Support Ticket Classification Using NLP

Group No.28
Kushal Patil, Mohammad Osama, Nikita Patil

Abstract

A support ticket is a description of an incident generated by end users or development engineers that disrupts an ideal workflow of a software or website. These tickets need to be assigned to appropriate people and that too quickly to ensure customer satisfaction and efficient management of the system. This process is still manual for majority of the organizations. With the help of natural language processing, this process can be automated, and ticket assignment time can be reduced from minutes to seconds.

1. Introduction

This project's goal is to create an automated ticket assignment system utilizing ML and DL so that the teams may concentrate on resolving the issues more quickly. To expedite incident resolution and cut down, it's critical that teams are effectively assigned to the tickets. This paper provides an overview of the project's specifics and the analysis carried out on the simulated dataset. The use case for creating an automated ticket assignment system and the business value of such a solution are further explained in the parts that follow.
To guarantee there is no influence on business operations, one of the main responsibilities of any IT department is to “keep the lights on.”
IT uses the incident management procedure to accomplish the goal.An unforeseen disruption to an IT serv ice or a decline in the quality of an IT service that has an impact on Users and Business is referred to as a n incident. In order to ensure no business damage, the primary objective of the incident management proc ess is to offer a quick resolution, workarounds, or solutions that resolve the disruption and restore the serv ice to its full capacity.

Many business and IT users, end users and vendors who have access to ticketing systems, as well as integrated monitoring systems and tools, are responsible for creating incidents in the majority of enterprises. To provide higher customer satisfaction and ensure better allocation of support resources, it is crucial to assign the incidents to the correct person or unit in the support team. In many IT organizations, assigning issues to the proper IT groups is still a manual process. The process of manually assigning incidents takes time and requires labor. Human error could lead to mistakes, and improper addressing causes resource consumption to be carried out inefficiently. However, manual assignment lengthens the response and resolution times, causing a decline in user satisfaction and subpar customer service.

2. Methodology

2.1 Dataset:

The dataset has 8500 rows and 4 columns. The Dataset Consists of the following columns.

- Short Description: Short description of the issue.
- Description: Detailed Description of the is- sue.
- Caller: Caller who has generated the ticket.
- Assignment Group: which group has to resolve the issue

Some Data Cleaning and EDA is performed on the dataset before any analysis can be done on the dataset.

2.2 Data cleaning:

It is observed there are lots of inconsistencies and issues with the current dataset. so, it has to be further processed before preforming any sort of analysis. The Dataset is first checked for any inconsistencies and NULL treatment is applied. In the process it was identified that 8 records had missing data. The null values were removed to avoid any data loss. So, the data in our hand is multilingual and it is quite not possible to derive embeddings for mix of multiple language. Nearly 41 Languages were detected in the Dataset. It's decided to translate the entire dataset into a single language of English. The dataset is translated using google translation package which supports multiple languages.

2.3 Preprocessing:

Other forms of cleansing functions applied are as follows:

- converting all letters to lower or upper case
- converting numbers into words or removing numbers
- removing punctuations, accent marks and other diacritics
- removing white spaces
- removing stop words, sparse terms, and particular words
- Stemming
- Lemmatization

The above process was achieved using a custom cleaning function and Spacy NLP pipelines. The dataset is first balanced using under sampling mentioned in the section 3 using LDA and oversampled. this will ensure that the models are more accurate on all classes and has sufficient training examples for the models. for the word-embeddings gensim package was used to build a custom doc2vec model. this model provides the necessary embedding weights for training the DL models. for the Classical ML models TF-IDF (Term Frequency - Inverse Document Frequency) was used using sklearn package.

2.4 EDA:

The Callers Distribution is also imbalanced as only one caller is generating all the tickets From Figure 1 it's clear that for certain groups certain members are creating major tickets.

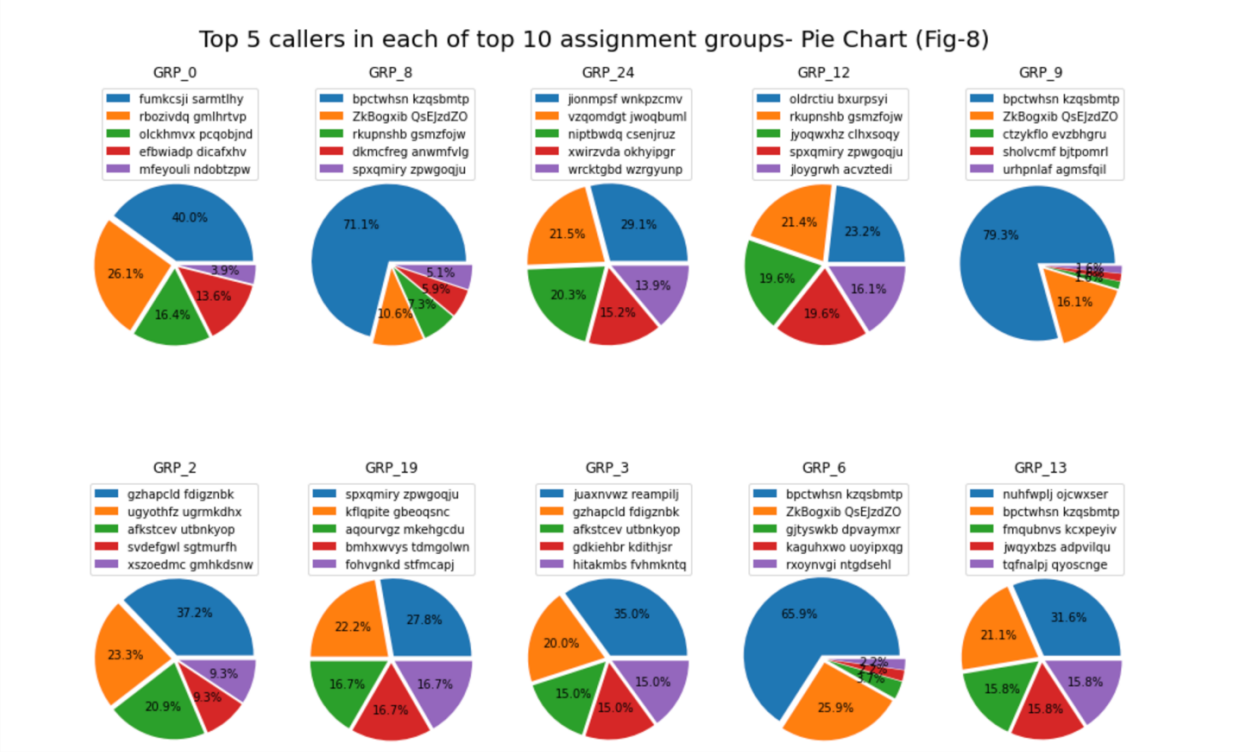

Fig. 1 Top 5 Callers in Top 10 Assignment Groups

N-Gram analysis was also performed to under- stand any form or patterns present in the words of the dataset and It's indicative from Figure 2 that the entire dataset speaks more about issues around

• password reset (1246 times)
• fail job scheduler (1614 times) • outlook (948 times)
• login (861 times)
• job fail (897 times)

According to an analysis of GRP 0, the group that receives tickets the most frequently, this group primarily deals with maintenance-related issues, such as password resets, account locks, login issues, ticket updates, etc. Most tickets from GRP 0 can be reduced by self-correcting itself by adding automatic scripts or procedures to help with resolving these frequent maintenance issues. By lowering the number of service tickets received, the company will need to employ less labor per client per hour and generate more revenue.

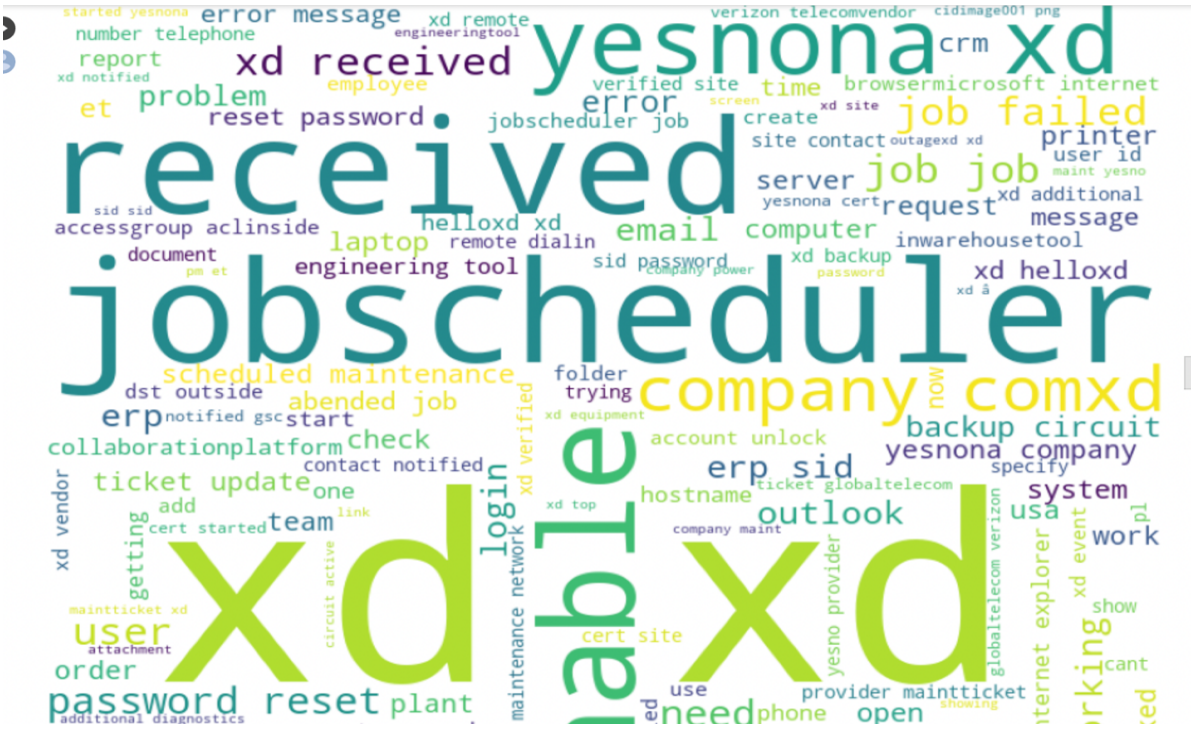

Figure 2. Word Cloud for ticket summary

The GRP 0 were now under sampled into the following categories for balancing the dataset. They are:

1. Communication Issue
2. Account/Password Reset 3. Access Issue
4. Other Issues

2.5 Feature Engineering:

Some Notable Features generated as part of the Analysis are as follows:

- word length: # of words in a document.
- lines: # of lines in a document.

To make it simple to feed into the ML models, all of these features have been combined into the Summary Feature.

3. Modelling

3.1 Initial training:

Table 1 contains a list of all the classical models that were applied during the modeling process. It is clear that, the traditional models are performing extremely accurately in this situation. On closer inspection, it becomes clear that the models have overfitted the dataset. After carefully evaluating the models' soundness, it is chosen to choose Recurrent Neural Networks with LSTM as the benchmark model for additional research.

Model	Accuracy (%)
Naïve Bayes	55
K-nearest	63
SVM	70
Decision Tree	60
Random Forest	64
CNN	53
RCNN	57
LSTM	55

Table 1. Models used with accuracy

3.2 Hyperparameter tuning and optimizations

Support Vector Machine (SVM) under statistical ML methods and neural networks outperform all other models in testing. The dataset's extreme imbalance was one of the clear causes of the models' extreme overfitting. There are 40 groups with an average number of tickets assigned per group of less than or equal to 30. The ratio of GRP 0 to all others is 47:53.

Table 2: DL Models Results

Model	Train accuracy (%)
Base Model	71.29
Iteration 1 (changing dropout)	84.22
Iteration 2 (increase LSTM units)	89.82
Iteration 3 (adding dense,dropout,batchnorm layers)	89.90
Iteration 4 (adding Adam) Iteration 1 (changing dropout)	89.90

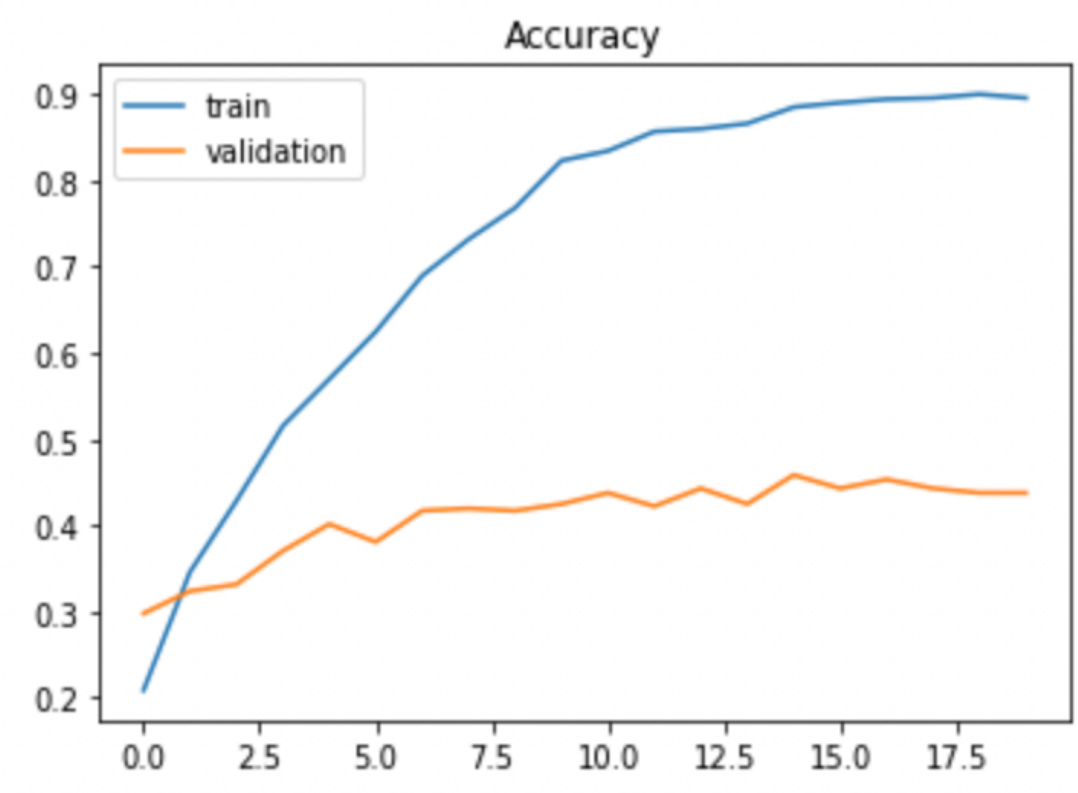

Fig 3. Train vs Validation Accuracy

From table 2, The accuracy of LSTM model can be seen. This is clear indicative of how LSTM, in the family of RNN is efficient of dealing with textual data.

- There is a bump in the model performance
- Making the dataset balanced, helped the model to be trained more accurately.
- Creating custom word embbeddings helped finding better representation of keywords of the corpus.

• By balancing the dataset, the model was trained more precisely.

• Developing unique word embeddings aided in improved representation of corpus keywords.

• The train vs. validation accuracy curve in fig.3 show that hyperparameter adjustment allowed for the discovery of a model with greater accuracy with some overfitting.

4 Conclusion

Because of the Data Modeling its very clear on how the ticketing system model can classify the tickets effectively. there are other benefits of such system which are listed below.

Benefits:

1. Increase in Customer Satisfaction.
2. Decrease in the response and resolution time.

3. Eliminate human error in Ticket Assignment. (Which was ~25% Incidents)
4. Avoid missing SLAs due to error in Ticket Assignment.
5. Eliminate any Financial penalty associated with missed SLAs.
6. Excellent Customer Service.

5. Statement of Contributions

Authors: Kushal Patil, Mohammad Osama, Nikita Patil

There are three members in the group. Each member will contribute equally to the project. Nikita Patil will be responsible for cleaning the data, EDA and Data Preprocessing, result analysis and report preparation. Mohammad Osama will be responsible for Base model implementation, result analysis and report preparation and Kushal Patil will be responsible for LSTM implementation, result analysis and report preparation.

References

1. "Googletrans api error - expecting value: line 1 column 1 (char 0)." https://stackoverflow. com/questions/49497391/googletrans-api- error-expecting-value-line-1-column-1- char-0.
2. abo Samoor, "Multilingual text (nlp) processing toolkit," 2019. https://github.com/ aboSamoor/polyglot.
3. Mor Kapronczay, "Text preprocessing in different languages for natural language processing in python," 2019. https: //medium.com/starschema-blog/text- preprocessing-in-different-languages-for-natural-language-processing-in- python-fb106f70b554.
4. Olga Davydova, "Text preprocessing in python: Steps, tools, and examples," 2018. https://www.kdnuggets.com/2018/11/text- preprocessing-python.html.