

Automatic Ticket Assignment

Capstone Project Report

AIML

# Index

[Index 2](#_Toc96278349)

[Team Details 4](#_Toc96278350)

[Summary of the problem statement, Data and findings 5](#_Toc96278351)

[Understanding the Business 5](#_Toc96278352)

[Risks Involved 5](#_Toc96278353)

[Background and Objective 5](#_Toc96278354)

[Data & Findings 6](#_Toc96278355)

[Data Fields 6](#_Toc96278356)

[Sample data 6](#_Toc96278357)

[Observations 7](#_Toc96278358)

[Summary of the approach to EDA and Pre-Processing 8](#_Toc96278359)

[Data Pre-Processing and Cleaning 8](#_Toc96278360)

[Data cleaning 8](#_Toc96278361)

[Lemmatization & Stop words removal 8](#_Toc96278362)

[Data Visualization 9](#_Toc96278363)

[High Imbalance 9](#_Toc96278364)

[Assignment Group Distribution 9](#_Toc96278365)

[Word Length Distribution 10](#_Toc96278366)

[Character Length Distribution 10](#_Toc96278367)

[Word Cloud After Data Pre-processing and Cleaning 12](#_Toc96278368)

[Charts 22](#_Toc96278369)

[Decide Model and Model building 26](#_Toc96278370)

[Word Embedding 26](#_Toc96278371)

[Text Augmentation 27](#_Toc96278372)

[Count Vectorization 27](#_Toc96278373)

[Tf idf Vectorization 27](#_Toc96278374)

[Bi-directional LSTM Model 27](#_Toc96278375)

[Machine Learning Models 28](#_Toc96278376)

[Deep Learning Models 31](#_Toc96278377)

[ML Model Performance 33](#_Toc96278378)

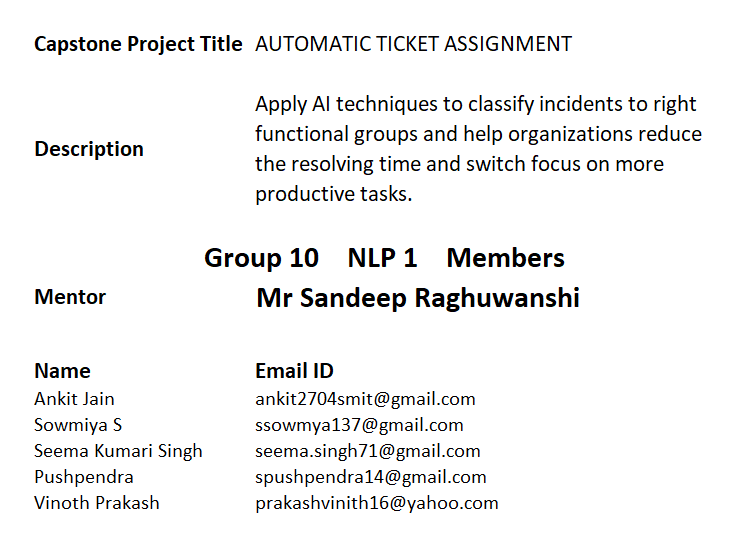
[DL Model Performance 33](#_Toc96278379)

[Comparison to benchmark 34](#_Toc96278380)

[Implications 34](#_Toc96278381)

[Limitations 34](#_Toc96278382)

[Closing Reflections 34](#_Toc96278383)

Team Details

# Summary of the problem statement, Data and findings

## Understanding the Business

IT leverages Incident Management process to ensure there is no disruption to business operations. Any unplanned disruption can cause interruption to business services. Incident management process helps in identification of issues or problems faced by users or operation teams. Manual assignment of incidents is time consuming and requires human intervention. There may be lag in incident resolution due to human errors or if incidents are not routed appropriately. On the other hand, manual assignment also increases the response and resolution times which result in user satisfaction deterioration / poor customer service.

## Risks Involved

The manual assignment of these incidents might have below disadvantages:

❖ More resource usage and expenses.   
❖ Human errors - Incidents get assigned to the wrong assignment groups   
❖ Delay in assigning the tickets   
❖ More resolution times   
❖ If a particular ticket takes more time in analysis, other productive tasks get affected for the Service Desk

## Background and Objective

To apply techniques and learnings to make ticket assignment more cost-effective, less resolution time so that service desk team can focus on other productive tasks. We are able to see that the current system is capable of assigning 70+% of the tickets correctly. Our target is to automatically classify tickets and directing them to appropriate groups at the earliest, helps in improving the throughput in the ticketing pipeline of an organization.

Data & Findings  
  
Data format CSV

Total Records 8500

## Data Fields

|  |  |
| --- | --- |
| Short description | A brief overview of the issue faced by the user |
| Description | Detailed description of the issue |
| Assignment group | GRP\_0 ~ GRP\_73 (total 74 classes of Assignment group) |

## Sample data

| **Short description** | **Description** | **Assignment group** |
| --- | --- | --- |
| login issue | -verified user details.(employee# & manager name)  -checked the user name in ad and reset the password.  -advised the user to login and check.  -caller confirmed that he was able to login.  -issue resolved. | GRP\_0 |
| Outlook | received from: hmjdrvpb.komuaywn@gmail.com  hello team,  my meetings/skype meetings etc are not appearing in my outlook calendar, can somebody please advise how to correct this?  kind | GRP\_0 |
| cant log in to vpn | received from: eylqgodm.ybqkwiam@gmail.com  hi  i cannot log on to vpn  best | GRP\_0 |

## Observations

1. High imbalance seen in data with GRP\_0 having 40%+ percent of representation
2. Many groups/classes are with very little representation.
3. Null values:

Short description 8

Description 1

Assignment group 0

1. Very few tickets have non-English descriptions
2. Four columns – Short Description, Description, Caller and Assignment group
3. 74 Assignment groups found - Target classes
4. Caller names in a random fashion (may not be useful for training data)
5. European non-English language also found in the data
6. Email/chat format in description
7. Symbols & other characters in the description
8. Hyperlinks, URLS & few image data found in the description
9. Blanks found either in the short description or description field
10. Few descriptions same as the short description
11. Few words were combined together
12. Spelling mistakes and typo errors are found

# Summary of the approach to EDA and Pre-Processing

## Data Pre-Processing and Cleaning

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

1. Dropped the caller field as the data was not found to be useful for analysis
2. Replaced Null values in short description & description with space.
3. Merged Short Description & Description fields for analysis
4. Contraction words found in the merged Description are removed for ease of word modelling
5. Changed the case sensitivity of words to lower case
6. Removed Hashtags, Hyperlinks, URLs, HTML tags, Emoji and non-ASCII symbols from merged fields.
7. Translating all languages (German) to English
8. Tokenization of merged data
9. Removal of Stop words and Meaningless words
10. Lemmatization
11. WordCloud created
12. Attempted to do spell check
13. Created Plot to understand the distribution of words
14. Removal of line breaks and tabs (\r\n\t)
15. Removal of special characters
16. Removal of punctuation
17. Removal of extra spaces
18. Missing value imputation

## Data cleaning

Data cleaning has been taken care by splitting the various functions which would remove and clean any unwanted or misleading information.

All words have been converted to lower case.   
Header and sender information from the emails have been removed.

All the numbers, non-dictionary characters, newline characters, hashtag, HTML entities, emojis, hyperlinks, extra spaces and unreadable characters have been removed.  
We have ensured to remove any caller names included in the description column.

## Lemmatization & Stop words removal

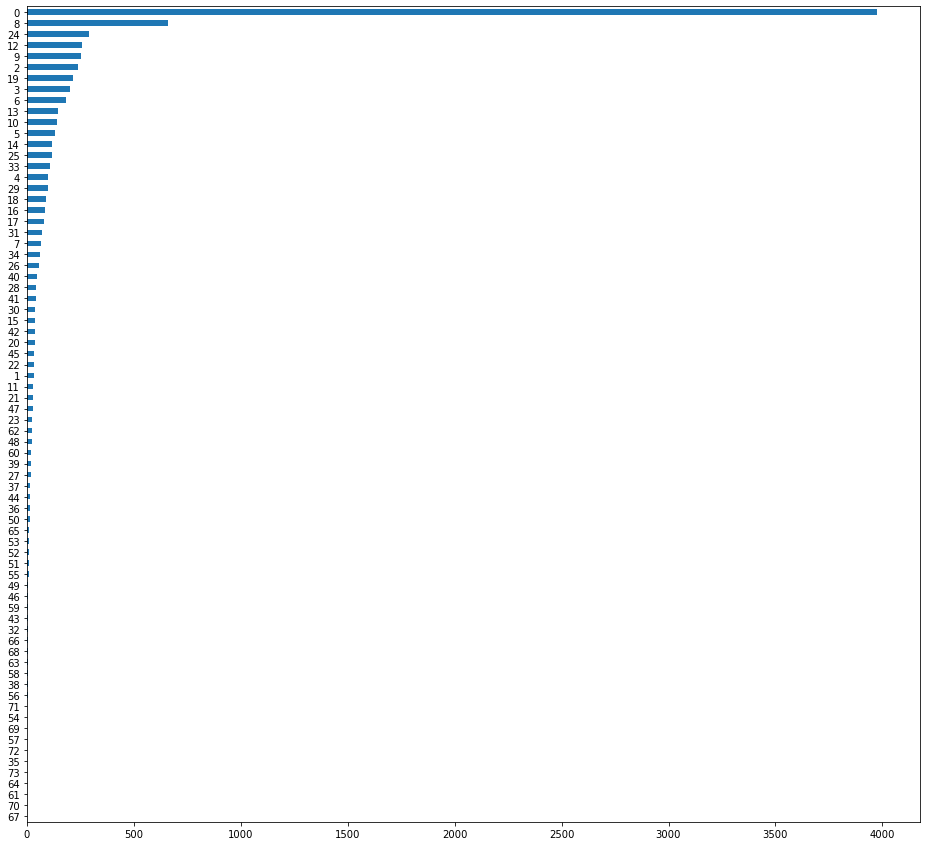
Stop words have been removed using nltk corpus modules.

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed 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.

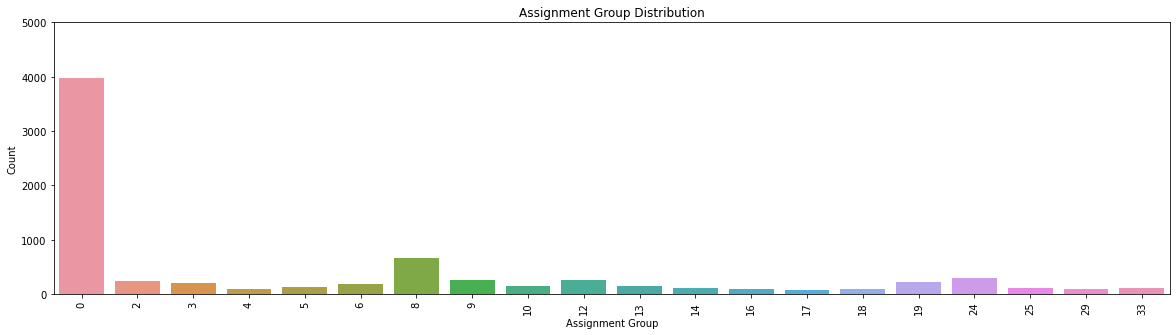
We have preferred Lemmatization over Stemming because lemmatization does morphological analysis of the words

# Data Visualization

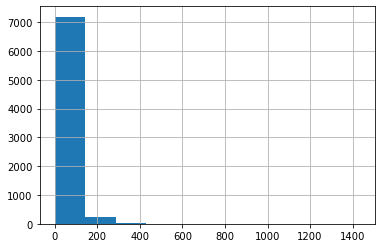
## High Imbalance



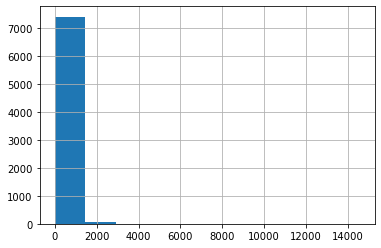
## Assignment Group Distribution



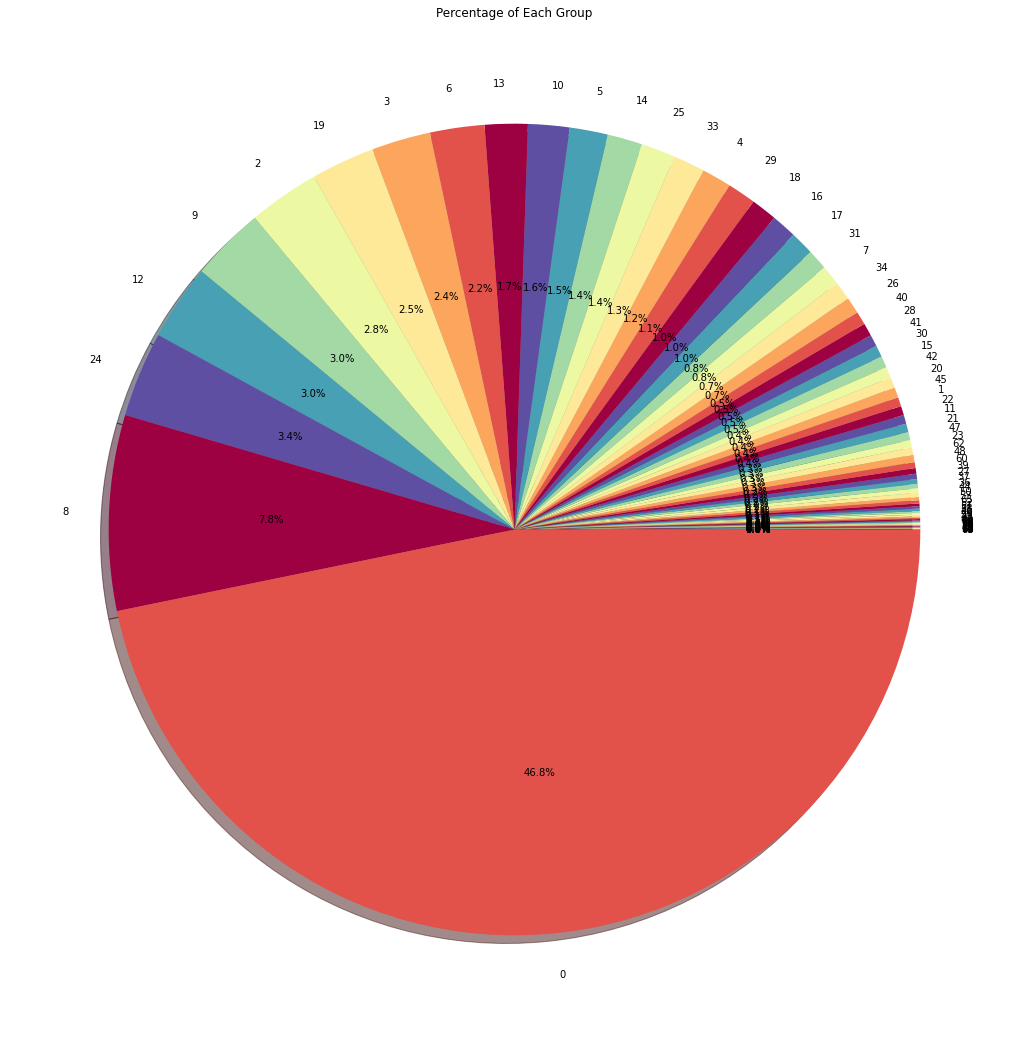
## Word Length Distribution



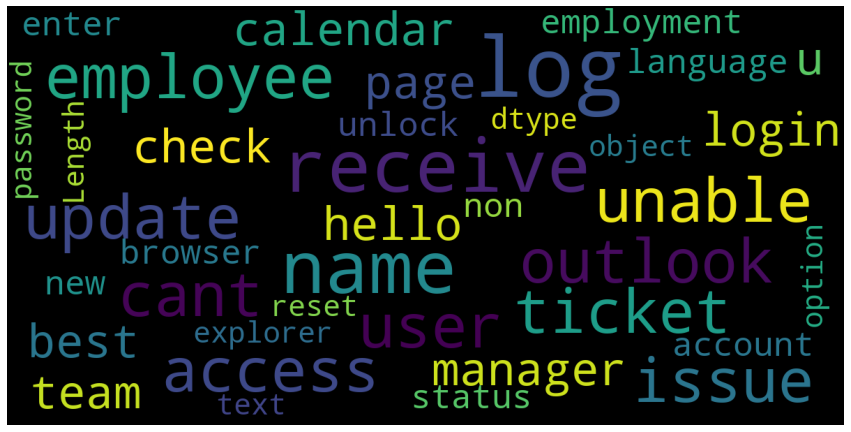
## Character Length Distribution

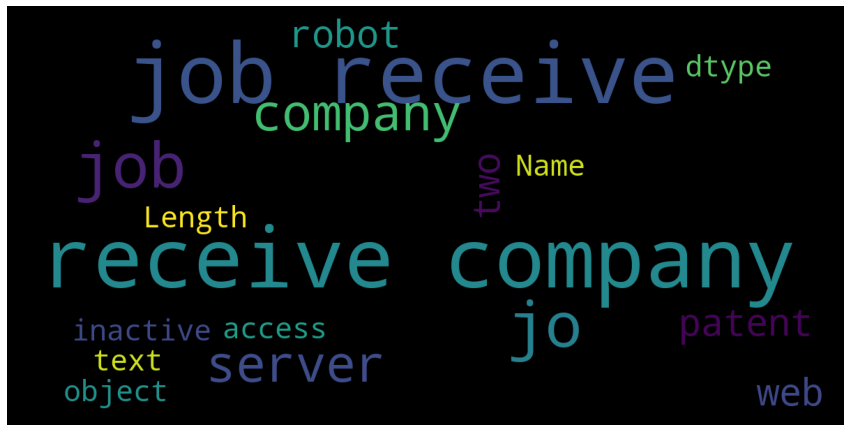


Percentage Split of Top 20 Groups Having 80+% of records



## Word Cloud After Data Pre-processing and Cleaning

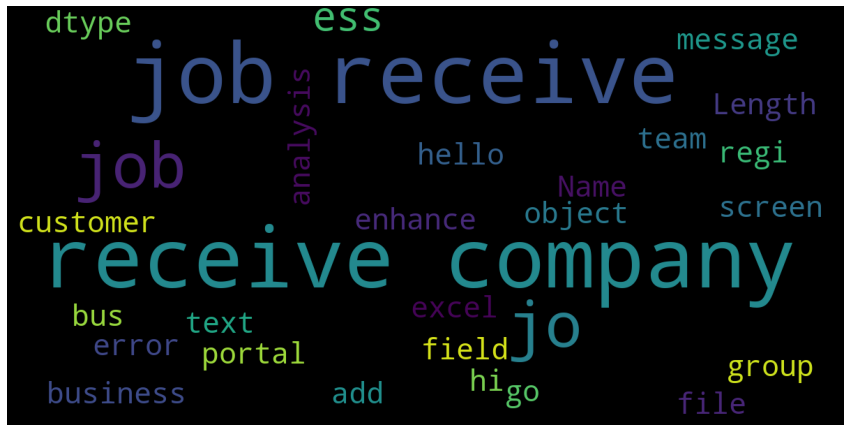
Grp\_0

Grp\_8

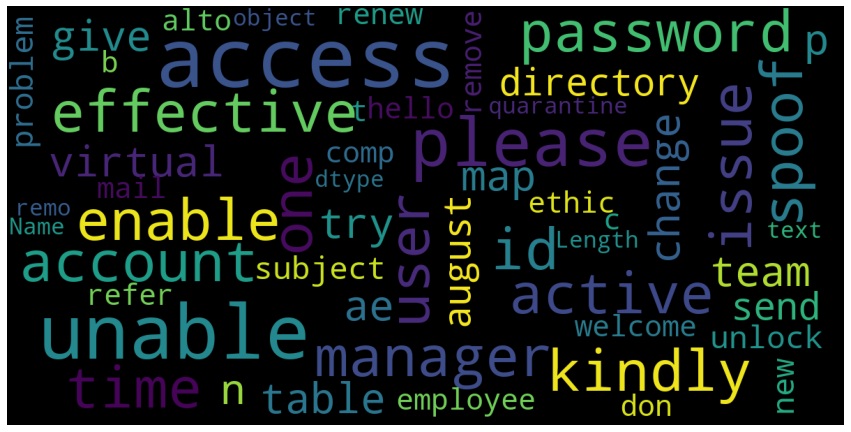
Grp\_12

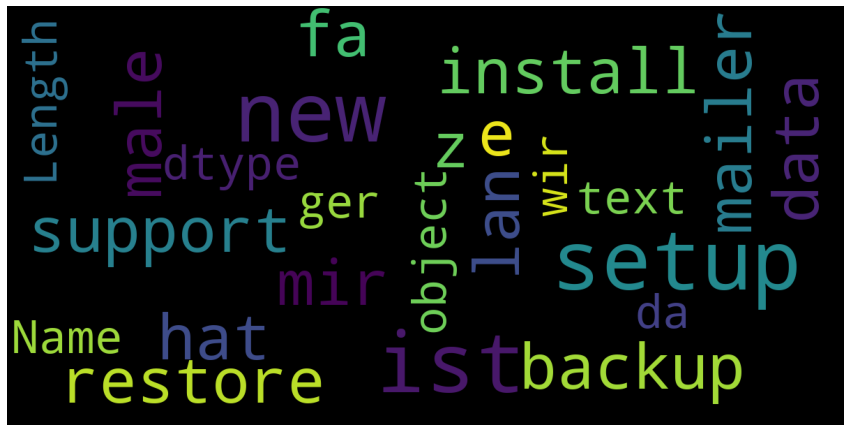


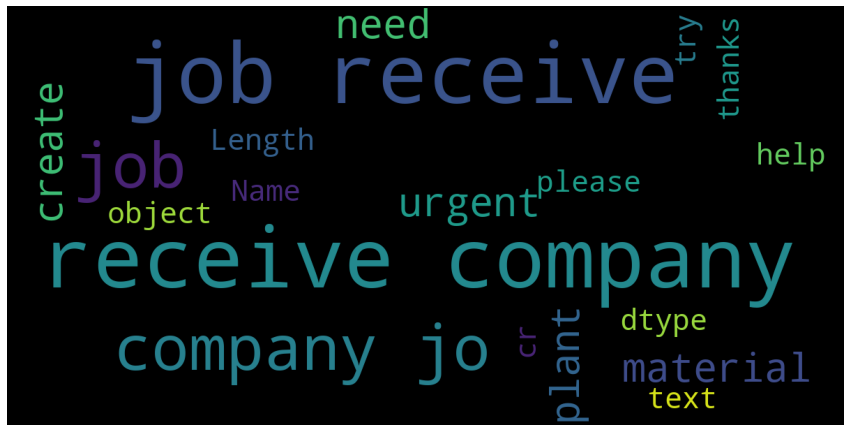
Grp\_09

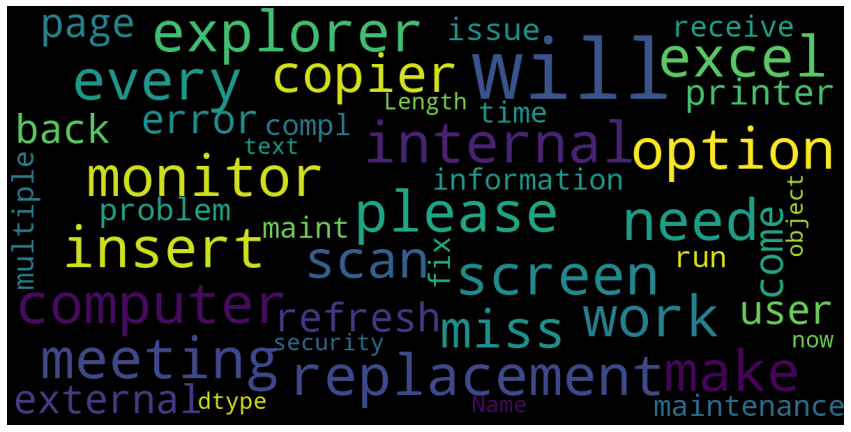


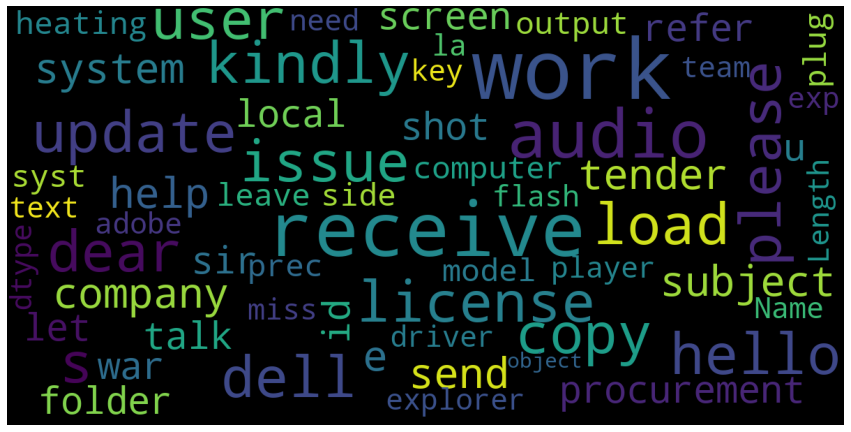
Grp\_02

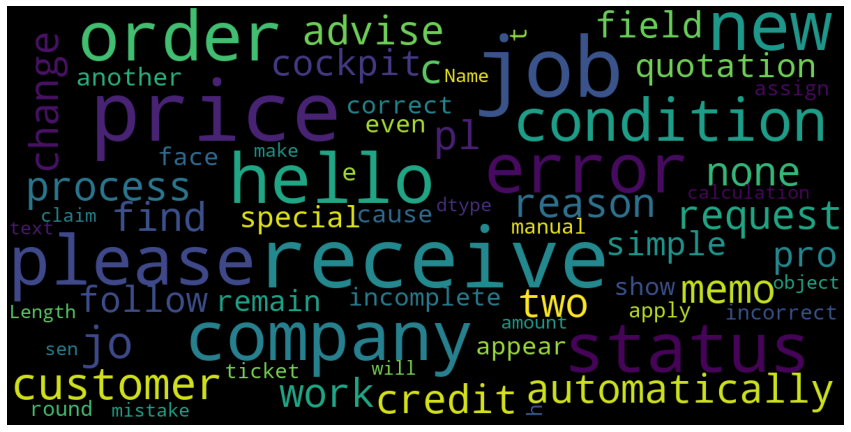


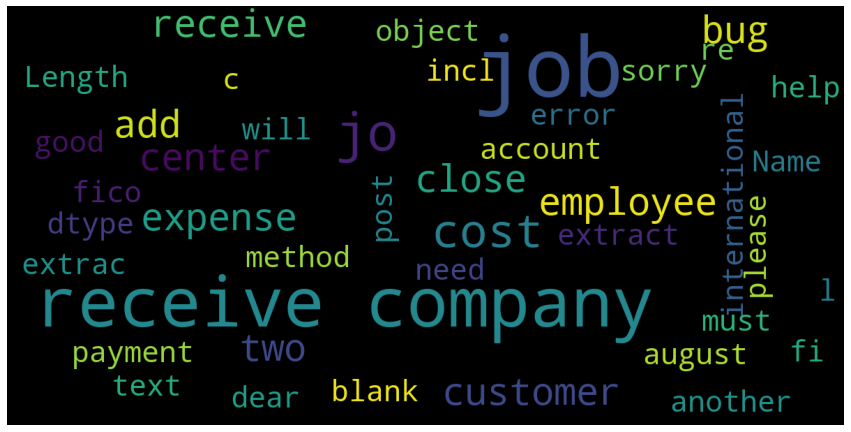
Grp\_24

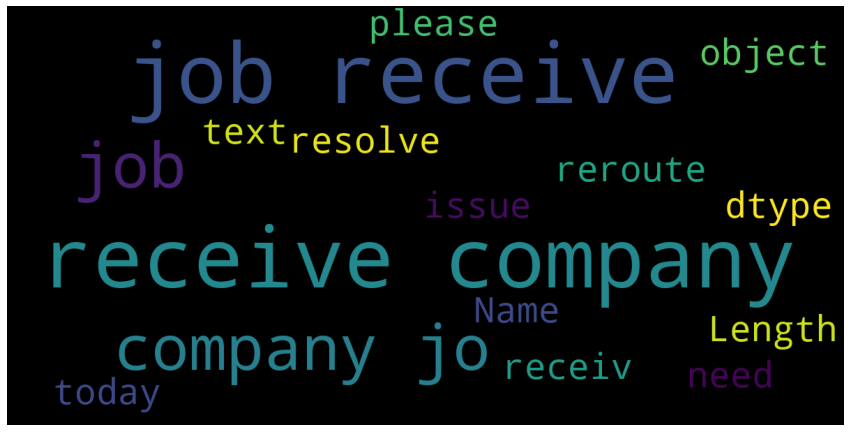
Grp\_06

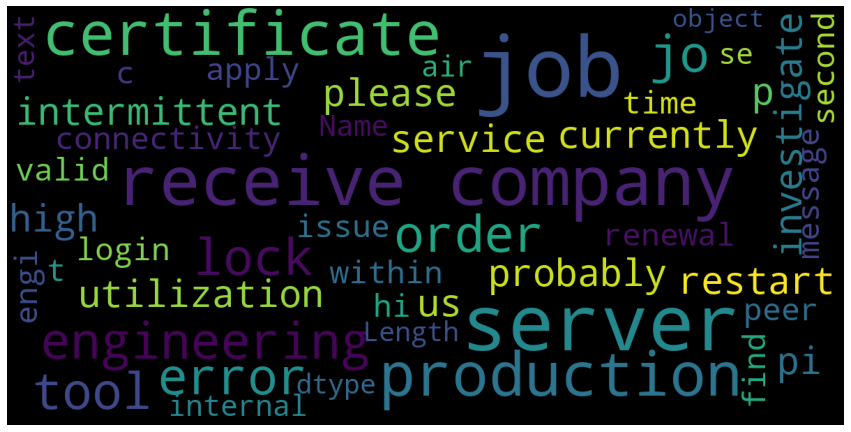
Grp\_03

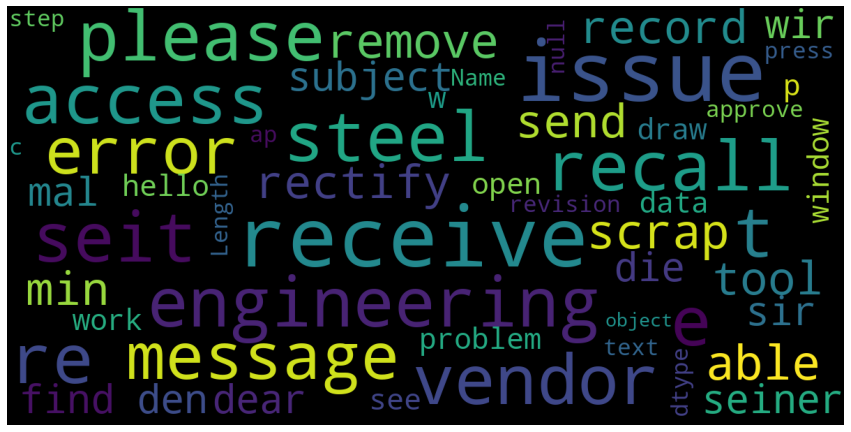
Grp\_19

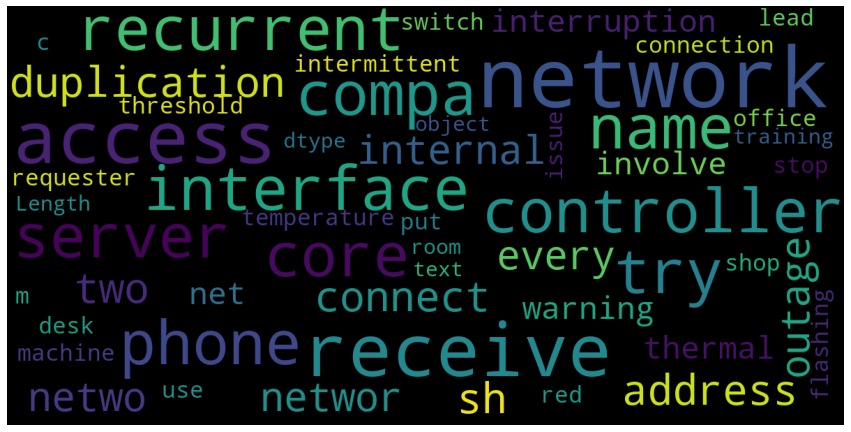
Grp\_13

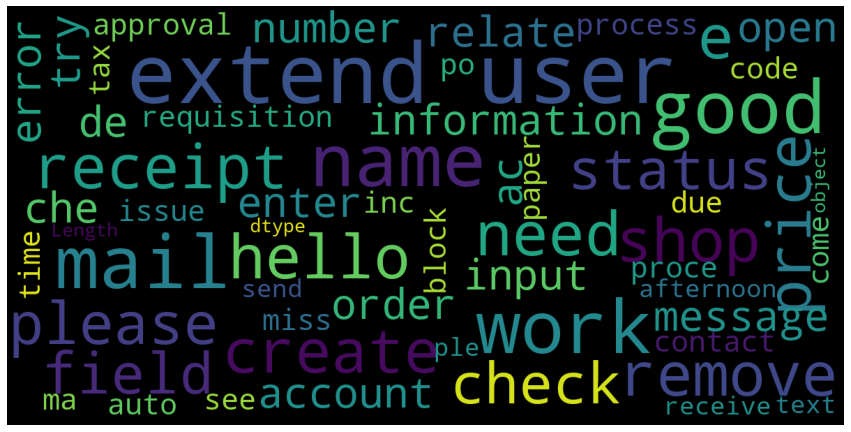
Grp\_10

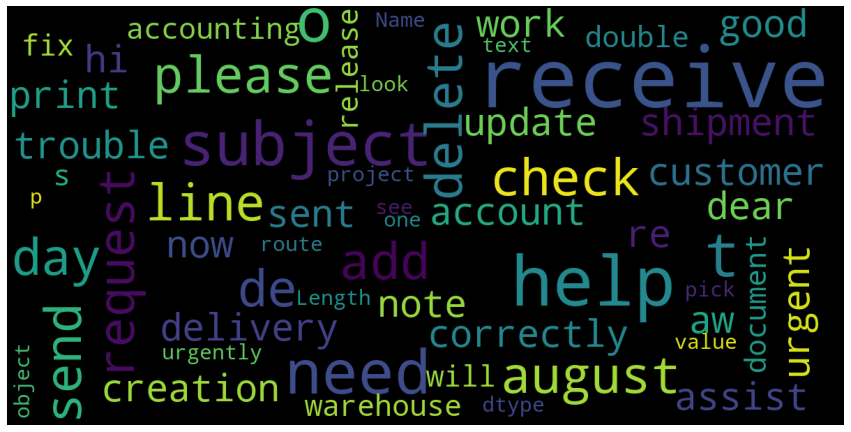
Grp\_05

Grp\_14

Grp\_25

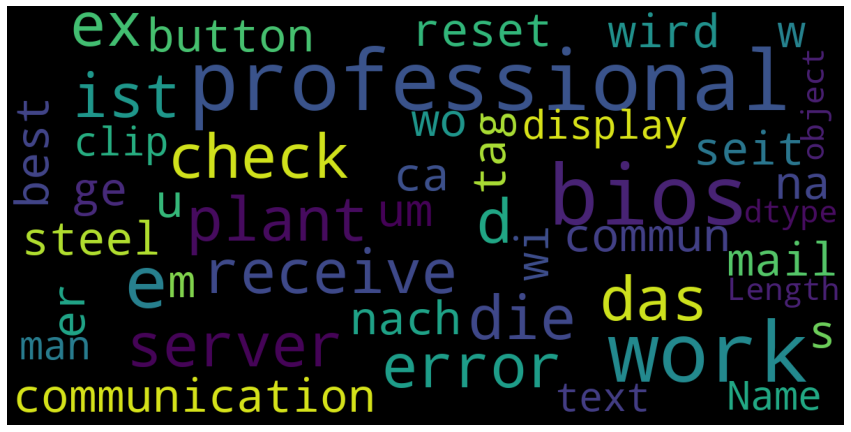
Grp\_04

Grp\_29

Grp\_18

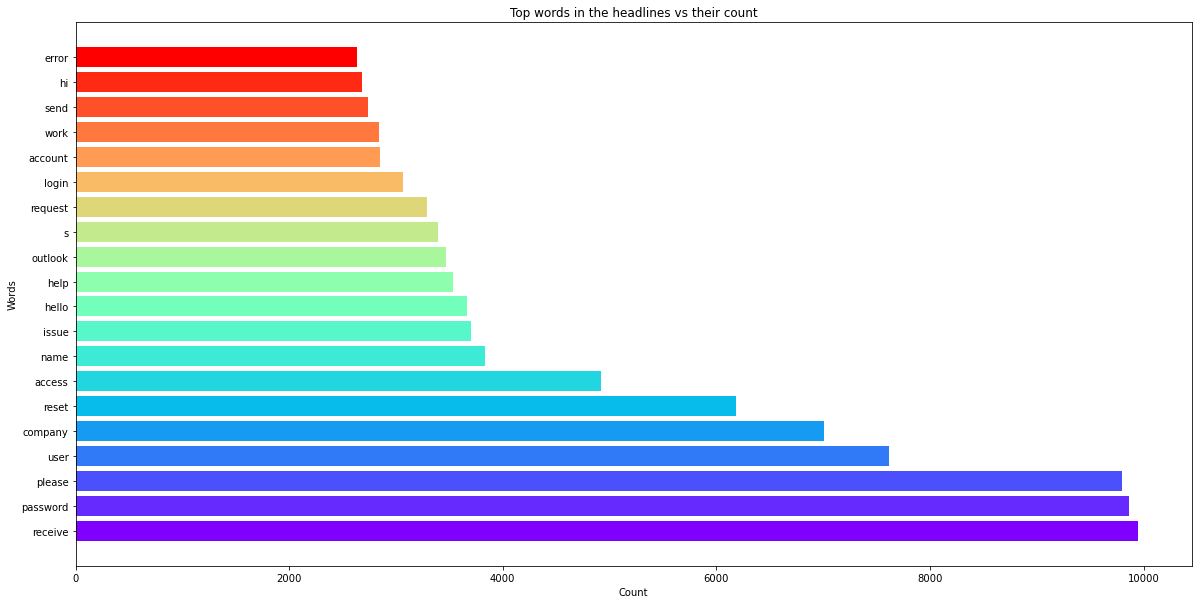
Grp\_17

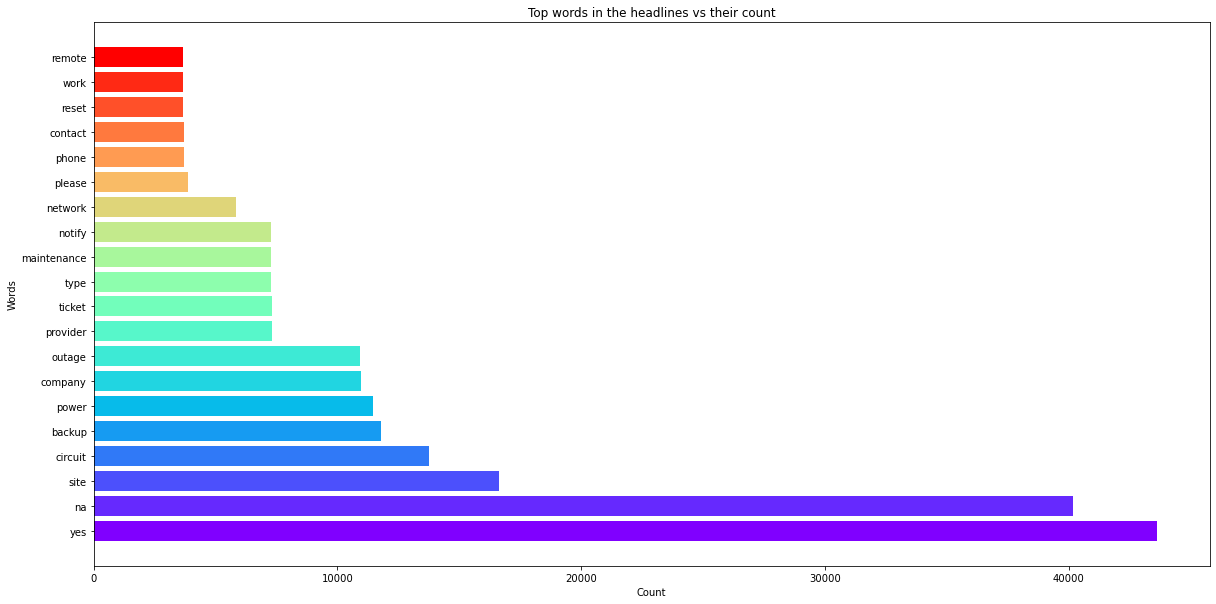
Grp\_16

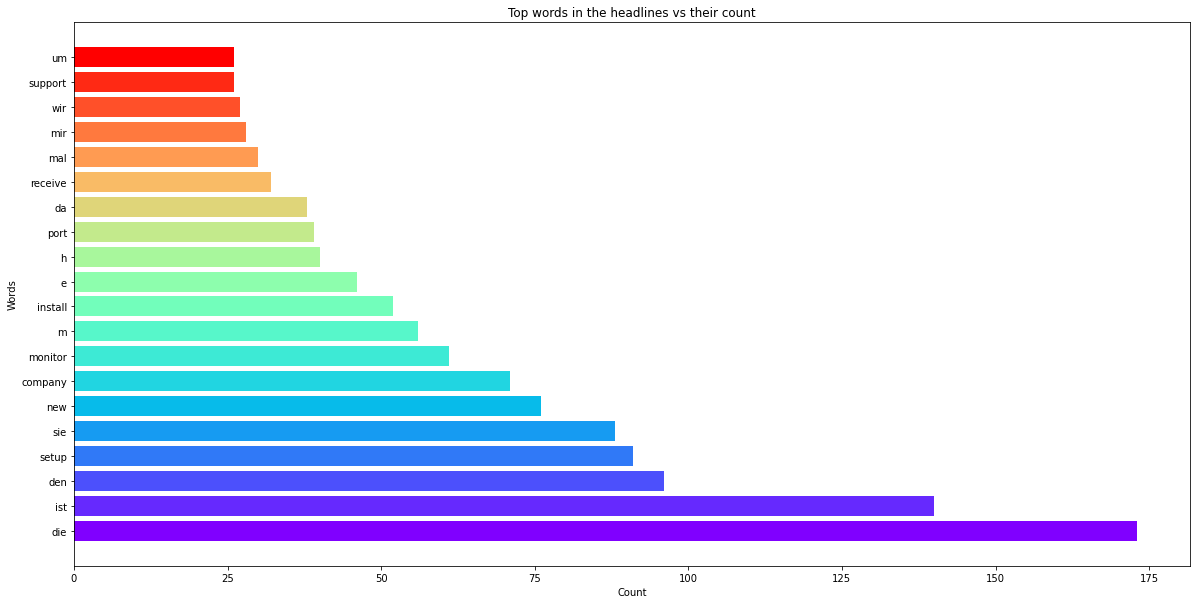
Grp\_33 

## Charts

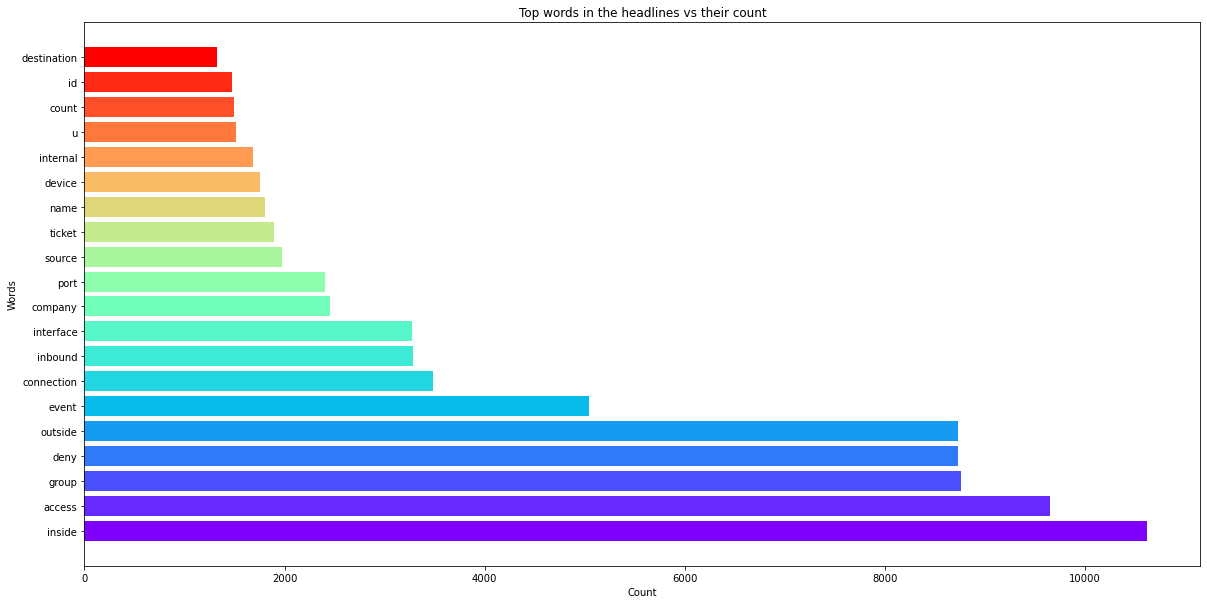
Frequent words used in Top 20 groups [Sharing charts for a few of them]

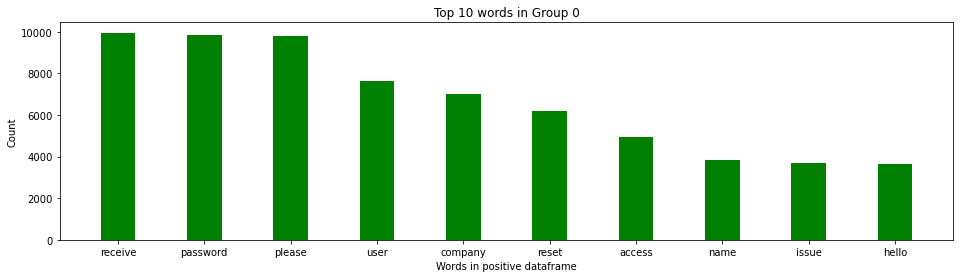
Grp\_0 

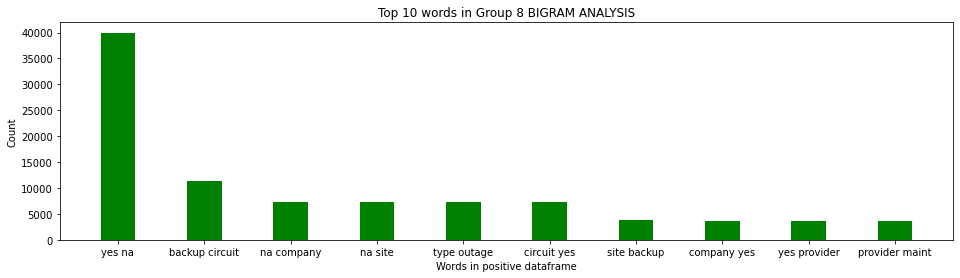
Grp\_8 

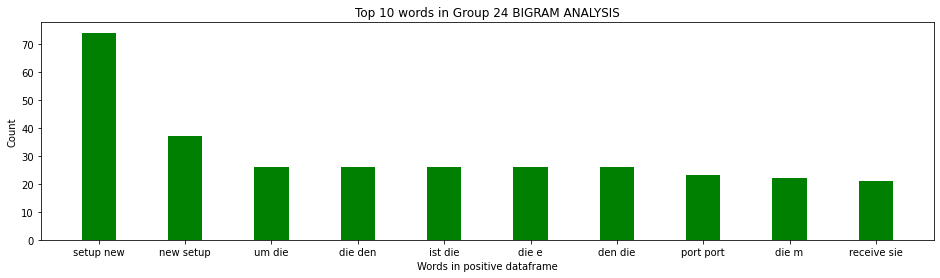
Grp\_24   


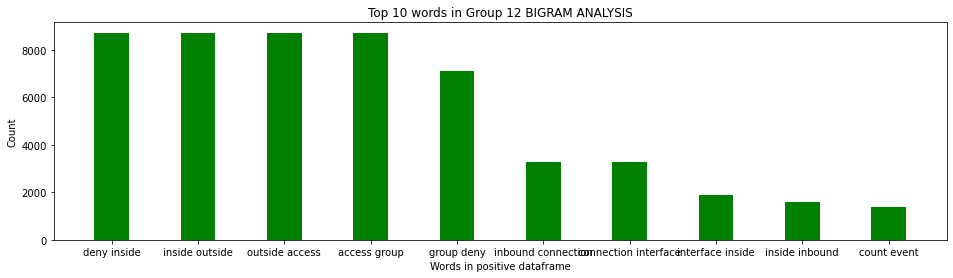
Grp\_12

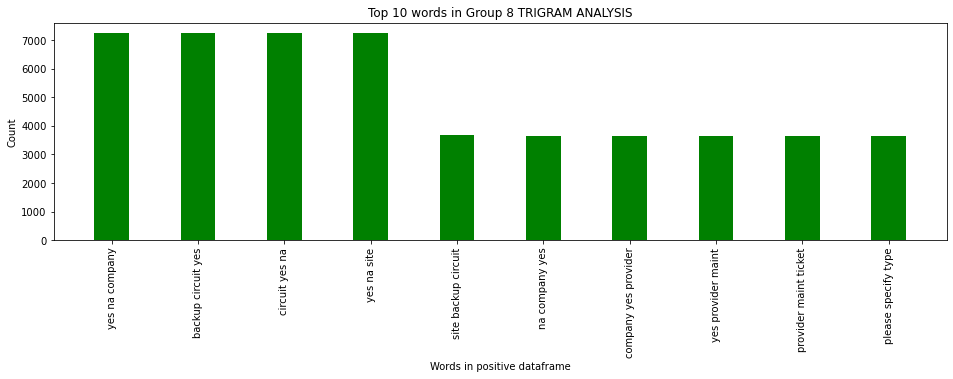


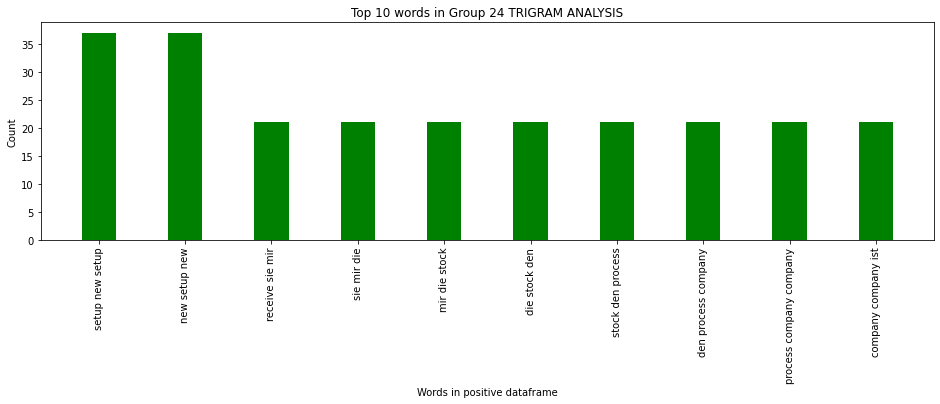
Unigram

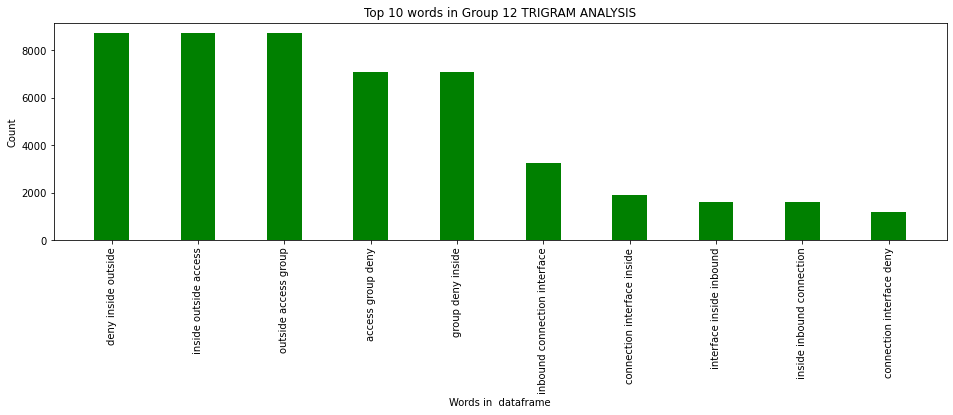
Bigram





Trigram

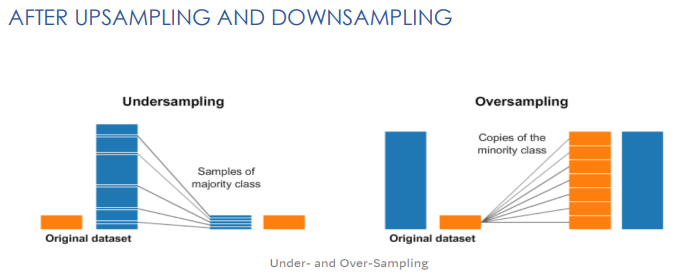




## Decide Model and Model building

As the target class is completely skewed, various models have been tried with the sampled datasets to compare each performance.

1. Unigram, Bi-gram and Tri-gram models
2. Glove Embedding
3. Count Vectorization
4. TF - IDF Vectorization
5. Text Augmentation
6. Naive Bayes
7. Logistic Regression
8. XGBoost
9. GRU model
10. RNN model
11. Random Forest
12. Bidirectional LSTM



Word Embedding   
As all our Machine Learning and Deep learning algorithms are incapable of processing strings or plain text in their raw form, word embedding has been used to convert the texts into numbers. There may be different numerical representations of the same text. It tries to map a word using a dictionary to a vector.

Tokenizing Text -> Representing each word by a number Mapping of original word to number is preserved in word index property of tokenizer Tokenized applies basic processing like changing it to lower case, explicitly setting that as False Lets keep all news to 300, add padding to news with less than 300 words and truncating long ones.

We have experimented with Glove embedding in our models.

**GloVe (Global Vectors) Embedding:**   
GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Text Augmentation

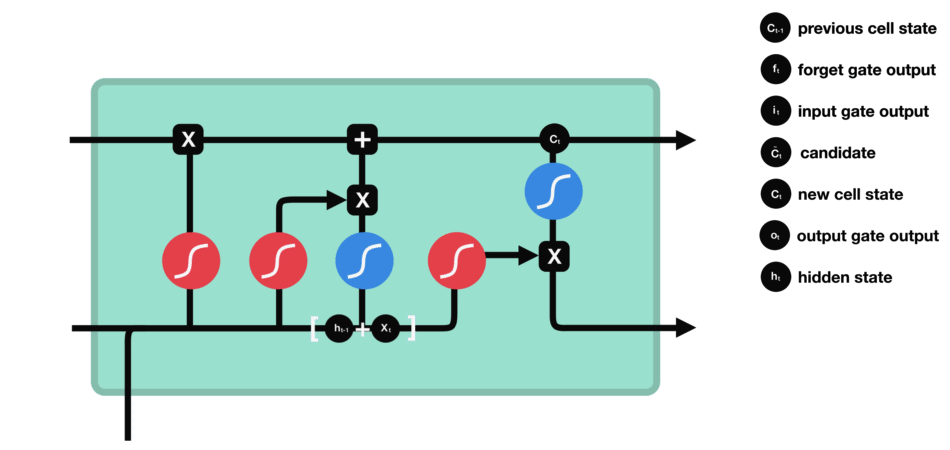
Data augmentation techniques are used to generate additional, synthetic data using the data you have. We used Synonym Replacement technique to up sample text Randomly choose n words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random.

Count Vectorization   
Count Vectorization involves counting the number of occurrences each word appears in a document (i.e., distinct text such as an article, book, even a paragraph!). Python’s Sci-kit learn library has a tool called Count Vectorizer to accomplish this.

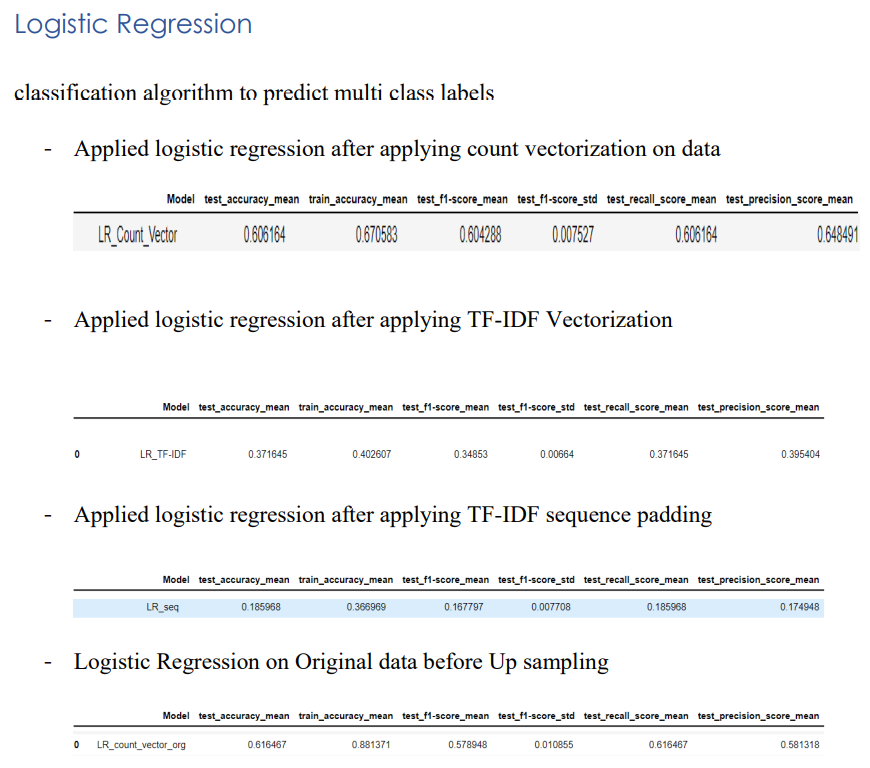
Tf idf Vectorization

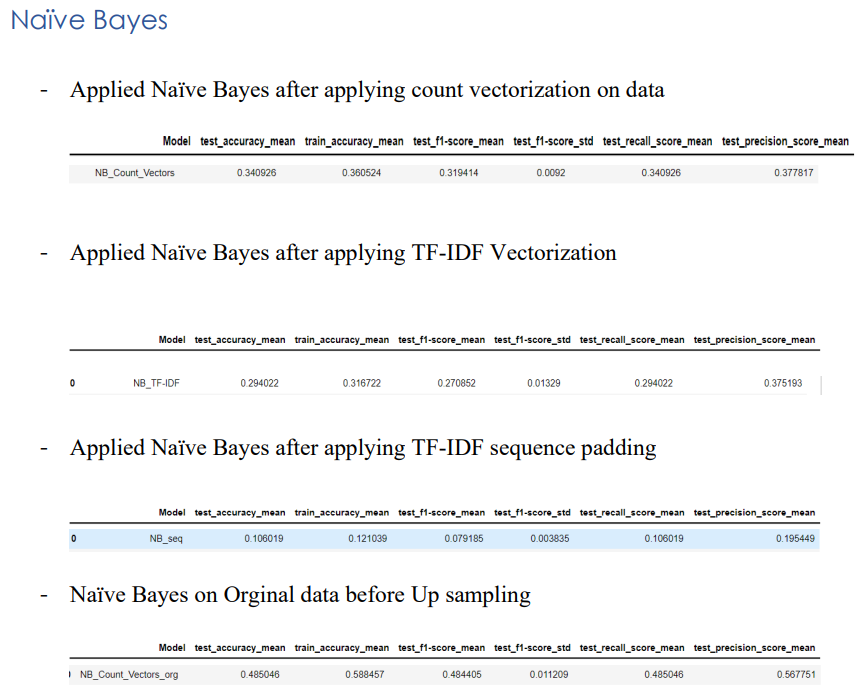
Tf idf Vectorizer - Transforms text to feature vectors that can be used as input to estimator.vocabulary\_Is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index

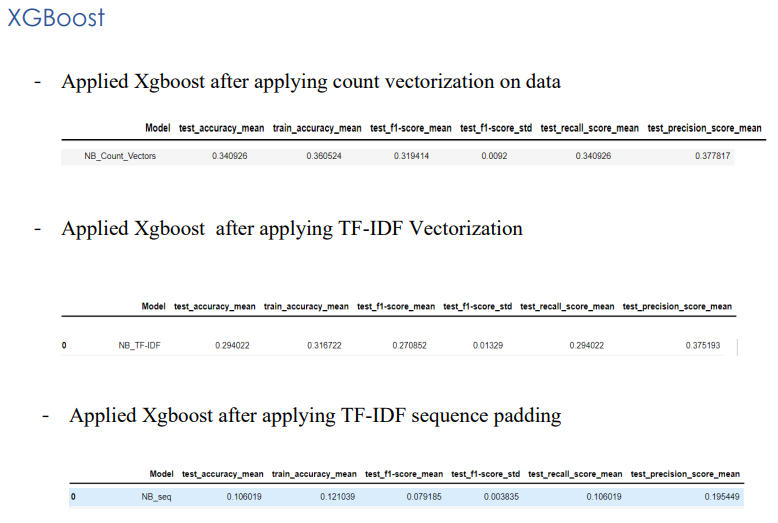
Bi-directional LSTM Model   
Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on classification problems. In problems where all timesteps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem.



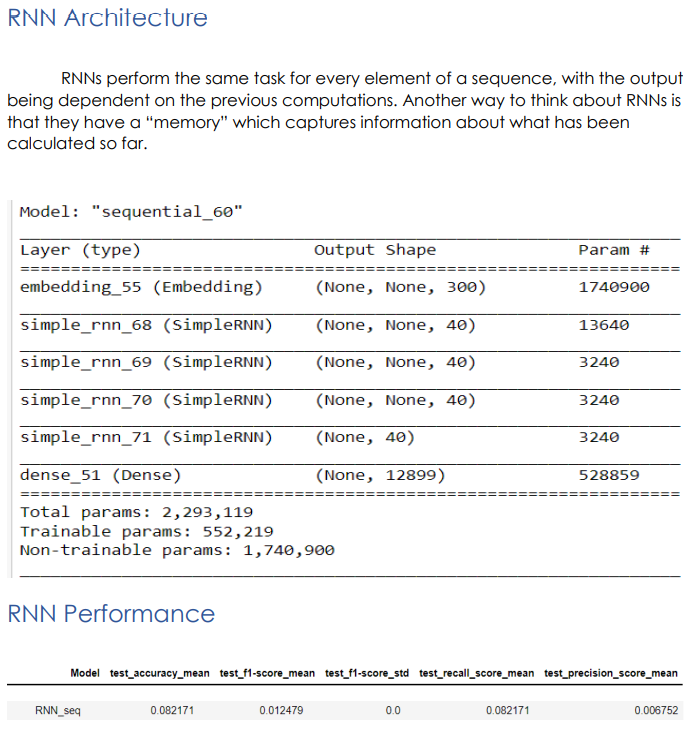
# Machine Learning Models

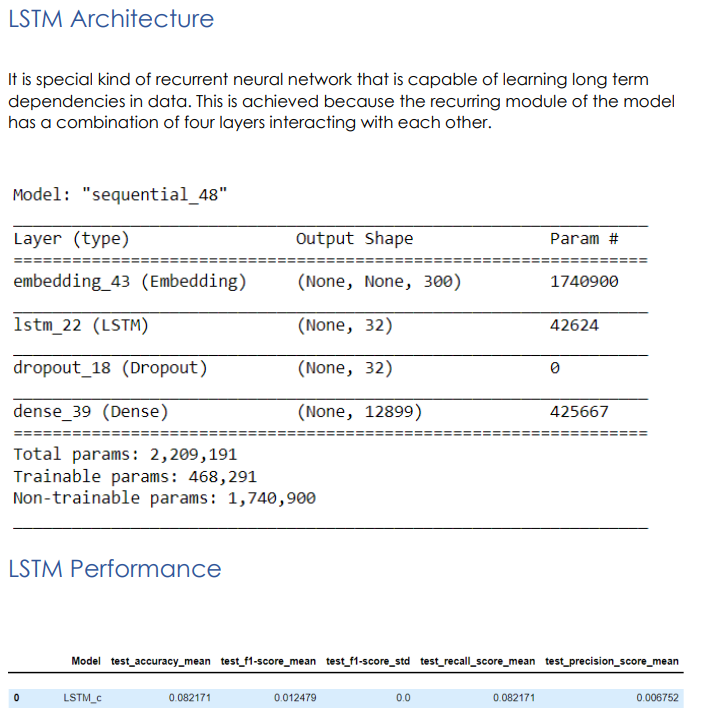


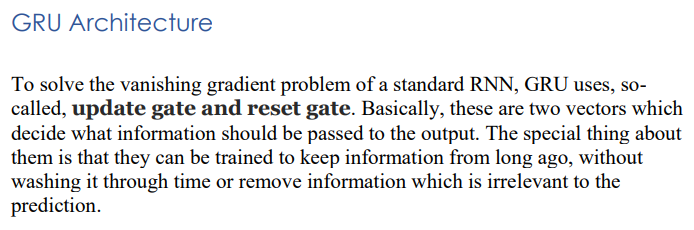


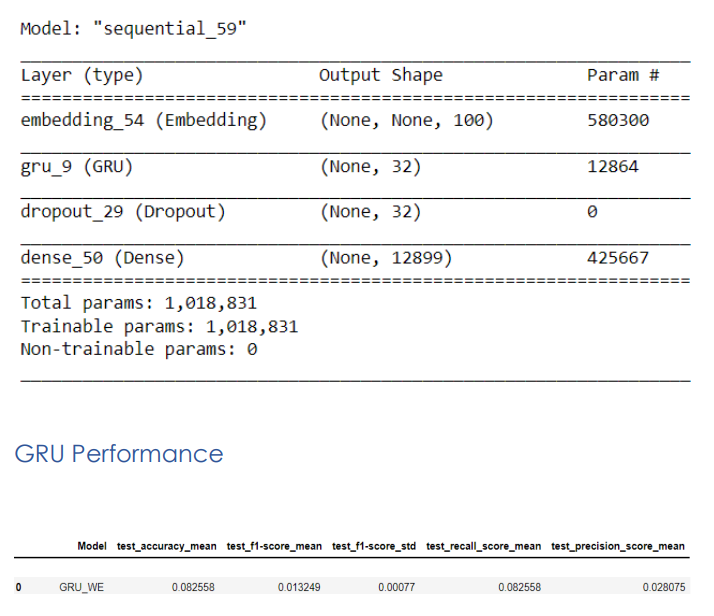


# Deep Learning Models

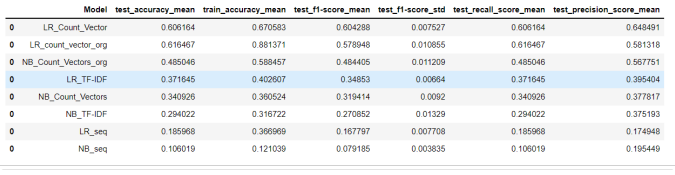




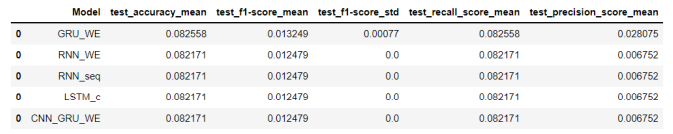




# ML Model Performance



# DL Model Performance



# Comparison to benchmark

From the given problem description, we could see that the current system is able to assign 75% of the tickets correctly. So, our objective here is to build an AI-based classifier model to assign the tickets to right functional groups by analysing the given description and achieve a higher degree of accuracy. From the prediction results we see that the GRU model based on is able to achieve an accuracy of 82.25% which is above than the benchmark.

# Implications

Although the GRU model can classify the IT tickets with 82.25% accuracy, to achieve better accuracy in the real world it would be good if additional data is collected for each group which may help reduce the data imbalance and skewness of the data.

# Limitations

As part of Data pre-processing, we have considered only the top 20 groups based on the count frequency. However, when applying this model in real world scenario, there could be quite a deviation in the model performance if the data is balanced across all assignment groups.

# Closing Reflections

We found the data was present in multiple languages and in various formats such as emails, chat, etc bringing in a lot of variability in the data to be analyzed.   
The Business can improve the process of raising tickets via a common unified IT Ticket Service Portal which reduces the above-mentioned variability.   
By doing this, the model can perform better which can help businesses to identify the problem area for relevant clusters of topics.