Industrial safety NLP based Chatbot

FINAL report group 3

|  |
| --- |
| Team mentor |
| Gaurav Srivastav |

|  |
| --- |
| Team members |
| Hima Bindu |
| Chandini |
| Akshay Ashok Bhamare |
| Kathambari |
| Vishnu |
| Soham Parik |

[Team Details 2](#_heading=h.gjdgxs)

[**1.**](#_heading=h.2et92p0) **Summary of the problem statement, Data and findings** 5

[1.1 Problem](#_heading=h.tyjcwt) Statement 5

[1.2 Data & Findings 5](#_heading=h.3dy6vkm)

[**2.**](#_heading=h.1t3h5sf) **Overview of the Final Process** 7

[**3.**](#_heading=h.4d34og8) **EDA and Pre-Processing** 8

[**4.**](#_heading=h.2s8eyo1) **Visualization** 10

[4.1 GRP\_0 10](#_heading=h.17dp8vu)

[4.2 GRP\_8 10](#_heading=h.3rdcrjn)

[4.3 GRP\_24 11](#_heading=h.26in1rg)

[4.4 Plot showing Frequency Distribution of words 11](#_heading=h.lnxbz9)

[4.5 Inter-topic Distance Map 11](#_heading=h.35nkun2)

[4.6 Heat Map 14](#_heading=h.1ksv4uv)

[**5**](#_heading=h.44sinio) **Model building and evaluation** 15

[5.1 Model Approach 15](#_heading=h.2jxsxqh)

[5.2 Model creation 15](#_heading=h.z337ya)

[5.2.1 Logistic Regression Model 15](#_heading=h.3j2qqm3)

[5.2.2 Random Forest Model 16](#_heading=h.1y810tw)

[5.2.3 KNN Model 16](#_heading=h.4i7ojhp)

[5.2.4](#_heading=h.2xcytpi) LSTM Model 16

[5.2.5](#_heading=h.1ci93xb) Bidirectional LSTM with Time Distributed 17

[5.3 Model Summary 17](#_heading=h.3whwml4)

[**6**](#_heading=h.2bn6wsx) **Bench Marking - Comparison of experiments** 18

[6.1 Traditional Models 18](#_heading=h.qsh70q)

[6.1.1 Output of Logistic Regression 18](#_heading=h.3as4poj)

[6.1.2 Output of Logistic Regression with Balanced Data 18](#_heading=h.1pxezwc)

[6.1.3 Output of Random Forest 19](#_heading=h.49x2ik5)

[6.1.4 Output of KNN 19](#_heading=h.2p2csry)

[6.1.5 Summary of ML Models 19](#_heading=h.147n2zr)

[6.2 NLP Models 20](#_heading=h.3o7alnk)

[Approach to select the final model based on the outcome 20](#_heading=h.23ckvvd)

[6.2.1 LSTM 20](#_heading=h.ihv636)

[6.2.2 Bidirectional LSTM with Time distributed 22](#_heading=h.32hioqz)

[6.2.3 NLP Model Summary 22](#_heading=h.1hmsyys)

[**7**](#_heading=h.41mghml) **Implication - business value derived** 23

[**8**](#_heading=h.2grqrue) **Limitations and Scope of improvement** 23

[8.1 Limitations of data 23](#_heading=h.vx1227)

[8.2 Scope of Improvement 23](#_heading=h.3fwokq0)

[**9**](#_heading=h.1v1yuxt) **Closing Reflection** 24

[**10**](#_heading=h.4f1mdlm) **Final Note** 24

[**11**](#_heading=h.2u6wntf) **Code and References** 24

# 

# 1.0 Summary of the problem statement, Data and findings

## **Problem Statement**

## **Project: Industrial safety. NLP based Chatbot**

The database comes from one of the biggest industries in Brazil and in the world. It is an urgent need for industries/companies around the globe to understand why employees still suffer some injuries/accidents in plants. Sometimes they also die in such environment

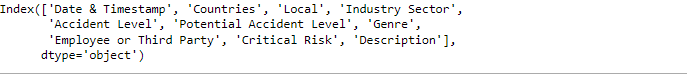
## **Abstract**

Applying traditional machine learning and neural network-based NLP to automatically classify records of accidents from 12 different plants in 03 different countries which every line in the data is an occurrence of an accident.

#### **Project Details:**

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

#### **Import Dataset and data columns**



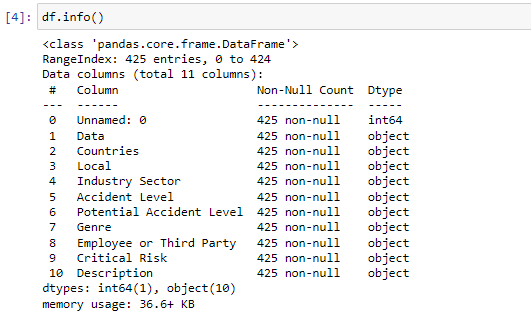
**Columns description:**

|  |  |
| --- | --- |
| **Dataset columns are below Data Set** | **Description** |
| Date | timestamp or time/date information |
| Countries | which country the accident occurred (anonymized) |
| Local | The city where the manufacturing plant is located (anonymized) |
| Industry Sector | which sector the plant belongs to |
| Accident Level | from I to VI, it registers how severe was the accident (I means not severe but VI means very severe) |
| Potential Accident Level | Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors involved in the accident) |
| Genre | 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 |

* The dataset is in csv format. Basic exploration of the data is as below

#### **Check for Shape, null values, Datatype &Missing Values**





* Non for the column have any null data.
* There are no missing values
* Total 10 columns and 425 rows of data.
* All the column datatype is object type.
* From the above output, we see that except first column all other columns data type is object

**Categorical columns**

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

Date column - 'Data'

**Given Data Summary**

There are about 425 rows and 11 columns in the dataset.

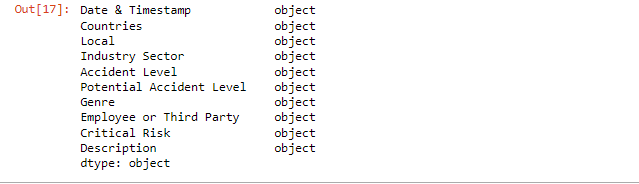
We noticed that except a 'date' column all other columns are categorical columns.

**Data Cleansing**

**Remove 'Unnamed: 0', Rename - 'Data' columns, Dropping Duplicates**

The field Unnamed: 0", is dropped and columns are renamed following

* ‘Data': Date & Timestamp



**Observation**

* Our first step was data pre-processing and in finding the right target variables.
* We observed that there are records of accidents from 1st Jan 2016 to 9th July 2017 in every month. So, there are no outliers in the 'Date' column.

▪ Winter seasons typically has more accidents.

▪ Many accidents are reported in Mining industry followed by Metal.

▪ Basis the accident data reported, Country-01 typically has more

mining plants, country-02 has more metal manufacturing plants

whereas country -03 has other industrial accidents reported.

▪ Locations or plants 01, 03, 04 and 05 have recorded more accidents than other locations.

▪ Female gender related incidents are seen more in metals industry.

* There are only three country types so there are no outliers in 'Countries' column.
* There are 12 Local cities where manufacturing plant is located and it's types are in sequence so there are no outliers in 'Local' column.
* There are only three Industry Sector types which are in sequence so there are no outliers in 'Industry Sector' column.
* There are only five Accident Level types which are in sequence so there are no outliers in 'Accident Level' column.
* There are only six Potential Accident Level types which are in sequence so there are no outliers in 'Potential Accident Level' column.
* There are only two Gender types in the provided data so there are no outliers in 'Gender' column.
* There are only three Employee types in the provided data so there are no outliers in 'Gender' column.
* There are quite a lot of Critical risk descriptions, and we don't see any outliers but with the help of SME we can decide whether this column has outliers or not.

**Data Cleansing Summary**

* Removed 'Unnamed: 0' column and renamed - 'Data' columns in the dataset.
* There are no outliers in the dataset.
* No missing values in dataset.
* We are left with 425 rows and 10 columns after data cleansing.

# 2.Overview of the Final Process

# 2.1 Salient Features of 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.

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

● Data: timestamp or time/date information

o 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 as the exact timestamp but as 00:00:00.

● Countries: which country the accident occurred (anonymised)

o Since three countries' data are recorded, country\_01 02 and 03 are the three values. It’s more of a categorical data. The data is presented for 3 different countries.

● Local: the city where the manufacturing plant is located (anonymised)

o 12 locations or plants from which accident data is captured. Again, a categorical column

● Industry sector: which sector the plant belongs to

o 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)

o This is the original recorded accident level. One being “not severe” to 6 being the “highest severity” one.

● 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)

o What could have been the severity of the accident level. Again, one being “not severe” to 6 being the “highest severity” one.

● Gender

o if the person is male of female

● Employee or Third Party

o if the injured person is an employee or a third party

● Critical Risk

o some categorisation of the risk factors involved based on the information of the accident

● Description

o Detailed description of how the accident happened. This will be used for

NLP pre-processing along with other parameters which predicts the

o 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. We will ascertain one final model which gives us accurate predictions.

# 2.2.EDA and Pre-Processing

1. Below are the Pre-Processing steps applied while performing Elaborate Data Analysis on the input data.
2. Verified the rows for Null values – no rows detected with null values
3. Checked if any column has missing value- There is no missing value found
4. Dropped the unused first column 'Unnamed'
5. Renamed the column with meaningful label
6. Visualization and preprocessing of each variable
7. Plotted the most frequent words with word cloud
8. Observed few rows of Description has trailing spaces which affects analysis. Hence removed the trailing spaces.
9. Observed punctuations, tags, tabs, digits and special characters that expects to make noise. Removed the same.
10. Converting the text to lower case, avoid any capital cases and also Converting apostrophe to the standard lexicons.
11. Using fast text library, the non-English descriptions are identified and removed the rows.
12. Punctuation marks part of English grammar are retained for the NL Model.
13. All special characters are removed for the traditional models
14. Using nltk - POS and Lemmatizer, we have further enhanced the ML\_Desription data to help the model to predict effectively.
15. In Description, English STOP words, all special characters are removed for the traditional models.
16. Used KNN model for unsampled data.
17. Using TFIDF and Glove Word Embeddings identified 400000-word vectors.
18. Performed Label encoding on Categorial Variables.
19. Defined seasons for months in Brazil.

# 3. Step-by-step walk through the solution

# The brief approach for the solution is given below

1. Solution requires model building based on the Classification model approach to design a ML/DL based chatbot utility which can help the professionals to highlight the safety risk as per the incident.
2. Data cleansing and Pre-Processing are important to have a good, cleaned input dataset for the model to predict the expected output. Hence the data cleansing and pre-processing steps are given in a detailed manner.
3. Visualization has been given to understand the dataset that feed into the model. This also helps to understand the structure of dataset
4. Two approaches of model creation is defined.

The first one is based on the conventional Machine learning algorithms Logical regression, Random Forest, KNN.

The second approach is based on the NLP algorithms LSTM and Bi-directional LSTM and Bert Model.

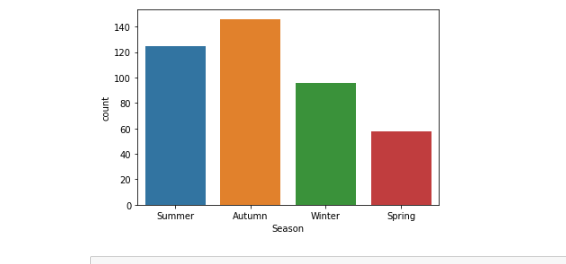
The evaluation approach is given as well.

1. The benchmarking of outcome has been captured. The performance of the model is tuned based on the different iterations with different parameters
2. The business derived value based on the outcome of the model is analyzed.
3. Limitations of the model and scope of improvement has been covered.
4. The lessons learnt on each of the step of the project is noted down and summary is provided as Learnings.

# 5.Visualization

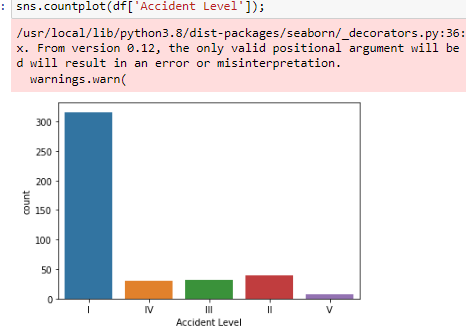
To better understand the data, I am extracting the day, month and year from Date column and creating new features such as weekday, week of year.

* As we know, this database comes from one of the biggest industry in Brazil which has four climatological seasons as below.
* https://seasonsyear.com/Brazil
* Spring: September to November
* Summer: December to February
* Autumn: March to May
* Winter: June to August
* We can create seasonal variable based on month variable.



**Univariate Analysis:**

**Checking the distribution of data based on accident levels**

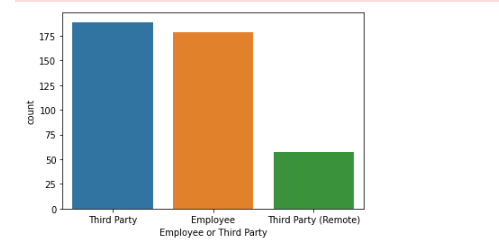


The distribution of Accident Levels is highly imbalanced in the dataset

**Distribution of the data based on country wise**

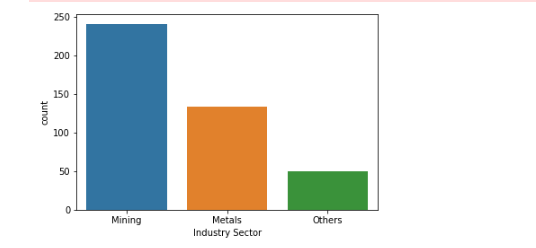


**Distribution of accidents by Employee Types**



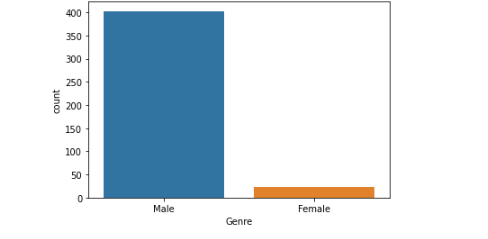
From the graph it is very clear that accidents have happened in almost equal proportions among permanent employees or third-party contractors, with third party contractors a bit on the higher side.

**Distribution of accidents as per industry sector.**



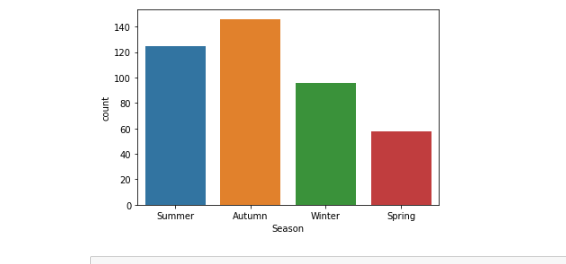
Majority of the accidents have happened in the mining sector, followed by metal industry and other type of industries.

**Distribution of accidents as per Gender**



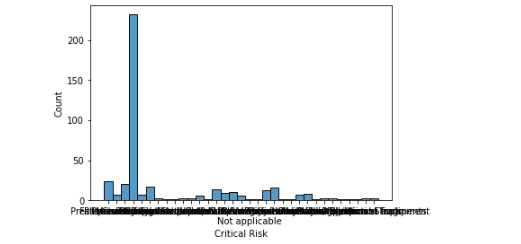
The distribution of accidents is imbalanced when checked by "Genre". The count of accidents in males is way higher than that in females.

**Distribution of Accident Asper season**



The count of accidents in Autumn is way higher than other seasons

**Distribution of Accident Asper season**



We can see from the graph that the Critical risk category "Others" have the most number of accidents. This means we are not clear about the exact risk factor associated with accidents in this dataset.

**Analyzing Description variable**

Chart, line chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

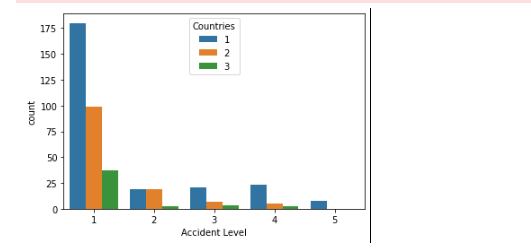
Average length of Description: 368.28

Maximum length of Description: 1029

Minimum length of Description: 94

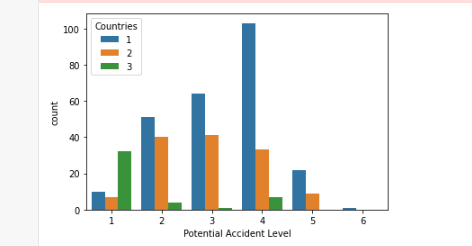
**Bivariate Analysis**

**Check the proportion of Accident in different countries**



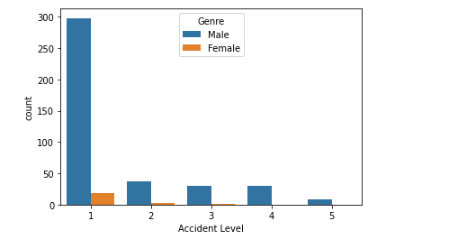
* Majority of the accident Level I accidents has occurred only in Country I.
* Maximum number of accidents in all countries are mainly of type Accident Level I.

**Check the Potential Accident level in different countries**



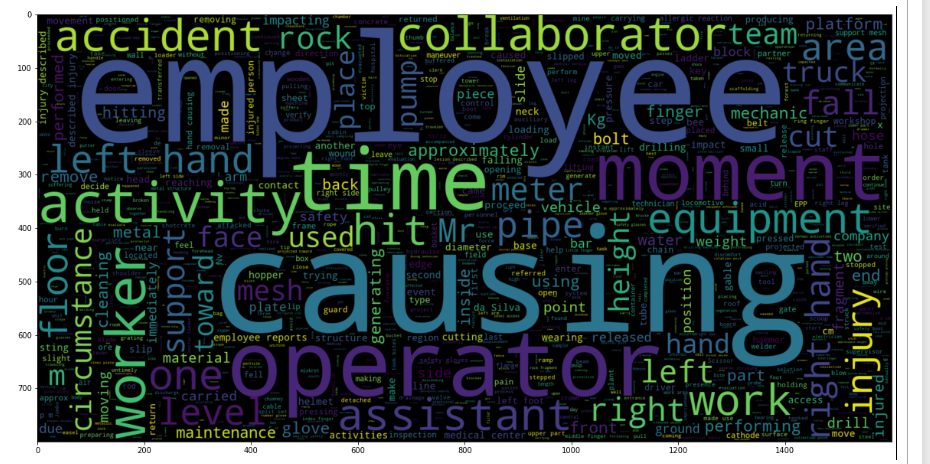
* Majority of the accidents has occurred only in Country I.

**Check the Accident level with Gender**



Majority of the accidents happened for Males.

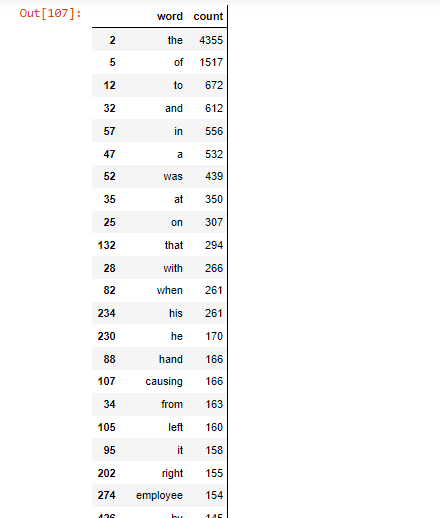
**Plotting the most frequent words with word cloud**



#### **Observations**

* 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

**Analyzing the frequent words**



# 5.Model building and evaluation

## **5.1 Model Approach**

Solution requires model building based on the Classification model approach to predict the safety risk as per the incident description.

We can approach solution with both Conventional model and using NLP.

In Conventional Model, we are using Logistic regression, Random Forest and NN models to predict the safety risk.

The second approach is using LSTM and Bert Models

## **5.2 Model creation**

Following Model and accuracy scores are given as per the initial interim stage.Further Model tuning and performance has been given in the next section

**i) Logistic Regression Model**

Using TfidfTransformer library in sklearn, bag of words is created to get the vocabulary (ngram 1,3)

Using Vectorizer transformation, features are mapped to training.

Logistic regression model is created and trained with 80-20 train test split.

**ii) Regression using Grid search**

**Accuracy: 81.1%**

Another Logistic model is also built after balancing the input data using imblearn library

**iii) Random Forest Model**

Random Forest Model created and trained with 80 – 20 Train Test Split. The parameters for the Random forest model are

max\_depth=15

max\_features

max\_leaf\_nodes

n\_estimators

**Accuracy: 76.4%**

**iv) Decision Tree**

**Accuracy: 64.7%**

**v) Decision Tree using grid Search**

**Accuracy:78.8%**

**vi) KNN Model**

Model created and trained using KneighboursClassifier using neighbours = 5

**KNN algorithm**

**Accuracy: 81.1%**

**vii) KNN Model using Grid Search**

**Accuracy: 80%**

**viii) LSTM Model**

Layer (type) Output Shape Param #

=================================================================

embedding\_1 (Embedding) (None, 95, 128) 256000

spatial\_dropout1d (SpatialD (None, 95, 128) 0

ropout1D)

bidirectional (Bidirectiona (None, 392) 509600

l)

dense\_5 (Dense) (None, 5) 1965

=================================================================

Total params: 767,565

Trainable params: 767,565

Non-trainable params: 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Accuracy: 80.1%**

**IX) LSTM Model with adding more layers**

Embedding model is created and trained using glove 6B 300 Dimensions

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

embedding\_2 (Embedding) (None, 95, 128) 256000

spatial\_dropout1d\_1 (Spatia (None, 95, 128) 0

lDropout1D)

bidirectional\_1 (Bidirectio (None, 392) 509600

nal)

dense\_6 (Dense) (None, 64) 25152

dense\_7 (Dense) (None, 32) 2080

dense\_8 (Dense) (None, 5) 165

=================================================================

Total params: 792,997

Trainable params: 792,997

Non-trainable params: 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

None

**Train Accuracy is 80%**

**Test Accuracy is 64%**

**X) RNN Model**

**Accuracy is 65.88%**

**5.1 Model Summary**

Summary of Model outputs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sl.No** | **Model** | **Accuracy** | **Recall** | **Precision** | **F1-score** |
| 1 | Random Forest | 0.764706 | 0.764706 | 0.736652 | 0.727011 |
| 2 | Decision Tree | 0.647059 | 0.647059 | 0.681448 | 0.663082 |
| 3 | KNN Model | 0.811765 | 0.811765 | 0.666807 | 0.73218 |
| 4 | KNN Grid Search | 0.8 | 0.8 | 0.66506 | 0.726316 |
| 5 | DT Grid Search | 0.788235 | 0.788235 | 0.65528 | 0.715635 |
| 6 | LR Grid Search | 0.811765 | 0.811765 | 0.658962 | 0.727426 |
| 7 | RF Grid Search | 0.811765 | 0.811765 | 0.658962 | 0.727426 |
| 8 | KNN with SMOTE | 0.329412 | 0.329412 | 0.665794 | 0.422824 |
| 9 | NN Model | 0.505882 | 0.505882 | 0.614286 | 0.554839 |
| 10 | LSTM | 0.8 | 0.8 | 0.647619 | 0.715789 |
| 11 | LSTM Updated | 0.8 | 0.8 | 0.64 | 0.711111 |
| 12 | RNN | 0.658824 | 0.658824 | 0.670296 | 0.662457 |
| 13 | Bert Model | 0.741176 | 0.741176 | 0.549343 | 0.631002 |

* pickling LSTM model as the best model.

## **Bench Marking - Comparison of experiments**

Our data set had only 418 records. There was a class imbalance for individual severity levels for both potential accident level and accident level data. If we had used this as it is, it would have resulted in lot of false positives for each of the severity levels. So, to avoid that problem, we had used binary classification based on text LSTM model to predict if an accident level is a low-grade severity level 1 and 2) or high-grade (severity level 3, 4 and 5). This resulted in good accuracy levels. This also processed the model faster.

For ML model too, since there are not many columns (since we reduced the text descriptors vector in x-array), the prediction of accident level was much better at 80% for LSTM Model.

This accuracy level is comparable to benchmark one. Had there been more data set, more diversified sampling of the various severity levels, then the possibility of better accuracy can be achieved.

# Implication - business value derived

**Observations:**

1. We have seven duplicate values in this dataset and dropped those duplicate values.

2. We have no outliers in this dataset.

3. We have no missing values in this dataset.

4. Extracted the day, month and year from Date column and created new features such as weekday, weekofyear and seasons.

5.LSTM model with an accuracy of 80% is our best model.

6.Class imbalance issue is handled using below methods and found out that, for this

particular dataset, with original data we have achieved the better results.

1. Resampling techniques: Oversampling minority class

2. SMOTE: Generate synthetic samples

# Limitations and Scope of improvement

**Limitations of data**

The limitations currently as we see it are:

• We have a smaller number of observations to analyse the cause of accidents correctly and rather we should collect a greater number of observations to get better results.

• Less number of features available in dataset.

• Lack of access to quality data. Model require more data sampling for all the severity levels. That could have helped in identifying exactly the severity level than currently telling them low or high-grade ones. A greater data observation can help. Over sampling methods like SMOTE etc. cannot help in solving this completely.

• We could have made more conversational chatbot. Getting the parameters through conversation would have enhanced the solution.

• Given the time, we could not work on better UI experience.

• We also can add continuous learning back to keep improving the model. But again, after certain accuracy level, the model will not learn but only apes the data set.

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

• Once we deploy the finalised model in Production, we feel we might get less f1-score as compared to productionalized model results.

• 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 is the “allowable limit” have to be worked with industry expert in the organisation.

**What can you do to enhance the solution?** -- Need to work on limitations

## Scope of Improvement

As mentioned in the limitation,

1. The pre-processing can be improved
2. Requesting business to give proper balanced data between the groups
3. Collecting more data for other Classes and reduce class difference within 15%
4. Equal and Standard unique words 200 or 500 on all the groups
5. Smote synthetic data.
6. Combine the similar classes with business decisions

The model is open for further improvement.

# Closing Reflection

**What did we learn from the process?**

* How to work on Data Science project to end-to-end.
* How to handle class imbalance data set.
* How to build different ANN and CNN model architectures for handling multi-class
* classification problems.
* How to build different NLP architectures for handling text data.
* It is an important industrial use case problem that requires solution. Many

organizations can have data, but to derive insights using the NLP techniques and also

the other ML models will go a long way in solving such problems.

* The learning also provided us how to deploy and how to utilise them in real world

with real use cases.

* Had a good opportunity to learn preprocessing of text using various modules
* Learnt and used Language detection modules and translation
* Visualization using Word cloud to visualize the different combination of words in a group
* Handling multi-classes distribution
* Handling imbalanced data and its implication

**What will we do differently next time**

* We spent lot of time in addressing the target variable. Once that clarity came in,

things became clearer. We will try to understand the core business problem right and then use learning to apply it better as per the problem.

* We will explore more feature engineering and feature selection techniques.
* We will build the real chatbot using Streamlit or some other applications

# 10.Final Note

Thanks to Great Learning team for the help to learn AIML and to do this Capstone Project

Thanks to Aditi who has helped in many ways to complete this course

Many thanks to our Gaurav Srivastava, his experience in the field of AIML has guided us in learning throughout this course. The team appreciates his patience. His practical knowledge gives us a lot of insights to tackle the issues.