INTERIM REPORT

Capstone Project AIML

Automatic Ticketing System

Mentor:

- Mr. Srihari

Project Group Members:

- Shripad Vijay Jangam
- Harshit Porwal
- Priyadharshini P J
- Deepanshu Dhingra
- Pranav Kokate

ABSTRACT:

One of the key activities of any IT function is to “Keep the lights on” to ensure there is no impact to the Business operations. IT leverages Incident Management process to achieve the above Objective. An incident is something that is unplanned interruption to an IT service or reduction in the quality of an IT service that affects the Users and the Business. The main goal of Incident 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. In most of the organisations, incidents are created by various Business and IT Users, End Users/ Vendors if they have access to ticketing systems, and from the integrated monitoring systems and tools.

Assigning the incidents to the appropriate person or unit in the support team has critical importance to provide improved user satisfaction while ensuring better allocation of support resources. The assignment of incidents to appropriate IT groups is still a manual process in many of the IT organisations.

Manual assignment of incidents is time consuming and requires human efforts. There may be mistakes due to human errors and resource consumption is carried out ineffectively because of the misaddressing. On the other hand, manual assignment increases the response and resolution times which result in user satisfaction deterioration / poor customer service.

BUSINESS DOMAIN VALUE:

In the support process, incoming incidents are analysed and assessed by organization’s support teams to fulfil the request. In many organizations, better allocation and effective usage of the valuable support resources will directly result in substantial cost savings.

Currently the incidents are created by various stakeholders (Business Users, IT Users and Monitoring Tools) within IT Service Management Tool and are assigned to Service Desk teams (L1 / L2 teams). This team will review the incidents for right ticket categorization, priorities and then carry out initial diagnosis to see if they can resolve. Around ~54% of the incidents are resolved by L1 / L2 teams. In case L1 / L2 is unable to resolve, they will then escalate / assign the tickets to Functional teams from Applications and Infrastructure (L3 teams). Some portions of incidents are directly assigned to L3 teams by either Monitoring tools or Callers / Requestors. L3 teams will carry out detailed diagnosis and resolve the incidents. Around ~56% of incidents are resolved by Functional / L3 teams. In case if vendor support is needed, they will reach out for their support towards incident closure.

L1 / L2 needs to spend time reviewing Standard Operating Procedures (SOPs) before assigning to Functional teams (Minimum ~25-30% of incidents needs to be reviewed for SOPs before ticket assignment). 15 min is 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.

OBJECTIVE:

Create a multi-class classifier that analyses text to classify tickets.

With the help of strong AI approaches that can classify occurrences into the appropriate functional groups, organisations may cut the time it takes to resolve issues and focus on more productive tasks.

DATASET DETAILS:

Number of Columns:- 4

Number of Rows:- 8500

Number of Rows with garbled Text:- 828

Number of Rows with non-English language:- 824

APPROACH AND DESIGN:

Imported all necessary packages, and installed all the libraries.

Loaded the dataset and performed various operations as per the problem statement.

LIBRARIES:

Following are the external libraries which are used in the implementation of the project.

Library	Purpose
Numpy	Numerical Calculation
Pandas	Data Handling
Plotly	Data Visualization
Seaborn	Data Visualization
Keras	Sequential Modelling
Sklearn	Tools & Evaluation Metrics
NLTK	NLP Toolkit
Mojibake	Text Pre-processing
Goslate	Google Translate API for Text Pre-processing

DESIGN:

1. Data Loading – Loading data from the dataset
2. Data Clean-up
3. Data Pre-processing
4. Exploratory Data Analysis (EDA)
5. Model Selection

Data Cleaning Flow:

1. Null Value Treatment
2. Special Character Handling
3. Text Translation
4. Merging Attributes

IMPLEMENTATION:

1. Loaded the dataset, and checked the first 5 and last 5 rows in the dataset.

Displaying first 5 records of Dataset

```
df = pd.read_excel('input_data.xlsx', )
df.head()
```

	Short description	Description	Caller	Assignment group
0	login issue	-verified user details.(employee# & manager na...	spxjnwir pjlcoqds	GRP_0
1	outlook	_x000D_\n_x000D_\nreceived from: hmjdrvpb.komu...	hmjdrvpb komuaywn	GRP_0
2	cant log in to vpn	_x000D_\n_x000D_\nreceived from: eylqgodm.ybqk...	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

Displaying last 5 records of Dataset

```
df.tail()
```

	Short description	Description	Caller	Assignment group
8495	emails not coming in from zz mail	_x000D_\n_x000D_\nreceived from: avglmrts.vhqm...	avglmrts vhqmtiua	GRP_29
8496	telephony_software issue	telephony_software issue	rbozivdq gmlhrtvp	GRP_0
8497	vip2: windows password reset for tifpdchb pedx...	vip2: windows password reset for tifpdchb pedx...	oybwdsgx oxyhwrfz	GRP_0
8498	machine nÃ£o estÃ¡ funcionando	i am unable to access the machine utilities to...	ufawcgob aowhxjky	GRP_62
8499	an mehreren pc`s lassen sich verschiedene prgr...	an mehreren pc`s lassen sich verschiedene prgr...	kqvbrspl jyzoklfx	GRP_49

2. Looked for the other info of the dataset.

Printing information about the data

```
print('No of rows:\033[1m', df.shape[0], '\033[0m')
print('No of cols:\033[1m', df.shape[1], '\033[0m')
```

```
No of rows: 8500
No of cols: 4
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8500 entries, 0 to 8499
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Short description  8492 non-null   object
 1   Description        8499 non-null   object
 2   Caller             8500 non-null   object
 3   Assignment group   8500 non-null   object
dtypes: object(4)
memory usage: 265.8+ KB
```

Printing description of the dataset with various summary and statistics

```
df.describe()
```

	Short description	Description	Caller	Assignment group
count	8492	8499	8500	8500
unique	7481	7817	2950	74
top	password reset	the	bpctwhsn kzqsbmtp	GRP_0
freq	38	56	810	3976

OBSERVATIONS:

1. The dataset consists of 8500 rows and 4 columns.
2. All columns are of type object containing textual information.
3. There are 8 null/missing values present in the short description and 1 null/missing values present in the description column.
4. Password reset is one of the most occurring tickets which reflects in the short description column

3. Looked for null and missing values and treated them accordingly.

Finding out the NULL values in each column

```
df.isnull().sum()
```

```
Short description    8
Description          1
Caller               0
Assignment group     0
dtype: int64
```

```
df[pd.isnull(df).any(axis=1)]
```

	Short description	Description	Caller	Assignment group
2604	NaN	_x000D_\n_x000D_\nreceived from: ohdrnswl.rezu...	ohdrnswl rezuibdt	GRP_34
3383	NaN	_x000D_\n-connected to the user system using t...	qftpazns fxpnytmk	GRP_0
3906	NaN	-user unable tologin to vpn._x000D_\n-connect...	awpcmsey ctdiuqwe	GRP_0
3910	NaN	-user unable tologin to vpn._x000D_\n-connect...	rhwsmefo tvphyura	GRP_0
3915	NaN	-user unable tologin to vpn._x000D_\n-connect...	hxripljo efzounig	GRP_0
3921	NaN	-user unable tologin to vpn._x000D_\n-connect...	cziadygo veiosxby	GRP_0
3924	NaN	name:wvqgbdhm fwchqjor\nlanguage:\nbrowser:mic...	wvqgbdhm fwchqjor	GRP_0
4341	NaN	_x000D_\n_x000D_\nreceived from: eqmuniov.ehxk...	eqmuniov ehxkcbgj	GRP_0
4395	i am locked out of skype	NaN	viyglzfo ajtfzpkb	GRP_0

OBSERVATIONS:

1. We have various ways of treating the NULL/Missing values in the dataset such as:
    a. Replacing them with empty string
    b. Replacing them with some default values
    c. Duplicating the Short description and Description values wherever one of them is Null
    d. Dropping the records with null/missing values completely.
2. We are not dropping any records because we do not want to lose any data.
3. We don't want to pollute the data by introducing default values or skew it by duplicating the description columns because we'll be concatenating the short description and Description columns for each record when feeding them into NLP.
4. As a result, our NULL/Missing value treatment simply replaces NaN cells with empty strings.

4. MOJIBAKE: Mojibake is the jumbled text that results from decoding text with an unexpected character encoding. As a result, symbols are often replaced by wholly unrelated symbols, often from a different writing system.

TEXT PRE PROCESSING:

Text pre-processing is the conversion of human-readable text into a machine-readable format for further processing. We begin text normalisation after obtaining a text. Normalization of text entails:

- 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
- Text canonicalization

OBSERVATIONS and RESULT of TEXT PRE PROCESSING:

- Entire dataset is converted into lower case
- Users email addresses will add NO value to our analysis, despite the fact that user id is given in the caller column. So, all email addresses are removed from the dataset
- All numerals are removed because they were dominating the dataset if we were converting them into their word representation otherwise.
- All punctuation marks are removed which used to be a hindrance in lemmatization.
- All occurrences of more than one blank spaces, horizontal tab spaces, new line breaks etc. have been replaced with single blank space.

STEMMING and LEMETIZATION:

In the discipline of Natural Language Processing, stemming and lemmatization are Text Normalization (or often termed Word Normalization) procedures that are used to prepare text, words, and documents for further processing.

The change of a word to communicate multiple grammatical categories such as tense, case, voice, aspect, person, number, gender, and mood is known as inflection in grammar. With a prefix, suffix, or infix, or another internal modification such as a vowel change, an inflection communicates one or more grammatical categories.

The technique of reducing inflection in words to their root forms, such as mapping a set of words to the same stem, even if the stem is not a valid word in the Language, is known as stemming.

Lemmatization properly lowers inflected words, guaranteeing that the root word is a part of the language.

Along with NLTK, which contains only one, but the best method for solving every Natural Language problem, spaCy is one of the most popular NLP libraries. The language model, which

is used to execute a range of NLP tasks, can be downloaded after it has been downloaded and installed.

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is data analysis approach/philosophy that uses a variety of techniques (mostly graphical) to:

- maximize insight into a data set
- uncover underlying structure
- extract important variables
- detect outliers and anomalies
- test underlying assumptions
- develop parsimonious models
- determine optimal factor settings

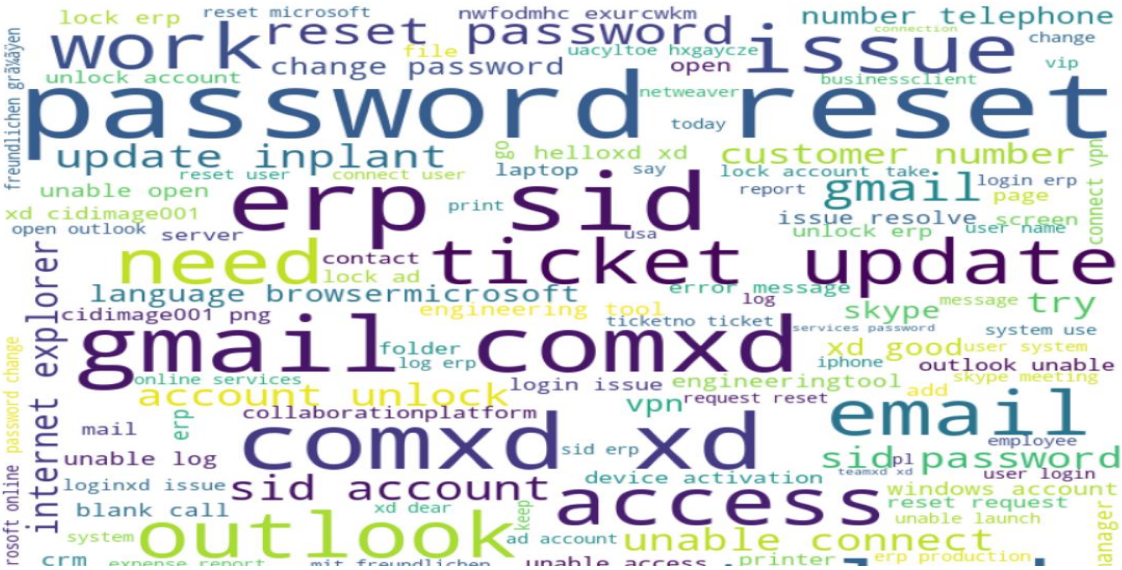
One of the most significant responsibilities in text mining is visually portraying the content of a text document. It assists in not just exploring the content of documents from various perspectives and at various levels of detail, but also in summarising a single document, displaying words and subjects, detecting events, and creating storylines.

The graphs and visualisations will be created with the plotly library. To connect plotly to a pandas dataframe and add the iplot method, we'll require cufflinks.

WORD CLOUD

A word cloud is a grouping of words that are shown in various sizes. The larger and bolder the term, the more frequently it appears in a document and the more essential it is.

Word clouds, also known as text clouds, are useful for extracting the most important bits of textual material. They may also be used to compare and contrast two separate pieces of text to uncover language similarities.



PREDICTION MODELS:

- Multinomial Naive Bayes
- K Nearest neighbour
- Support Vector Machine
- Decision Tree
- Random Forest
- Deep Neural Network
- Convolutional Neural Network
- Recurrent Neural Network
- Recurrent Convolutional Neural Network
- RNN with LSTM

Observations:

All of the following Statistical Machine learning algorithms we tried, are heavily overfitted. The training accuracy for all of them is coming to be more than 90%

The test accuracies are as follows:

- Multinomial Naive Bayes- 53.24%
- K Nearest Neighbour- 62.88%
- Linear SVM- 68.35%
- SVC with RBF kernel- 62.29%
- Decision Tree- 57.41%
- Random Forest- 62.71%

We'll be fine tuning the models and reduce the overfitting in next iteration.