Capstone Project Interim Report

*Industry Data Chatbot (NLP-2)*

Group: AIML Feb’21D Group 5 (AIML Capstone G2)

*Date: 06-Mar-2022*

Capstone Project Submitted by:

*Himanshu Kohli*

*Mridul Miglani*

*Vishnudas Raveendran*

*Hemlata Channe*

*Saurabh Singh*

Domain:

Industrial safety. NLP based Chatbot.

Context:

The database comes from one of the biggest industry 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.

1. Summary of Problem Statement, Data and findings

**Problem Statement**

* Since industrial accidents across various sectors are quite common, it is important to know common causes of such incidents so that employees are aware and these can be prevented to save them from injuries, health issues and deaths.
* It would be useful identify an approach through which we can derive the level of accident impact that would is likely to occur if we know some details of the incident including keywords like machine used, steps involved, type of movements involved through various body parts, type of industry, etc

**Abstract**

* Exploratory data analysis, NLP pre-processing, Machine learning and Deep Learning techniques would be applied on the descriptions so that we can categorize impact/ level of accident that can happen.
* Chatbot interface to be created to feed input keywords and get accident level as output.

**Data and findings**

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

**Columns and descriptions:**

**‣ 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)

**‣ Gender:** if the person is male of female

**‣ Employee or Third Party:** if the injured person is an employee or a third party

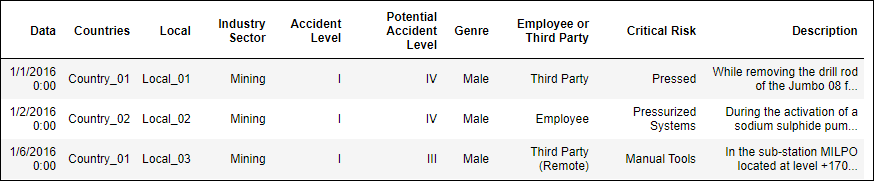
**‣ Critical Risk:** some description of the risk involved in the accident

**‣ Description:** Detailed description of how the accident happened.

**Datatypes:**All columns are of ‘Object’ type

**Data Format:**csv file

**Sample data:**



**Shape of file:**

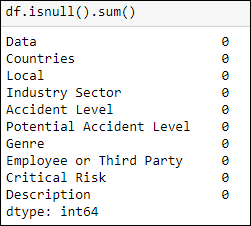
*425 rows, 11 columns*

***Date Range:***

Jan-2016 to Jul-2017

**Missing values / Nulls:**

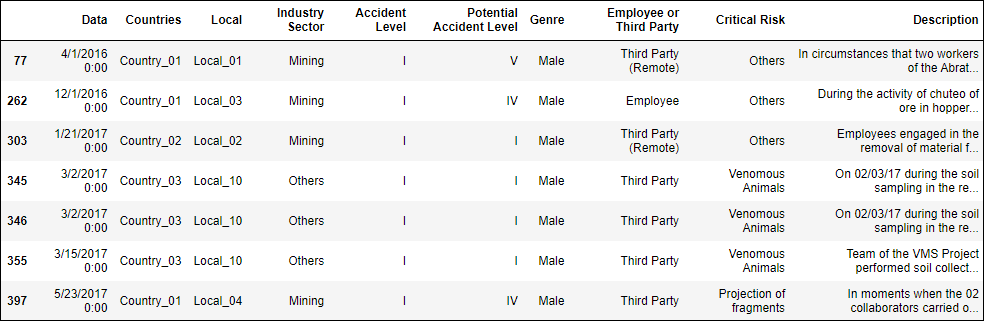
No null values are present in any column



**Duplicate values:**

7 duplicate records are present in data. These were removed.

After removing these, 418 rows remained.



**Class balancing:**

Target variable ‘Accident level’ have highly imbalanced classes:

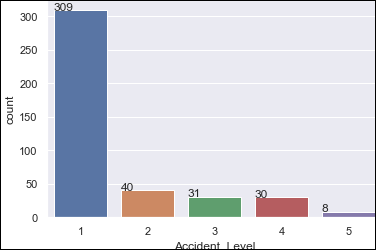
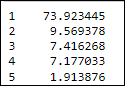


\*Oversampling techniques to be applied on dataset.

1. Summary of Approach to EDA and Pre-processing

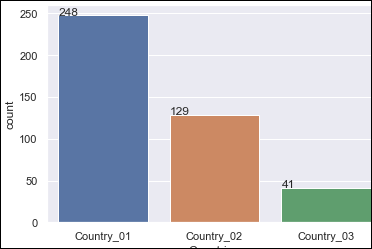
**Univariate Analysis:**

Count by Accident Level:

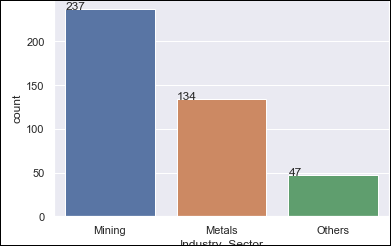
**Inference:** Accident level 1 is most frequently occurring accident.

Count by Country:

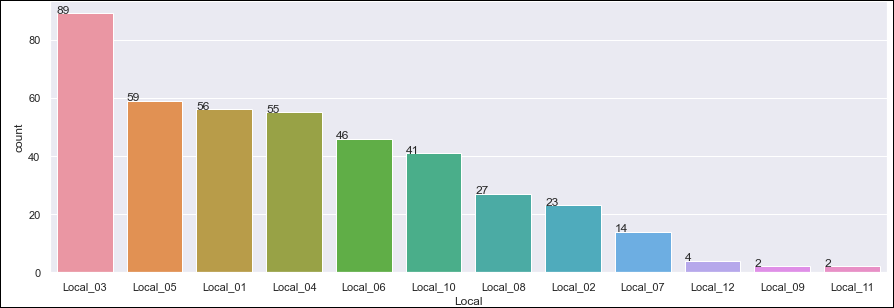
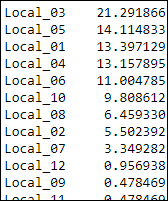
**Inference:** More accidents are occurring in Country\_01 and least in Country\_03.

Count by Industry Sector:

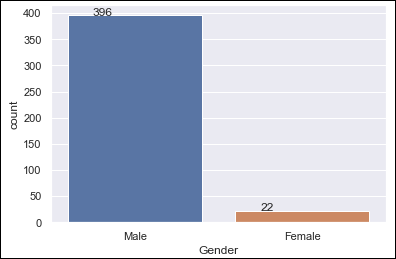
**Inference:** Highest accidents occur in Mining sector.

Count by Local:

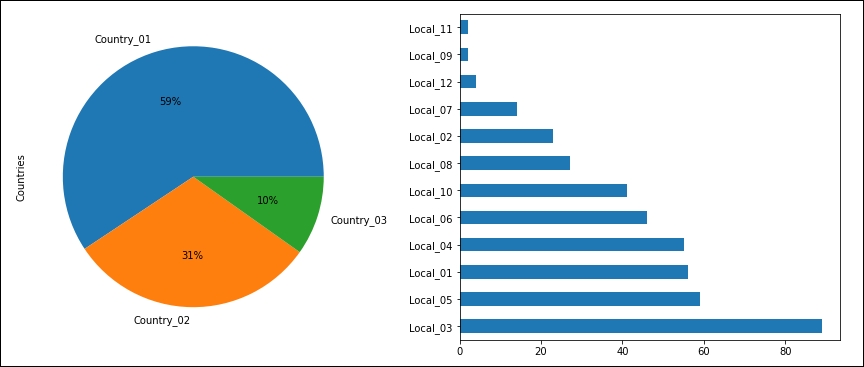
**Inference:** More accidents are occurring in Local\_03 site and least in Local\_11 site.

Count by Gender:

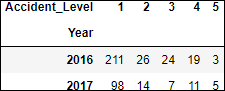
 

**Inference:** Most of accidents have occurred for Males (around 95%).

Distribution of accidents in Countries and Local:

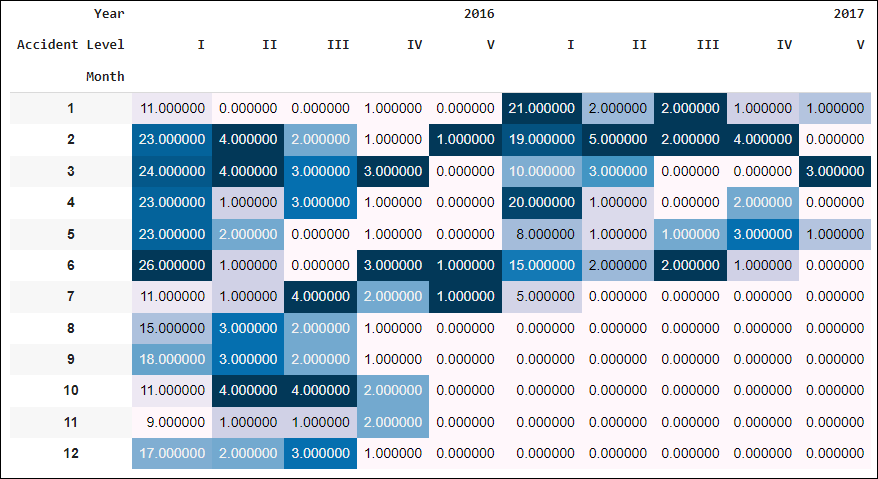


Accident Level Counts by Year:



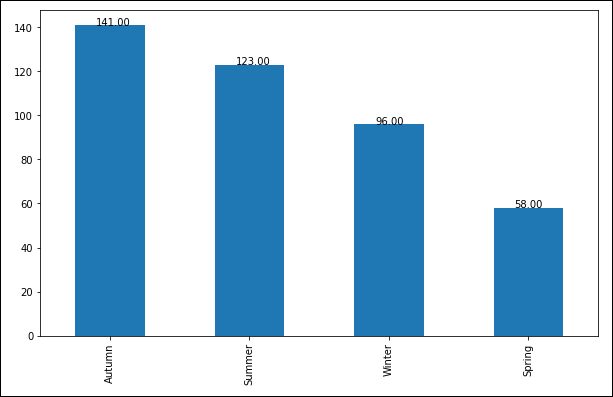
**Inference:** More accidents occurred in 2016 and highest occurrence is of Accident Level 1 in both years.

Accident Level Counts by Month:



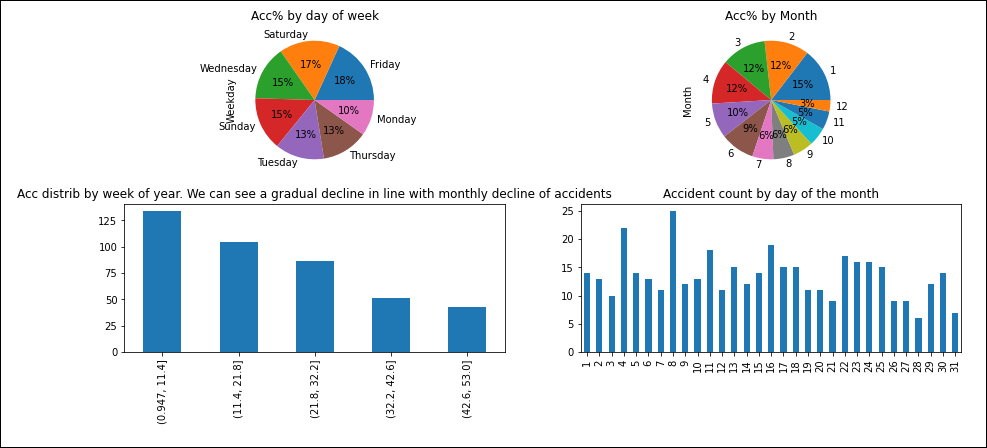
**Inference:** Level 1 accidents have frequent occurrence across all months for both the years.

Distribution by Seasons:



**Inference:** Highest accidents occur in Autumn and lowest in Spring.

Distribution by day of week and month:

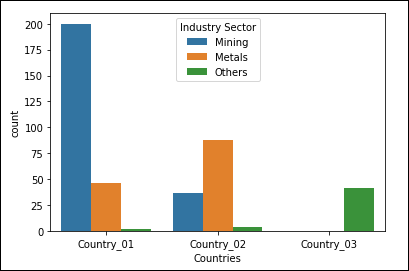


**Inference from above 2 plots:**

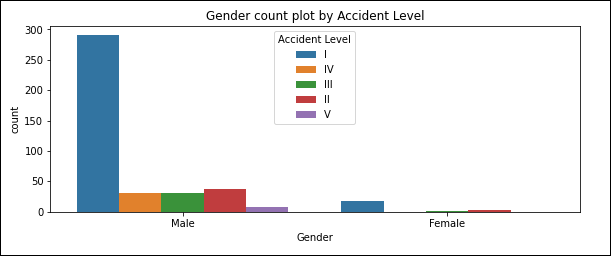
* + We see a difference of 8% in accidents between Monday and Friday. This could be due to accidents caused due to workers hurrying to finish the job for the weekend.
  + The accidents are most in the first \*four\* months and gradually decreases to half by July. By December, the number of accidents have fallen to 3% as compared to 15% in January (Evidenced also by week of year distribution)
  + If we consider the sudden spike of accidents on the 4th and 8th day of the month as outliers, we can see a curve building-up to a peak mid-month before gradually receding.

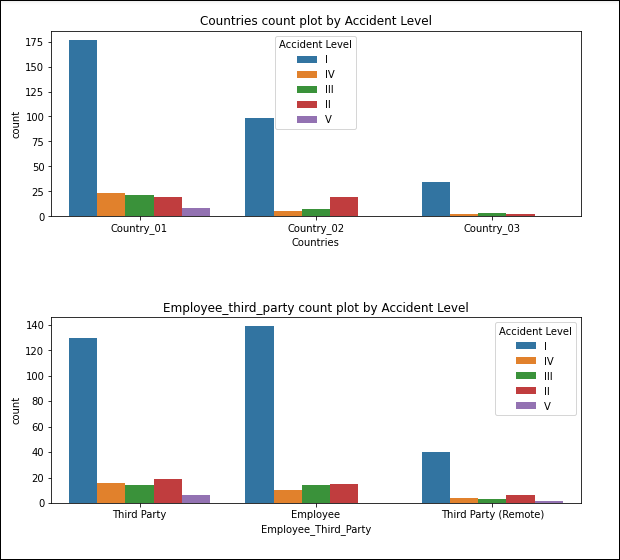
**Bi-variate Analysis:**

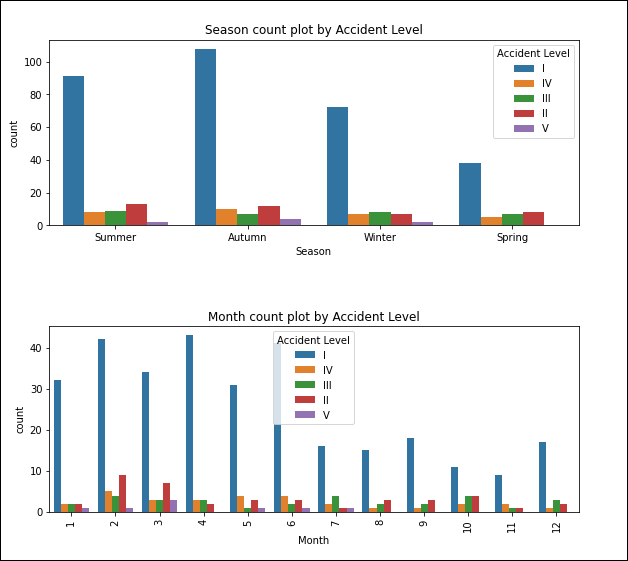
Counts by Country and Industry sector:



For 2016, Level 1 accident counts are higher for Jan to June months and all accident levels are reducing towards year end.







**Inferences:**

Bivariate plots of columns 'Gender','Country','Employee\_Third\_Party','Industry\_sector','Season','Month' w.r.t. Accident level shows that:

- Male has more number of accidents in accident level I.

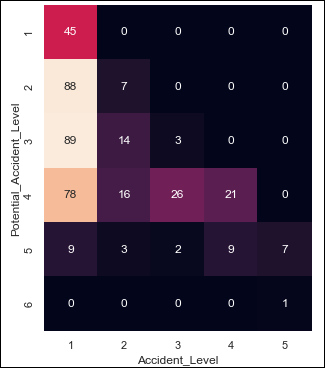
- All countries have more accidents in Level I.

- All type of employees have accident level I as higher count.

- In all industry sectors, seasons and months accident level I as higher count.

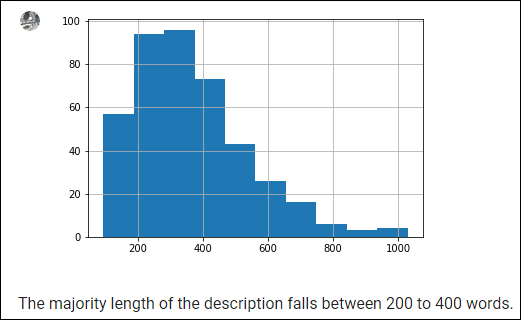
* For males, most accidents are for Employees, then third party and leas for third party remote (due to obvious reasons as they are very less exposed to vulnerable working conditions)
* Female employees have majorly level 1 accidents and more are employee and third party.
* For males, accidents are high in both Mining and metals sector (highest in mining).
* For females, accidents are more in metals sector (reason could be due less females working in Mining sector).

Correlation between Accident Level and Potential Accident Level:



**Inferences:** Frequency is high for Accident Level 1 and Potential Accident Level 1,2,3,4. Even with Accident Level 1, higher levels of Potential Accident are associated for many records.

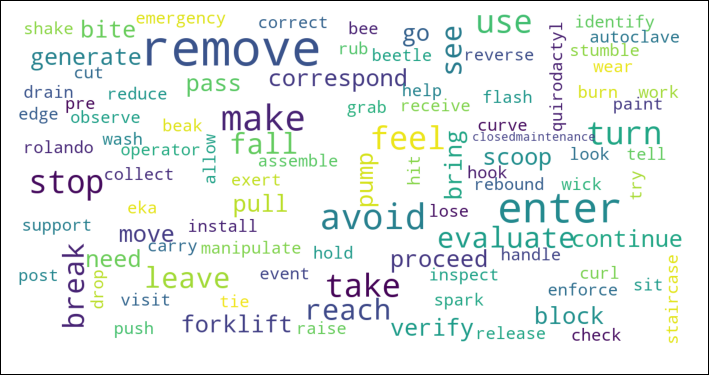
Analysis of Description:



**Word Cloud**

In word cloud, we can see below categories of words:

1. **Word Cloud of Verbs:**





**Inference:**

We see that the top causes are

`Removal, Fork-lift, Fall, Bite, reach, take etc`

The possible victims are

`Operator, Assistant, Employee, Collaborator, Worker, Mechanic`

Most body-injuries happen to

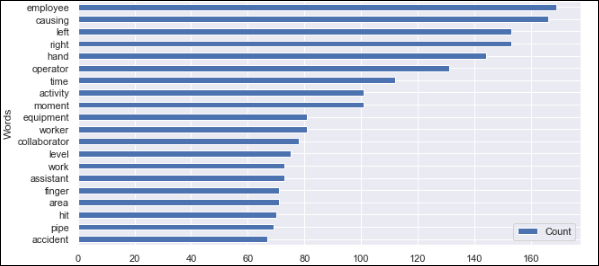
`Hand, Finger, Neck, face`

Most accident-prone equipment/area are

`Truck, Drill, Pump, Ladder, Platform, Tube, Pipe, Mesh`

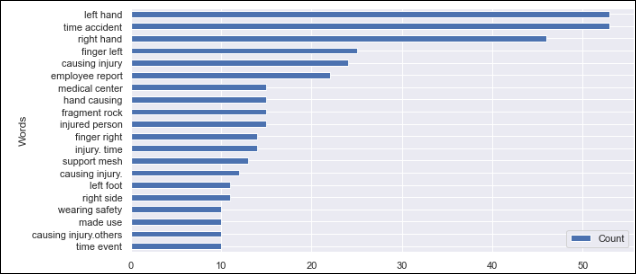
**Ngrams-**

Top 20 Uni-grams



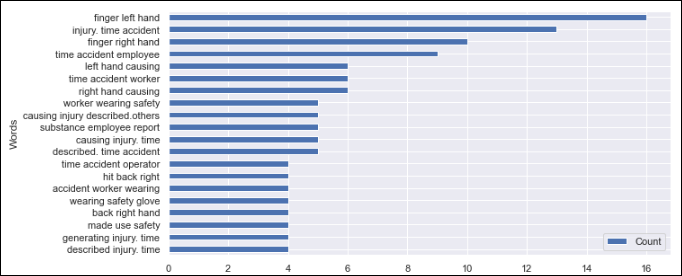
**Inference:** Top uni grams include words like employee, causing, left, right, hand, etc.

Top 20 bi-grams:



**Inference:** Top bi-grams include words like left hand, time accident, right hand, finger left, etc.

Top 20 tri-grams:



**Inference:** Top tri-grams include words like finger left hand, injury time accident, finger right hand, time accident employee, left hand causing, etc.

**After applying NLP pre-processing, description looks like below:**

Converting to lower case, strip (removing leading and trailing spaces), removal of special characters, lemmatization

