**Interim Project Report**

Project: - Automatic Ticket Assignment

**Submitted by**:

NLP Group 7

* 1. [Introduction](#_TOC_250003) 3
  2. Summary of problem statement, data and findings 4
     1. Problem statement summarization………………………………. 4
     2. Data Analysis and findings………………………………………... 4
  3. Summary of the approach to EDA and pre-processing 5
     1. EDA…………………………………………………………………….. 5
     2. Data preprocessing steps 6
  4. Deciding models and model building 6
     1. Options for NLP classification model 6
     2. Model selection 7
  5. Model performance 7
     1. Improvements to model performance xx
  6. Summary xx

[References xx](#_TOC_250000)

Appendix x

# Introduction

# 

# Summary of problem statement, data and findings

This section details about the **“Automatic Ticket Assignment”** NLP problem statement. We will discuss the dataset used for this project and will present a few findings from this data.

# 2.1. Problem statement summarization

Incident management is critical to every software organization. Users typically face issues while using the software and resolution time of such issues is a critical parameter in assessing the software usability.

Typically the incidents are reported to the service desk, which analyzes the problem and assigns to concerned L1/L2 support groups. This activity is manual and error prone and hence it takes a long time for resolution of user reported incidents.

To reduce this cycle time it is proposed to employ the NLP techniques to automatically assign the incoming incidents to the concerned groups based on the description of the incidents. This makes the problem as a classic example of classification based NLP machine learning problem.

# 2.2. Data analysis and findings

# The dataset used for this project can be found at [dataset link](https://drive.google.com/file/d/1OZNJm81JXucV3HmZroMq6qCT2m7ez7IJ/view). Here is some initial analysis -

# The dataset consists of 8500 rows and 4 columns. Each row represents information about an incident. The columns are Short description, Description, Caller and Assignment group.

# The Short description column summarizes the incident description and is short text.

# The Description column is the in detail description of the incident detailing the issue.

# The Caller column identifies the user with the first name and last name of the user.

# The Assignment group is the support group to which a particular incident was assigned to.

# Findings from the data

# Target column – The Assignment group is the target column of this problem.

# The caller column does not affect the target column directly and hence can be safely removed from the model building process.

# 

# 

# Summary of the approach to EDA and pre-processing

# This section details the data preprocessing and EDA on the dataset.

# 3.1. EDA

# Below are the observations when exploratory data analysis was performed on the given dataset –

# There are null values present for Short description and Description columns in the dataset. Specifically 8 incidents do not have values for Short description and 1 entry does not have a value for Description column.

# There are in total 74 unique support groups named from GRP\_0 to GRP\_73 present in the database. Analysis shows that there are 3976 incidents assigned to GRP\_0 and the rest 4524 among the rest 73 groups. This makes the data highly disproportionate. The ticket distribution for top 10 assignment groups is shown below -

# 

# There are 2950 unique users that have reported the incidents and the topmost user has 810 number of incidents raised. The ticket distribution for top 10 users is as shown below -

# 

# The description column for most of the tickets also consists of text like “Reported by: emailid”. This is unnecessary and can be removed as callers email id has no relationship with the target column.

# 3.2. Data preprocessing

# The fundamental aspect of machine learning that determines how good the ML model can perform is Data. If data is insufficient or is not processed correctly to target the problem, we cannot expect the model to perform well.

# We have identified following preprocessing steps for the given dataset –

# The first step is to get rid of null values from the dataset. As noted earlier there are 9 rows that null values for either of Short Description and Description columns. We will drop these rows as we cannot impute these null values.

# As noted earlier, we can get rid of text like “Reported by emailid” from the Description column. To build the email id for each user we can use the first name and last name information from caller column.

# The caller column does not affect the target column and hence can be safely removed in the end.

# The data is highly imbalanced as a lot of entries correspond to GRP\_0. Hence we will down sample the dataset for GRP\_0. Also there are 6 groups for which only one entry is present. This needs to be handled appropriately as well.

# The Description column also contains a lot of Non alphanumeric characters. We can get rid of these values and work only on alphanumeric text values.

# The text in the Description column can be converted to all lowercase.

# Numbers in the text for Description column do not imply the target column hence we can safely remove them.

# Stop words can cause a lot of noise during the word embedding’s hence we can remove them as well.

# Deciding models and model building

# This sections details the research work on choosing the right model for the given problem. Based on the model selected we explore the different parameters available to tune the model.

# 4.1. Options for NLP classification model

# There are three kind of models for NLP classification viz.

# RNN

# LSTM

# ULMFit

# 4.2. Model selection

# Model performance

# 5.1. Improvements to model performance

# 6. Summary