|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| **Capstone Project - NLP CHATBOT**  **Interim report** |
| |  |  |  | | --- | --- | --- | |  |  |  | |

**Mentor:**

***Raamanathan Gururajan***

**Team:**

***Ayon Arunkumar Basu***

***Chandrasekaran Manickam***

***Mohnish Kewlani***

***Rajaraman Loganathan***

Table of Contents

[1. Problem Description 2](#_Toc67860337)

[2. Description of given dataset 2](#_Toc67860338)

[3. High Level Approach/Plan 3](#_Toc67860339)

[4. Exploratory Data Analysis 3](#_Toc67860340)

[4.1 Basic dataset analysis 3](#_Toc67860341)

[4.2 Univariate Analysis 4](#_Toc67860342)

[4.3 Bivariate Analysis 11](#_Toc67860343)

[4.4 Multivariate Analysis 15](#_Toc67860344)

[5. Feature Engineering and PCA 17](#_Toc67860345)

[6. Data Augmentation 20](#_Toc67860346)

[7. NLP Pre-processing 20](#_Toc67860347)

[8. NLP n-gram Analysis 21](#_Toc67860348)

[9. NLP Feature Engineering 26](#_Toc67860349)

[9.1 Count vectors 26](#_Toc67860350)

[9.2 TF-IDF vectors 26](#_Toc67860351)

[9.3 Word2Vec 27](#_Toc67860352)

[9.4 Glove 28](#_Toc67860353)

[10.Model Building 29](#_Toc67860354)

[10.1 Naive Bayes Classifier 30](#_Toc67860355)

[10.2 Linear Classifier 32](#_Toc67860357)

[10.3 Support Vector Machine 34](#_Toc67860358)

[10.4 Bagging Models 36](#_Toc67860359)

[10.5 Boosting Models 38](#_Toc67860360)

[10.6 Basic Neural Networks 40](#_Toc67860361)

[10.7 Deep Neural Networks 40](#_Toc67860362)

[10.7.1 Convolutional Neural Network (CNN + RNN) 40](#_Toc67860363)

[10.7.2 Long Short-Term Model (LSTM) 40](#_Toc67860364)

[10.7.3 Gated Recurrent Unit (GRU) 40](#_Toc67860365)

# 1. Problem Description

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.

The objective of this Capstone project is to design a ML/DL based chatbot utility which can help the professionals to highlight the safety risk as per the incident description.

For this task, we have been given a data set which comes from one of the biggest industry in Brazil and in the world.

# 2. Description of given dataset

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

**Columns description:**

‣ Data: 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.

As per the project requirement, we need to find out the safety risk levels. Potential accident level depends not only on accident levels, but it also considers other factors involved. From industrial knowledge, to predict the safety risk, we need to find out the 'potential' accident level. Hence, this is a classification problem where the target is Potential Accident Level.

# 3. High Level Approach/Plan

1. Identify the type of problem – Classification or Regression

2. Analyze the basic structure of the dataset, e.g. number of records, features, missing data, correlations etc. and identify the target column

3. Perform EDA

4. Identify and perform data augmentation as required

5. Perform NLP preprocessing and feature engineering

6. Perform Model training and tuning to achieve high accuracy (using multiple ML, DL and NLP algorithms)

7. Build Chatbot UI

# 4. Exploratory Data Analysis

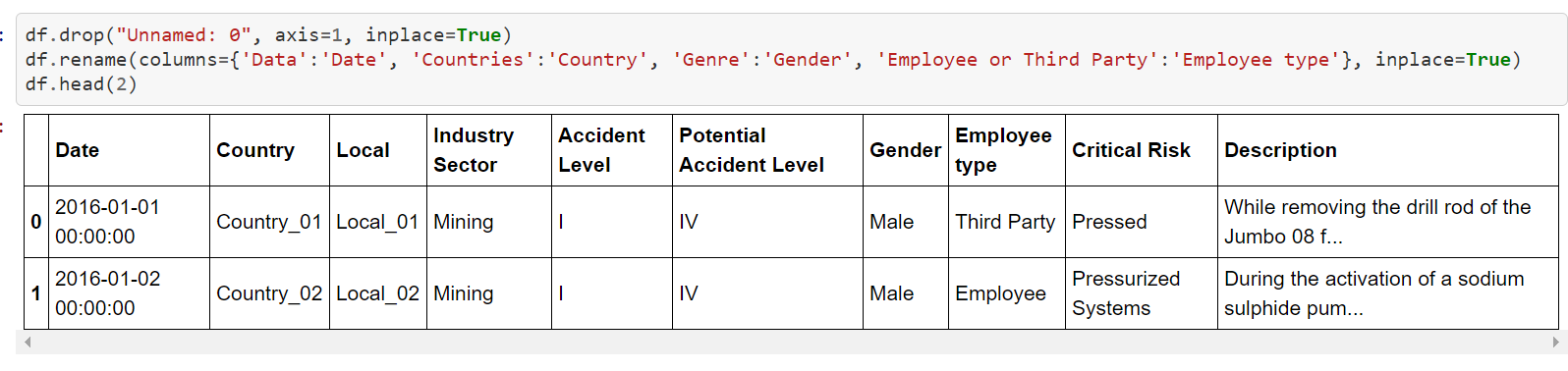
## 4.1 Basic dataset analysis

A. The data set contains 10 columns, out of which 8 are categorical, 1 is date and 1 is text.

B. There is no null value in the dataset.

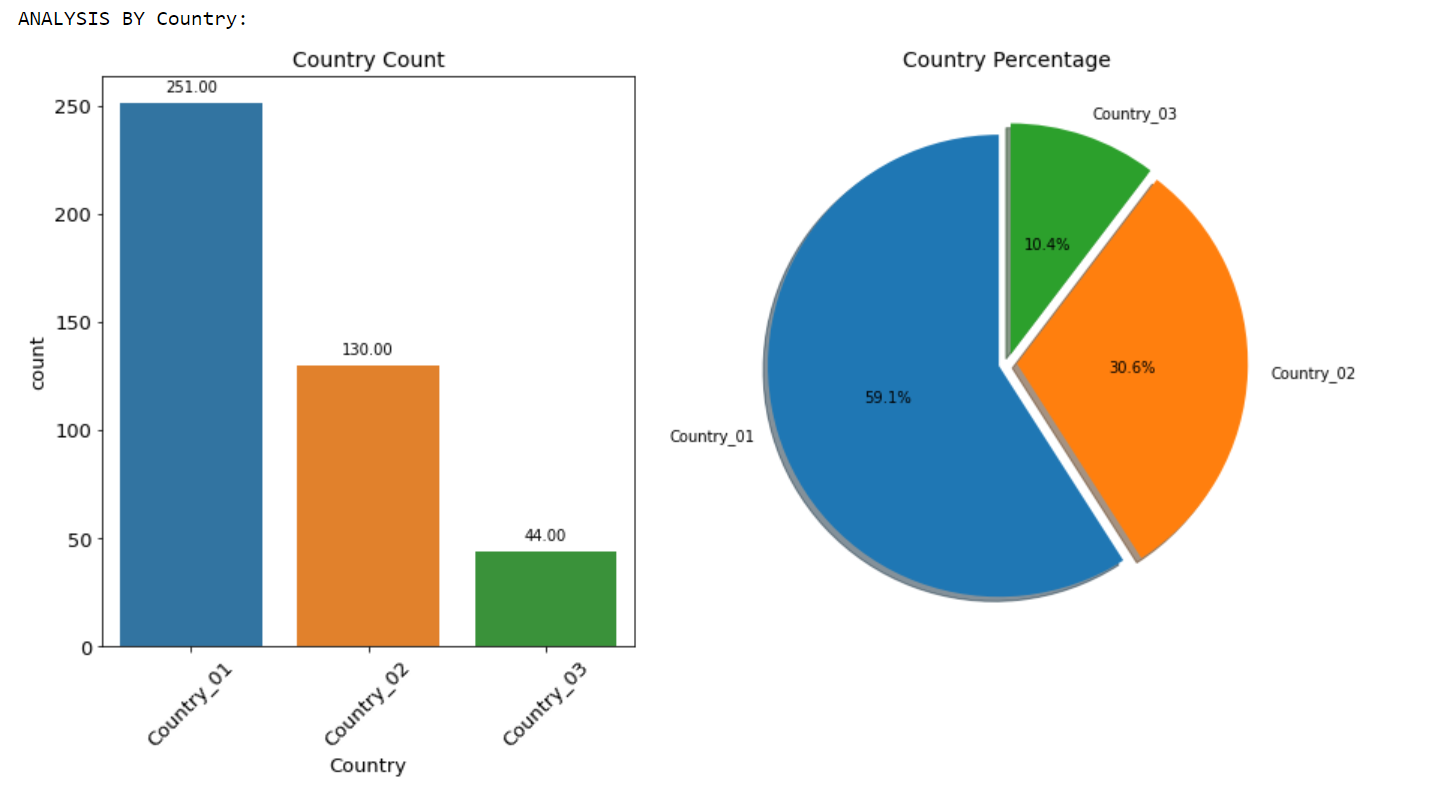
C. “Unnamed” column is just a serial number field, hence dropped.

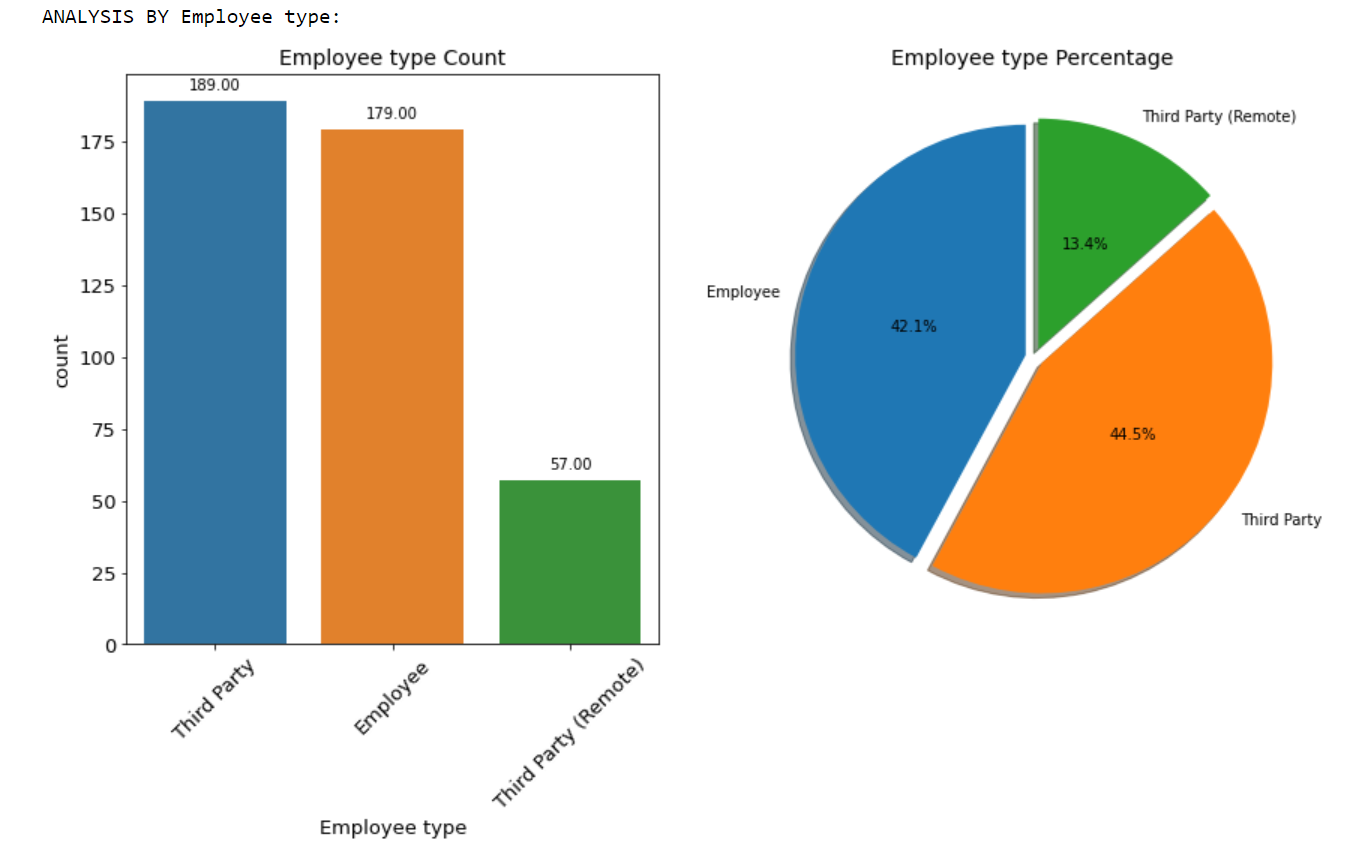
D. Renamed the labels of each column into English. The data set looks as follows now –

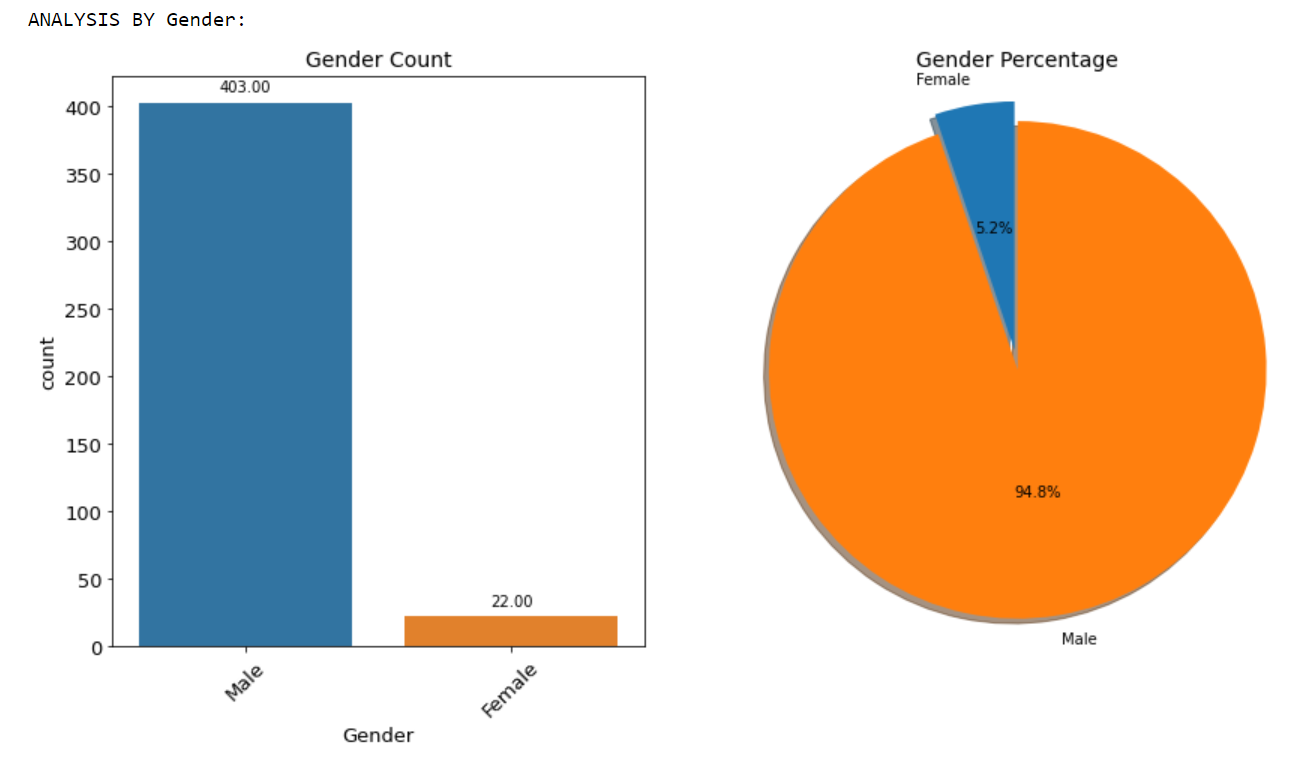


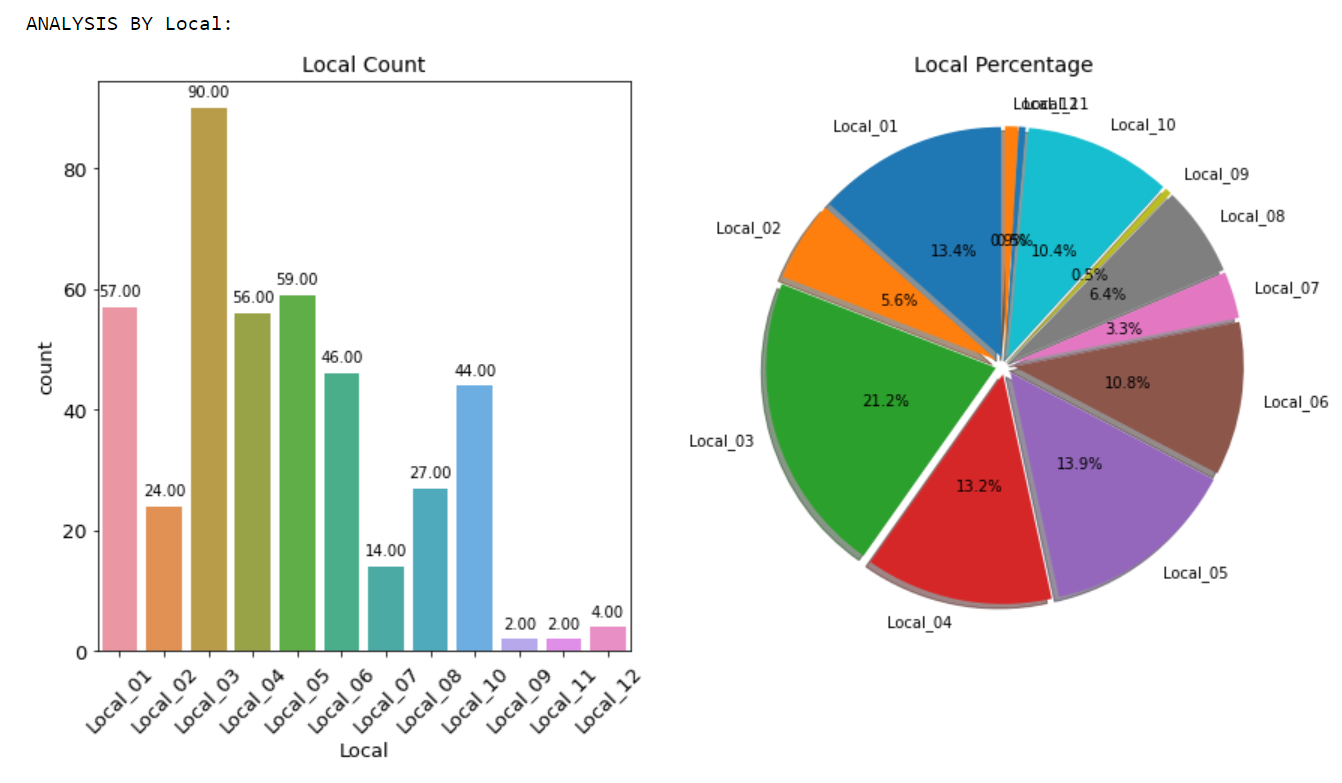
E. Split date into multiple numeric columns – Day, Year, Month, Week of Year, Weekdays and Season.

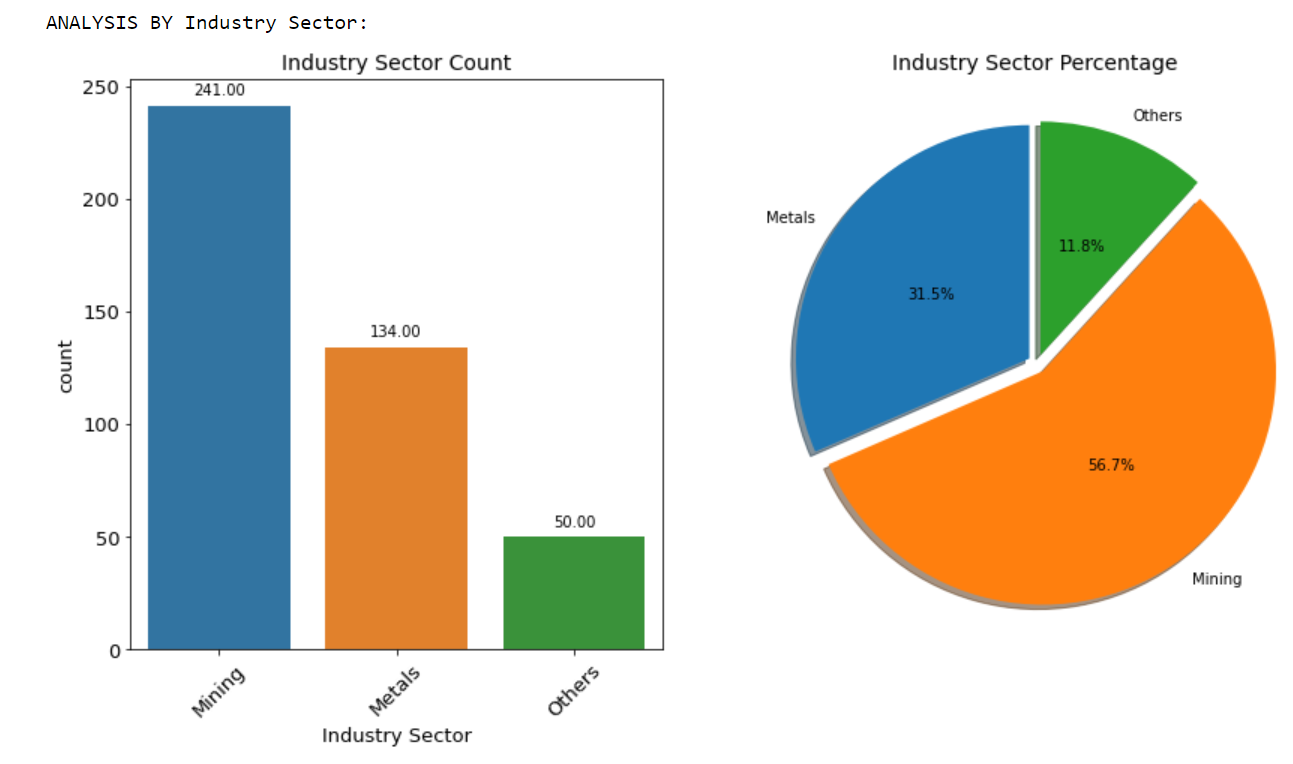
## 4.2 Univariate Analysis

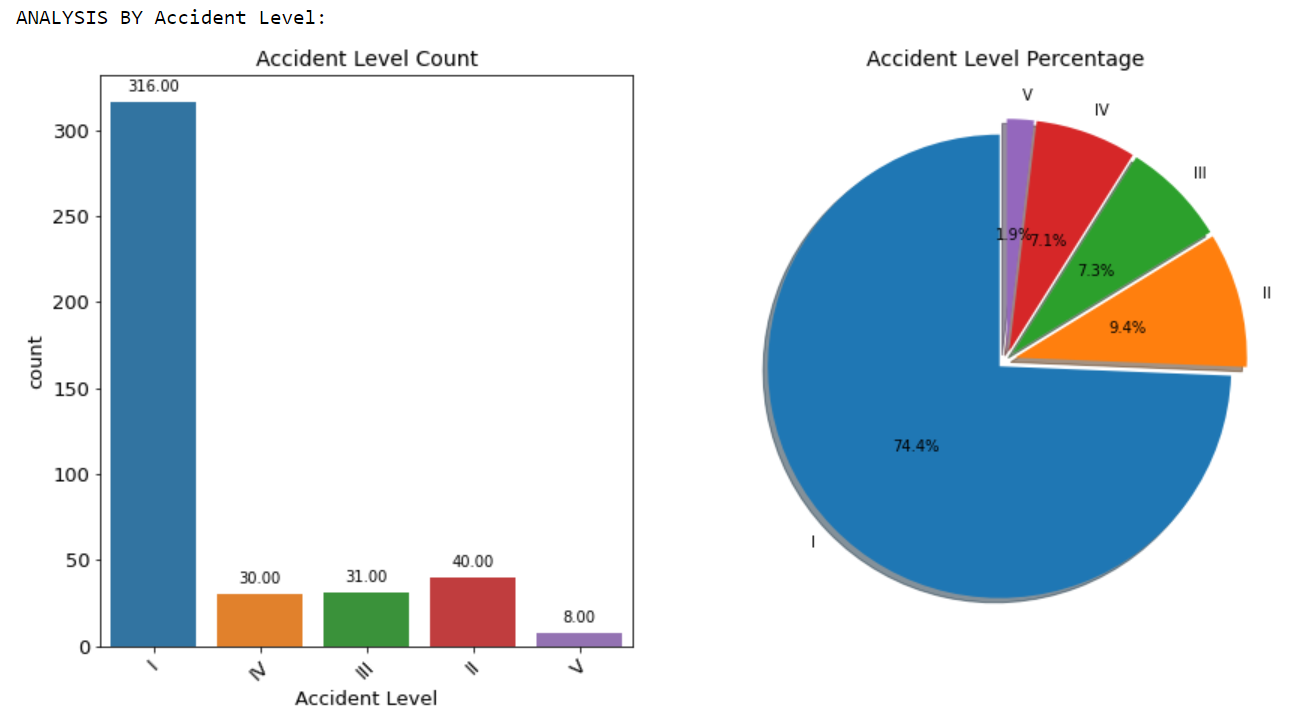


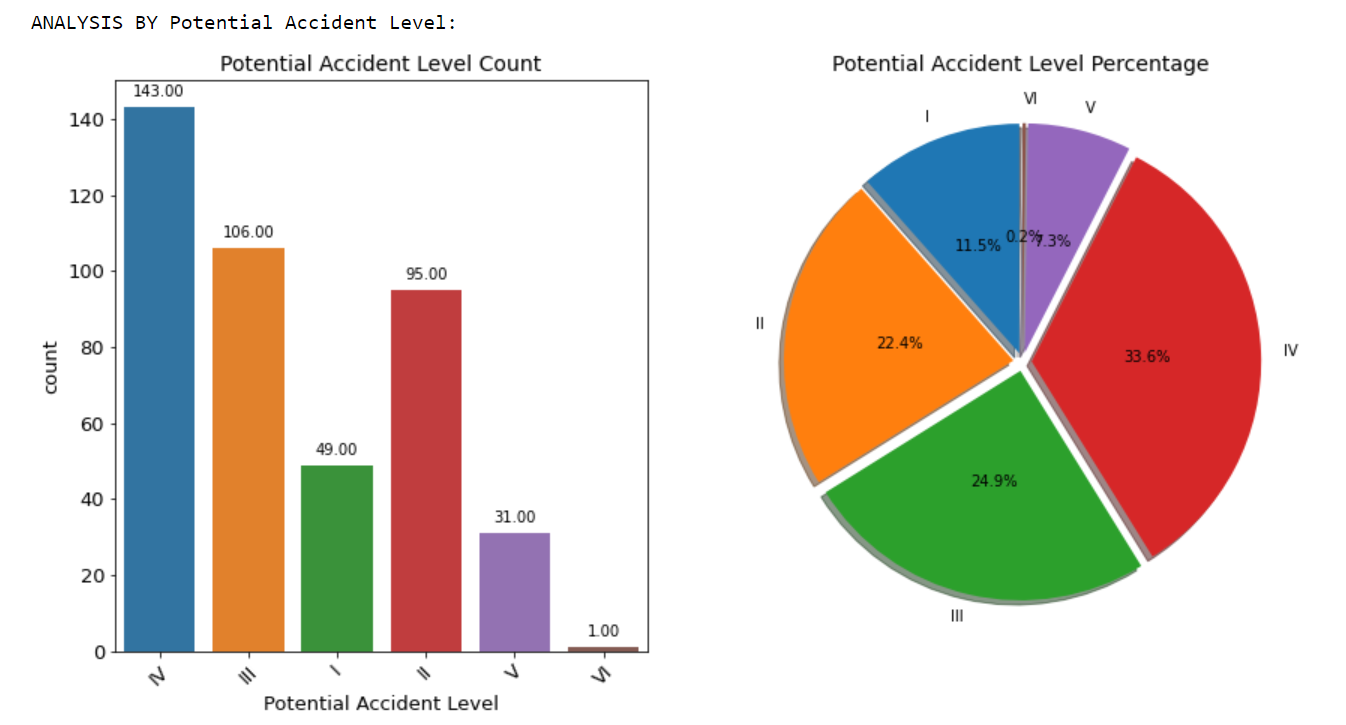


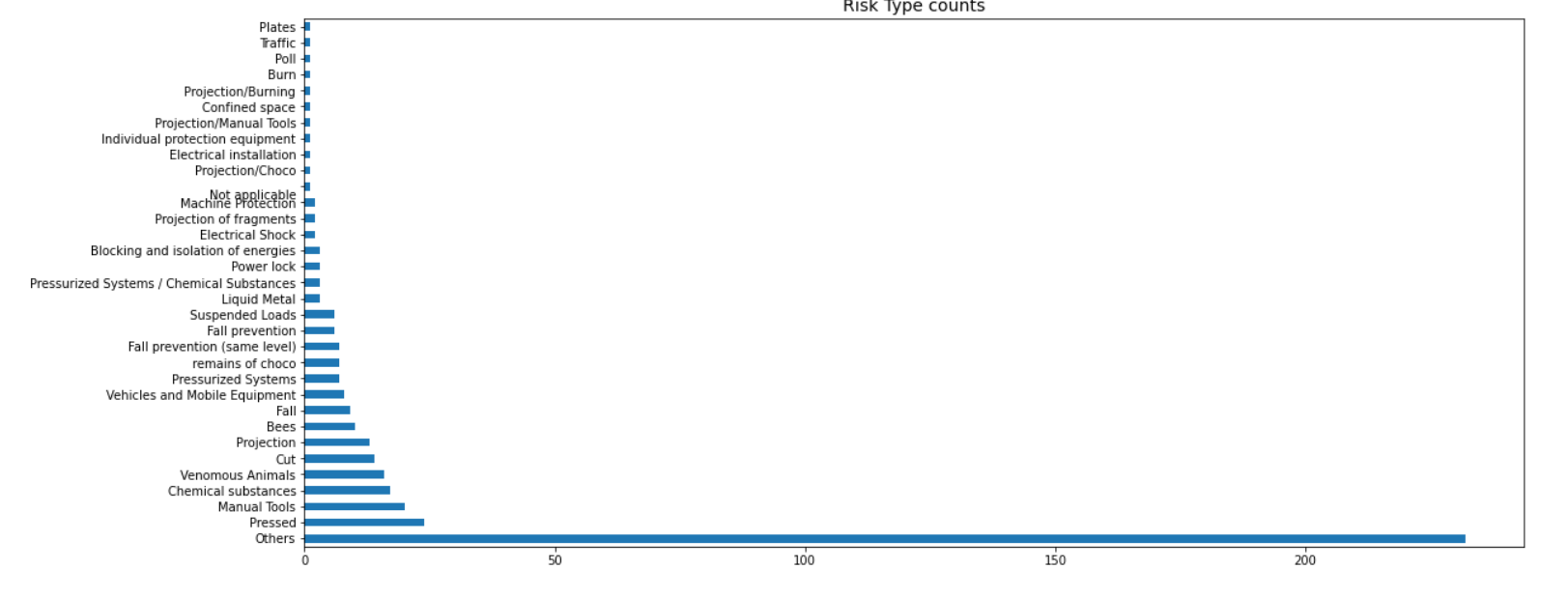


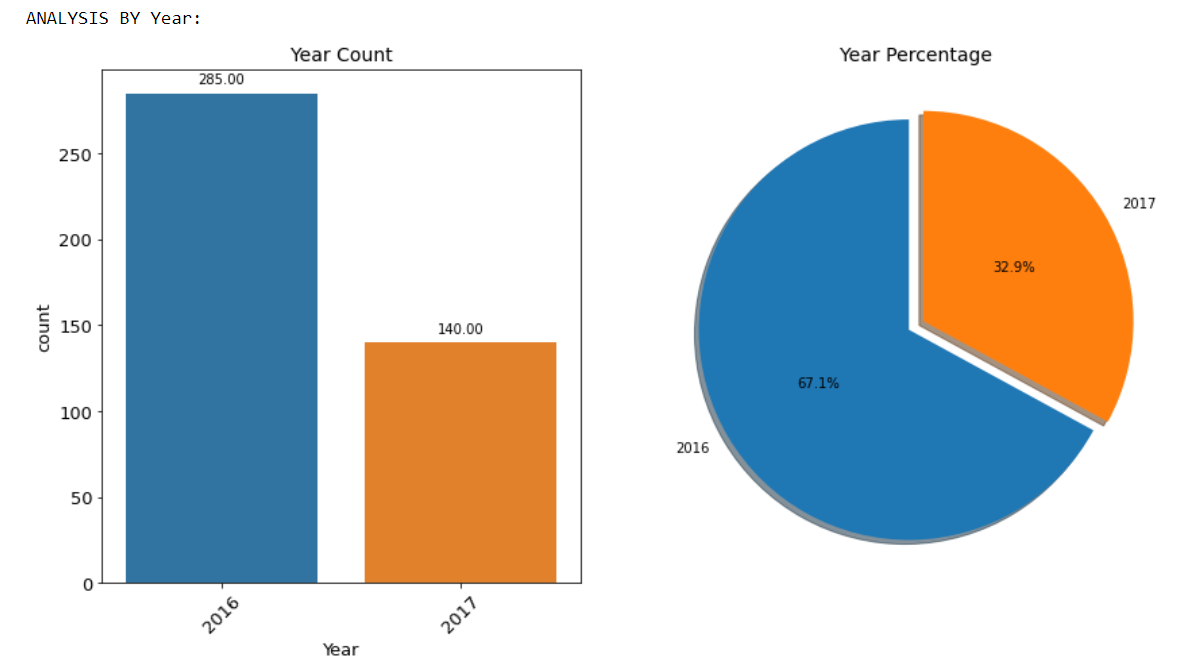


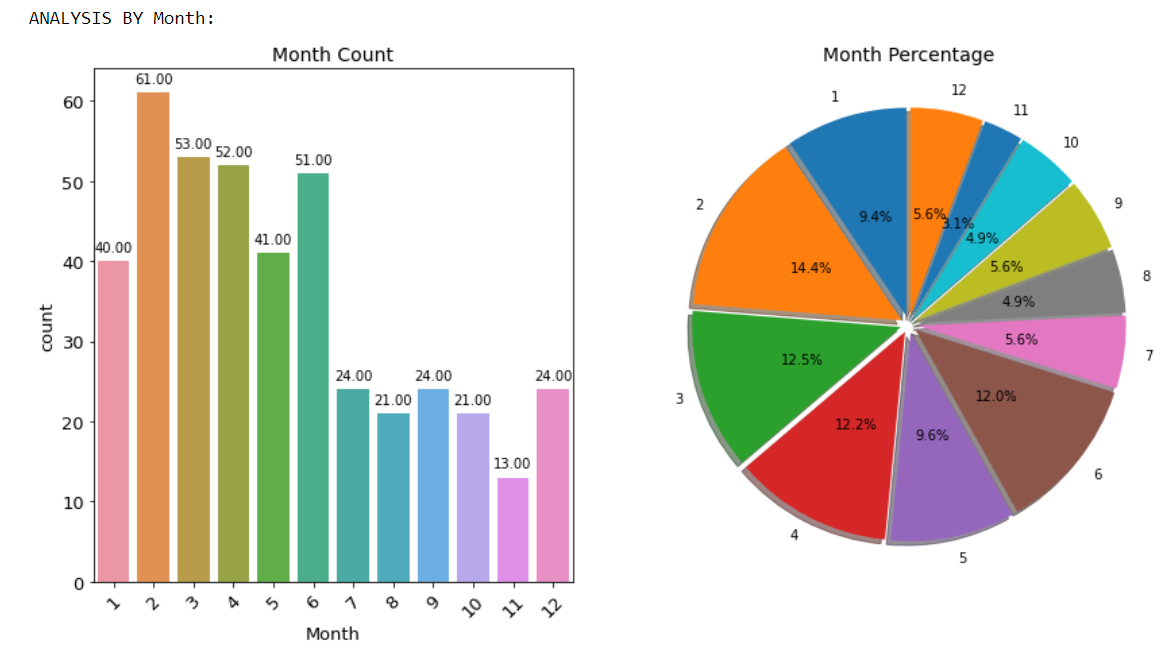


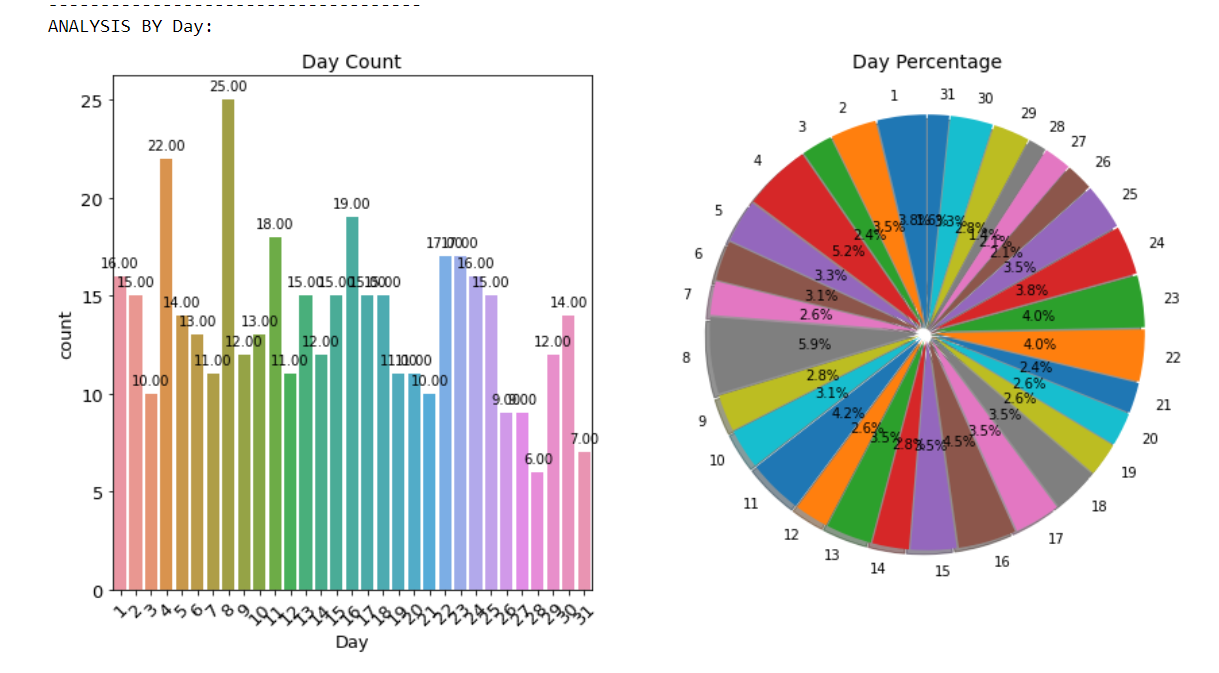


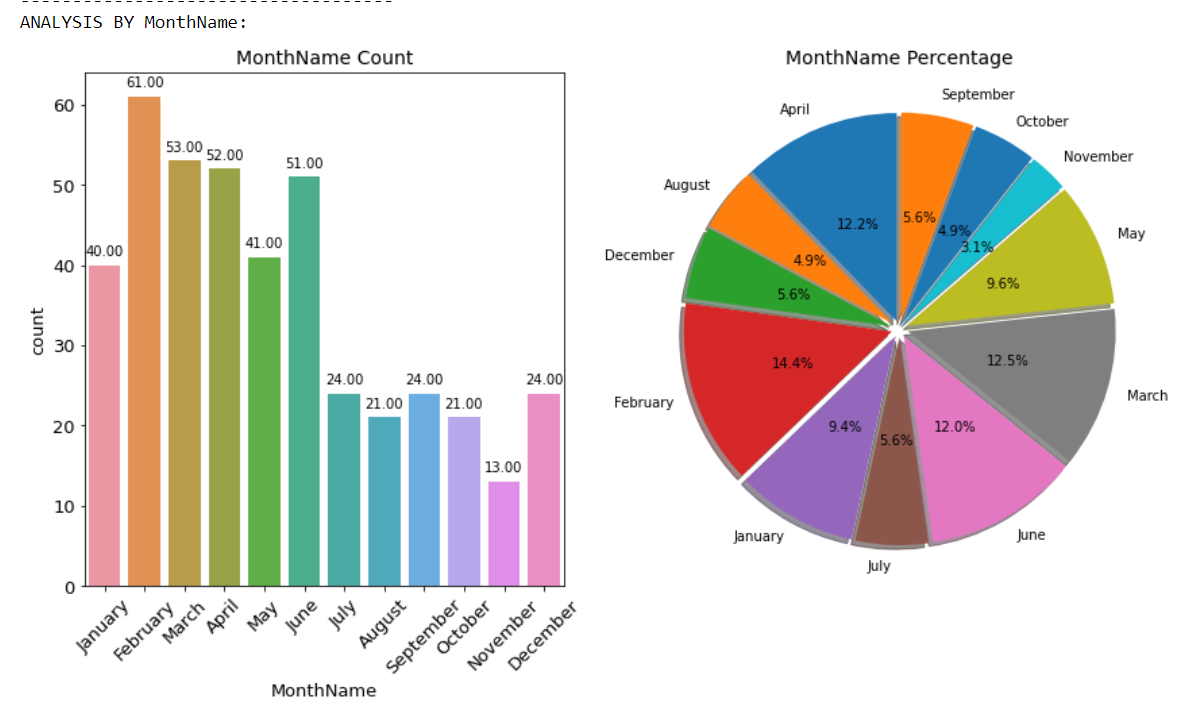


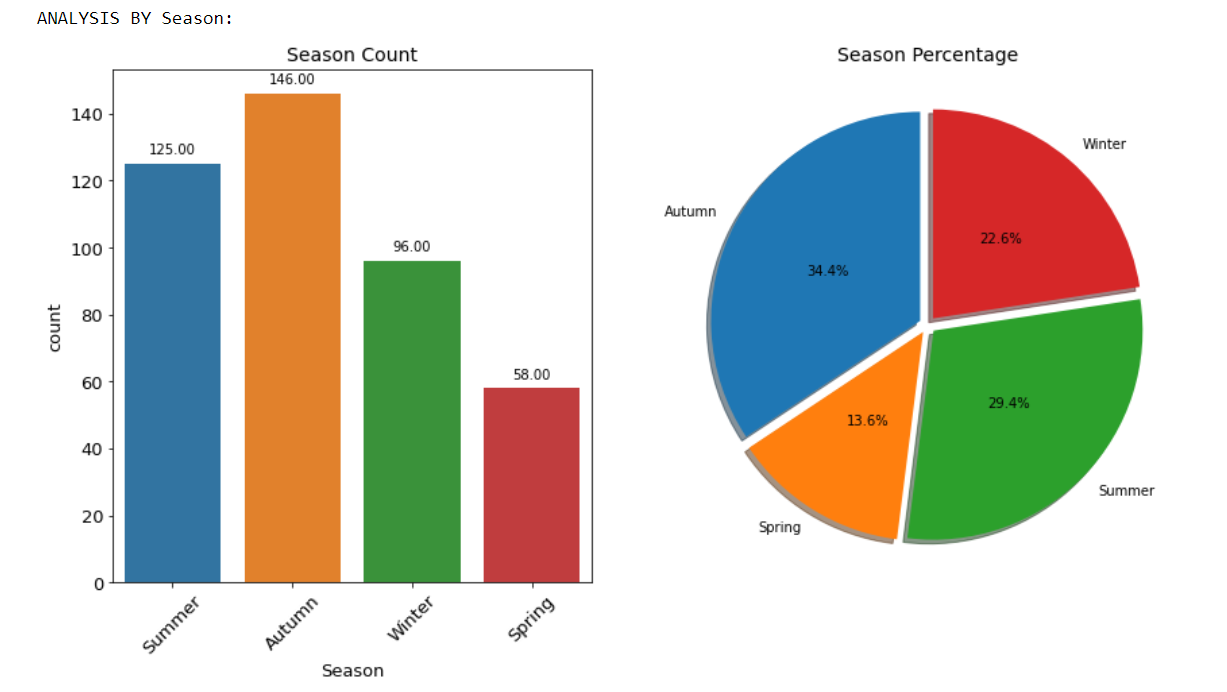








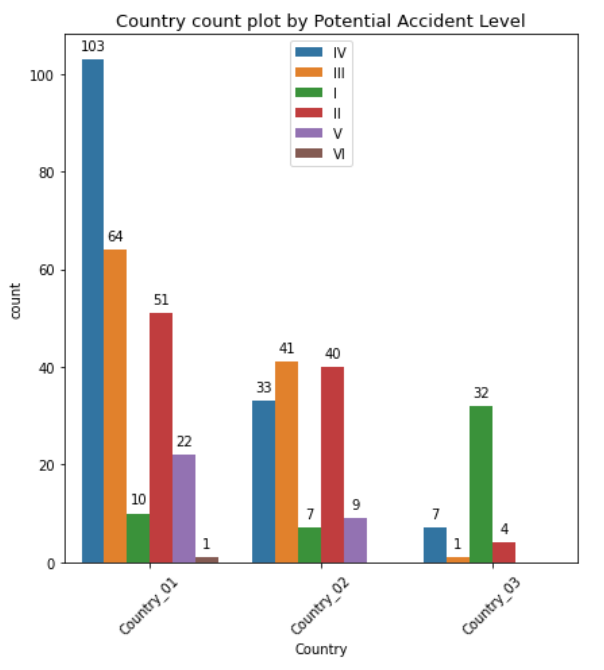




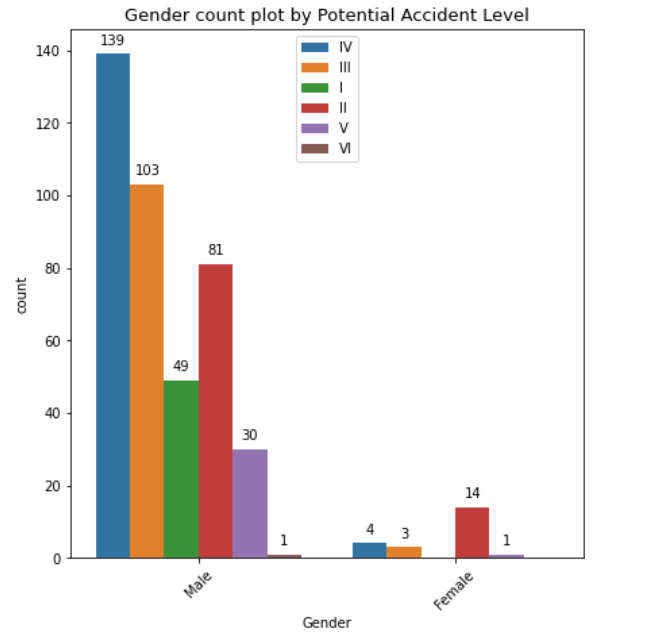
Observations :

* Country\_01 recorded with highest incident (59%) amongst all 3 countries
* Country\_03 recorded lowest (10.4%) amongst all 3 countries
* Total 12 localities data from 3 countries captured
* Local\_03 recorded with highest incident amongst all 12 countries
* 9 out of 12 localities have double digit incidents
* 3 localities have single digit low incidents
* Total 3 Industrial sectors covered
* Mining recorded with highest incident - 56.7%
* Metals stands high next to Mining Sector - 31.5%
* Other sectors contribute lowest incidents - 11.8%
* Accident Level - I occured at high frequency 74.4%
* Accident Level - V occured at low frequency 1.9%
* High Level accidents are lower and Low Level accidents are higher
* Predicted Potential Accident Level - IV is at high frequency 33.6%
* Predicted Potential Accident Level - VI is just 1 occasio - 0.2%
* Actual Accident levels are different than Predicted Potential Accident Level
* Males are high prone to exposed to Accident (94.8% compare to Females (5.2%)
* Third Party Employees and Direct Employees are more prone to Accident than Remote working third party
* Accidents are categories into 33 types of Risk Types
* Undefined other type of Risks contribute to 54.6% accident
* It indicates deep investigation and proactive approaches - FMEA/RCA are needed to identify the type of risks
* Y-2017 Accident is halved Y-2016 - Good controls in Accidents and safety performance
* In general Accident occurs through the year irrespective of months
* However in the month of Feb and Apr accident level is high
* July, Oct & Dec have 50% of reduced accident level than Feb & Apr
* Beginning of every month have more accidents.
* End of the the month relatively accidents are at lower side
* Except Sunday all other days there is a accident
* However Sunday also Accident occurs
* Almost equal amount of accident occurs across all seasons except Spring

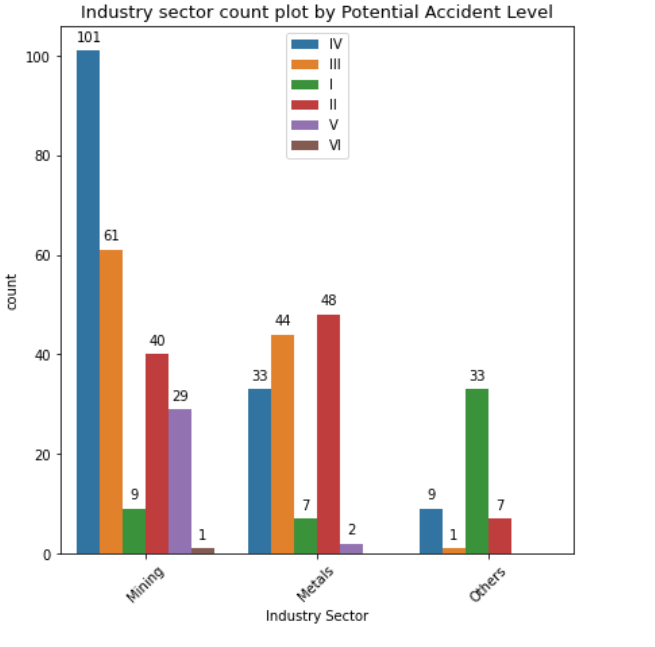
## 4.3 Bivariate Analysis



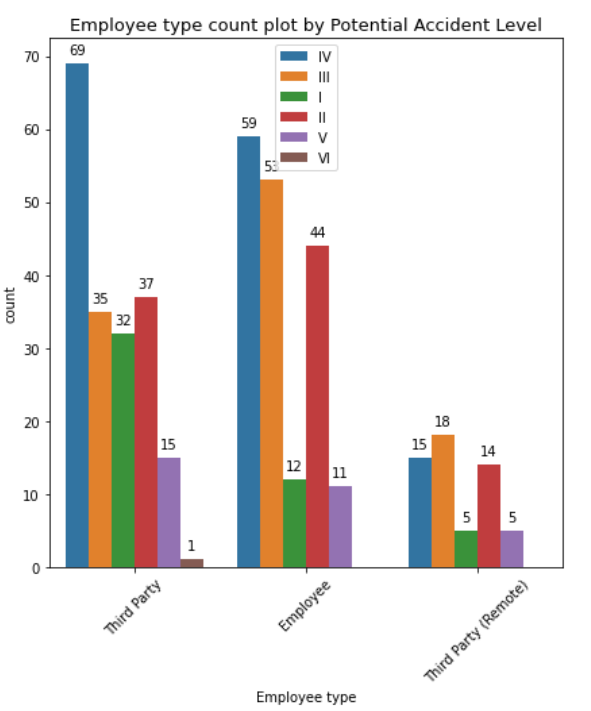
* PAL IV, III and II cases are the highest in country\_01
* PAL III and II are almost equal in Country\_02
* PAL I is the highest in Country\_03



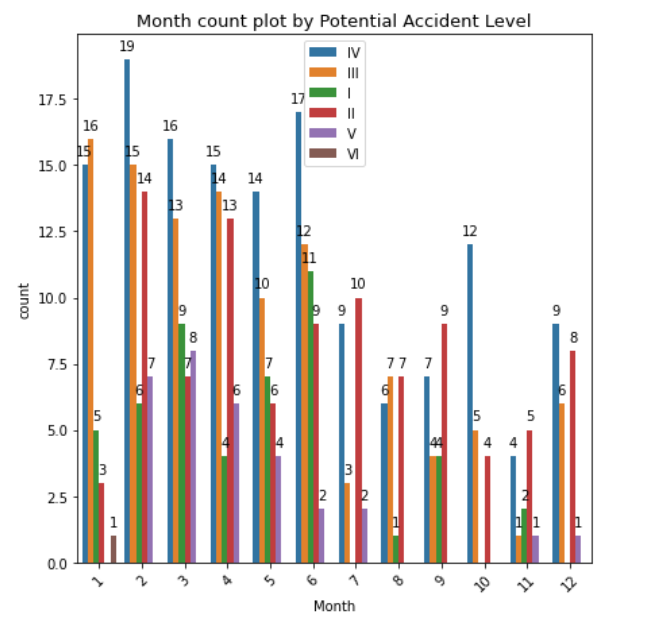
* Males are more prone to PAL IV accidents
* Females are more prone to PAL II accidents



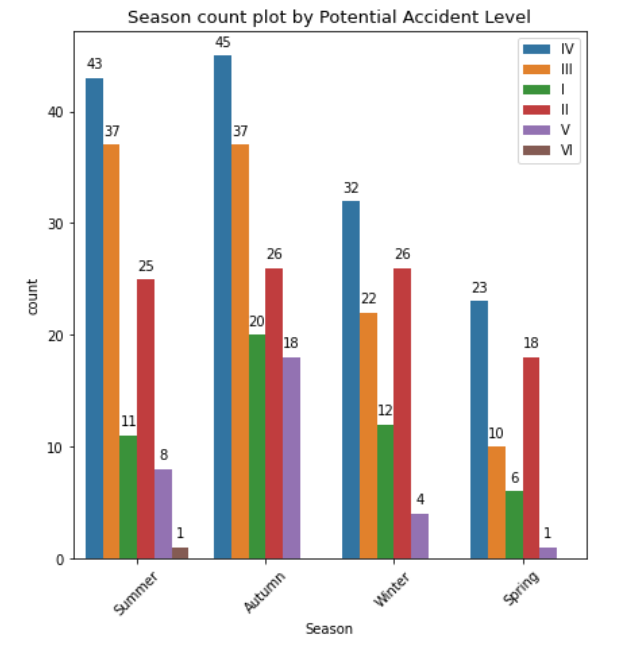
* Mining sector has the highest count of accidents and the PAL IV and III are the highest.
* Metals has the second highest count of accidents and PAL II and III are the highest.
* For others, PAL I is the highest.



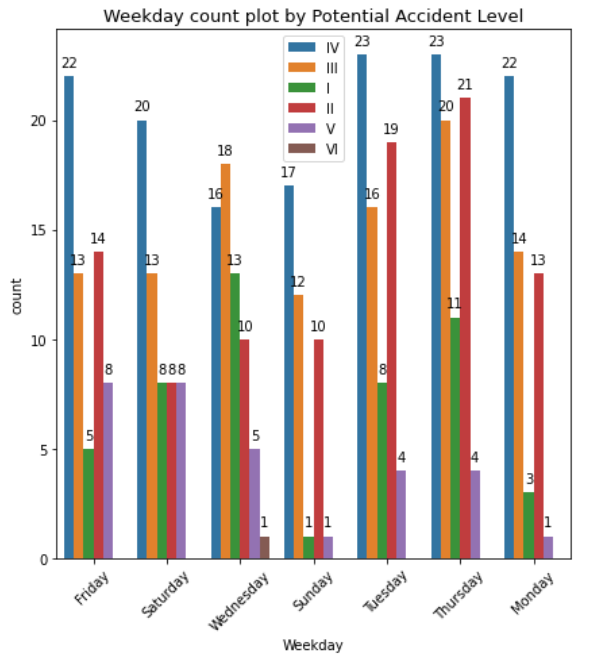
* For both Third party and Employees, PAL IV has the highest count



* Most of the months have PAL IV as the highest level of accidents and PAL V as the least level of accidents

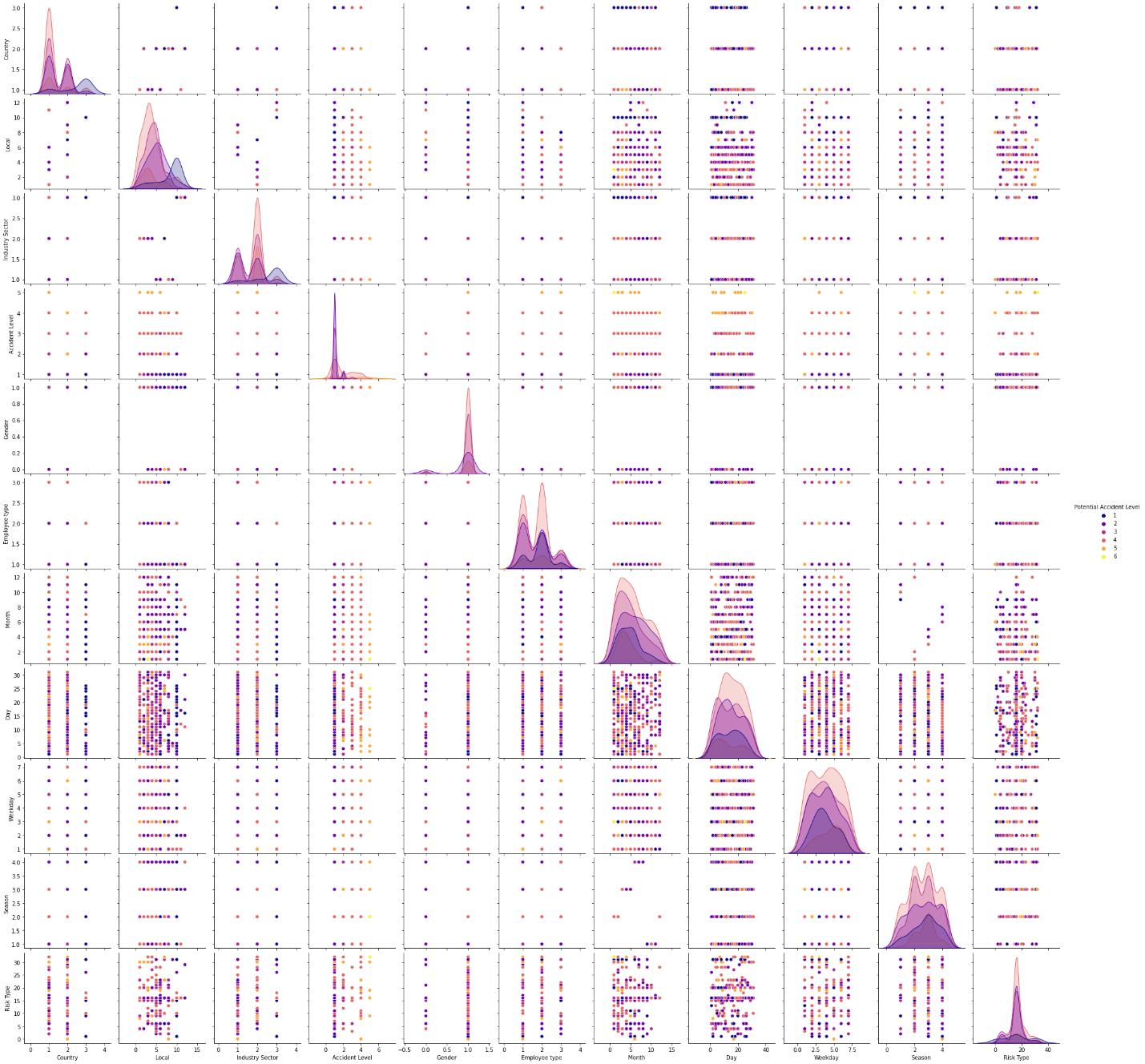


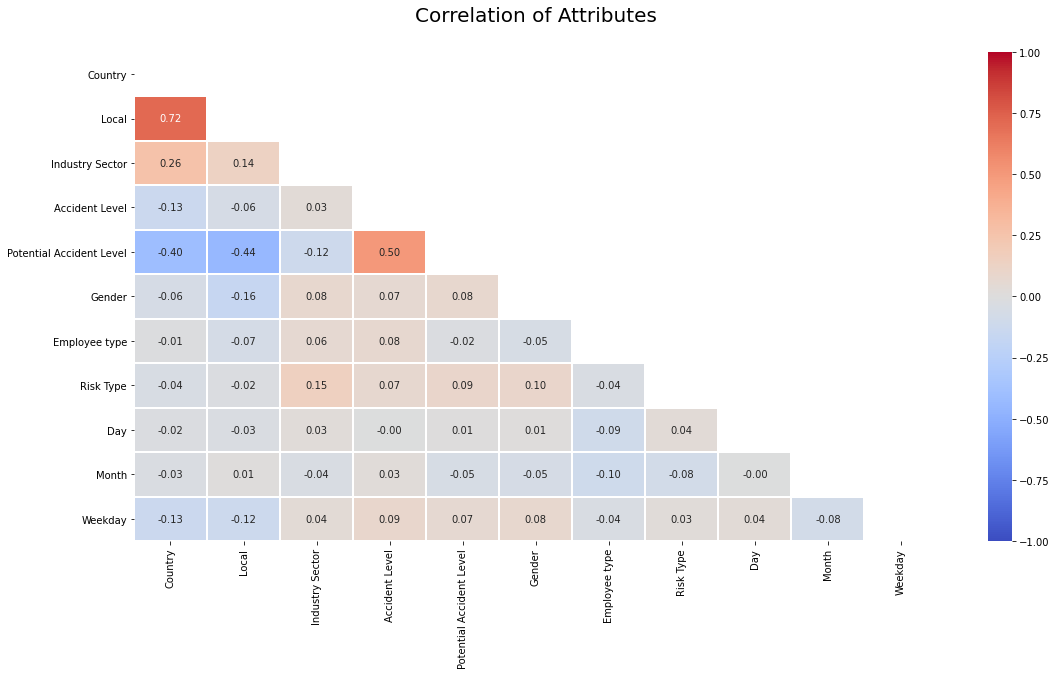
* PAL IV is the highest across all seasons
* Distribution of PAL levels is similar across seasons

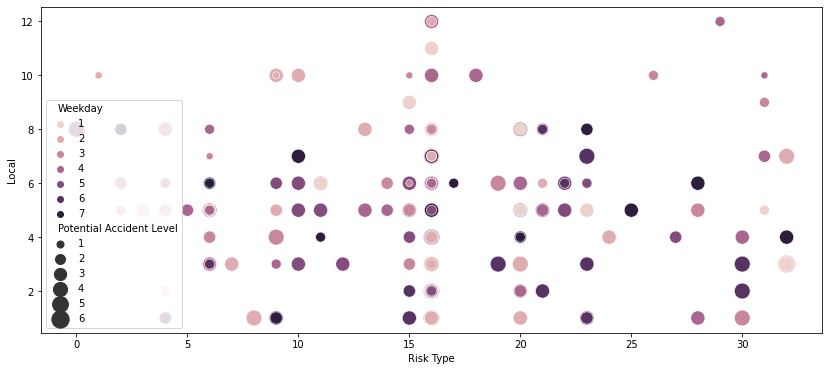


* PAL IV is the highest across the week except Wednesday

## 4.4 Multivariate Analysis







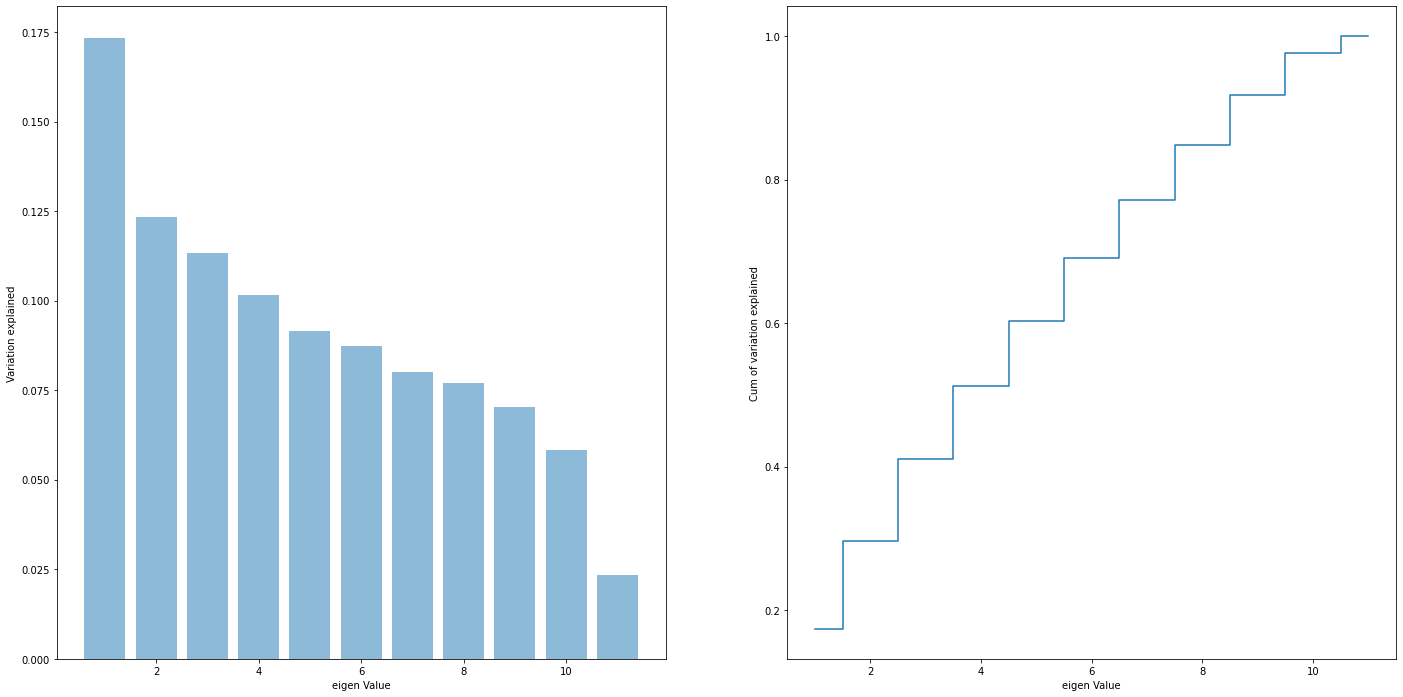
1. Country\_01 recorded the highest (59.1%) number of accidents. This is not necessarily significant as the employees in Country\_01 could be more than other two Countries
2. Male Gender recorded 94.8% of accidents compared to Female Gender with 5.2%, with no gender ratio of overall employees, this may not be significant. Could be number of Male employees are much more than Female employees
3. 54.6% of the accidents fall in Others category of Critical Risk, this could be due to limitations on valid options on data input/collection
4. Third Party (44.5%) and Employee (42.1%) are prone to overall 87% of accidents
5. Of the 12 Localities(cities), 3 localities (Local\_09, Local\_11 and Local\_12) record very low accidents compared to other Localities. Could be these Localities have highly safe workspace. Or there could be lesser number of employees in these localities. Possibility of no proper reporting of accidents in these Localities
6. Accident Level categorized to 6 Levels (I to VI), with I as 'not severe' and VI as 'very severe'. No accidents reported under Level VI. Level V (8) is reported with least number of accidents and 75% of accidents reported under Level I
7. Potential Accident Level categorized to 6 Levels (I to VI), with I as 'not severe' and VI as 'very severe'. 1 accident reported under Level VI and Level IV (143) is reported with more number of accidents
8. Months of February, March, April and June record the most accidents and Thursday record the highest number of accidents

# 5. Feature Engineering and PCA

After doing the EDA and looking at the insights from EDA, we want to check the feature importance using PCA and Decision tree and also the accuracy of Decision tree classifier on the given data set.

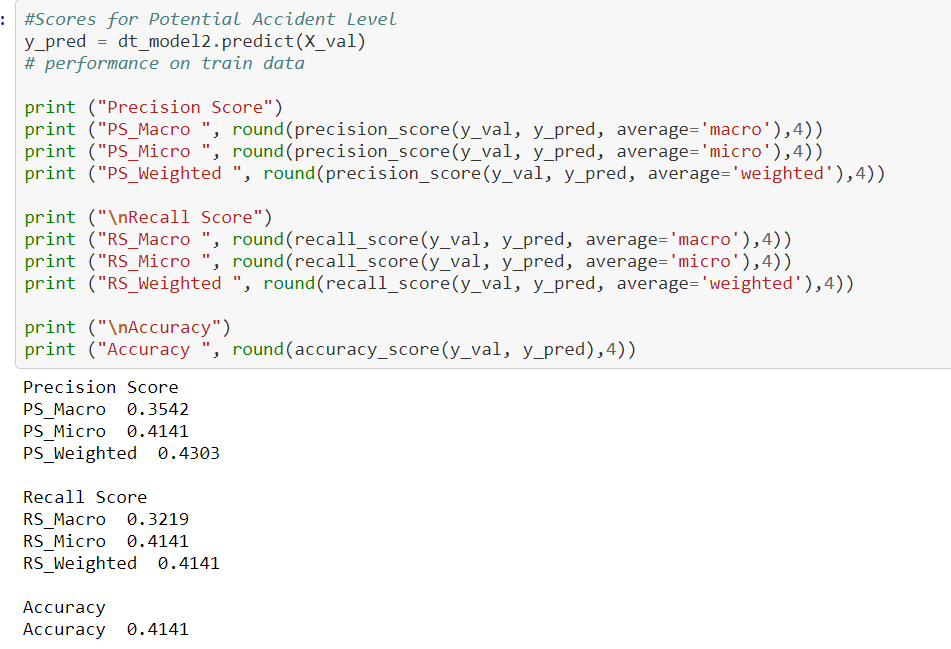
**PCA:**

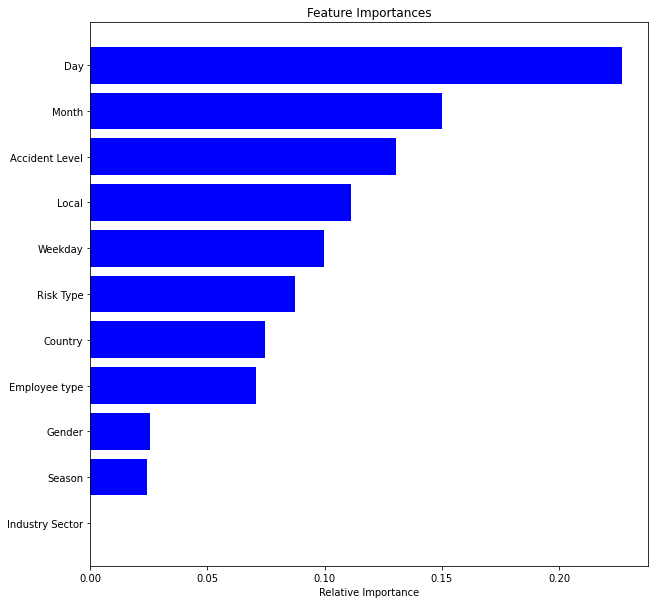
From PCA, we observed that the principle components do not explain the target to a great extent and the components have the explained variation ratio of 0.17 or less.

****

**Decision Tree:**

Decision tree classifier model also does not produce good accuracy.





Looking at the results of PCA and decision tree, the qualitative/categorical columns do not appear to be great predictors of Potential Accident Level. Hence, we proceed with NLP pre-processing and NLP feature engineering based on Description column. Hereafter, we consider the prediction of Potential accident Level as a text classification problem based on Description of the accident.

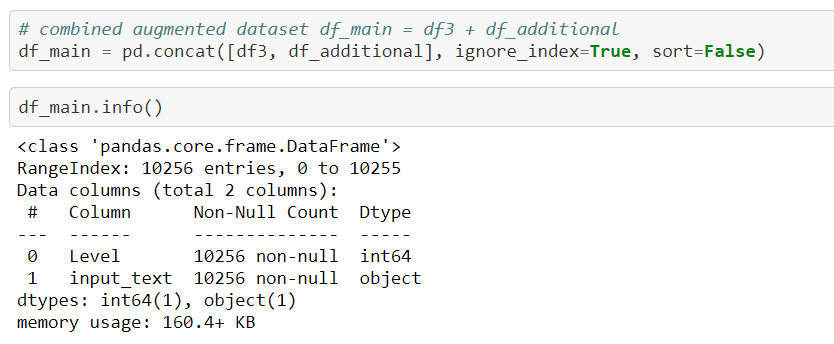
PS: This also resonates with the Project Objective given to us – “*Design a ML/DL based chatbot utility which can help the professionals to highlight the safety risk as per the incident description*.”

# 6. Data Augmentation

The given dataset contains only 424 data points which is not sufficient for Deep learning algorithms. Therefore, we took approach to augment the data set with more records.

We took the “Severe Injury Reports” data set which is available in public domain - <https://www.osha.gov/severeinjury>. This data set contains accident and injury details from industrial setup which is similar to data in our project. 2 columns in this dataset are of interest for our scenario. Column name “Final Narrative” is same as our descriptions. And column name “Hospitalized” has 6 levels which is similar to our potential accident level.

We augmented our given data set of 10K records with two features. i.e. potential accident level and descriptions.



# 7. NLP Pre-processing

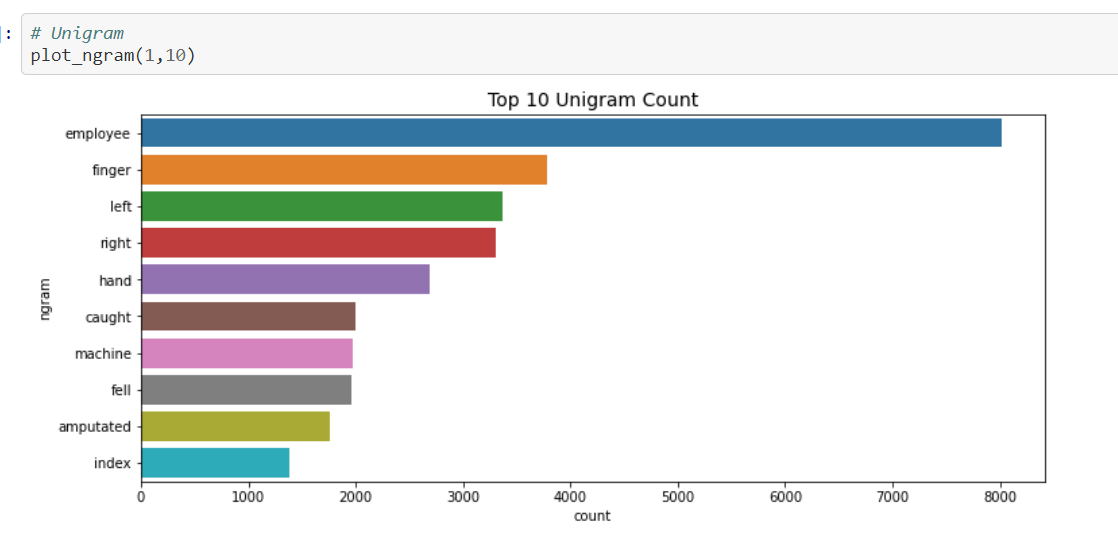
For NLP preprocessing, we used NLTK and RE libraries and performed the following-

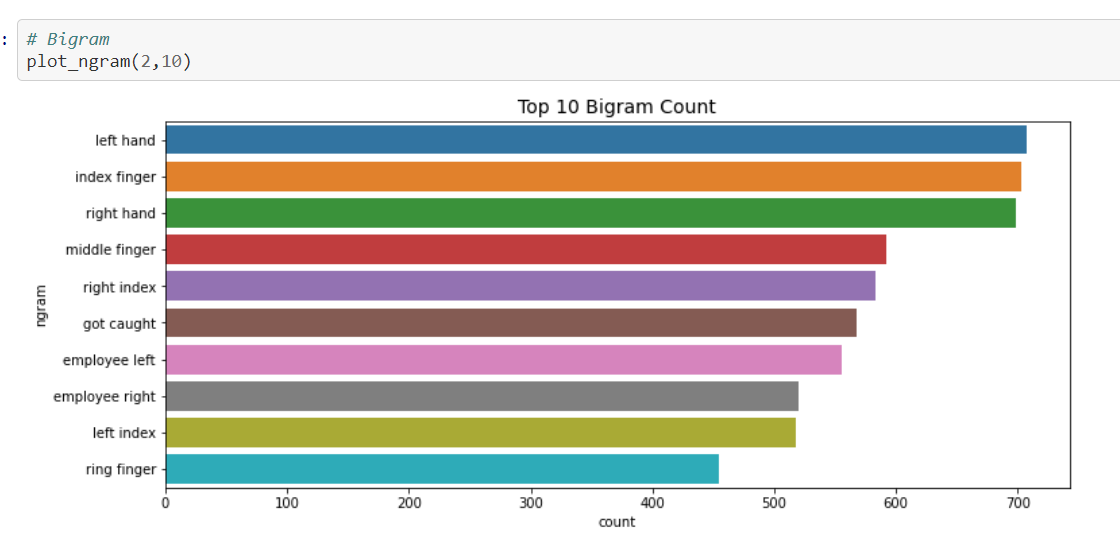
* Removal of stop words
* Converting to accented characters to Unicode
* Remove punctuations
* Remove unwanted space and new lines
* Convert words to lower cases
* Lemmatization

After our data is cleaned, we exported our csv “df\_main\_capstone.csv” which we will use in our next step for NLP feature engineering and model building.

# 8. NLP n-gram Analysis

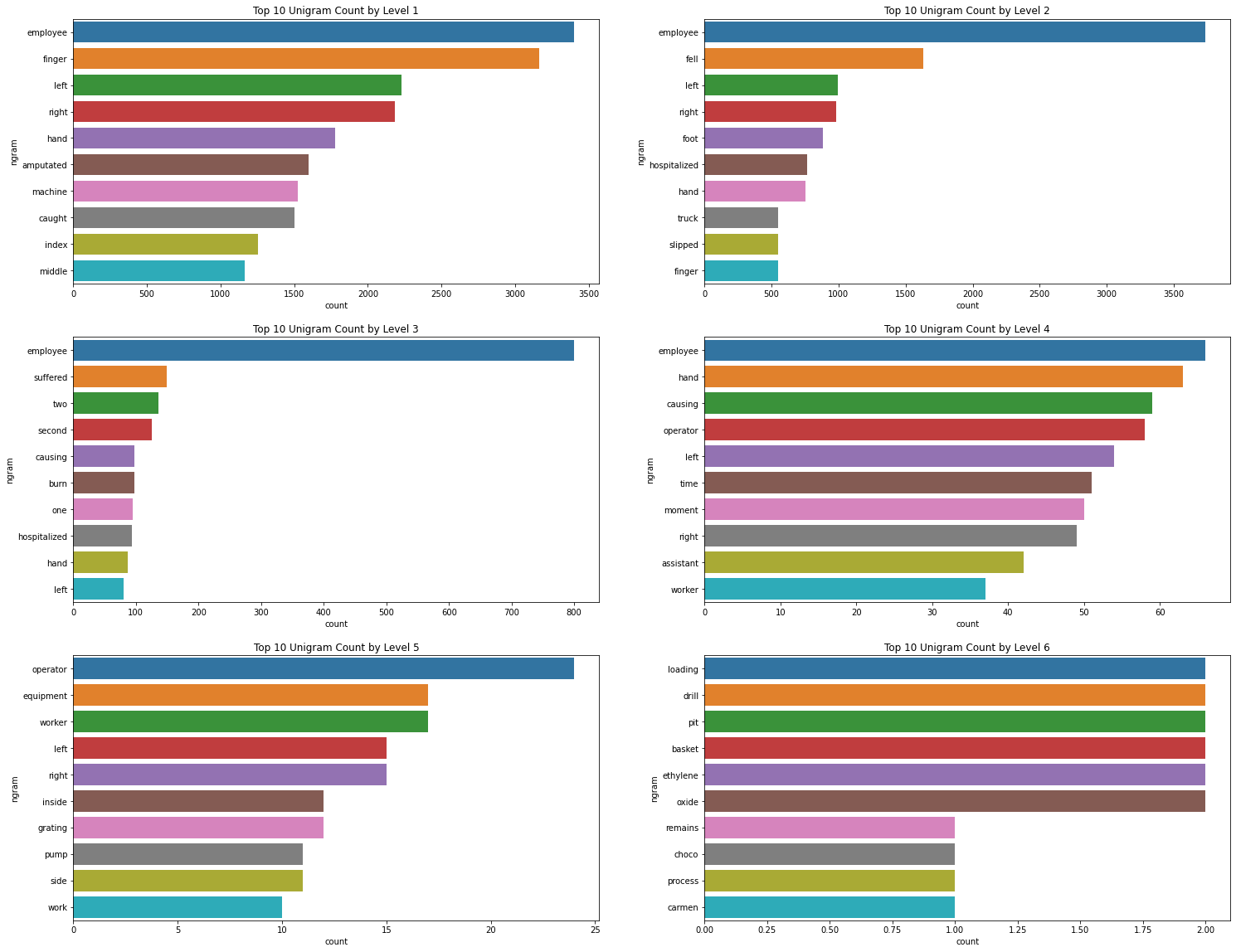
Firstly , analysis of n-grams :



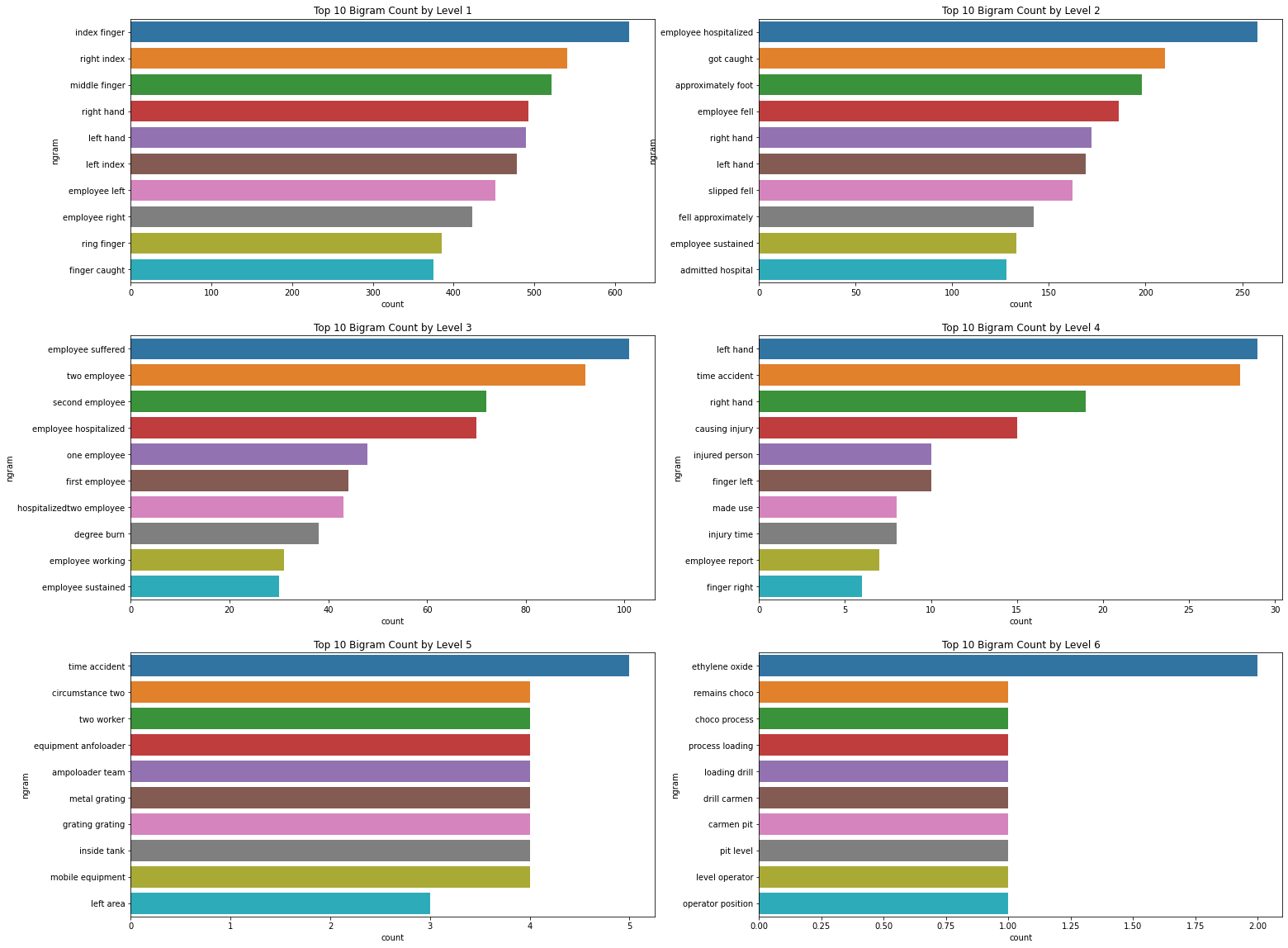




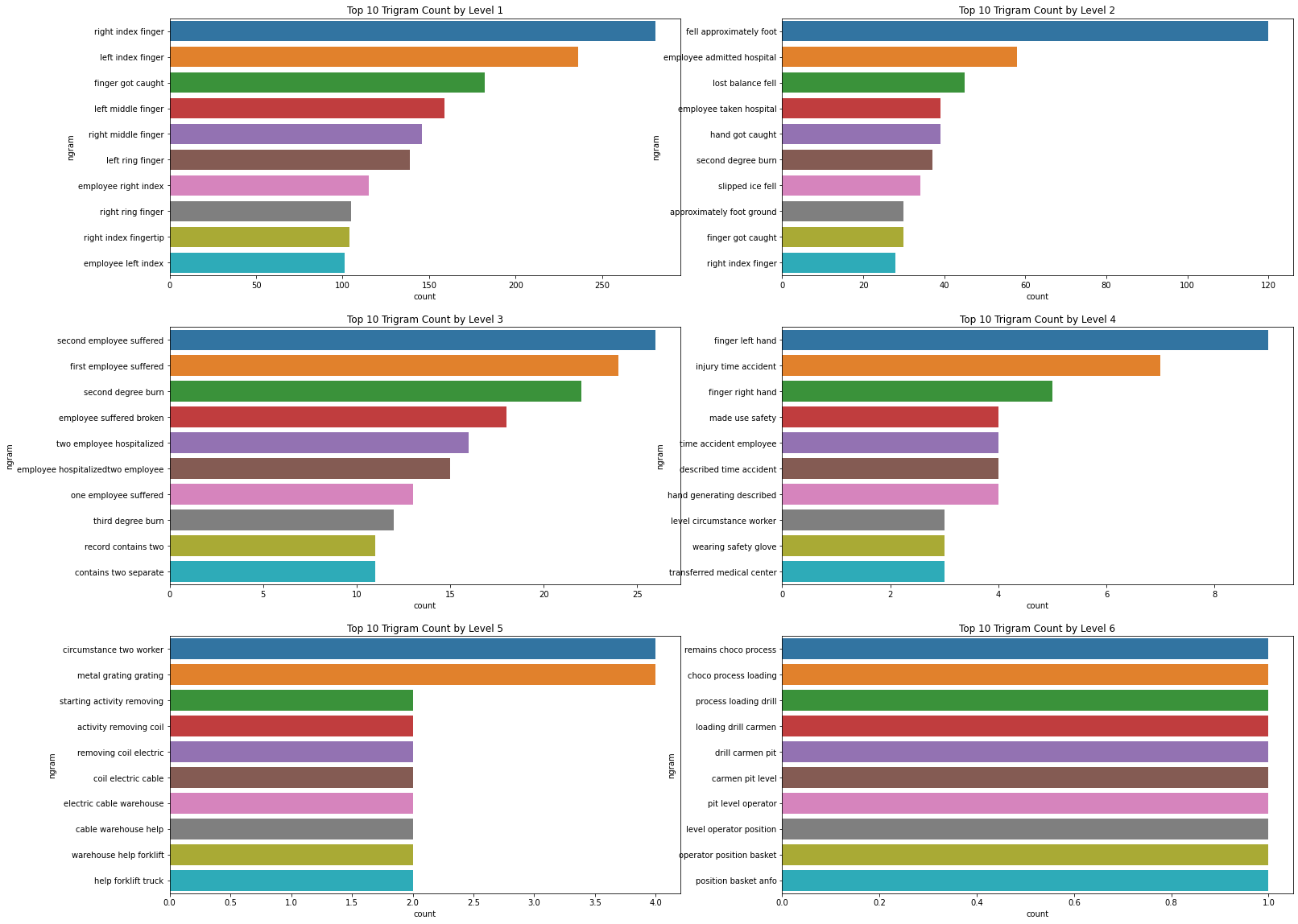
Second, Analysis of n-grams per Level



Analysis of bigram by individual levels



Analysis of trigram by levels,



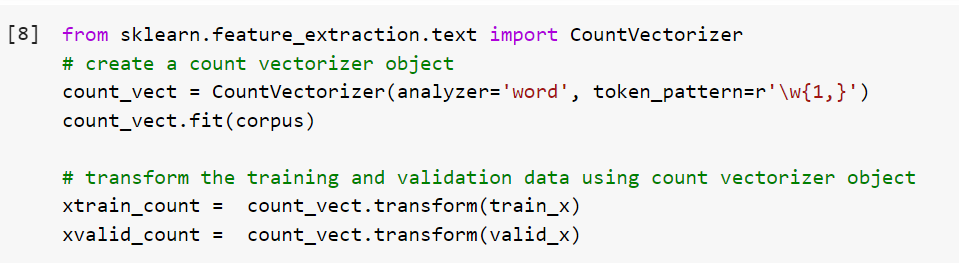
# NLP Feature Engineering

For NLP problem, we are converting text to vectors. We are considering count vectors, TFIDF vectors, Word 2 Vector using genism , Stanford Glove Vectors:

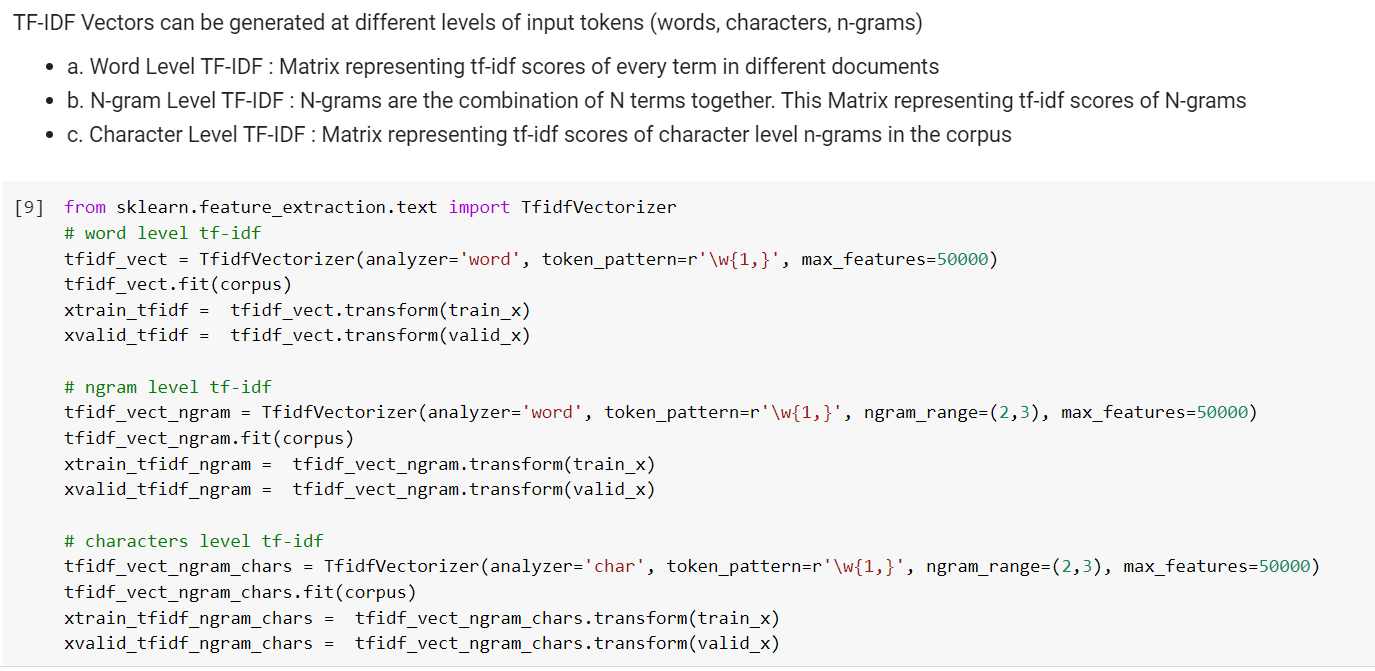
Before doing that, we are loading text and label as two columns. We also create a corpus of our text. We split the train and validation set.

## 9.1 Count vectors

Using CountVectorizer for creating xtrain\_count and xvalid\_count:

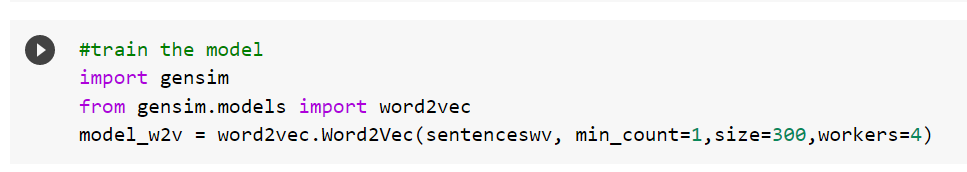


## 9.2 TF-IDF vectors

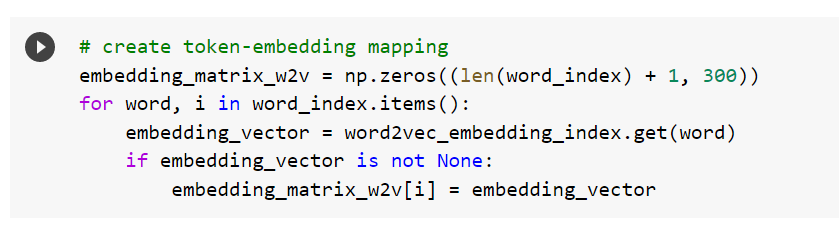


## 9.3 Word2Vec

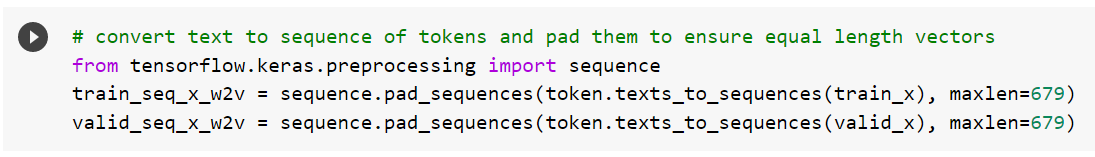
Using genism model to create word2vec:



Using subsequent steps create word2vec embedding matrix to be used in neural networks:

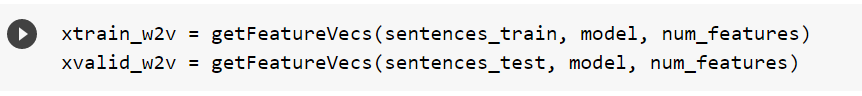


Created padded sequence for loading in neural network in upcoming steps:



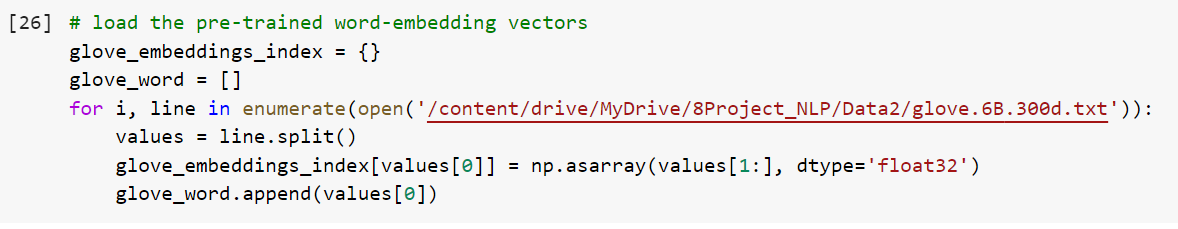
We have used custom functions and word2vec embeddings to create training and test sets. Please refer to notebook for coding steps.

Thus xtrain\_w2v and xvalid\_w2v for machine learning models have been created.

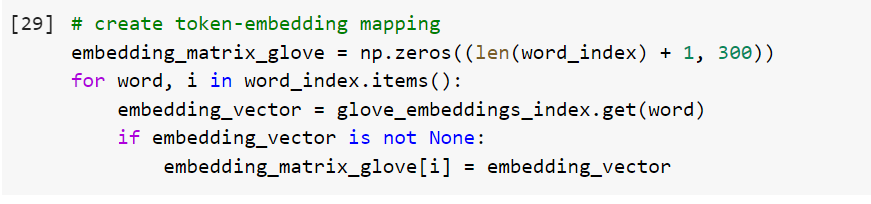


## 9.4 Glove

For glove we use the same steps as Word2vec section. Only difference is we are using Stanford glove embeddings for this purpose.



Using subsequent steps create Glove embedding matrix to be used in neural networks:

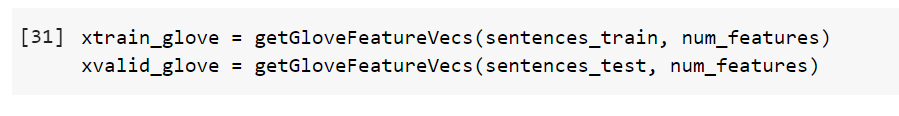


Created padded sequence for loading in neural network in upcoming steps:



**We have used custom functions and glove embeddings to create training and test sets. Please refer to notebook for coding steps.**

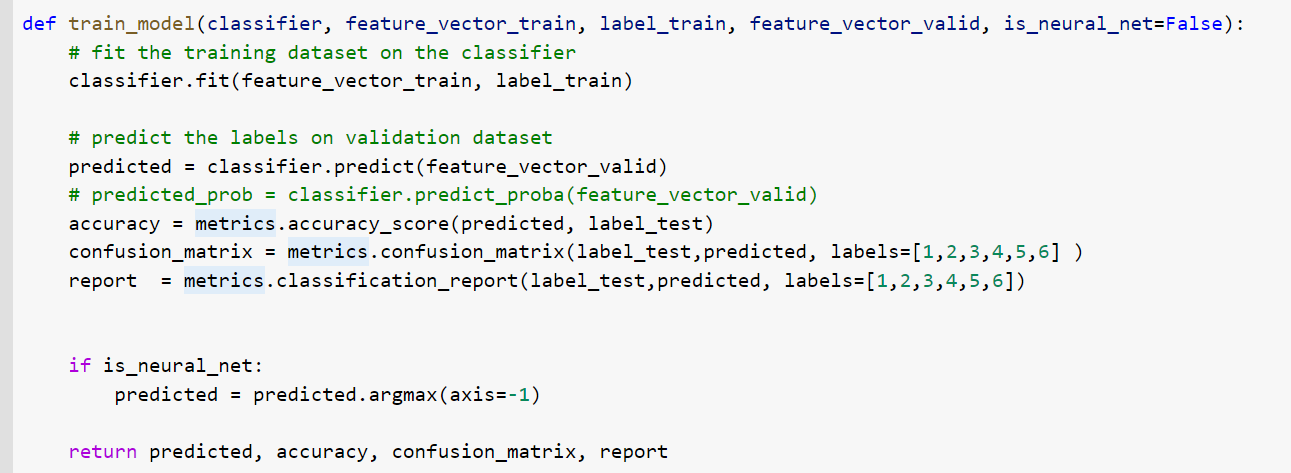
Thus xtrain\_glove and xvalid\_glove for machine learning models have been created.



## 10.Model Building

We will implement following different classifiers for this purpose:

Before creating our machine learning models, we have created a train\_model function which takes our training features and returns the output values.



From next page we will see the model building results:

## 

## 10.1 Naive Bayes Classifier

Accuracy summary for NB is as follows:

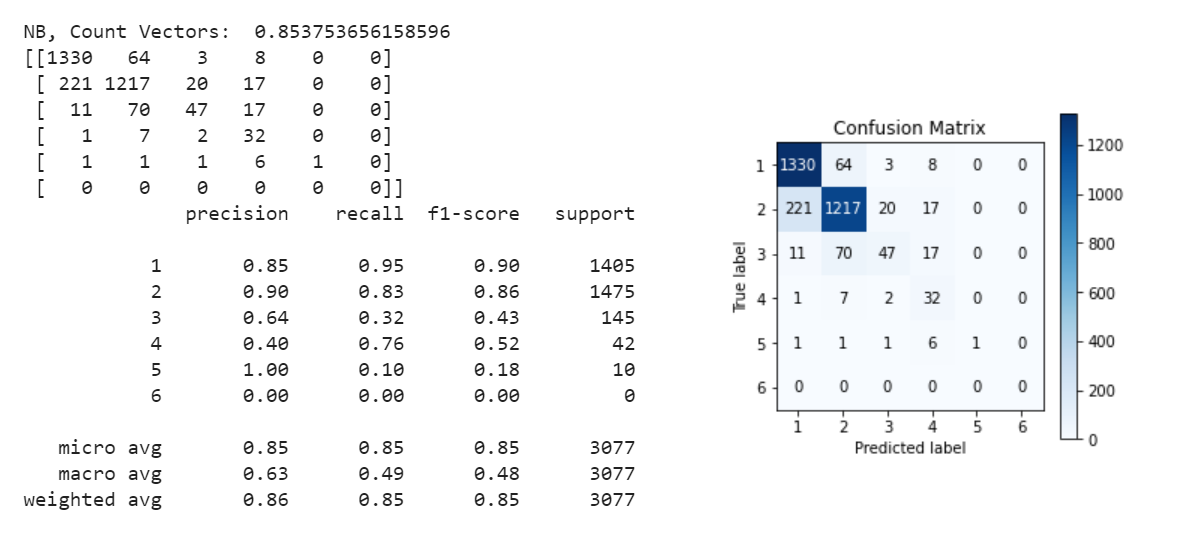
NB, Count Vectors: 0.853753656158596

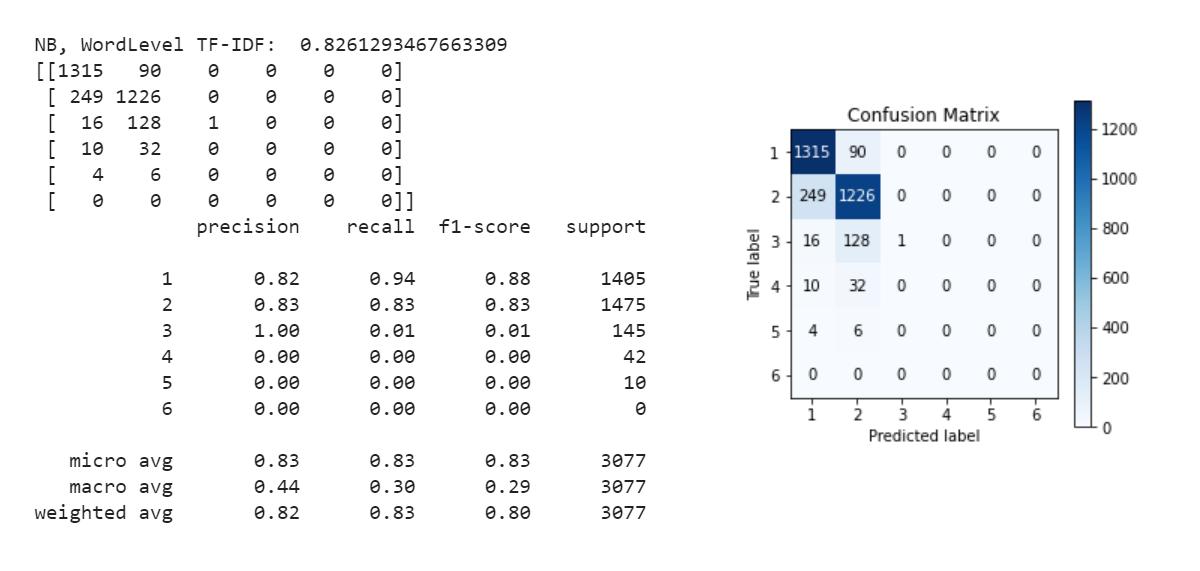
NB, WordLevel TF-IDF: 0.8261293467663309

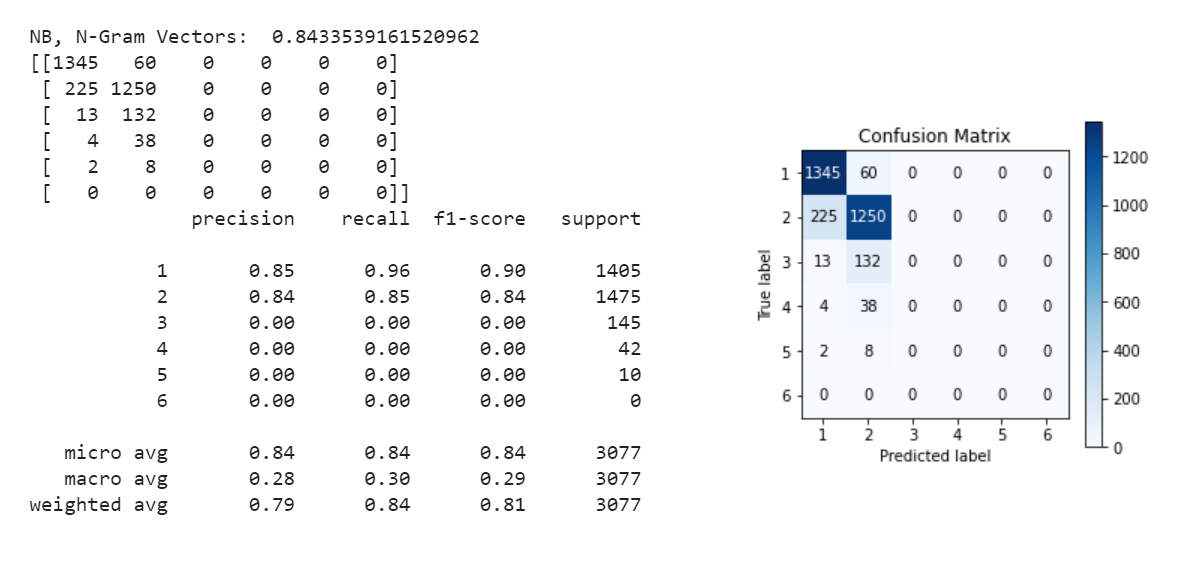
NB, N-Gram Vectors: 0.8433539161520962

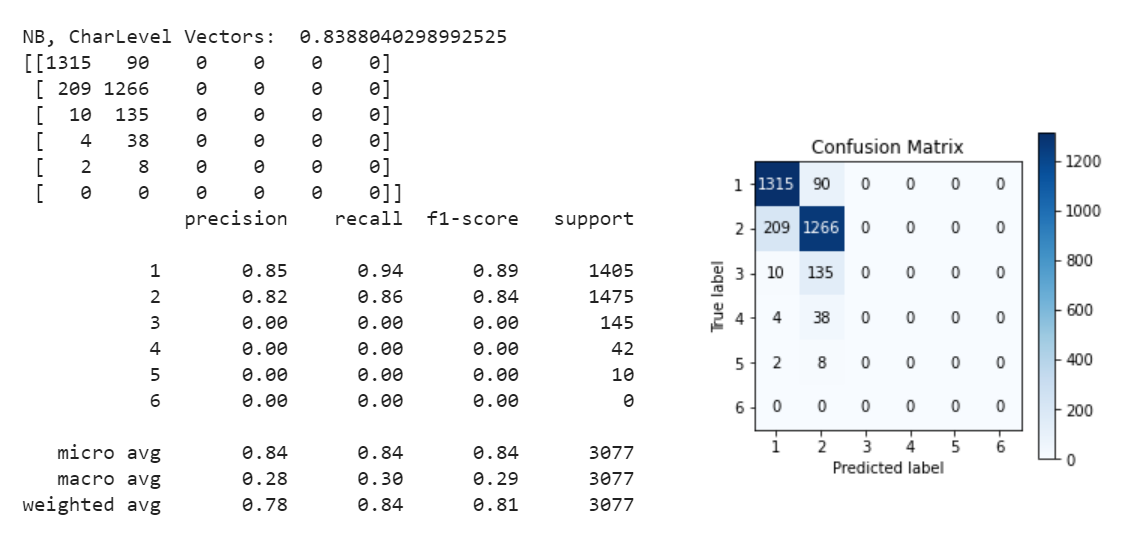
NB, CharLevel Vectors: 0.8388040298992525

Results, confusion Matrix and Classification report in following screenshots:









Please Note: Naïve Bayes cannot be used with matrix factorization hence we are not using them with word2vec or glove feature sets

## -

## 10.2 Linear Classifier

Accuracy summary for LR is as follows:

LR, Count Vectors: 0.8765030874228145

LR, WordLevel TF-IDF: 0.8674033149171271

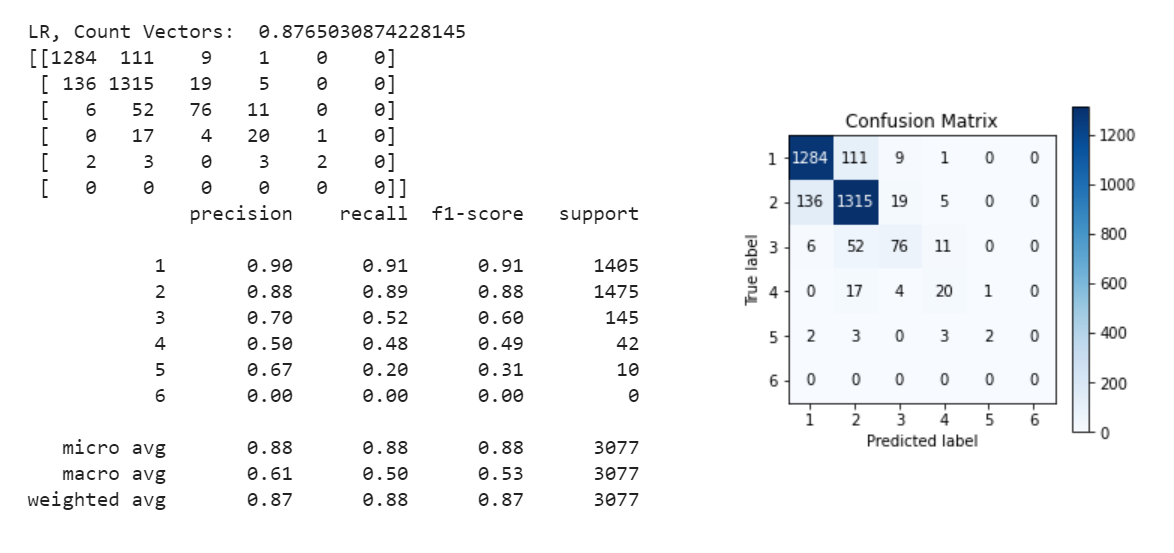
LR, N-Gram Vectors: 0.8609034774130647

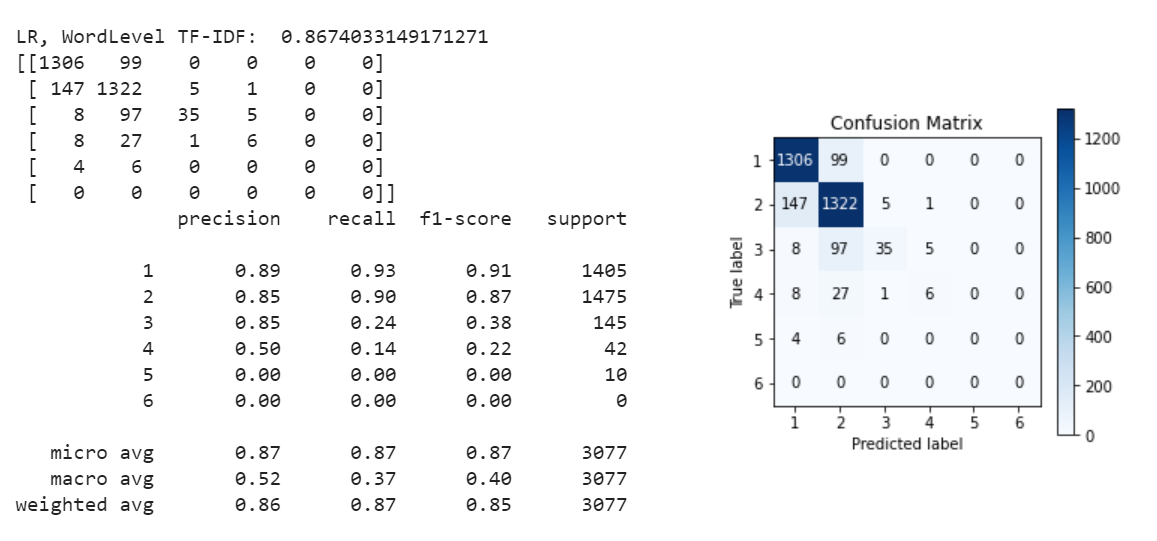
LR, CharLevel Vectors: 0.8735781605459864

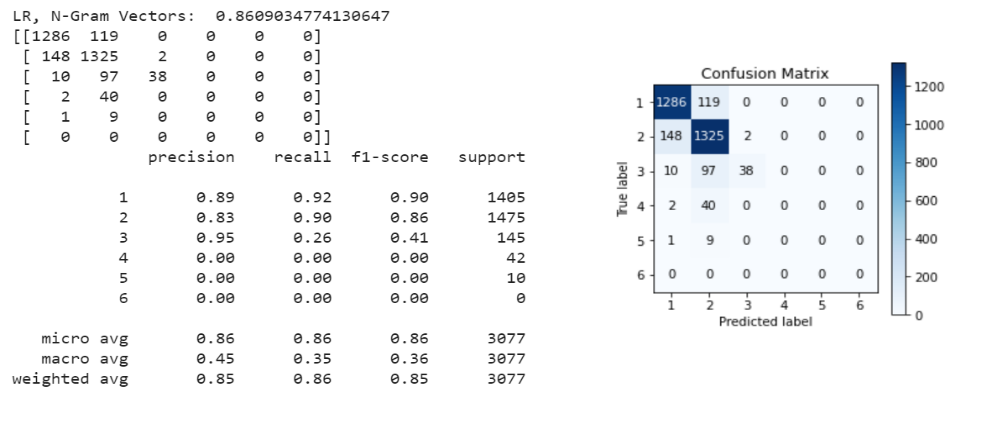
LR, Gensim Word2vec: 0.8215794605134872

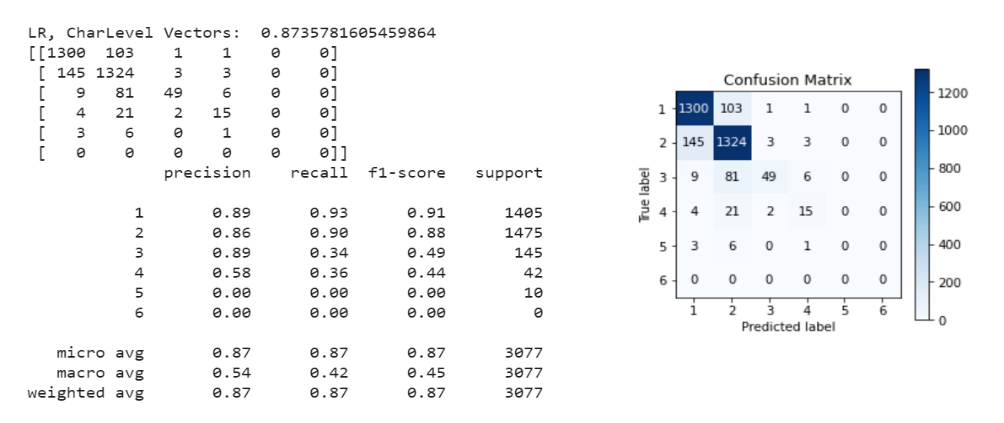
LR, GloVe vector: 0.867078323041924

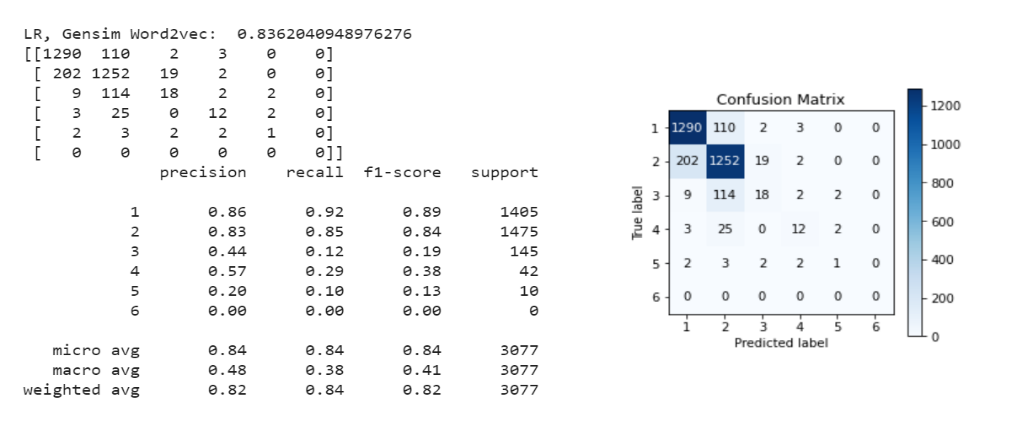
Results, confusion Matrix and Classification report in following screenshots:

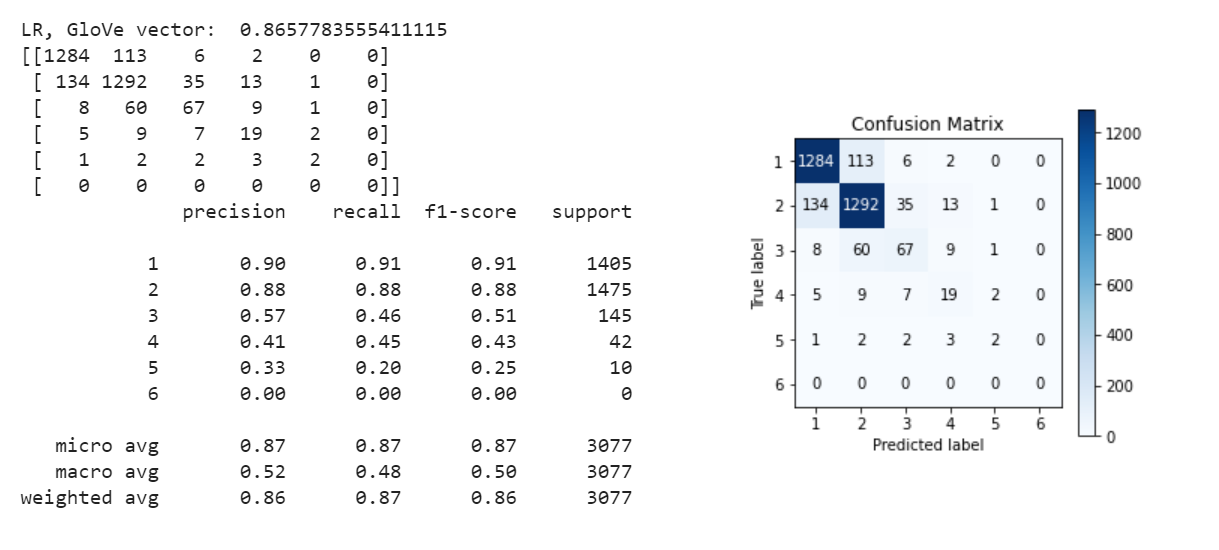












## 10.3 Support Vector Machine

Accuracy summary for SVM is as follows:

SVM, Count Vectors: 0.8810529736756582

SVM, WordLevel TF-IDF: 0.872603184920377

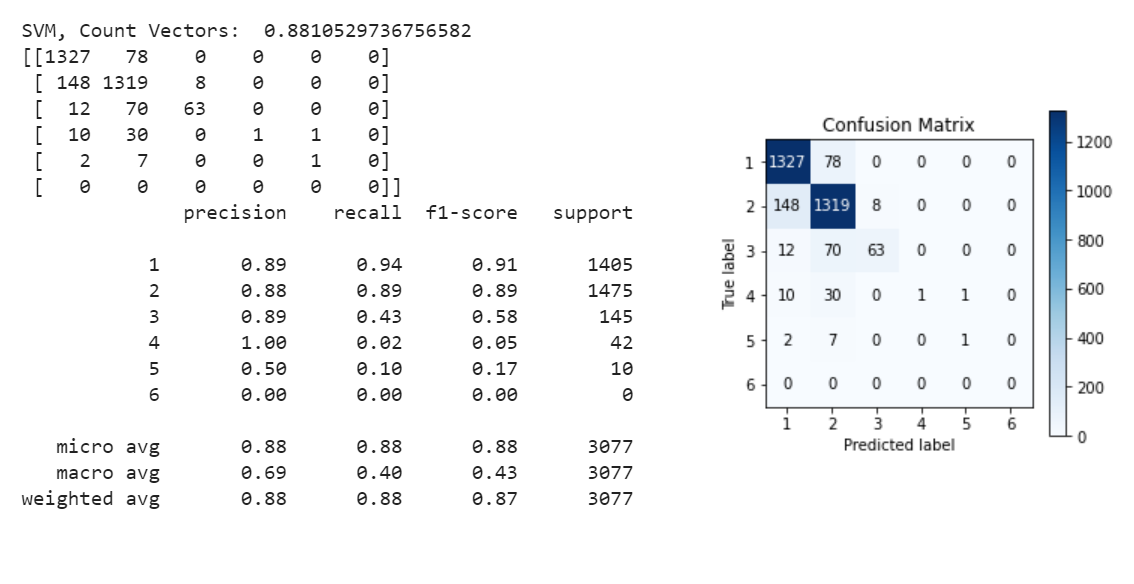
SVM, N-Gram Vectors: 0.8648033799155022

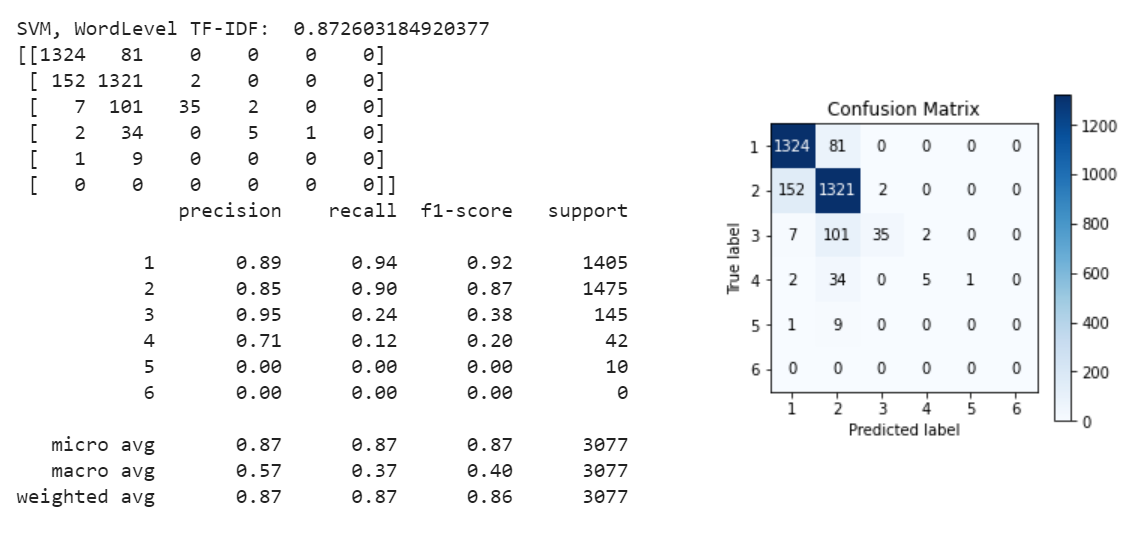
SVM, CharLevel Vectors: 0.8865778355541112

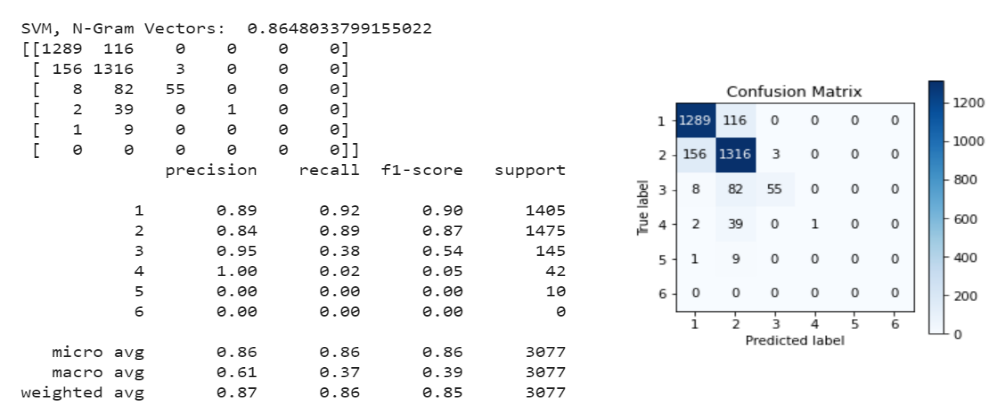
SVM, Gensim Word2vec: 0.823854403639909

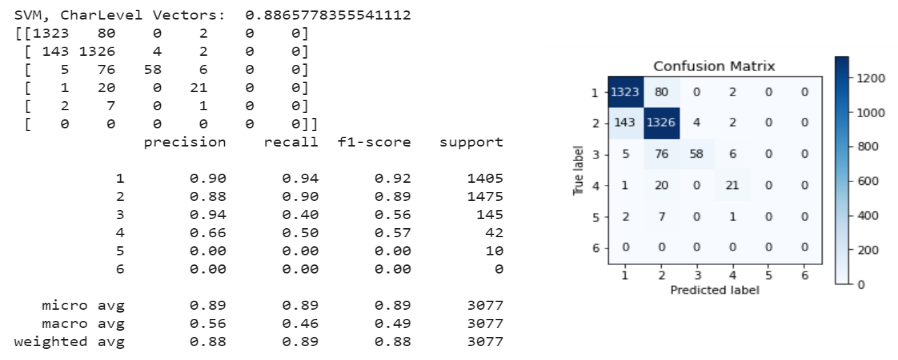
SVM, GloVe vector: 0.8713032174195645

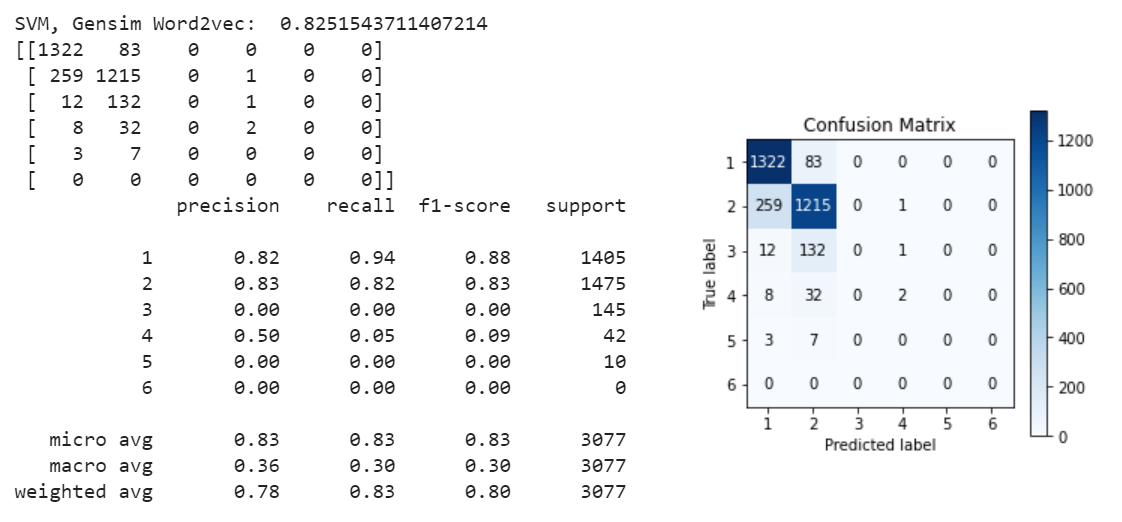
Results, confusion Matrix and Classification report in following screenshots:

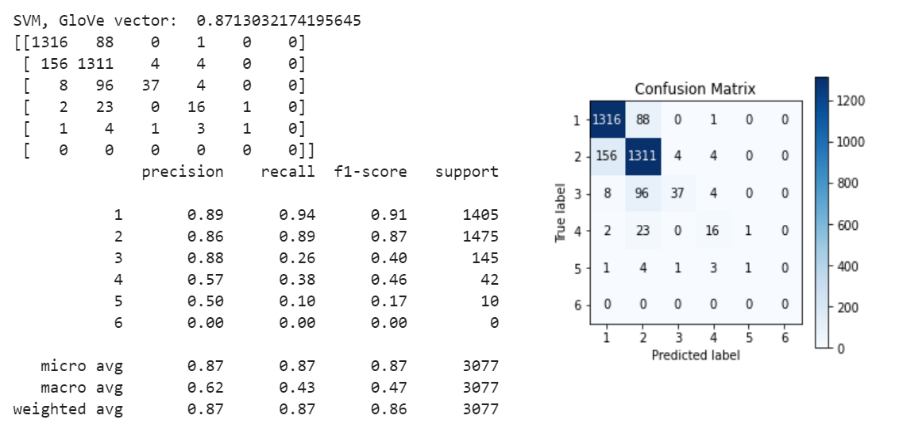












## 10.4 Bagging Models

Accuracy summary for Random Forest is as follows:

Random Forest, Count Vectors: 0.8674033149171271

Random Forest, WordLevel TF-IDF: 0.8690282742931427

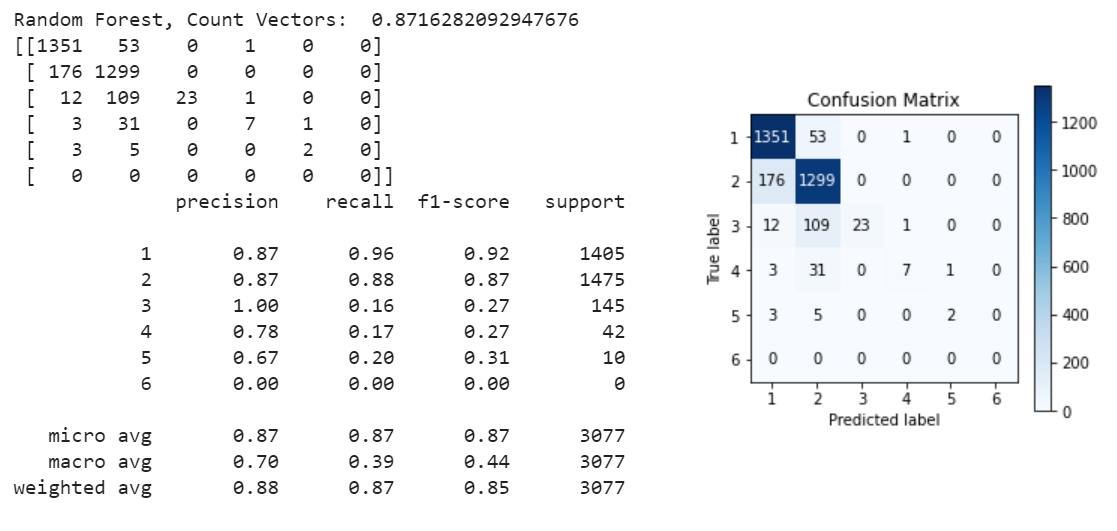
Random Forest, N-Gram Vectors: 0.864478388040299

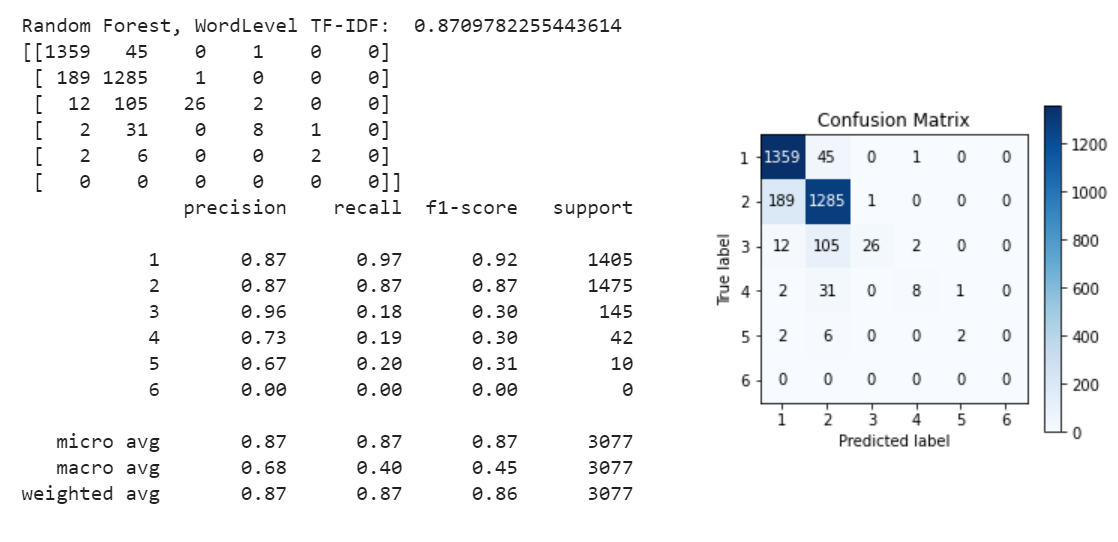
Random Forest, CharLevel Vectors: 0.8771530711732207

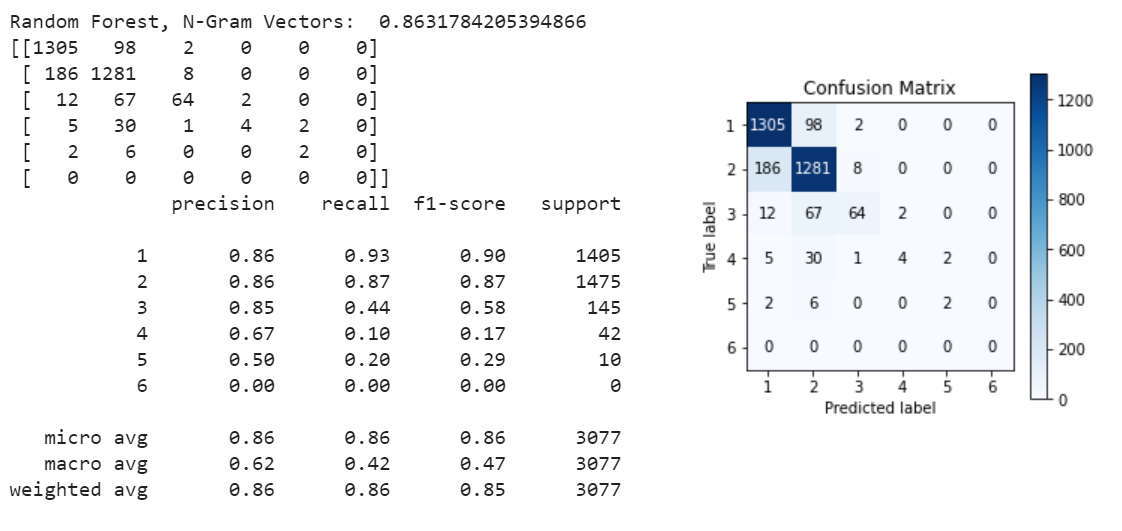
Random Forest, Gensim Word2vec: 0.8316542086447839

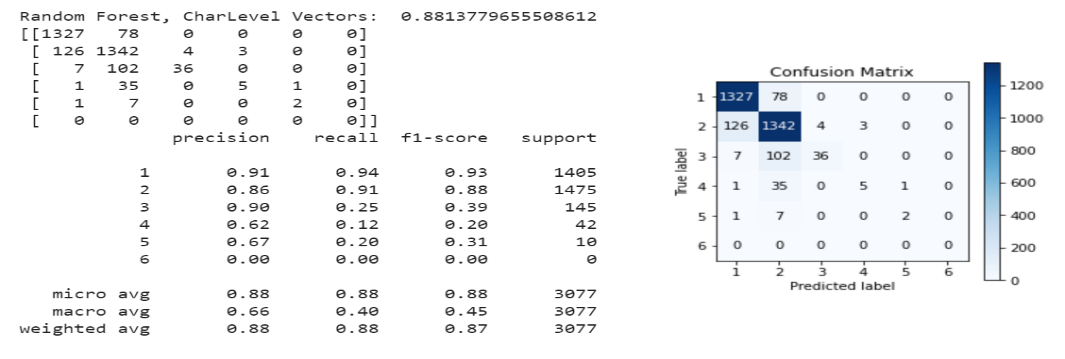
Random Forest, GloVe vector: 0.8355541111472213

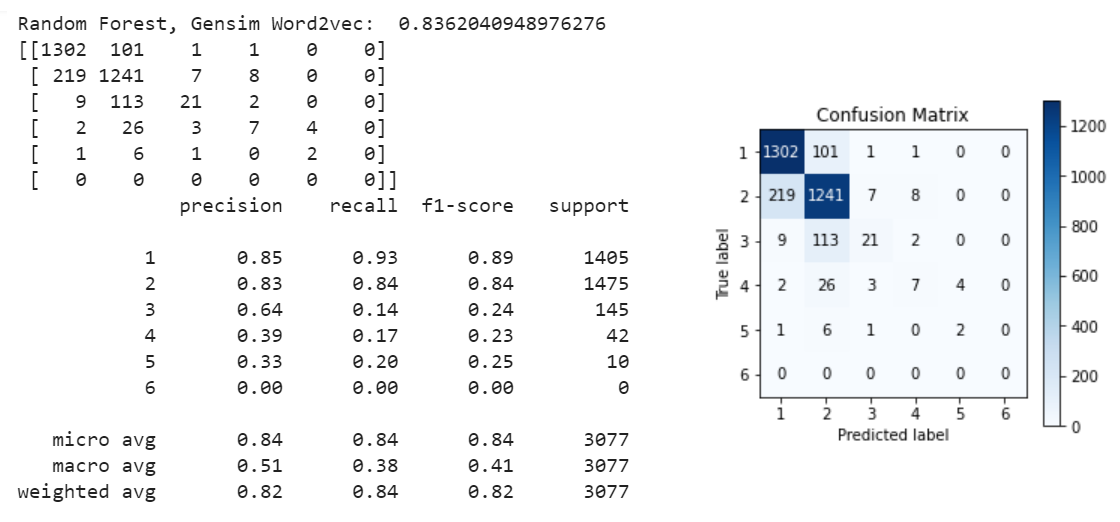
Results, confusion Matrix and Classification report in following screenshots:

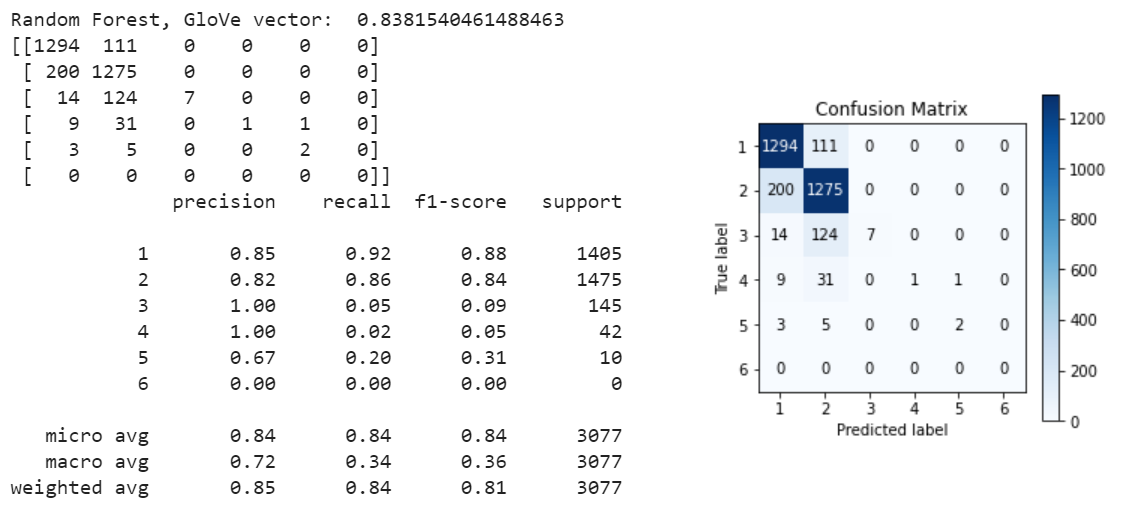












## 10.5 Boosting Models

Accuracy summary for XGB is as follows:

XGB, Count Vectors: 0.8820279493012675

XGB, WordLevel TF-IDF: 0.8843028924276893

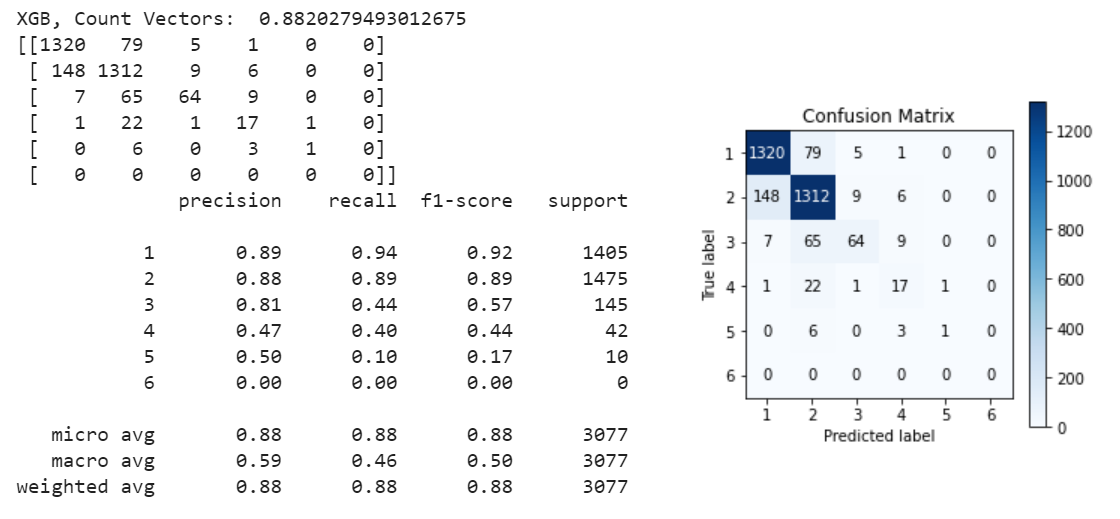
XGB, N-Gram Vectors: 0.8332791680207995

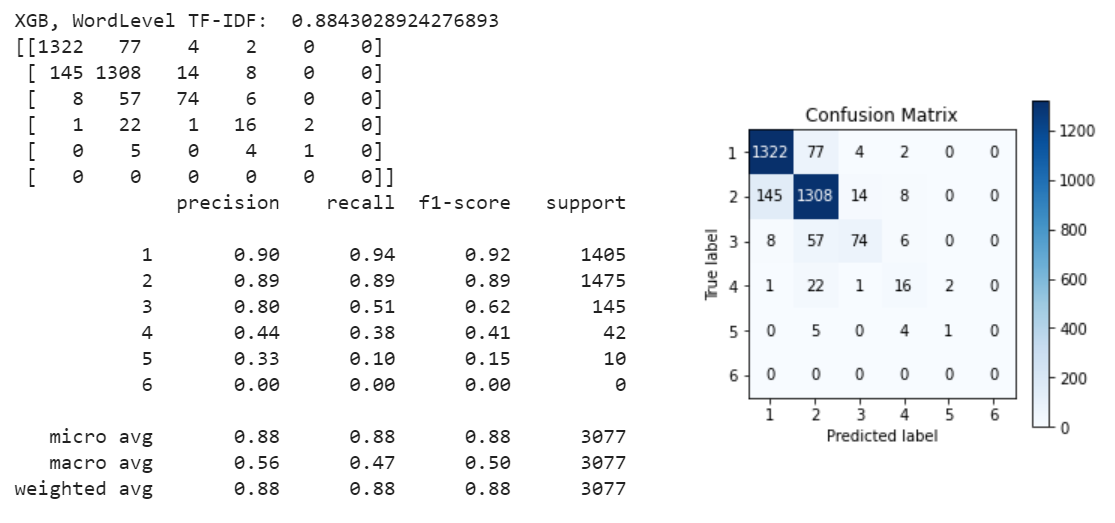
XGB, CharLevel Vectors: 0.8901527461813454

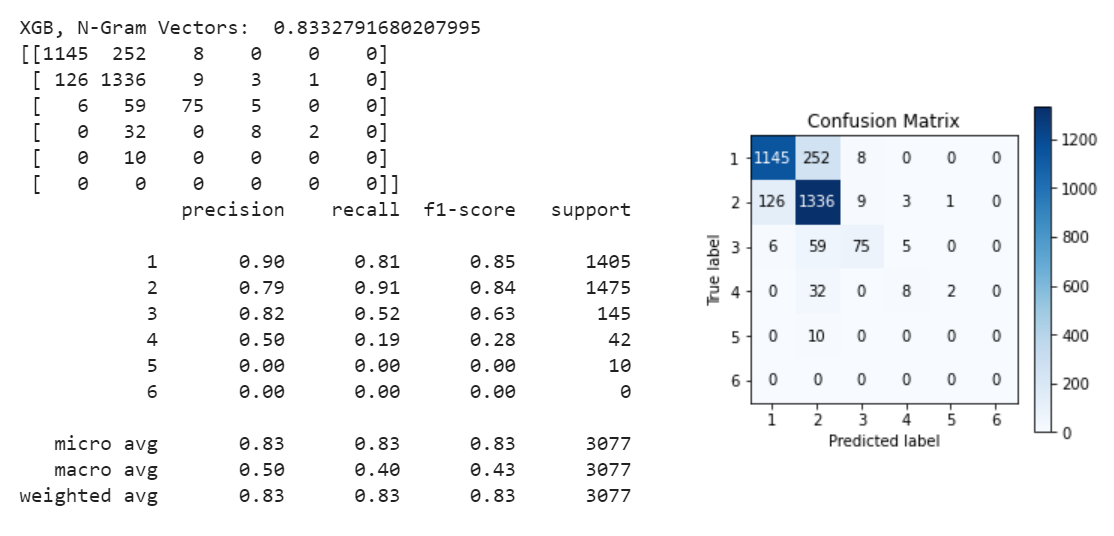
XGB, Gensim Word2vec: 0.8332791680207995

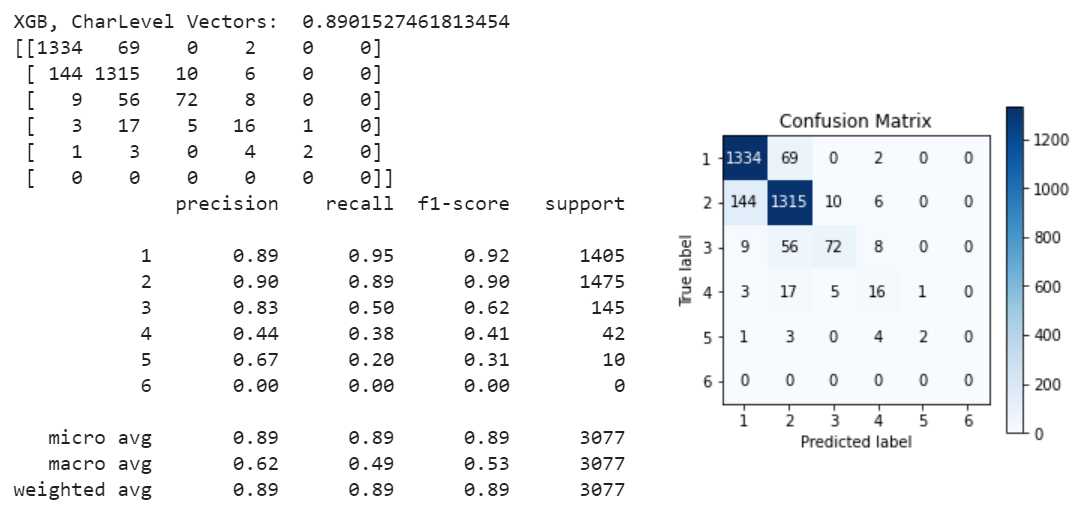
XGB, GloVe vector: 0.8498537536561586

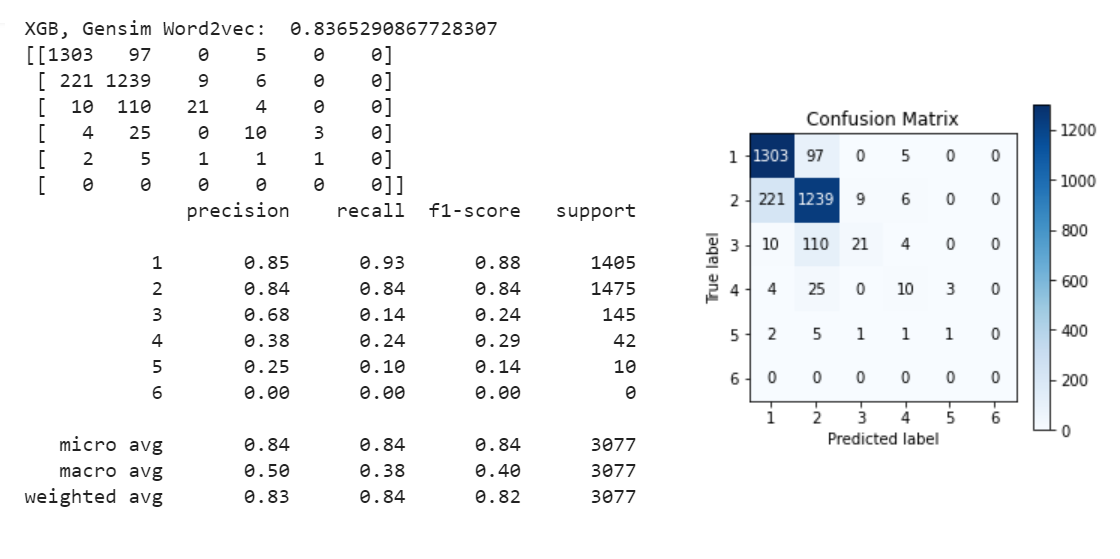
Results, confusion Matrix and Classification report in following screenshots:













## 10.6 Basic Neural Networks

## 10.7 Deep Neural Networks

## 10.7.1 Convolutional Neural Network (CNN + RNN)

## 10.7.2 Long Short-Term Model (LSTM)

## 10.7.3 Gated Recurrent Unit (GRU)