**INTERIM REPORT**

**AI Enabled IT Ticketing Service Tool**

*Submitted By*

*Disha Palan*

*Gloria Kamal*

*Parita Desai*

*Vivek UImale*

Table of Contents

[**1.** Summary of problem statement, data and findings 3](#_Toc74988490)

[**Problem Statement** 3](#_Toc74988491)

[**Objective** 4](#_Toc74988492)

[Observations from the given Dataset 4](#_Toc74988493)

[**2.** Overview of the final process 4](#_Toc74988494)

[Observations from Target Class 4](#_Toc74988495)

[Data Pre-processing 4](#_Toc74988496)

[**3.** Step-by-step walk through the solution 5](#_Toc74988497)

[**Data** 5](#_Toc74988498)

[Summary of the Approach to EDA and Pre-processing 6](#_Toc74988499)

[EDA 6](#_Toc74988500)

[**Findings** 8](#_Toc74988501)

[Other findings 10](#_Toc74988502)

[Further Data Analysis 11](#_Toc74988503)

[Word cloud for Short description 11](#_Toc74988504)

[Word cloud for Description 12](#_Toc74988505)

[Assignment group distribution of top 10 callers as compared to other callers 12](#_Toc74988506)

[Assignment group distribution amongst top 10 callers 13](#_Toc74988507)

[Assignment group distribution 13](#_Toc74988508)

[Feature Engineering 15](#_Toc74988509)

[Data Pre-processing 16](#_Toc74988510)

[Data Cleaning 16](#_Toc74988511)

[NER and POS Tagging 16](#_Toc74988512)

[Deciding Models and Model Building 17](#_Toc74988513)

[Models 18](#_Toc74988514)

[Traditional Models 18](#_Toc74988515)

[Sequential Models 19](#_Toc74988516)

[How to improve your model performance? 20](#_Toc74988517)

[**4.** Comparison to benchmark 20](#_Toc74988518)

[**5.** Implications 21](#_Toc74988519)

[**6.** Limitations 21](#_Toc74988520)

[**7.** Closing Reflections 21](#_Toc74988521)

# Summary of problem statement, data and findings

## **Problem Statement**

In any of the IT industry, incident management plays an important role in delivering quality and timely support to its customers across the globe.

The incidents are generally created by various stakeholders like end users, vendors, IT users, etc. They might not have right information as to which team the ticket should go to. Hence, to improve and retain customer satisfaction, it is very important that the ticket is assigned to the right group of people for faster and appropriate resolution. In many Organizations this is still a manual process. There are few problems with the manual process:

1. Manual assignment of incidents is time consuming
2. It requires human efforts
3. There may be mistakes due to human errors and resource consumption is carried out ineffectively because of the misaddressing
4. Manual assignment increases the response and resolution times which result in user satisfaction deterioration / poor customer service

L1 / L2 needs to spend time to review Standard Operating Procedures (SOPs) before assigning to Functional teams (Minimum 25–30% of incidents needs to be reviewed for SOPs before ticket assignment).

15 mins are being spent for SOP review for each incident. Minimum of 1 FTE effort needed only for incident assignment to L3 teams.

During the process of incident assignments by L1 / L2 teams to functional groups, there were multiple instances of incidents getting assigned to wrong functional groups.

Around 25% of Incidents are wrongly assigned to functional teams. Additional effort needed for Functional teams to re-assign to right functional groups

During this process, some of the incidents are in queue and not addressed timely resulting in poor customer service and loss of business.

### **Objective**

We are building an AI solution which will enable organizations to classify incidents to the right functional group by implementing the best suited machine learning model and leading to customer satisfaction.

Guided by AI, organizations can reduce the resolution time and focus on more productive tasks. This will overcome and save time with below losses:

1. Time latency due to review of SOPs before assigning to right functional group
2. Incorrect assignments to functional groups

## Observations from the given Dataset

* Four columns – Short Description, Description, Caller and Assignment group
* 74 Assignment groups found - Target classes
* Caller names in a random fashion (may not be useful for training data)
* European non-English language also found in the data
* Email/chat format in description
* Symbols & other characters in the description
* Hyperlinks, URLS & few image data found in the description
* Blanks found either in the short description or description field
* Few descriptions same as the short description
* Few words were combined together
* Spelling mistakes and typo errors are found

# Overview of the final process

## Observations from Target Class

* The Target class distribution is extremely skewed
* A large no of entries for GRP\_0 (mounting to 3976) which account for ~50% of the data
* There are groups with 1 entry also. We could merge all groups with small entries to a group to reduce the imbalance in the target. This may reduce the imbalance to some extent.

## Data Pre-processing

Below steps have been performed for initial pre-processing and clean up of data.

* Dropped the caller field as the data was not found to be useful for analysis
* Replaced Null values in Short description & description with space.
* Merged Short Description & Description fields for analysis
* Contraction words found in the merged Description are removed for ease of word modelling
* Changed the case sensitivity of words to the common one
* Removed Hashtags and kept the words, Hyperlinks, URLs, HTML tags & non-ASCII symbols from merged fields.
* Translating all languages (German) to English
* Tokenization of merged data
* Removal of Stop words
* Lemmatization
* WordCloud created for all available 50 groups to have more information specific to Assignment groups
* Attempted to do spell check
* Created Plot to understand the distribution of words

# Step-by-step walk through the solution

## **Data**

Reference: <https://drive.google.com/open?id=1OZNJm81JXucV3HmZroMq6qCT2m7ez7IJ>

The given dataset has below four columns:

1. Short description
2. Description
3. Caller
4. Assignment group

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column | | Description | Data type | | |
| **Short description** | | Short description on the problem for which incident is being  raised | | 8492 | non-null object |
| **Description** | | Detailed description of the problem | | 8499 | non-null object |
| **Caller** | | Email id of the User who raised the problem | | 8500 | non-null object |
| **Assignment Group** | | IT Support Group to which the Incident log is been assigned to | | 8500 | non-null object |

Out of above four columns we have 3 features namely, short description, description and caller and one target group namely assignment group

The dataset is divided into two parts, namely, **feature matrix** and the **response vector**.

* Feature matrix contains all the vectors (rows) of dataset in which each vector consists of the value of **dependent features**. In above dataset, features are *Short description*, *Description* and *Caller*.
* Response vector contains the value of **class variable** (prediction or output) for each row of feature matrix. In above dataset, the class variable name is *Assignment group*.

Top 10 records of our dataset :

Graphical user interface, text

Description automatically generated

# Summary of the Approach to EDA and Pre-processing

## EDA

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to:

* + Maximize Insight Into A Data Set;
  + Uncover Underlying Structure;
  + Extract Important Variables;
  + Detect Outliers And Anomalies;
  + Test Underlying Assumptions;
  + Determine Optimal Factor Setting

Visually representing the content of a text document is one of the most important tasks in the field of text mining.

We have used SweetViz and Panda profiling to Visualize and analyse our dataset. Below are attached reports for and observations from reports are in next sections. The reports can be viewed in GitHub link - https://github.com/dishapalan02/AI-Enabled-IT-Service-Ticketing-tool/tree/main

**SweetViz**:

Graphical user interface, application

Description automatically generated

**Pandas Profiling:**

A picture containing chart

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

Chart, bar chart

Description automatically generated

## **Findings**

From Above two reports we have below observations:

1. Shape of the data - { Rows : 8500, Columns : 4 }
2. Total features - 3

2.1. Short Description - Text

2.2. Description - Text

2.3. Caller - Text

1. Target Column - 1

3.1 Assignment Group - Categorical

1. There are 84 duplicate records in total. Strategy to handle duplicates and the approach taken is defined in the pre-processing section below.
2. New features are required or not needs to be analysed further and also to check if below hidden patterns can be figured out:

   A. Common Issues -> user can be trained if possible   
   B. Common Caller -> May be user needs training or help with hardware or software  
   C. To find if customer is happy with service or needs further imporvement and assistance

Now let's have a look at individual features:

1. **Short description**

A. Total values - 8492 ( > 99% )

B. Missing values - 8 ( < 1% )

C. Distinct values - 7481 (88%)

D. Mostly occurring value - password reset ( 0.4% )

E. We can also see the number of times each value is being repeated

F. Max length of statement - 159

G. It contains:

 Characters -> Lowercase Letter, Punctuation,   
 Uppercase Letter, Decimal Number,   
 Math Symbol, Math Symbol,   
 Modifier Symbol, Other Number,   
 Other Symbol, Currency Symbol  
   
 Scripts -> Common(ASCII) and Latin

H. Point G indicates that we have to translate the texts in the dataset based on the scripts as part of data pre-processing.

1. **Description**

A. Total values - 8499 ( > 99% )

B. Missing values - 1 ( < 1% )

C. Distinct values - 7817 ( 92% )

D. Mostly occurring value - it shows "the" ( 0.7% ) but will analyse further after the removal of stop words. But we consider the next which is windows password reset ( 0.3% )

E. We can also see the number of times each value is being repeated

F. Max length of statement - 13001

G. It contains:

 Characters -> Lowercase Letter, Punctuation,   
 Uppercase Letter, Decimal Number,   
 Math Symbol, Math Symbol,   
 Modifier Symbol, Other Number,   
 Other Symbol, Currency Symbol  
   
 Scripts -> Common(ASCII) and Latin

H. Point G indicates that we have to translate the texts in the dataset based on the scripts as part of data pre-processing.

1. **Caller**

A. Total values - 8500 ( 100% )

B. Missing values - no missing value

C. Distinct values - 2950 ( 35% )

D. Mostly occurring value - bpctwhsn kzqsbmtp (10%)

E. We can also see the number of times each value is being repeated

F. Max length of statement - 30

G. It contains:

 Characters -> Lowercase Letter, Space Separator,   
 Uppercase Letter, Connector Punctuation  
   
 Scripts -> Common(ASCII) and Latin

H. Point G indicates that we have to translate the texts in the dataset based on the scripts as part of data pre-processing.

1. **Assignment Group**

A. Total values - 8500 ( 100% )

B. Missing values - no missing value

C. Distinct values - 74

D. Mostly occurring value - GRP\_0 (47% ~ nearly half of the data --> Hence we can say that target class is highly imbalanced, so needs a strategy to be employed to reduce the bias here)

E. We can also see the number of times each value is being repeated

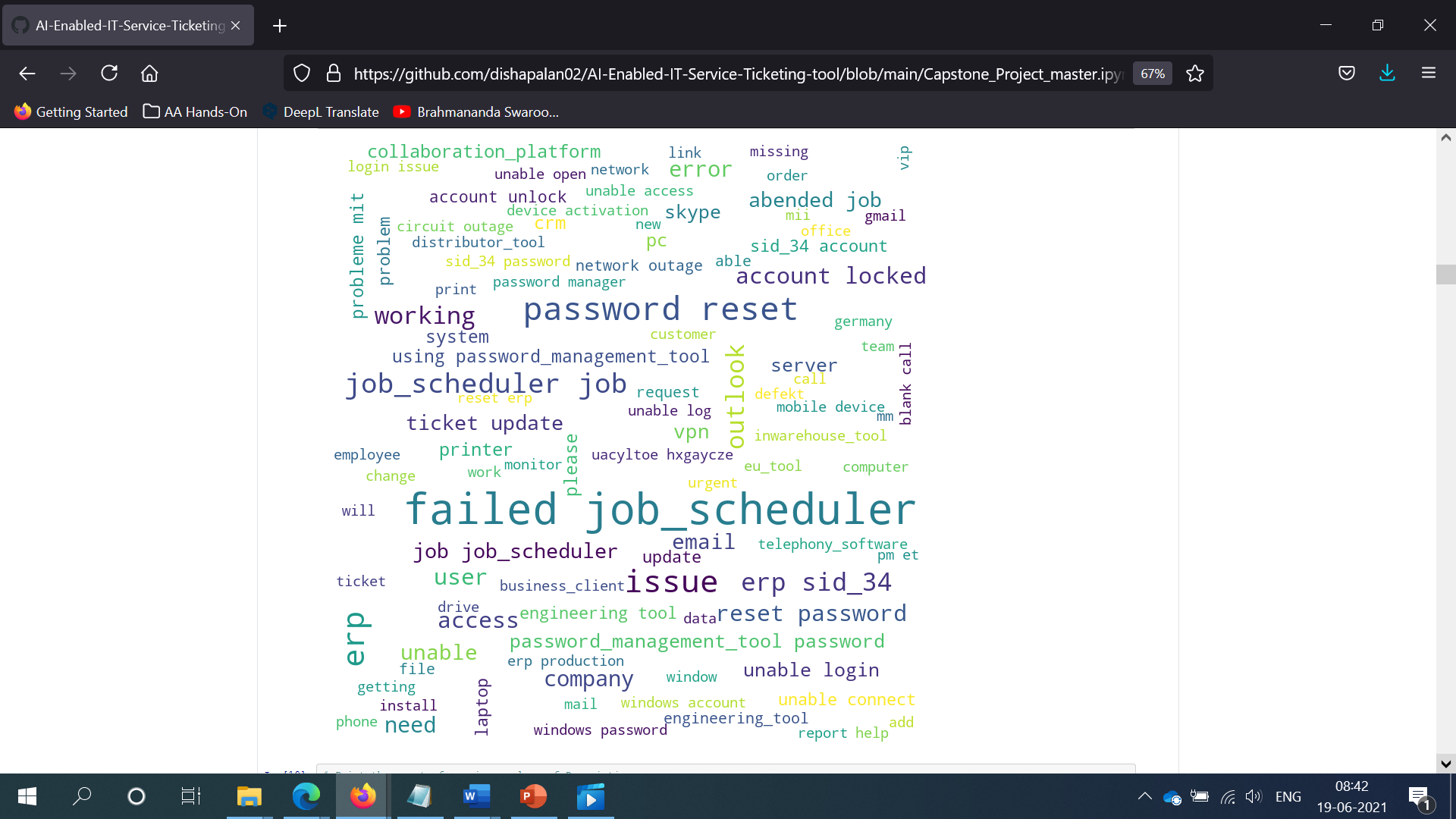
G. This indicates we can merge few assignment groups with smaller percentage to reduce overall number of categories.

### Other findings

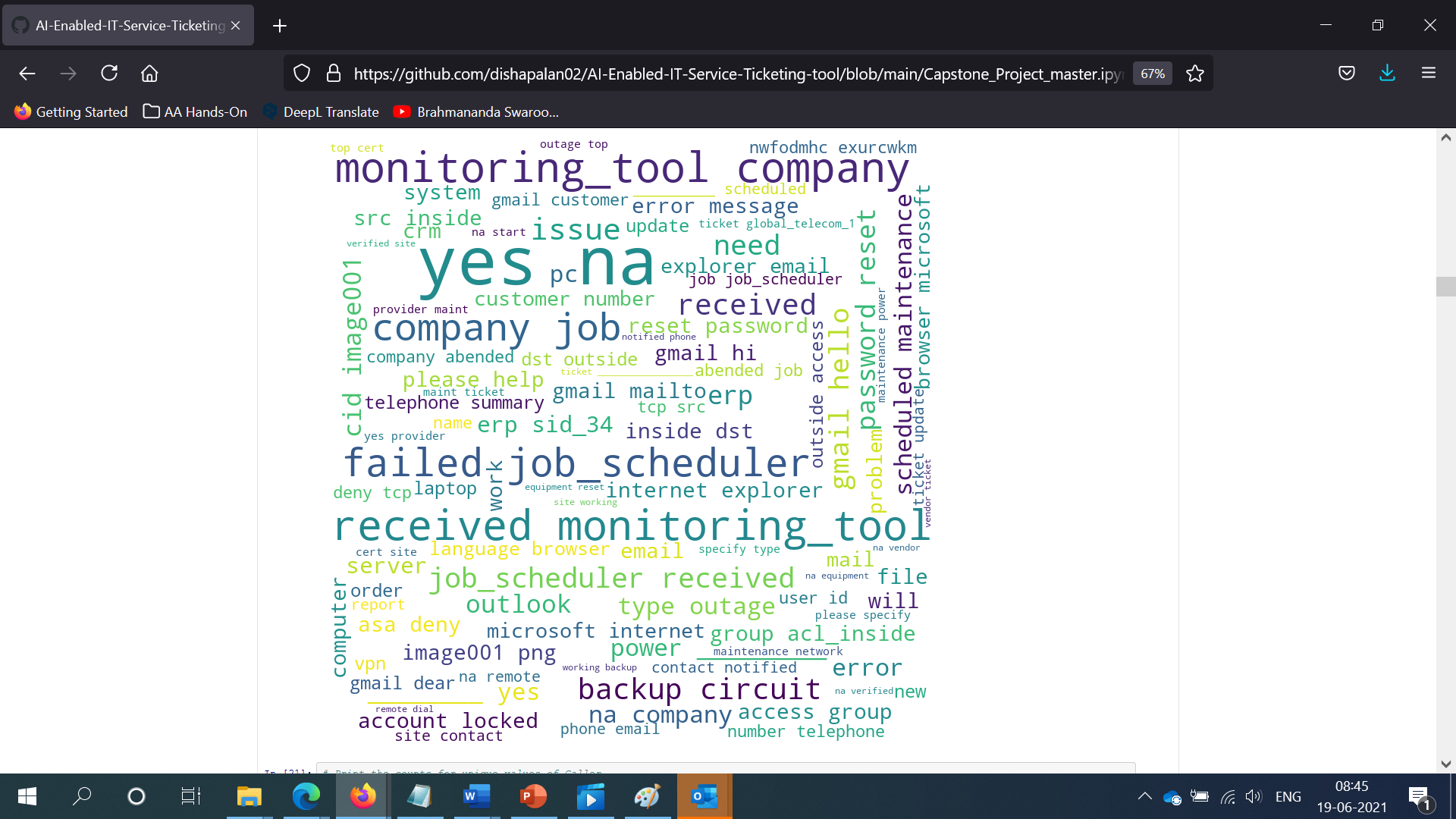
1. There are duplicates which needs to be tackled
2. There are mojibake texts in the description and short description which needs to be processed
3. There are texts belonging to different languages which needs translations
4. There are email ids, blank spaces, dates, numbers which needs to be processed
5. There are missing values to be treated

## Further Data Analysis

### Word cloud for Short description



### Word cloud for Description



### Assignment group distribution of top 10 callers as compared to other callers

Graphical user interface, text

Description automatically generated with medium confidence

### Assignment group distribution amongst top 10 callers

Graphical user interface, application

Description automatically generated with medium confidence

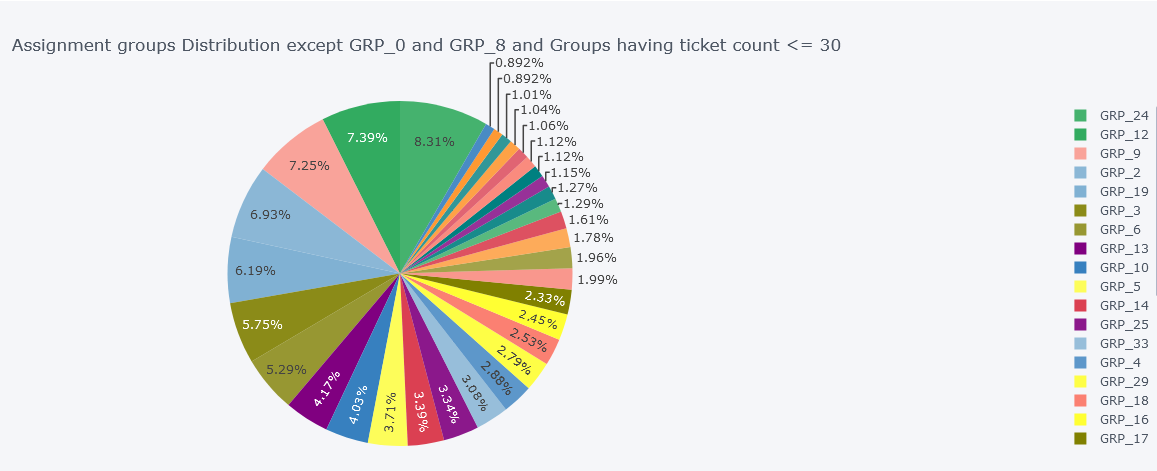
### Assignment group distribution

Graphical user interface

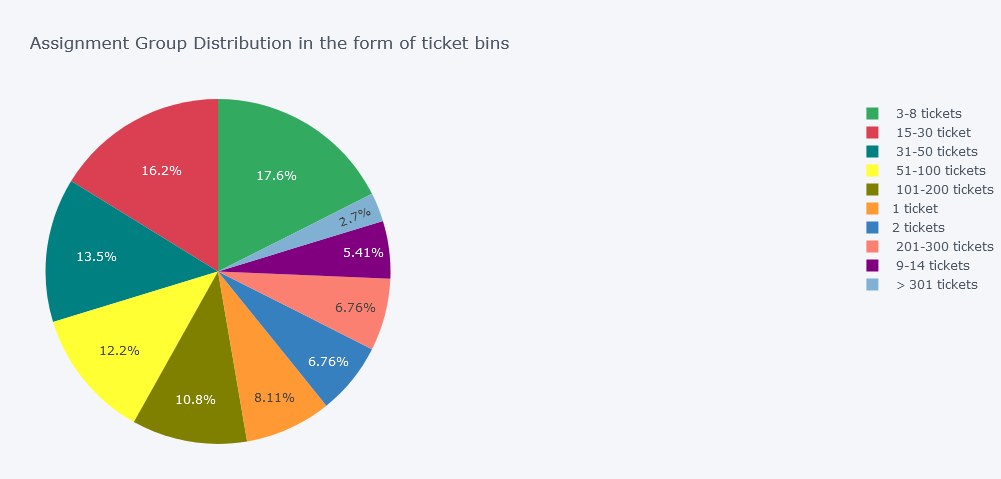
Description automatically generated

A picture containing table

Description automatically generated  
From above we see that there is significant class imbalance with GRP\_0 being the majority class. Now let’s see the distribution of classes other than two majority classes GRP\_0 and GRP\_8 an d also for those classes which have ticket count less than or equal to 30 ( The number 30 is chosen using central limit theorem here)



Now we will see if we can ignore some classes or merge them into some other assignment group considering that other group has capability to resolve the ticket for this assignment group. For this we first divided the classes in some ticket bins and found the distribution as



We see above that majority of assignment groups are one with tickets between 3 and 8 i.e. 17.6%. Also, from the above chart, we can see that Assignment group <= 2 tickets contributes to 14.87% i.e. (8.11 + 6.76%).

Also, after comparing the descriptions and understanding that other groups have capability to resolve same tickets. We found that we can either merge or ignore below assignment groups :

A picture containing text, bird, plant

Description automatically generated

## Feature Engineering

***Null value treatment***

***Removal of Duplicates***

***N-gram analysis***

***Word cloud for each assignment group***

## Data Pre-processing

***Translation***

***Detecting different languages***

***Fix Encoding***

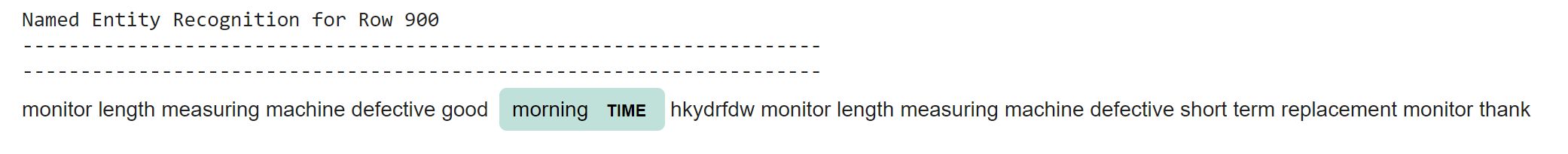
## Data Cleaning

***Removal of stop words***

***Lemmatization***

***Tokenization***

## NER and POS Tagging



Diagram

Description automatically generated

# Deciding Models and Model Building

Before building the model below steps were performed

Label Encoding

Splitting into train and test (80:20)

Treating class imbalance

The Assignment group before and after class balancing is shown as below:

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

## Models

As this is a classification problem where we need to classify the assignment groups using the ticket description and short descriptions. We have selected below traditional models and sequential models

As the target class is completely skewed, various models have been tried with the below set of datasets to compare each performance. Datasets used for each model are:

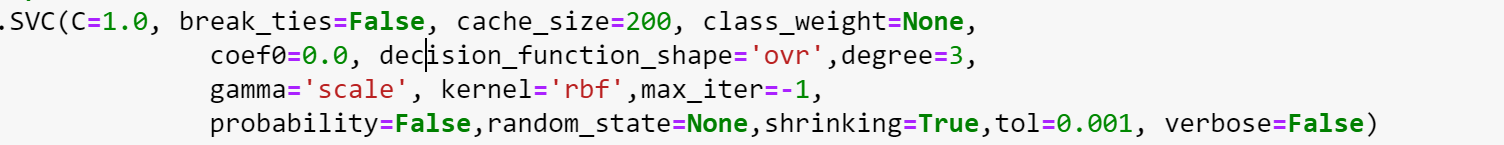
* Raw data with the target class without any sampling
* Resampled data where all the target classes are sampled with a count of 660. (Eg. Grp\_0 is down sampled and other groups are up sampled)
* Model with Two datasets: Model 1 with Grp\_0 & Model 2 with all other groups except Grp\_0 and Model 2 is resampled

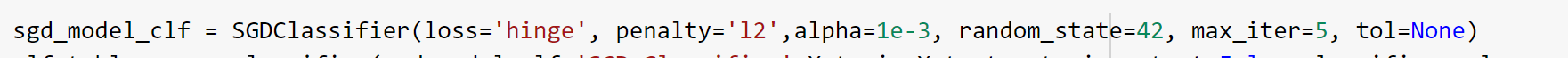
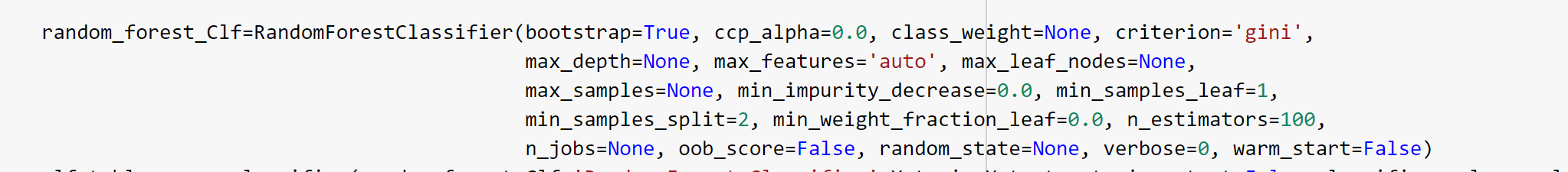
### Traditional Models

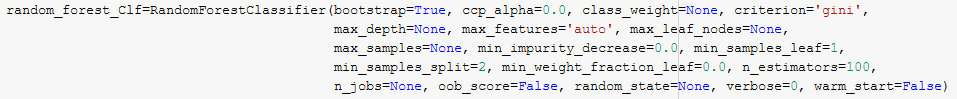
1. Multinomial NB Classifier



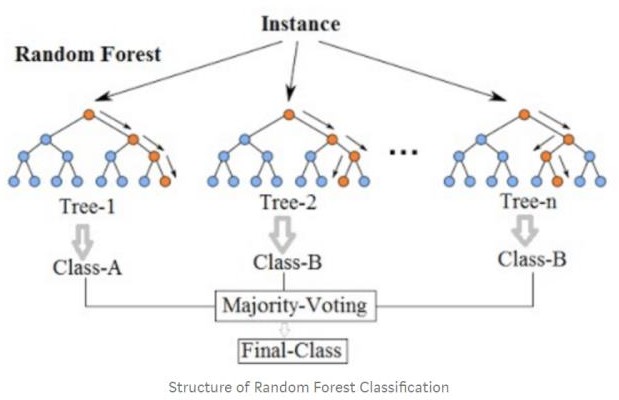
1. SVC Classifier

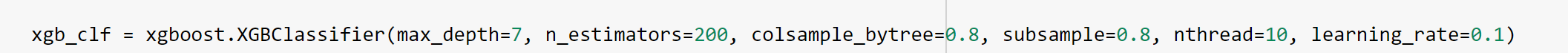


1. KNN Classifier
2. SGD Classifier
3. Random Forest Classifier



Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. It creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.



1. XGBOOST

We got below training and test accuracies and F1 score

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Train Accuracy** | **Test Accuracy** | **F1 Score** |
| Multinomial NB Classifier | 69.57% | 59.89% | 70.91% |
| SVC Classifier | 90.01% | 66.72% | 71.98% |
| KNN Classifier | 76.33% | 62.62% | 65.60% |
| SGD Classifier | 72.95% | 62.38% | 70.78% |
| Random Forest Classifier | 95.36% | 63.75% | 71.36% |
| XGBOOST | 92.70% | 64.52% | 69.78% |
|  |  |  |  |

### Sequential Models

1. CNN



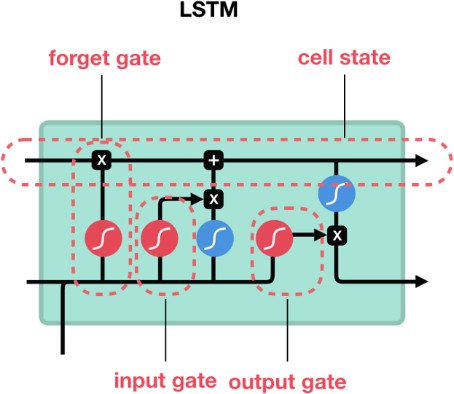
1. LSTM



**Bi-directional LSTM Model**

Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on classification problems.

In problems where all timesteps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem.



|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Train Accuracy** | **Test Accuracy** | **F1 Score** |
| CNN |  |  |  |
| LSTM |  |  |  |

# How to improve your model performance?

* Still we are working with balancing the classes ( We are also trying SMOTE for this)
* We want to use PCA as there seems to be too much difference between training and test accuracies
* Plotting the accuracy graphs is in progress

# Comparison to benchmark

From the given problem description, we could see that the existing system is able to assign 75% of the tickets correctly.

So our objective here is to build an AI-based classifier model to assign the tickets to right functional groups by analysing the given description with an accuracy of at least 85%.

From the prediction results we see that the GRU model based on the resampled data is able to achieve an accuracy of 91.24% which is above our benchmark.

# Implications

Although this model can classify the IT tickets with 91.24% accuracy, to achieve better accuracy in the real world it would be good if the business can collect additional data around 300 records for each group.

# Limitations

As part of Data pre-processing, we had grouped all assignment groups with less than 10 entries as one group (misc\_grp) which had reduced the Target class from 74 to 50 groups. While applying this model in the real world there could be additional intervention required to classify the tickets if it has been classified as misc\_grp by our model. Since the number of elements reported under misc\_grp are less, we expect this intervention to be done less often.

# Closing Reflections

We found the data was present in multiple languages and in various formats such as emails, chat, etc bringing in a lot of variability in the data to be analyzed. The Business can improve the process of raising tickets via a common unified IT Ticket Service Portal which reduces the above mentioned variability. By doing this, the model can perform better which can help businesses to identify the problem area for relevant clusters of topics.