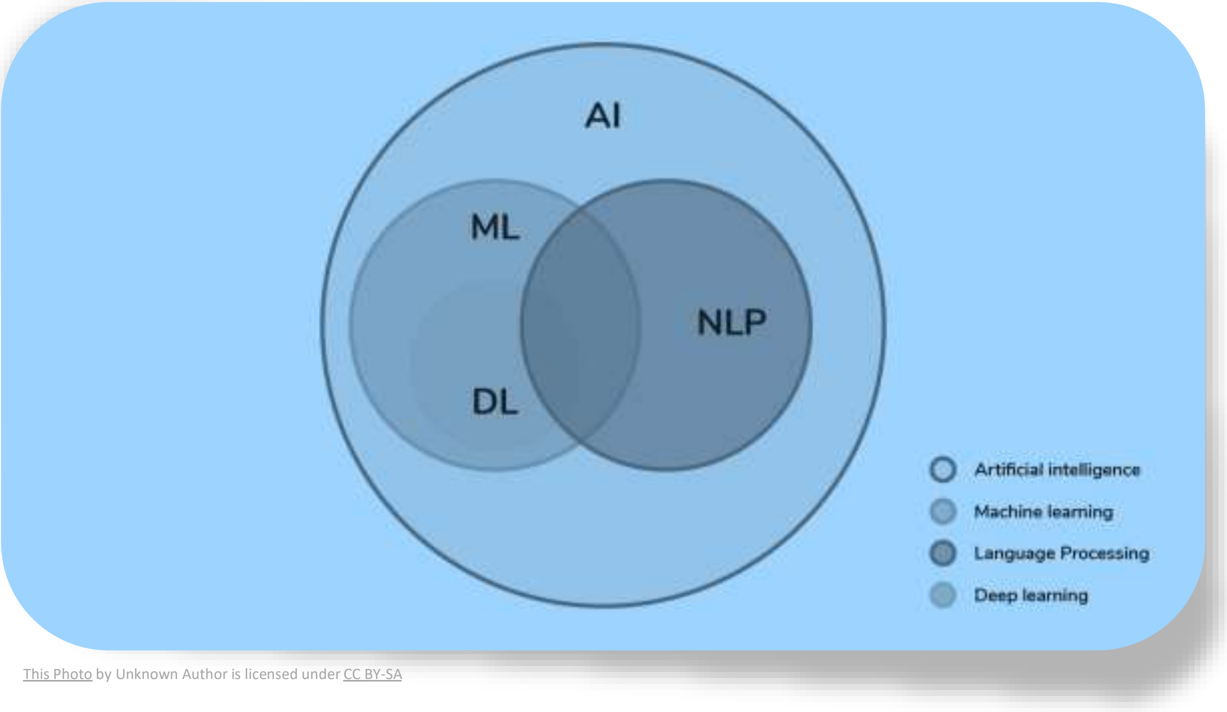
AI ML Capstone Project – NLP 2 Interim Report

EDA Analysis for

NLP based Chatbot project for safety department of Industrial companies



Project Interim Report prepared by

Amol Bulbule
Dr. Kaushik Sarkar
Manoj Vardhan
Suresh Rajagopalan
Venkatesh Chandiran

Presented to Great learning Mentor **Mr. Srikanth Girijala**

Date: 08-Aug-2021

Disclaimer: This project is a collective effort by our team and has been done for capstone evaluation by great learning only. This is for strict internal use and confidentiality. This is not to be shared or used for external purposes. The images in front and last page are from online sources. We do not own the copyrights for them.

Table of Contents

1. Scope of the work / Objective of this project 3

1.1 Summary of Problem Statement 3

1.2 About Data and our findings on the data 3

1.3 Brief summarisation of our data findings and implications 3

2. EDA and Text Preprocessing analysis 5

2.1 Approach to EDA 5

2.2 Knowing more about Dataset: 5

2.3 Visualisation Insights (Bivariate analysis) and Statistical Analysis: 11

2.4 Statistical Tests 21

2.5 Conclusion of EDA: 24

2.6 Text Data Pre-processing: Description Variable or Field 24

2.7 Conclusion of Text preprocessing or analysis on description field: 40

3. Model decision analysis 41

3.1 Sampling 41

3.2 Machine Learning Classifier Models 41

3.3 Neural Network Classifiers 42

3.4 RNN and LSTM Classifiers 42

4. Model performance improvement - Insights 42

4.1 ML Classifiers model performance improvement 42

4.2 NN Classifiers model performance improvement 42

4.3 RNN Classifiers model and LSTM performance improvement 43

1. Scope of the work / Objective of this project

1.1 Summary of Problem Statement

The objective of this project is to design a ML/DL based chatbot utility for the Industrial safety domain/department of any industrial companies/organizations. There is an urgent need for industries/companies around the globe to understand why employees still suffer some injuries/accidents in plants. Sometimes, they even die in such an environment. The industrial safety department is an important function of any industries/companies with plants where the work environment can be dangerous, unsafe and prone to accidents such as fire, hazardous chemicals etc. They define policies and procedures to provide a safe and hazardous-free work environment for their employees and third-party contractors. This chatbot utility will be designed and developed with the intent of helping the professionals / employees/ third party-contractors to highlight the safety risk as per the incident description.

Chatbots gained popularity when NLP algorithms and models have improved over the past decade. One of the many applications of chatbot is to automate some of the functions of their call centre or L1 helpdesk teams. They are aimed at bringing in efficiencies and productivity improvements of a L1 helpdesk team where repetitive queries can be handled as part of first level touch point. Companies across the globe implement chatbot programs as part of their digital automation initiatives to have first level of resolution to their employees/contractors based on some of the common repeatedly occurring incidents/problems. For the majority of the incidents, they are more of a request or not knowing where the information is present. For such cases, Chatbot utilities are designed to give directions to the possible resolutions. For complex problems where human intervention will be required, they can be designed with an in-built call to action (CTA) function to direct the employees towards resolution.

In this problem statement case, we will design a chatbot utility to help the professionals highlight the safety risk based on some of the incident descriptions that the users provide. Based on the safety risk, they can potentially take the next course of action.

1.2 About Data and our findings on the data

The data set comes from one of the biggest industries in Brazil and in the world. This dataset basically records accidents/incidents from 12 different plants taken in three different countries over a period of 18 months between 2016 and 2017. Every data record is an accident or an incident occurred and reported and extracted in the form of a CSV file to us.

1.3 Brief summarisation of our data findings and implications

The various columns that are captured as part of the data set is as below:

- Data: timestamp or time/date information
    - This is a date field with a timestamp. Most of the timestamps recorded are from the database and the time of the incident is not recorded in this timestamp. It is more of the recorded timestamp
- Countries: which country the accident occurred (anonymised)

  - Since three countries' data are recorded, country_01 02 and 03 are the three values. It's more of a categorical data/
- Local: the city where the manufacturing plant is located (anonymised)
  - 12 locations or plants from which accident data is captured. Again, a categorical column
- Industry sector: which sector the plant belongs to
  - Industry sectors are dangerous and accident-prone ones like metals (metal processing or manufacturing ones) and mining where operations happen in dangerous and chemical hazardous mines. There is third value as others whichever is not fallen into metals and mining
- Accident level: from I to VI, it registers how severe was the accident (I means not severe but VI means very severe)
  - This is the original recorded accident level. One being "not severe" to 6 being the "highest severity" one. This is our target variable.
- 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)
  - What could have been the severity of the accident level. Again, one being "not severe" to 6 being the "highest severity" one. This can also be one of the target variables.
- 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. This will be used for NLP pre-processing along with other parameters which predicts the possible accident level and accident levels.

The potential target variable can be Accident level or Potential Accident level. Using our EDA and pre-processing analysis, we will ascertain which parameters will help in predicting these closely. Since this column will be having multiple values, we will be employing multi classification ML models, later use deep learning models like RNN and finally employ LSTM models.

In this document, the following sections include the Exploratory data analysis (EDA) and also NLP data analysis on the incident description columns. The EDA will help us find which parameters influence the target variable to a larger extent. We will also use visualisation techniques to provide better insights into the various factors or parameters. The NLP analysis will help us in predicting the possible accident level or safety level based on the user's incident description. We will also analyse which model will help us in most accurately predicting the target variables.

Note: At the submission stage of this interim report document, we have finalised only the possible models that we will employ. The real model building and running is yet to be done. The results of the same will be shared in the subsequent reports.

2. EDA and Text Preprocessing analysis

This section comprise of two major elements:

- EDA - Exploratory data analysis on categorical columns
    - In this, we will do univariate, bivariate and statistical significant hypothesis tests to determine which parameters drives our target variable prediction
- NLP text preprocessing on Description column
    - Here we will employ various NLP preprocessing models like n-gram, word cloud and find words that will help in predicting the target variable.

As earlier mentioned, we will have to employ a multi classification models for our prediction.

2.1 Approach to EDA

We will do the following

- Pre-Processing steps:
    - Getting to know the data set better. Assessing the data using various EDA techniques, to know about dataframes, the null values if any, removing duplicates have been carried out.
- Univariate analysis
- Bivariate analysis
- Statistical tests
- Concluding on EDA part

2.2 Knowing more about Dataset:

a) Shape of the dataframe:

The data frame consists of 425 rows and 11 columns

```
#   Column                       Non-Null Count Data type
---  ------                  --------------  -----
 0  Unnamed: 0                  425 non-null   int64
 1  Data                        425 non-null   object
 2  Countries                   425 non-null   object
 3  Local                       425 non-null   object
 4  Industry Sector             425 non-null   object
 5  Accident Level              425 non-null   object
 6  Potential Accident Level    425 non-null   object
 7  Genre                       425 non-null   object
 8  Employee or Third Party     425 non-null   object
```

```
9   Critical Risk                425 non-null    object
10  Description                  425 non-null    object
```

All the variables are of object or categorical type. Of them the "Data" variable is a datetime variable, yet put in place in the dataframe as an object variable. Description is a free flow text whereas the remaining attributes like country, local, critical risk, Gender, Employee type and Accident level (potential included) are all categorical in nature. We will first conduct EDA analysis on categorical columns followed by text preprocessing or analysis of Description column.

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

Note: **Data** variable seems to be an object, but in real sense it is a Date column

- The given dataset has 425 records and 11 variables.
- There are no null values.
- One variable "Unnamed" is a redundant variable.

There are errors in variable names: "Data" and "Genre", which needs correction. Genre is Gender column. Data is Date column. The column of Description contains textual data that needs to be analysed using NLP.

b) Dropping & renaming Data Variables:

From the dataset, we understand

'Unnamed' column needs to be dropped as this doesn't contribute to the analysis

Renaming variables:

- Data → Date
- Countries → Country
- Genre → Gender
- Employee or Third PArty → Employee Type

c) Duplicate Removal:

From the dataset of 425 rows and 11 columns, after carrying out the drop duplicate, we infer the clean dataset of 418 rows and 10 columns.

Note: Sl. No: 77, 262, 303,345,346,355 and 397 are duplicate fields

d) Check Missing Values

The dataset doesn't contain any missing Values

e) Secondary Variable creation:

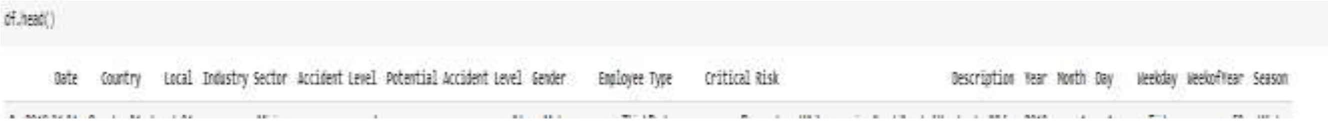
Since the Date column has a timestamp which is only updated as 00:00:00, we will create secondary variables to identify any specific time period that has an impact on the accidents that occurred. So in order to check that we will create the secondary variables and split a date into day, month, week and year. Accidents may increase or decrease throughout the year or month, so we need to add datetime features such as year, month and day.

Further, countries where the dataset was collected are all located in South America. So, in this analysis, let's assume the dataset was collected in Brazil.

Here Brazil has four climatological seasons as below.

- Spring : September to November
- Summer : December to February
- Fall: March to May
- Winter : June to August

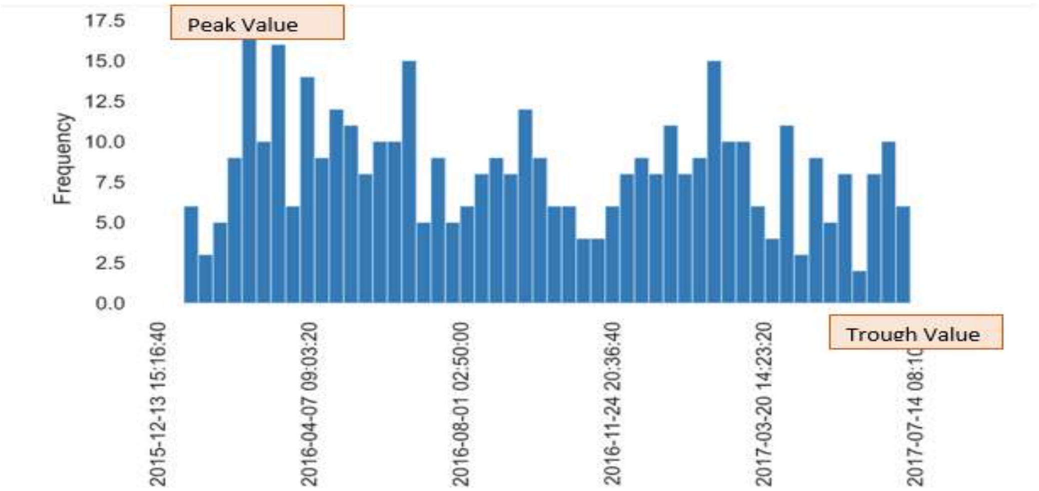
We can create seasonal variables based on month variables. After the creation of the secondary variables, the new data frame has this set of columns:



```
df.head()

     Date   Country   Local  Industry Sector  Accident Level  Potential Accident Level  Gender   Employee Type   Critical Risk        Description  Year  Month  Day   Weekday  WeekofYear  Season
```

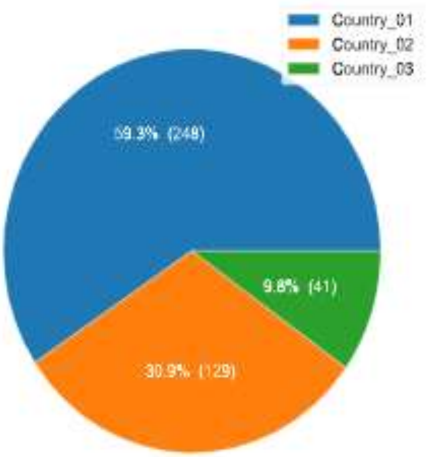
f) Basic Variable analysis

The below is a histogram of peak frequency of accidents date wise. We can see peaks on some days and troughs on some days. It is evident that they must have taken some actions which might have led to this. Having said that, let us analyse date wise here.



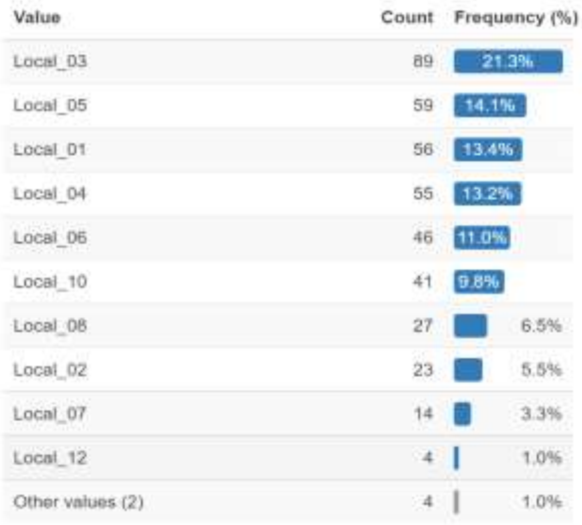
We cannot ascertain at this point if any actions were taken to lead that led to increased or decreased accident levels on that day. However, looking at other patterns, it is clear that date cannot be a factor in predicting the variable. But it is important for recording purpose.

Country wise data profiling:



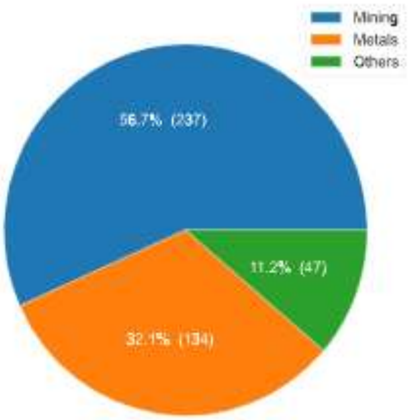
Country_01 has the maximum occurrence of accidents followed by Country_02 and Country_03.

Local variable: The following is the distribution of the local variable across 12 locations.



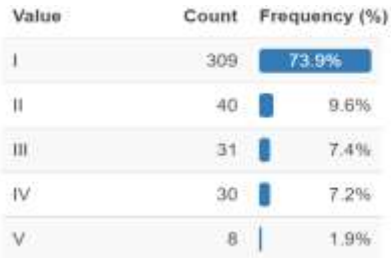
Value	Count	Frequency (%)
Local_03	89	21.3%
Local_05	59	14.1%
Local_01	56	13.4%
Local_04	55	13.2%
Local_06	46	11.0%
Local_10	41	9.8%
Local_08	27	6.5%
Local_02	23	5.5%
Local_07	14	3.3%
Local_12	4	1.0%
Other values (2)	4	1.0%

Industry sector wise:



It is clear that the maximum incidents occurred in the Mining industry followed by Metals and others. This clearly indicates Mining has more dangerous zones and accident-prone work environments. Metal Industry also has a significant number of 32% which indicates industrial plants will have exposure to unsafe practices.

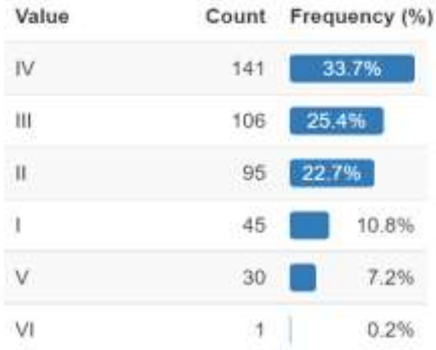
Reported Accident Levels (recorded accident levels):



Value	Count	Frequency (%)
I	309	73.9%
II	40	9.6%
III	31	7.4%
IV	30	7.2%
V	8	1.9%

Accident level I being the least severe and V being the most severe ones, it is clear from this distribution that Severity - I have been recorded the most. The less severe and least severe categories contribute to 80% of the accidents that happen. There seems to be a data imbalance here. But are they really the lowest severity ones? Let us analyse potential accident levels.

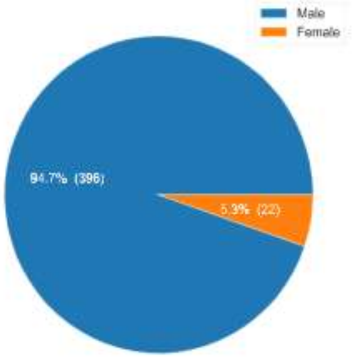
Potential Accident Levels:



Value	Count	Frequency (%)
IV	141	33.7%
III	106	25.4%
II	95	22.7%
I	45	10.8%
V	30	7.2%
VI	1	0.2%

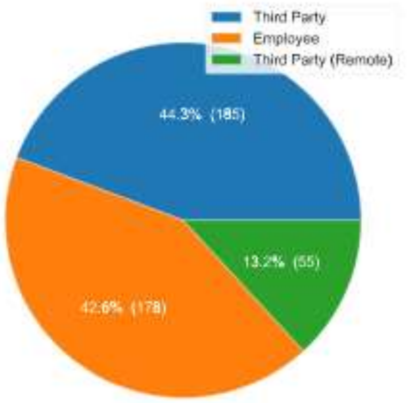
Based on the analysis, it was recorded later that what could have been the accident levels later. This shows a very different picture. If we count the high (III), very high (IV) and the most severe levels (V), it adds up to more than 65% of the cases. This indicates to us that this may be a better target variable than the accident levels as it was more skewed towards the lower end.

Gender:



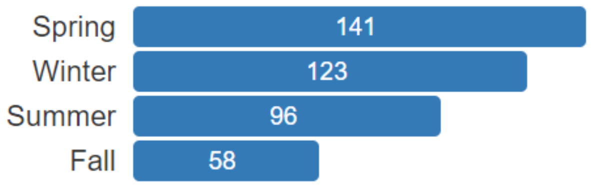
Gender clearly shows data imbalance. We will do statistical significance later to check if it really aids in our prediction.

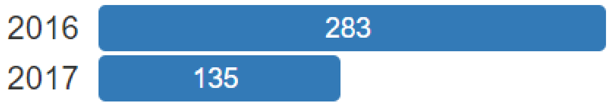
Employee Type:

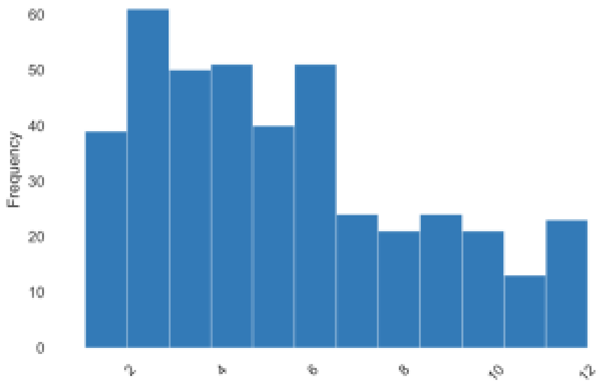


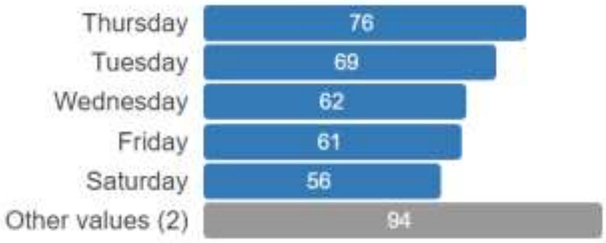
Third parties and employees have had equal frequency of accidents. But one important observation is that remote locations contributed to ~13% of incidents. This clearly indicates that some mining sites have accident prone work areas.

Date - Day, week, month, year and seasons:



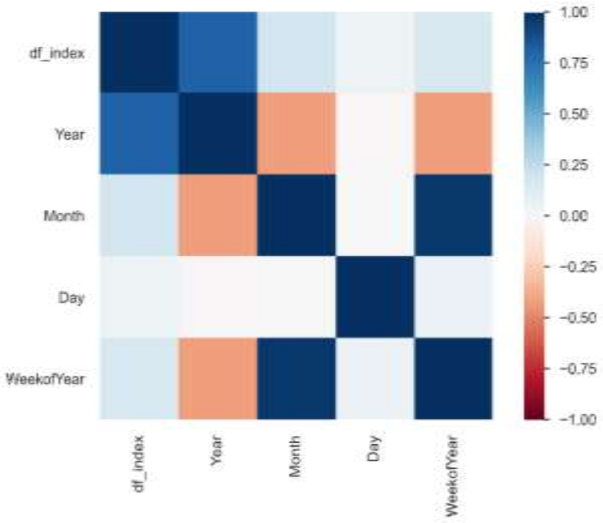






Between 2016 and 2017, the recorded data is for the full year of 2016 whereas 6 months for 2017. So, most of the individual analysis will always have this imbalance. If any seasonality affects the accidents, we can check in statistical tests.

We did correlations for various date columns and checked. It was a weak correlation.

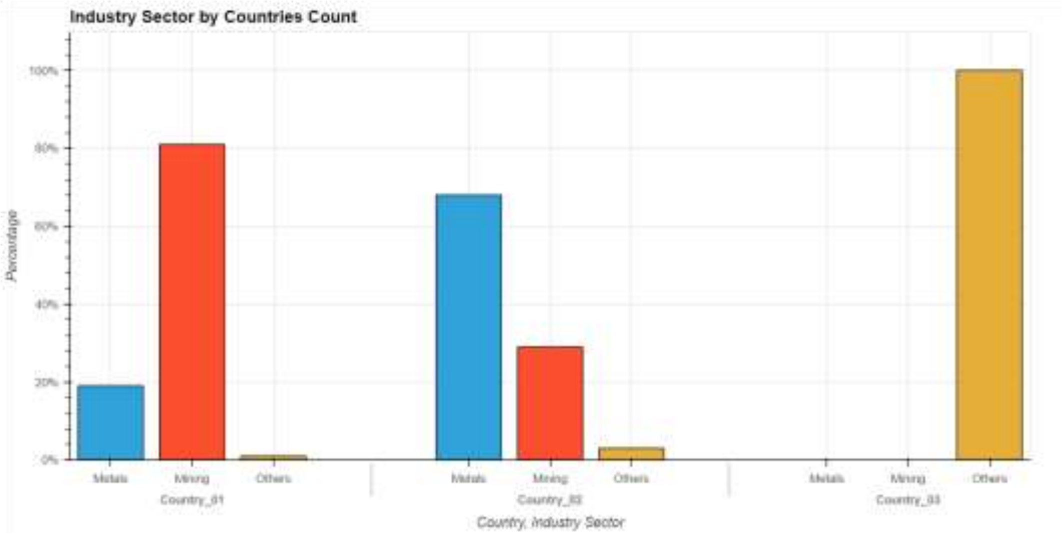


2.3 Visualisation Insights (Bivariate analysis) and Statistical Analysis:

2.3.1 Checking other bivariate relationships

- Country_01 : Mining
- Country_02 : Metals
- Country_03 : Others

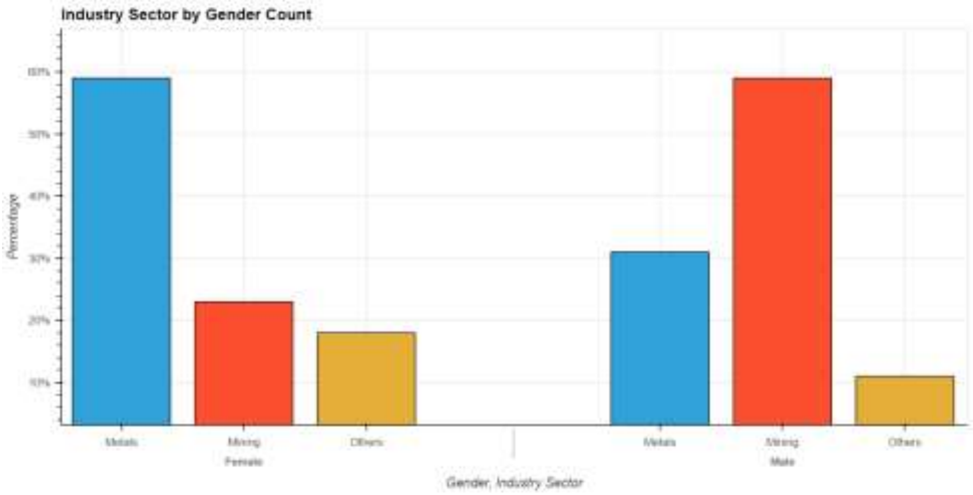
The below is the distribution of industry segments for each of the country data.
Mining is highly present in Country 01 for which accident levels are recorded.



Let’s try slicing the data Gender-wise with respect to the Industry Sector.

We observe:

- Mining activity is majorly carried out by Male and Metals by females. In other words, Mining environments are largely male dominated because of their remote work areas. The Metal industry is plant or factory set up; hence the presence of females is fairly higher there. It is just a possible conjecture but no significant evidences seen or observed through data.

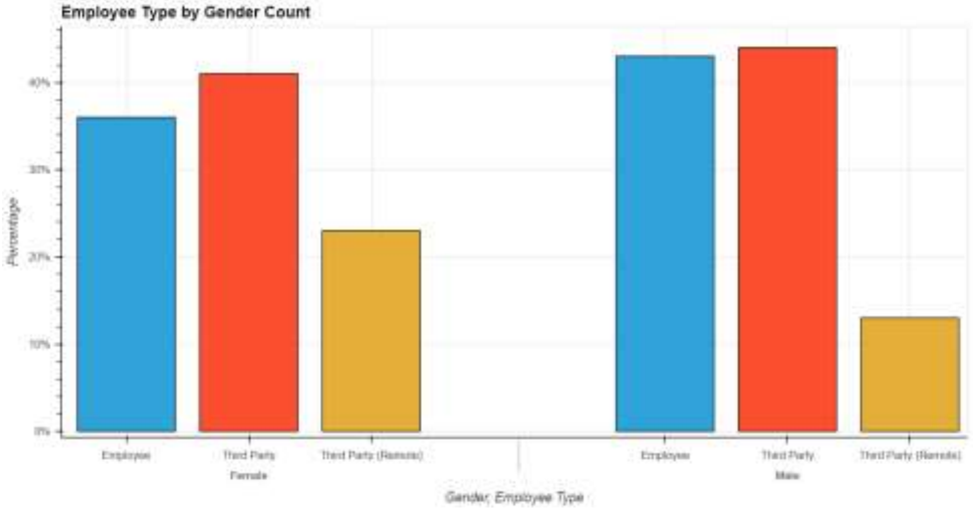


Let's further slice Employee Type with respect to Gender:

We observe:

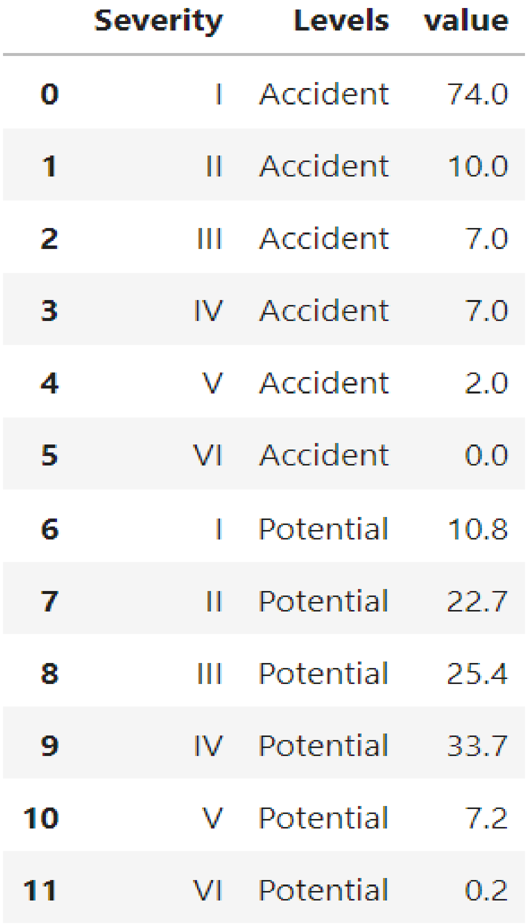
- There is no significant difference in ratio of employee types to gender
- The proportion of females with Third Party (Remote) is slightly higher than that of males.

- It could be a conjecture that companies tend to employ female contractors in difficult or dangerous work environment than using their own female employees

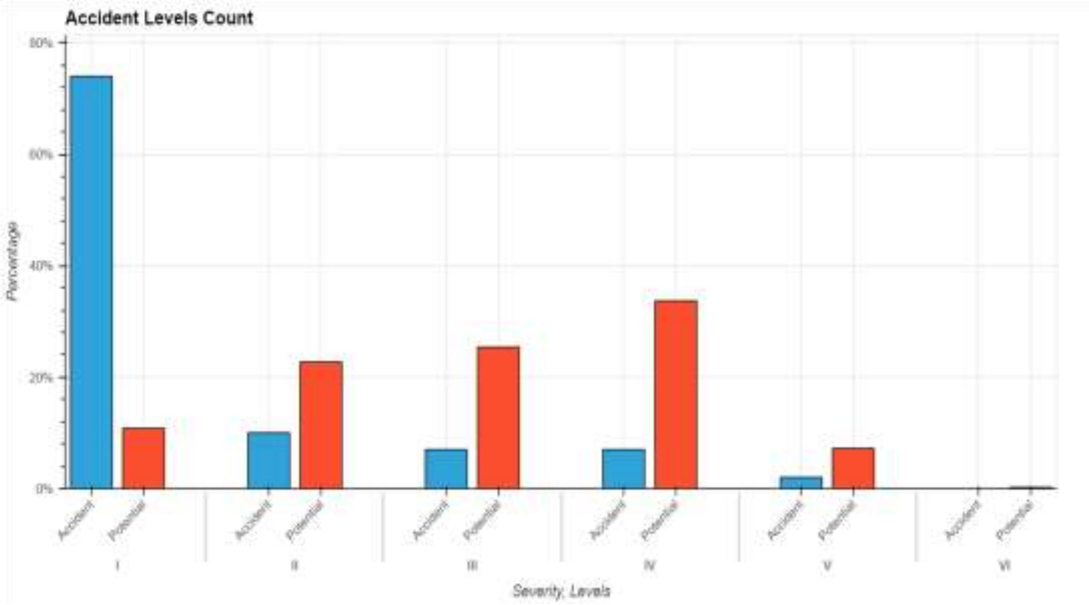


2.3.2 Exploring association with Potential Accident Level and Accident Level

In the last section, we saw some contrasting finds on accident level and potential ones.



	Severity	Levels	value
0	I	Accident	74.0
1	II	Accident	10.0
2	III	Accident	7.0
3	IV	Accident	7.0
4	V	Accident	2.0
5	VI	Accident	0.0
6	I	Potential	10.8
7	II	Potential	22.7
8	III	Potential	25.4
9	IV	Potential	33.7
10	V	Potential	7.2
11	VI	Potential	0.2



Assessing the trend between Potential accident level and Accident level, we can infer

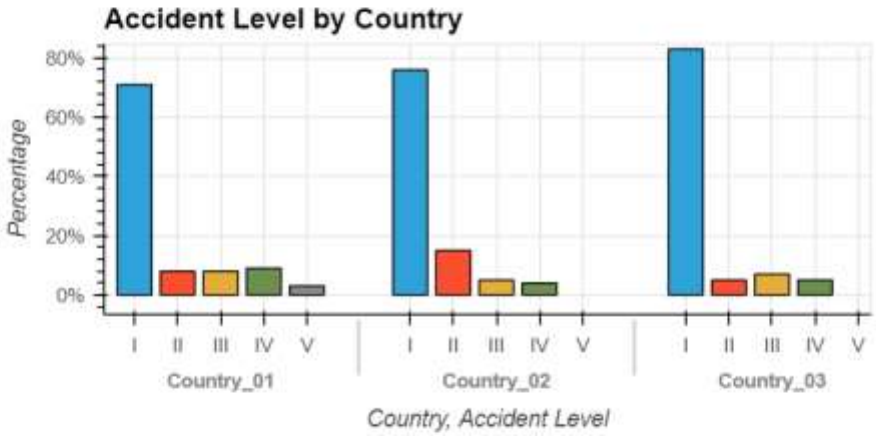
- The accident levels being recorded as “Low” may in turn be severe ones. That is an important observation for us.

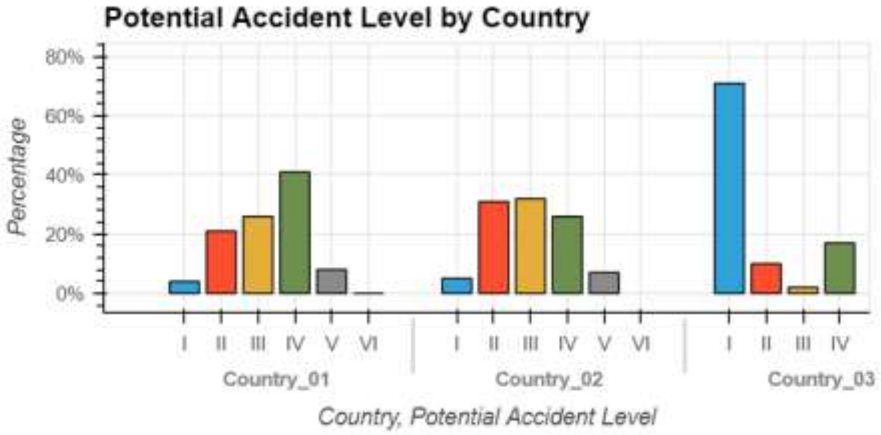
- This may be due to “Potential Accident Level” being overlooked and potentially high-risk accidents are possible.

The other way of looking at this data point is: The accident level is recorded as per the incident reported. Whereas Potential accident level is what might have been the real severity. The below data point shows that for most of the low severe accident levels recorded, the potential accident levels might have been higher. This is true for all accident levels. That means, potential accident level is a much better variable to be predicted. This can be our target variable.

We will choose *Potential Accident Level* as our *Target variable.* *We are looking at a Multi classification problem here, since Potential Accident level variable has more than 2 values.*

Exploring the association of all the variables with regards to Accident level and Potential Accident Levels

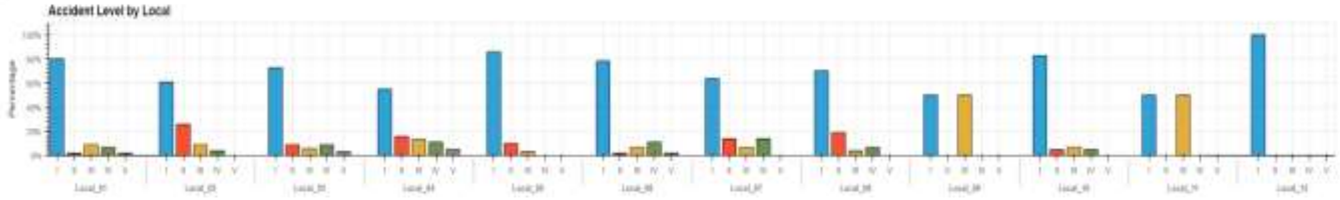


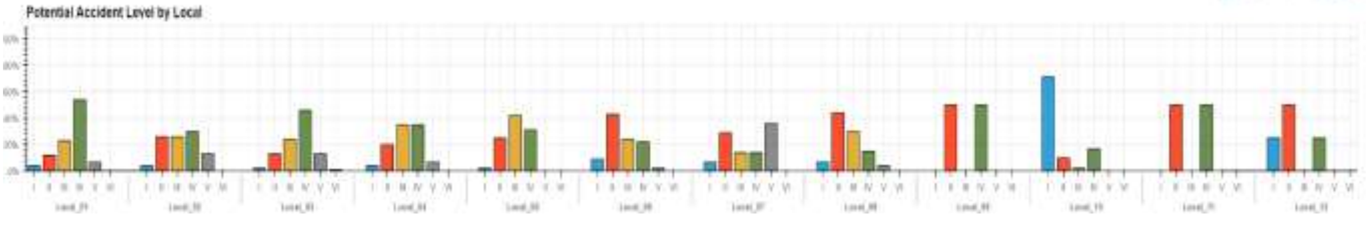


Inferences:

- All countries have significantly higher Severity - I (Lowest severe) counts, whereas if you see potential accident levels, it is leaning more towards Severity Level III and IV.
- This again validates our observation that the potential accident level is a much more valid target.

<u>Locals and Accident levels</u>



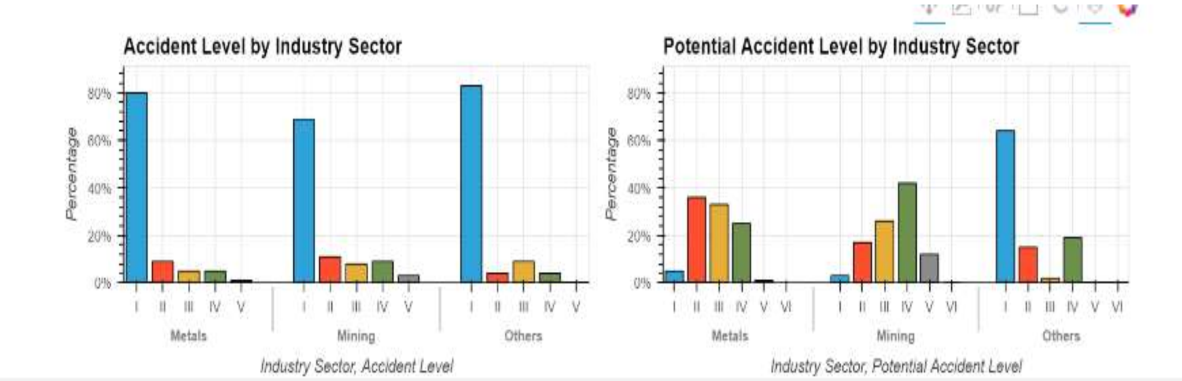


Inferences:

- Similar inferences as above can be observed and derived.

- The local_09 and local_11 has an equal number of recorded III incidents. This is also changed in potential accident levels to IV.
- Local_10 seems to be incurring the maximum least severe incidents. Are they imparting better practices or are their training programs better or are their work conditions safe? We do not have that much data to validate or observe such outlier behaviour.

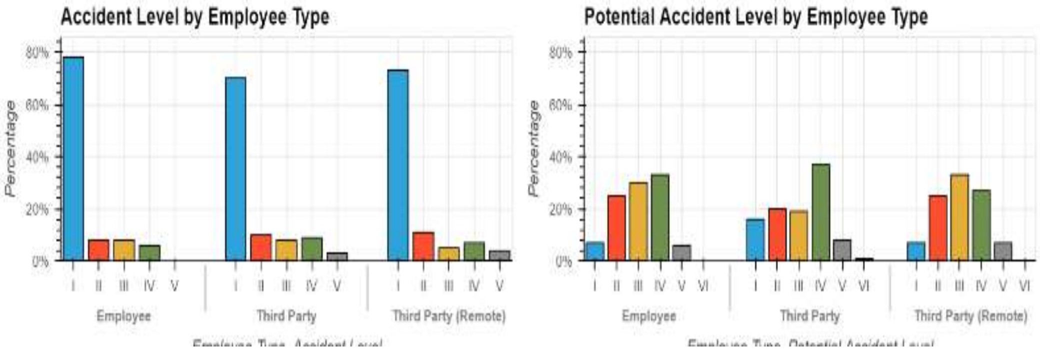
Industry Sector



Inferences:

- The potential Accident levels III and IV are higher in the Mining area whereas in Metals, Levels II and III are higher.
- Other Industries do not encounter many higher severe accidents than mining and metals. Mostly (60%) it is all Least severity ones.

Employee Type:



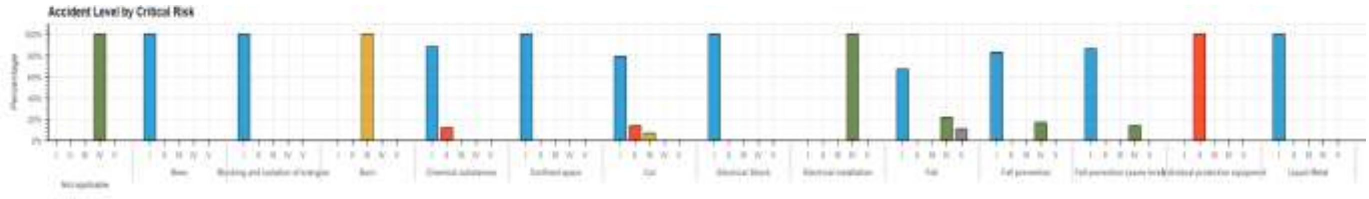
Inferences:

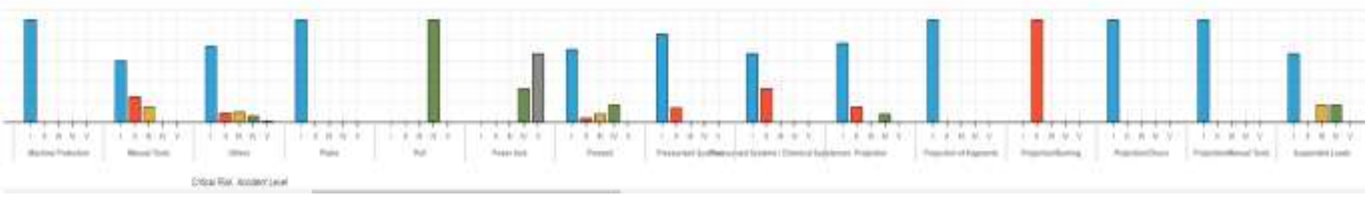
- Potential Accident levels are almost equally distributed for severity level II, III and IV for both third party contractors and employees. The remote third party also encounters more higher severity incidents.
- One thing can be inferred is employee type is important to be recorded but incidents occur across various types. There aren't specific observations

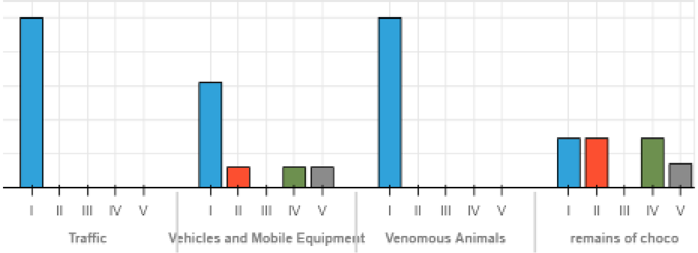
which lead us to believe employees have a lower rate of severe incidents than third parties or vice versa.

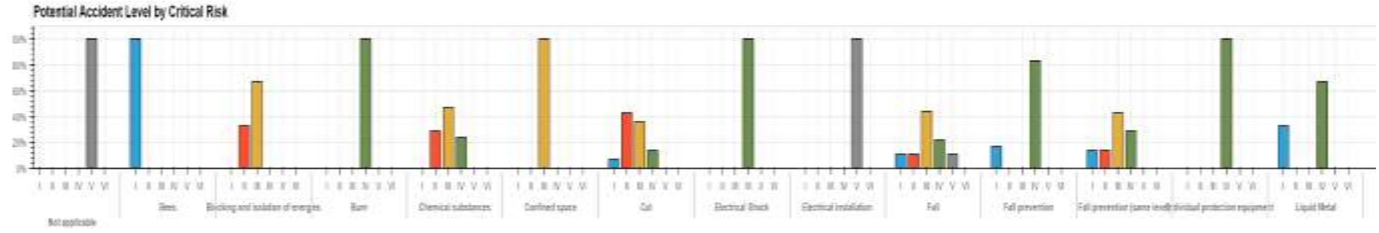
- This also may indicate a possible conjecture that the training program imparted to any employee is almost the same. But focus should be given to improving the training program for any employee on how to work safely in such a dangerous environment.

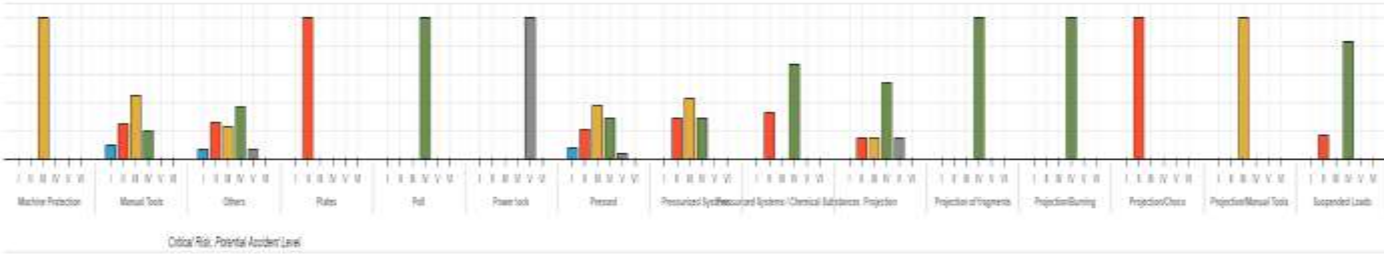
Critical risk factors

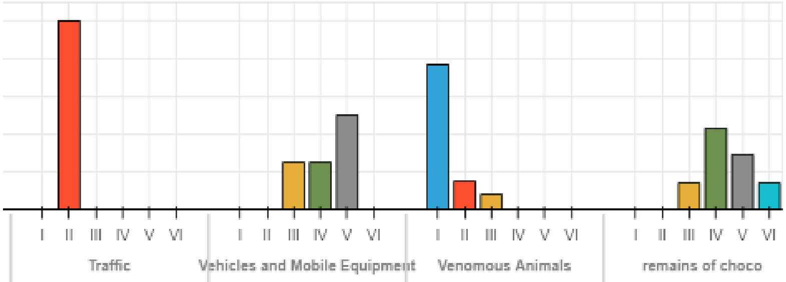












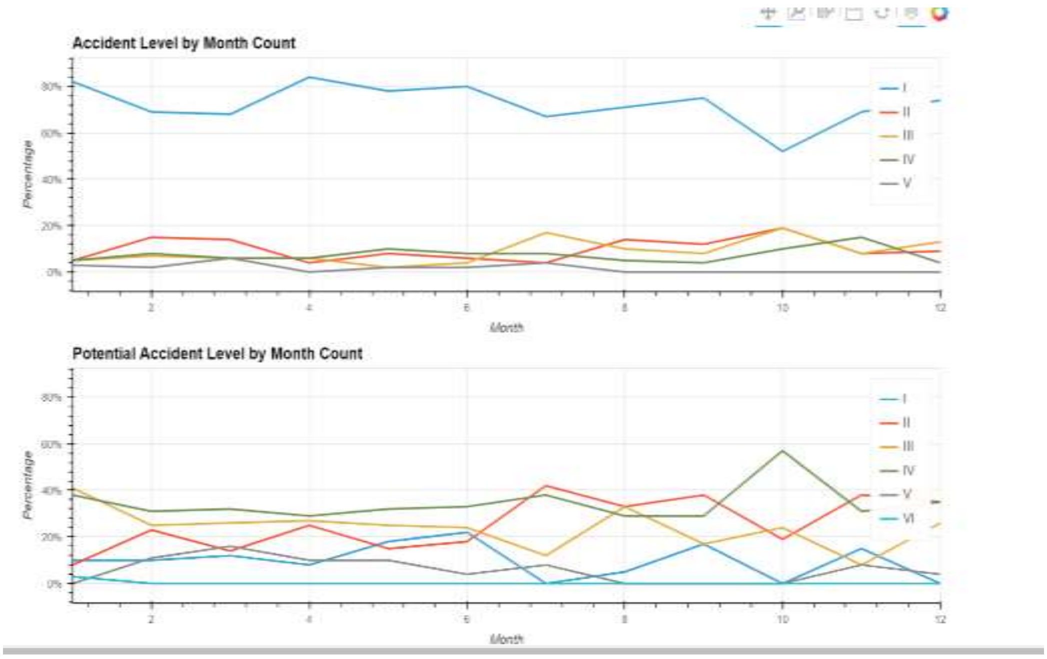
Inferences:

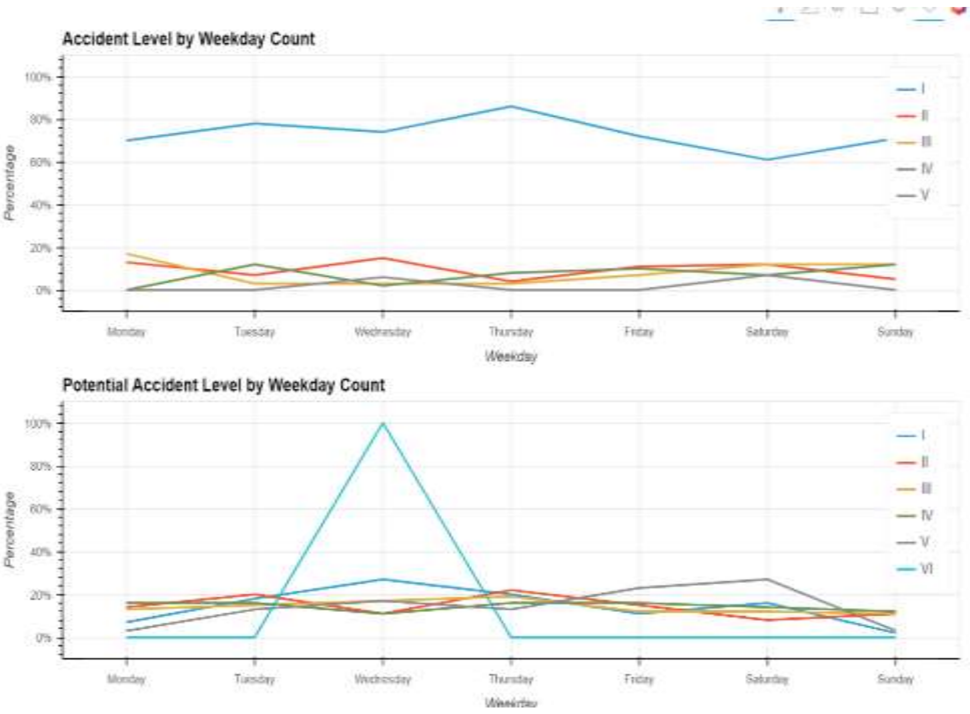
- Many not applicable categories have had potentially severe accidents than what was recorded. This may be derived from text processing which we will analyse in the below sections.
- Bees attack typically suggest its a lower grade severity
- Burns category - can be potentially higher severe ones than what was recorded. The severity of burn degree may be derived from text analysis.
- Some Chemical Substances can lead to higher severe accidents in potential accident levels.
- Electrical shock - higher severe incidents. The recorded severity levels are lowest. It certainly requires an update to SOP or better training programs.
- Confined Spaces, Cut or a Fall category is generally recorded as low severe apart from few serious cases whereas in reality, this can be a potentially higher severe one
- The Liquid Metal category has potentially higher severe category ones than the recorded ones.
- *The Poll category has been recorded and observed in the right way. Is it that the systems or SOPs are better maintained for this type of incident?*
- *Powerlock category is typically observed as potentially the highest severe* incident type. Even some of the recorded observations validate that. As in the above point, *are the systems or SOPs better maintained for this type of incident?*
- Electrical installation - again a higher severe category was observed than what was recorded. Even in recorded cases, it was all severity - 4 ones. Maybe a little better training or updating SOP may rectify this error. It is a possible conjecture.
- Venomous animals can have slightly higher severe categories. This need to be observed from text processing
- Suspended loads also are treated as higher severe incidents than what was recorded. Again few changes to SOP and text analysis can lead us to right accident level determination.

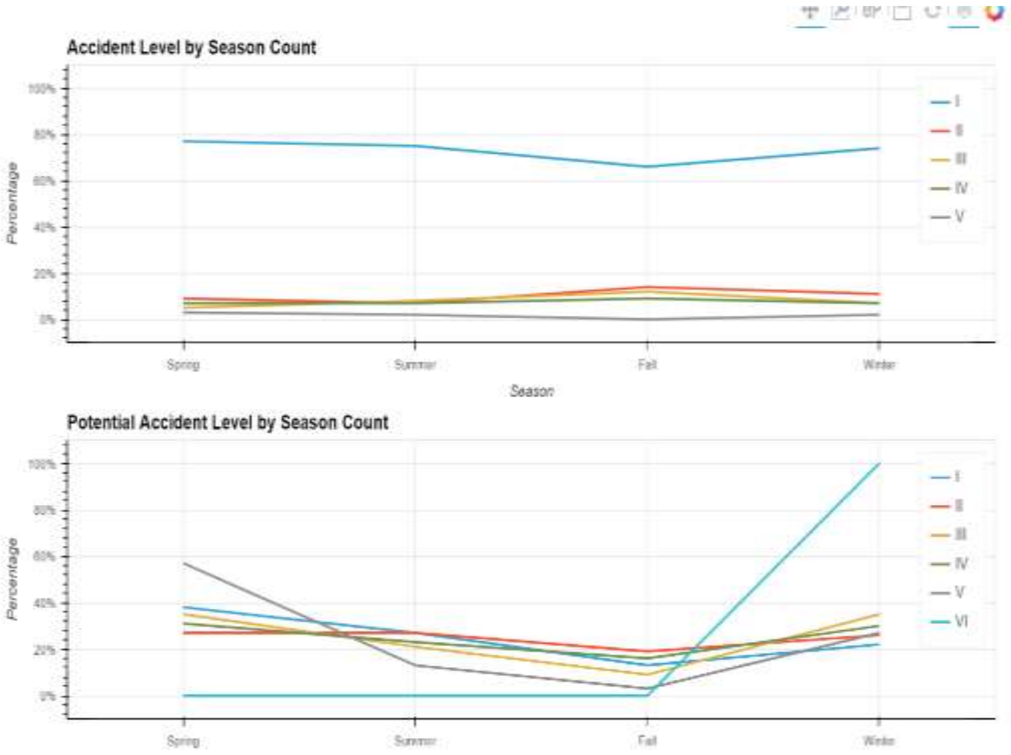
This critical risk factor parameter is really important. It tells some important observations with regards to either limitations in training program/ SOPs or the analysis at first level recording information was poor.

This can be a potential case to be automated using text analysis and running through our models to have better prediction of accident levels. Once the accident levels are identified closer to the right level, the treatment towards it can be addressed. Lot of data analysis to such levels can potentially save the people working in such an environment.

Date columns







Inferences:

- There are no observable patterns with regards to month or week day or seasons. Only potential accident level VI shows some higher variations in the winter season. But only one record is there for potential accident level VI, so we need not worry too much on these parameters.
- As earlier said, the date column is important for recording purposes but may not have a significant impact on accident levels or incidents occurring. We are not seeing a pattern here to have deeper analysis. Nor if any higher or lower incident pattern seen can be validated or analysed through other data points.
- One study that can emerge from this analysis of date columns: We can go back and see the logs in specific locals or countries and record if they did something different that particular day, week or month. But that is not the scope of this document or project.

Conclusion of Bivariate analysis:

- Association of various parameters like Date, Country, Locations, Employee type, Gender, Critical Risk factor on both accident and potential accident levels throw some important observations which we have highlighted in this section.
- Are these mere parameters or are they significantly influence or impact accident levels will be addressed through statistical tests in the next section?

2.4 Statistical Tests

Since all are categorical variables, we will use the chi-square test for finding the statistical significance of each of them on the potential Accident levels. We will try to formulate null and alternate hypothesis and run statistical tests using chi-square and determine whether to reject or accept null hypotheses.

Potential Accident Level & Countries:

- Ho = The proportions of **Potential Accident Level** is not different in different **countries**
- Ha = The proportions of **Potential Accident Level** is different in different **countries**
- Decide the significance level: alpha = 0.05

We observe:

It is observed that the p-value is 4.1017395785840915e-34

At 0.05 significance, as p-value<level of Significance, we reject the null hypothesis.

i.e., The proportions of Potential Accident Level is different in different countries.

we also have observed that Higher potential accident level is occurring more in Country_01. Statistically significant difference between countries.

Our Conclusion: This shows that countries are an important parameter for our potential accident level prediction. This will be a part of our analysis.

Potential Accident Level & City:

- Ho = The proportions of **Potential Accident Level** is not different in different **City**
- Ha = The proportions of **Potential Accident Level** is different in different **City**
- Decide the significance level: alpha = 0.05

We observe:

It is observed that the p-value is 3.728976101772464e-28

At 0.05 significance, as p-value<level of Significance, we reject the null hypothesis.

i.e., The proportions of Potential Accident Level is different in different City

Potential accident level (PAL) is different from location to location.

- PAL I is unusually higher in L10
- PAL II is high L3, L4, L5, L6, L8
- PAL III, IV are high in L1, L3, L4, L5
- PAL V is unusually high in L7

Our Conclusion: City or locations is an important feature that will be required for analysis.

Industry Sector & Potential Accident Level:

- Ho = The proportions of **Potential Accident Level** is not different in different **Industry Sector**
- Ha = The proportions of **Potential Accident Level** is different in different **Industry Sector**
- Decide the significance level: alpha = 0.05

We observe:

It is observed that the p-value is 1.992985715278311e-36

At 0.05 significance, as p-value<level of Significance, we reject the null hypothesis.

i.e., The proportions of Potential Accident Level is different in different Industry Sector

Higher potential accident levels are observed in mining, intermediate in metals sector compared to other sectors. Statistically significant difference between industry sector

Conclusion: Industry sector is an important feature to be considered for analysis.

Gender & Potential Accident Level:

- Ho = The proportions of **Potential Accident Level** is not different in different **Gender**
- Ha = The proportions of **Potential Accident Level** is different in different **Gender**
- Decide the significance level: alpha = 0.05

It is observed that the p-value is 0.0003681748355215793

At 0.05 significance, as p-value<level of Significance, we reject the null hypothesis.

i.e., The proportions of **Potential Accident Level** is different in different **Gender**

Conclusion: Gender will still be an important parameter despite the data imbalance for each type.

Critical Risk & Potential Accident Level:

- Ho = The proportions of **Potential Accident Level** is not different in different **Critical Risk**
- Ha = The proportions of **Potential Accident Level** is different in different **Critical Risk**
- Decide the significance level: alpha = 0.05

We observe:

It is observed that the p-value is 6.463580982130032e-22

At 0.05 significance, as p-value<level of Significance, we reject the null hypothesis.

i.e., The proportions of **Potential Accident Level** is different in different **Critical Risk area**

Conclusion: Critical Risk area is also important parameter for analysis

Weekday & Potential Accident Level:

- Ho = The proportions of **Potential Accident Level** is not different in different **Weekday**
- Ha = The proportions of **Potential Accident Level** is different in different **Weekday**
- Decide the significance level: alpha = 0.05

We observe:

It is observed that the p-value is 0.30204663583713837

At 0.05 significance, as p-value>level of Significance, we accept the null hypothesis.

i.e., The proportions of **Potential Accident Level** is not different in different **Weekday**

All Potential Accident levels are typically low on Sundays, yet the difference is not statistically significant

Conclusion: While week day or date is not important and since dates recorded for 2016 for a full year whereas only 6 months data recorded for 2017 . Week day or Date may not be that significant parameter for analysis. Date is needed for recording the accident information. It does not have much significance beyond that. Nor any pattern seen towards specific days or time or seasonal variation is affecting the accident.

Month & Potential Accident Level:

- Ho = The proportions of **Potential Accident Level** is not different in different **Month**
- Ha = The proportions of **Potential Accident Level** is different in different **Month**
- Decide the significance level: alpha = 0.05

We observe:

It is observed that the p-value is 0.30204663583713837

At 0.05 significance, as p-value>level of Significance, we accept the null hypothesis.

i.e. The proportions of **Potential Accident Level** is not different in different **Month**

Like above point, not specific analysis or statistical significance seen for Month column. But it is required to be captured as part of date column.

2.5 Conclusion of EDA:

- EDA analysis has thrown some important decisions on our parameters along with potential accident levels. This will help us in building our chatbot utility to see which parameter will be significant and should be asked.
- We concluded that Potential accident level is our target variable.
- Like we said above, “Date” does not add much value but can be used for only recording purposes.
- Gender can be important parameter although it has data imbalances.
- Higher potential accident level is occurring more in Country_01. Statistically significant difference between countries.
- Statistically significant difference between various locations or locals.
- Higher potential accident level is more in mining, intermediate in metals sector compared to other sectors. Statistically significant difference between industry sectors.
- All potential accident levels were more in males. Statistically significant difference between gender.
- Having observed many of these and while these parameters are important, these alone cannot predict the potential accident level. They are still categorical and there is no sure way of using such categorical variables to predict the accident levels.
- Hence, we require text processing or analysis on the description to enhance our prediction of potential accident levels. This will be taken in the next section.

- **Why and how EDA is important for our Chatbot utility: Chatbot will be the final output that would be visible to end users. All these statistical analysis shows an important aspect for our Chatbot utility. This will aid in designing our chatbot questions or GUI questions to take input on all these parameters that can potentially impact or predict the potential accident levels.**

2.6 Text Data Pre-processing: Description Variable or Field

This section will detail the entire text preprocessing of the “Description” field or variable. There are a lot of insights that can be derived from text analysis. Our conjecture is reading through detailed descriptions of the accidents might have enabled the data collection team to add a potential accident level column manually later. Hence, it is all the more important and essential to analyse the “description” column to infer more about the accident level and potential accident levels.

NLP preprocessing steps are taken sequentially before applying the model on the data. The sequential processes are:

- Remove stop words
- Tokenize the words
  - Convert all words to lowercase ones,avoid any capital cases
  - Converting apostrophe to the standard lexicons
  - Removing punctuations

- Lemmatization & Stemming

    - stemming technique only looks at the form of the word whereas lemmatization technique looks at the meaning of the word. It means after applying lemmatization, we will always get a valid word
- Removing non-alphabetical characters like '(', '.' or '!'
    - We have also observed many punctuations, alphanumeric characters and special characters. These add no significance to prediction. Hence, removing them will help

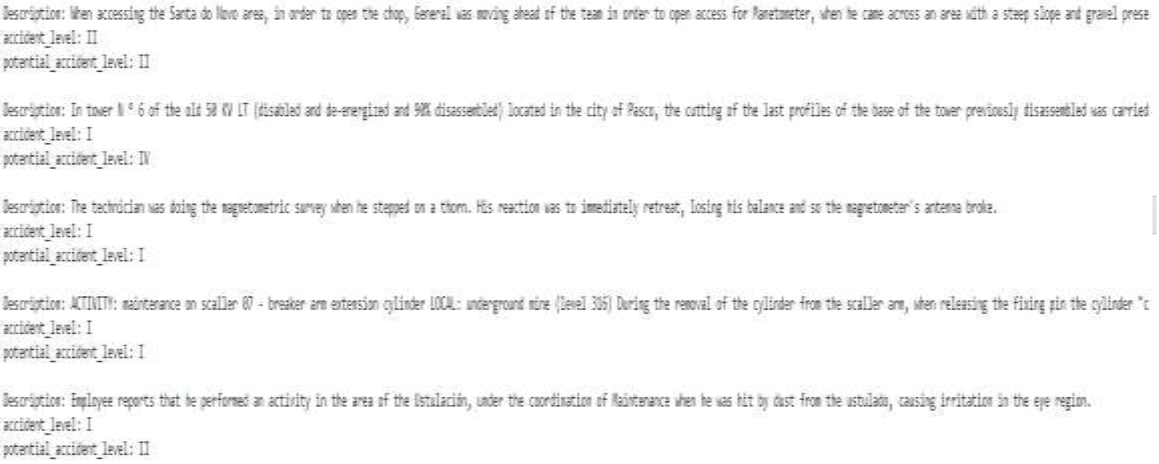
The first step in any text processing is removing of stop words. These stop words are commonly found in any vocabulary and do not amount to any prediction of target variable. Our target variable is potential accident level.

We removed all these stop words from the description column.

{"didn't", 'more', 'she', 'here', 'if', 'he', 'we', "aren't", 'a', "hasn't", "she's", "they'd", "you're", 'its', 'further', "he'd", 'which', 'kg', 'an', 'had', "don't", 'ever', 'too', 'after', 'until', 'herself', 'why', 'yours', 'this', 'of', "he'll", "she'll", "we've", "it's", 'has', 'and', 'does', 'his', 'been', 'me', "won't", 'since', 'am', 'during', "wouldn't", "i'm", 'with', 'himself', 'i', 'but', 'r', 'few', 'just', 'out', 'the', 'by', "that's", 'should', 'in', 'to', 'at', "i'll", 'down', 'once', 'my', 'else', "we're", 'your', 'up', 'did', 'k', 'yourselves', 'same', 'each', 'most', 'nor', 'only', "they'll", "i've", "they've", 'ourselves', 'both', "haven't", 'there', "you'll", 'be', 'like', 'ours', 'under', "how's", 'having', 'is', 'over', "here's", "shan't", 'these', "i'd", 'where', 'him', 'were', "shouldn't", 'into', 'being', "mustn't", 'below', 'as', 'they', 'was', 'all', 'otherwise', 'while', "where's", 'because', "let's", 'those', 'do', "he's", 'off', 'own', 'not', 'whom', 'also', 'however', 'who', 'pm', 'so', "when's", 'yourself', "couldn't", 'get', 'cannot', 'when', 'again', 'http', 'for', 'from', 'how', "we'd", 'then', 'doing', 'about', 'cm', "wasn't", 'than', 'myself', 'other', "what's", "who's", 'above', 'it', 'can', 'itself', "why's", 'on', 'com', "isn't", 'very', 'ought', 'through', 'some', 'would', "she'd", "hadn't", 'shall', 'no', "there's", 'against', "can't", 'their', 'you', 'our', 'such', 'are', 'them', 'themselves', "you've", 'could', "they're", 'before', 'any', 'her', 'www', "weren't", "doesn't", "we'll", 'that', 'what', "you'd", 'between', 'or', 'have', 'hers', 'theirs'}

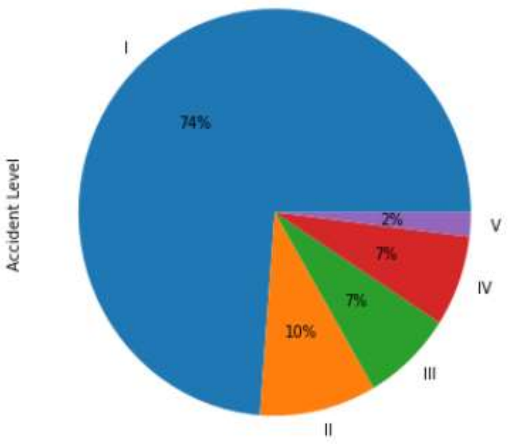
2.6.1 Observations basis the analysis of the error description and distribution of accident level

1. Let us analyse a random sample of 5 accident descriptions where the length of the accident description is greater than 100.

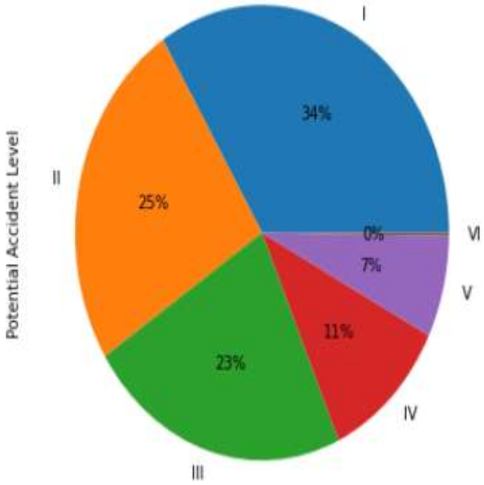


```
Description: When accessing the Santa do Novo area, in order to open the chop, General was moving ahead of the team in order to open access for Banetometer, when he came across an area with a steep slope and gravel prese
accident_level: II
potential_accident_level: II

Description: In tower N ° 6 of the old 50 KV LT (disabled and de-energized and 90% disassembled) located in the city of Pasco, the cutting of the last profiles of the base of the tower previously disassembled was carried
accident_level: I
potential_accident_level: IV

Description: The technician was doing the magnetometric survey when he stepped on a thorn. His reaction was to immediately retreat, losing his balance and so the magnetometer's antenna broke.
accident_level: I
potential_accident_level: I

Description: ACTIVITY: maintenance on scaller 07 - breaker arm extension cylinder LOCAL: underground mine (level 305) During the removal of the cylinder from the scaller arm, when releasing the fixing pin the cylinder "c
accident_level: I
potential_accident_level: I

Description: Employee reports that he performed an activity in the area of the Ustulación, under the coordination of Maintenance when he was hit by dust from the ustulado, causing irritation in the eye region.
accident_level: I
potential_accident_level: II
```

```
----------------------------------------------------------------------
Distributon of accident_level where the length of Description is > 100
----------------------------------------------------------------------
```



```
---------------------------------------------------------------------------
Distributon of potential accident_level where the length of Description is > 10(
---------------------------------------------------------------------------
```



Inferences:

- If you see some description texts, they are referring to some machine models with numerical digits. In singularity, the numerical or model number may not be useful for prediction. However, the entire machine model along with Machine name may be useful for prediction. We will use n-gram and other models below in coming sections.
- A random analysis on length of description column over 100 words, shows some 74% cases are recorded as accident level - I.
- The same is reduced for potential accident level reduced to 34%.
- Although this has no direct correlation, what one can see is whether a deeper description over many words will yield the results differently. It is not seen or observed. We will head to other analysis

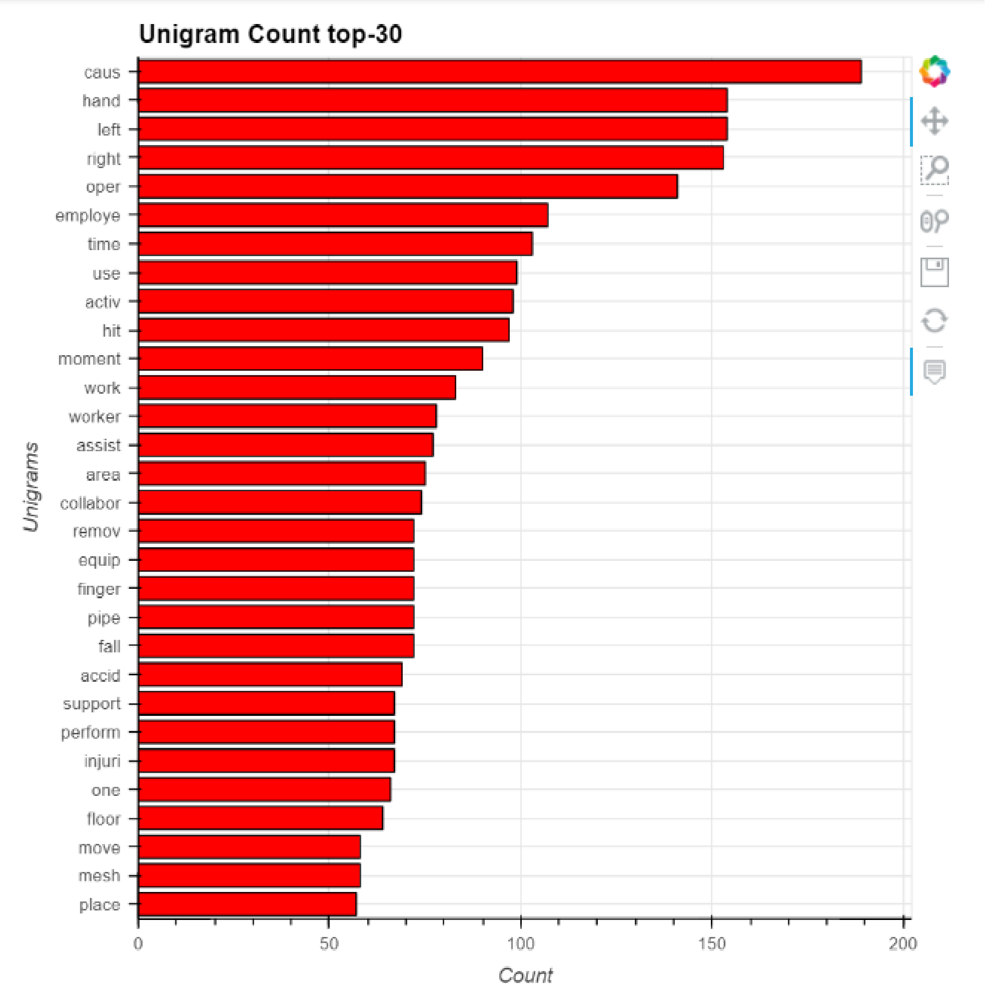
2.6.2 Text Pre-processing and Exploration

N-Gram Analysis:

Description of accidents is important to understand the cause of accidents, so we need to discover characteristic words or phrases indicating the situation when accidents occurred.

N-grams is a very important tool in NLP exploration for finding sets of common co-occurring words, where *n* refers to the number of consecutives occurring words. For instance, *'left', 'hand', 'finger'* would be a 3-word n-gram or trigram.

Using the Unigram:

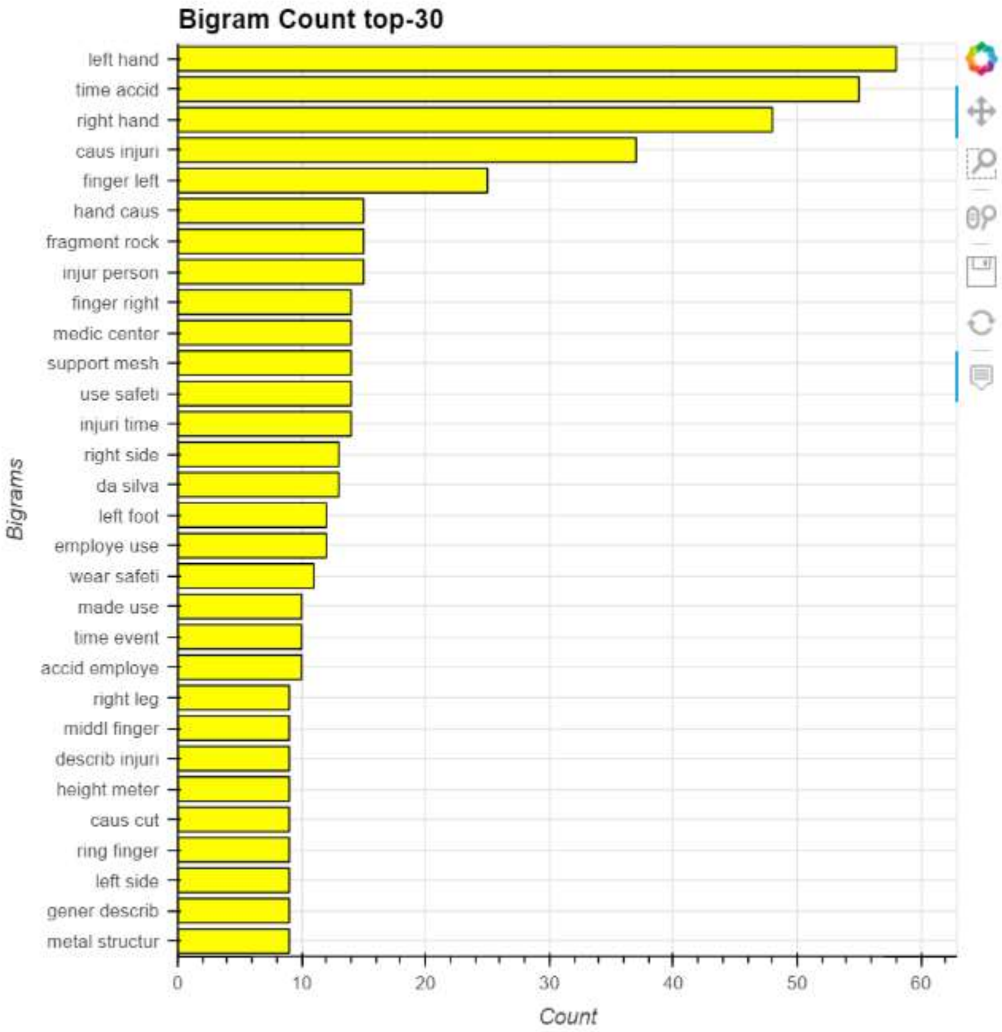


Inferences:

- We observe that top 5 words are <u>'cause', 'hand', 'left', 'right', 'oper'.</u>
  - A hand in itself may not be a great indicator or predictor. So we may require a bigram model to check if the right hand or left hand has any significance.
- There are several words related to hands. For example, <u>left, hand, right and finger</u>. How these all help us in predicting. So should we drop this? Such an analysis will be done using n-gram models.

- Moreover, there are several words related to movement of something. For example, hit, remove, fall and move.
- Unigram model has so many top words but they are all very generic and in singularity, may not be of much use. Let us use bigram model and see if any better words can be used for our analysis.

Using the Bigram:

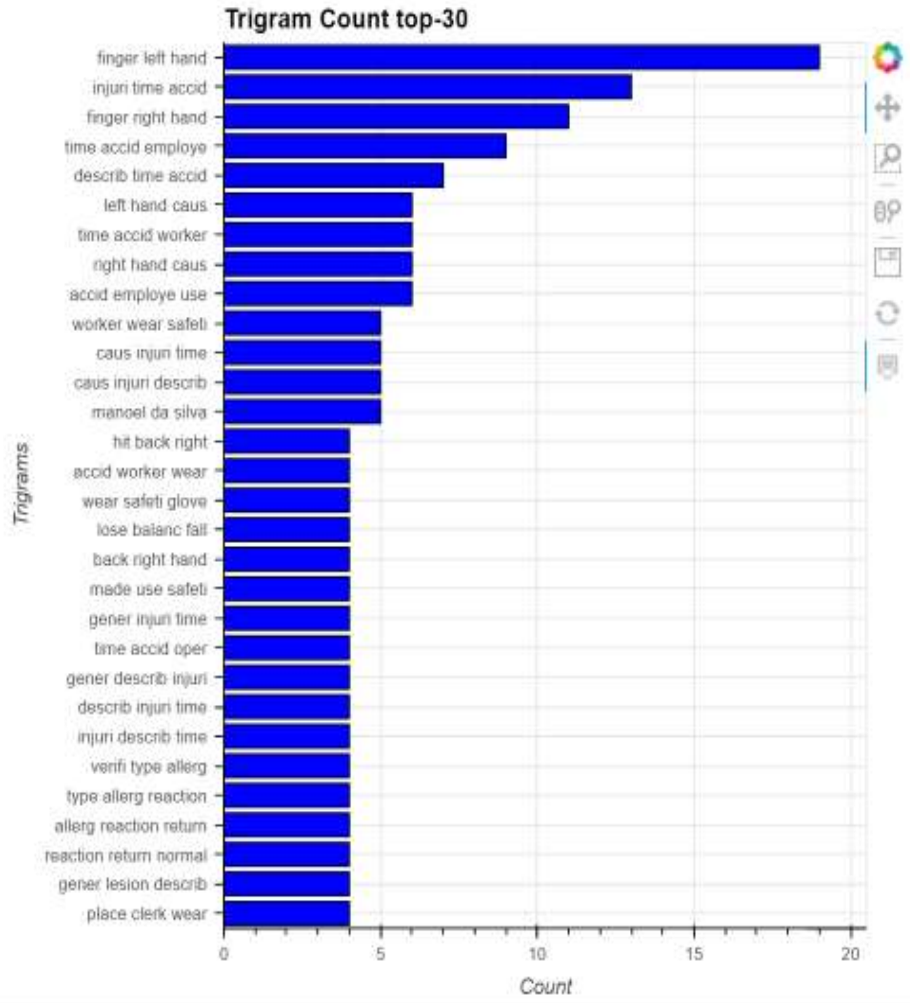


Inferences:

- There are so many phrases which are related to hands. For example: left hand, right hand, finger left, finger right, middle finger and ring finger.
- There are also some phrases which are related to other body parts. For example: left foot and right leg, right side
- How is this relevant for prediction apart from being a body part. In itself, the body parts only depict the area where injury occured but it tells nothing about severity levels. Having said, these n-gram models help us in visualising how many times a particular body part is affected across. Here across 400 odd accidents reported, the left

hand is the major one that got injured the most with 70 times. That is useful for more numerical analysis but not significant in predicting the severity levels.

Using the Trigram:

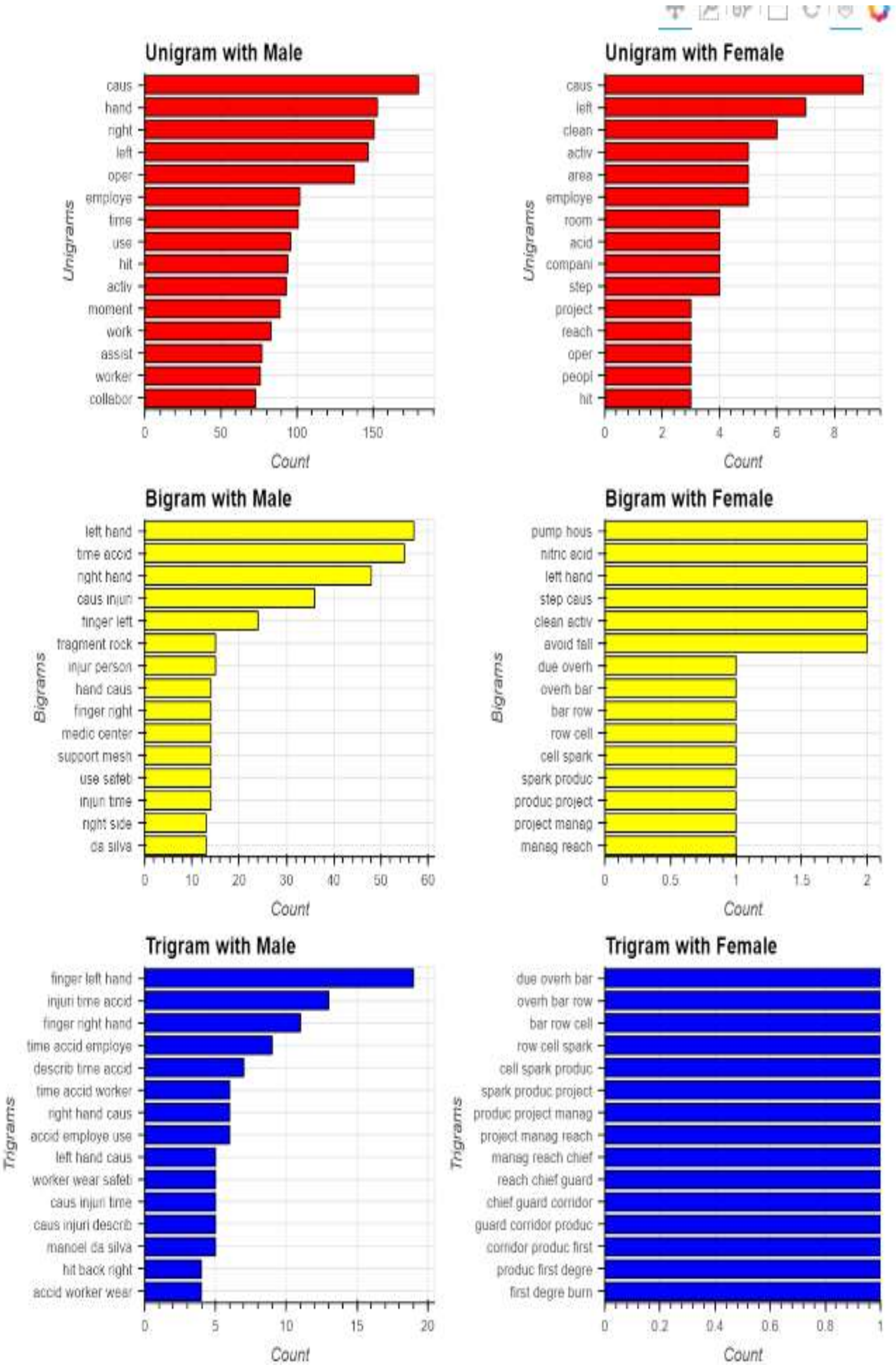


Inferences:

- Like what we found in Unigram and Bigram, there are also many phrases which are related to hands or other body parts and the confidence level increases on the body parts in which the injury is reported more. For example: - finger left hand, finger right-hand, left-hand cause, right hand cause, one hand glove
- According to Ngram analysis, we can say that operations related to hands are where the number of injuries or accidents have occurred more.
- So it is important for us to see this n-gram analysis along with other variables. We will analyse for any patterns that can help us in going towards the accident level severity prediction.

Bivariate n-gram analysis:

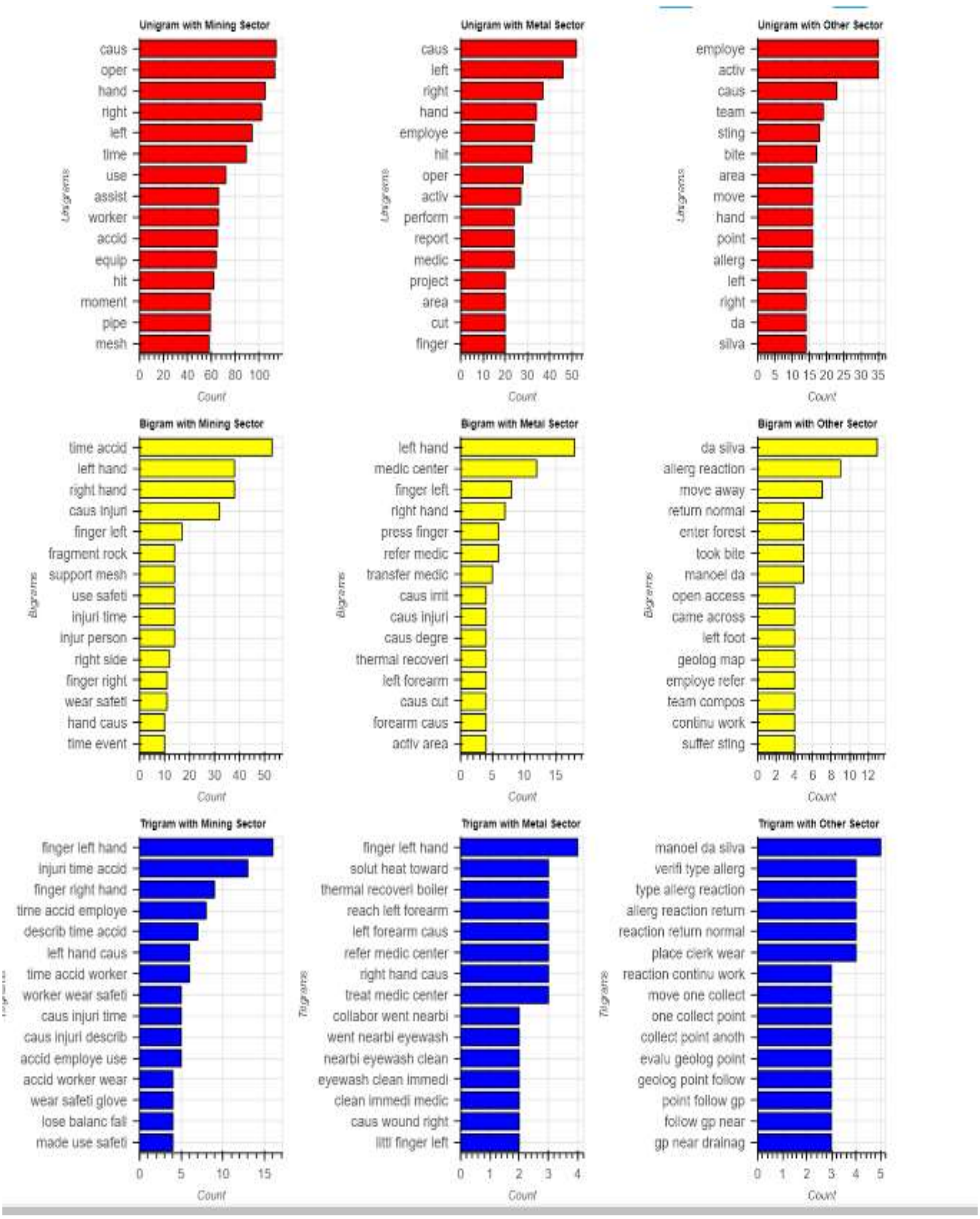
Now we will try to do the n-gram analysis along with other variables like Gender, industry, accident levels etc.



Inferences:

- It is evident from the data that the number of injured people is male and the typical injuries are related to left-hand, left-hand finger, right hand finger. There is no pattern apart from the body parts are time
- We observe that very few females are getting injured and in the similar categories like the left hand finger. The female tri-gram also shows its more to do with spark or injury caused during production related operations.
- We observe that the injury nature within the female category is distributed evenly than in males.

N-gram analysis against the Industrial Sector as variable:

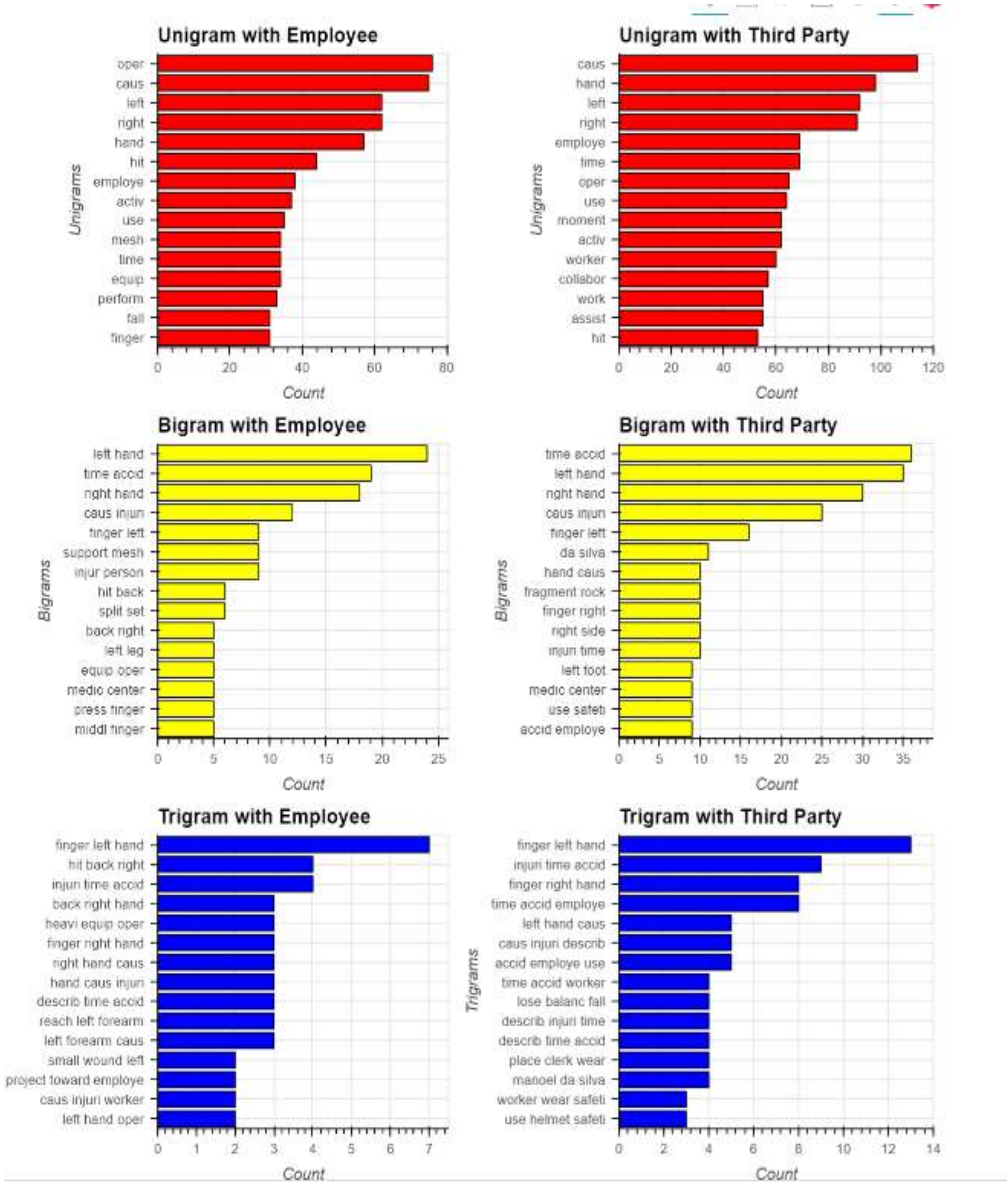


Inferences:

- Mining sector has more accidents and injuries related to left hand finger, right hand finger, workers not wearing safety uniform, workers not wearing glove
- Metal sector has less accidents and the injuries and accidents are related to left hand, solution heat, thermal recovery boiler, related to eye -wash cleaning

- Other sector injuries are very less and largely related to allergic reactions, entering the forest and took bite etc., as per the n-gram analysis.
- Mining sector uses Pipe and equipment, Metal sector shows thermal recovery boilers, eye related injuries/accidents and other sectors the accidents are observed due to employees entering forest, bite related and sting related and also related to open access/ area.

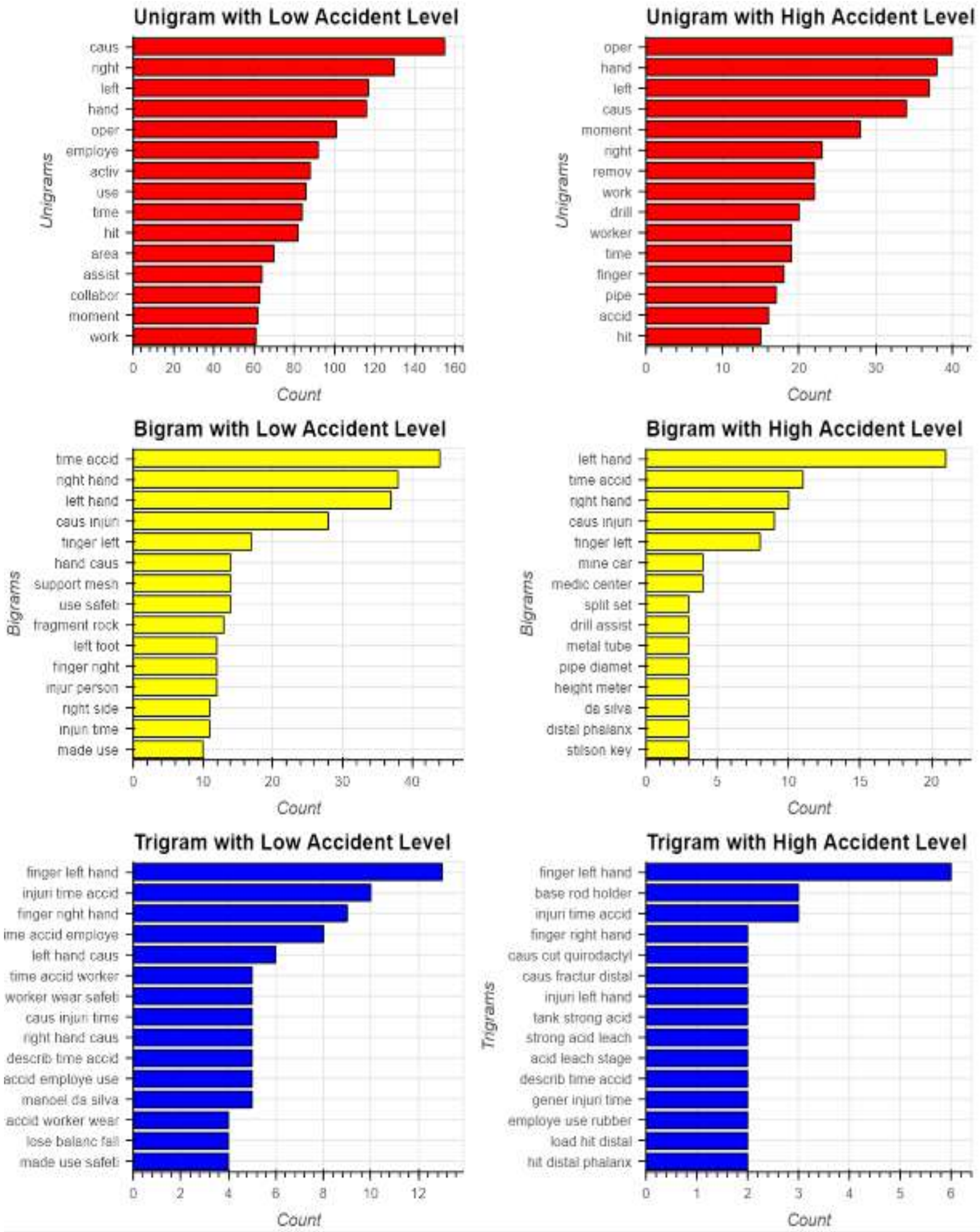
N-gram analysis with the employee type (Employee or Third Party)



Inferences:

- The observations indicate that there are more accidents in Third Party but the injury nature is similar to Employee.
- Both the types of employees face similar problems of left hand and right-hand finger injuries.
- The n-gram analysis indicates that Third Party involved in Other sectors and more injury prone due to insect bite or sting.

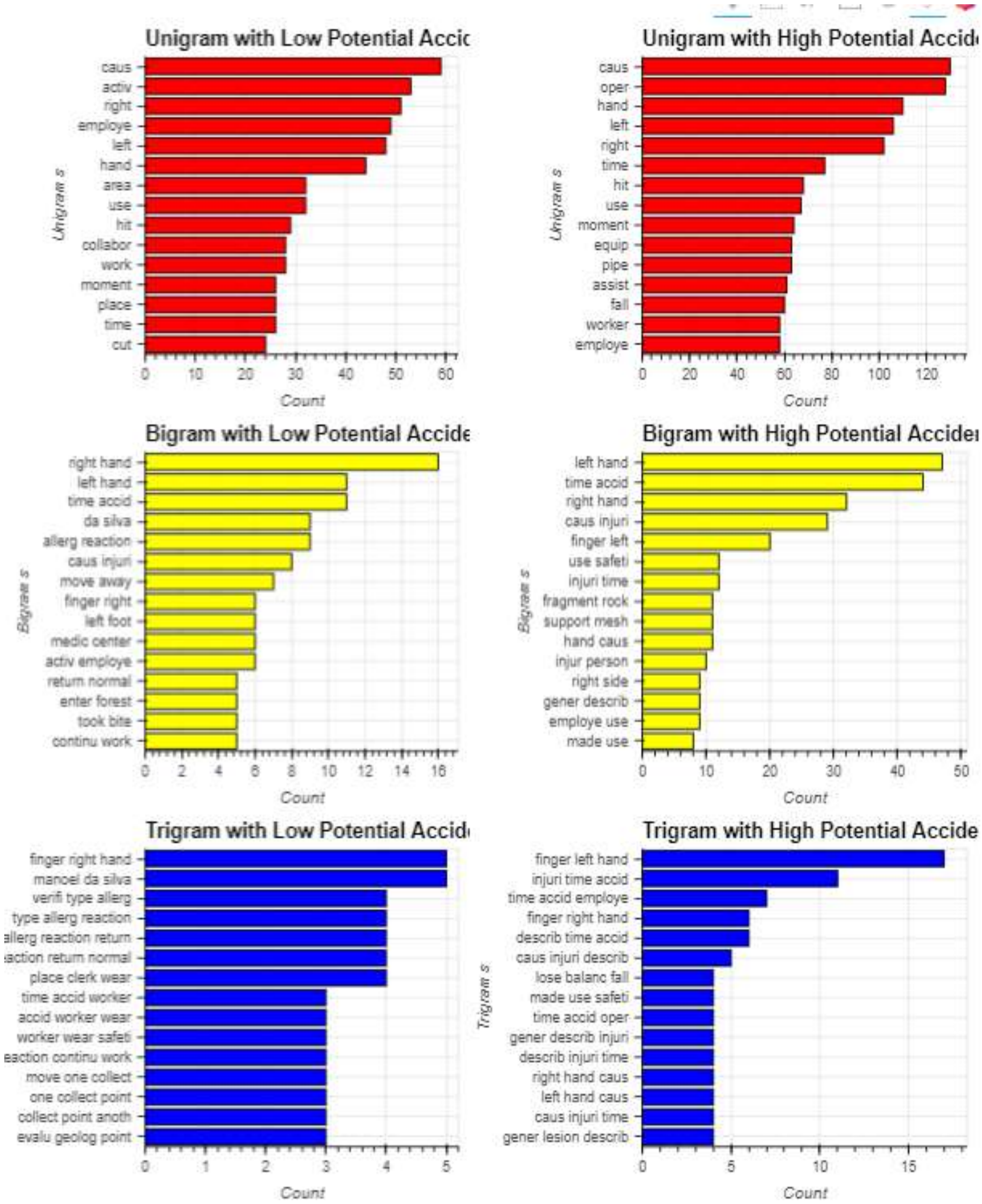
N-gram analysis with accident level as variable



Inferences:

- Let us classify the accident level as low (accident level I, II) & high (accident level II, IV, V)
- Count of Low accident levels is higher than high accident level and also it is observed that both accidents are related to left, right hand finger related injuries as the major contributors.
- High accident levels are due to Metal sector and the words such as metal rod, acid etc., are indicative of such injuries
- Here one thing to be observed is that finger left hand is present for high accident levels and also for low accident levels. That means, we need more words or more elements along with severity of accidents like a burn or a minor injury to predict the levels.

N-gram analysis with potential accident level as variable

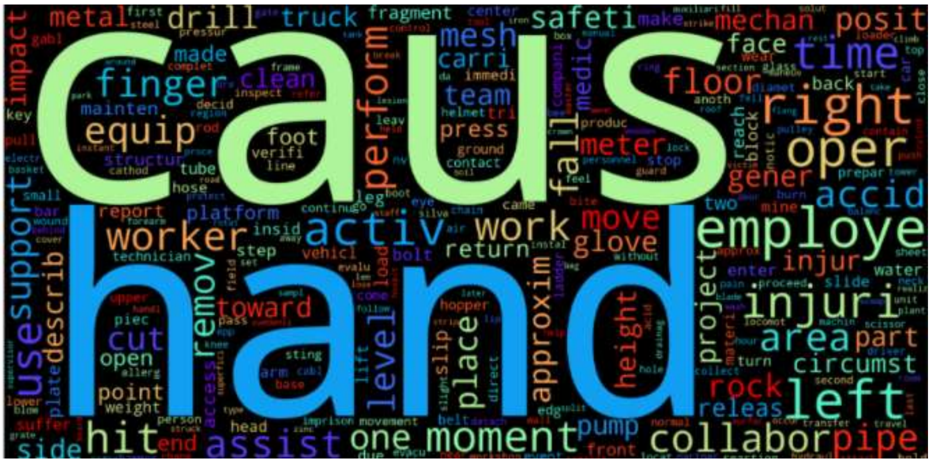


Inferences:

- Basis the n-gram analysis, it is observed that low potential level is due to allergic reaction, injuries to left hand finger, right hand finger
- Same observation as accident levels.

2.6.3 Observations on Word Cloud

There are many body-related, employee related, movement-related, equipment-related and accident-related words.

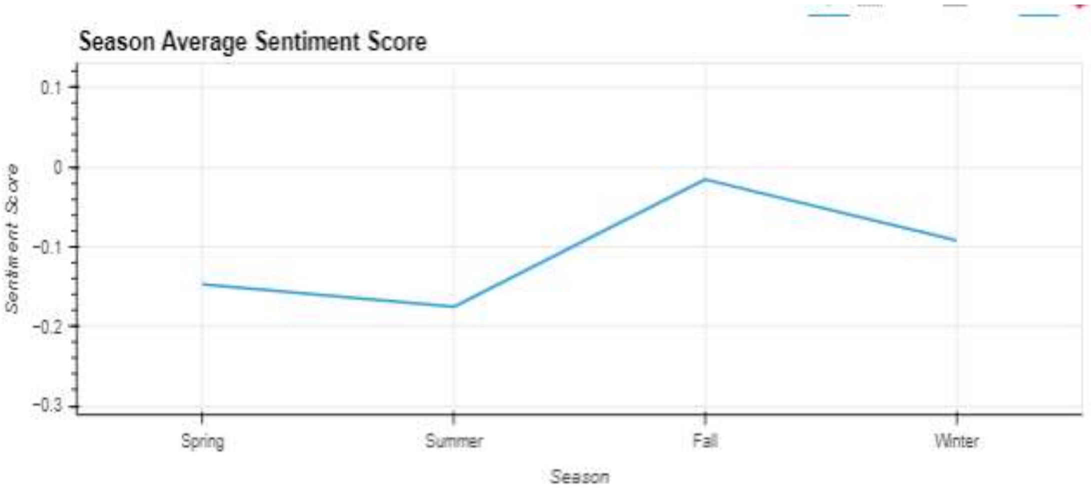


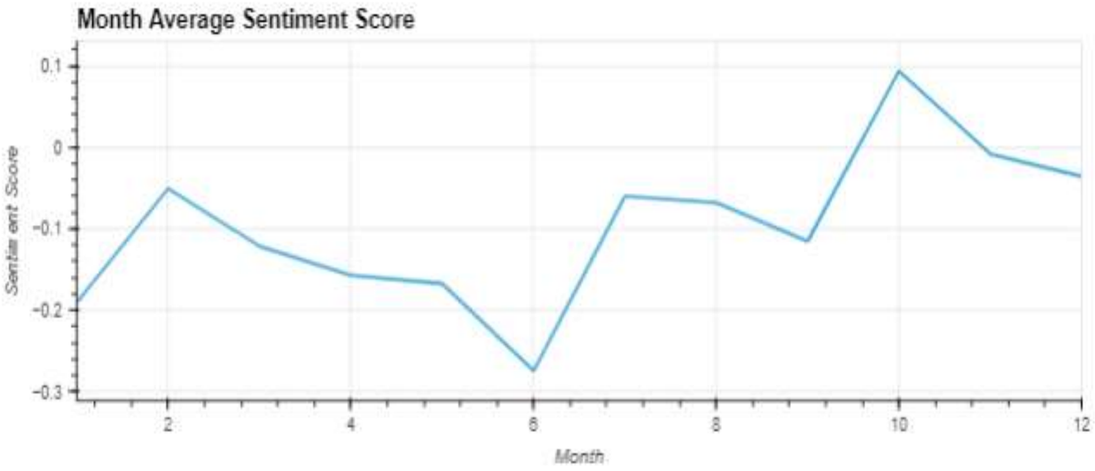
Inferences:

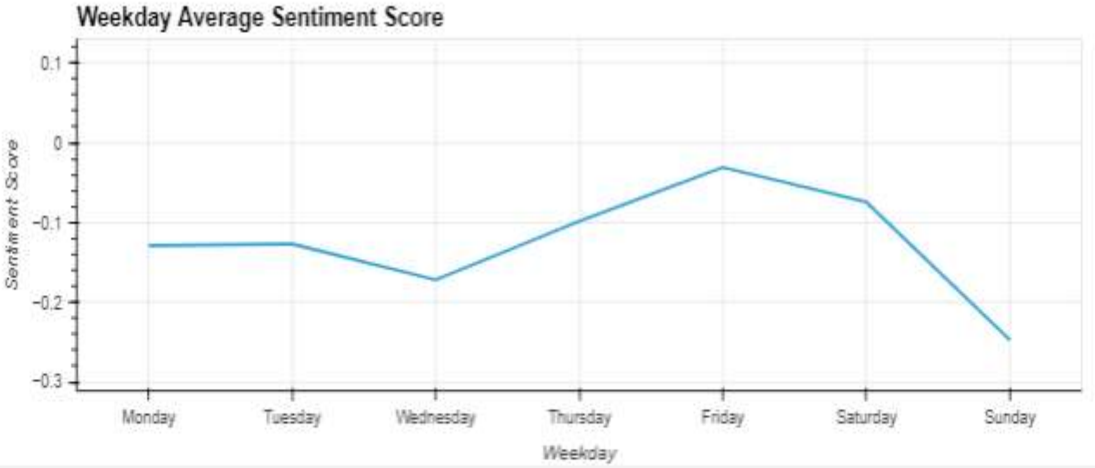
- Body-related: left, right, hand, finger, face, foot and glove
- Employee-related: employee, operator, collaborator, assistant, worker and mechanic
- Movement-related: fall, hit, lift and slip
- Equipment-related: equipment, pump, meter, drill, truck and tube
- Accident-related: accident, activity, safety, injury, causing
- Word-cloud also suggest the same set of words that have been occurring the maximum number of times.

2.6.4 Sentiment Analysis:

Based on the input variable date of the accident, a sentiment analysis was performed to determine the correlation between the accidents with day of the week, season of the year, month of the year and the accident. We built the sentiment analysis through *SentimentIntensityAnalyser* function and calculated the sentiment score through *polarity scores.*







Inferences:

- ***Sentiment basis the month of the year:***

  The sentiment score is very low in June and is observed that the number of accidents is very high during that month and October the sentiment is very high and positive indicating a smaller number of accidents.

- ***Sentiments basis the season of the year*:**

Based on the low sentiment score, it is believed that the number of accidents is very high in summer and estimated to be low in fall season.

- ***Sentiments basis the day of the week:***

Based on the average sentiment score for each day of the week, it is evident that Friday being less accident prone and Sunday indicating more employee risk due to the number of accidents being high.

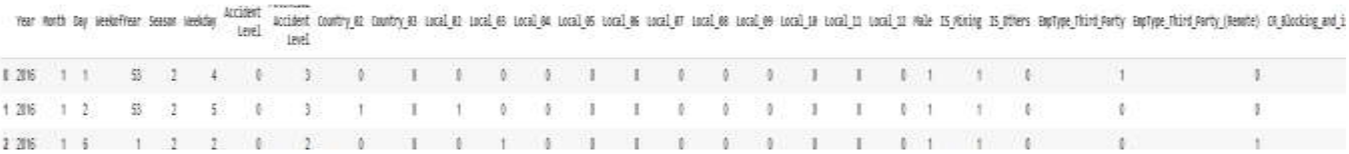
2.6.5 Feature Engineering - Embedding Models:

Since we had done sentiment analysis and n-gram analysis, we will use TF-IDF vectorisation and other embedding models like Glove and word2vec to analyse further on which words or vector of words help in predicting the right or closest potential accidental level. We take all the top n-gram words for our analysis. Why we need to use the word embedding models is that they help us really better in understanding the language semantics and in turn summarizing.

Word embeddings are a family of natural language processing techniques aiming at mapping semantic meaning into a geometric space. This is done by associating a numeric vector to every word in a dictionary, such that the distance (e.g. L2 distance or more commonly cosine distance) between any two vectors would capture part of the semantic relationship between the two associated words. The geometric space formed by these vectors is called an *embedding space*. Word embeddings are computed by applying dimensionality reduction techniques to datasets of co-occurrence statistics between words in a corpus of text. This can be done via neural networks (the "word2vec" technique), or via matrix factorization.

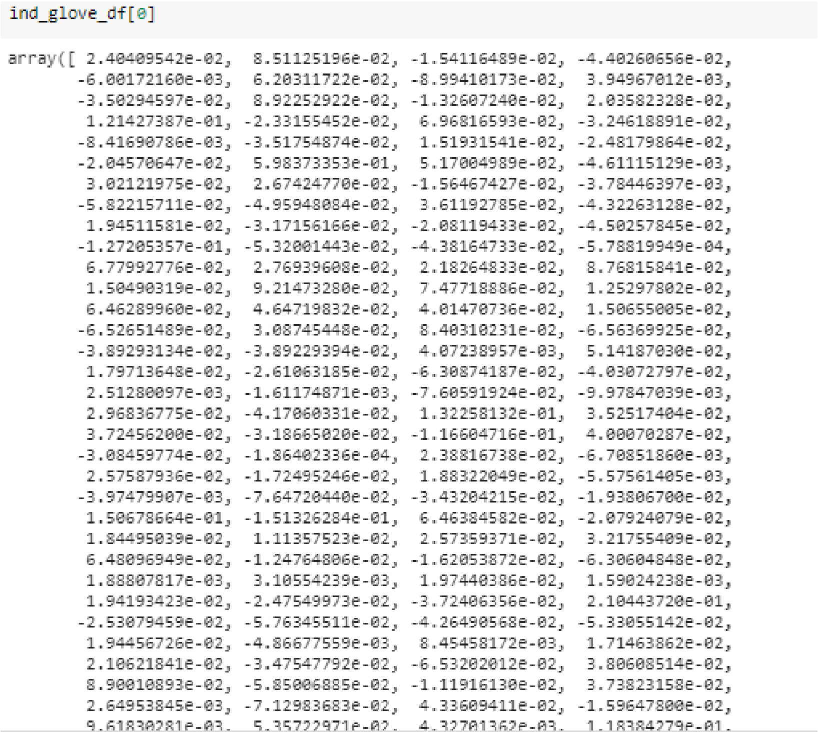
The two of the most common word embeddings are: **Word2Vec** and **GloVe,** and both of them are equally popular. But GloVe("Global Vectors for Word Representation") as the name suggests is better for preserving the **global contexts** as it creates a global co-occurrence matrix by estimating the probability a given word will co-occur with other words.

- Use TF IDF transformation model on the data frame
- We remove hyphenated words in critical risk factors
- We create the dummies for each of the columns and create one-hot encoding as shown below. The list of columns is large, so we will use word2vec at a later stage.



	Year	Month	Day	WeekofYear	Season	Weekday	Accident Level	Potential Accident Level	Country_02	Country_03	Local_02	Local_03	Local_04	Local_05	Local_06	Local_07	Local_08	Local_09	Local_10	Local_11	Local_12	Male	IS_Mining	IS_Others	EmpType_Third Party	EmpType_Third Party (Remote)	CR_Blocking and i
0	2016	1	1	53	2	4	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	
1	2016	1	2	53	2	5	0	3	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	
2	2016	1	6	1	2	2	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	

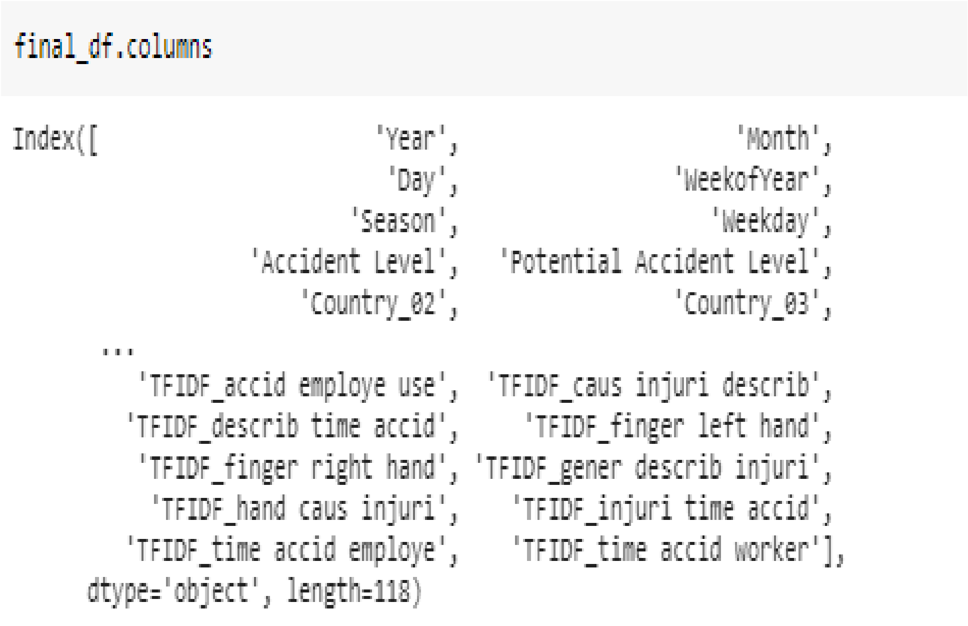
- We then use glove embeddings to get the number of words. We found 40K word vectors for the description vocabulary set.
- For each of these sentences, we create normalised vector.
- Normalised array for a glove embedded dataframe is shown as below:



```
ind_glove_df[0]
```

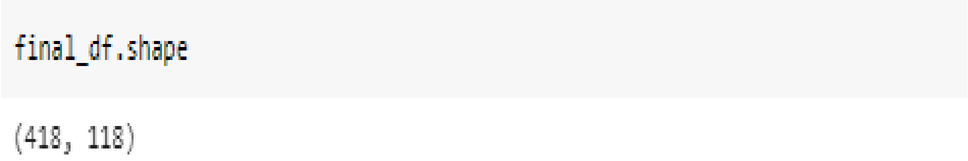
```
array([ 2.40409542e-02,  8.51125196e-02, -1.54116489e-02, -4.40260656e-02,
       -6.00172160e-03,  6.20311722e-02, -8.99410173e-02,  3.94967012e-03,
       -3.50294597e-02,  8.92252922e-02, -1.32607240e-02,  2.03582328e-02,
        1.21427387e-01, -2.33155452e-02,  6.96816593e-02, -3.24618891e-02,
       -8.41690786e-03, -3.51754874e-02,  1.51931541e-02, -2.48179864e-02,
       -2.04570647e-02,  5.98373353e-01,  5.17004989e-02, -4.61115129e-03,
        3.02121975e-02,  2.67424770e-02, -1.56467427e-02, -3.78446397e-03,
       -5.82215711e-02, -4.95948084e-02,  3.61192785e-02, -4.32263128e-02,
        1.94511581e-02, -3.17156166e-02, -2.08119433e-02, -4.50257845e-02,
       -1.27205357e-01, -5.32001443e-02, -4.38164733e-02, -5.78819949e-04,
        6.77992776e-02,  2.76939608e-02,  2.18264833e-02,  8.76815841e-02,
        1.50490319e-02,  9.21473280e-02,  7.47718886e-02,  1.25297802e-02,
        6.46289960e-02,  4.64719832e-02,  4.01470736e-02,  1.50655005e-02,
       -6.52651489e-02,  3.08745448e-02,  8.40310231e-02, -6.56369925e-02,
       -3.89293134e-02, -3.89229394e-02,  4.07238957e-03,  5.14187030e-02,
        1.79713648e-02, -2.61063185e-02, -6.30874187e-02, -4.03072797e-02,
        2.51280097e-03, -1.61174871e-03, -7.60591924e-02, -9.97847039e-03,
        2.96836775e-02, -4.17060331e-02,  1.32258132e-01,  3.52517404e-02,
        3.72456200e-02, -3.18665020e-02, -1.16604716e-01,  4.00070287e-02,
       -3.08459774e-02, -1.86402336e-04,  2.38816738e-02, -6.70851860e-03,
        2.57587936e-02, -1.72495246e-02,  1.88322049e-02, -5.57561405e-03,
       -3.97479907e-03, -7.64720440e-02, -3.43204215e-02, -1.93806700e-02,
        1.50678664e-01, -1.51326284e-01,  6.46384582e-02, -2.07924079e-02,
        1.84495039e-02,  1.11357523e-02,  2.57359371e-02,  3.21755409e-02,
        6.48096949e-02, -1.24764806e-02, -1.62053872e-02, -6.30604848e-02,
        1.88807817e-03,  3.10554239e-03,  1.97440386e-02,  1.59024238e-03,
        1.94193423e-02, -2.47549973e-02, -3.72406356e-02,  2.10443720e-01,
       -2.53079459e-02, -5.76345511e-02, -4.26490568e-02, -5.33055142e-02,
        1.94456726e-02, -4.86677559e-03,  8.45458172e-03,  1.71463862e-02,
        2.10621841e-02, -3.47547792e-02, -6.53202012e-02,  3.80608514e-02,
        8.90010893e-02, -5.85006885e-02, -1.11916130e-02,  3.73823158e-02,
        2.64953845e-03, -7.12983683e-02,  4.33609411e-02, -1.59647800e-02,
        9.61830281e-03,  5.35722971e-02,  4.32701362e-03,  1.18384279e-01,
```

- Before we head to final model building, this is our embedded model shape



```
final_df.columns
```

```
Index([                   'Year',                      'Month',
                           'Day',                 'WeekofYear',
                        'Season',                    'Weekday',
                'Accident Level',   'Potential Accident Level',
                    'Country_02',                 'Country_03',
       ...
         'TFIDF_accid employe use',  'TFIDF_caus injuri describ',
        'TFIDF_describ time accid',     'TFIDF_finger left hand',
         'TFIDF_finger right hand', 'TFIDF_gener describ injuri',
           'TFIDF_hand caus injuri',    'TFIDF_injuri time accid',
        'TFIDF_time accid employe',    'TFIDF_time accid worker'],
      dtype='object', length=118)
```



```
final_df.shape
```

```
(418, 118)
```

2.7 Conclusion of Text preprocessing or analysis on description field:

- We had done categorical column analysis on EDA section, whereas in Text processing, we took the description column and tried to do NLP model analysis.
- We used the n-gram analysis for finding the multi words that go together for our analysis
- We also did lot of feature engineering and we used glove embedding and TF IDF and came out with the finalised list of one hot encoded parameter which will be used for the model building which will be done in next section.
- **How this will be used for Chatbot utility:** Our chatbot utility will be required to take some parameters as input like Country, Locals, Gender, Industry, Employee type and critical risk factor etc. We also would want them to describe their injuries. In this field, text processing will have to be done. Hence, this is how it will be used in final GUI or NLP Chatbot utility program.

3. Model decision analysis

The models are run based on the clean samples build. First we will try and build those samples from preprocessing analysis done in earlier sections.

3.1 Sampling

We use SMOTE methods to solve some of the class imbalance distribution. SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances. We have not done PCA for dimensionality reduction since many values are not linear. Through SMOTE, we understood 47 out of 118 parameters are actually being used for prediction. The remaining are not adding much value.

We also used SHAP values to understand the decisions the model is making. Also used in better visualisation. It is more of understanding the "Why" part of predictive models.

3.2 Machine Learning Classifier Models

We will use the following multi-classification models - Machine learning models:

- Logistic regression
- KNN
- Gaussian Naive Bayes
- SVC
- Ridge and Lasso
- Decision Trees
- Randomforest

From Ensemble models, we will use the following:

- StackC
- Bagging
- AdaBoost
- LGBM Classifier
- XGB Classifier

Cross validation techniques. - K-Fold and others. Internally, we will use SMOTE for re-sampling and PCA algorithms for feature engineering. Significance of SMOTE

In our model building, we have used Lazyclassifier library which runs all the ML classifier models and give as one output. Even performance tuning is done and run across all the models

3.3 Neural Network Classifiers

Neural Networks or Deep learning

3.4 RNN and LSTM Classifiers

Typical steps of RNN and LSTM classifier.

- An Embedding which maps each input word to a n-dimensional vector. The embedding can use pre-trained weights (more in a second)
- The heart of the network: a layer of LSTM cells with dropout to prevent overfitting. Since we are only using one LSTM layer, it *does not* return the sequences, for using two or more layers, make sure to return sequences.
- A fully-connected Dense layer with relu activation. This adds additional representational capacity to the network.
- A Dropout layer to prevent overfitting to the training data.
- A Dense fully-connected output layer. This produces a probability for every word in the vocab using softmax activation.
- The model is compiled with the Adam optimizer (a variant on Stochastic Gradient Descent) and trained using the categorical_crossentropy loss.

4. Model performance improvement - Insights

4.1 ML Classifiers model performance improvement

The following steps will be taken to ensure our model’s accuracy is consistently improving.

- If you feed poor quality data in then the model will yield poor results. So, we had covered proper data as part of EDA and text processing. We have also ensured we used SMOTE for sampling and feature selections.
- Hyperparameters are parameters of the models that can be input as arguments to the models.
  - Identify Best Parameter Value Using Validation Curves like K-fold CV etc.
  - One may use Grid Search to optimise Hyper parameter combination
- Continuously fine tuning the parameter till the best accuracy is arrived at for each model.
- Choose the best model

4.2 NN Classifiers model performance improvement

- Start with *learning rate*
- Then try *number of hidden units*, *mini-batch size* and *momentum term*;

- Lastly, tune *number of layers* and *learning rate decay*.

4.3 RNN Classifiers model and LSTM performance improvement

- Tuning the Epoch size and batch sizes of every run
- Tuning the number of neurons
- Optimise the Drop out layers, Activation functions either using Relu.
- Optimisation of Algorithms and Loss functions

This Photo by Unknown Author is licensed under CC BY