

Automatic Ticket Assignment

Interim Report

# 

# Index

[Team Details 2](#_Toc37546120)

[Index 3](#_Toc37546121)

[Summary of the problem statement, Data and findings 4](#_Toc37546122)

[Problem Statement 4](#_Toc37546123)

[Data & Findings 4](#_Toc37546124)

[Summary of the approach to EDA and Pre-Processing 5](#_Toc37546125)

[Visualization 5](#_Toc37546126)

[Word Cloud 5](#_Toc37546127)

[Charts 5](#_Toc37546128)

[Decide Model and Model building 6](#_Toc37546129)

[Model performance - Approaches to improve model 7](#_Toc37546130)

[Code Snippet 8](#_Toc37546131)

[Finalized results 9](#_Toc37546132)

[Link to code and references 10](#_Toc37546133)

# Team Details

|  |  |
| --- | --- |
| Charan Teja J  Harsha V Chunduri  Jayapavan P Deshpande  Richa Anand  Soumendra Bagh  Visakh Vijayaprakash | [Charantej588@gmail.com](mailto:Charantej588@gmail.com)  [harshu244@gmail.com](mailto:harshu244@gmail.com)  [jayapavanpd@gmail.com](mailto:jayapavanpd@gmail.com)  richa.anand15@gmail.com  [soumendra.bagh@gmail.com](mailto:soumendra.bagh@gmail.com)  visakhvijayaprakash@gmail.com |

# **Summary of the Problem Statement, Data and Findings**

## Problem Statement

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

## Abstract

Applying traditional machine learning and neural network-based NLP to automatically classify tickets and assign them to the right owner in a timely manner to save effort, increase user satisfaction and improve throughput in the ticketing pipeline of an organization.

## Data & Findings

1. The dataset comprises of **8500 rows** and **4 columns**
2. All columns are of type object containing textual information.
3. There are **8 null/missing values** present in the Short description and **1 null/missing values** present in the description column
4. **Password reset** is one of the most occuring tickets which reflects in the Short description column.
5. The top occuring Description in the dataset is only the text **'the'**, which absolutely doesn't make any sense. hence by looking at the Short description of such rows reveals that these are also a category of Password reset.

### Data provided in format

XLSX/CSV

### Total Records

8500

### Data Fields

|  |  |
| --- | --- |
| Short description | A summary of the issue faced by the user |
| Description | Detailed description of the issue |
| Assignment group | GRP\_0 ~ GRP\_73 (total 74 classes of Assignment group) |

### Sample data

| **Short description** | **Description** | **Assignment group** |
| --- | --- | --- |
| login issue | -verified user details.(employee# & manager na... | GRP\_0 |
| outlook | \r\n\r\nreceived from: hmjdrvpb.komuaywn@gmail... | GRP\_0 |
| cant log in to vpn | \r\n\r\nreceived from: eylqgodm.ybqkwiam@gmail... | GRP\_0 |

### Distribution of classes and Observation

1. High imbalance seen in data for target column in our dataset with GRP\_0 having highest percent of representation
2. Many classes with very little representation.
3. Null values in Data:
   1. Short description **8**
   2. Description **1**
   3. Assignment group **0**
4. Observed few Ticket having Non-English ticket descriptions

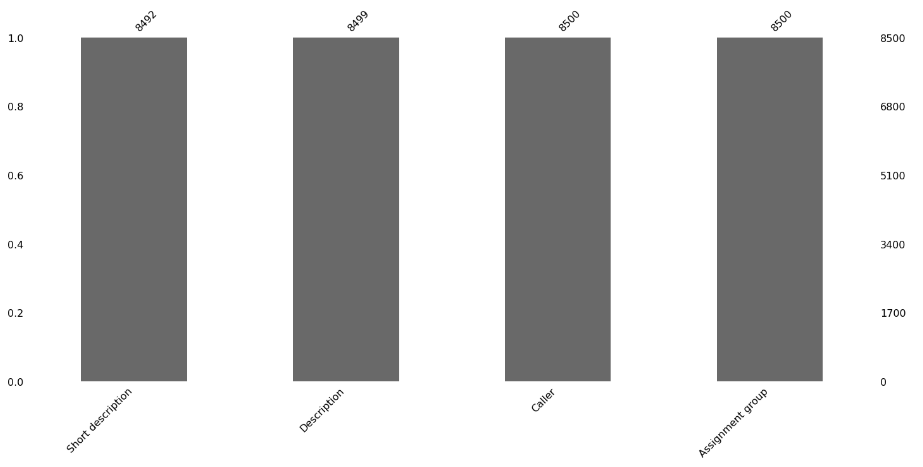
## Summary of the approach to EDA and Pre-Processing

### Cleaning processes applied

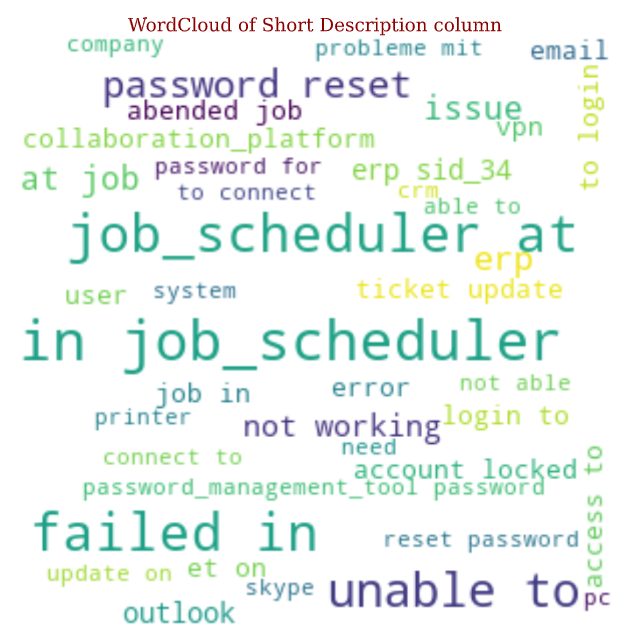
1. Removal of trailing spaces, line breaks and tabs (\r\n\t), removal of special characters and removal of extra spaces with **Regex Library**.
2. Convert the data having garbled text such as mojibake using **Ftfy Library**.
3. Replace NaN in short description with Short Description of other records having similar description’s Key Phrases. The Key phrases are extracted via **Rake Library**.
4. Foreign language detection and translation –
   1. Identified around 28 languages,
   2. Translated using **Goslate , Google Translate API** and **Azure APIs**
   3. Identified better translation accuracy using **Azure translator API**
5. Identified the data distribution to be imbalanced, the groups with records less than 20 records are handled by creating deterministic rules.
   1. Able to completely eliminate 6 groups based on these rules.
6. Sub divided Group 0 to further sub groups
7. Performed text summarisation using Rake.

## Visualization

**Visualization of Missing Values via Missingno Library**



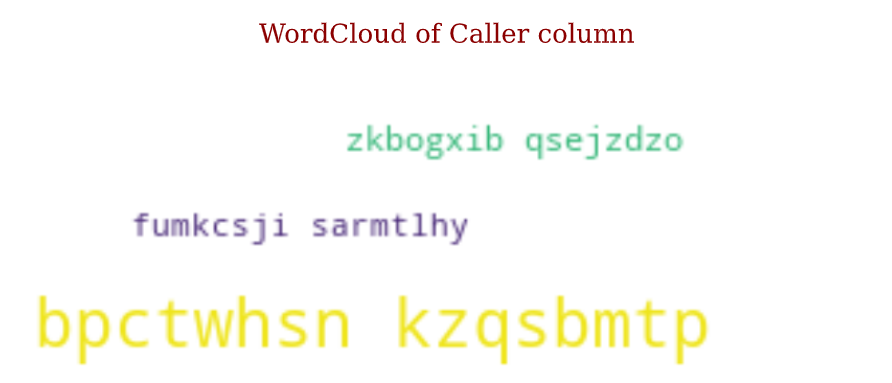
**Word Cloud Visualization for ‘Short description’ Feature:**



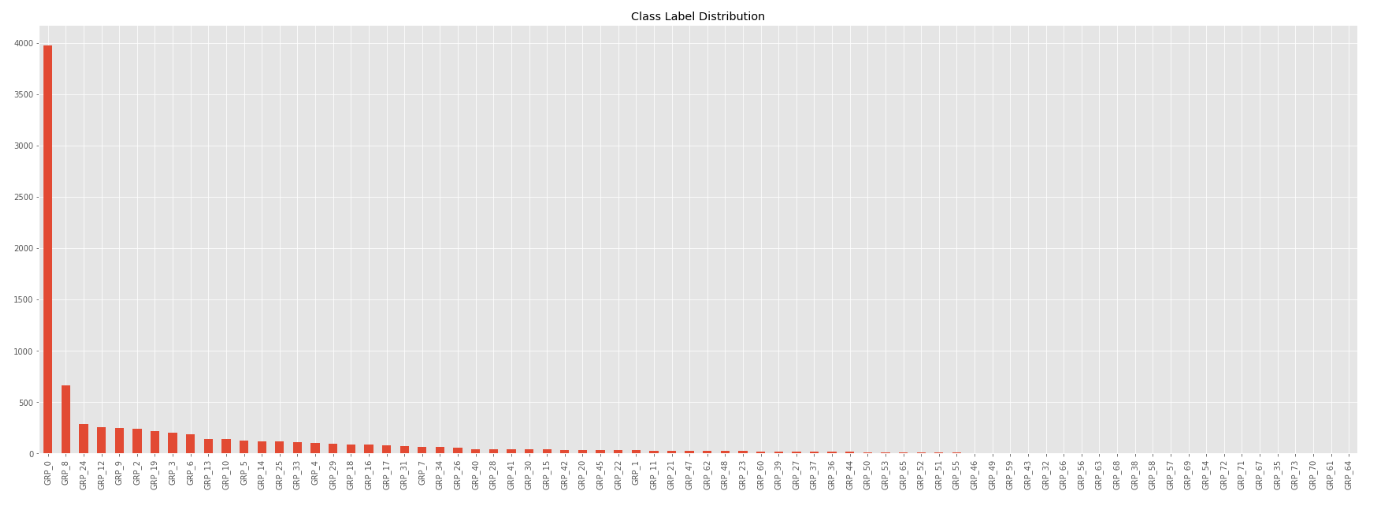
**Word Cloud Visualization for ‘Description’ Feature:**



**Word Cloud Visualization for ‘Caller’ Feature:**

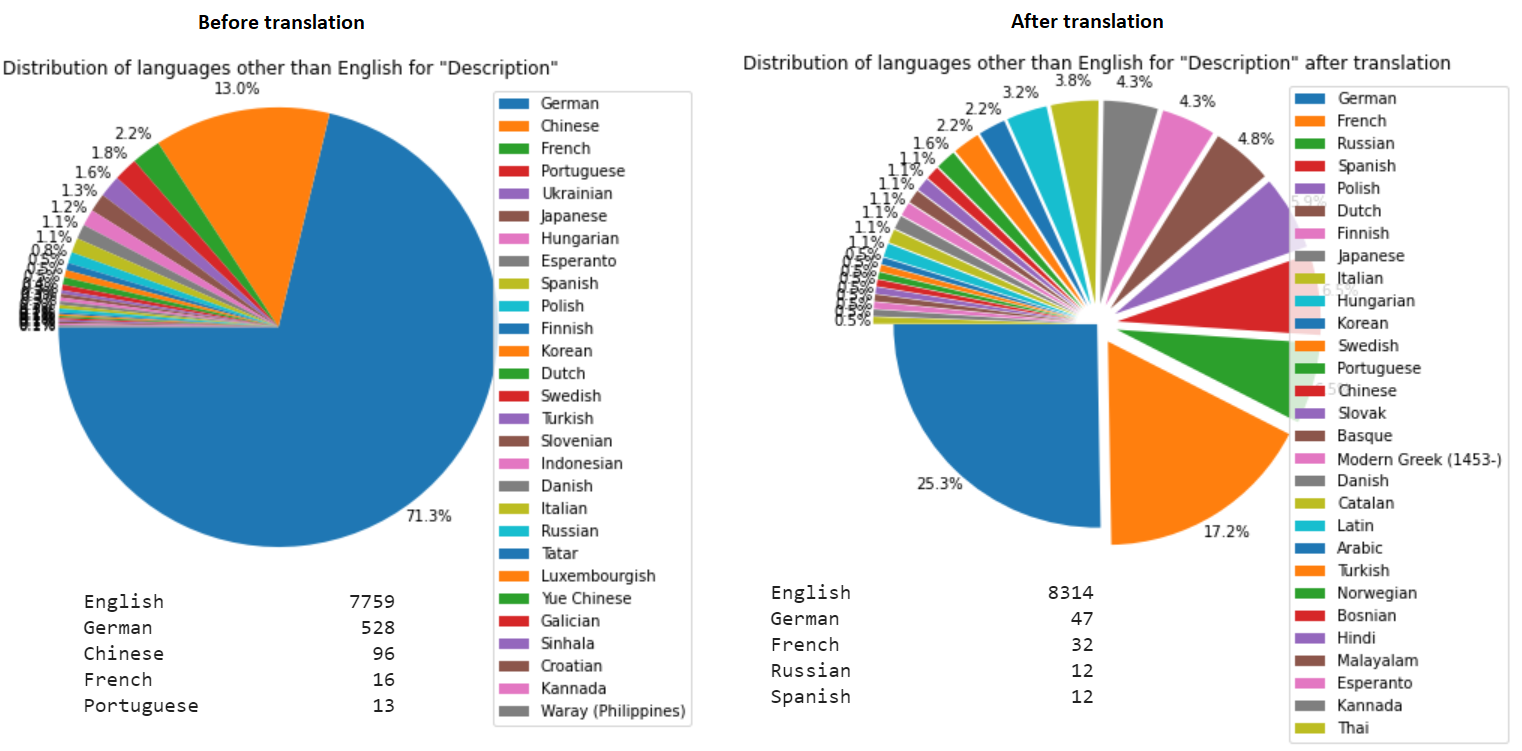


**Assignment Group distribution for Target Feature:**



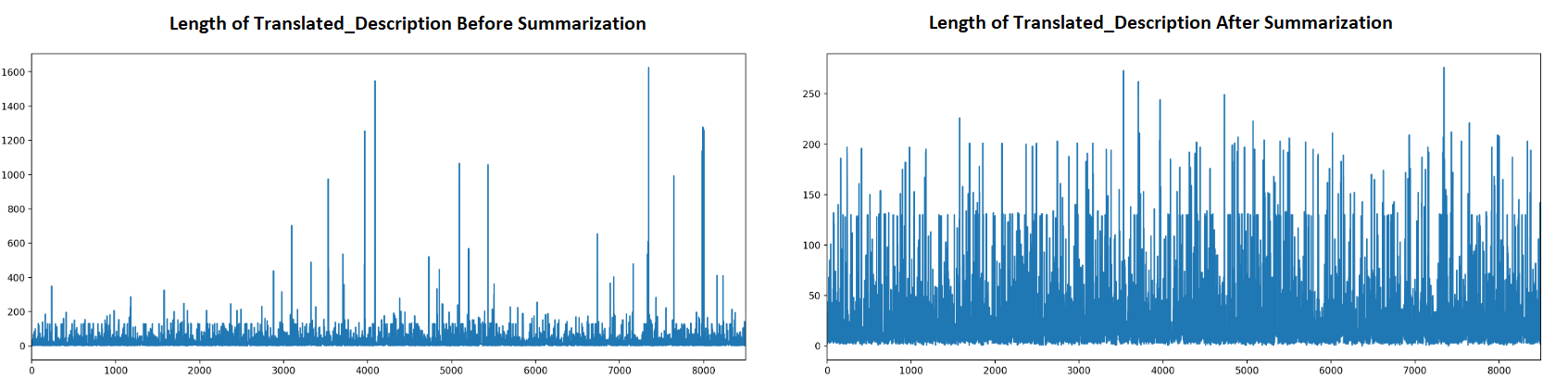
**Language Distribution:**

Distribution of foreign languages other than English before and after translation.

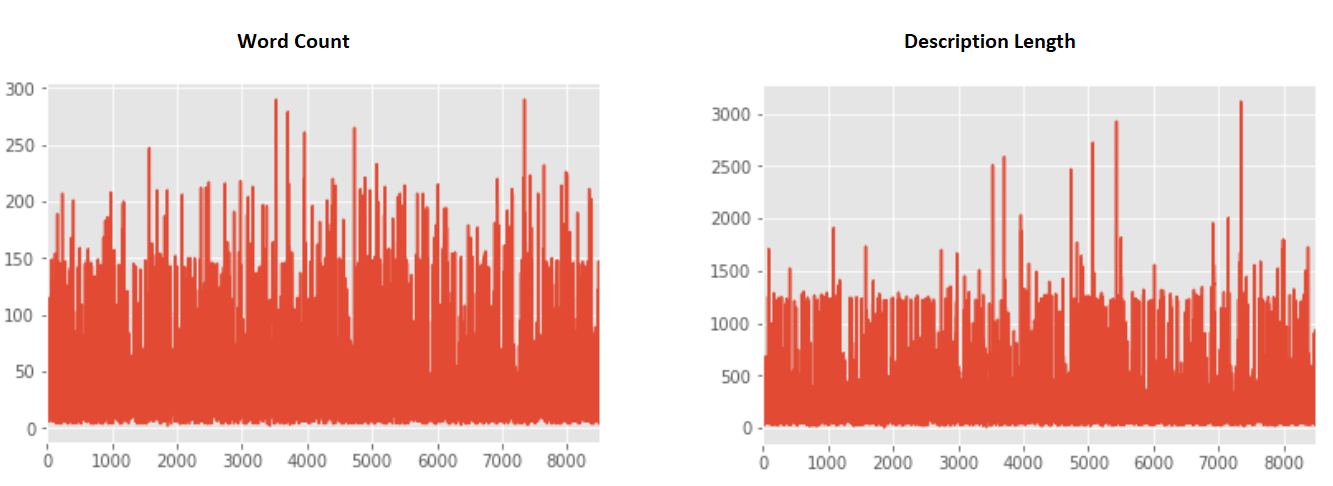


**Rake summarization:**

Length of text Description before and after Summarization



**Word count & Description length distribution:**



# Decide Model and Model building

We will try the problem with both Classical ML models (e.g. SVM, NB Classifier and Ensemble Models like Random Forest) and advanced Deep Learning based NLP methods(ex. Plain FC NN, RNNs and some Hybrid CNN-Seq. NN).

For classifying the assignment group, below model architectures have been tried.

Basic ML classification models such as

1. **Support Vector Machine (SVM)**
2. **Naïve Bayes Classifier**
3. **Random Forest**

Deep learning ANN

1. **Recurrent Neural Network (RNN)**

|  |  |
| --- | --- |
| **Model Name** | **Accuracy Score** |
| Support Vector Machine | 67.6% |
| Naïve Bayes | 54.5% |
| Random Forest | 63.88 |
| RNN | 57.17% |

**The key observation from use of these multiple models are as follows:**

* Simple ML Models are performing marginally better compared to heavy siblings DL Models.
* The main reason is the amount of training data are not enough to train huge number of params in NN based models.
* Because of less data the Models is simple memorizing training data (which can be validated by near 100% training accuracy) instead of calibrating/modelling any non linear relationships between input and target features.
* This is causing it to perform poorly on validation/test data(which can be validated by peer validation accuracy).
* And the problem of class imbalance in dataset is further compounding the problem, countering which will be our agenda for coming weeks.
* Because of skewed and sparse nature of dataset we will use other techniques like upscaling, down sampling, deterministic rule based preliminary classification to aid our overall models robustness and accuracy.

# Model performance - Approaches to improve model

As the dataset was highly imbalanced, models were overfitted and affected the performance.

Approaches to improve model performance:

1. **Upsampling and Downsampling of Groups** : We will sub-clustering Group-0 and other dominant Groups into smaller bins/sub clusters via downsampling to enable the model to give equal importance to other classes/groups which are scarce in datasets. We will further minimize this imbalance in proportionality by upscaling the very sparse groups in target column to a certain value.
2. **Feature Engineering** : By introducing feature Engineering we will transform the input data into a more perceptive form from model’s point of view model. The more the model will be able to understand the artificially transformed (via feature engineering) data, the better predictions it will be able to deliver. Some of the feature engineering currently tested are Topic Modelling, Document Clustering, TF-IDF and Word Embedding.
3. **Hyper parameter tuning in ML and DL models**. As it will fine tune our model to closely emulate the non linear relationships between all independent features i.e. Short description, Description and Caller and the Target/Dependent feature which has to be predicted successfully by our model.
4. **Addition of Deterministic Rule based classification** for Groups having very scarce presence in dataset. By adopting this non-ML approach we will be able to improve prediction of such non-dominant Group’s prediction. The deterministic rule are those relationships between dependent features and target feature that be deduced manually and are very unique belonging to a specific group. This deterministic rule part of the program will be implemented just before applying any ML/DL approach on the same data so the data will be checked first by this and then after failing to achieve any conclusion will have to be passed into more complex ML/DL models.

# Link to code and references

<https://github.com/jayapavandeshpande/nlp_project>