Industrial safety NLP based Chatbot

Interim report group 3

|  |
| --- |
| Team mentor |
| Gaurav |

|  |
| --- |
| Team members |
| Hima Bindu |
| Chandini |
| Akshay Kumar Bhamare |
| Kathambari |
| Vishnu |
| Soham Parik |

**Table of Contents**

1. Overview
   * Project Detail
2. Column’s description
3. Import libraries
4. Load the dataset
5. Pre-processing
   * NLP Pre-processing
6. [EDA](https://www.kaggle.com/mysh7989/industrial-accident-causal-analysis#5.-EDA)
   * Univariate Analysis
   * Multivariate Analysis
   * NLP Analysis
7. Modeling
   * Feature Engineering
   * Case1 : Accident Level
   * Case2 : Potential Accident Level
8. Conclusion
   * Task Submission
9. References

# 

# Summary of the problem statement, Data and findings

## **Problem Statement**

## **Project: Industrial safety. NLP based Chatbot**

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 environment

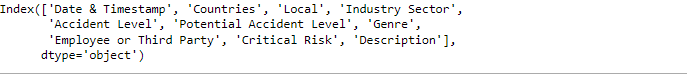
## **Abstract**

Applying traditional machine learning and neural network-based NLP to automatically classify records of accidents from 12 different plants in 03 different countries which every line in the data is an occurrence of an accident.

#### **Project Details:**

This The database is basically records of accidents from 12 different plants in 03 different countries which every line in the data is an occurrence of an accident.

#### **Import Dataset and data columns**

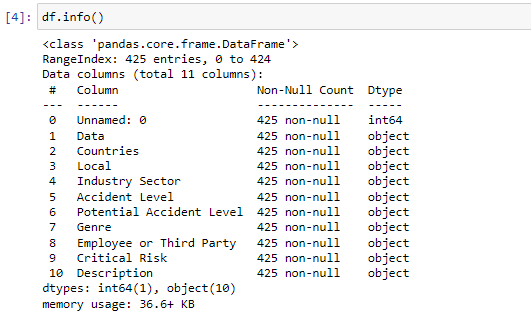


**Columns description:**

* 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.
* The dataset is in csv format. Basic exploration of the data is as below

#### **Check for Shape, null values, Datatype &Missing Values**





* Non for the column have any null data.
* There are no missing values
* Total 10 columns and 425 rows of data.
* All the column datatype os object type.
* From the above output, we see that except first column all other columns data type is object

**Categorical columns**

'Countries', 'Local', 'Industry Sector', 'Accident Level', 'Potential Accident Level', 'Genre', 'Employee or Third Party', 'Critical Risk', 'Description'

Date column - 'Data'

**Given Data Summary**

There are about 425 rows and 11 columns in the dataset.

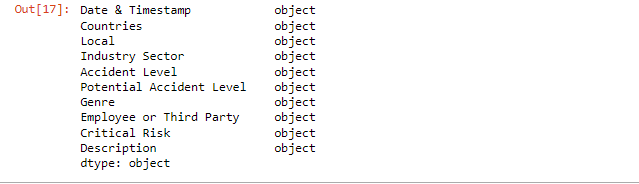
We noticed that except a 'date' column all other columns are categorical columns.

**Data Cleansing**

**Remove 'Unnamed: 0',Rename - 'Data' columns, Dropping Duplicates**

The field Unnamed: 0", is dropped and columns are renamed following

* ‘Data': Date & Timestamp



**Observation**

* We observed that there are records of accidents from 1st Jan 2016 to 9th July 2017 in every month. So there are no outliers in the 'Date' column.
* There are only three country types so there are no outliers in 'Countries' column.
* There are 12 Local cities where manufacturing plant is located and it's types are in sequence so there are no outliers in 'Local' column.
* There are only three Industry Sector types which are in sequence so there are no outliers in 'Industry Sector' column.
* There are only five Accident Level types which are in sequence so there are no outliers in 'Accident Level' column.
* There are only six Potential Accident Level types which are in sequence so there are no outliers in 'Potential Accident Level' column.
* There are only two Gender types in the provided data so there are no outliers in 'Gender' column.
* There are only three Employee types in the provided data so there are no outliers in 'Gender' column.
* There are quite a lot of Critical risk descriptions, and we don't see any outliers but with the help of SME we can decide whether this column has outliers or not.

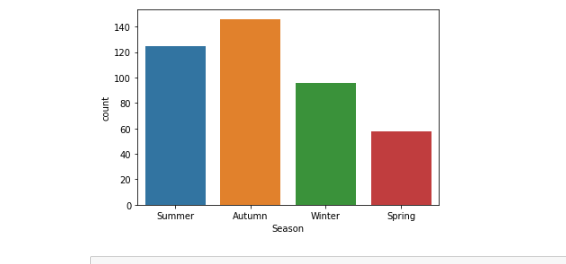
**Data Cleansing Summary**

* Removed 'Unnamed: 0' column and renamed - 'Data' columns in the dataset.
* There are no outliers in the dataset.
* No missing values in dataset.
* We are left with 425 rows and 10 columns after data cleansing.

**Data Pre-processing**

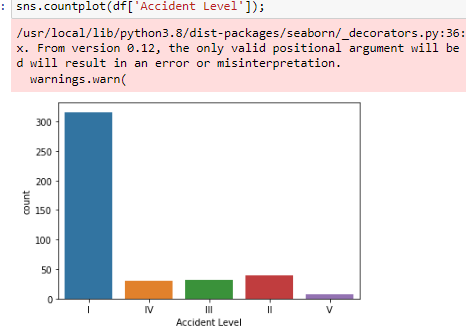
To better understand the data, I am extracting the day, month and year from Date column and creating new features such as weekday, week of year.

* As we know, this database comes from one of the biggest industry in Brazil which has four climatological seasons as below.
* https://seasonsyear.com/Brazil
* Spring: September to November
* Summer: December to February
* Autumn: March to May
* Winter: June to August
* We can create seasonal variable based on month variable.



**Univariate Analysis:**

**Checking the distribution of data based on accident levels**

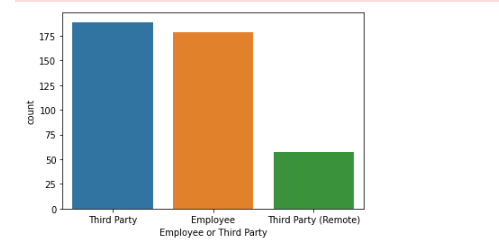


The distribution of Accident Levels is highly imbalanced in the dataset

**Distribution of the data based on country wise**

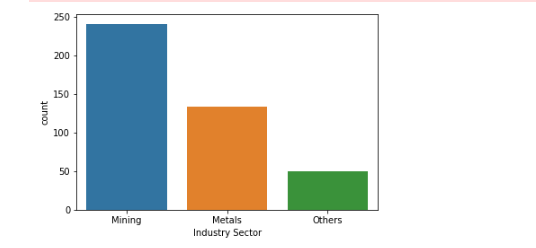


**Distribution of accidents by Employee Types**



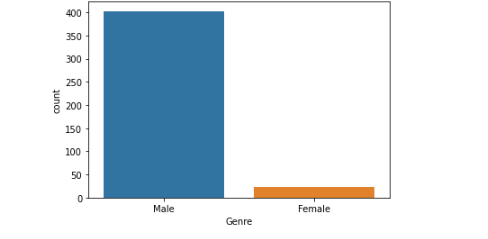
From the graph it is very clear that accidents have happened in almost equal proportions among permanent employees or third-party contractors, with third party contractors a bit on the higher side.

**Distribution of accidents as per industry sector.**



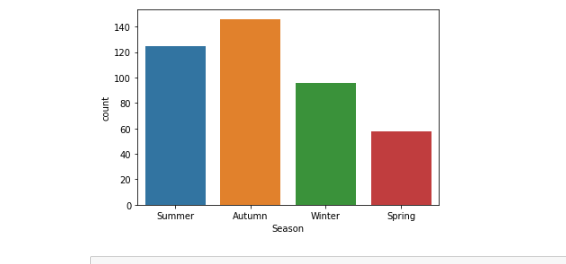
Majority of the accidents have happened in the mining sector, followed by metal industry and other type of industries.

**Distribution of accidents as per Gender**



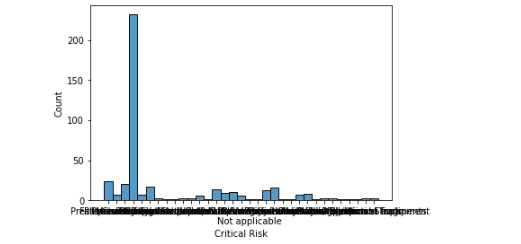
The distribution of accidents is imbalanced when checked by "Genre". The count of accidents in males is way higher than that in females.

**Distribution of Accident Asper season**



The count of accidents in Autumn is way higher than other seasons

**Distribution of Accident Asper season**



We can see from the graph that the Critical risk category "Others" have the most number of accidents. This means we are not clear about the exact risk factor associated with accidents in this dataset.

**Analyzing Description variable**

Chart, line chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

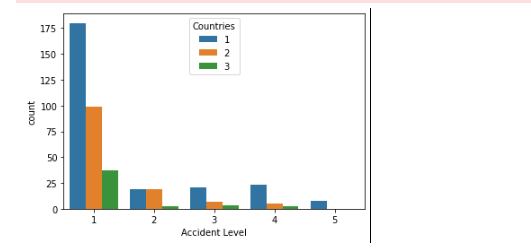
Average length of Description: 368.28

Maximum length of Description: 1029

Minimum length of Description: 94

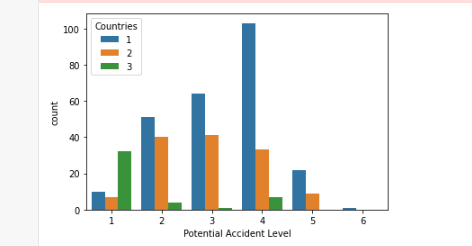
**Bivariate Analysis**

**Check the proportion of Accident in different countries**



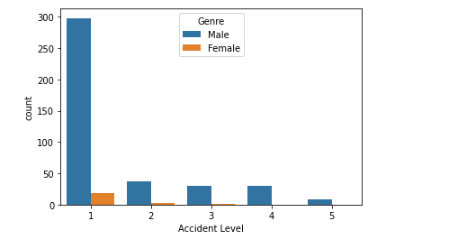
* Majority of the accident Level I accidents has occurred only in Country I.
* Maximum number of accidents in all countries are mainly of type Accident Level I.

**Check the Potential Accident level in different countries**



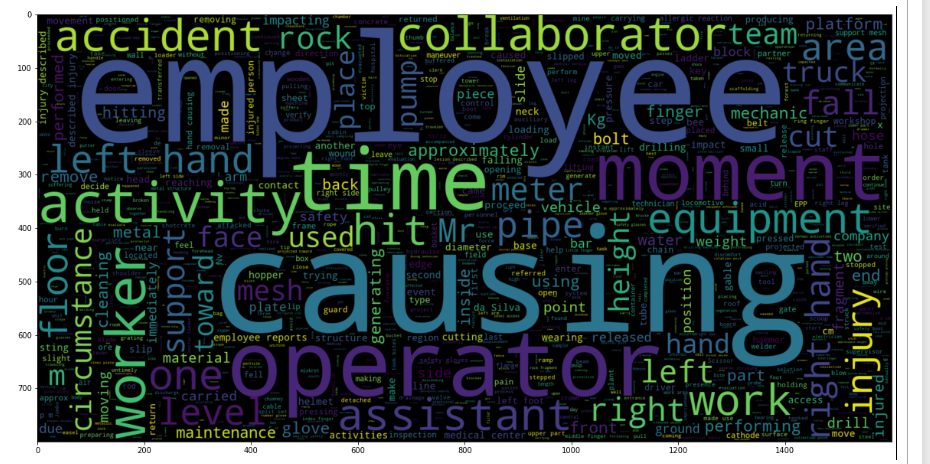
* Majority of the accidents has occurred only in Country I.

**Check the Accident level with Gender**



Majority of the accidents happened for Males.

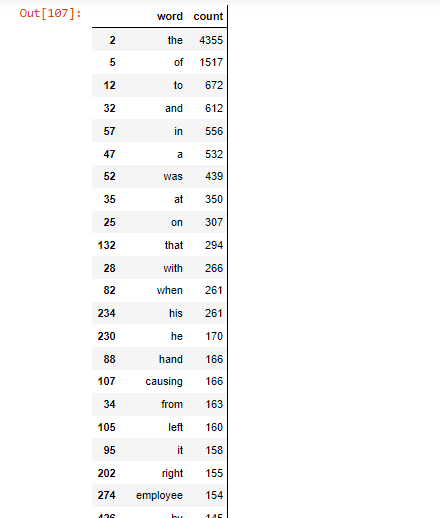
**Plotting the most frequent words with word cloud**



#### **Observations**

* There are many body-related, employee related, movement-related, equipment-related and accident-related words.
* Body-related: left, right, hand, finger, face, foot and glove
* Employee-related: employee, operator, collaborator, assistant, worker and mechanic
* Movement-related: fall, hit, lift and slip
* Equipment-related: equipment, pump, meter, drill, truck and tube
* Accident-related: accident, activity, safety, injury, causing

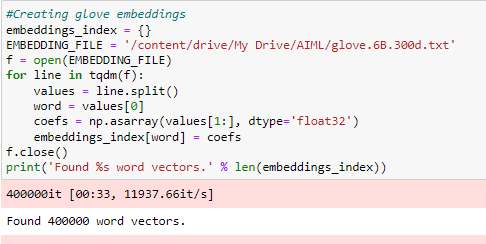
**Analyzing the frequent words**



**NLP Pre-processing Summary:**

* 74% of data where accident description > 100 is captured in low accident level.
* 34% of data where accident description > 100 is captured in high medium potential accident level.
* 25% of data where accident description > 100 is captured in medium potential accident level.
* 23% of data where accident description > 100 is captured in low potential accident level.
* Few of the NLP pre-processing steps taken before applying model on the data
* Converting to lower case, avoid any capital cases
* Converting apostrophe to the standard lexicons
* Removing punctuations
* Lemmatization
* Removing stop words
* After pre-processing steps:
* Minimum line length: 64
* Maximum line length: 672
* Minimum number of words: 10
* Maximum number of words: 98

#### **Glove Word Embeddings**



Identified 400000-word vectors

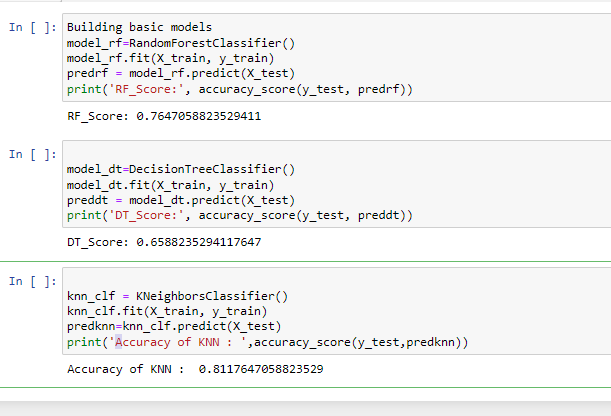
**Deciding Models and Model Building**

**Basic Model**

For modelling both machine learning and deep learning algorithms have been used, compared performances of all the models and chosen the model with highest score. Objective is predict both ‘Accident Level’ & ‘Potential Accident Level’ labels, based on description models for both these variables are generated and model giving best accuracy scores for both these variables are chosen.

We have build the ML model Random forest, Decision tree KNN Classifier

We have less number of inputs and no outliers.



KNeighbors Classifier class. The neighbors number is important in this method. Selecting the right number of neighbors provides the more accurate results.

train the KNN classifier with the dataset for n=10 neighbors and see how much accuracy

KNN accuracy score is better than Random Forest and decision tree.

**Model performance Improvements**