

NLP CAPSTONE PROJECT - Final Report

NLP- 2 Semi Ruled Chatbot

CONTENT

1. Summary
2. Overview of the final process
3. Step by step walkthrough the solution
4. Model evaluation
5. Comparison to benchmark
6. Visualizations
7. Implications
8. Limitations
9. Closing reflection

1. Summary

Given a dataset with records of accidents from 12 different plants in 03 different countries describing different accidents. We have developed a chatbot which can accurately predict the accident level using the description the user provides. We have tried different vectorizers and different classification models to predict the accident level. This chatbot can be used to effectively predict the accident level and provide required assistance as per the incident description.

GitHub Link:- <https://github.com/SaiCharan99/GL-Capstone-NLP2>

2. Overview of the final process

2.1 Data Cleaning

Shape: There are 425 records and 11 attributes/columns

```
: (425, 11)
```

Missing values: There are no missing values in the data

```
: Date                0
   Country             0
   Local              0
   Industry Sector     0
   Accident Level      0
   Potential Accident Level 0
   Gender              0
   Employee type       0
   Critical Risk       0
   Description         0
   dtype: int64
```

Type of attributes: All columns are object type. There is no numerical attribute present.

```
: Date          object
Country        object
Local          object
Industry Sector object
Accident Level object
Potential Accident Level object
Gender         object
Employee type  object
Critical Risk  object
Description    object
dtype: object
```

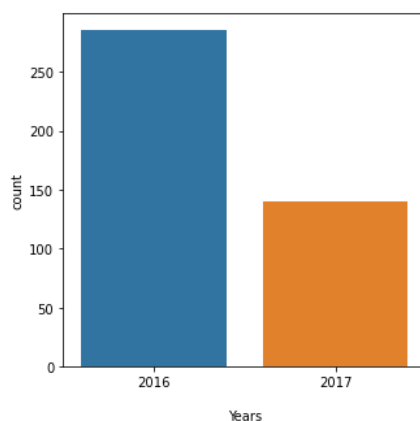
2.2 EDA

Univariate Analysis: In this, each attribute is individually analysed.

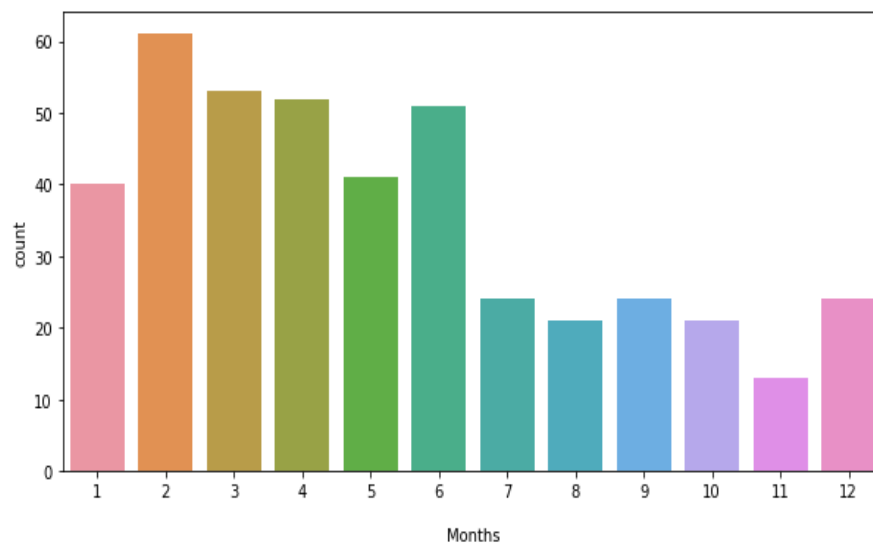
DATE:

Here, the column name is renamed from Data to Date. Also, the year, month and day variables are extracted from the date column to find if there are any seasonality co-occurrences of accident.

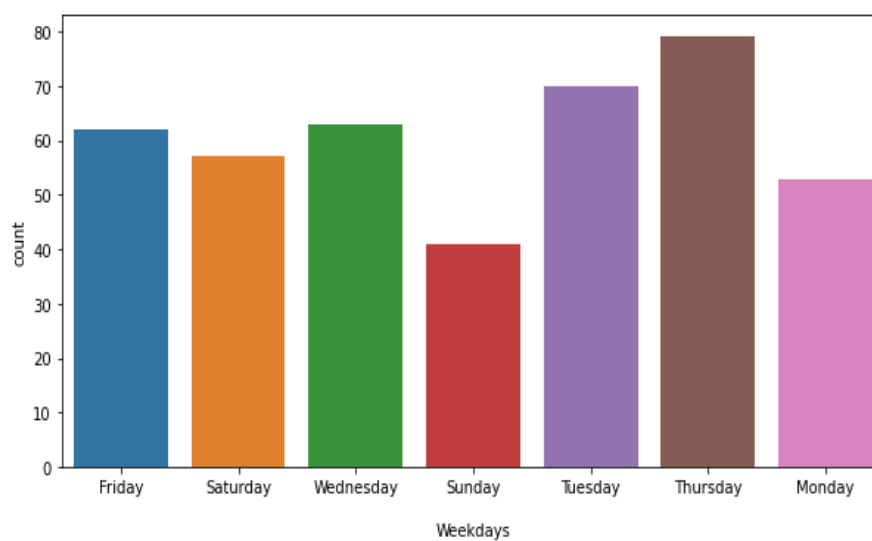
Year-



Month-



Day-

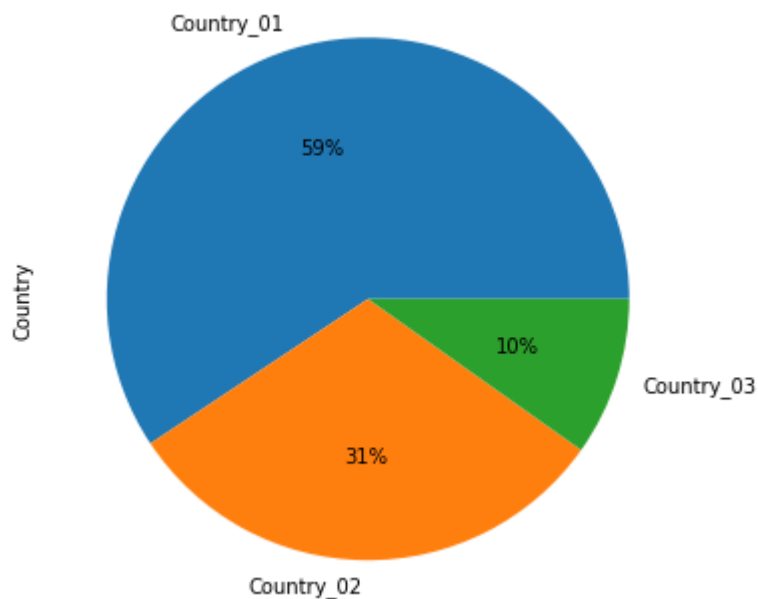


Observations:

- There are lots of records for the year 2016, contributing to approximately 67%, whereas only approximately 33% for the year 2017.
- Most of the incidents are recorded during the first 6 months than the last six months of a year
- Thursday has recorded the highest incidents

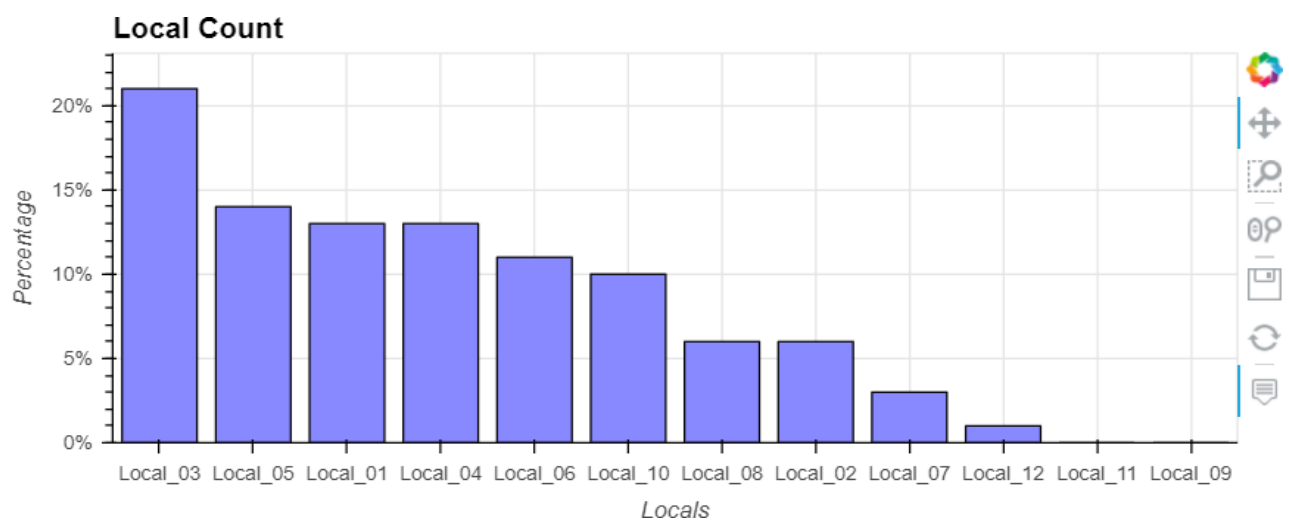
COUNTRY:

Here, the column name is renamed from Countries to Country



Observation: The most affected country from the above dataset is Country_01 with around 59% of the accidents, whereas Country_02 and Country_03 counts for 31% and 10% respectively.

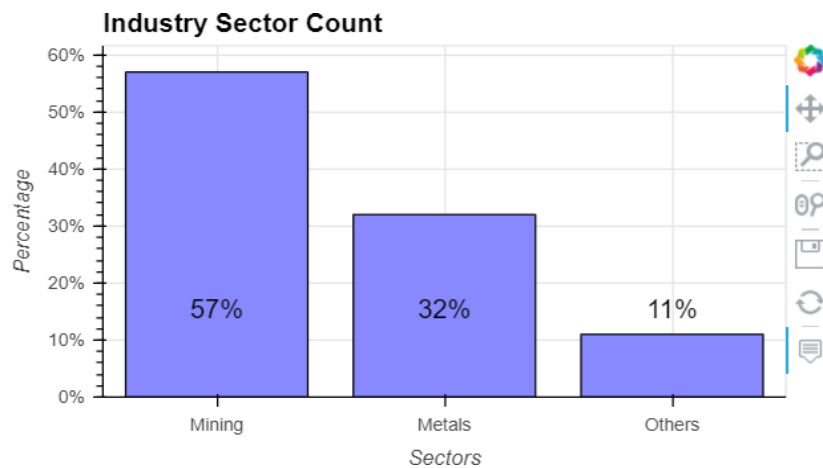
LOCAL:



Observations:

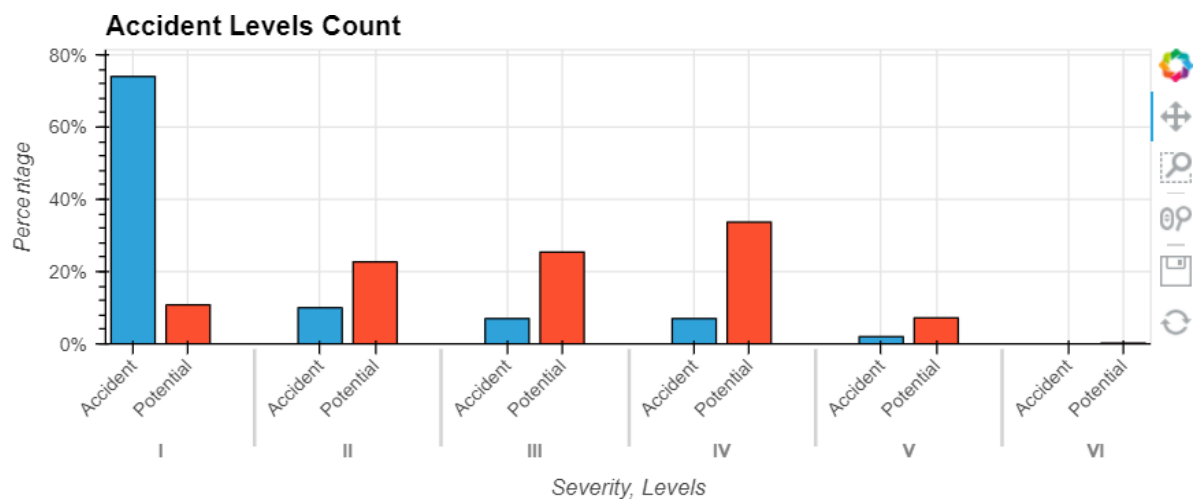
- Highest manufacturing plants are located in Local_03 city.
- Lowest manufacturing plants are located in Local_09

INDUSTRY SECTOR:



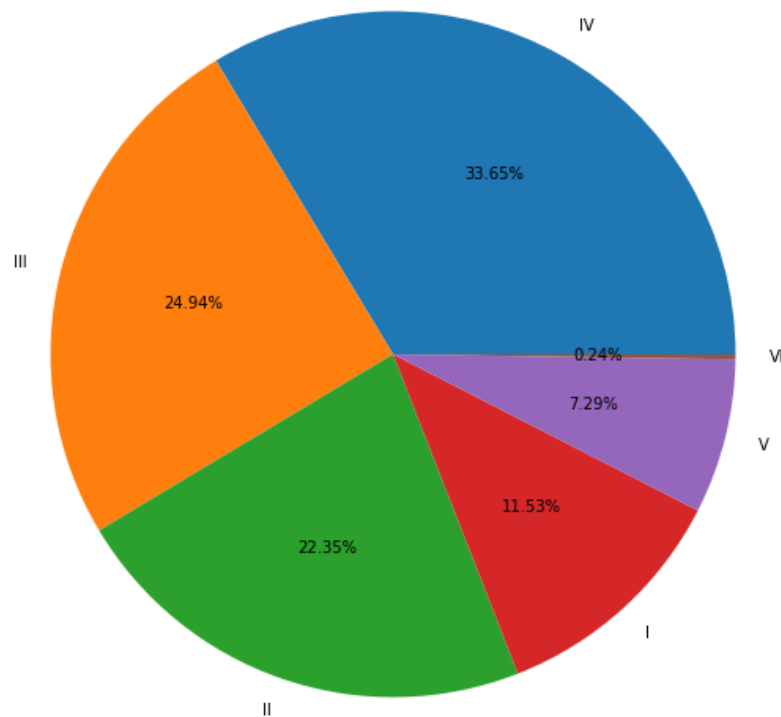
Observation: Most of the manufacturing plants belongs to Mining sector.

ACCIDENT LEVEL:



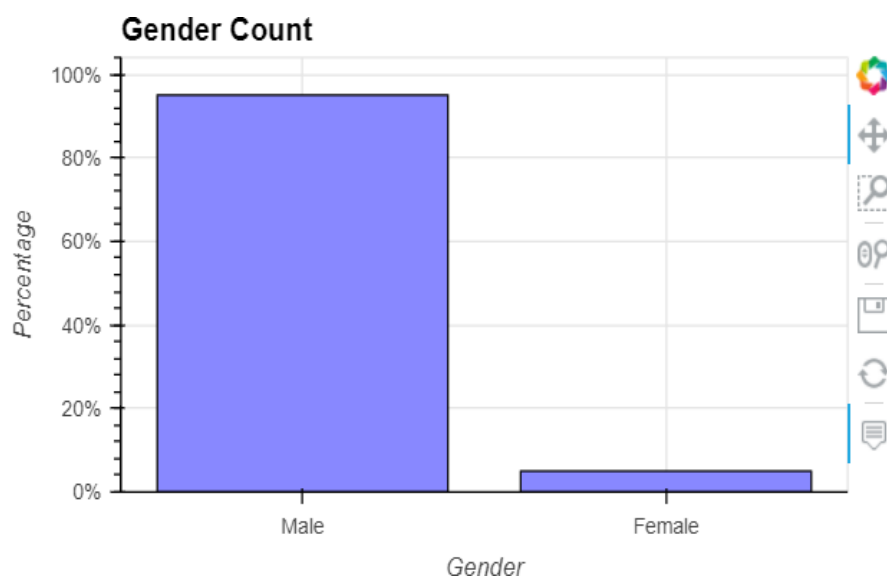
Observation: Most accidents belongs to "Accident Level" I .Its count is 316 which is equivalent to 74.35%% of total accidents

POTENTIAL ACCIDENT LEVEL:



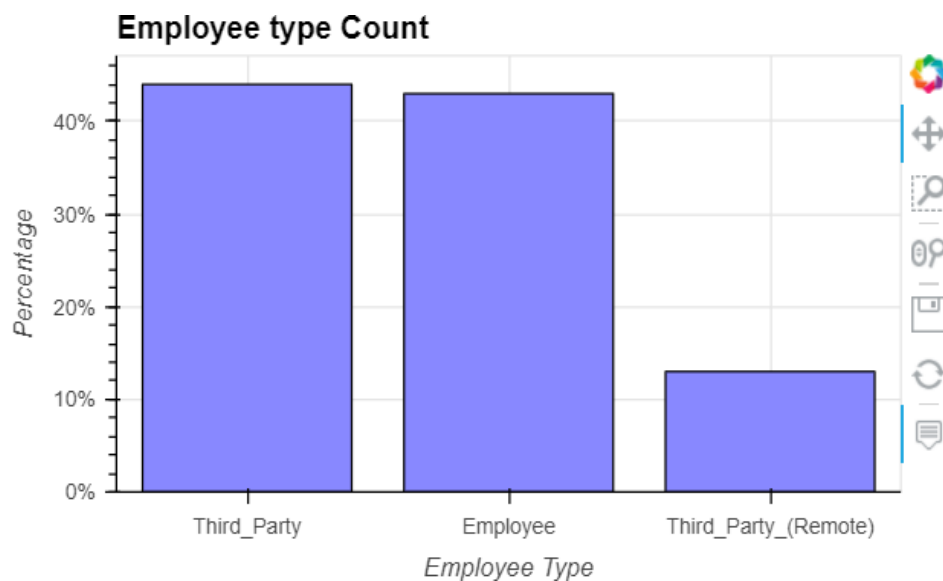
Observation: Most "Potential Accident Level" belongs to level IV .Its count is 143 which is equivalent to 33.65% of total potential accidents.

GENDER:



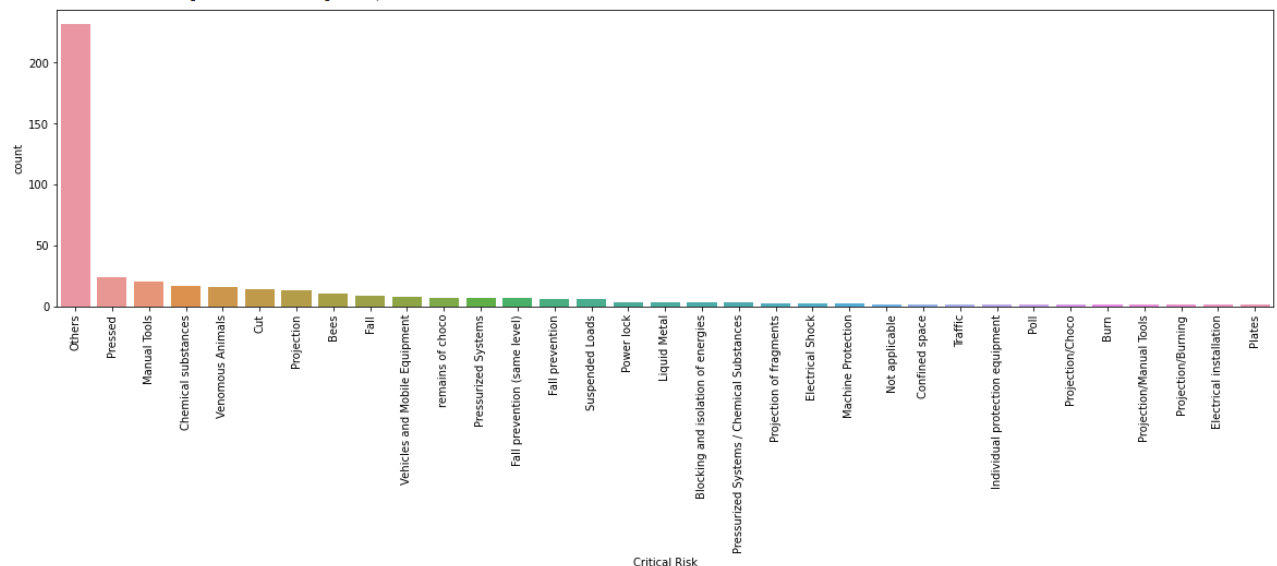
Observation: There are more men working in this industry as compared to women.

EMPLOYEE TYPE:



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

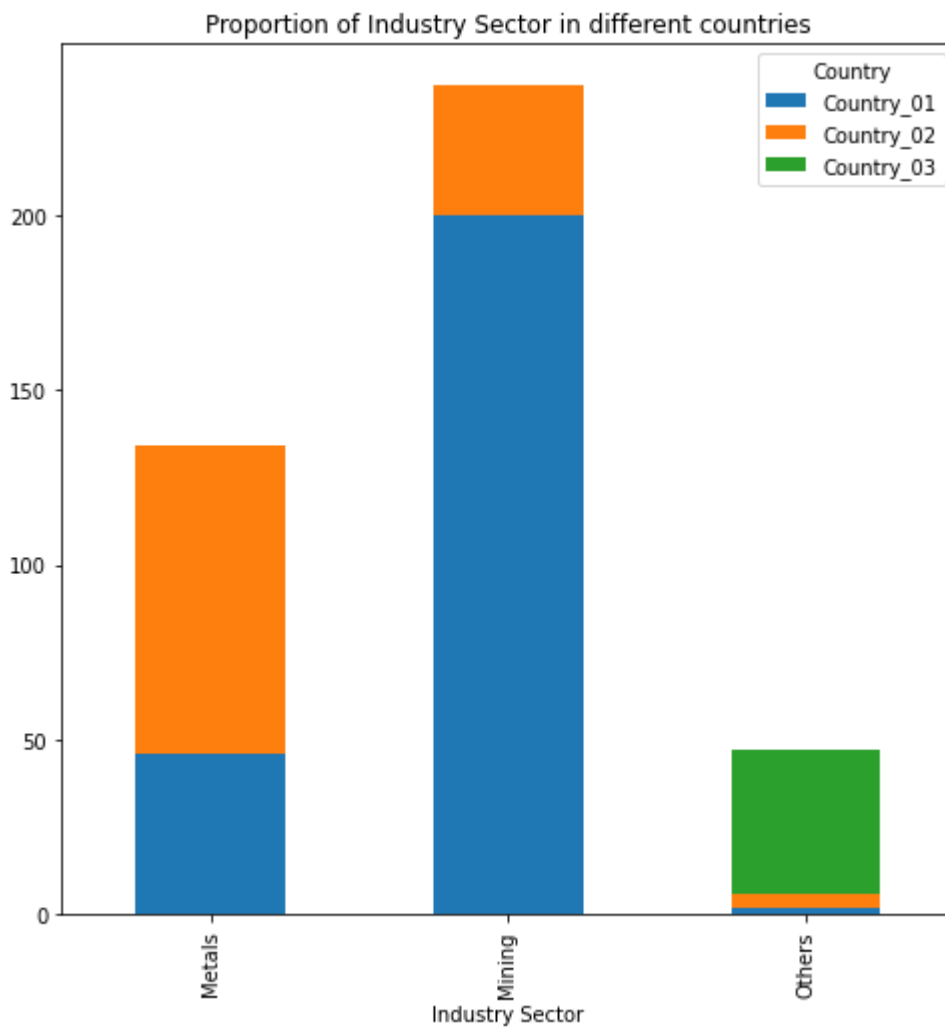
CRITICAL RISK:



Observation: Most of the incidents are registered as 'Others', it takes lot of time to analyse risks and reasons why the accidents occur.

Bivariate Analysis:

Country and Industry Sector:

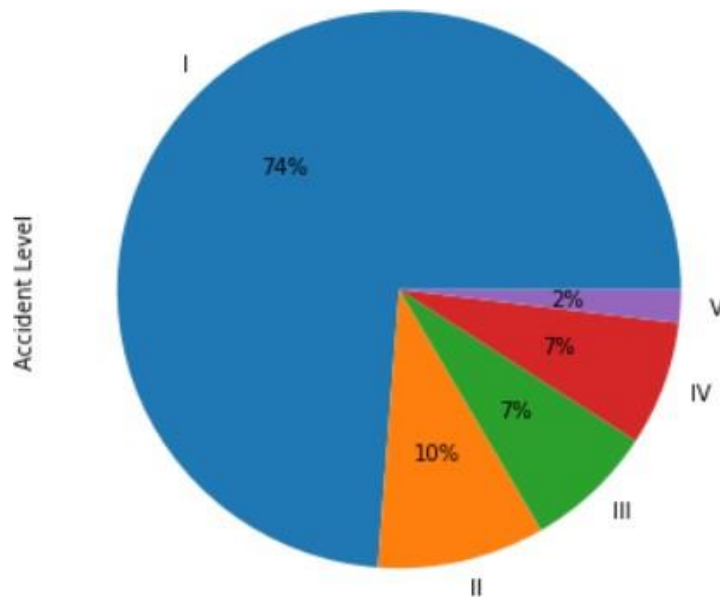


Observations:

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

2.3 NLP Analysis

- Distribution of accident level where the length of Description is greater than 100



- 74% of data where accident description > 100 is captured in low accident level.
- Based on some random headlines seen above, it appears that the data is mostly lower-cased. Pre-processing such as removing punctuations and lemmatization can be used.
- There are few alphanumeric characters like 042-TC-06, Nv. 3370, CX 212 captured in description where removing these characters might help.
- There are digits in the description for e.g. level 326, Dumper 01 where removing the digits wouldn't help.

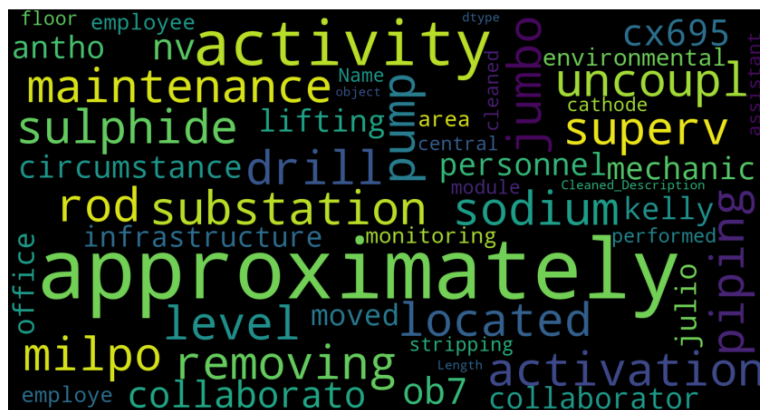
2.4 NLP Pre-processing

Here, the main task is to process a language given in Description column in a right way to get the best results.

- Converting to lowercase for all words
- removing numbers
- removing hyperlinks
- removing hashtags
- only removing the hash # sign from the word
- removing stopwords
- removing punctuation

WordCloud Analysis

Wordcloud for cleaned description:



Observation: Most words are related to maintenance, accident, employee, equipment, infrastructure.

3. Step by step walkthrough the solution

To start with data cleaning, we have checked for duplicate rows and missing values and handled them. After that, EDA was performed followed by NLP analysis and NLP pre-processing. We analysed the 'Description' feature and cleaned it by removing unwanted characters, stopwords and tokenizing the remaining words.

We used multiple vectorizers like TF-IDF, Count and Glove with word2vec and found that Glove was giving better results. We tried various machine learning models like Naïve Bayes, SVC, Random classifier, Bagging classifier, LSTM and BiLSTM and found that Random classifier and Bagging classifier gave the best accuracy.

Though we got around 80% validation accuracy with them, we observed that parameters like precision, recall and f1 score depicted that the models were skewed to 1 class. It occurred due to imbalanced target data. Hence, we used SMOTE to overcome this and got a better result on the other parameters.

4. Model Evaluation

We picked **Bagging Classifier with SMOTE using glove** as our final model. SMOTE is used to overcome the issue with imbalanced data.

We evaluated mainly using parameters like precision, recall & f1-score and found that this model gave better results as compared to other models that we tried.

Model performance:

Bagging Classifier with SMOTE classification report:-


	precision	recall	f1-score	support
--	-----------	--------	----------	---------

1	0.84	0.79	0.82	68
2	0.25	0.33	0.29	6
3	0.00	0.00	0.00	5
4	0.40	0.50	0.44	4
5	0.00	0.00	0.00	2
accuracy			0.68	85
macro avg	0.30	0.33	0.31	85
weighted avg	0.71	0.68	0.70	85

5. Comparison to benchmark

Initial benchmark was 80% in terms of accuracy, but due to imbalanced data the other parameters were low. We were able to improve on other parameters in our final model using SMOTE.

Chat bot results on Bagging Classifier with SMOTE (Final model)

 Accident Chatbot ACB

Hi! Welcome to the accident report portal.

My name is Accident Chatbot ACB.
Please describe the accident in the below box:-

You -> Approximately 1:40 p.m. in circumstances that shotcrete was launched in the Nv. 1680 BP 255 of the OB2B, after finishing the launch of the first mixkret 113, the assistant of the alpha, Mr. Albertico asks the operator of the mixkret 113, Mr. Jhony to move the mixkret 116, so that access, finding in the cockpit of the mixkret the operator of the Launcher team

ACB -> The accident is of level: V

You -> During maintenance of the Flyght pump rotor, the oil pressure of the the lubrication chamber caused the chamber cover to be projected towards the employee's face, striking him superficially on the forehead, causing injury.

ACB -> The accident is of level: II

You -> During the refurbishment work of the HDPE pipes (4 "of diameter) with two workers, when the worker who secured the pipe with a chain, standing on the basket of the amploader - raised to a height of 3.0 m from the ground -, the pipe slipped and impacted his arm right, causing an injury to the radius of the right arm.

ACB -> The accident is of level: IV

Send

Chat bot results on Random Classifier without SMOTE

Accident Chatbot ACB

Hi! Welcome to the accident report portal.

My name is Accident Chatbot ACB.
Please describe the accident in the below box:-

You -> Approximately 1:40 p.m. in circumstances that shotcrete was launched in the Nv. 1680 BP 255 of the OB2B, after finishing the launch of the first mixkret 113, the assistant of the alpha, Mr. Albertico asks the operator of the mixkret 113, Mr. Jhony to move the mixkret 116, so that access, finding in the cockpit of the mixkret the operator of the Launcher team

ACB -> The accident is of level: I

You -> During maintenance of the Flyght pump rotor, the oil pressure of the the lubrication chamber caused the chamber cover to be projected towards the employee's face, striking him superficially on the forehead, causing injury.

ACB -> The accident is of level: I

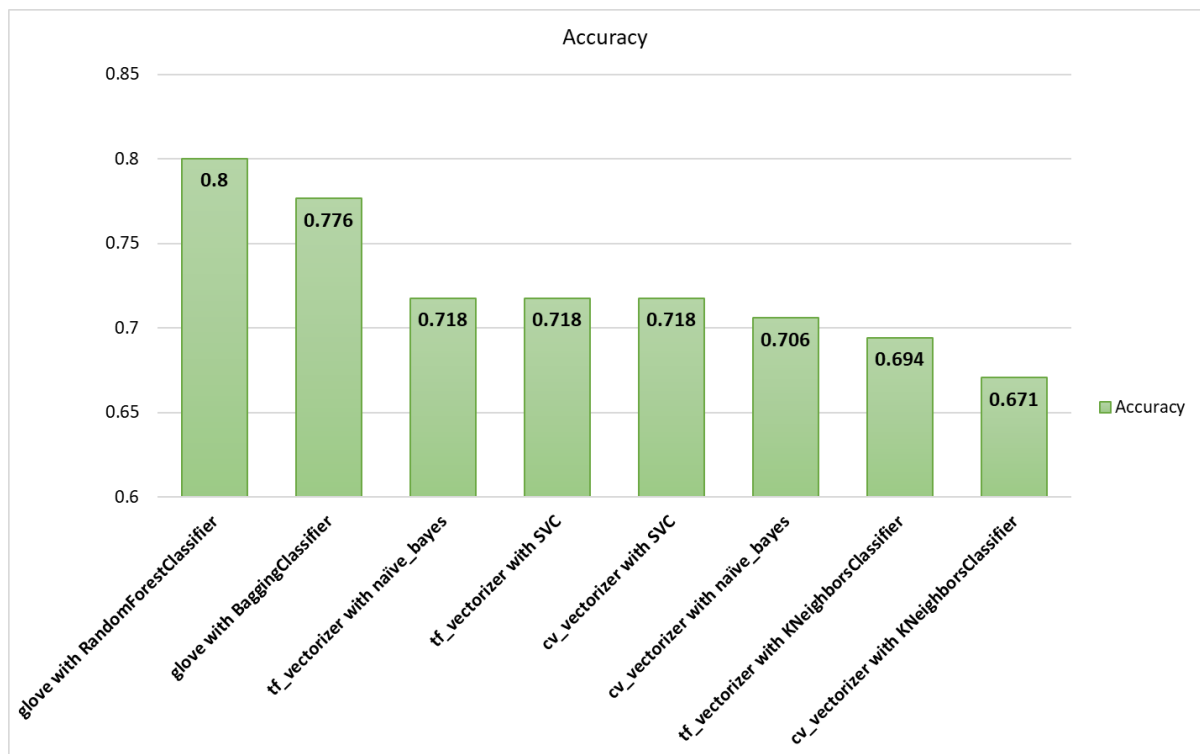
You -> During the refurbishment work of the HDPE pipes (4 "of diameter) with two workers, when the worker who secured the pipe with a chain, standing on the basket of the amploader - raised to a height of 3.0 m from the ground -, the pipe slipped and impacted his arm right, causing an injury to the radius of the right arm.

ACB -> The accident is of level: I

Send

6. Visualizations

The following bar chart displays performance of all other models in descending order:



7. Implications

Our solution solves the problem of providing accurate help during accidents in the mechanical industries prone to such incidents.

We would recommend to gather more data for each type of target and improve the model with it.

8. Limitations

Following are few limitations:

- i) The manual work required by the worker to type in the description. This delay could be critical in the real world. To enhance this, we could introduce text to speech and inculcate the model into an automated system which could provide effective help.
- ii) The model is trained on very small dataset. Many other real-world accidents are not considered here. Hence this could be eliminated by collecting more data and building a more robust model on it.

9. Closing reflection

We have learned the end to end process of converting a database entry into a working chatbot that can effectively predict the accident level when we enter the accident's description.

During this process, we have also learned to try various vectorizers and use them to build numerous other ML models. Many other latest models like BERT, RoBERTa, XLNet etc can also be tried and tested.

-----END-----