GREAT LAKES
INSTITUTE OF MANAGEMENT

&

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Post-Graduation Program in Artificial Intelligence & Machine Learning

Batch: Jan 20-B

Capstone Project Group: NLP Group 6

Great Lakes and TEXAS McCombs

Automatic Ticket Assignment

Submitted By:
Neha Vyas
Arpita Jain
Dibakar Paul
Kunal Rathod
Rohit Mundlapati

Mentor:
Saurabh Bansal

Submission Date: December 18, 2020
Submitted in Partial Fulfilment of the requirements for PGP in AIML

Table of Contents

1. Introduction

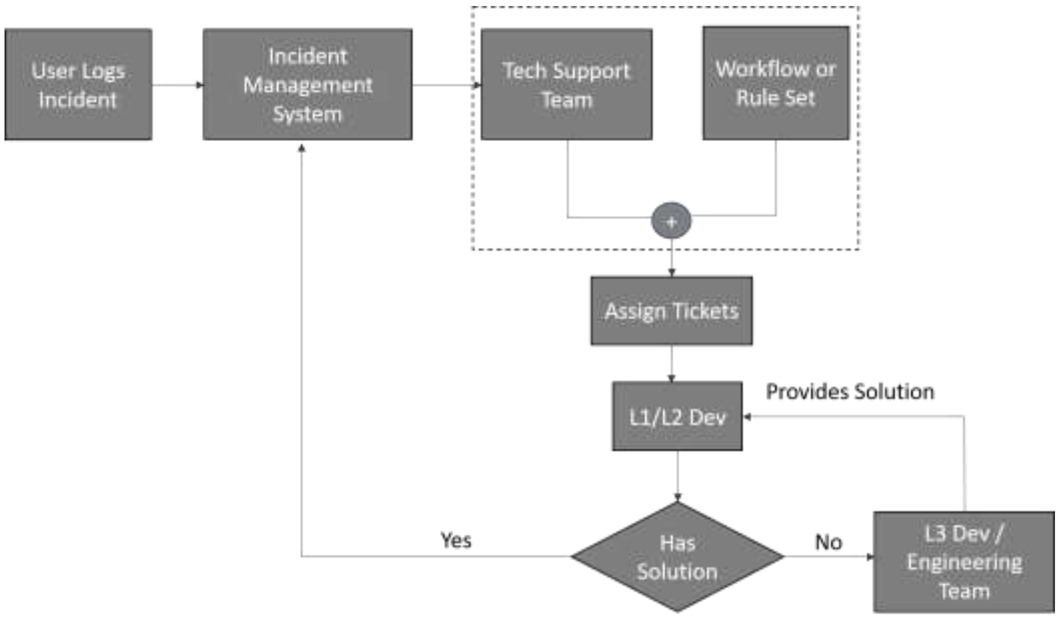
1.1 Background

The secret to business success is customer service. Good customer interactions create loyalty to the brand, drive more sales, and produce positive word-of-mouth. However, because of poor customer service, 66% of consumers switch products or services, and the key reasons for customer dissatisfaction are lack of efficiency and lack of speed. It's no wonder that consumers appreciate personalised, timely, and efficient customer service experiences in this age of technological innovation. However, Companies are falling short in this since they don't have appropriate workflows to manage all their customer questions in location.

1.2 Business Problem Statement

Consumers meet firms, from social media networks and review pages to email and live chats, any time of the day, wherever they are, by using different channels. Moreover, the growing number of users of smartphones makes it easier for customers to get in touch than ever before and you can begin to imagine how tickets for customer service are starting to pile up.

When a support ticket drops into the help desk, first it needs to be processed and assigned a group or category so that it's routed to the correct team member. This involves reading the ticket, so that agents know which category to choose.



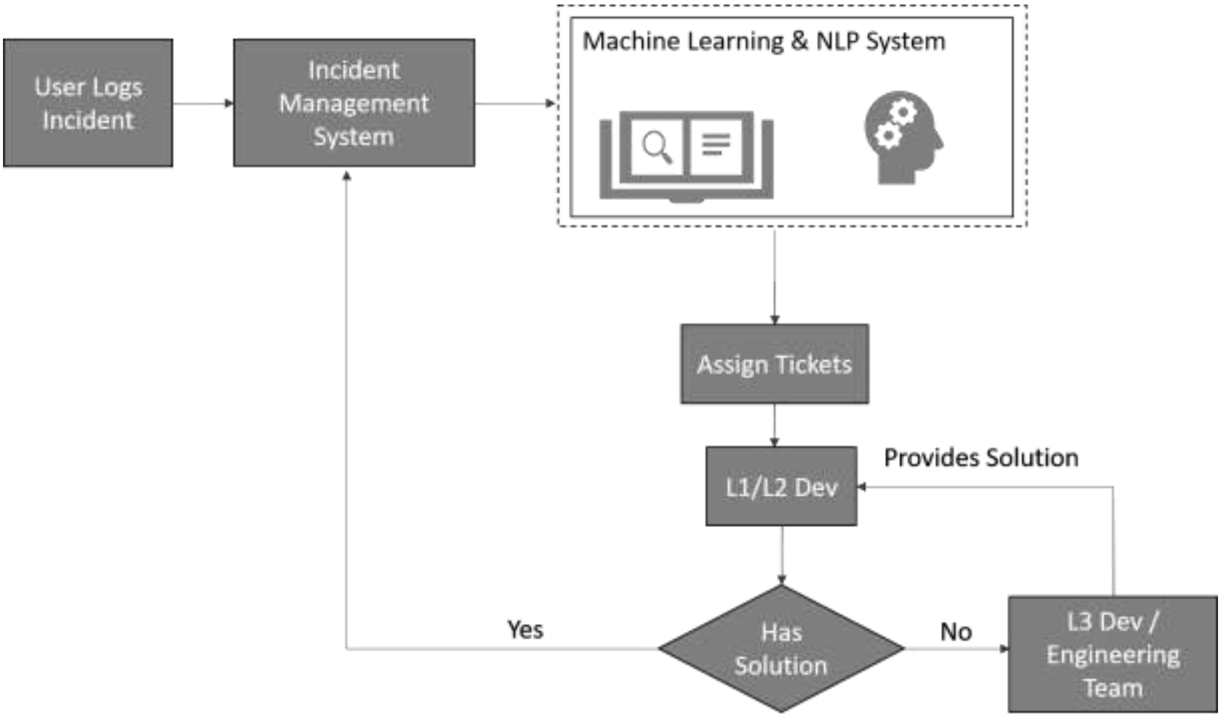
For any customer support team, the process of going through incoming tickets and allocating them to the agents best qualified to manage them is important, which suffers from below pain points in traditional incident management systems.

- Manually triaging high numbers of tickets is time consuming and extremely costly.
- It requires human efforts which may lead to inaccurate allocation of customer service agents due to human errors.

- Misaddressing of tickets leads to ineffective resource consumption.
- Around 25% of incidents are assigned to wrong functional groups.
- Additional effort needed for functional teams to reassign to right functional teams, during this process some of the incidents are in queue and are not addressed timely.
- Manual assignment increases the response and resolution times which result in user satisfaction deterioration / poor customer service.
- Cost involved in maintaining a team which works 24x7, including training cost for forming the team.

1.3 Proposed Solution

When businesses receive more customer questions through multiple channels, keeping up and maintaining lengthy queue is more difficult for support agents. Thus automatic incident assignment comes to rescue leveraging machine learning capabilities. Some of the major benefits of automatic classification of incidents are highlighted below:



1.4 Benefits of the Proposed Solution

Scalability: Categorize millions of incidents at a fraction of the cost of manual methods, save time so that agents can focus on more fulfilling tasks, and avoid inundating teams with heavy and repetitive workloads.

Availability: Available round the clock so incidents can get assigned immediately and can send a response in real-time, 24/7.

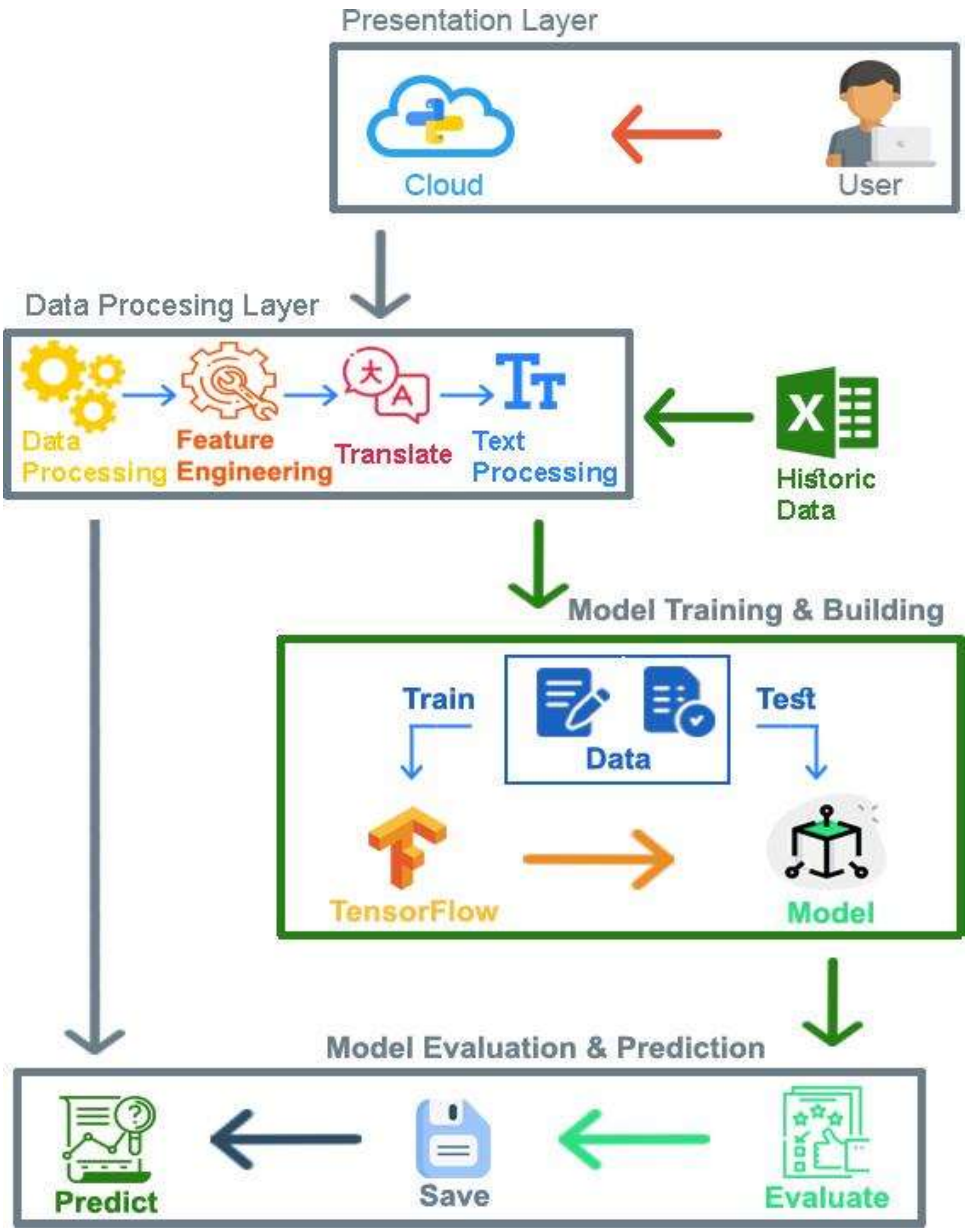
Real-time Analysis: Incident categorization data can provide valuable insights, to streamline processes of routing tickets to correct team members, and prioritizing tickets that are more urgent.

Consistent Criteria: Incident classification with machine learning enables grouping incidents accurately because it applies the same criteria to measure each set of data, plus a machine will never be subjective, lack alertness, and rush through tickets without understanding them properly.

Cost Saving: Saves cost to employ multiple employees or losing business due to customer dissatisfaction.

1.5 Architecture of the Proposed Solution

The proposed solution can further be integrated with any of the ITSM service based tools for end to end automation. A Flask microservice would be developed to deploy the model based classification as a Web Service which can further be exposed using a Restful API to communicate with any ITSM client service tool.



1.6 Data Source

The details of the data to build this classification model that can classify the tickets is available at the below link:

https://drive.google.com/open?id=1OZNJm81JXucV3HmZroMq6qCT2m7ez7IJ

The Data source consists of a single spreadsheet with different attributes of the generated tickets. The shape of this historical data is **(8500, 4)** i.e. **8500 rows & 4 columns**.

A sample data is shown below:

```
Original Data Shape: (8500, 4)
```

	Short description	Description	Caller	Assignment group
0	login issue	-verified user details.(employee# & manager na...	spxjnwir pjlcoqds	GRP_0
1	outlook	\r\n\r\nreceived from: hmjdrvpb.komuaywn@gmail...	hmjdrvpb komuaywn	GRP_0
2	cant log in to vpn	\r\n\r\nreceived from: eylqgodm.ybqkwiam@gmail...	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
10	engineering tool says not connected and unable...	engineering tool says not connected and unable...	badgknqs xwelumfz	GRP_0
11	hr_tool site not loading page correctly	hr_tool site not loading page correctly	dcqsolkx kmsijcuz	GRP_0
12	unable to login to hr_tool to sgxqsuojr xwbeso...	unable to login to hr_tool to sgxqsuojr xwbeso...	oblekmrw qltgvspb	GRP_0
13	user wants to reset the password	user wants to reset the password	iftldbmu fujslwby	GRP_0
14	unable to open payslips	unable to open payslips	epwyvjsz najukwho	GRP_0
15	ticket update on inplant_874743	ticket update on inplant_874743	fumkcsji sarmtlhy	GRP_0
16	unable to login to company vpn	\n\nreceived from: xyz@company.com\n\nhi,\n\ni...	chobktqj qdamxfuc	GRP_0
17	when undocking pc , screen will not come back	when undocking pc , screen will not come back	sigfdwcj reofwzlm	GRP_3
18	erp SID_34 account locked	erp SID_34 account locked	nqdyowsm yqerwtna	GRP_0
19	unable to sign into vpn	unable to sign into vpn	ftsqkvre bqzrupic	GRP_0
20	unable to check payslips	unable to check payslips	mrzgjdal whnldmef	GRP_0

2. Exploratory Data Analysis

2.1 Data Pre-Processing

2.1.1 Explore Data

On exploring the data source in greater detail we were able to derive below insights that will further drive the solution of this problem statement.

Inference

Short Description & Description:

These columns contain the issue description about the ticket raised. It also has various special characters, HTML tags, email ids, text from multiple languages etc. which needs to be handled as a part of the Data Pre-processing steps

Caller:

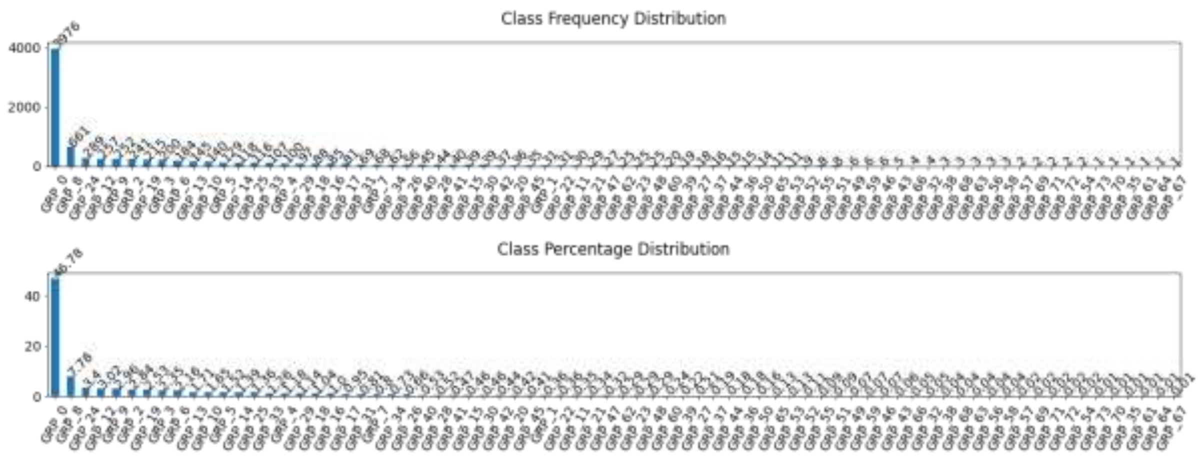
It has values that don't signify any particular feature. There is no specific pattern observed with remaining columns. For e.g. text "afkstcev utbnkyop" is getting related with multiple assignment groups like GRP_0, GRP_12, GRP_16, GRP_19, GRP_2, GRP_3, GRP_30, GRP_31, GRP_33, GRP_39, GRP_47, GRP_50 and GRP_69. Thus it doesn't help in inferencing anything significant hence we can drop it from the dataset.

Assignment group:

It has the list of groups to which the tickets are actually getting allotted. It will be our Target column since we need to assign the incoming tickets to these groups.

2.1.2 Explore Target Column

To solve any problem statement we have to understand what is actual outcome that is expected based on which the proposed solution is designed. In our case, the Target is Assignment Group which we have to predict from the available data. We will understand it further with the help of a graph



Inference

- After analyzing the graph & value counts it is evident that the Target column distribution is extremely skewed. There are 74 distinct assignment groups available in the dataset
- GRP_0 is the most dominant assignment group with accounts for 46.78% of data

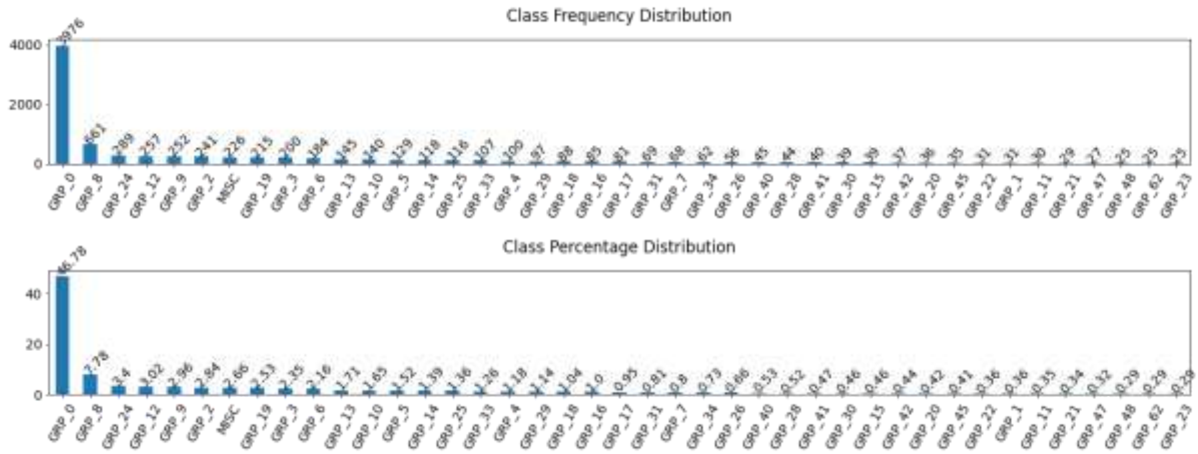
- There are 26 groups that have less than 10 assigned to them. Out of which 11 groups like GRP_58, GRP_57, GRP_69, GRP_70, GRP_67, GRP_71, GRP_72, GRP_64, GRP_61 and GRP_35 etc. have only 1 or 2 entries
- Thus, it implies that we have a very imbalanced data in hand

Challenge

Reduce the imbalance in data

Solution

We merge the groups with small entries which are contributing less than 0.25 % (or < 25 entries) to the data into a miscellaneous group. This will reduce the skewness to an extent. This has a business implication wherein the miscellaneous group has to be manually managed by someone and assign the tickets to the individual groups that are merged. Post merging, we have the following outcome.



We were able to bring down the distinct number of groups from 74 to 41.

2.1.3 Duplicate Data

Challenge

Data consist of good amount of duplicate rows as shown below

	Short description	Description	Assignment group
51	call for ecwtrjnq jpecxuty	call for ecwtrjnq jpecxuty	GRP_0
81	erp SID_34 account locked	erp SID_34 account locked	GRP_0
123	unable to display expense report	unable to display expense report	GRP_0
157	ess password reset	ess password reset	GRP_0
229	call for ecwtrjnq jpecxuty	call for ecwtrjnq jpecxuty	GRP_0
235	erp SID_34 account unlock and password reset	erp SID_34 account unlock and password reset	GRP_0
242	windows password reset	windows password reset	GRP_0
274	windows account locked	windows account locked	GRP_0
301	windows password reset	windows password reset	GRP_0
312	erp SID_34 account unlock	erp SID_34 account unlock	GRP_0
333	windows password reset	windows password reset	GRP_0
380	unable to login to erp SID_34	unable to login to erp SID_34	GRP_0
391	password reset request	password reset request	GRP_0
393	password reset	password reset	GRP_0
422	password reset	password reset	GRP_0

GRP_0	547
GRP_8	16
GRP_17	13
MISC	4
GRP_24	4
GRP_21	1
GRP_6	1
GRP_4	1
GRP_15	1
GRP_5	1
GRP_12	1
GRP_19	1

Solution

We exclude these rows from all further processing. There are 591 such duplicate rows in the dataset which are excluded. This reduces the shape of original data to (7909, 3)

2.1.4 Missing Data

Challenge

The data has very few missing values which can create issues later while processing. There are 6 missing values in the dataset as shown below

	Short description	Description	Assignment group
2604	NaN	\r\n\r\nreceived from: ohdrnswl.rezuibdt@gmail...	GRP_34
3383	NaN	\r\n-connected to the user system using teamvi...	GRP_0
3906	NaN	-user unable tologin to vpn.\r\n-connected to...	GRP_0
3924	NaN	name:wvqgbdhm fwchqjor\nlanguage:\nbrowser:mic...	GRP_0
4341	NaN	\r\n\r\nreceived from: eqmuniov.ehxkcbgj@gmail...	GRP_0
4395	i am locked out of skype	NaN	GRP_0

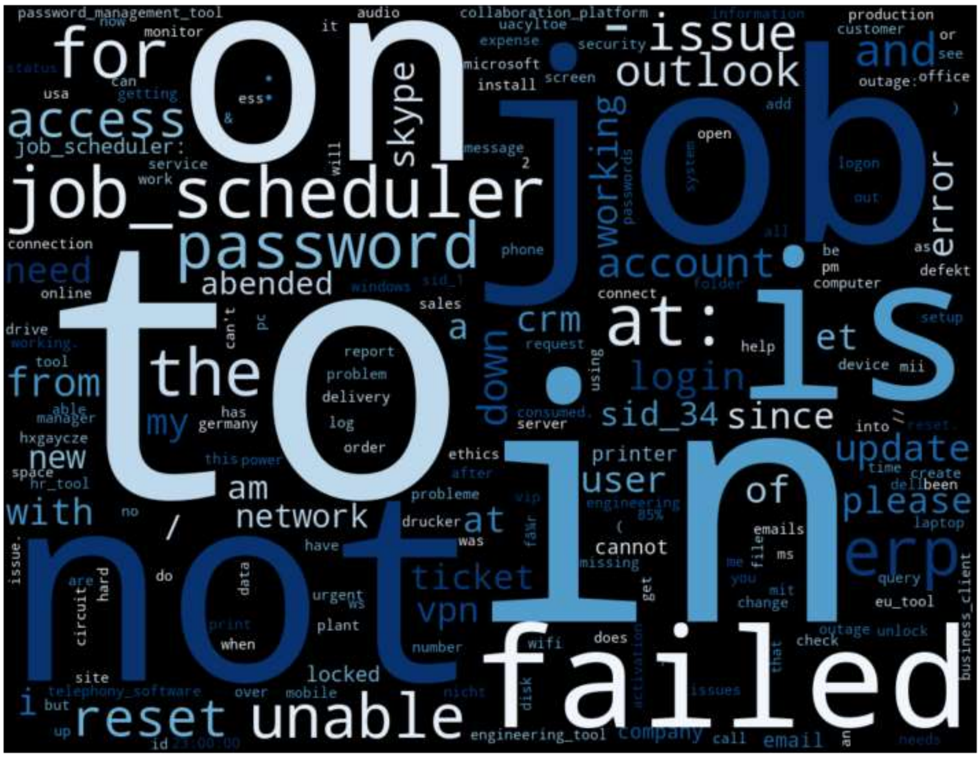
Solution

We replace the missing values with empty string. This enables us to handle the missing values as well as retain the information available within these rows.

2.2 Text Features

2.2.1 Short Description

It consists of brief description about the ticket which could also be considered as the title of the ticket. We can visualize its features using a word cloud.



The graph indicates that the top words in Short Description are stop words i.e. to, in, at, on, is etc. The other prominent words are job, job_scheduler, failed, unable, reset, erp etc.

2.2.2 Description

It consists of detailed description of the issue raised via ticket. In some cases, the short description & description have the same text as well. We can visualize its features using a word cloud.

The graph indicates that the top words in Description also include stop words i.e. to, in, the, from, is etc. The other prominent words include received, company, password, outlook, gmail etc.

Challenge

It is evident from the word clouds that both the columns i.e. Short Description & Description are focusing on different set of words (excluding the stop words). Indicating that both are required to predict the correct assignment group.

Solution

Looking at the data we could imply that both columns gives some information about the ticket. In some cases both are same & in some cases short description has proper text & description doesn't have any meaningful text. Hence, we merge Short Description & Description as one column.

This merging leads to emergence of duplicate words which we will eliminate first to make the Full Description look more appropriate and clean. The word cloud for this merged column is given below.

The word cloud of Full Description clearly show that it is an amalgamation of both Short Description & Description. Along with stop words, we can also observe unicode characters, numbers and special characters in the Word Cloud.

2.3 Text Pre-Processing

Pre-processing is mandatory to get the data in a consistent format. We will perform the following list of pre-processing on the merged text column Full Description (refer it as Description)

- Handle Unicode Characters
- Make text lowercase
- Remove hyperlinks
- Remove punctuation
- Remove numbers
- Remove special characters
- Remove HTML tags, unwanted spacing
- Specific words etc.

2.3.1 Unicode Characters

Challenge
The text column in the dataset contains Unicode characters as shown in the example below

é’å²›å…´åˆæœºç”µshipment notificationé‚®ç®±è®¾ç½® from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dearï¼Œ pls help to update customer 4563729890 shipment notification email address : abcdegy@gmail.com b.

Solution

We have used a python library i.e. **ftfy** to convert these unicode characters into their respective non-English language format.

青岛兴合机电shipment notification邮箱设置 from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dear, pls help to update customer 4563729890 shipment notification email address : abcdegy@gmail.com b.

2.3.2 Translation

Challenge

The text column also has data in other non-English languages along the converted text from unicode characters mentioned earlier.

青岛兴合机电shipment notification邮箱设置 from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dear, pls help to update customer 4563729890 shipment notification email address : abcdegy@gmail.com b.
an mehreren pc`s lassen sich verschiedene prgramdntyme nicht öffnen. bereich cnc.
无法登陆hr_tool考勤系统 显示java插件无法加载,所需版本1.8.0.-45或更高版本。

Solution

We have used another python library i.e. **google_translate_new** for language translation from non-English to English words. This also improves the overall features since we have more information about the ticket after translation.

Qingdao Xinghe Electromechanical shipment notification email setting from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dear, pls help to update customer 4563729890 shipment notification email address: abcdegy@gmail.com b.
Different programs cannot be opened on several pc's. area cnc.
Unable to log in to the hr_tool attendance system. It shows that the java plug-in cannot be loaded. The required version is 1.8.0.-45 or higher.

2.3.3 Text Cleansing

Challenge

As mentioned earlier, the text column are having numbers, email ids, special characters, hyperlinks, punctuations, HTML tags, unwanted space and keywords like to:, from:, received from: and subject etc. These are not useful in predicting in fact it can lead to extra word count which might affect the model performance.

Qingdao Xinghe Electromechanical shipment notification email setting from: sent: friday, october 28, 2016 7:20 am to: nwfodmhc exurcwkm subject: re: dear, pls help to update customer 4563729890 shipment notification email address: abcdegy@gmail.com b.
Different programs cannot be opened on several pc's. area cnc.
Unable to log in to the hr_tool attendance system. It shows that the java plug-in cannot be loaded. The required version is 1.8.0.-45 or higher.

Solution

We have cleaned the data using **Regular Expressions** as well as with the help of a python library named **nostril**. This library enables us to remove some of the random words but not all.

qingdao xinghe electromechanical shipment notification email setting from sent friday october am to re dear pls help to update customer shipment notification email address b
different programs cannot be opened on several pc area cnc
unable to log in to the hr tool attendance system it shows that the java plug in cannot be loaded the required version is or higher

2.3.4 Lemmatization

Challenge

In the description text we can see different transformation of the same word. E.g. shows, showed, showing etc. We as human can understand that all these words are related to the act of show but for machine these are separate words and are treated differently.

prpf instead of prir for usa location and order operation mii showed the in status looking at erp was partially confirmed followed by automatic pause which adds change log shows all entries were done miiadmin it weird since erp is showing expected status provided but itself does not make sense reporting this issue so team can do deep dive come up with solution

Solution

We can resolve these particular scenario by performing Lemmatization of data. It is the process of grouping together the inflected forms of a word so they can be analyzed as a single item. We used a python library called **spacy** for this task. Thus, facilitates in reducing the number of unique word counts and eventually increase the overall performance of the model.

prpf instead of prir for usa location and order operation mii show the in status look at erp be partially confirm follow by automatic pause which add change log show all entry be do miiadmin it weird since erp be show expect status provide but itself do not make sense report this issue so team can do deep dive come up with solution

2.3.5 Stop words

Challenge

As observed in the word cloud and examples shared earlier, the text contains words like to, from, in, the, is, on, and, for, at etc. These are the most frequent words across the whole corpus of data available. In case of text classification, these words are of very little help in classifying the correct class. In fact it increases the computation time of the model.

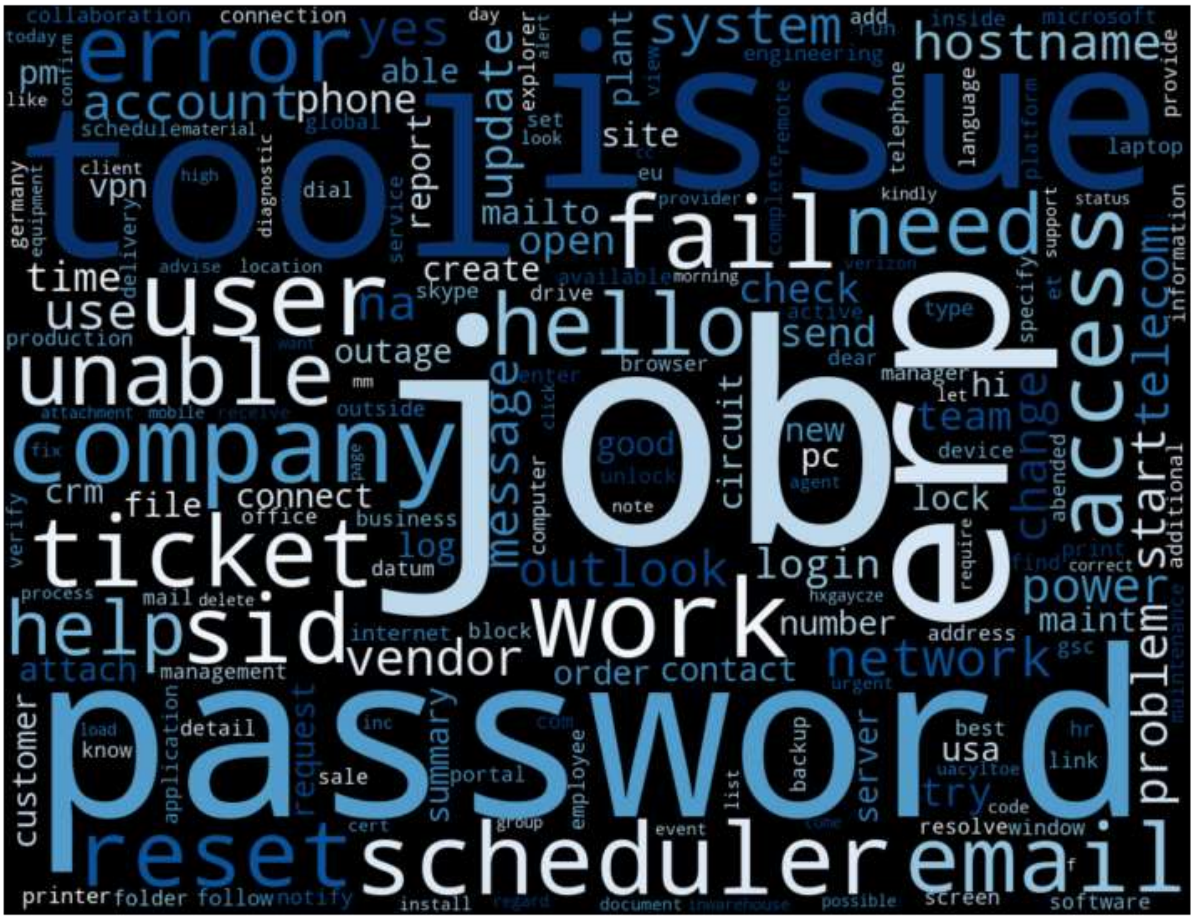
prpf instead of prir for usa location and order operation mii show the in status look at erp be partially confirm follow by automatic pause which add change log show all entry be do miiadmin it weird since erp be show expect status provide but itself do not make sense report this issue so team can do deep dive come up with solution

Solution

In text processing, these words are known as Stop words. We have excluded these words using the same python library **spacy**. It tokenizes each sentence into words and identifies that each token/word is a stop word or not.

prpf instead prir usa location order operation mii status look erp partially confirm follow automatic pause add change log entry miiadmin weird erp expect status provide sense report issue team deep dive come solution

The word cloud after performing various text preprocessing looks as follows



Now, it looks much cleaner than the original one. The important words like job, password, scheduler, tool, issue, ticket, work, email, reset etc. are correctly visible in the figure. To conclude, the text preprocessing task is completed successfully and the actual objective is achieved.

2.4 Data Resampling

Challenge

We are still facing the challenge of imbalanced data. The distribution is very much skewed and it needs to be rectified before passing to models.

Solution

We have implemented a strategic solution to carry out resampling of data. In this we are following a 3-step approach

1. Divide the original dataset into 2 i.e. GRP_0 dataset & other groups dataset
2. Resample both the datasets separately
    a. Under Sample the GRP_0 dataset by 70% i.e. we will take only 30% of its data for processing
    b. Over Sample the other groups dataset based on maximum value count of a group within that dataset.
3. Concatenate both the datasets to create a single sampled dataset

It has helped us reduce the imbalance of data on GRP_0 to 6.76% from the original value of 45.78%. The distribution graph depicts the same.

3. Model Building

3.1 Nature of Problem

With reference to the Business Problem statement for this exercise, we are to analyze the text input from an end user describing an issue and thereafter predict the appropriate support group to assign the ticket automatically. So it is prudent that the nature of the problem is Multiclass Text Classification utilizing NLP capabilities.

Text Classification (a.k.a. text categorization or text tagging) is the process of classifying documents into predefined categories based on their content. It is the automated assignment of Natural Language texts to predefined categories.

Text Classifiers can be used to organize structure and categorize pretty much any kind of text. It takes text as an input, analyze its content and then automatically assign relevant tags.

Machine Learning, Natural Language Processing (NLP) and other AI-guided techniques are being used to automatically classify text in a faster, more cost-effective and more accurate manner.

3.2 Models

Multiclass Text Classification can be achieved using a number of different ML or DL based models for achieving optimal classification output.

3.2.1 Machine Learning Models

3.2.1.1 Logistic Regression

It is a supervised learning classification algorithm used to predict the probability of a target variable. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, text classification etc. Logistic Regression is the most preferred ML algorithm when the dependent variables are Categorical.

We will use Multinomial type of Logistic Regression in which the dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. It is easy to implement and does not require too many computational resources.

3.2.1.2 K-Nearest Neighbors

It is a supervised machine learning algorithm that can be used to solve both classification and regression problems. The algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

Very effective for text datasets and it naturally handles the multi-class classification problem but computationally this model is very expensive.

3.2.1.3 Support Vector Machine

SVMs are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. They are extremely popular because of their ability to handle multiple continuous and categorical variables. They are generally used in classification problems.

Robust against overfitting problems especially for text dataset due to its representation in high dimensional space. It also results in lack of transparency and memory complexity. Still usually offers good accuracy and uses less memory.

3.2.1.4 Naive Bayes

It is a classification technique based on applying Bayes' theorem with a strong assumption that all the predictors are independent to each other i.e. the presence of a feature in a class is independent to the presence of any other feature in the same class.

We will use Multinomial Naïve Bayes classifier for our problem. It works very well with text data. Very easy to implement and converges faster. It requires less training data and it highly scalable in nature.

3.2.1.5 Decision Trees

Decision tree analysis is a predictive modelling tool which can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. The two main entities of a tree are **decision nodes**, where the data is split and **leaves**, where we got outcome.

Very fast algorithm for both learning and prediction. It can easily handle categorical features. But it is extremely sensitive to data and can over fit easily.

3.2.1.6 Random Forest

Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Very flexible, less variance then decision trees, works well with large range of data and possess very high accuracy. Complexity is high, constructing it is harder, more computational resources required and prediction process is time-consuming.

3.2.1.7 Bagging

It is an ensemble learning method that combines the weak learners. It often considers homogeneous weak learners, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process or sits on top of the majority voting principle. It is also known as Bootstrap Aggregating.

The recursive nature of picking the samples at random with replacement can improve the accuracy of an unstable machine learning model. Additionally, it prevents overfitting and makes your model generalize better on unseen data. On the downside, it has large computational complexity and requires careful tuning.

3.2.1.8 Boosting

It is also an ensemble learning method that combines the weak learners. It often considers homogeneous weak learners, learns them sequentially in a very adaptive way (a base model depends on the previous ones) and combines them following a deterministic strategy.

The core concept of boosting focuses on those specific training samples that are hard to classify. When a weak-classifier misclassifies a training sample, the algorithm then uses these very samples to improve the performance of the ensemble. It is known to decrease bias. On the downside, it also has large computational complexity.

We will use Gradient Boosting & XG Boosting Classifier for model building.

3.2.2 Deep Learning Models

3.2.2.1 LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

LSTM networks are well-suited to classifying, processing and making predictions. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.

3.2.2.2 Bi-Directional LSTM

Bidirectional LSTM are really just putting two independent RNNs together. This structure allows the networks to have both backward and forward information about the sequence at every time step.

Unlike LSTMs, Using bidirectional will run your inputs in two ways, one from past to future and one from future to past. In this way both the layers are co-trained simultaneously thus helping to maintain the context of input.

3.2.2.3 GRU

GRU supports gating and a hidden state to control the flow of information. Unlike LSTM, GRU does not have an output gate and combines the input and the forget gate into a single update gate.

GRU is less complex than LSTM and is significantly faster to compute and the results are almost similar to LSTMs.

3.3 Model Selection

Model selection is the process of choosing one among many candidate models for a predictive modeling problem. There may be many competing concerns when performing model selection beyond model performance, such as complexity, maintainability, and available resources. There are two main classes of techniques to approximate the ideal case of model selection

3.3.1.1 Probabilistic Measures

It involves analytically scoring a candidate model using both its performance on the training dataset and the complexity of the model. A hold-out test set is typically not required. For e.g. a highly biased model like the linear regression algorithm is less complex and on the other hand, a neural network is very high on complexity.

A fair bit of disadvantage however lies in the fact that probabilistic measures do not consider the uncertainty of the models and has a chance of selecting simpler models over complex models.

3.3.1.2 Resampling Methods

It estimate the performance of a model on out-of-sample data. These are simple techniques of rearranging data samples to inspect if the model performs well on data samples that it has not been trained on. In other words, resampling helps to understand if the model will generalize well.

3.3.1.2.1 Random Splits

Random Splits are used to randomly sample a percentage of data into training and testing sets. The advantage of this method is that there is a good chance that the original population is well represented in all the three sets. In more formal terms, random splitting will prevent a biased sampling of data.

We are going to use the train/test split method which is random splits technique under resampling methods for model selection.

3.4 Model Evaluation

Models can be evaluated using multiple metrics. However, the right choice of an evaluation metric is crucial and often depends upon the problem that is being solved. A clear understanding of a wide range of metrics can help the evaluator to chance upon an appropriate match of the problem statement and a metric.

Evaluation metrics considered for our problem statement are as follows

Accuracy

It is the simplest metric and can be defined as the number of test cases correctly classified divided by the total number of test cases. It can be applied to most generic problems but is not very useful when it comes to unbalanced datasets.

Precision (Specificity)

Precision is the metric used to identify the correctness of classification. It is the ratio of correct positive classifications to the total number of predicted positive classifications. The greater the fraction, the higher is the precision, which means better is the ability of the model to correctly classify the positive class.

Recall (Sensitivity)

Recall tells us the number of positive cases correctly identified out of the total number of positive cases. Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions. It gives a measure of how accurately our model is able to identify the relevant data.

F1-Score

F1 score is the harmonic mean of Recall and Precision and therefore, balances out the strengths of each. It is useful in cases where both recall and precision can be valuable – like in the identification of plane parts that might require repairing. Here, precision will be required to save on the company's cost (because plane parts are extremely expensive) and recall will be required to ensure that the machinery is stable and not a threat to human lives.

3.5 Model Performance

Using the above evaluation metrics we have evaluated all the models. The performance of each model is shown in the tables below

Machine Learning Models

The score table below shows that the Random Forest is the best performing classifier with accuracy of 92.32% and f1 score of 92.24% closely followed by Support Vector Machine with 91.91% and Gradient Boosting classifier with 91.64%.

	Accuracy Score	Precision Score	Recall Score	F1 Score	Training Score	Testing Score
Logistic Regression	0.883585	0.899501	0.883585	0.883433	0.94989	0.883585
K-Nearest Neighbors	0.623478	0.904646	0.623478	0.677002	0.739534	0.623478
Support Vector Machine	0.919143	0.923554	0.919143	0.918736	0.980687	0.919143
Naive Bayes	0.811982	0.89913	0.811982	0.821002	0.888193	0.811982
Decision Trees	0.90112	0.901714	0.90112	0.89723	0.991022	0.90112
Random Forest	0.923283	0.926593	0.923283	0.922423	0.991022	0.923283
Bagging	0.905504	0.906302	0.905504	0.902087	0.986742	0.905504
Gradient Boosting	0.916464	0.916813	0.916464	0.915082	0.98883	0.916464
XG Boosting	0.841695	0.844947	0.841695	0.838281	0.907715	0.841695

Deep Learning Models

The result & score table below shows that GRU model has the best validation accuracy and testing accuracy with 85.75% and 86.97% respectively. Also, the f1 score for GRU is 86.72%. It is closely followed by Bi-directional LSTM model with 86.33% and LSTM model is also not very far with 85.02% accuracy.

	Training Loss	Training Accuracy	Val. Loss	Val. Accuracy	Testing Loss	Testing Accuracy
LSTM	0.442924	0.867023	0.612583	0.847077	0.681609	0.850219
Bi-Directional LSTM	0.40091	0.880465	0.633488	0.856994	0.601689	0.863371
GRU	0.385964	0.882031	0.696433	0.857516	0.674826	0.869703

	Accuracy Score	Precision Score	Recall Score	F1 Score
LSTM	0.850219	0.843768	0.850219	0.843581
Bi-Directional LSTM	0.863371	0.859708	0.863371	0.859575
GRU	0.869703	0.870307	0.869703	0.867245

4. Future Scope & Enhancements

- Data Augmentation for Resampling and check how the model performs
- Hyper Parameter tuning to improve overall model performance
- Try out other Resampling methods for model selection and monitor overall model performance
- Explore other evaluation metrics especially for deep learning models
- With reference to the given business problem statement, it has been mentioned that around 54% of the incidents are being resolved by the L1/L2 team and remaining unresolved issues gets assigned to the L3 group. From the historical data we have seen that GRP_0 and GRP_8 contribute to 54% of the incidents which are being assigned to L1/L2 teams. Hence we would take an alternate approach of first classifying the Incident tickets into a small classification group of L1, L2 and L3 classes and then we would further classify the incidents into the given assignment groups