1. **Summary of the problem statement,**

**Every good abstract describes briefly what was intended at the outset,**

The database is of industrial accidents from 12 different plants in Brazil and a couple more countries. The project intends to develop a chatbot that would help highlight the safety risk concerning the incident description given by the user.

**summarizes findings and implications.**

**data**

The database on hand has particulars regarding the sector and location of the incident which shows the industry of the facility ( Mining and metals) and the country and the city where the incident occurred.

The severity of the accident is recorded in the “accident level” parameter, where the user inputs the severity of the accident from a scale of I to VI. “Potential accident level” is registered based on the database regarding the potential of the incident.

The particulars of the person injured are recorded in the Gender and “Employee or third party” columns to record if the person’s gender and the relation of the person to the facility.

**Findings**

1. The original data set had a total of 425 rows and 11 columns and after the data processing and cleaning ended up with 418 rows and 10 columns.
2. “Unnamed: 0” was dropped due to its redundancy.
3. “Data” column was dropped to be replaced by five more columns, namely “Weekday”, “Week of Year”, “Year”, “Month” and “Day”. The newly added columns have been extracted from the “Data” column and bring more insight to the data.
4. A total of 7 duplicate rows were removed.
5. No null values were found.
6. **Overview of the final process**

**Briefly describe your problem methodology. Include information about the salient features of your data, data pre-processing steps, the algorithms you used and how you combined techniques.**

Treating clean Description as input for predicting Accident Level

data pre-processing:- treating

Select only alphabets

Convert text to lowercase

Strip unwanted spaces

Remove stopwords

Lemmatization

Replace empty strings with Null

Final algorithms used :- bidirectional LSTM

1. **Step-by-step walk through the solution**

**Describe the steps you took to solve the problem. What did you find at each stage, and how did it inform the next steps? Build up to the final solution.**

* EDA: -

First we have explore the dataset. In that we have seen multiple chart which help us get insight

* pre-processing:

In pre-processing there is our input column name as Description which need to process properly. For description column we done various operation selecting Select only alphabets, Convert text to lowercase, Strip unwanted spaces, Remove stopwords, Lemmatization, Replace empty strings with Null. For target column which is Accident Level first we had to map to number from roman letter and letter did one hot encoding on that .this pre-processing in must required in order to good model prediction

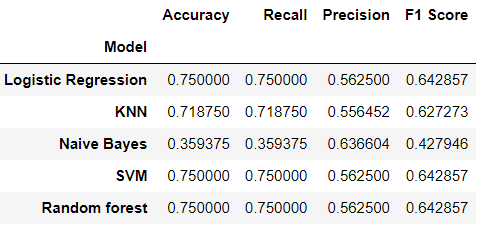
* creating embedding:

As we are working on text data it’s important to get sequential contents meaning, suppose there is one sentence, and we want to predict next word for that, the previous words are important. We as humans can predict the next word a person will say. It’s difficult for machines to do that. but there are some model which addresses to this issue

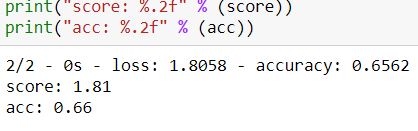
In order to able work on text data embedding is required in lay man term its create text data in numerical form. Doing that we make our data ready for able to use in model

* using multiple model

First we tried using classical machine learning models. their result as follow

****

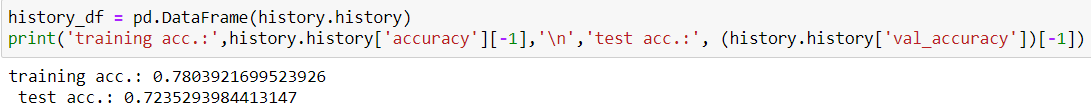
As we are working on sequential data its better to use LSTM or RNN model .we decided to use LSTM model here are result



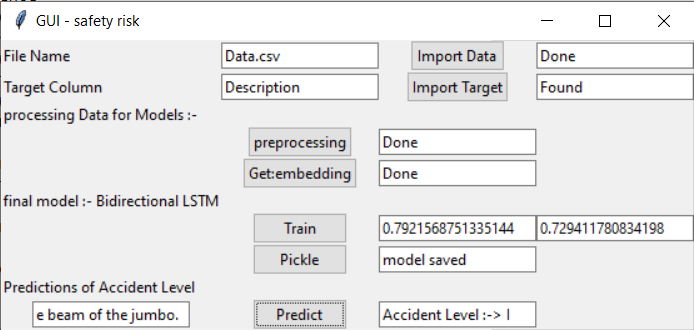
This also not accepted we need to tuning model

* select final model and tune it

Here we have used bidirectional LSTM which got use good result. here are result as given below



* create GUI for all process in one place



In this GUI we added all processing in that

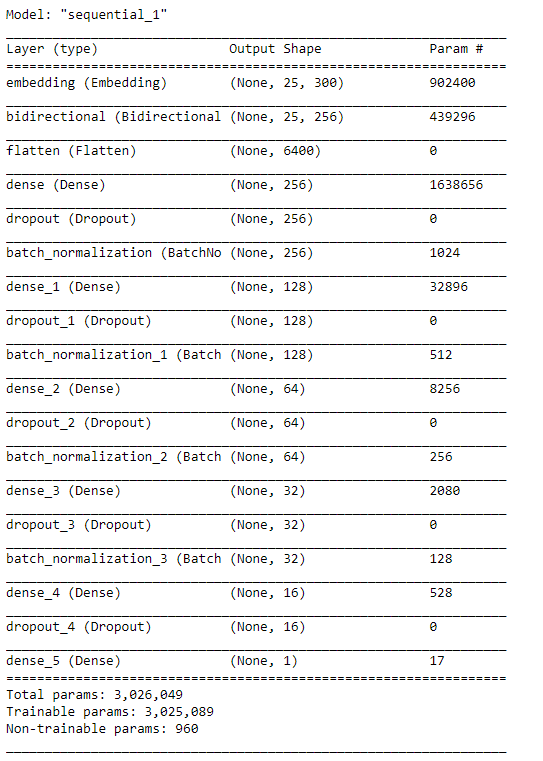
1. we have added import data button which will look for file if found we result as Done
2. we have added import target button which will for target column in dataset if found give result as found
3. for processing we have added to button pre-processing and get embedding which get our data ready for model
4. add our final LSTM model here added to button. first button is train which train our model taken ready data
5. second button is pickle button which will save model in same dir.
6. Final predictions as given sentence in box as we click on predict button it will give us prediction value as show GUI image(predicted accident level 1)

1. **Model evaluation**

**Describe the final model in detail. What was the objective, what parameters were prominent, and how did you evaluate the success of your models?**

Final model in detail :- bidirectional LSTM

Following image is summary of model

****

As we are working on text data that why LSTM give us good result.in final model we have added embedding layer, bidirectional layer, flatten, dense, dropout, batch normalization.

If we see the above image able to get complete structure about model

1. **Comparison to benchmark**

**How does your final solution compare to the benchmark you laid out at the outset?**

All our classical machine learning Model only able to give accuracy near 75% but our final model give use 80% which is good improvement

We able come with this solution because of bidirectional LSTM which got us final result

**Did you improve on the benchmark?**

Yes, From 75 % to above 80% but still not enough. since we working on human safety we required to get 100% as its about human life.

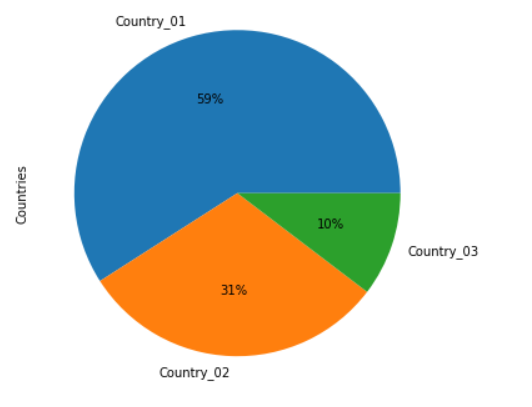
**Why or why not?**

We have got 80% we wanted improve more but due to less data not able improve that much. If we work on limitation from next we able to better result

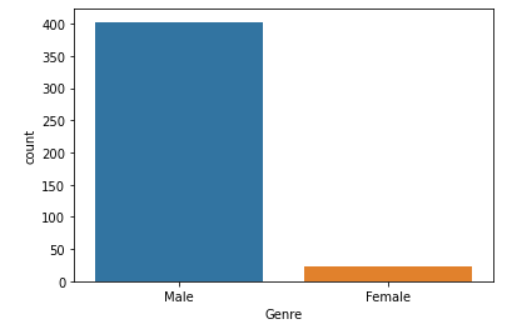
1. **Visualizations**

**In addition to quantifying your model and the solution, please include all relevant visualizations that support the ideas/insights that you gleaned from the data.**

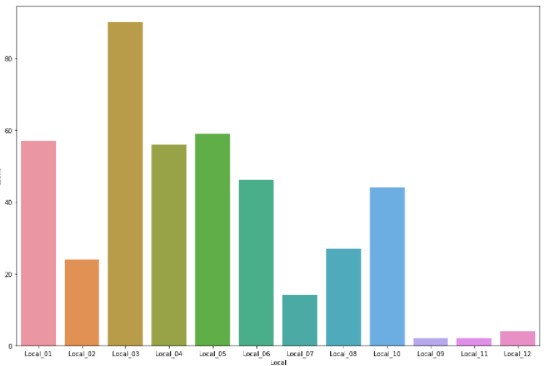
* All Chart

****

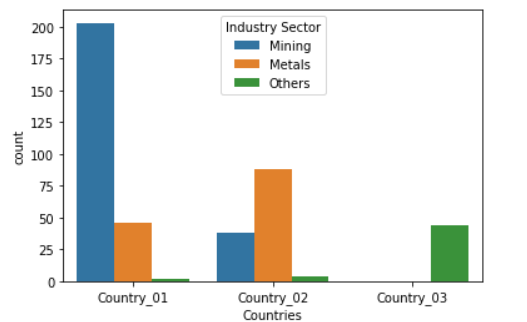
The count plot shows Country\_01 has had close to 250 incidents which is 59% of the total whereas Country\_02 and Country\_03 have reported about 129 and 41 accidents which contribute to 31% and 10% respectively.



Under the gender demographics, there is a high level of incidents among the male populace when compared to females, which amounts to less than 10% of the cases.



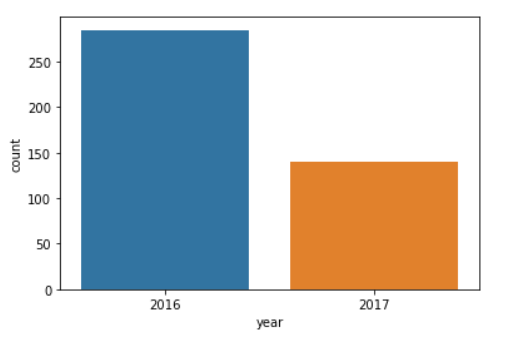
Observations regarding the cases spread across the cities, Local\_03 has the highest number of cases at above 80 and Local\_10, Local\_11 and Local\_12 amount to the least number of cases at less than 10 each.



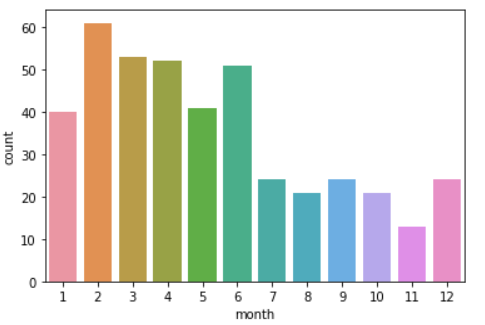
In country 1 : The mining sector has recorded the most number of cases across all incident levels, followed by metals and “others”, where it is good to see there lesser number of cases across the sectors with the increasing incident levels

In country 2: metal sector has number the most number of cases across all incident levels, followed by mining and other

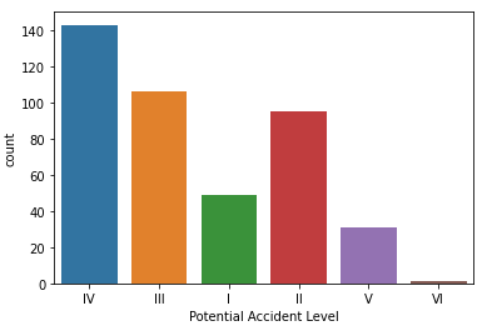
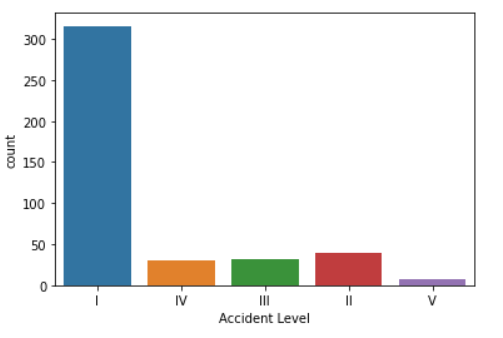
In country 2: only other sector has cases



There was close to double the number of incidents reported in 2016, compared to 2017.



The first half of the year contributed to the bulk of the cases when compared to the second part of the year, this could be due to the seasonal and weather patterns.



It’s worth mentioning that the “Potential Accident Levels” are spread across the board, whereas the “Accident Level” are mostly in Level “I” and then spread across the chart at a diminishing rate.

* + This could be due to the expert handling of the situation by the employees when an incident happens in order to reduce damages, or.
  + The incorrect projections for the “Potential Accident Levels”.

Word cloud

****

observation on word cloud:-

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

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

1. **Implications**

**How does your solution affect the problem in the domain or business?**

Using our solution we would able to come highlight the safety risk. which able to prevent any accident or injury

**What recommendations would you make ?**

* In this project, we discovered that the main causes of accidents are mistakes in hand-operation and time-related factor.
* To reduce the occurrences of accidents, more stringent safety standards in hand-operation will be needed in period when many accidents occur.
* we realized that the detailed information of accidents like 'Description' is so useful to analyze the cause.
* With more detailed information such as machining data (ex. CNC, Current, Voltage) in plants, weather information, employee's personal data (ex. age, experience in the industry sector, work performance ), we can clarify the cause of accidents more correctly.
* With more number of observations than current number of records = 425 so that we can feed more data into ML/ANN/NLP models to train, evaluate the performance of those models and get the better results.

1. **Limitations**

**What are the limitations of your solution?**

1. the data which we got are only 500 row which for not sufficient for deep learning mode
2. only one column is good features of prediction we needed more good features model
3. quality of data is not good
4. after creating model on this data model not able give good result on production

**Where does your model fall short in the real world?**

Since we are predicting the accident level, we need to be 100% sure or at least close to 100% so that we can prevent the lot of accidents in industry.

**What can you do to enhance the solution?**

Need to work on limitations.

1. **Closing Reflections**

**What have you learned from the process?**

* How to work on Data Science project to end-to-end.
* How to build different NLP architectures

**What you do differently next time?**

* Perhaps I will explore more feature engineering and feature selection techniques.