

Automatic Ticket Assignment

Capstone Project Report

AIML

# Index

[Index 2](#_Toc96268090)

[Team Details 4](#_Toc96268091)

[Summary of the problem statement, Data and findings 5](#_Toc96268092)

[Understanding the Business 5](#_Toc96268093)

[Risks Involved 5](#_Toc96268094)

[Background and Objective 6](#_Toc96268095)

[Data & Findings Data format 6](#_Toc96268096)

[Total Records 8500 6](#_Toc96268097)

[Data Fields 6](#_Toc96268098)

[Sample data 6](#_Toc96268099)

[Observations 7](#_Toc96268100)

[Summary of the approach to EDA and Pre-Processing 8](#_Toc96268101)

[Data Pre-Processing and Cleaning 8](#_Toc96268102)

[Data Visualization 8](#_Toc96268103)

[Assignment Group Distribution 8](#_Toc96268104)

[Percentage Split of Top 20 Groups Having 80+% of records 9](#_Toc96268105)

[Word Cloud After Data Pre-processing and Cleaning 10](#_Toc96268106)

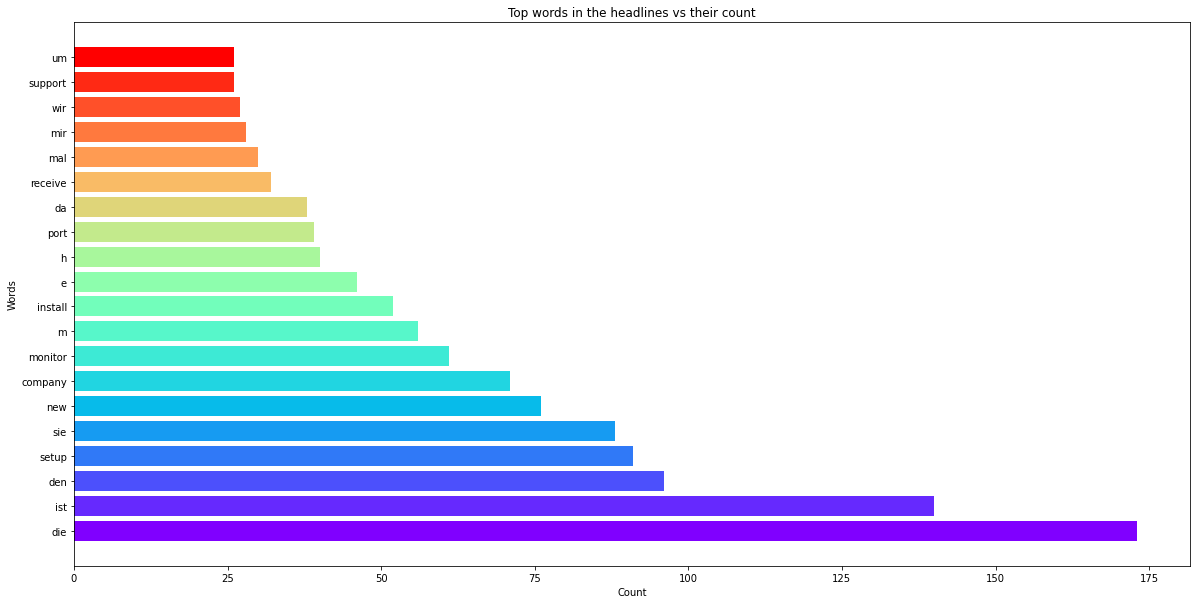
[Grp\_12 11](#_Toc96268107)

[Grp\_09 11](#_Toc96268108)

[Grp\_02 12](#_Toc96268109)

[Charts 20](#_Toc96268110)

[Frequent words used in Top 20 groups [Sharing charts for a few of them] 20](#_Toc96268111)

[Grp\_24  21](#_Toc96268112)

[Grp\_12 21](#_Toc96268113)

[Decide Model and Model building 24](#_Toc96268114)

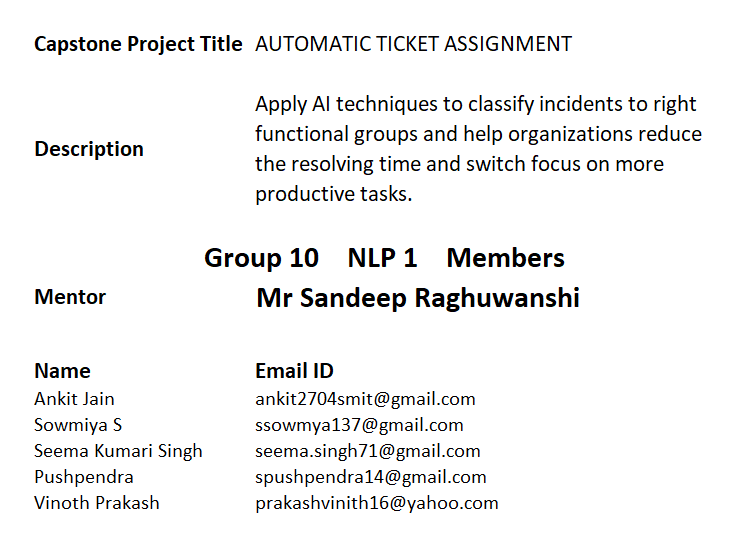
[Train and Test Accuracy for each model 26](#_Toc96268115)

[Model performance - Approaches to improve model 27](#_Toc96268116)

[Limitations 27](#_Toc96268117)

[Link to code and references 28](#_Toc96268118)

# 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:

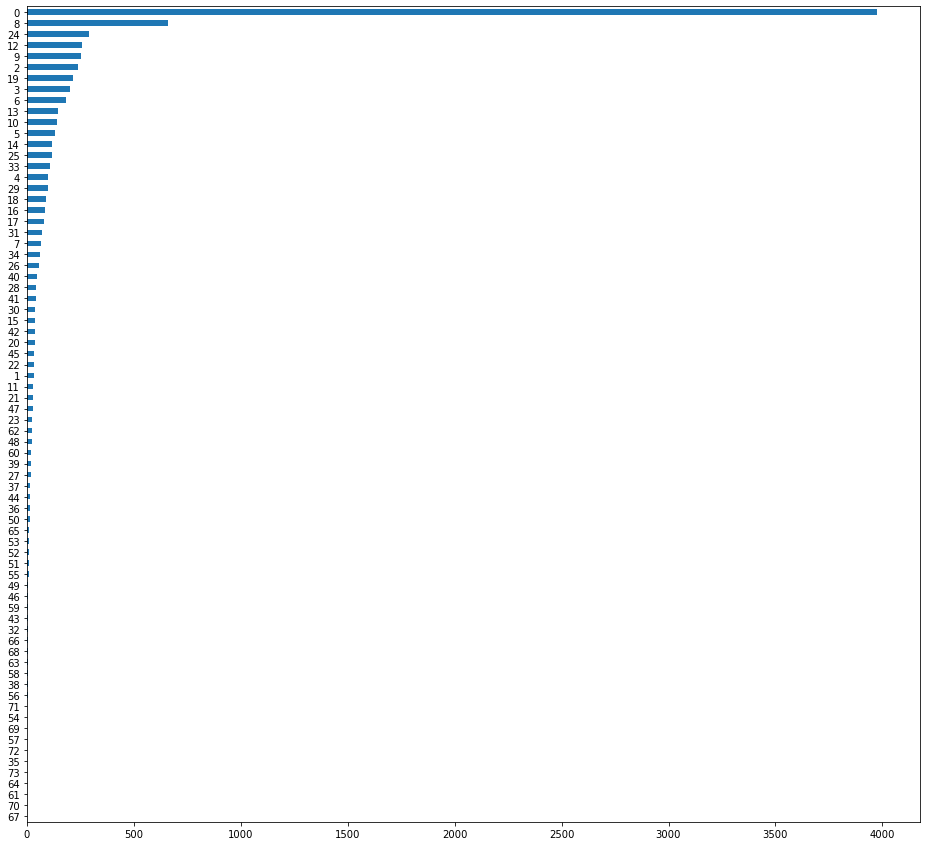
1. Dropped the caller field as the data was not found to be useful for analysis
2. Replaced Null values in short description & description with space.
3. Merged Short Description & Description fields for analysis
4. Contraction words found in the merged Description are removed for ease of word modelling
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

## Algorithms/Models Used

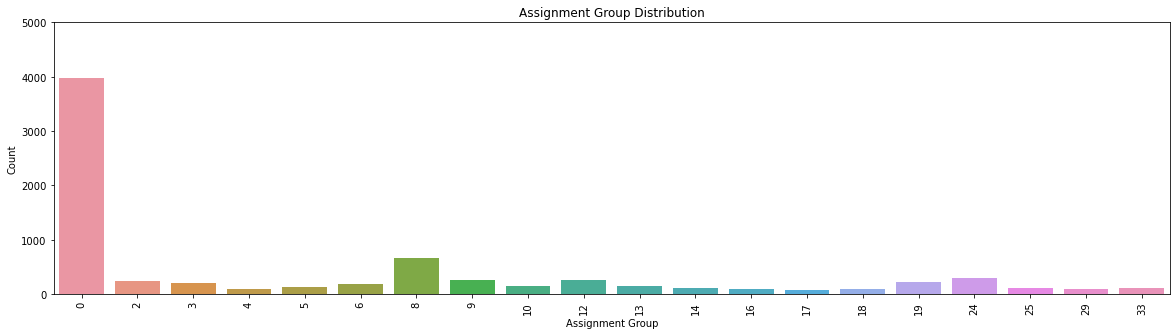
1. Built Bi-gram and Tri-gram models from the bag of words
2. Glove Embedding
3. Count Vectorization
4. TF - IDF Vectorization
5. Text Augmentation
6. Naive Bayes
7. Logistic Regression
8. XGBoost
9. GRU model
10. RNN model
11. Random Forest
12. Bidirectional LSTM

# Data Visualization

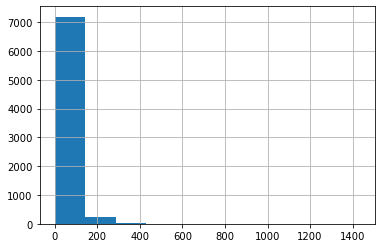
## High Imbalance



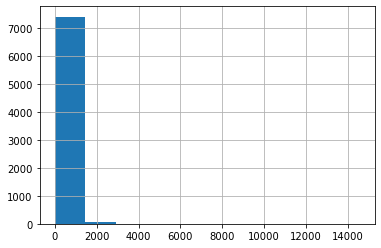
## Assignment Group Distribution



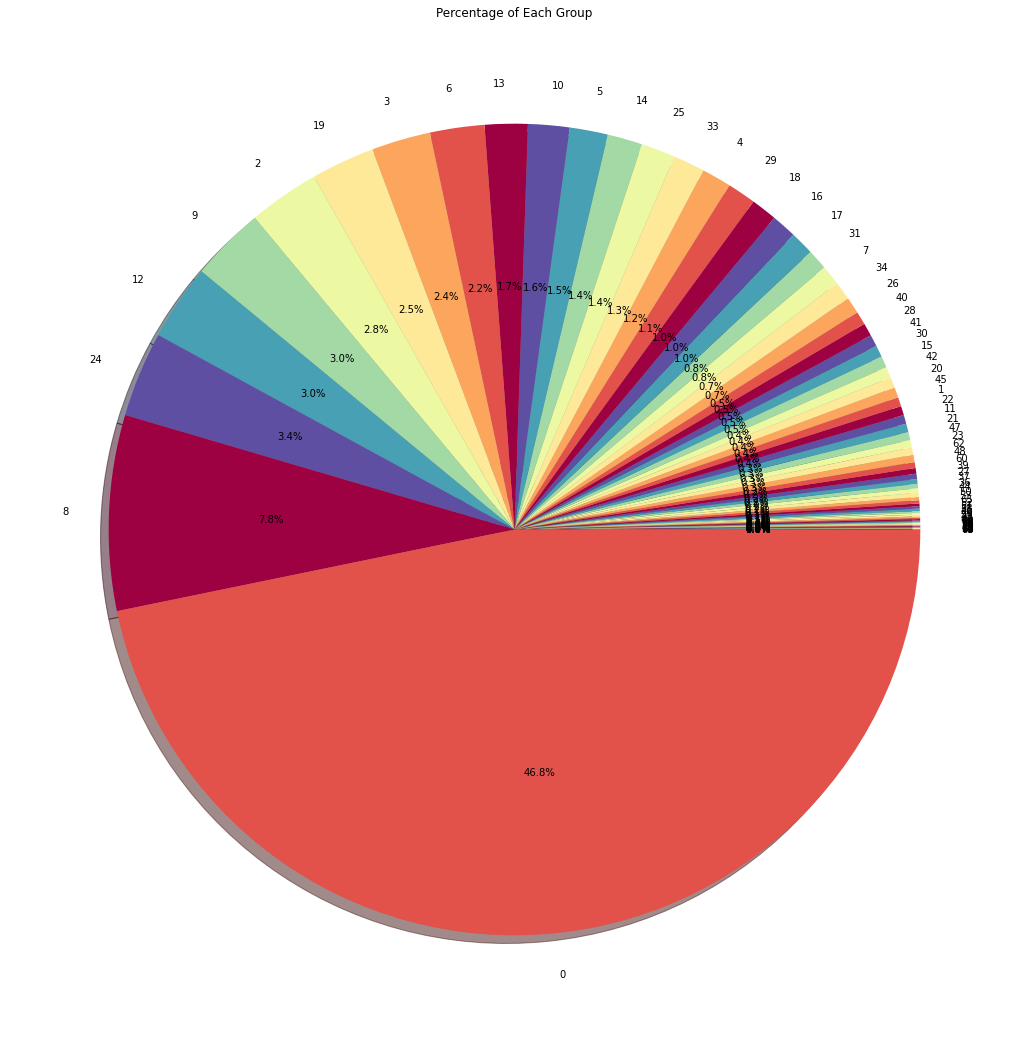
## Word Length Distribution



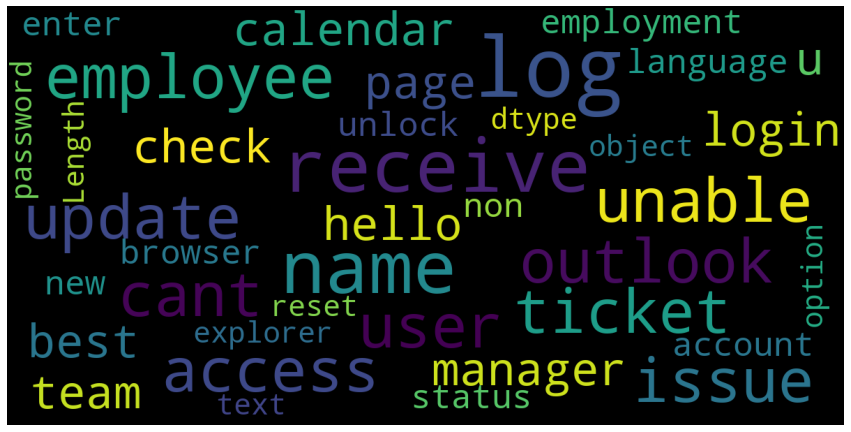
## Character Length Distribution

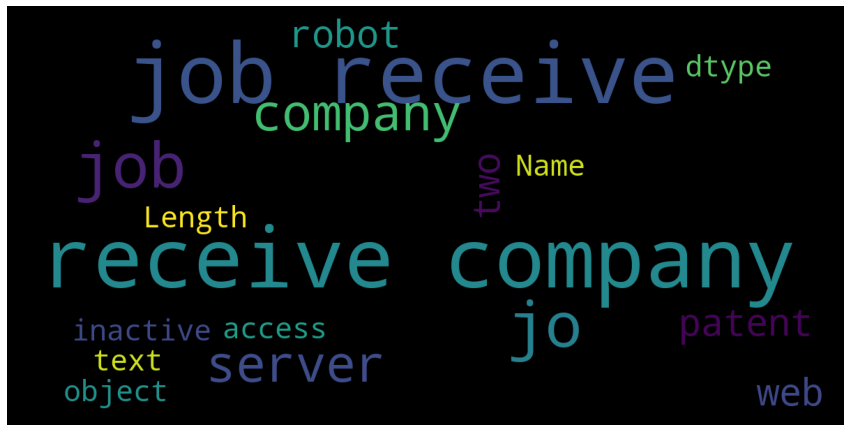


Percentage Split of Top 20 Groups Having 80+% of records



## Word Cloud After Data Pre-processing and Cleaning

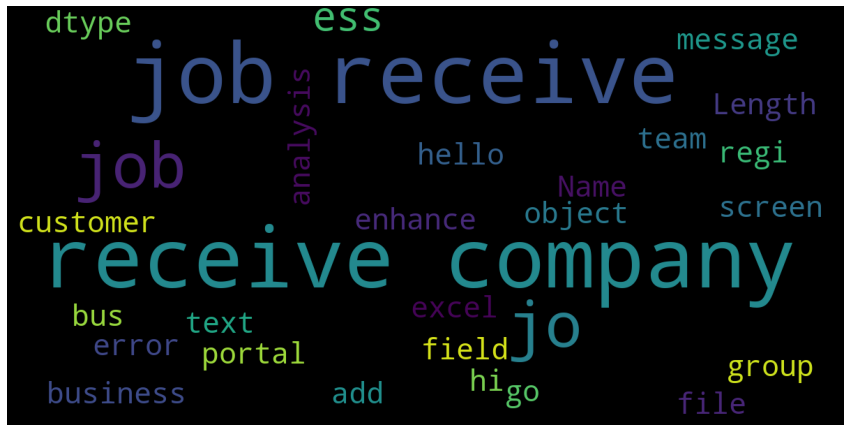
Grp\_0

Grp\_8

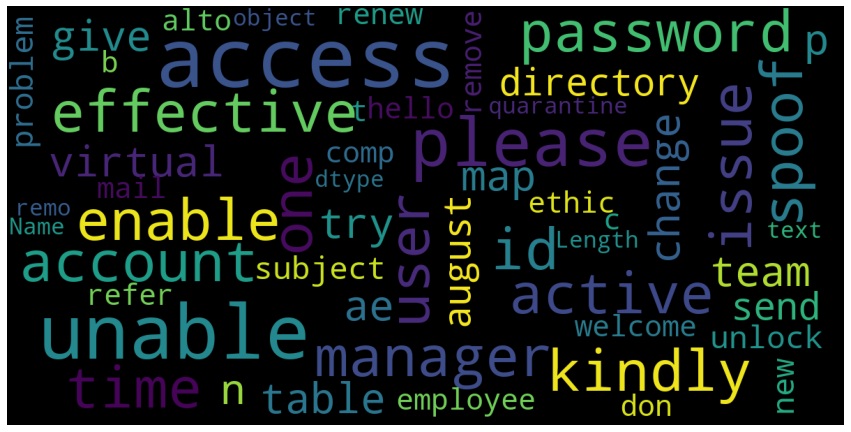
## Grp\_12

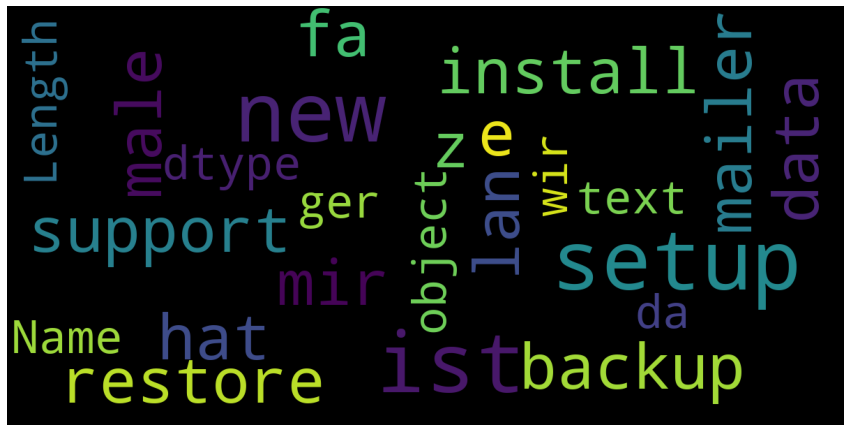


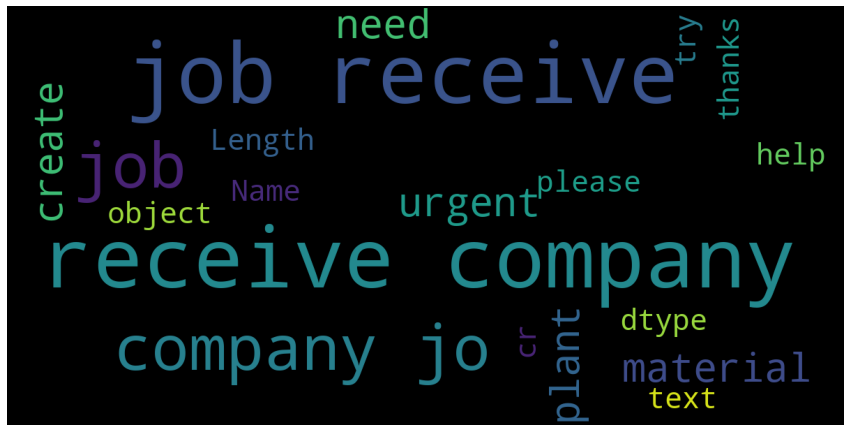
## Grp\_09

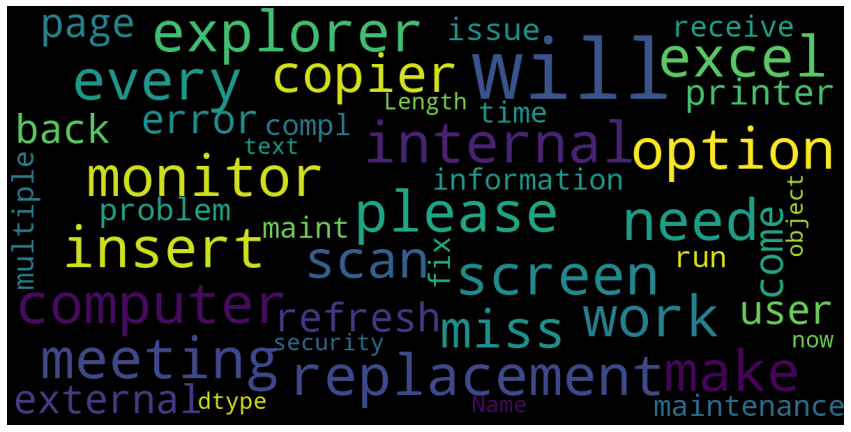


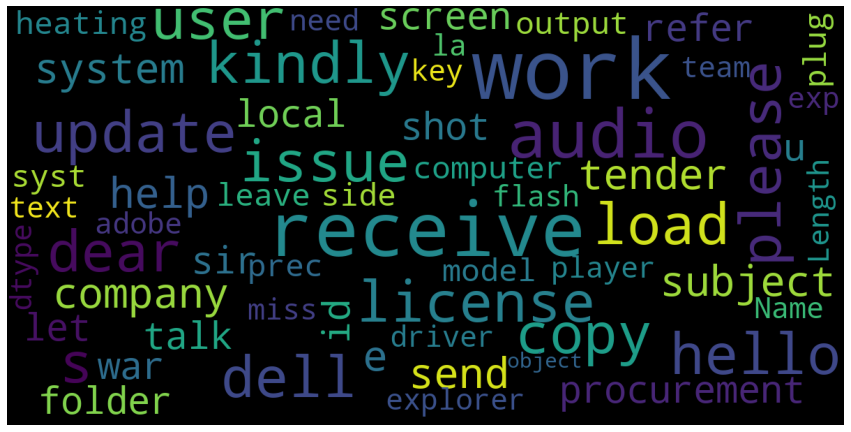
## Grp\_02

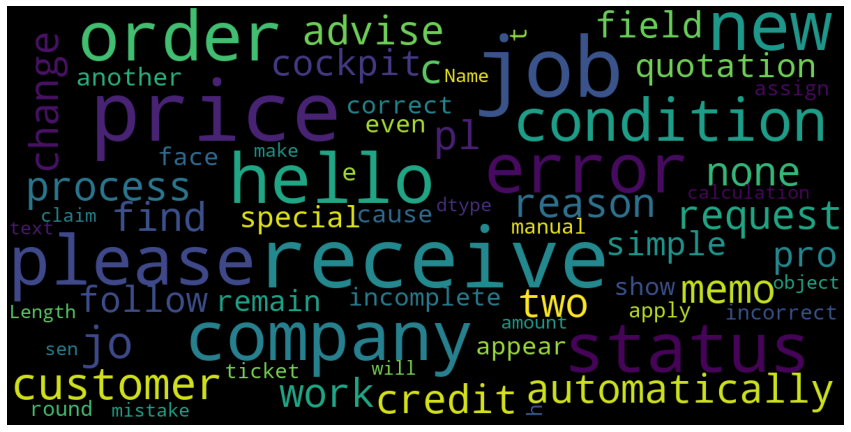


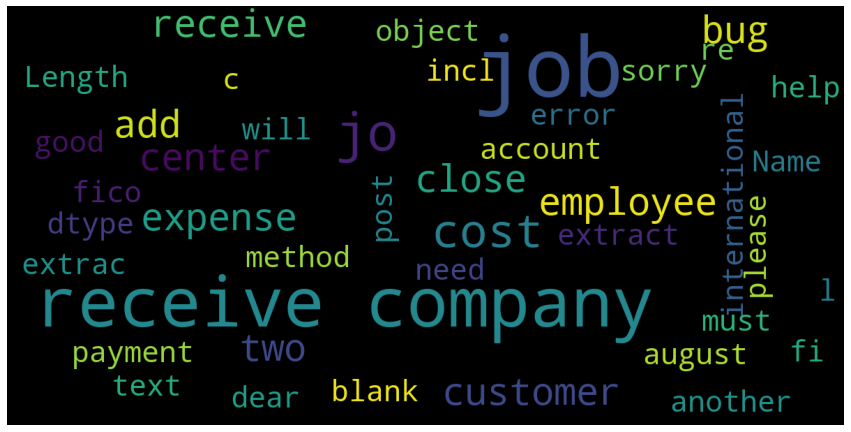
Grp\_24

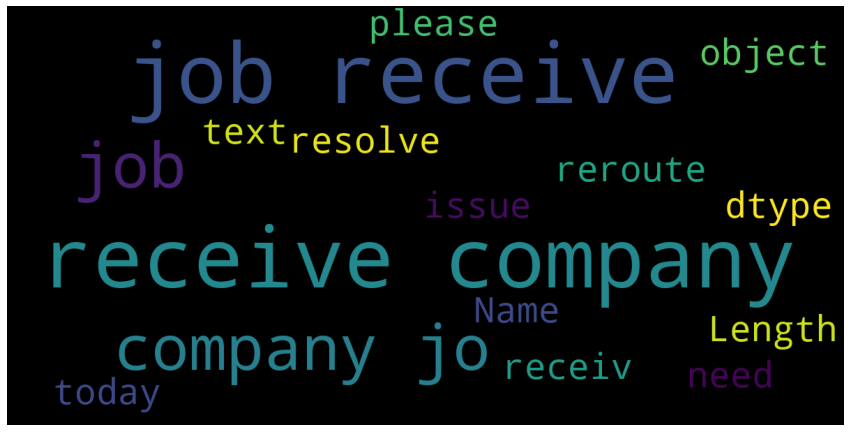
Grp\_06

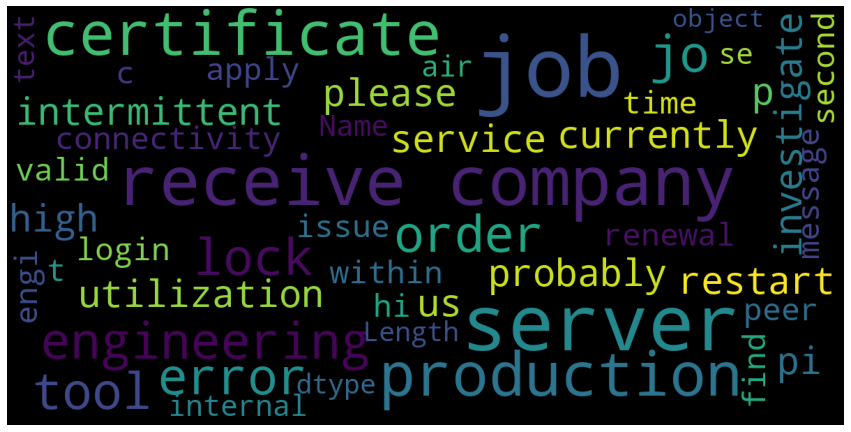
Grp\_03

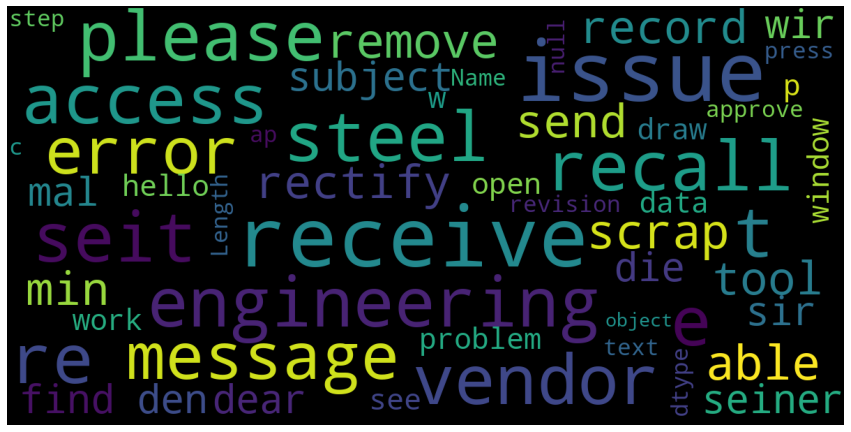
Grp\_19

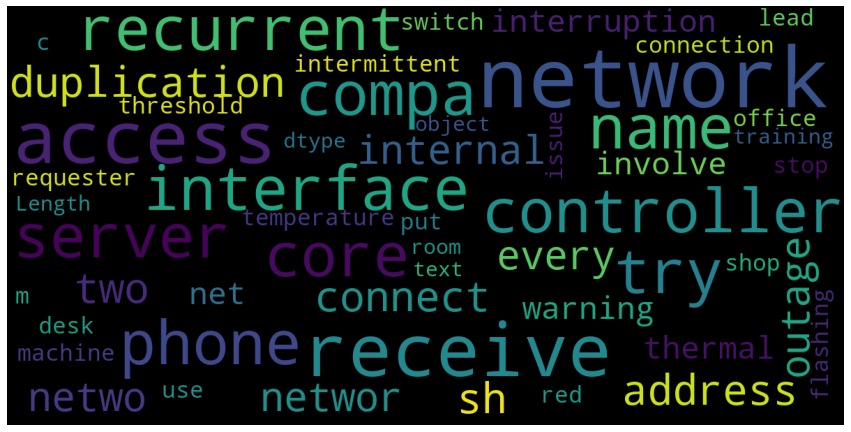
Grp\_13

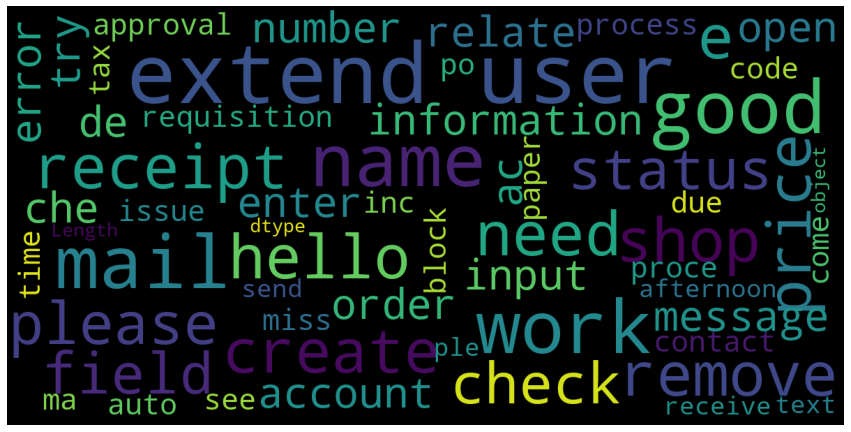
Grp\_10

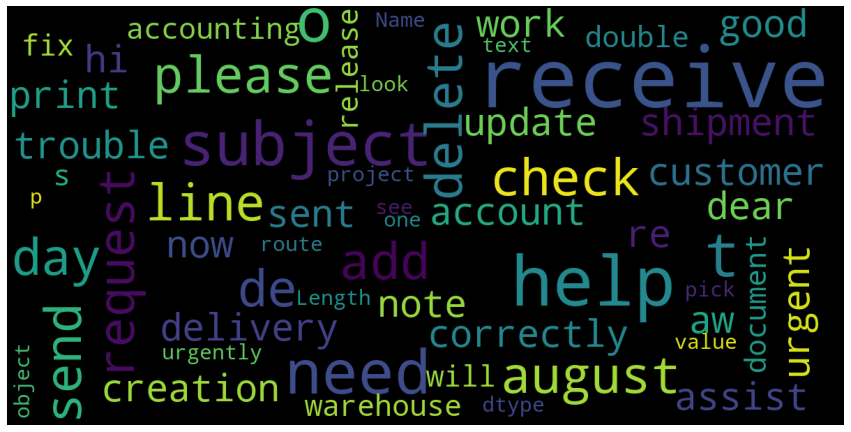
Grp\_05

Grp\_14

Grp\_25

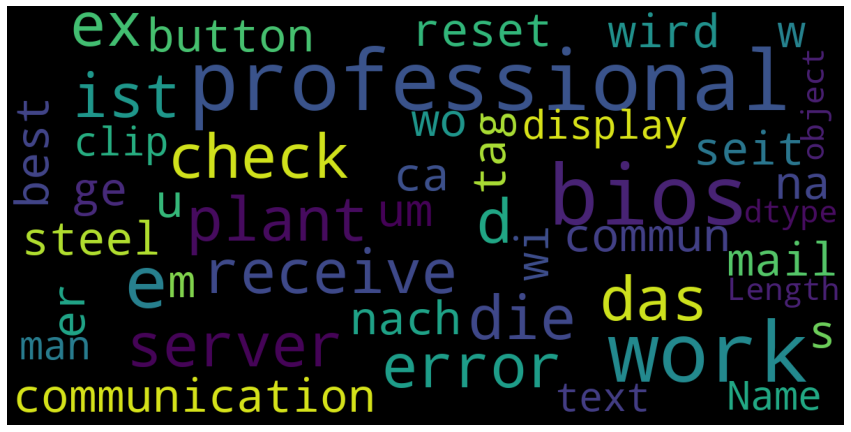
Grp\_04

Grp\_29

Grp\_18

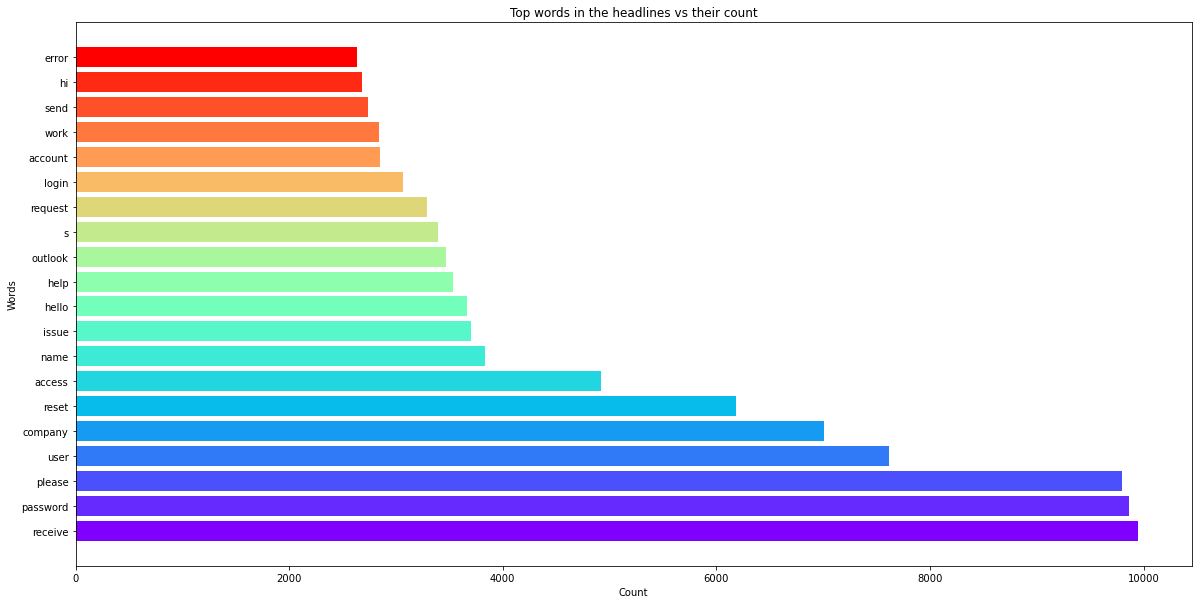
Grp\_17

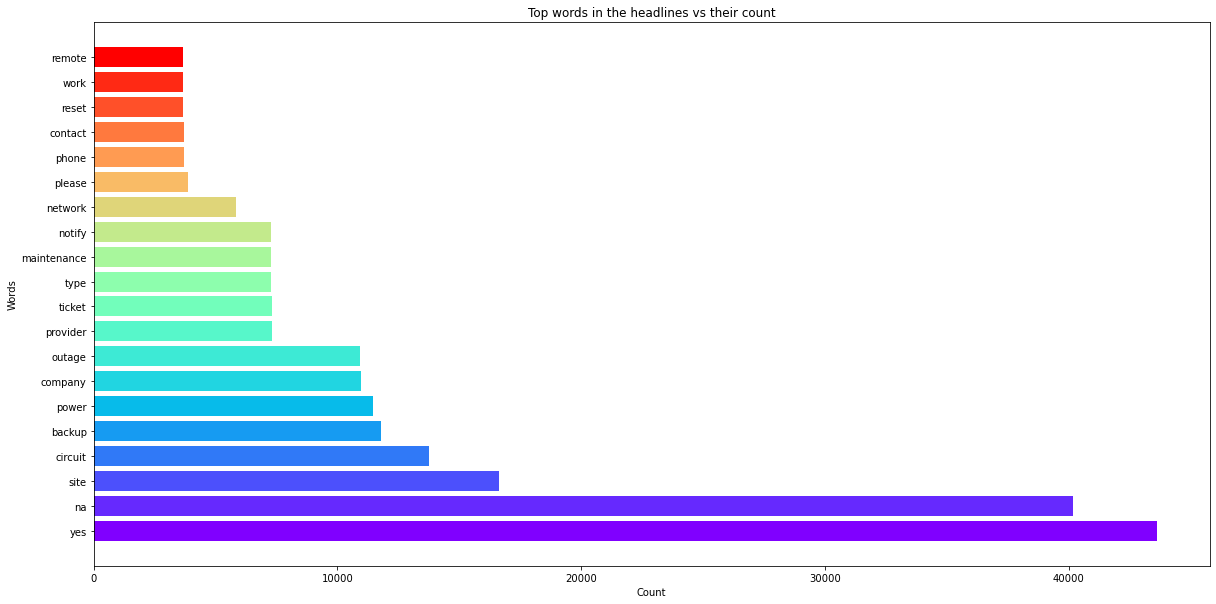
Grp\_16

Grp\_33 

## Charts

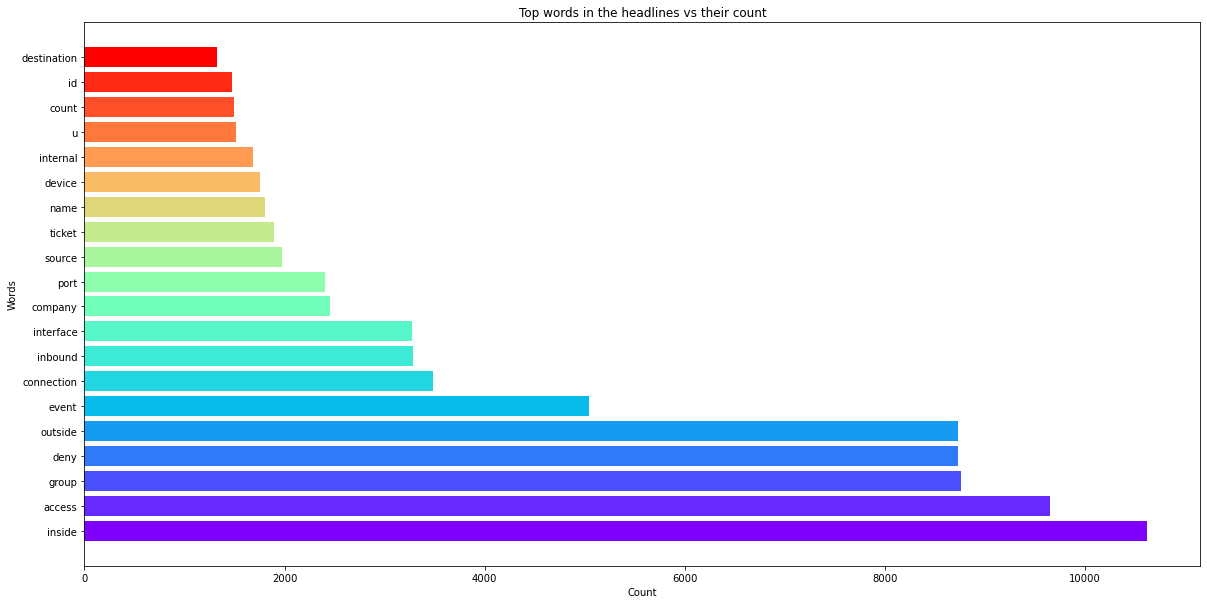
## Frequent words used in Top 20 groups [Sharing charts for a few of them]

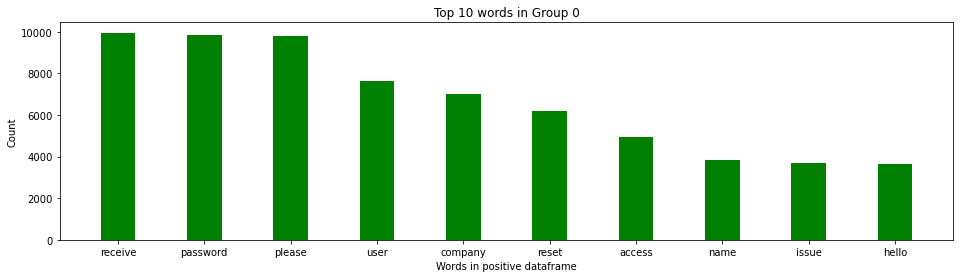
Grp\_0 

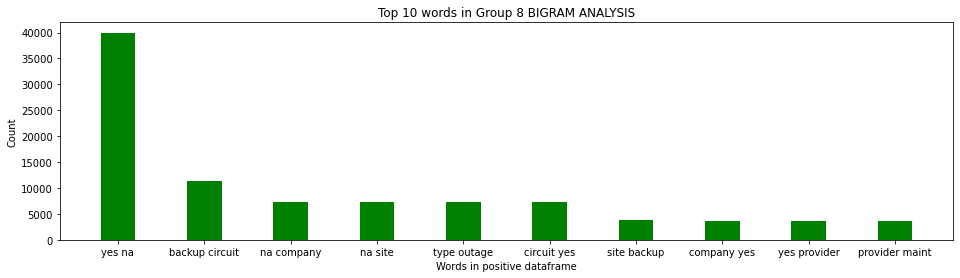
Grp\_8 

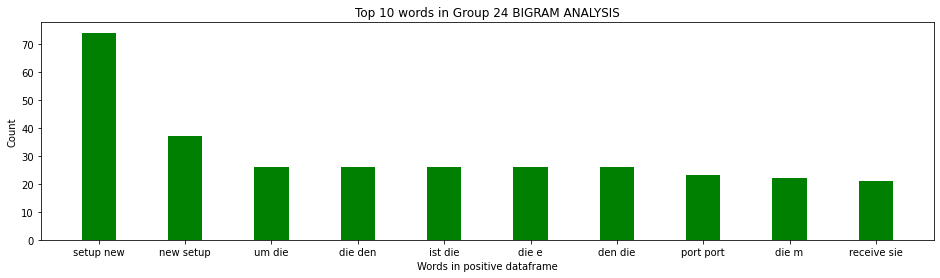
## Grp\_24

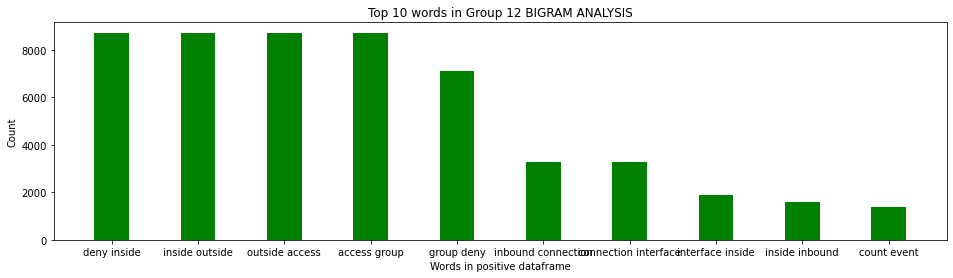
## Grp\_12

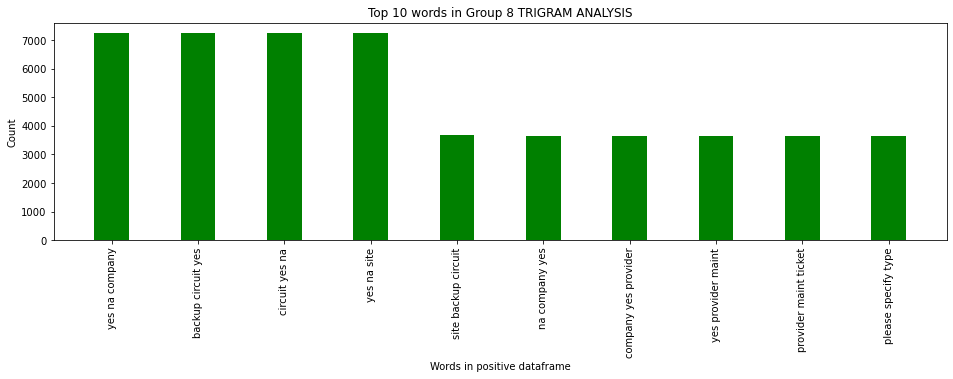


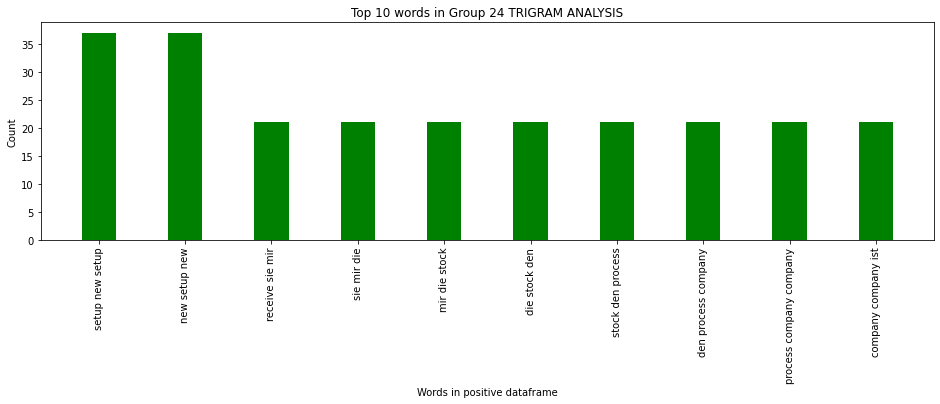
Unigram

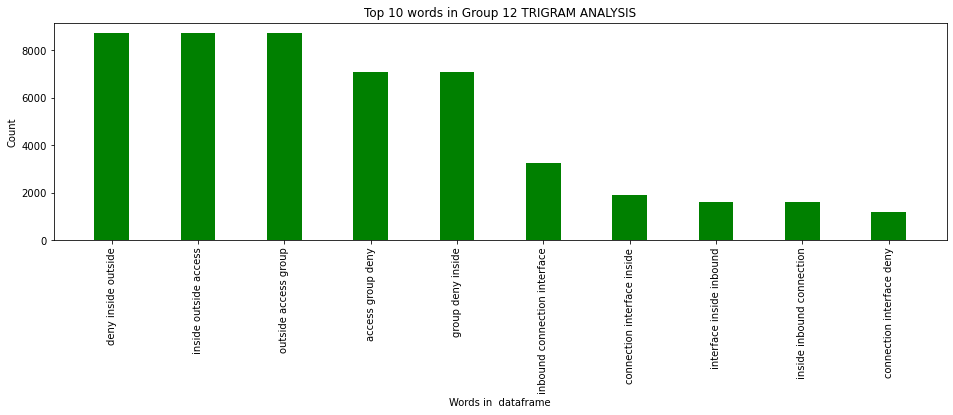
Bigram



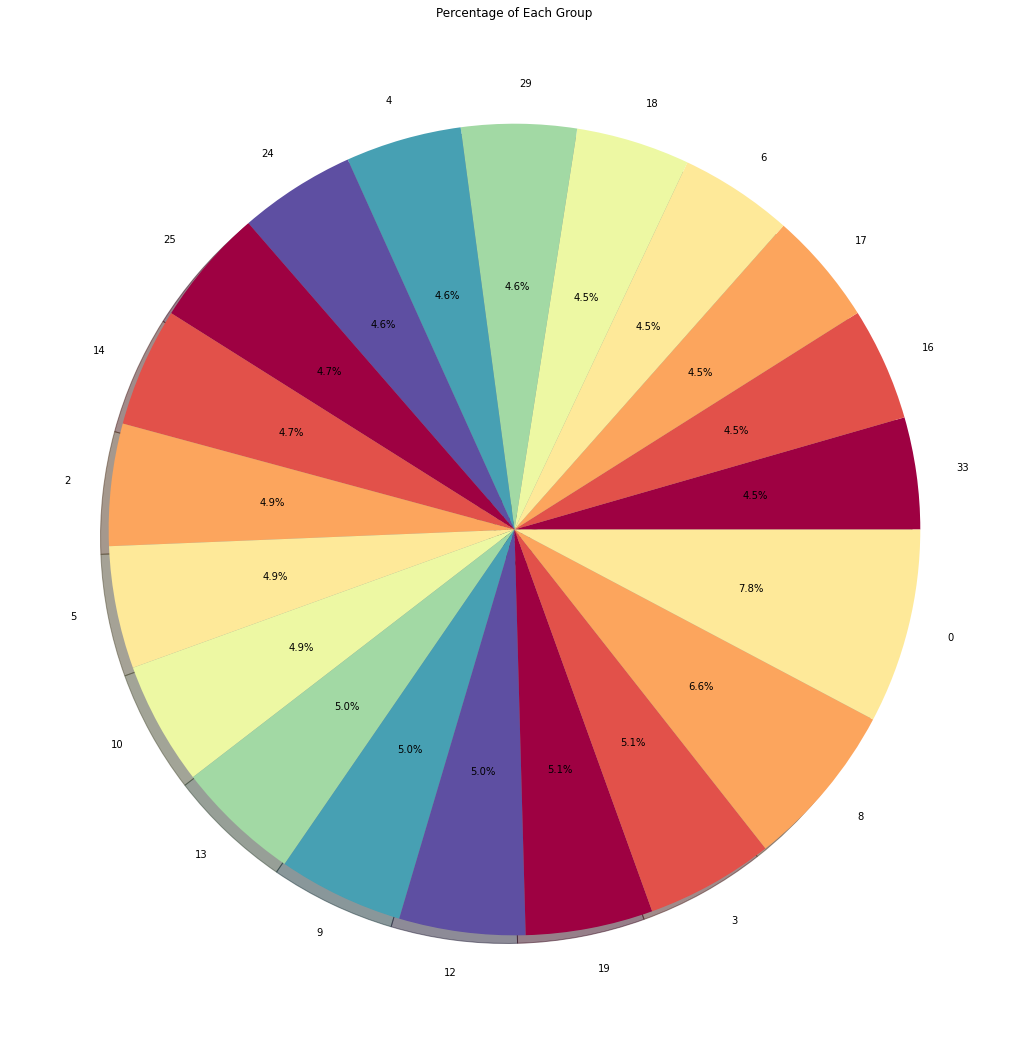


Trigram





Down sampling majority and Up sampling minority classes



# 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.

Count Vectorization

TF - IDF Vectorization

Word Embedding

Naive Bayes

Logistic Regression

XGBOOST

GLOVE EMBEDDING

PRETRAINED EMBEDDING MATRIX

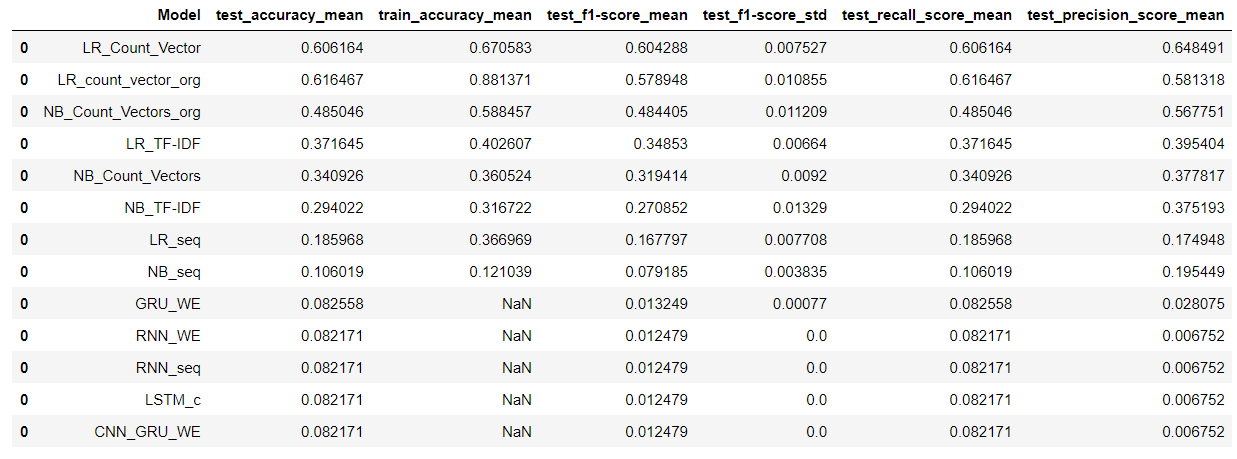
RNN MODEL

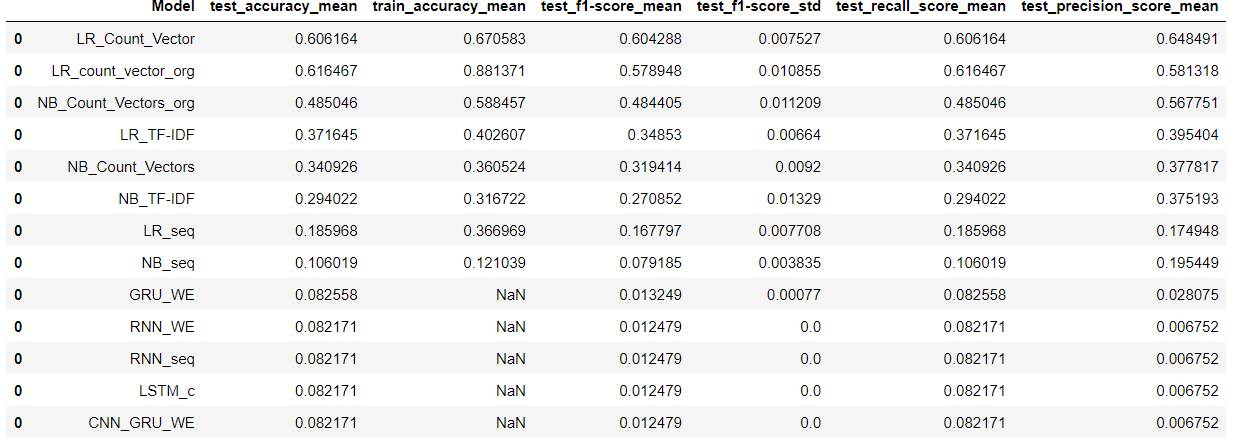
LSTM

CNN GRU MODEL

GRU MODEL

# Train and Test Accuracy for each model





# Model performance - Approaches to improve model

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 higher than that of the current.

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.

# Link to code and references