**Customer Feedback Analysis of Apple Products**

Submitted by

**Void Programmers**

Anirud Vem A20199565

Maruthi Sankar Nanduri A20224542

Monish Vallamkonda A20213053

Nikhil Gunti A20220906

Oklahoma State University

MSIS5600- Programming for Data Science (R and Python)

December 10th, 2019

Table of Contents

[Executive Summary 1](#_Toc26894868)

[Project Schedule 2](#_Toc26894869)

[Updated GANTT Chart (Deliverable 2) 2](#_Toc26894870)

[Initial GANTT Chart (Deliverable 1) 3](#_Toc26894871)

[Resource Management 3](#_Toc26894872)

[Time and Resource allocation 4](#_Toc26894873)

[Statement of Scope 5](#_Toc26894874)

[Data Preparation 6](#_Toc26894875)

[Data Access 6](#_Toc26894876)

[Data Consolidation 6](#_Toc26894877)

[Data Cleaning 7](#_Toc26894878)

[Data Transformation 7](#_Toc26894879)

[Data Reduction 7](#_Toc26894880)

[Data Dictionary 8](#_Toc26894881)

[Visualizations: 10](#_Toc26894882)

[Descriptive Statistics and Analysis: 13](#_Toc26894883)

[Topic Extraction 13](#_Toc26894884)

[Text Mining and Sentiment Analysis 15](#_Toc26894885)

[Text Mining 15](#_Toc26894886)

[Sentiment Analysis 17](#_Toc26894887)

[Comparison of words in sentiments 20](#_Toc26894888)

[Competitor Analysis 21](#_Toc26894889)

[Named Entity Recognition 22](#_Toc26894890)

[Conclusions 24](#_Toc26894891)

[References 24](#_Toc26894892)

[Appendix 25](#_Toc26894893)

[Reduced Columns 25](#_Toc26894894)

[Parts of Speech Tagging 26](#_Toc26894895)

# Executive Summary

Prospective modern-day customers read the product reviews and feedback given by experts and other users on social media before the purchase of new gadgets. This make companies cautious about their tech products and address the concerns and incorporate the recommendations of buyers in the next version of their products. An Analysis of the tech review and customer feedback on twitter, provided an insight into what the customers seek from Apple’s iPhone Series.

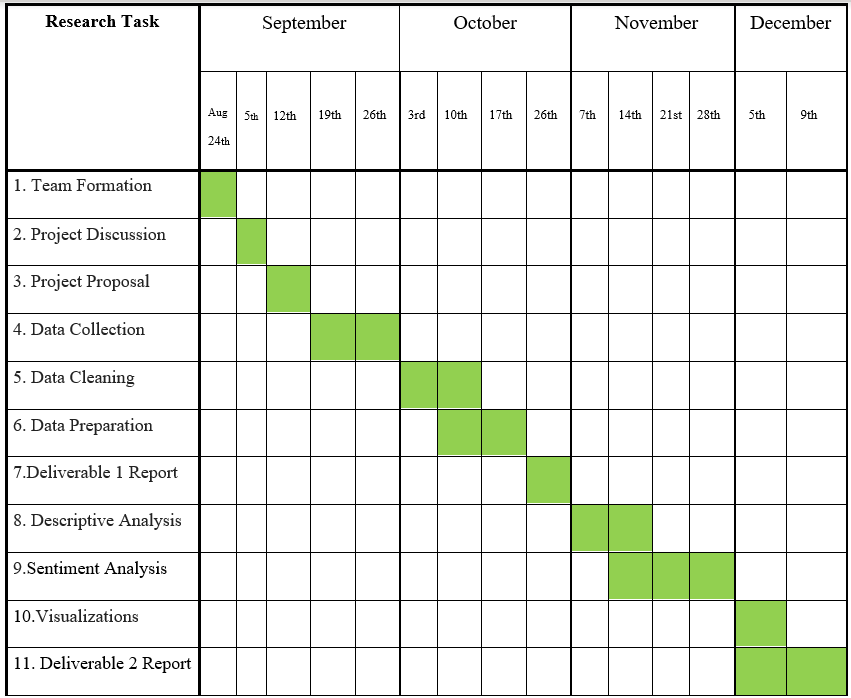
The process started with extracting tweets from twitter using Twitter API. The tweets are collected by filtering from frequently used hashtags like #iPhone11, #iPhone11Pro etc. and famous twitter handles such as the Apple CEO Tim Cook. Post collection, data is cleaned to make it devoid of duplicates, punctuation flaws, numbers, URL’s etc which cleared the ground for data analysis. This is followed by text mining that started with tokenization of the text and removing the redundancies from the text by using stop words, most and least frequent words and stemming.

After tokenization, the data is categorized into 10 unique sentiments - positive, anger, sadness, joy, trust, fear, surprise, disgust and anticipation. Top words in each of the mentioned sentiment categories are listed and then identified the user opinion on key features. Users are positive about camera, colour and apprehensive about the glass and durability. The tweets are also classified into broad topics by using topic modelling techniques like Latent Dirichlet Allocation (LDA) and Non-negative Matrix factorization. This provided insight into the themes of different topics covered by the tweets. Finally, a comparative analysis of the key features from iPhone and its competitors like Google and Samsung is performed to understand the gap in iPhone’s features. This analysis is a key takeaway in identifying the areas of concern that need more focus.

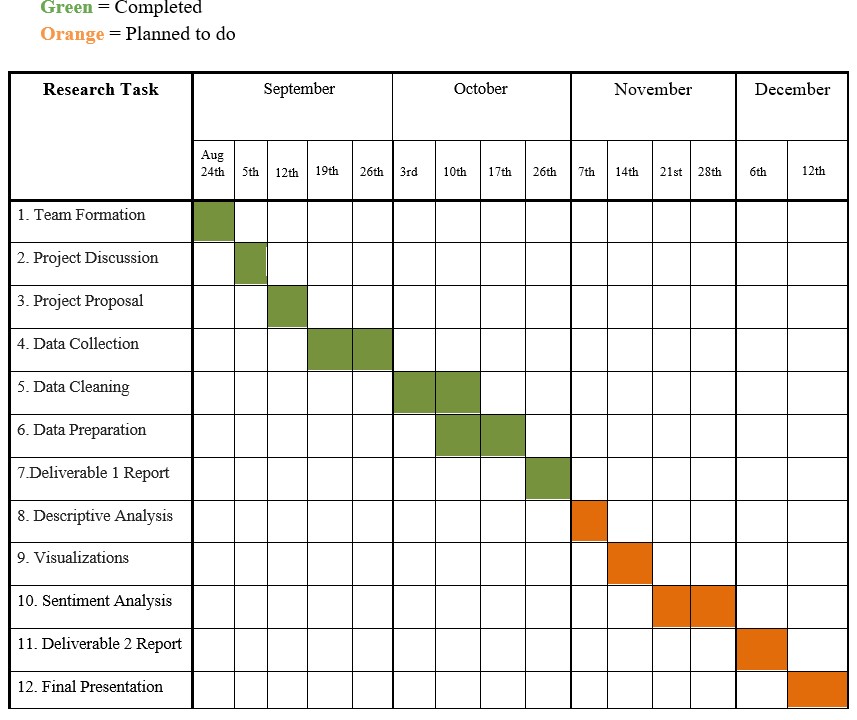
# Project Schedule

The project is completed by second week of December, within 15 weeks duration. The initial 3 weeks were spent on the onboarding, idea of the project and project proposal. As the data collection includes web scraping, two weeks were allocated to get a hold on this technique to collect ample data for analysis. Since the analysis dealt with unorganized text data, data cleaning and data preparation were completed in two weeks. This marked the completion of first phase of the project. Post data preparation, there was minimal deviation from the original plan due to adhoc academic tasks. In addition to the academic work, events such as ‘Analytics Case Competition’ in the first two weeks of November have caused the changes in project schedule. Because of this change, the timelines of all the tasks starting from ‘Descriptive Analysis’ are stretched by a week when compared with the original project plan.

## Updated GANTT Chart (Deliverable 2)



## Initial GANTT Chart (Deliverable 1)



Resource Management**:**

The team met every Thursday for 3 hours (5 to 8 PM) to work on the deliverables. The tasks were divided, and the work done was aggregated at the end of the session. The agenda for consecutive meetings were scheduled promptly and required pre-requisites were acquired accordingly. The weekly timings were adjusted during the fall-break and thanksgiving holidays. The time and resource allocation to each task are tabulated below.

## Time and Resource allocation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | **Task Name** | **Description of work** | **Number of**  **Resources at work** | **Working hours** |
| 1 | Team Formation | Team onboarding |  |  |
| 2 | Project Discussion | Discussed on the business idea to work on | 4 (Full Team) | 6 |
| 3 | Project Proposal | Prepared the proposal, finalized the project, listed down the tasks to be done | 4 (Full Team) | 5 |
| 4 | Data Collection | Scrapped the data from twitter | 4 (Full Team) | 6 |
| 5 | Data Cleaning | Removing unnecessary links, punctuations and stop words | 2(Anirudh and  Nikhil) | 8 |
| 6 | Data Preparation | Merging all the data into single data frame, Transforming into structured data | 2(Maruthi and  Monish) | 6 |
| 7 | Deliverable 1 report | Preparing the document of the work done till then | 2(Full Team) | 8 |
| 8 | Descriptive Analysis of Data | Performing basic analysis of the data and draw inferences | 4(Full Team) | 3 |
| 9 | Sentiment Analysis | Perform Sentiment analysis using libraries  in R and draw inferences | 4(Full Team) | 8 |
| 10 | Visualizations | Preparing charts, word clouds of the data | 2(Anirudh and Nikhil) | 4 |
| 11 | Deliverable Report 2 | Prepare a report of all the work done and work on any further analysis | 2(Maruthi and Monish) | 10 |

# Statement of Scope

This objective is achieved through accomplishing the below tasks:

* Web scrapping the customer review tweets on iPhone11, iPhone11Pro and iPhone11Pro Max from the data from twitter handles through API provided by Twitter
* Clean the text data obtained from web scraping and format it accordingly
* Analyze the final data obtained in the above step and Generate visualization graphs to explain the patterns in customer feedback through R/Python packages.

The scope of the analysis includes:

1. Exploration of customer demographics

2. Identify the on-demand iPhone11 features

3. Sentiment analysis of the tweets to capture the customer response on various iPhone 11 features

4. Evaluate the iPhone11 performance in comparison to the recent phones of other brands like Samsung and Google

During the project, we analyzed the data corresponding to 12000 tweets, which included features related to the tweets, retweets, user demographics, location, and the number of followers to a user to understand the impact that each tweet on its followers.

# Data Preparation

Data Access**:**

Twitter is the primary source of data in this project. Since customers use Twitter as a platform to express their feedback and reviews on technology products, it is appropriate to use the twitter data in analyzing the feedback of customers. Twitter API was used to scrape the tweets related to the recently launched iphone11 and iphone11Pro using keywords like iphone11, iphone11pro, iphone11promax, selfies, #iphone11 #iphone11, etc. The code used for extracting the data is in the Appendix.

Website used: [https://twitter.com](https://twitter.com/home)

Data Consolidation**:**

Data corresponding to multiple weeks of tweets was collected based on the keywords and hashtags mentioned in the above step. The twitter data is majorly retrieved from two files. Original file ‘unique\_tweets.csv’ consists of only the text data corresponding to the tweets from the first week of October. The second file, ‘final\_data\_tweets.csv,’ consists of the data with the twitter ‘object’ elements as columns. This dataset consists of all the demographics, location, and the followers of a user who tweeted the text. This dataset is created by extracting the tweets from the second and third week of October, along with the tweets and replies in the twitter handle of Apple CEO Tim Cook. After data cleaning, the tweets in both the above data sets were consolidated into a third data file ‘total\_text\_final.csv’ to perform the sentiment analysis. The corresponding Code and data files are included in the Appendix.

Data Cleaning**:**

Since the data obtained from twitter is random text with a lot of irregularities and irrelevant content, data had to be made ready for further analysis. The tweets consisted of duplicate tweets, additional website/blog links, Punctuation marks, irregular font and text, repeated words.

To fix these problems in data, the following tasks were performed in unique\_tweets.csv and final\_data\_tweets.csv:

• Eliminated duplicate tweets

• Removed the additional website/blog links in the data

• Transformed all the text data into lowercase

• Eliminated the punctuations from the data

• Removed generic “stop words” and most frequently repeated words from the data.

Data Transformation**:**

There are no numeric columns to be transformed in the analysis. Most of the data transformation with respect to text is completed in the data cleaning step.

Data Reduction**:**

There were three datasets used in the analysis where two of them has only one text column. The dataset ‘final\_data\_tweets.csv’ consists of the variables related to the user and tweet details. Out of the 89 variables, only 18 variables are relevant for the analysis. The rest of the features are reduced from the final data frame ‘df\_final\_data’. A list of the reduced attributes is mentioned in the appendix.

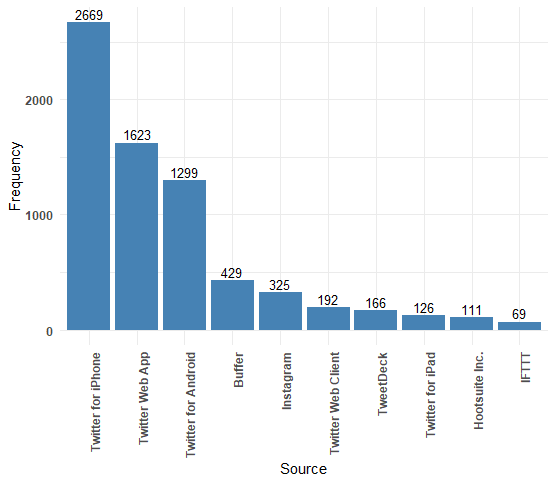
Data Dictionary **(df\_final\_data):**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Attribute Name** | **Type** | **Description** |
| 1 | user\_id | Int64 | unique twitter name |
| 2 | created\_at | String | date of tweet created |
| 3 | screen\_name | String | The screen name(unique), handle, or alias that this user identifies themselves with. |
| 4 | text | String | Tweet text |
| 5 | source | String | device from which the tweet was tweeted |
| 6 | is\_retweet | Boolean | Boolean flag indicating if the tweet is retweet to any other tweet |
| 7 | favorite\_count | Int | [Indicates approximately how many times this Tweet has been liked by Twitter users.](https://developer.twitter.com/rest/reference/post/favorites/create) |
| 8 | retweet\_count | Int | the number of times this tweet is retweeted |
| 9 | reply\_count | Int | The number of replies given to this tweet |
| 10 | hashtags | Array of  Hashtag Objects | Represents hashtags which have been parsed out of the Tweet text. |
| 11 | place\_full\_name | String | Full human-readable representation of the place’s name |
| 12 | place\_type | String | The type of location represented by this place |
| 13 | country | String | Name of the country containing this place. |
| 14 | country\_code | String | Shortened country code representing the country containing this place |
| 15 | location | String | The user-defined location for this account’s  profile. |
| 16 | description | String | The user-defined UTF-8 string describing their account. |
| 17 | followers\_count | Int | The number of followers this account currently has. |
| 18 | account\_lang | String | The language in which the user account operates |

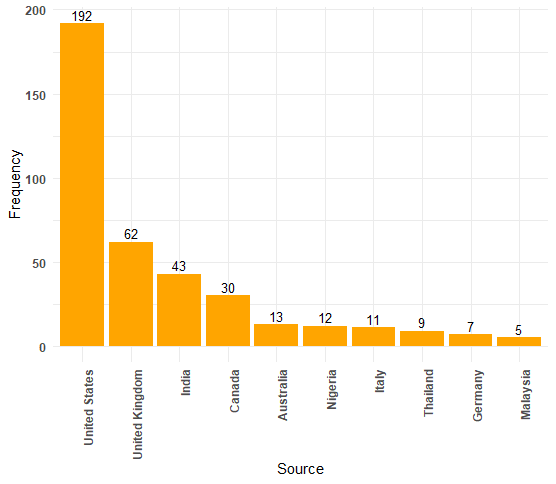
# Visualizations:

1. Tweets based on the source

The below bar graph explains the number of tweets from each unique source – mobile, web, android, iOS, windows etc. used to tweet. This graph is plotted based on the information present in ‘final\_data\_tweets.csv’

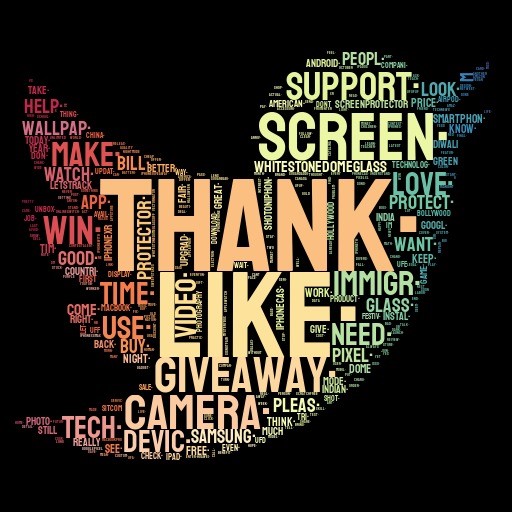


1. Frequency of tweets based on Geographical location:



To look at the number of tweets from each country, “United states” is the highest with 192 tweets followed by “United Kingdom” with 62 and “India” with 43.

1. Word Cloud of the Text Data



The above visualization is the world cloud of all the tweets about iPhone from the data obtained by scrapping from twitter. We see that mostly people discuss about Camera, Screen, Glass, Video, Support, Like, Thank, Giveaway when they express their opinion about iPhone in twitter. Where Thank in the word cloud means customers are tweeted about the Thanks update program. It is also observed that people express their opinion comparing with other competitive models of Samsung phones and Pixel phones which is clearly seen in the world cloud.

# Descriptive Statistics and Analysis:

As the analysis is about the data extracted from twitter, there are no predefined sentiments tagged to the tweets in the data. So, Sentimental Analysis is performed. Since there is no scope for building models based on training and validation data sets, topic modelling techniques are used on the tweets. Latent Dirichlet allocation (LDA) and Non-Negative Matrix Factorization (NMF) are the two topic modelling techniques used in the analysis.

LDA is a natural language processing technique that creates topics based on the co-occurrence of words in documents. If a set of words appear more frequently together in a set of documents, then that denotes a topic. Each document (or tweet in our example) can contain multiple topics. Some topics are represented less in a document than others depending on word frequency.

Non-negative matrix factorization (NMF or NNMF), also non-negative matrix approximation algorithms in multivariate analysis and linear algebra where matrix is factorized into usually two matrices W and H, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. NMF has an inherent clustering property i.e., it automatically clusters the columns of the input.

## Topic Extraction

1. Latent Dirichlet Allocation (LDA).

There were 5 topics created out of the words using the LDA method and the top words in each topic are mentioned above. For example, Topic 0 has the words related to the iPhone competitors like pixel, Samsung, tech and camera which is a competitive feature



1. Non-negative Matrix Factorization (NMF):

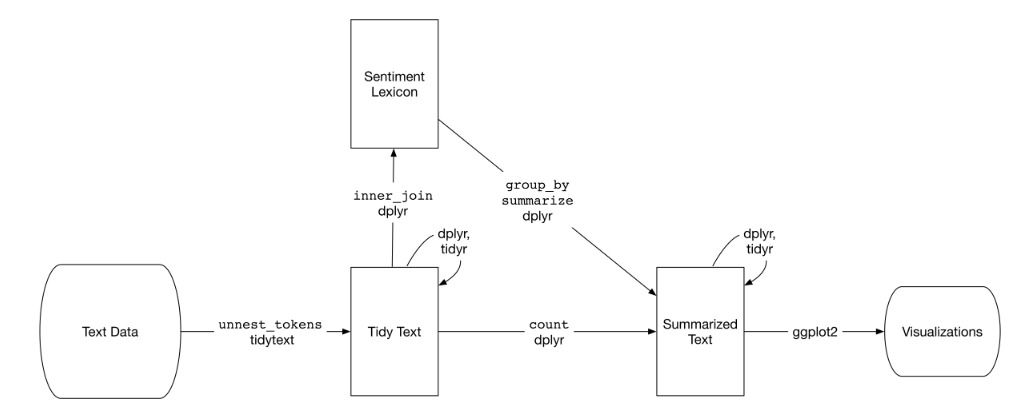
Five topics extracted using NMF technique describe the different categories of tweets by customers. For example, topics in the above figure explain that customers are interested about the durability of screen and the glass and the associated device protection plans.



# Text Mining and Sentiment Analysis

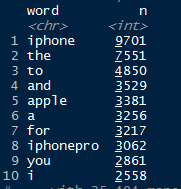
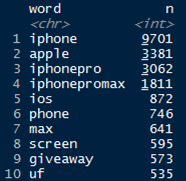
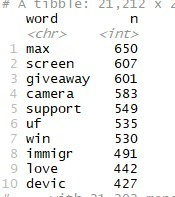
## Text Mining

The sentiment analysis of the customer’s tweets is done using the package ‘tidytext’ in R. The flow chart of the sentiment analysis which is performed through tidytext is given below.



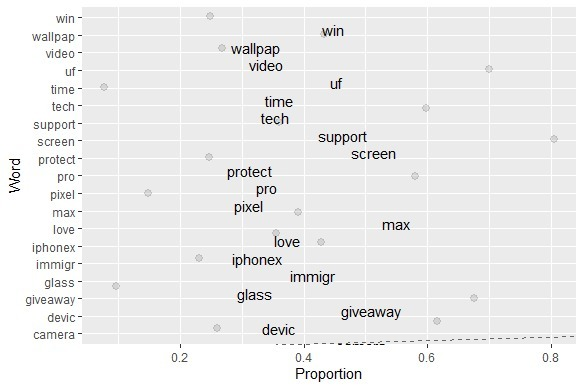
The process starts with the tokenization of text data. It is done by splitting the text into meaningful smaller words or phrases along with a count of the same. Tokenization is complimented with the below steps:

1. Generation of tokens – words or phrases
2. Removing the stop words
3. Removing the digits from the text data
4. Removing the white and blank spaces
5. Removing most frequent and least frequent terms
6. Stemming the tokens

Generation of Tokens > Removal of stop words > After Stemming

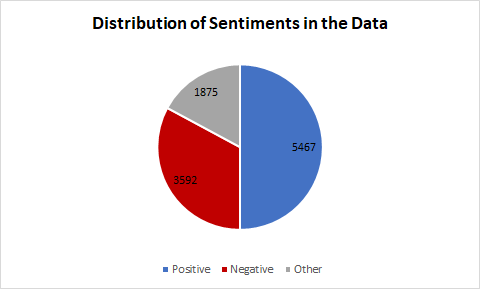
Proportion of words after stemming:



Post this process, the data ready for sentiment analysis.

## Sentiment Analysis

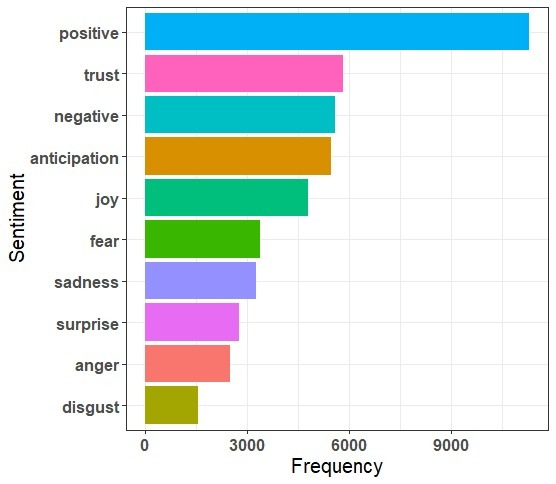
Once the data is cleaned and it was used for sentiment analysis. Broadly the sentiments are categorized into positive, negative and other and their distribution is charted below.



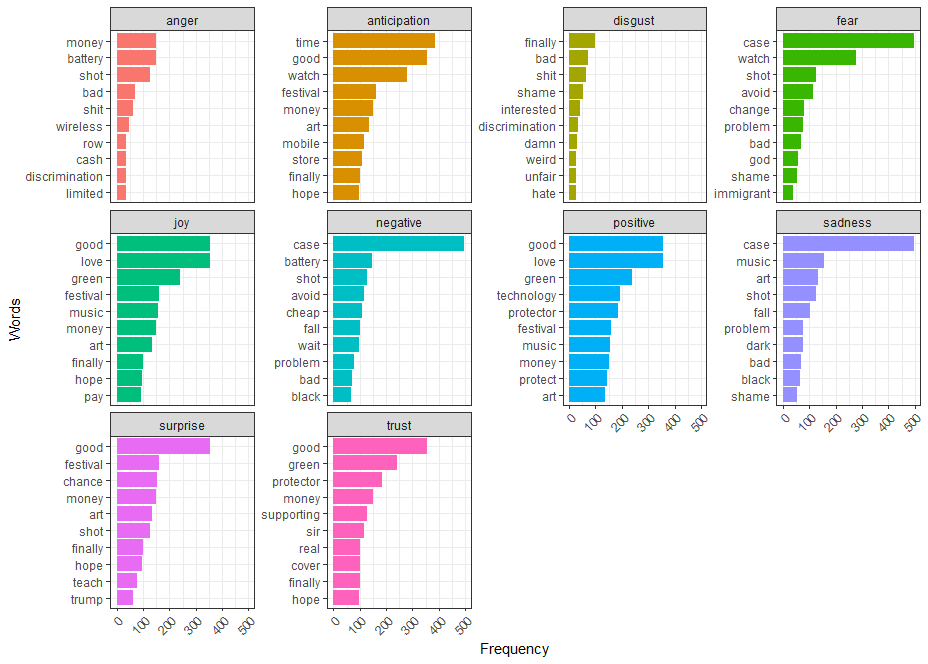
It is observed that:

1. 50% of total data which is 5467 tweets have positive sentiment.
2. 32.8% of total data which is 3592 tweets have negative sentiment.
3. 17.2% of total data which is 1875 tweets have other sentiments like Joy, Anger, Surprise etc

The words in other are split into further categories of sentiments such as joy, sadness, anger etc. and corresponding word frequency is plotted below:



The top 10 frequent terms in each sentiment are identified and visualized as below:

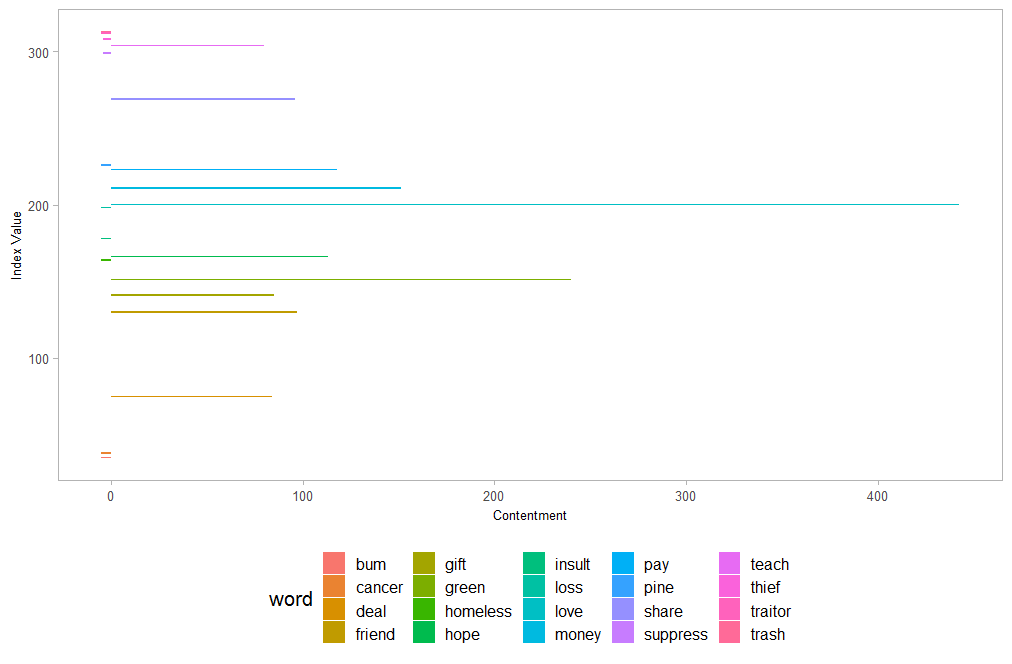


Summary about top words in each sentiment:

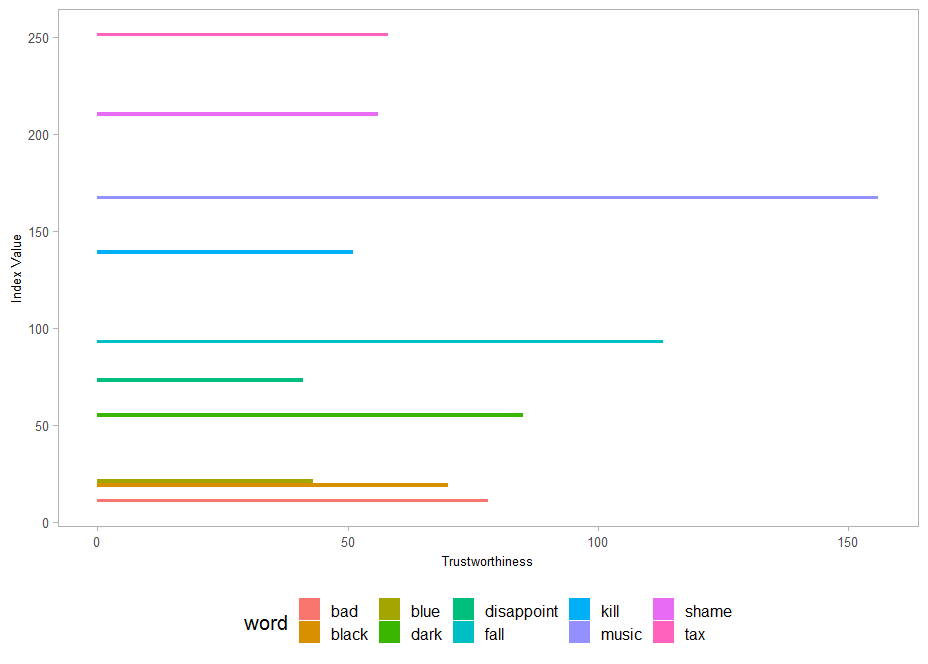
1. It is observed that words such as money, battery and wireless are tagged to the sentiment **anger** indicating customers are concerned in those aspects.
2. Words such as good, finally, hope, time and store are tagged in **anticipation** implying that the customers are eagerly looking forward to buying the iPhone11
3. iPhone11 customers are noted to be afraid with respect to the hardware, camera etc. as we see the words case, shot etc in sentiment **fear**
4. The customers are joyfully waiting for the green colour iPhone11 where the set of words in **joy** sentiment are comparable to those of **anticipation.**
5. In all the positive sentiments customers are happy about the technology, the good features of iPhone and expressed their love for the apple products.
6. Few customers expressed their sadness and concerns about the case and its pricing which is observed from the **sadness** sentiment

## Comparison of words in sentiments

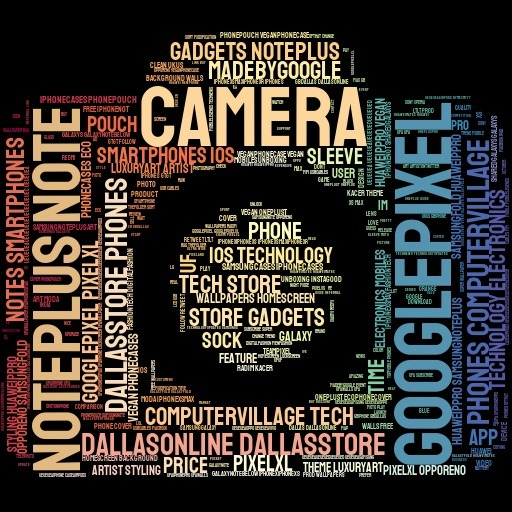
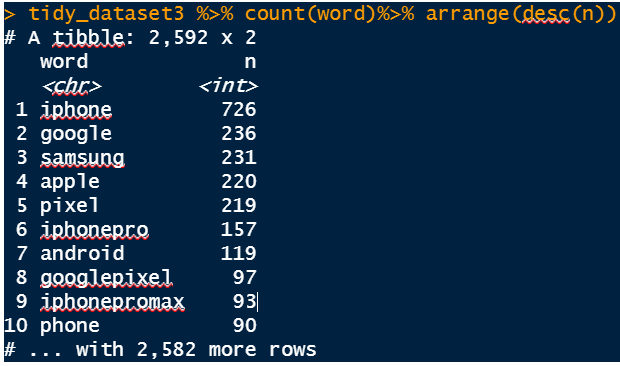
**Joy vs Sadness**



**Sadness vs Surprise**



## Competitor Analysis

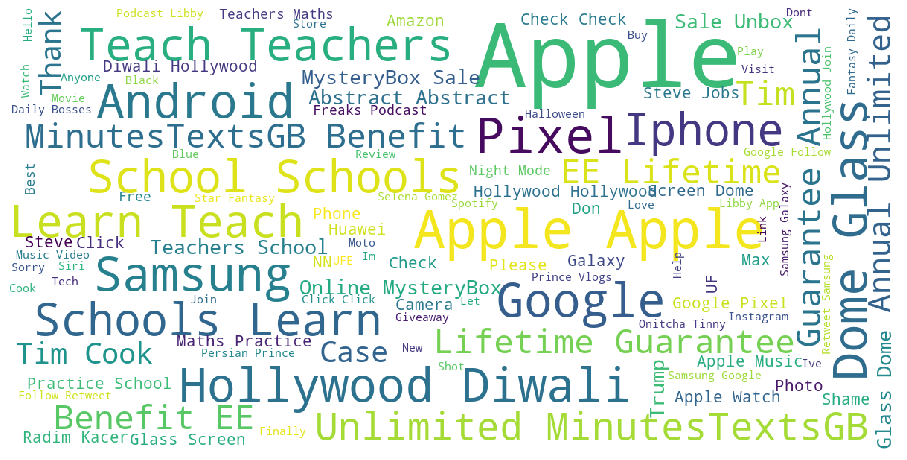


It is important to analyse the features of iPhone competitors to understand the area of improvement. The word frequency table with top 10 rows and the word cloud above attempt to explain the aspect of competitor brand comparison. Key takeaways from the word cloud:

1. Samsung Note, Google Pixel are the competitors with which the features of iPhone11 are compared to.
2. Camera is the most discussed feature and most compared feature explaining that the customers are giving top preference to camera in the digital world.
3. Customers feel the price of iPhone11 is high when compared to other brand mobile phones of similar league.

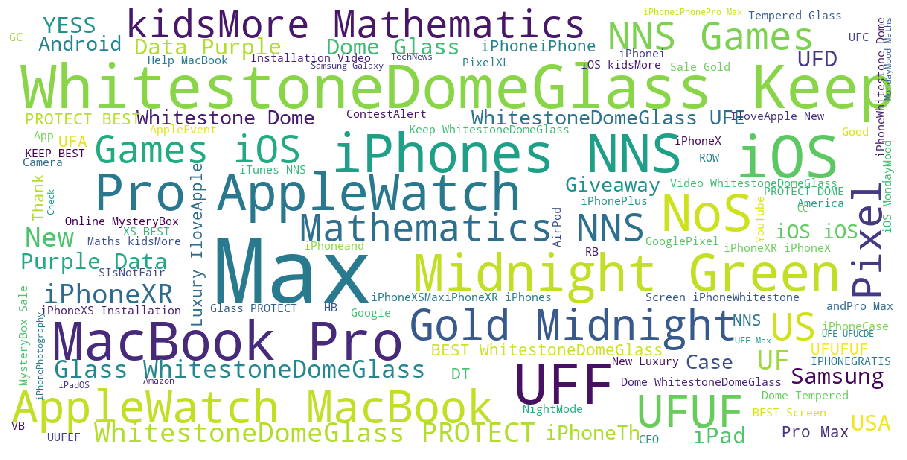
## Named Entity Recognition

1. Person:



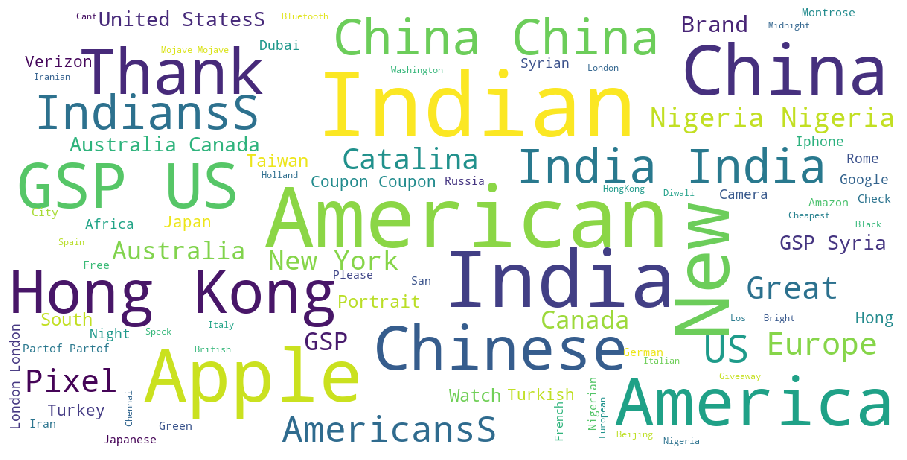
Observation : Entities like Apple, Samsung, Pixel, iPhone, Android Standout. previous findings that customers are comparing apple iPhone with competitive mobile brand like Samsung and Google is re-iterated. Also, iOS was compared to android.

1. Organization



**Observations:** Entities that standout are Whitestone Dome Glass, MacBook, Green, Max Organizations are not categorized correctly by NER in this field. Though it was identified before that customers talked about Glass a lot and are happy about it.

1. Location:



**Observations:** Entities that standout are American, India, US, China, Hongkong. We have observed that most of tweets in data are from the customers from the Countries like United States of America, India.

# Conclusions

* Overall customers are very happy and joyful about the new green color iPhone launched for iPhone 11 or iPhone 11 pro models.
* Customers compared the iPhones with other competitive models of Samsung and Google.
* Mostly the comparisons are done with specific model of Samsung Note and Google Pixel phones.
* The features like Camera and overall technology is compared the most. Also there are many tweets about the price comparison.
* Most of the tweets were posted from customers of United States, United Kingdom and India.
* There are more positive tweets in the data collected than negative or other.

References

1) <https://www.tidytextmining.com/sentiment.html>

2) <https://twitter.com>

# Appendix

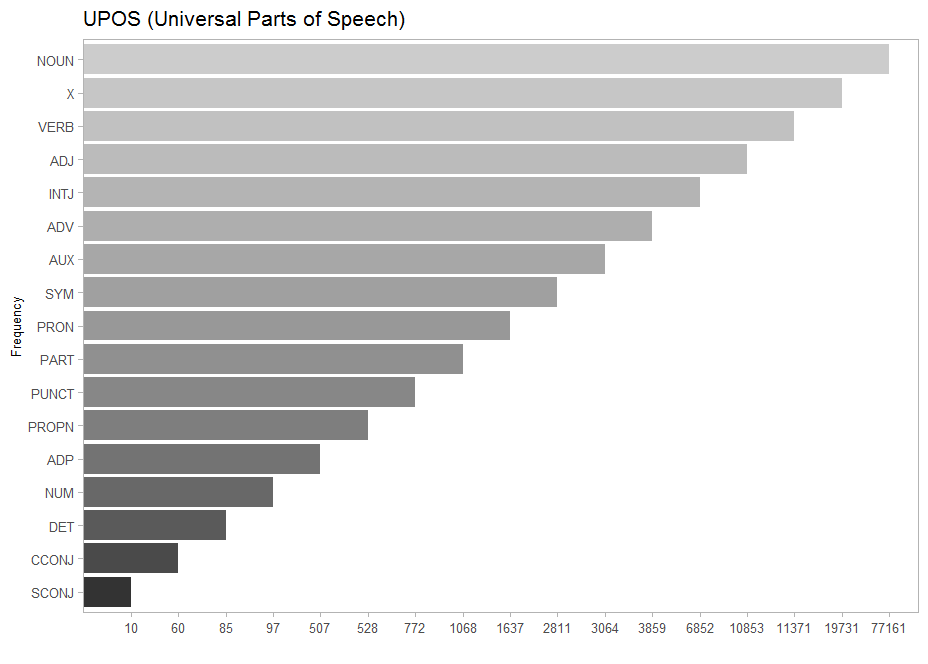
## Reduced Columns

|  |  |  |  |
| --- | --- | --- | --- |
| status\_id | ext\_media\_type | quoted\_verified | geo\_coords |
| display\_text\_width | mentions\_user\_id | retweet\_status\_id | coords\_coords |
| reply\_to\_status\_id | mentions\_screen\_name | retweet\_text | bbox\_coords |
| reply\_to\_user\_id | lang | retweet\_created\_at | status\_url |
| reply\_to\_screen\_name | quoted\_status\_id | retweet\_source | Name |
| is\_quote | quoted\_text | retweet\_favorite\_count | url |
| quote\_count | quoted\_created\_at | retweet\_retweet\_count | protected |
| symbols | quoted\_source | retweet\_user\_id | friends\_count |
| urls\_url | quoted\_favorite\_count | retweet\_screen\_name | listed\_count |
| urls\_t.co | quoted\_retweet\_count | retweet\_name | statuses\_count |
| urls\_expanded\_url | quoted\_user\_id | retweet\_followers\_count | favourites\_count |
| media\_url | quoted\_screen\_name | retweet\_friends\_count | account\_created\_at |
| media\_t.co | quoted\_name | retweet\_statuses\_count | verified |
| media\_expanded\_url | quoted\_followers\_count | retweet\_location | profile\_url |
| media\_type | quoted\_friends\_count | retweet\_description | profile\_expanded\_url |
| ext\_media\_url | quoted\_statuses\_count | retweet\_verified | profile\_banner\_url |
| ext\_media\_t.co | quoted\_location | place\_url | profile\_background\_url |
| ext\_media\_expanded\_url | quoted\_description | place\_name | profile\_image\_url |

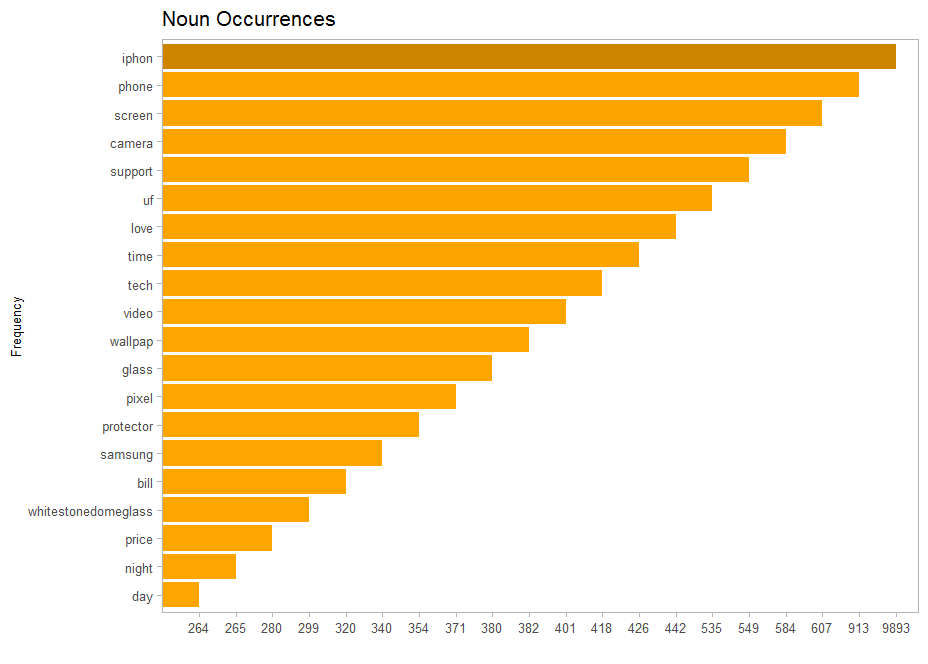
## Parts of Speech Tagging

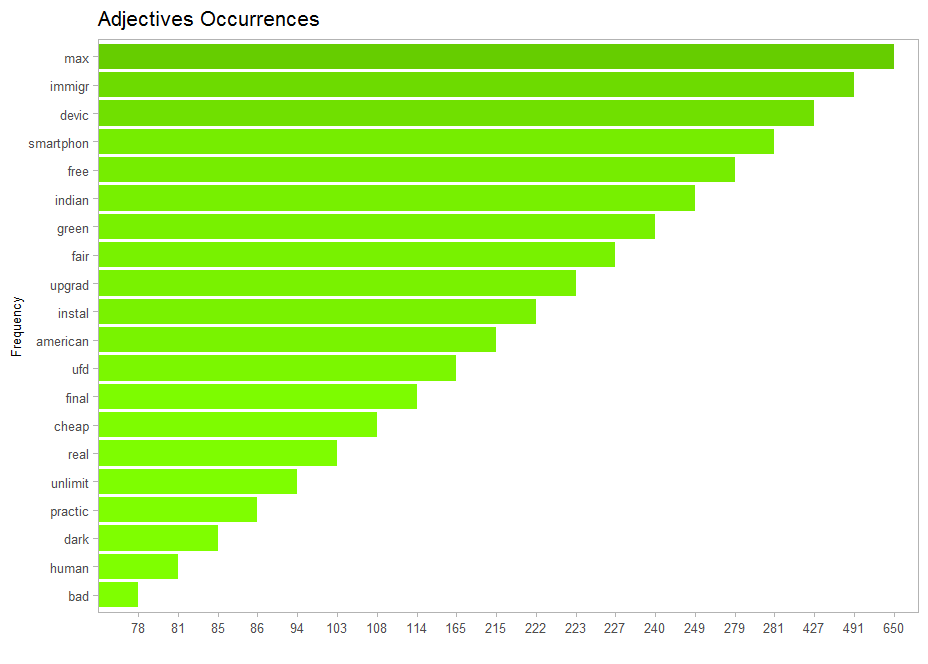
Part of Speech Tagging (POST) is a natural language processing technique in which terms are given additional meaning and context by tagging them to identify terms by their grammatical components.

The bar chat shows the distribution of Parts of Speech of all tweets in the data. It is observed that there are more Nouns, Verbs and Adjectives in all the tweets in the data.



The bar chart shows the top noun words in the tweets. It is obvious customers talk about iPhone which is the top frequent noun. Screen, Camera and Support are other topics customers expressing their opinion about.





The above chart shows the top adjective words in the tweets. Top 3 adjectives observed are about iPhone max, device and smartphone.

The bar chart below shows the top verb occurrences in the data. The top 3 of those are customers talking about iPhone giveaway program, protection and iWatch

