INTERIM REPORT

AI Enabled IT Ticketing Service Tool

Submitted By

Disha Palan

Parita Desai

Gloria Preet

Vivek UImale

Table of Contents

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

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

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

Top 10 records of our dataset :

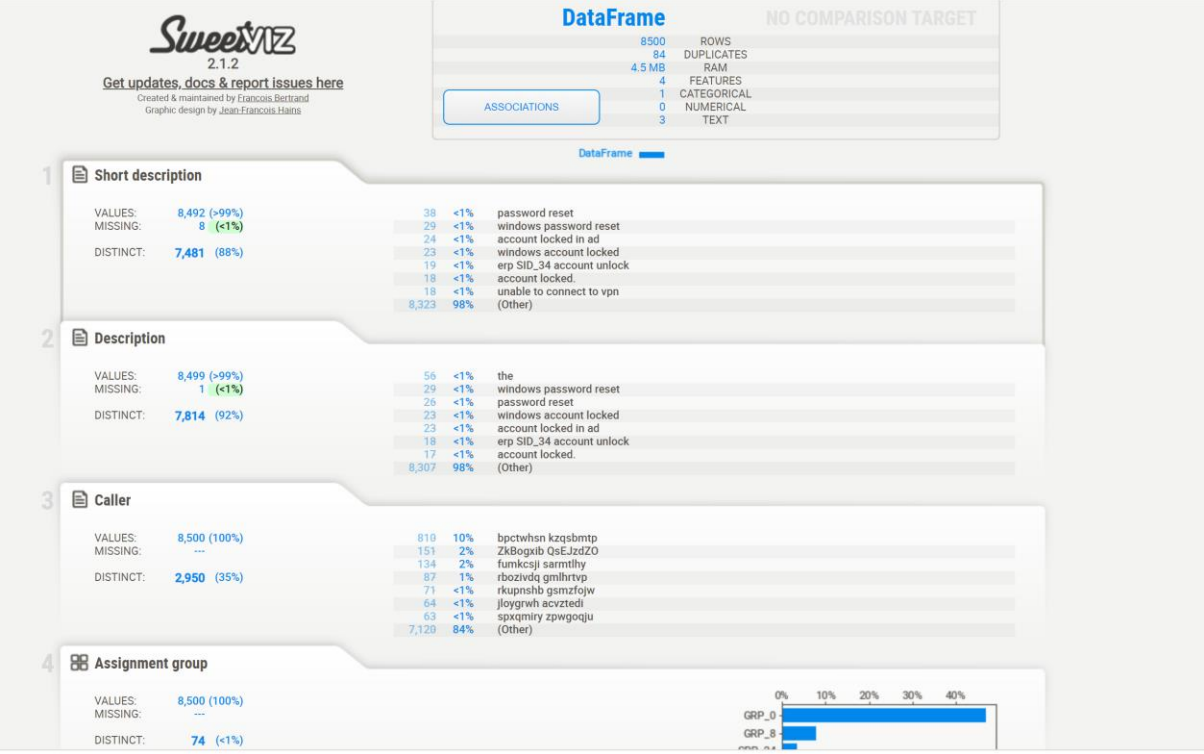
	Short description	Description	Caller	Assignment group
0	login issue	-verified user details.(employee# & manager na...	spxjnwir pjlcoqds	GRP_0
1	outlook	\n\nreceived from: hmjdrvpb.komuaywn@gmail.com...	hmjdrvpb komuaywn	GRP_0
2	cant log in to vpn	\n\nreceived from: eylqgodm.ybqkwiam@gmail.com...	eylqgodm ybqkwiam	GRP_0
3	unable to access hr_tool page	unable to access hr_tool page	xbkucsvz gcpydteq	GRP_0
4	skype error	skype error	owlgqjme qhcozdfx	GRP_0
5	unable to log in to engineering tool and skype	unable to log in to engineering tool and skype	eflahbxn ltdgrvkz	GRP_0
6	event: critical:HostName_221.company.com the v...	event: critical:HostName_221.company.com the v...	jyoqwxhz clhxsoqy	GRP_1
7	ticket_no1550391- employment status - new non-...	ticket_no1550391- employment status - new non-...	eqzibjhw ymebpoih	GRP_0
8	unable to disable add ins on outlook	unable to disable add ins on outlook	mdbegvct dbvichlg	GRP_0
9	ticket update on inplant_874773	ticket update on inplant_874773	fumkcsji sarmtlhy	GRP_0

Summary of the Approach to EDA and Pre-processing

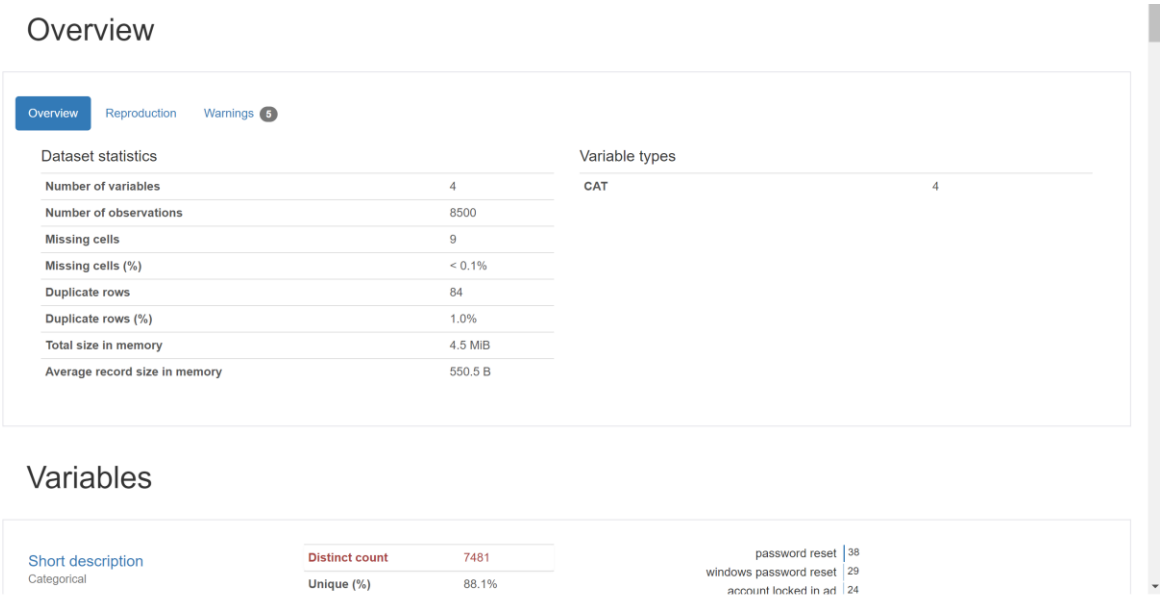
EDA

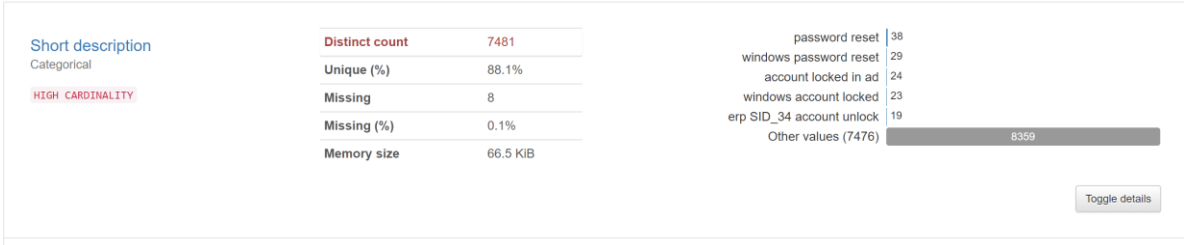
We have used SweetViz and Panda profiling to Visualize and analyse our dataset. Below are attached reports for and observations from reports are in next sections. The reports can be viewed in GitHub link - https://github.com/dishapalan02/AI-Enabled-IT-Service-Ticketing-tool/tree/main

SweetViz:

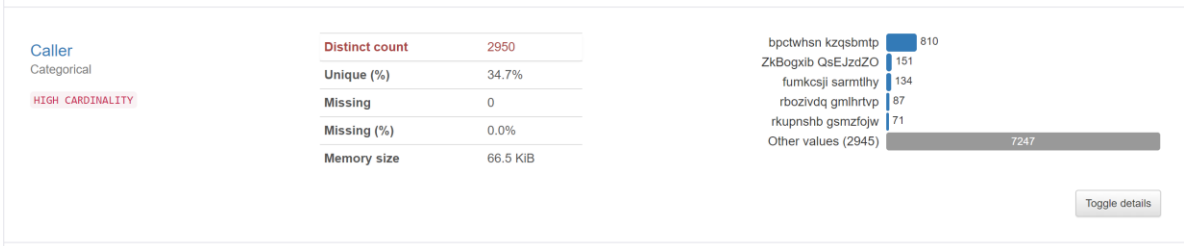


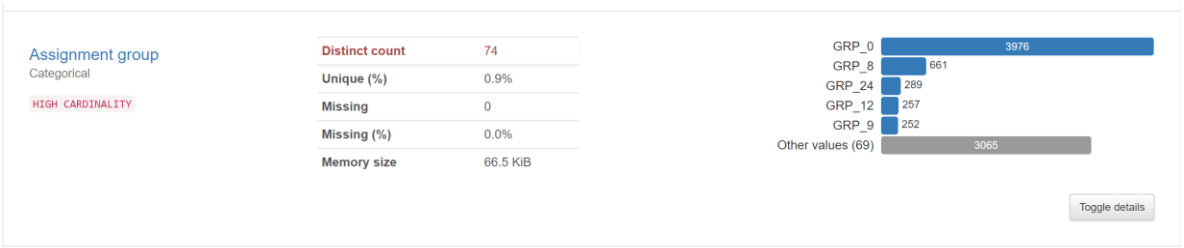
Pandas Profiling:

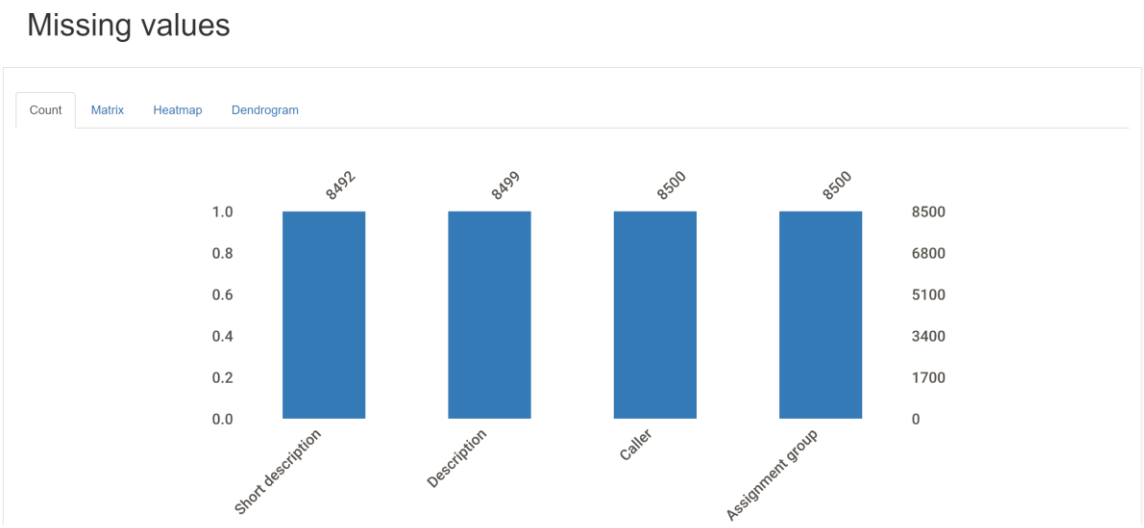

Overview
Overview
Reproduction
Warnings 5
Dataset statistics
Number of variables 4
Number of observations 8500
Missing cells 9
Missing cells (%) < 0.1%
Duplicate rows 84
Duplicate rows (%) 1.0%
Total size in memory 4.5 MiB
Average record size in memory 550.5 B
Variable types
CAT 4
Variables
Short description
Categorical
Distinct count 7481
Unique (%) 88.1%
password reset 38
windows password reset 29
account locked in ad 24


Short description
Categorical
HIGH CARDINALITY
Distinct count 7481
Unique (%) 88.1%
Missing 8
Missing (%) 0.1%
Memory size 66.5 KiB
password reset 38
windows password reset 29
account locked in ad 24
windows account locked 23
erp SID_34 account unlock 19
Other values (7476) 8359
Toggle details


Description
Categorical
HIGH CARDINALITY
Distinct count 7814
Unique (%) 91.9%
Missing 1
Missing (%) < 0.1%
Memory size 66.5 KiB
the 56
windows password reset 29
password reset 26
windows account locked 23
account locked in ad 23
Other values (7809) 8342
Toggle details


Caller
Categorical
HIGH CARDINALITY
Distinct count 2950
Unique (%) 34.7%
Missing 0
Missing (%) 0.0%
Memory size 66.5 KiB
bpctwhsn kzqsbmtp 810
ZkBogxib QsEJzdZO 151
fumkcsji sarmtlhy 134
rbozivdq gmlhrtvp 87
rkupnshb gsmzfojw 71
Other values (2945) 7247
Toggle details


Assignment group
Categorical
HIGH CARDINALITY
Distinct count 74
Unique (%) 0.9%
Missing 0
Missing (%) 0.0%
Memory size 66.5 KiB
GRP_0 3976
GRP_8 661
GRP_24 289
GRP_12 257
GRP_9 252
Other values (69) 3065
Toggle details



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
3. Target Column - 1

   3.1 Assignment Group - Categorical
4. There are 84 duplicate records in total. Strategy to handle duplicates and the approach taken is defined in the pre-processing section below.
5. 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. #ToDo To find if issue is controllable or not --> Check if possible .
   D. 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.

2. 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.

3. 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.

4. 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

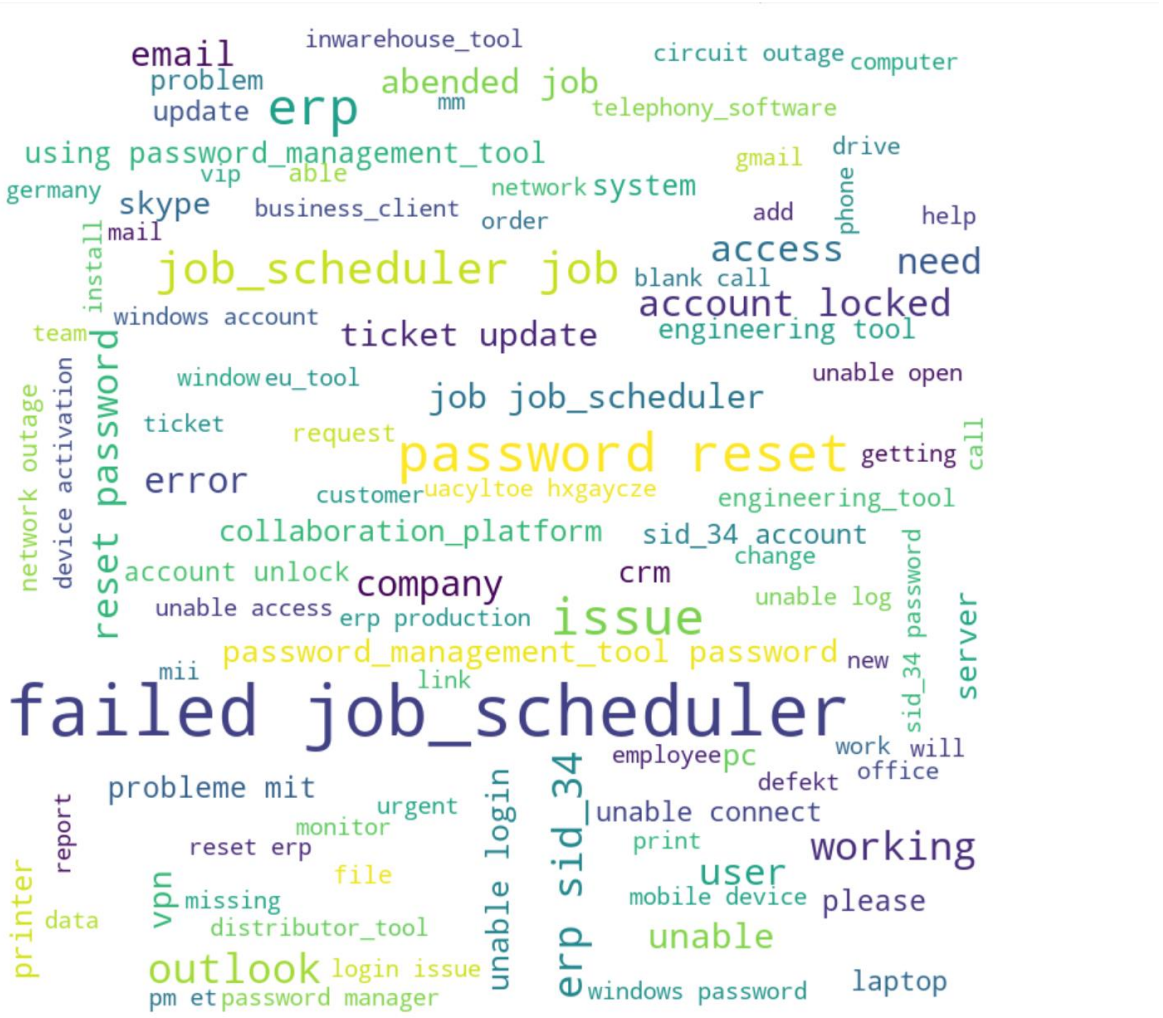
   G. This indicates we can merge few assignment groups with smaller percentage to reduce overall number of categories.

Other findings

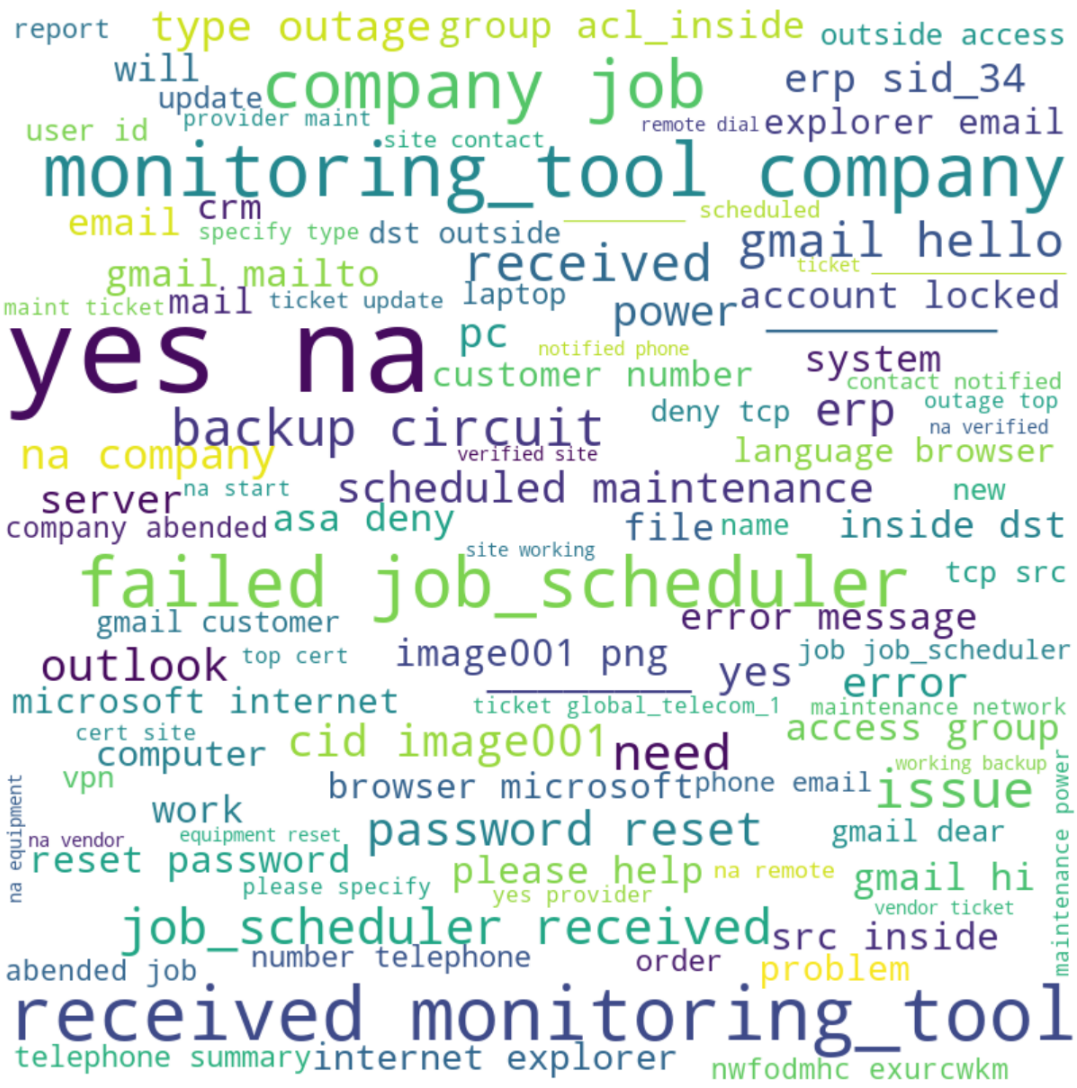
1. There are duplicates which needs to be tackled
2. There are mojibake texts in the description and short description which needs to be processed
3. There are texts belonging to different languages which needs translations
4. There are email ids, blank spaces, dates, numbers which needs to be processed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

Further Data Analysis

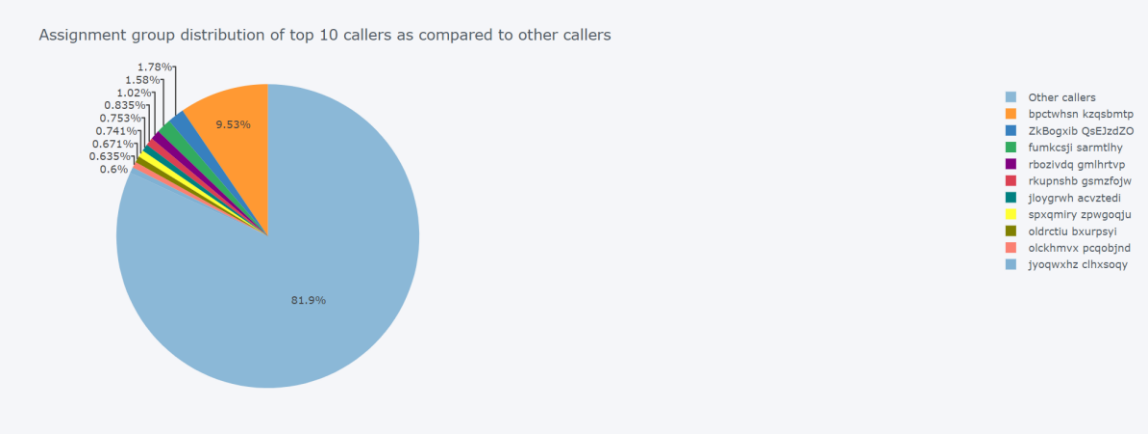
Word cloud for Short description



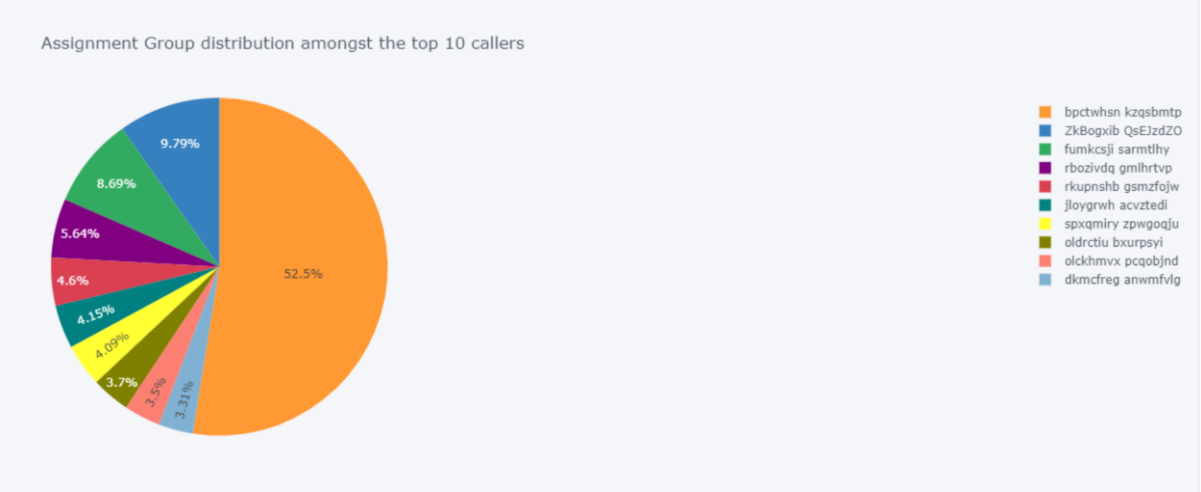
Word cloud for Description



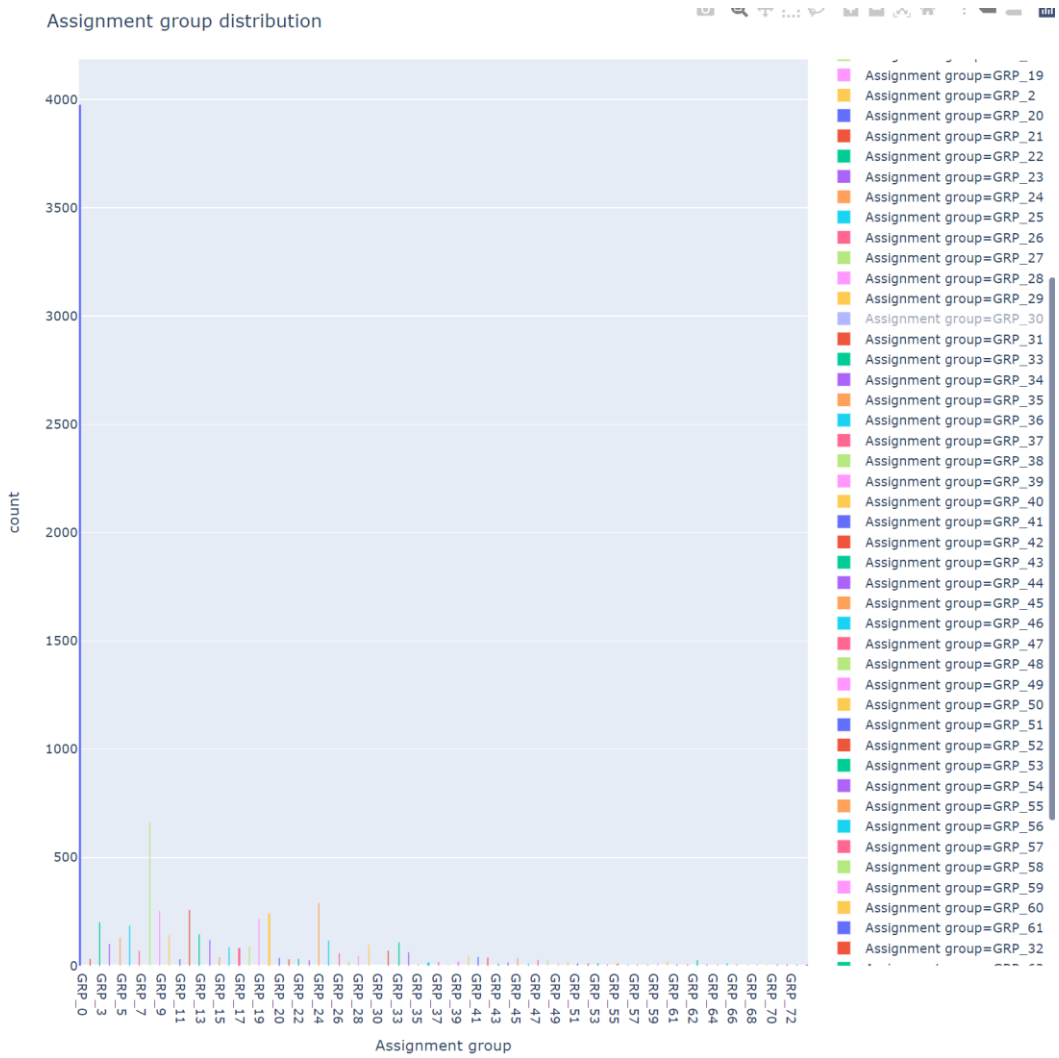
Assignment group distribution of top 10 callers as compared to other callers

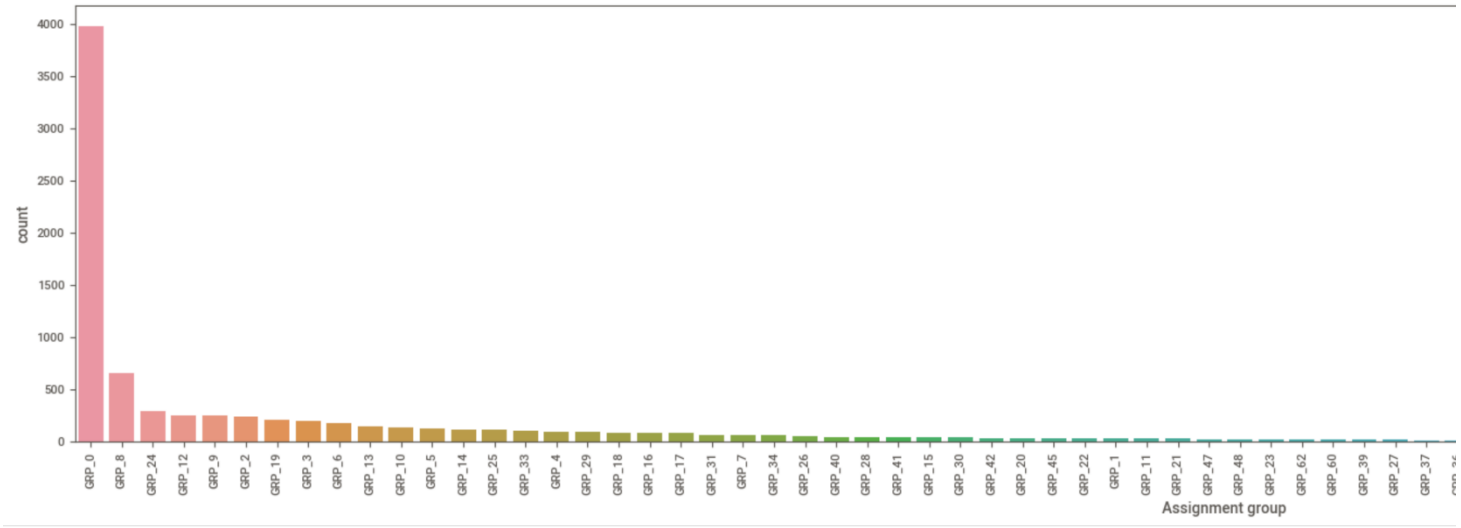


Assignment group distribution amongst top 10 callers

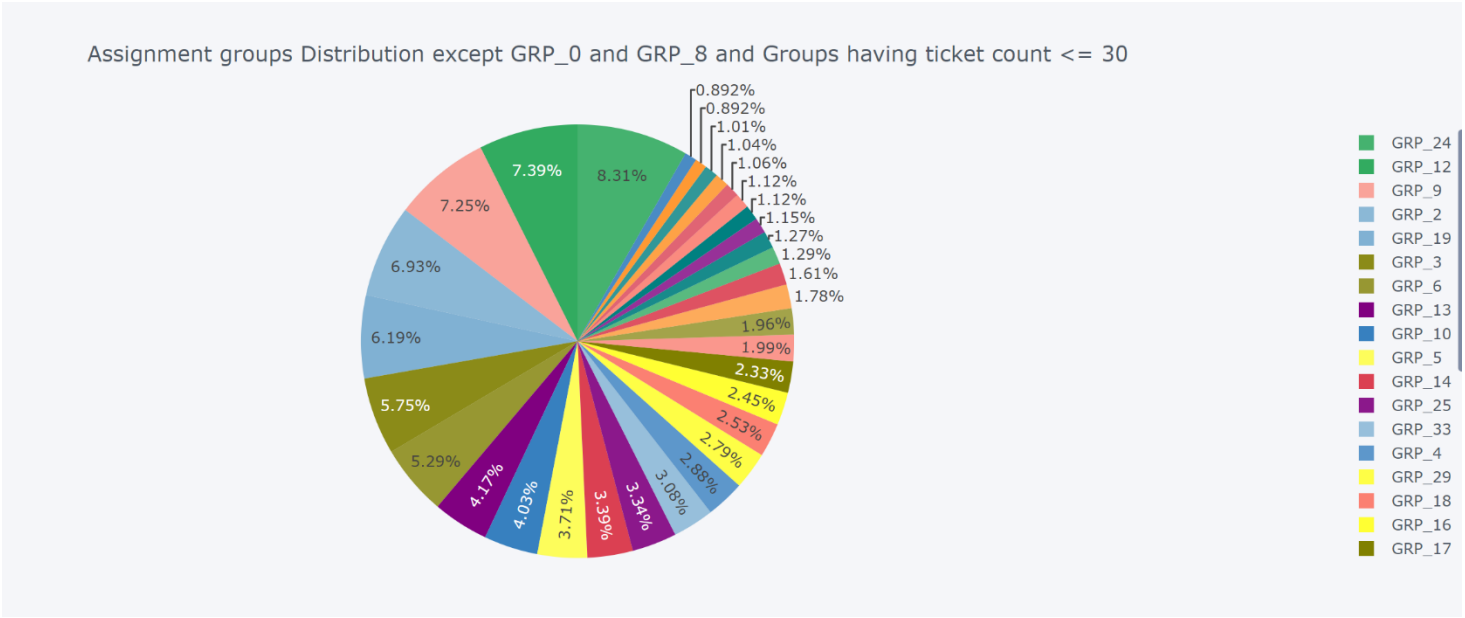


Assignment group distribution

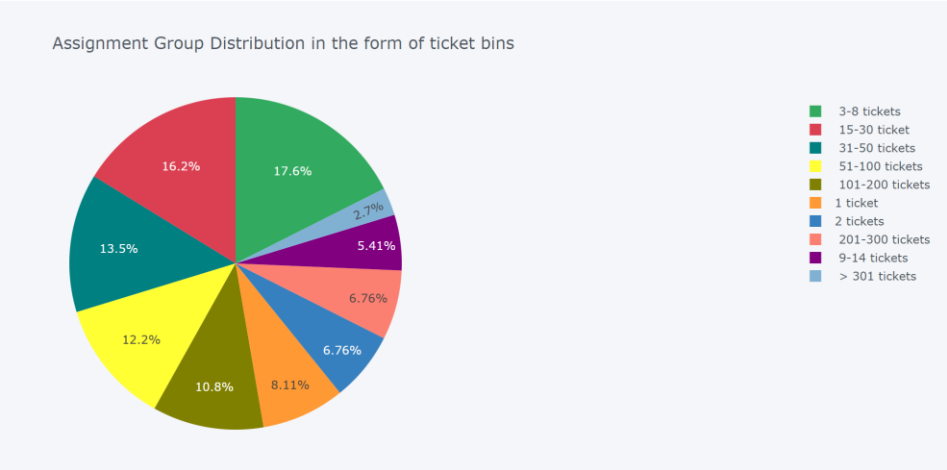




From above we see that there is significant class imbalance with GRP_0 being the majority class. Now let's see the distribution of classes other than two majority classes GRP_0 and GRP_8 an d also for those classes which have ticket count less than or equal to 30 (The number 30 is chosen using central 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 :

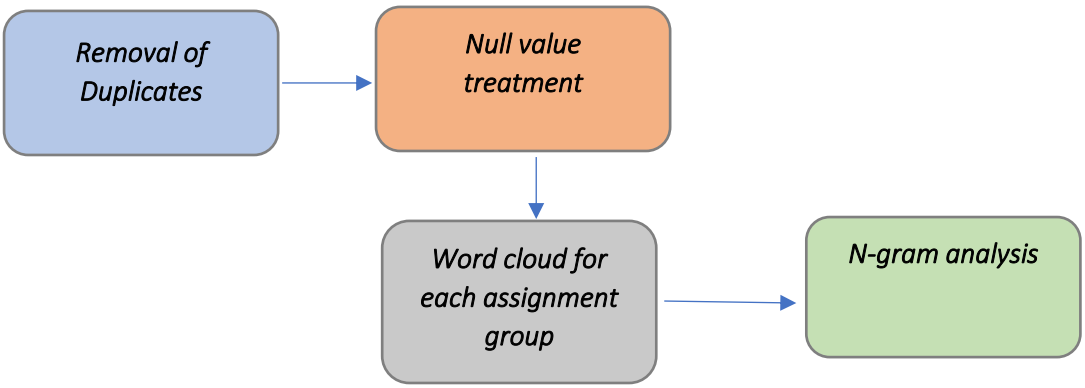
So finally, we see that below groups can be ignored based on above analysis. We will decide this later post feature engineering

1. GRP_72
2. GRP_54
3. GRP_57
4. GRP_69
5. GRP_67
6. GRP_35
7. GRP_73

But, we shall not ignore below assignment groups:

1. GRP_71
2. GRP_70
3. GRP_61
4. GRP_64

Feature Engineering



Data Pre-processing

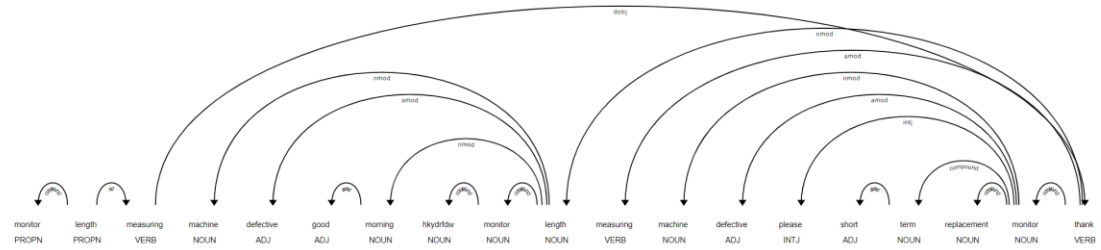
Data Cleaning

NER and POS Tagging

```
Named Entity Recognition for Row 900
------------------------------------------------------------------------
------------------------------------------------------------------------
```

monitor length measuring machine defective good morning TIME hkydrfdw monitor length measuring machine defective short term replacement monitor thank

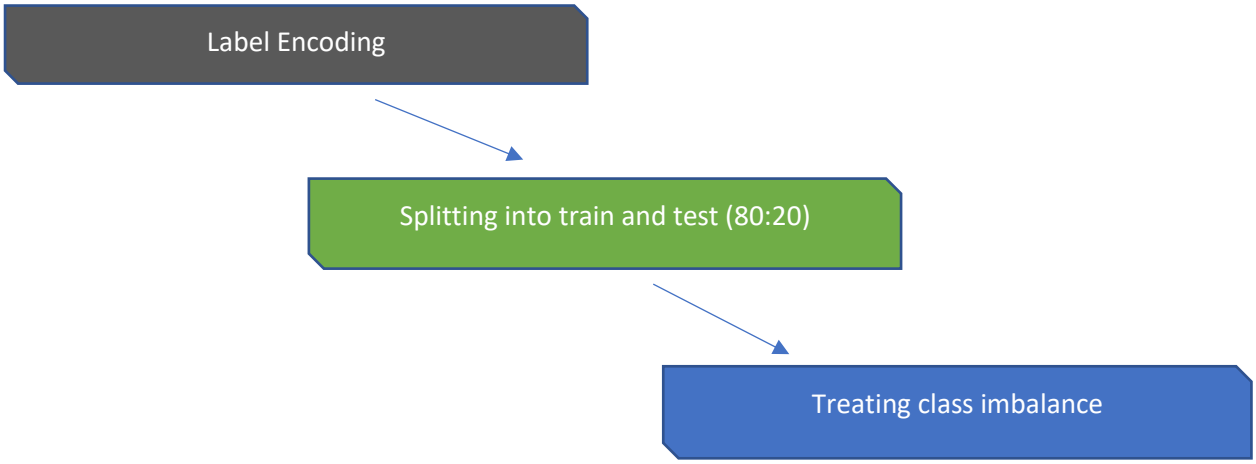
```
Dependencies for Row 900
------------------------------------------------------------------------
------------------------------------------------------------------------
```



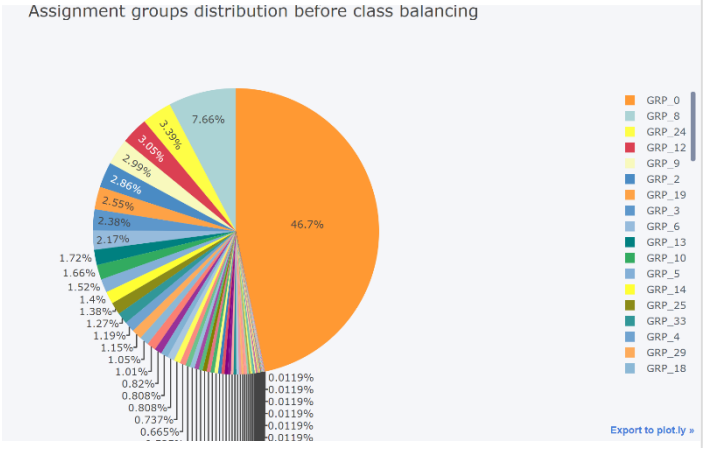

```
Word in sentence, Parts-of-speech and Lemmatized sentence 900
------------------------------------------------------------------------
------------------------------------------------------------------------
[('monitor', 'PROPN', 'monitor'), ('length', 'PROPN', 'length'), ('measuring', 'VERB', 'measure'), ('machine', 'NOUN', 'machine'), ('defe
ctive', 'ADJ', 'defective'), ('good', 'ADJ', 'good'), ('morning', 'NOUN', 'morning'), ('hkydrfdw', 'NOUN', 'hkydrfdw'), ('monitor', 'NOU
N', 'monitor'), ('length', 'NOUN', 'length'), ('measuring', 'VERB', 'measure'), ('machine', 'NOUN', 'machine'), ('defective', 'ADJ', 'def
ective'), ('short', 'ADJ', 'short'), ('term', 'NOUN', 'term'), ('replacement', 'NOUN', 'replacement'), ('monitor', 'NOUN', 'monitor'),
('thank', 'VERB', 'thank')]

[(monitor, 'O', ''), (length, 'O', ''), (measuring, 'O', ''), (machine, 'O', ''), (defective, 'O', ''), (good, 'O', ''), (morning, 'B',
'TIME'), (hkydrfdw, 'O', ''), (monitor, 'O', ''), (length, 'O', ''), (measuring, 'O', ''), (machine, 'O', ''), (defective, 'O', ''), (ple
ase, 'O', ''), (short, 'O', ''), (term, 'O', ''), (replacement, 'O', ''), (monitor, 'O', ''), (thank, 'O', '')]

{'morning': 'TIME'}
```

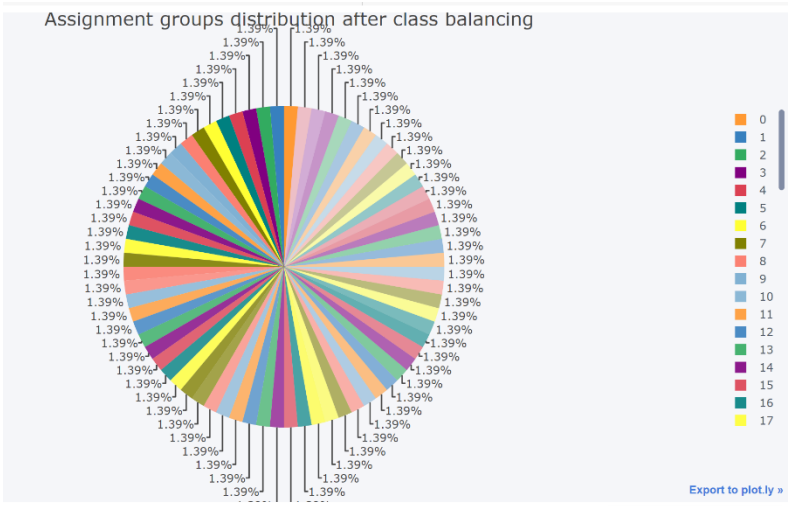
Deciding Models and Model Building

Before building the model below steps were performed



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





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

Traditional Models

1. Multinomial NB Classifier

```
multi_nb_clf = MultinomialNB(alpha=0.25)
multi_nb_clf = OneVsRestClassifier(multi_nb_clf)
```

2. SVC Classifier

```
.SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None,
             coef0=0.0, decision_function_shape='ovr',degree=3,
             gamma='scale', kernel='rbf',max_iter=-1,
             probability=False,random_state=None,shrinking=True,tol=0.001, verbose=False)
```

3. KNN Classifier

```
knn_model_clf = KNeighborsRegressor(algorithm='auto', leaf_size=30, n_jobs=None, n_neighbors=3, p=2, weights='uniform')
knn_model_clf = OneVsRestClassifier(knn_model_clf)
```

4. SGD Classifier

```
sgd_model_clf = SGDClassifier(loss='hinge', penalty='l2',alpha=1e-3, random_state=42, max_iter=5, tol=None)
```

```
random_forest_Clf=RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini',
                                         max_depth=None, max_features='auto', max_leaf_nodes=None,
                                         max_samples=None, min_impurity_decrease=0.0, min_samples_leaf=1,
                                         min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100,
                                         n_jobs=None, oob_score=False, random_state=None, verbose=0, warm_start=False)
```

5. Random Forest Classifier

6. XGBOOST

```
xgb_clf = xgboost.XGBClassifier(max_depth=7, n_estimators=200, colsample_bytree=0.8, subsample=0.8, nthread=10, learning_rate=0.1)
```

We got below training and test accuracies and F1 score

Classifier	Train Accuracy	Test Accuracy	F1 Score
Multinomial NB Classifier	69.77%	59.89%	70.91%
SVC Classifier	91.64%	66.72%	71.98%
KNN Classifier	75.34%	62.63%	65.60%
SGD Classifier	72.56%	62.39%	70.78%
Random Forest Classifier	95.47%	63.34%	71.10%
XGBOOST	92.60%	64.53%	69.78%

Sequential Models (In-progress)

1. CNN
2. LSTM

Classifier	Train Accuracy	Test Accuracy	F1 Score
CNN			
LSTM			

As per above model evaluation result SVC classifier performs best with 66.72% accuracy and 71.98% F1 score on test data set. This can be further improved with below strategies.

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