**Capstone Project**

**Automatic Ticket Assignment**

**Prepared By: Shivalik Chhabra, Nishant Joshi, Srikanth Rammnath, Mansi Garg**

Summary of problem statement, data and findings

**PROBLEM STATEMENT**

In any IT industry, Incident(an unplanned interruption to an IT service or reduction in the quality of an IT service that affects the Users and the Business) Management plays an important role in delivering quality support to customers. The main goal of this management process is to provide a quick fix / workarounds or solutions that resolves the interruption and restores the service to its full capacity to ensure no business impact . Whenever an incident is created, it reaches the Service desk team and then it gets assigned to the respective teams to work on the incident.

The manual assignment of these incidents might have below disadvantages:

* Time consuming and requires human efforts
* Increases human errors and resource consumption , as it is carried out ineffectively because of the misaddressing.
* Increases the response and resolution times which result in user satisfaction deterioration / poor customer service

If this ticket assignment is automated, it can be more cost-effective, less resolution time and the Service Desk team can focus on other productive tasks.

**OBJECTIVE**

The goal is to build a classifier that can classify the tickets by analysing text.

**DATA DESCRIPTION**

The given dataset consists of the following four attributes:

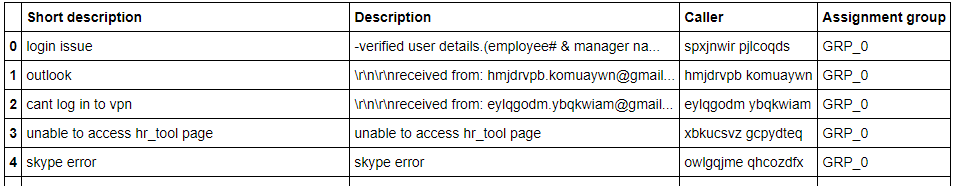
1. Short Description (a summary of the issue faced by the user)

2.Description ( detailed description of the issue)

3.Caller ( ID of the caller)

4.Assignment group(GRP\_0 ~ GRP\_73 i.e., total 74 classes of Assignment group) – Target class

**SAMPLE DATA**



**FINDINGS**

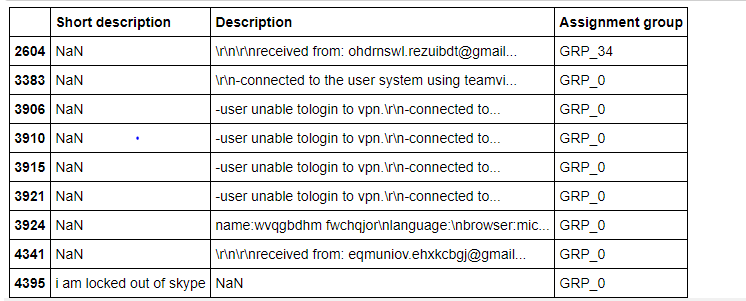
Followings are the general observation from the given dataset :

* Caller ID are present in a random manner (may not be useful for training data)
* Languages other than English for example- German, etc. are also present in the dataset
* Non-English languages are also found in the data
* Email/chat format with symbols in description
* Hyperlinks and URLS are found in the description
* Blank records are present in either short description or description
* Few descriptions are replica of the short description
* Few words were combined together
* Spelling mistakes and typo errors are found

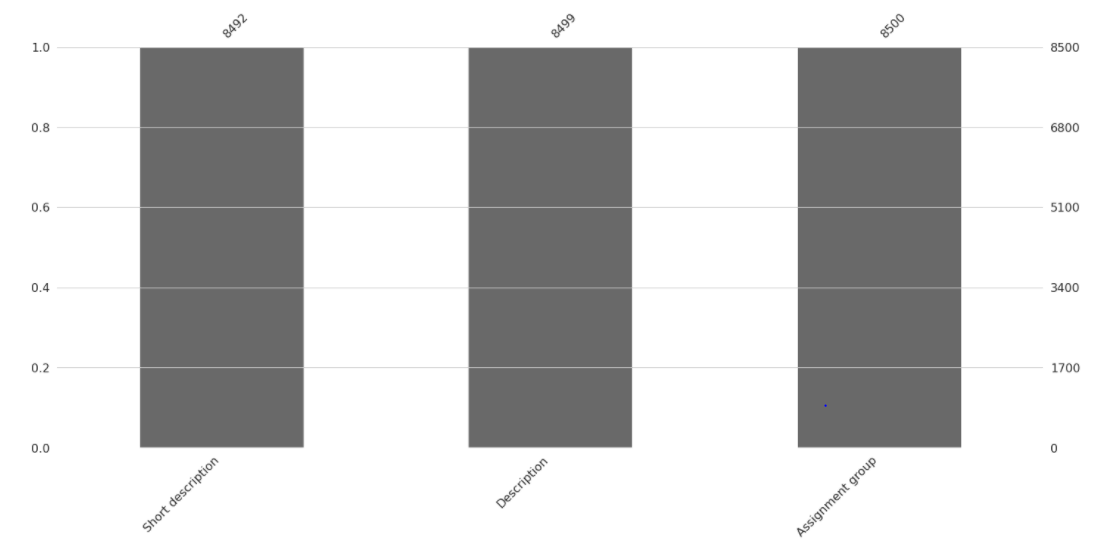
Summary of Approach to EDA and Pre-processing

**Step-by-step walk through the solution with the observations**

* Loaded the input csv file into pandas data frame.
* EDA has been performed on the dataset and following are the observations from that:
* All columns are of type object containing textual information.
* Assignment group is our predictor / target column with multiple classes. So, this is a Multiclass Classification problem.
* There are **8 null/missing values** present in the Short description and **1 null/missing values** present in the description column.



Visually the number of missing values are as shown below:

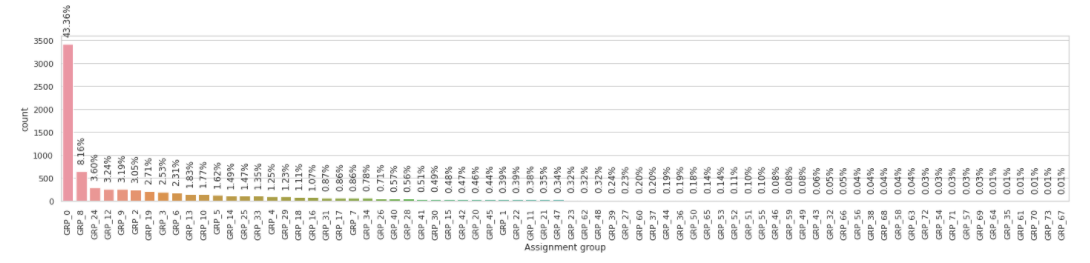


* Dropped the duplicates entries of the incidents. Thus, the size of the dataset gets reduced to (7909,4)
* Dropped the caller attribute as the data was not found to be useful for analysis
* Replaced Null values in Short description & description with space.
* Merged Short Description & Description fields for analysis

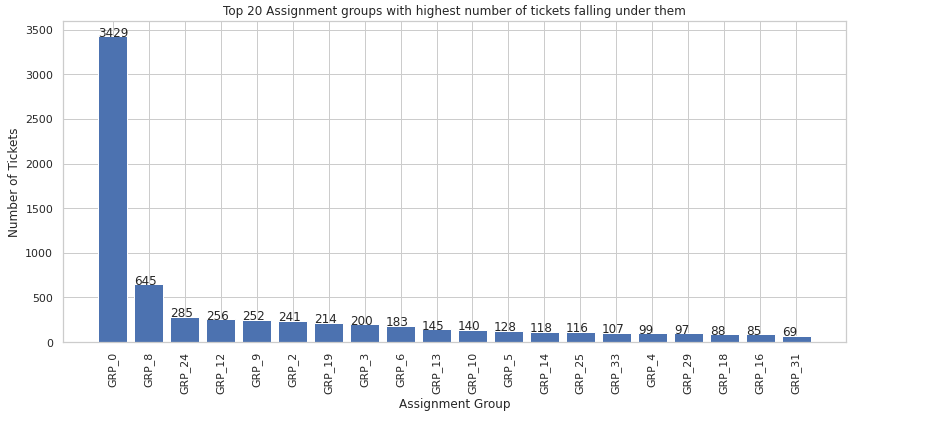
**OBSERVATIONS REGARDING TARGET CLASS**

* A large number of entries belonged to GRP\_0 (mounting to 3429 which account for ~ more than 40% of the data )

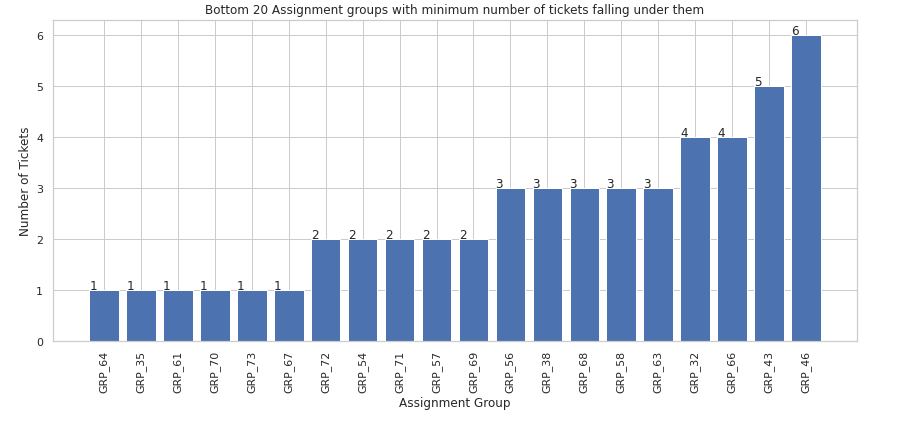
**Distribution of Target class for all groups**



* The Target class distribution is extremely skewed
* The data is too much biased towards a single group and seems to be highly imbalanced, with majority of incidents are from Group 0 followed by Group 8 , 24 , 12 , 9 , 2 and so on
* There are few classes which just have less than 10 incidents per class and even classes with just 1 or 2 incidents (samples), need to see if we can drop those rows due to the lack of samples representing those classes. They might not be of much help as a predictor
* Top 20 Assignment groups having the highest number of tickets for training the data.



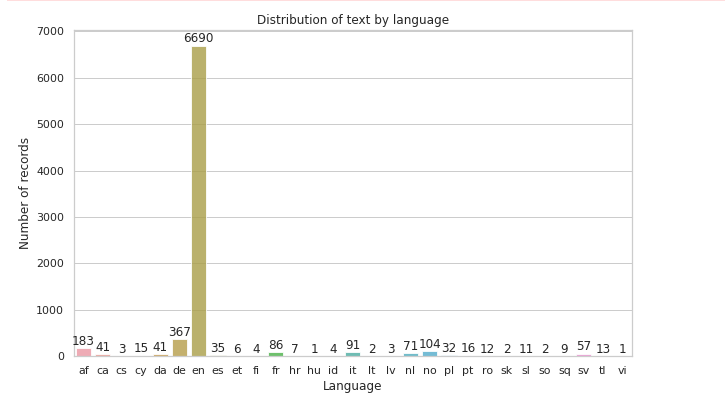
* Following are the Tickets with less number of tickets per Assignment groups.



**DATA PRE-PROCESSING**

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

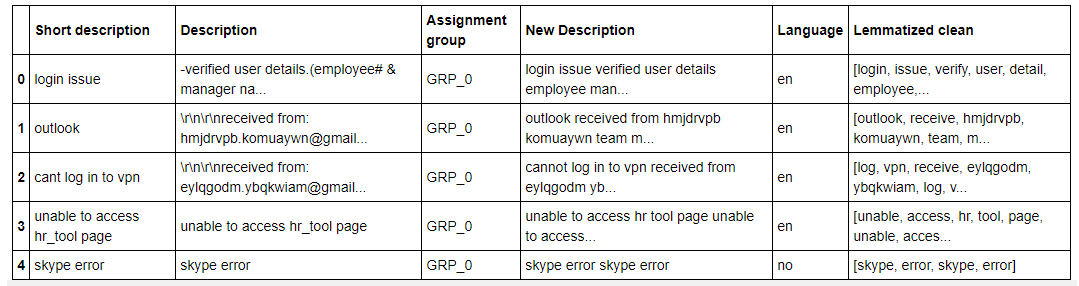
* Replaced the gibberish text using **FTFY**
* Contraction words found in the merged Description are removed for ease of word modelling
* Changed the case sensitivity of words to the common one
* Removed Hashtags and kept the words, Hyperlinks, URLs, HTML tags & non-ASCII symbols from merged fields.
* There were quite few entries with languages different from English.



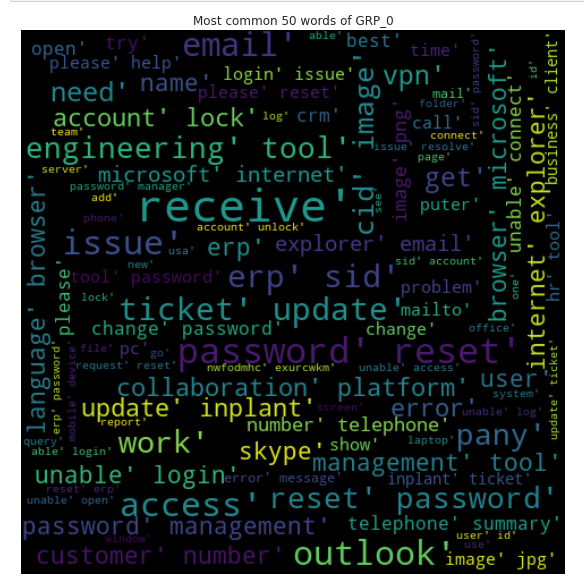
We can see that most of the tickets are in English, followed by tickets in German language.

* **Tokenization** of merged data
* **Stop words** have been removed using nltk corpus modules.
* **Lemmatization** is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is similar to Stemming but it brings context to the words. So, it links words with similar meanings to one word.

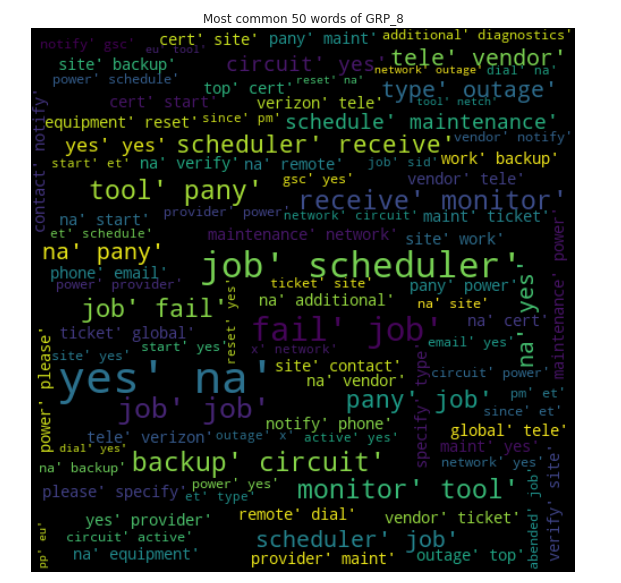
Here we have preferred Lemmatization over Stemming because lemmatization does morphological analysis of the words.



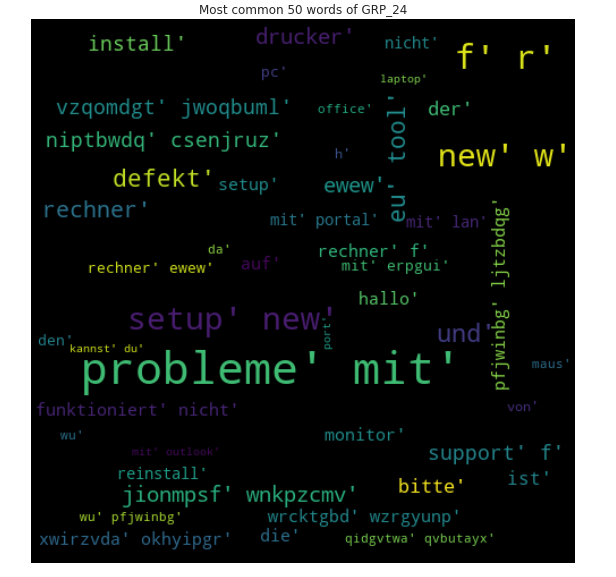
* **WordCloud** created for all available 50 groups to have more information specific to Assignment groups



**Analysis on GRP\_0 which is the most frequent group to assign a ticket to reveals that this group deals with mostly the maintenance problems such as *password, receive, reset* ,outlook, *account lock* , *login issue* , *ticket update* etc.**



**GRP\_8 seems to have tickets related *to scheduler, job failures, monitoring tool* etc.**



**GRP\_24 - Tickets are mainly in german.**