# **Customer Support on Twitter**

Social Media has evolved a lot in last 10 years platforms like Facebook, WhatsApp, and Twitter are changing the way of business. There are millions of tweets being tweeted every day. It is a large dataset of tweets and replies to aid innovation in natural language understanding and conversational models, and for study of modern customer support practices and impact.

We are trying to explore the subset of data generated from twitter platform being used for customer service. Which companies are leading on twitter in resolving or at least recognizing the concerns. How are the response time and the sentiment analysis of customers.

**Data Source :** <https://www.kaggle.com/soaxelbrooke/customer-support-on-twitter/data>

The dataset is a CSV, where each row is a tweet. The columns metadata is given below. Every conversation included has at least one request from a consumer and at least one response from a company. Which user IDs are company user IDs can be calculated using the inbound field.

|  |  |
| --- | --- |
| **Column** | **Data type** |
| tweet\_id | numeric |
| author\_id | string |
| inbound | boolean |
| created\_at | string |
| text | string |
| response\_tweet\_id | string |
| in\_response\_to\_tweet\_id | string |

|  |  |  |
| --- | --- | --- |
| File Name | Format | Size |
| twcs.csv | csv | 492 MB |

**Exploration Ideas:**

1. Top and bottom 10 companies by replies on tweets.
2. Compare average time required to respond on tweet by top & bottom 10 companies.
3. Are the responses from companies made by bots or in automated manner?
4. Does the automated responses help in resolving customer’s issue?
5. Identifying the problems associated with automated response.
6. Identifying the issue with bottom companies.

Understanding the data

* Loaded the dataset (twcs.csv) file into HDFS.
* Using PySpark read the file into pyspark.sql.Dataframe named tweets
* Register it as temp table and cache the dataset
* Run basic queries to understand the data

Exploratory Question

Which are the leaders in response on twitter resolving the customer queries

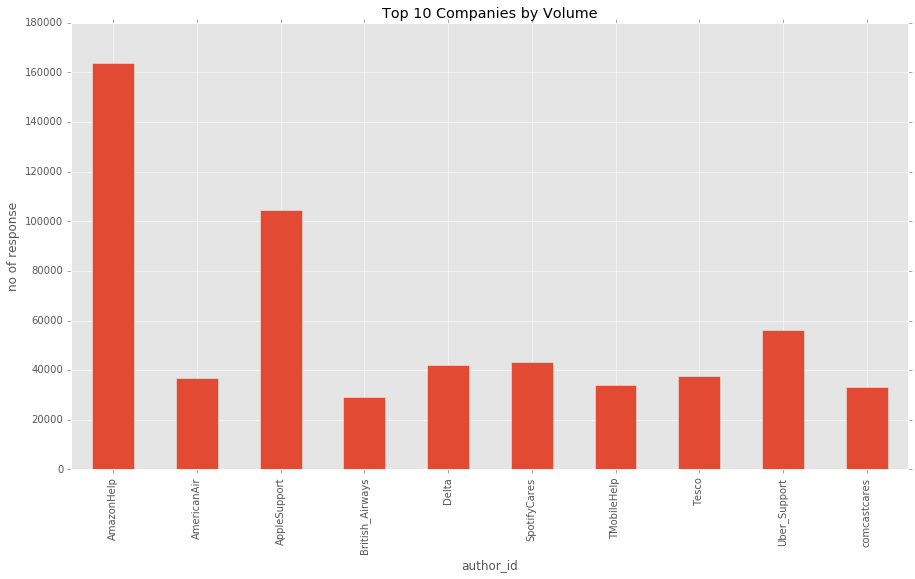
Approach

The following steps were followed to answer the business question:

* Get top 10 organizations by total count of tweets made
* Plot graph using matplotlib

|  |  |
| --- | --- |
| **author\_id** | **no\_of\_response** |
| AmazonHelp | 163753 |
| AppleSupport | 104523 |
| Uber\_Support | 56191 |
| SpotifyCares | 43255 |
| Delta | 42100 |
| Tesco | 37797 |
| AmericanAir | 36748 |
| TMobileHelp | 34023 |
| comcastcares | 32997 |
| British\_Airways | 29311 |

Clearly Amazon and Apple are leading in resolving customer issues on twitter. Amazon is market leader in online retail which makes sense for having lot of customer queries and issues. What’s interesting is Apple leading in here does it mean the apple customer are not that happy with product or services?



Also let’s take a look at bottom 10 originations.

|  |  |
| --- | --- |
| **author\_id** | **no\_of\_response** |
| HotelTonightCX | 152 |
| CarlsJr | 189 |
| AskDSC | 210 |
| OfficeSupport | 218 |
| MTNC\_Care | 242 |
| JackBox | 266 |
| OPPOCareIN | 283 |
| mediatemplehelp | 297 |
| AskRobinhood | 431 |
| AskSeagate | 506 |

Exploratory Question

What was the response time on the query made.

Approach

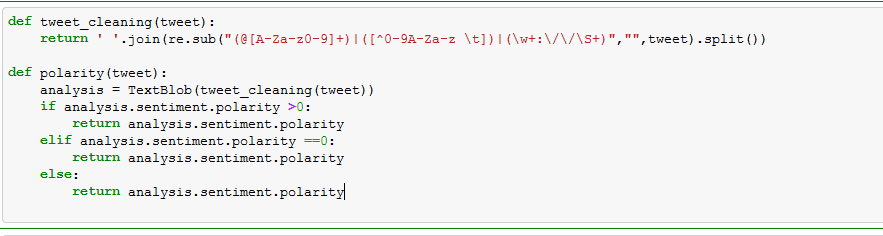
* Get all the tweets made by users towards the organization these will be the inboundTweets pyspark.sql.DataFrame.
* Join this dataframe with original dataset to get the response tweets as well these will be called as originalAndResponseTweets
* Calculate difference between created time of both which would be the response\_duration
* Plot a boxplot for comparing organizations response time and does it impact the sentiment of users.

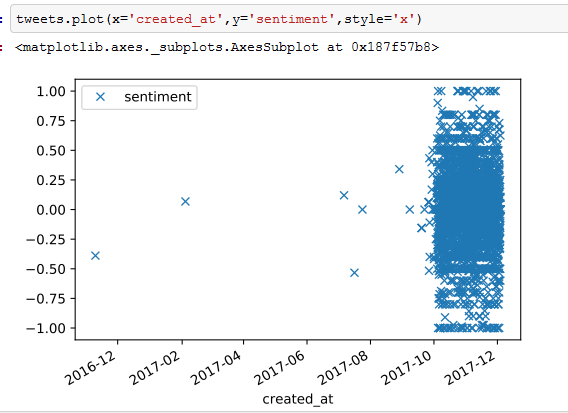
Exploratory Question

Sentiment analysis of the tweets

Approach

* Used SentimentIntensityAnalyzer from from nltk.sentiment.vader package
* We will be using the inbound\_text column from originalAndResponseTweets dataframe to calculate the polarity
* Plotted this polarity by brand
* We will be cleaning up tweets first then calculating the polarity. If polarity is positive then it is considered as positive sentiments and so on.





Exploratory Question

Is there any pattern in tweeting of customers?

Approach

* Using inbound dataframe
* 