Capstone Project Interim Report

*Industry Data Chat-bot*

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 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 (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

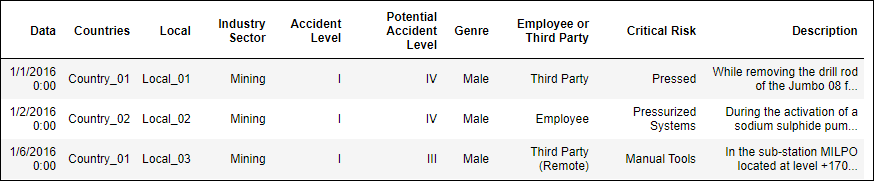
‣ Critical Risk: some description of the risk involved in the accident

‣ Description: Detailed description of how the accident happened.

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

**Data Format:**csv file

**Sample data:**

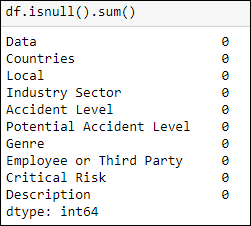


**Shape of file:**

*425 rows, 11 columns*

**Missing values / Nulls:**

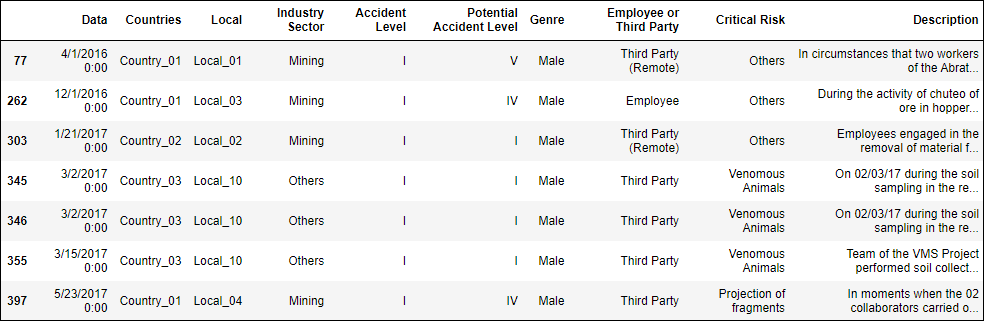
No null values are present in any column



**Duplicate values:**

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

After removing these, 418 rows remained.



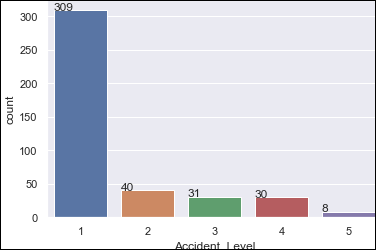
**Class balance:**

Target variable ‘Accident level’ have highly imbalanced classes:

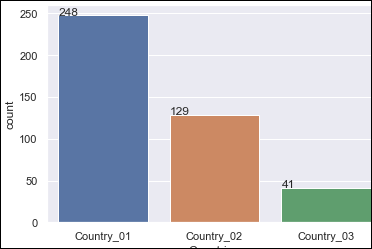


1. Summary of Approach to EDA and Pre-processing

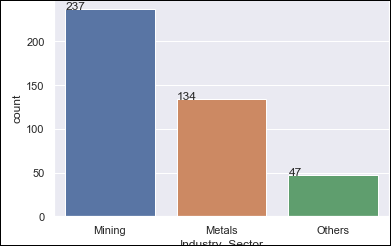
Accident level 1 is most frequently occurring accident.



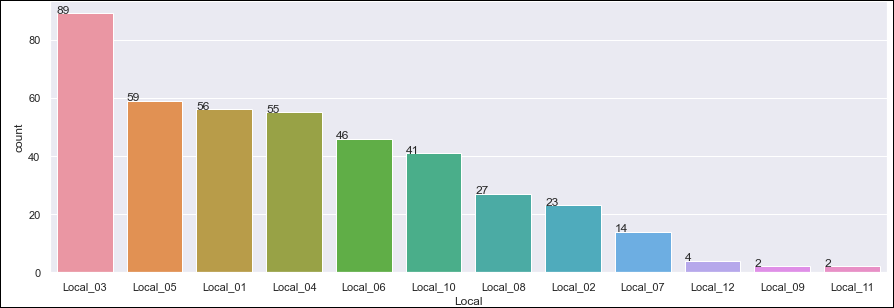
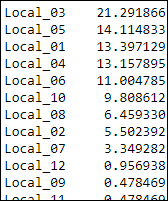
More accidents are occurring in Country\_01 and least in Country\_03.



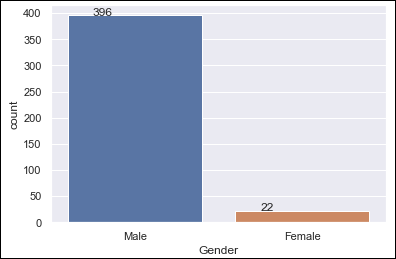
Highest accidents occur in Mining sector.

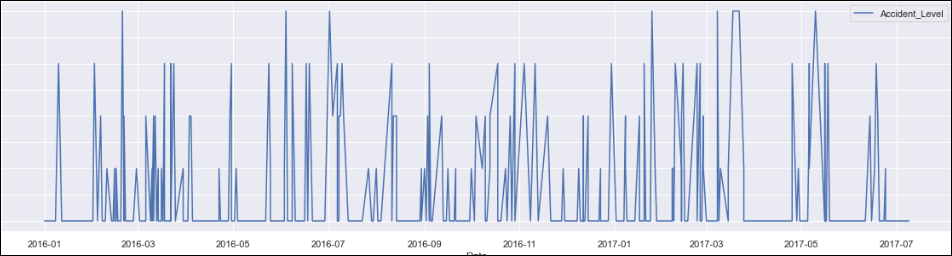
More accidents are occurring in Local\_03 site and least in Local\_11 site.

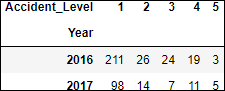
Most of accidents have occurred for Males (around 95%).

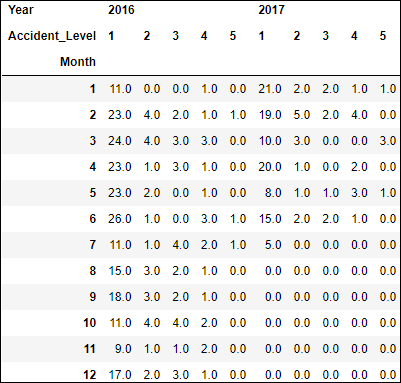
Accidents appear to be some seasonality in occurrence as there are peaks and no accidents on certain dates.



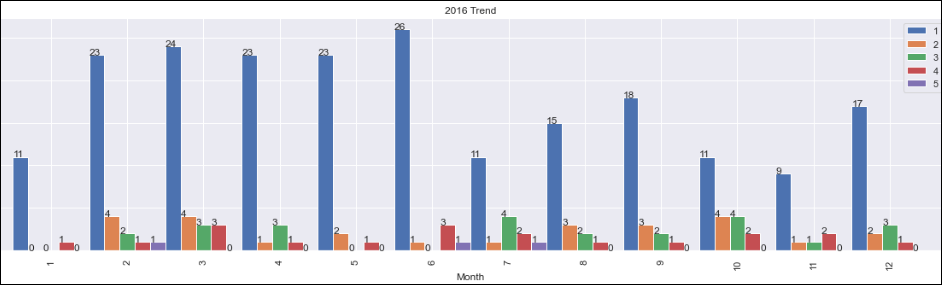
More accidents occurred in 2016 and highest occurrence is of Accident Level 1 in both years.



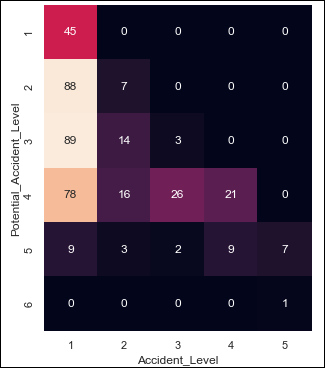
Level 1 accidents have frequent occurrence across all months for both the years.



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

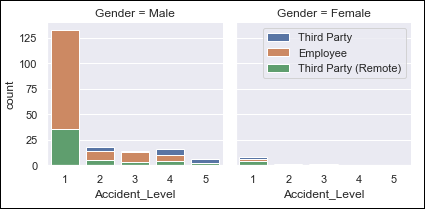


Even with Accident Level 1, higher levels of Potential Accident are associated for many records.



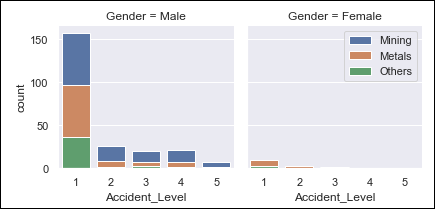
For males, most accidents are for Employees, then third party and leas for third party remote (due to obvious reasons as they are very less exposed to vulnerable working conditions)

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).



**Word Cloud**

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

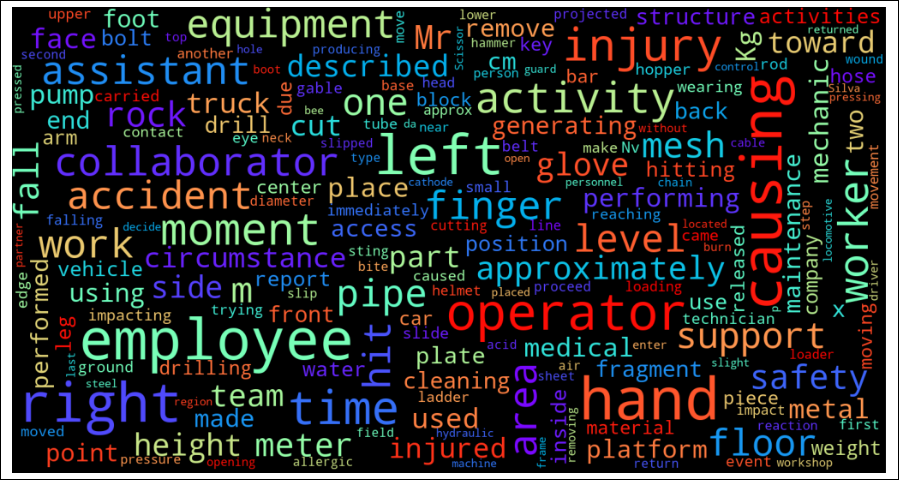
\* Type of action: Causing, opening, fall, drilling, used, support, cut, remove, cleaning, moving, slide, slip, released

\* Body parts, position/ direction, equipment: finger, glove, hand, leg, foot, left, right, helmet

\* Machine part / type: Hydraulic, pump, bolt, ladder, truck, vehicle, cathode, collaborator, belt, plate

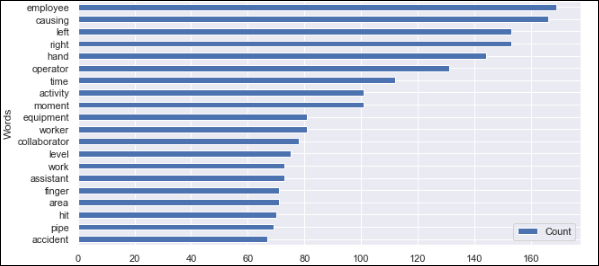
\* Impact type: Sting bite, injured, burn, pressed, allergic

\* Type of role: Operator, technician

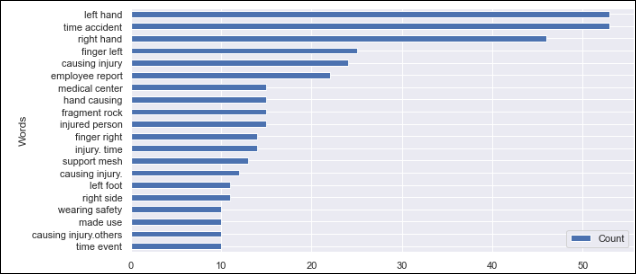


**Ngrams-**

Top Uni-grams



Top bi-grams



Top tri-grams

