13% Third party(Remote) employees working in this industry.

Gender + Employee type

- Proportion of third party employees in each gender is equal, third party(remote) employees in each gender are not equal and own employees in each gender are not equal.

Gender + Industry Sector

- Proportion of Metals, Mining and Others sector employees in each gender is not equal

Gender + Accident Levels

- Males have a higher accident level than females.
- There are many low risks at general accident level, but many high risks at potential accident level.

Accident Levels + Employee type

- For both accident levels, the incidence of Employee is higher at low accident levels, but the incidence of Third parties seems to be slightly higher at high accident levels.

Accident Levels + Calendar

- Accidents are recorded from 1st Jan 2016 to 9th July 2017 every month, there are a high number of accidents in 2016 and less in 2017.
- Number of accidents is high in the beginning of the year and it keeps decreasing later.
- Number of accidents are very high in particular days like 4, 8 and 16 in every month.
- Number of accidents increased during the middle of the week and declined since the middle of the week.
- Both of the two accident levels have the tendency that non-severe levels decreased throughout the year, but severe levels did not change much, and some of these levels increased slightly in the second half of the year.
- Both of the two accident levels are thought that non-severe levels decreased in the first and the last of the week, but severe levels did not change much.

Critical Risk

- Most of the critical risks are classified as Others.

NLP PRE-PROCESSING:

Data pre-processing involves the following steps:

- Punctuation Removal
- Lowercasing
- Lemmatization
- Alphanumeric value check

WORD- CLOUD

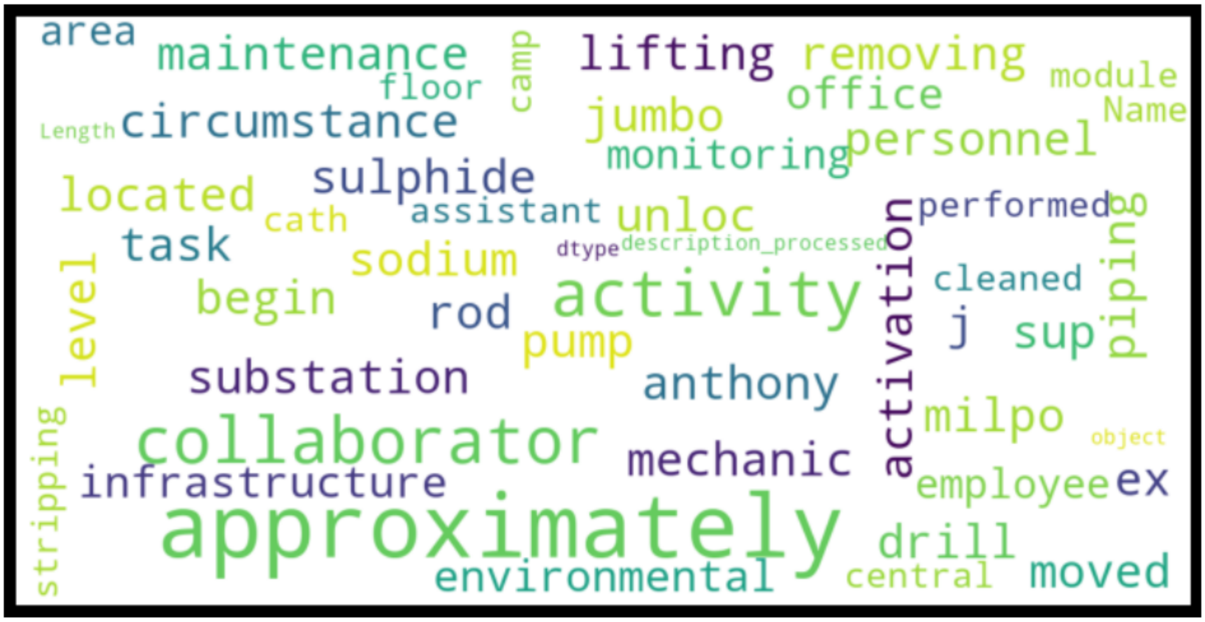


FIG 14: Word Cloud Analysis for common words

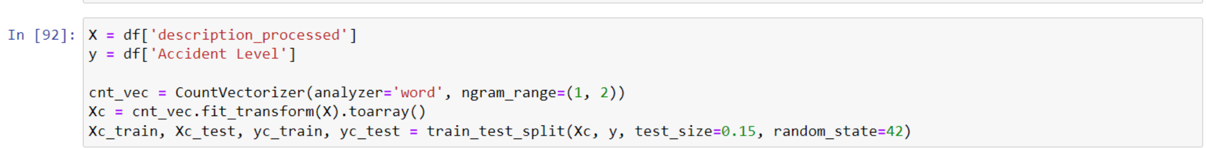
MODEL DEVELOPMENT AND EVALUATION

For predicting the accident level and the potential accident level based on the input description, various models were trained, evaluated and compared. Amongst them, the best performing model was chosen for further use.

MACHINE LEARNING MODELS

TEXT VECTORIZATION

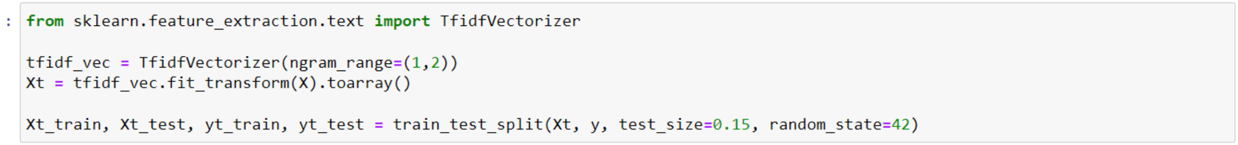
For this particular problem, two text vectorization techniques have been used- Count Vectorization and TF-IDF Vectorization.



```
In [92]: X = df['description_processed']
         y = df['Accident Level']

         cnt_vec = CountVectorizer(analyzer='word', ngram_range=(1, 2))
         Xc = cnt_vec.fit_transform(X).toarray()
         Xc_train, Xc_test, yc_train, yc_test = train_test_split(Xc, y, test_size=0.15, random_state=42)
```

Fig 15: Count Vectorization



```
: from sklearn.feature_extraction.text import TfidfVectorizer

  tfidf_vec = TfidfVectorizer(ngram_range=(1,2))
  Xt = tfidf_vec.fit_transform(X).toarray()

  Xt_train, Xt_test, yt_train, yt_test = train_test_split(Xt, y, test_size=0.15, random_state=42)
```

Fig 16 : TF-IDF Vectorization

MODEL TRAINING

The objective is to train and pickle two machine learning models, one to predict the Accident Level and the other to predict the Potential Accident Level based on the descrption given.

Following models have been trained and compared for accuracies attained:

COUNT VECTORIZER

Models/ Prediction	SVC	Random Forest	Gradient Boosting	XGBoost	Neural Network
Accident Level	76.19%	76.19%	73.02%	73.02%	76.19%
Potential Accident Level	44.44%	34.92%	38.10%	-	11.1%

TF-IDF VECTORIZER

Models/ Prediction	SVC	Random Forest	Gradient Boosting	XGBoost	Neural Network
Accident Level	76.19%	76.19%	76.19%	71.43%	76.19%
Potential Accident Level	44.44%	47.62%	41.27%	44.44%	11.11%

WORD2VEC

Models/ Prediction	LSTM
Accident Level	75%
Potential Accident Level	32%

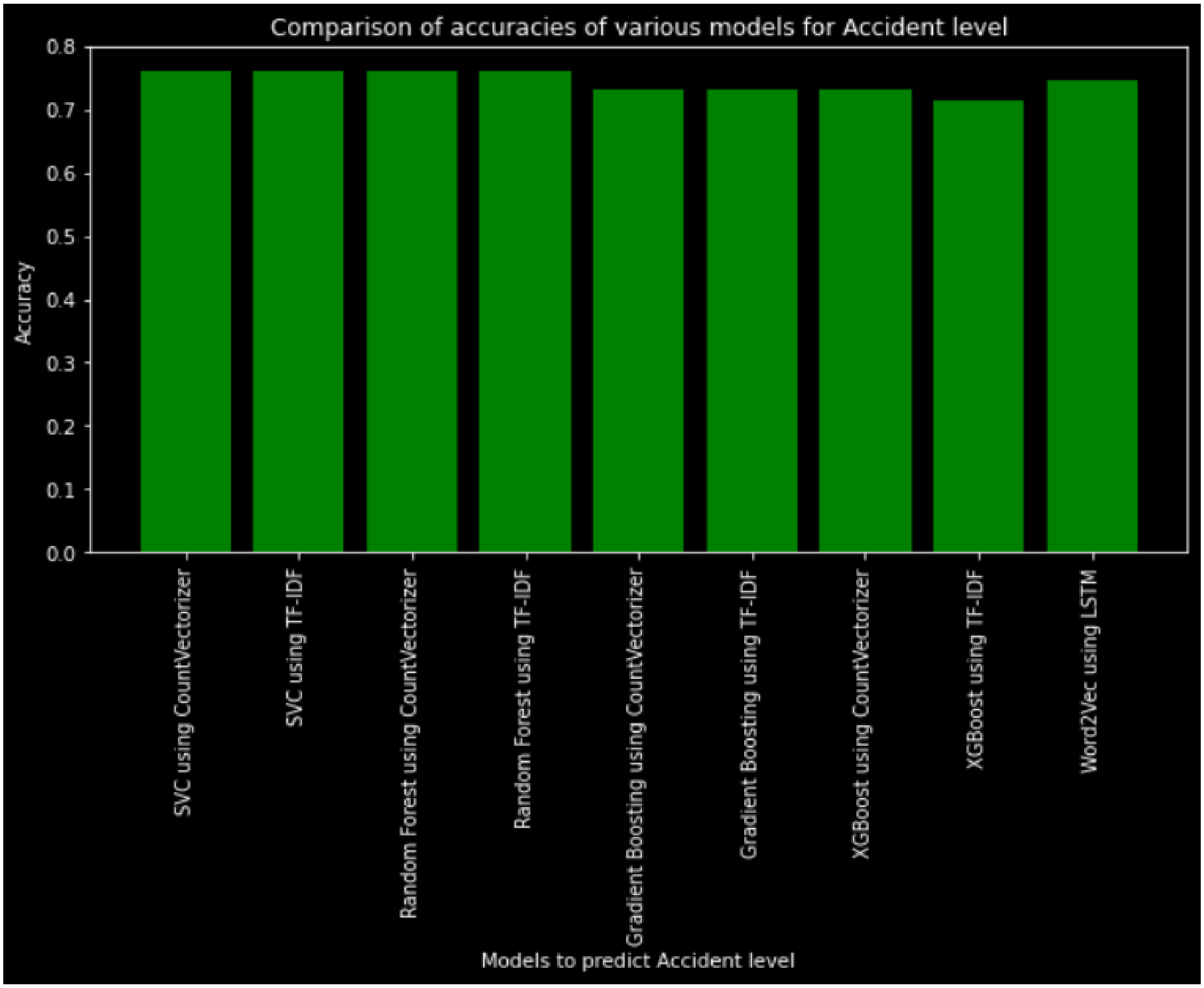


Fig 17: Comparison of Accuracies of various models for Accident Level

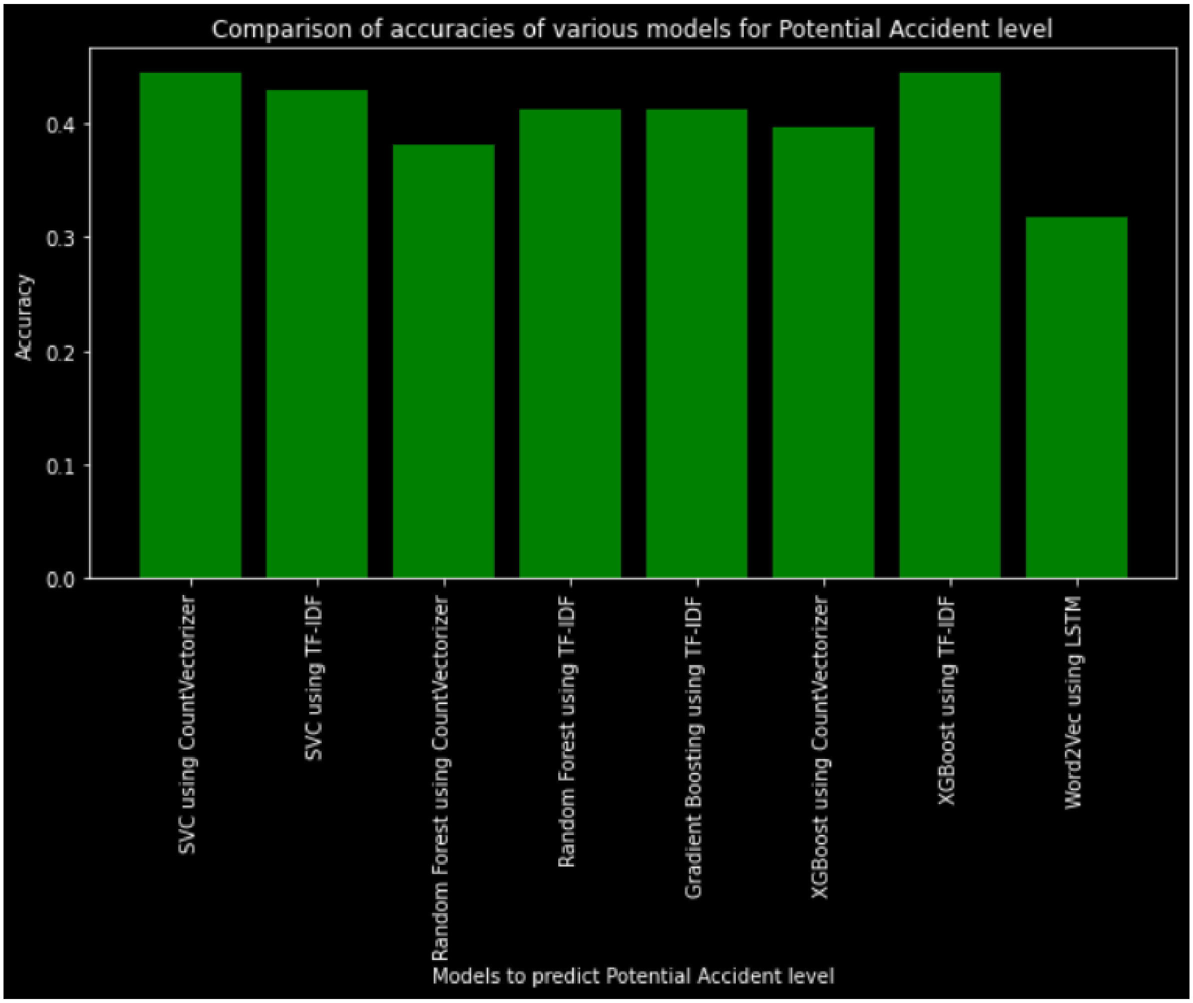


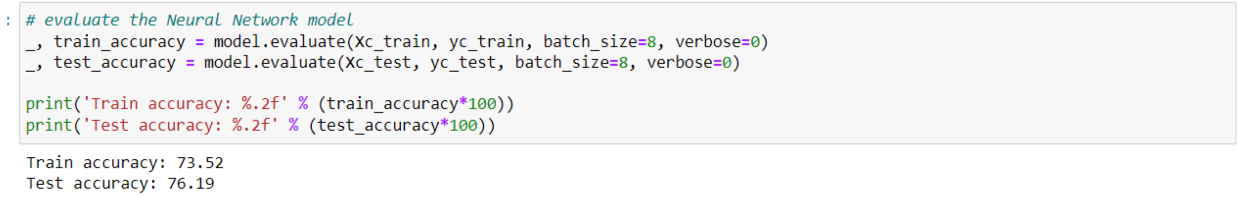
Fig 18: Comparison of Accuracies of various models for Potential Accident Level

In comparison to all the above models for target label Accident Level and Potential accident level, we can say that SVC using CountVectorizer is having better accuracies than others.

We can choose SVC using CountVectorizer, since it's common for both the targe labels.

NEURAL NETWORK MODEL

Since the accuracy of the Machine learning model was not satisfactory, hence neural network model was tried on the dataset to check the performance of the model. As expected, the Neural Network model performed better as compared to the ML models.



```
# evaluate the Neural Network model
_, train_accuracy = model.evaluate(Xc_train, yc_train, batch_size=8, verbose=0)
_, test_accuracy = model.evaluate(Xc_test, yc_test, batch_size=8, verbose=0)

print('Train accuracy: %.2f' % (train_accuracy*100))
print('Test accuracy: %.2f' % (test_accuracy*100))
```

```
Train accuracy: 73.52
Test accuracy: 76.19
```

The train accuracy of the model was 73.52 % and the test accuracy is 76.19.

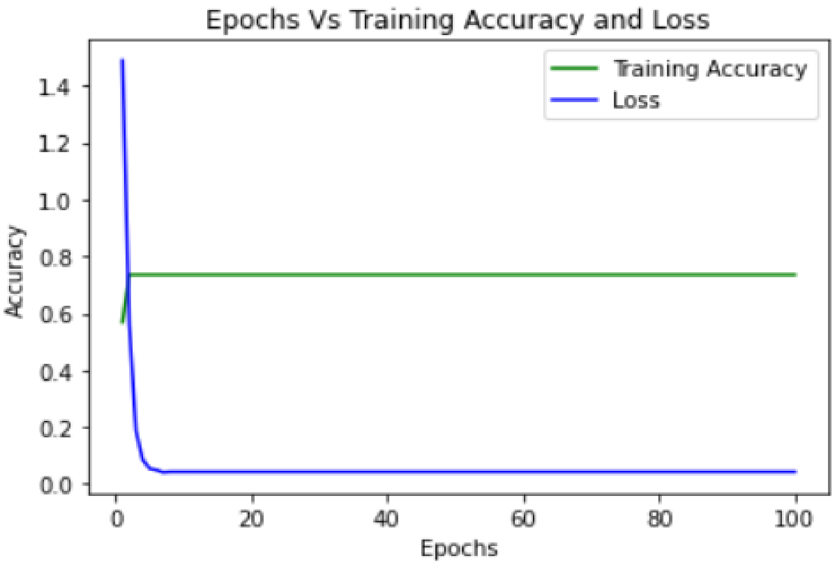
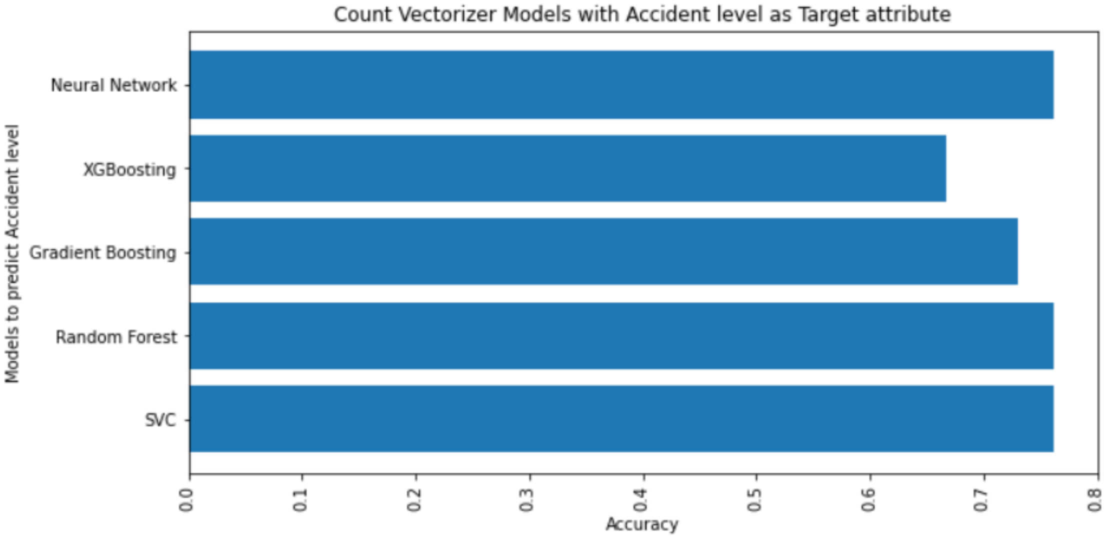
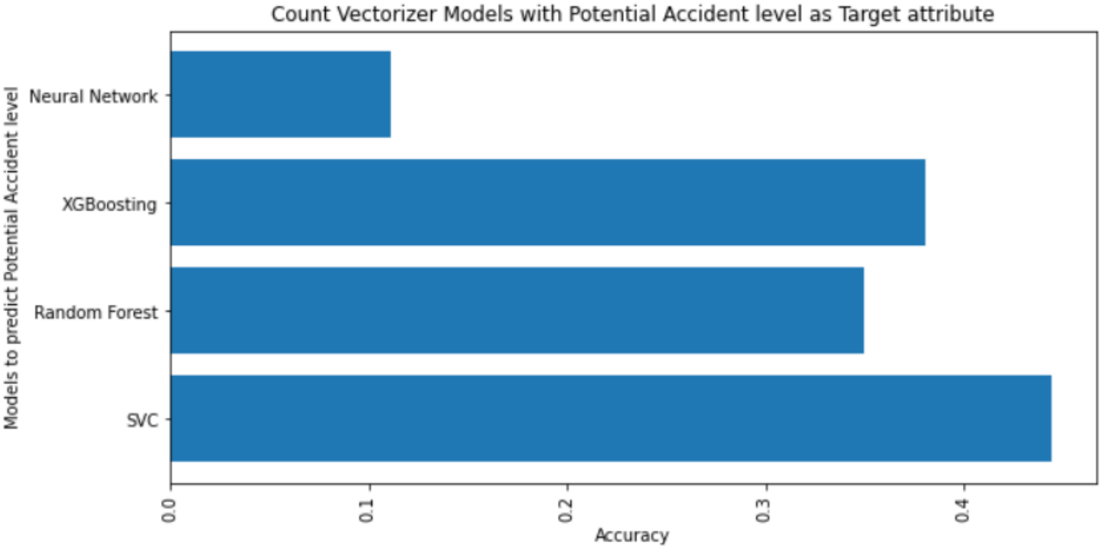
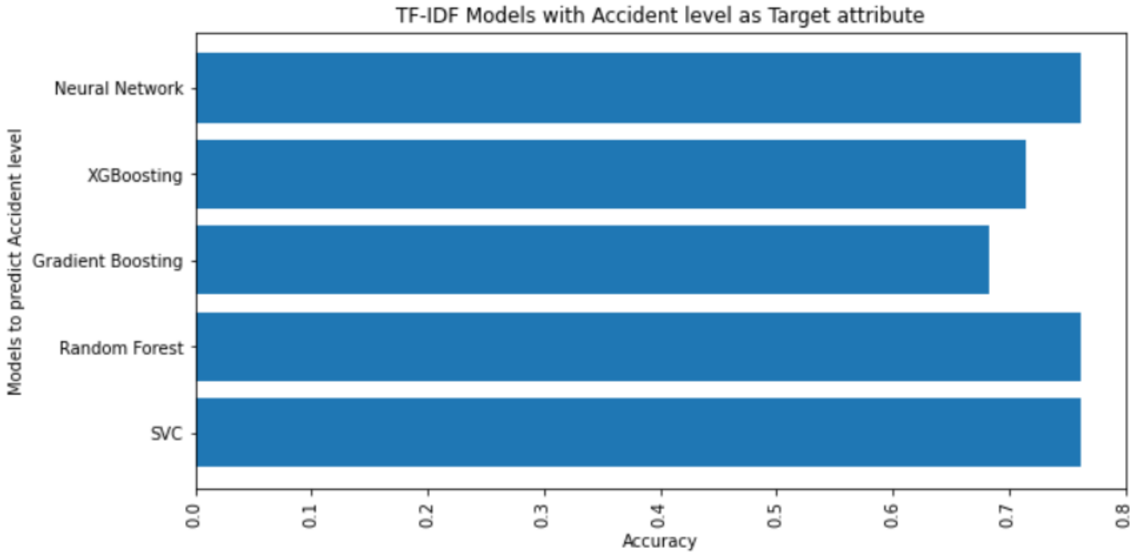


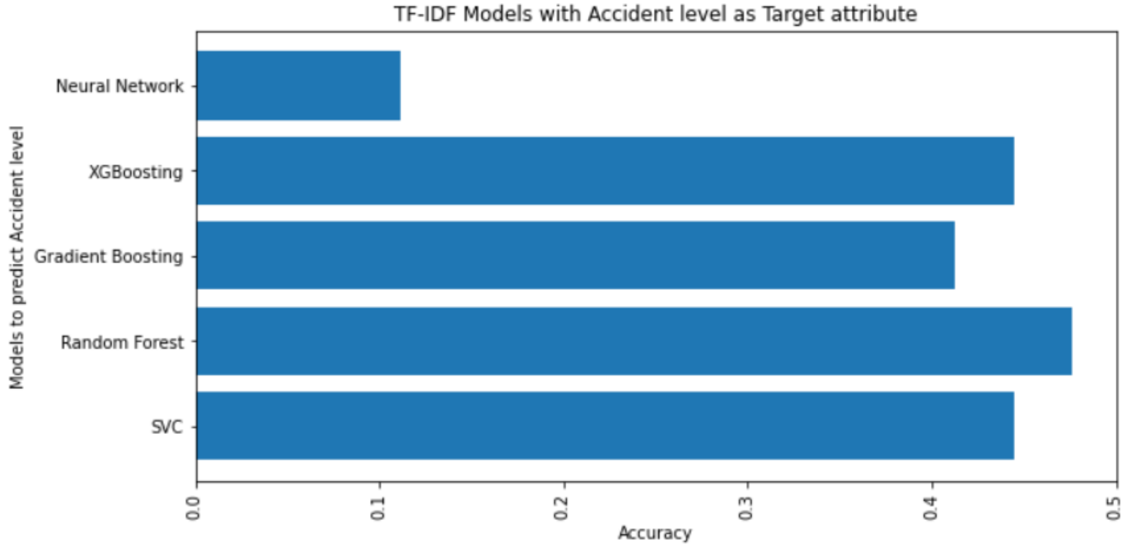
Fig 19: Epochs V/S Accuracy and Loss

MODEL COMPARISON GRAPHS



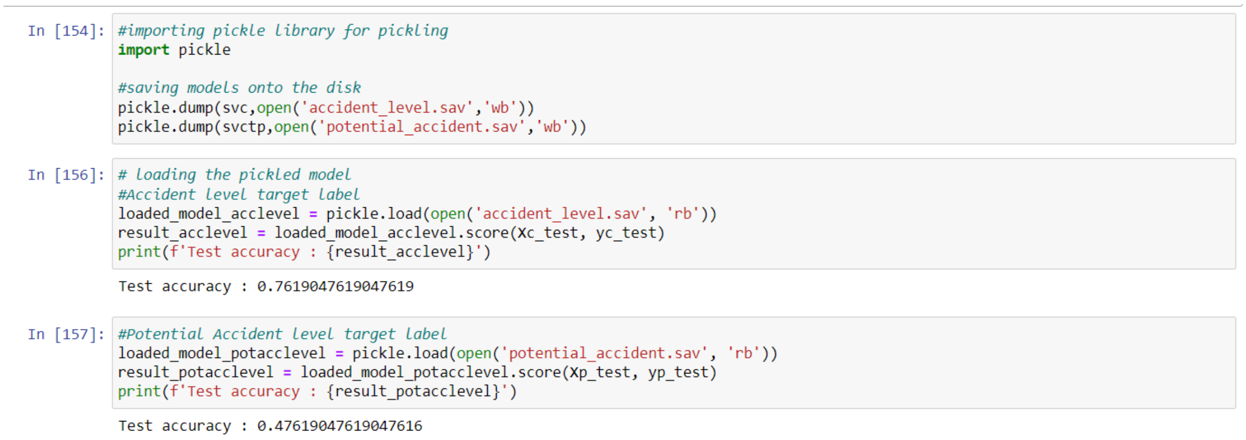






Pickling

Pickling is done to save the model onto the disk. We serialize the model providing the best accuracy and then load the saved model into memory whenever necessary.Comparing all the above models for Accident Level and Potential Accident Level attributes, we can see that Neural Network Model with Count Vectorizer is the most accurate. We can choose SVC using CountVectorizer, since it's common for both the targe labels. Pickling and saving this model for further use



```
In [154]: #importing pickle library for pickling
          import pickle

          #saving models onto the disk
          pickle.dump(svc,open('accident_level.sav','wb'))
          pickle.dump(svctp,open('potential_accident.sav','wb'))

In [156]: # loading the pickled model
          #Accident level target label
          loaded_model_acclevel = pickle.load(open('accident_level.sav', 'rb'))
          result_acclevel = loaded_model_acclevel.score(Xc_test, yc_test)
          print(f'Test accuracy : {result_acclevel}')

          Test accuracy : 0.7619047619047619

In [157]: #Potential Accident level target label
          loaded_model_potacclevel = pickle.load(open('potential_accident.sav', 'rb'))
          result_potacclevel = loaded_model_potacclevel.score(Xp_test, yp_test)
          print(f'Test accuracy : {result_potacclevel}')

          Test accuracy : 0.47619047619047616
```

Fig: Model Pickling

CHATBOT

The aim is to create a chatbot that can provide industrial accidents related information to human beings and help to spread awareness and warn about scenarios that can lead to severe accidents in future.

CHATBOT GUI

The chatbot graphical user interface was created using the tkinter library in python. The design was kept simple for user-friendly interaction.

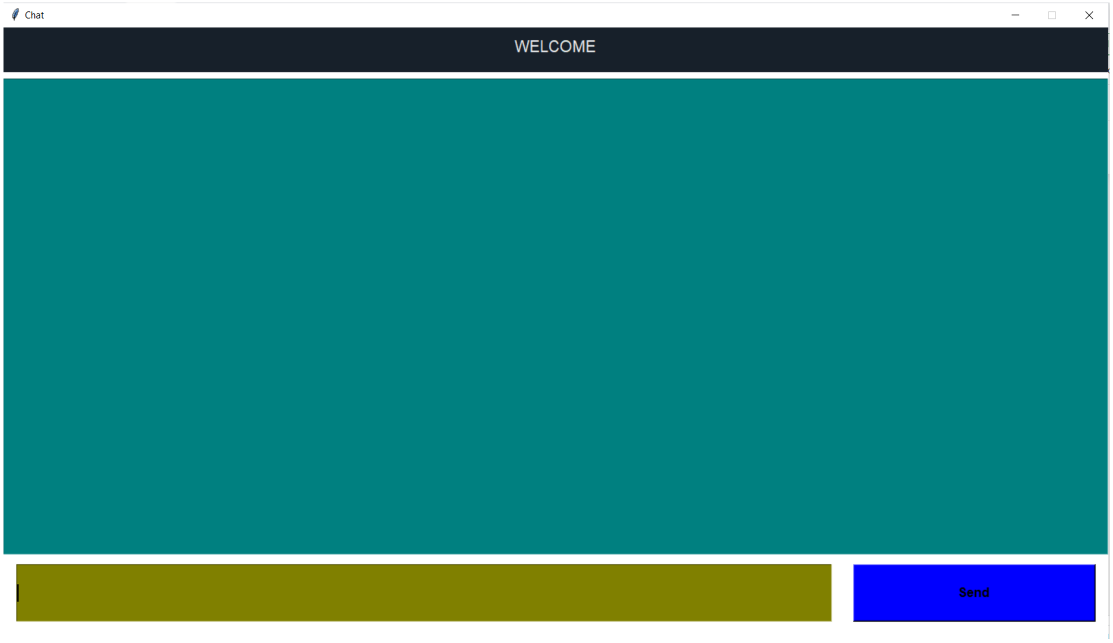
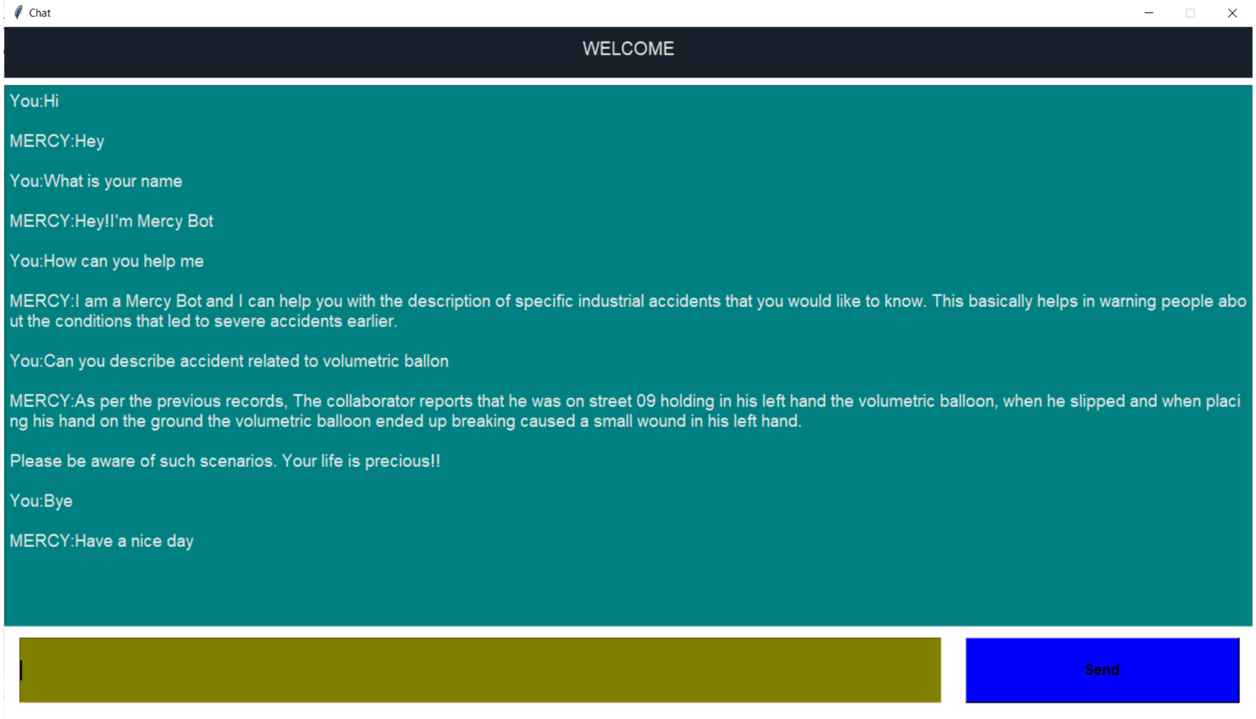


Fig 20: CHATBOT GUI

CHATBOT CAPABILITIES

The chatbot is capable of answering the below type of questions:



You:Hi

MERCY:Hey

You:What is your name

MERCY:Hey!I'm Mercy Bot

You:How can you help me

MERCY:I am a Mercy Bot and I can help you with the description of specific industrial accidents that you would like to know. This basically helps in warning people about the conditions that led to severe accidents earlier.

You:Can you describe accident related to volumetric ballon

MERCY:As per the previous records, The collaborator reports that he was on street 09 holding in his left hand the volumetric balloon, when he slipped and when placing his hand on the ground the volumetric balloon ended up breaking caused a small wound in his left hand.

Please be aware of such scenarios. Your life is precious!!

You:Bye

MERCY:Have a nice day

Fig 21: Demo Chat with Mercy Bot

Future Scope of Study:

1. Collect more data Currently there are 425 records. This count is not sufficient to train the models. Having more data will allow better accuracy of results
2. Collect additional information about the accidents Additional information about the equipment that led to the accident, such as equipment name, voltage passed to the equipment, the temperature of the equipment at the time of the accident, will greatly help in predicting a wider range of incidents and the accuracy of the model will also be higher
3. Use Named Entity Recognition Having NER will allow us to identify equipments that led to an accident or the persons involved in the accident
4. Sentiment analysis Sentiment analysis can be used to detect the severity of the accident description and assign relevant accident level or potential accident level

Summary

Industrial accidents cause huge damage to human lives as well as the environment. It is important to take timely precautions in order to avoid such severe accidents in future. The Natural Language processing technique can help in analyzing previous accidents and the reason of their occurance. The current project aims to use NLP technology for developing a chatbot that can understand the given set of accident descriptions and warn humans about the scenarios that could lead to severe accidents. The chatbot is capable of understanding input text messages provided by the user in the form of various intents and it replies accordingly.

The below flowchart depicts the entire process flow of the project.

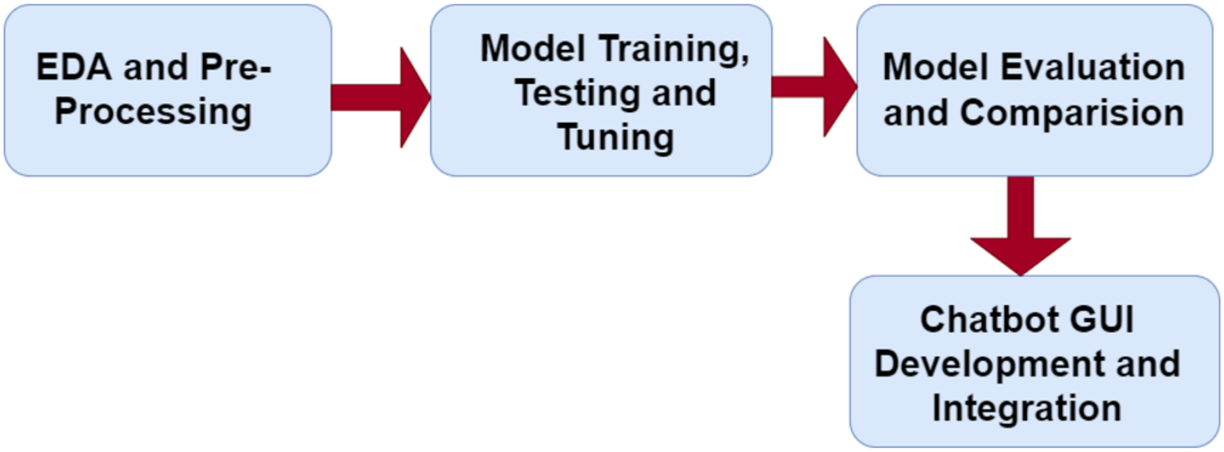

EDA and Pre-Processing
Model Training, Testing and Tuning
Model Evaluation and Comparision
Chatbot GUI Development and Integration

Fig 22: NLP CHATBOT PROCESS FLOW

References

1. https://www.ilo.org/moscow/areas-of-work/occupational-safety-and-health/WCMS_249278/lang--en/index.htm
2. https://www.ijstr.org/final-print/jun2020/Industrial-Accident-Report-Analysis-Using-Natural-Language-Processing.pdf
3. https://www.kaggle.com/ihmstefanini/industrial-safety-and-health-analytics-database
4. https://www.ibm.com/cloud/learn/natural-language-processing
5. https://www.malvicalewis.com/post/capstone-project-report-on-text-classification
6. https://www.ijstr.org/final-print/jun2020/Industrial-Accident-Report-Analysis-Using-Natural-Language-Processing.pdf
7. https://www.kaggle.com/daguirreag/analytics-industrial-safety-and-health
8. https://tel.archives-ouvertes.fr/tel-01230079/document
9. https://demos.co.uk/wp-content/uploads/2019/10/Jisc-OCT-2019-2.pdf
10. https://towardsdatascience.com/introduction-to-data-visualization-in-python-89a54c97fbed
11. https://www.boldbi.com/blog/data-visualization-importance-and-benefits

12. https://www.ibm.com/cloud/learn/natural-language-processing

PROJECT CONTRIBUTION CHART

TASKS	PERFORMED BY
Loading Dataset	Nishtha, Barnali
Understanding Dataset	Nishtha, Barnali
Data Cleaning	Nishtha, Barnali
EDA	Nishtha, Barnali
Preprocessing	Nishtha, Barnali
EDA Rework	Nishtha, Barnali, Suraj
ML MODELS	Nishtha, Barnali
LSTM	Nishtha, Barnali
Neural Networks	Suraj
Future Scope	Suraj

ChatBot Intents	Nishtha,Barnali, Suraj
ChatBot GUI	Nishtha, Barnali
ChatBot Integration with Dataset Description	Nishtha,Barnali, Suraj