# **Project Report for IAL621: Content Analysis for Social Network Data**

**Extract & Convert Social Network (Twitter) content for sentiment analysis**

Bala Mallampati Thursday, December 9, 2021

Master’s in informatics and Analytics

University of North Carolina, Greensboro, USA

[b\_mallampat@uncg.edu](mailto:b_mallampat@uncg.edu)

**Abstract:**

I understand the Social Network (Twitter, Facebook) and analyzed using NodeXL software. And then extracted twitter text information and did sentiment analysis of topics/tweets by following all privacy, regulatory & access policies.

In this project, I answered the following data research questions:

1. Twitter network analysis to identify their networks as part of exploratory analytics.
2. Extract information for UNCG top tweets & topics from Twitter site and convert to numbers as part of Data wrangling.
3. Classify UNCG tweets/hashtags using Machine Learning Models(Linear, KNN, Decision Tree, Random Forest..)
4. Do sentiment analysis of UNCG Tweets/Topics and present extracted insights to corresponding organization.

**Introduction and Motivation:**

Now a days everyone using social media for News, follow Trends and share their feelings and get in touch with their friends & family/Employers/Learning. Social media can create impacts on every organization if management is not up to date on latest trends/news.

Social media include all the ways people connect to people through computation. Mobile devices, social networks, email, texting, micro-blogging and location sharing are just a few of the many ways people engage in computer-mediated collective action. As people link, like, follow, friend, reply, retweet, comment, tag, rate, review, edit, update, and text one another (among other channels) they form collections of connections. These collections contain network structures that can be extracted, analyzed, and visualized. The result can be insights into the structure, size, and key positions in these networks.

I love to play with data, now “The world’s most valuable resource is no longer oil, but data” as per The Economist article (<https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>).

If we can understand how these social network works and can read & extract useful information, then we can predict how it will impact our organization by doing sentiment analysis on Tweets/topics.

In this project, I used MassMine framework for twitter data extraction and analysis to predict impacts on organization and take preventive actions.

**Data Source and description:**

Tweets are the basic atomic building block of all things Twitter. Tweets are also known as “status updates.” The Tweet object has a long list of ‘root-level’ attributes, including fundamental attributes such as id, created\_at, and text. Tweet objects are also the ‘parent’ object to several child objects. Tweet child objects include user, entities, and extended\_entities. Tweets that are geo-tagged will have a place child object.

Data extracted from below social media site. I had extracted 50K UNCG tweets from past one week data using Massmine and summarized to 2165 tweets/topics with below attributes.

<https://twitter.com/>

<https://developer.twitter.com/en/portal/dashboard>

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Alias in Project** | **Type** | **Description** |
| created\_at | tweet.datetime | String | UTC time when this Tweet was created. Example:"created\_at": "Wed Oct 10 20:19:24 +0000 2018" |
| text | text | String | The actual UTF-8 text of the status update. See twitter-text for details on what characters are currently considered valid. Example:"text":"To make room for more expression, we will now count all emojis as equal—including those with gender‍‍‍ ‍‍and skin t… https://t.co/MkGjXf9aXm" |
| user.description | user.description | String | User Object: "description": "The Real Twitter API. Tweets about API changes, service issues and our Developer Platform. Don't get an answer? It's on my website.", |
| user.favourites\_count | total.favorites | Int64 | User Object: "favourites\_count": 31, |
| user.followers\_count | total.followers | Int64 | User Object: "followers\_count": 6129794, |
| user.friends\_count | total.friends | Int64 | User Object: "friends\_count": 12, |
| user.location | user.location | String | User Object: location: "San Francisco, CA", |
| user.name | user.name | String | User Object: "name": "Twitter API", |
| user.screen\_name | screen.name | String | User Object: "screen\_name": "Twitter API", |
| user.statuses\_count | total.statuses | Int64 | User Object: "statuses\_count": 3658, |
| entities.user\_mentions.0.screen\_name | tweet.mentions | String | Entities:"user\_mentions":[], |
| retweeted\_status.user.screen\_name | retweet.mentions | String | Retweets can be distinguished from typical Tweets by the existence of a retweeted\_status attribute. This attribute contains a representation of the original Tweet that was retweeted. Note that retweets of retweets do not show representations of the intermediary retweet, but only the original Tweet. (Users can also unretweet a retweet they created by deleting their retweet.) |
| entities.urls.0.expanded\_url | url | String | Entities: "urls":[], |
| retweeted\_status.user.entities.url.urls.0.expanded\_url | retweet.url | String | Retweets can be distinguished from typical Tweets by the existence of a retweeted\_status attribute. This attribute contains a representation of the original Tweet that was retweeted. Note that retweets of retweets do not show representations of the intermediary retweet, but only the original Tweet. (Users can also unretweet a retweet they created by deleting their retweet.) |
| retweet\_count | total.retweets | Int64 | Number of times this Tweet has been retweeted. Example:"retweet\_count":160 |

**Methodology:**

Data processing is an important step for in the data analysis. Data science involves methods of analyzing massive amounts of data for the purposes of knowledge extraction. It evolved from statistics and traditional data management. Data comes in many shapes and forms, and many times we need to get it ready to be able to analyze it. The phrase “garbage-in and garbage-out” is particularly applicable to text mining to Train and Test Data.

In this project I have used the following languages, frameworks, tools, libraries, packages from Twitter data extraction to Sentiment analysis and Tweets data classification/regression.

*Technologies & Libraries*: R(ndjson, tidyverse, tm, lubridate, fliptime, stringr,ggmap, maps, quantenda, readtxt, textplots, widyr), Python(numpy, pandas, matplotlib, seaborn, sklearn, google.colab, )

*Framework/Softwares*: Massmine, NodeXL, Oracle VM virtual box 6.1.26, Ubuntu 20.04

*Tools*: Rstudio, Colab for python, Power BI, Tableau

*Methods*: R Selenium, Machine Learning(Linear, Logistic, SVM, RandomForest, KNN, SVM)

**Data Analysis:**

Social media networks form in Twitter around a wide range of terms. People talk about the news of the day, celebrities, companies, technology, entertainment, and more. As each person uses Twitter they form networks as they follow, reply and mention one another. These connections are visible in the text of each tweet or by requesting lists of the users that follow the author of each tweet from Twitter.

**Logo, company name

Description automatically generated with medium confidence**

I came to learn below NodeXL software to analyze Social Networks by input below values. NodeXL enables the automatic execution of a five-step data workflow that starts with data collection from a variety of network data sources, through storage, analysis, visualization and finally publication.

<https://nodexl.com/guides/how-to-install-nodexl-pro>

**A screenshot of a computer

Description automatically generated with medium confidence**

[**http://nodexlgraphgallery.org/Pages/Default.aspx**](http://nodexlgraphgallery.org/Pages/Default.aspx) **–** It has different network analysis graphs to analyze data with graphs(ex: Covid19, Twitter networks, Climate changes..etc)

**Calendar

Description automatically generated**

**Graphical user interface, diagram, application

Description automatically generated**

<https://nodexlgraphgallery.org/Pages/InteractiveGraph.aspx?graphID=267938>

**Chart

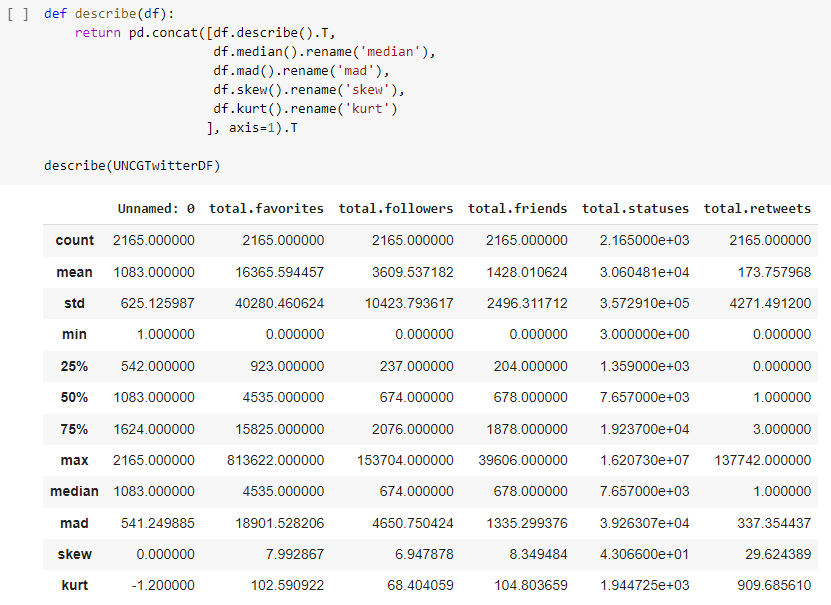
Description automatically generated**

***Exploratory Analysis:*** Below Python code shows 15 features and 2165 observations.It has 5 metrics with integer data type and rest of them are object type(String).

**Table

Description automatically generated**

Following are the stats about twitter data set metrics with min, max, std, count, mean, mad, skew and kurt.

****

I also ran correlation plot between Total Retweets with other features. Its highly correlated with Statuses, friends, followers.

**Chart, treemap chart

Description automatically generated**

Below R Plot shows trend of UNCG Tweets for past one week. Dec 1st 2021 shows high tweets than other dates.

Chart, line chart

Description automatically generated

I also verified all metrics comparison using tableau tool, below shows comparison between UNCG Total Tweets with Status, Friends Followers & favorites.

A picture containing chart

Description automatically generated**s**

***Day wise comparison:*** Below bar graph shows Tweets per day for past one-week (Nov/Dec 2021) UNCG twitter data. I see Dec 5th has many tweets posted, it might be due to System generated and outliers.

Chart

Description automatically generated

***Hourly comparison***: At 5AM ET, it shows high tweets but strongly suspects its outlier. Next most tweets happened between 7PM to 11PM ET.

Chart, line chart

Description automatically generated

***Tweets Mentions vs count Comparison***: SpiderVerse & Dernswift has highest tweet counts than others.

Chart, bar chart

Description automatically generated

***Tweets By User Comparison*:** Below Tableau reports shows Nabil & Memphis, UNCG Athletics tweeted more than other users.

**Chart, bar chart

Description automatically generated**

Similar Tweets by user in different view using Power BI radar view, Jonathan & UNCG are seems outliers.

**Chart, radar chart

Description automatically generated**

**Timeline

Description automatically generated**

**Results:**

I have included parameters based on above exploratory analysis, Below R graph shows co-occurrence relationships for mentioned usernames, it’s using quanteda R package text plot.It shows as Decentralized(B) UNCG network.

Diagram

Description automatically generated

**Machine Learning:**

I have taken Randomforest Classifier Importance feature extraction for UNCG Tweets,, it shows that Mentions, text, URL, Date & Time are highest importance features to predict Tweets count.

**Chart, bar chart

Description automatically generated**

One of my research questions related to Tweets classification, which I ran with below Machine Learning classifiers, which shows 40% accuracy, this might be due to text data.

**Graphical user interface, application, Word

Description automatically generated**

And executed Linear regression with correlated & highly statistically significant variables to predict tweet count. Below model shows RMSE=9.42, MSE=88.65, MAE=7.87.

**Graphical user interface, text, application, email

Description automatically generated**

*Linear Regression Formula for Tweets prediction:*

UNCG Tweet(Y|X)= 18.17-0.085Mentions+0.004Text-0.0005URL

The following graph shows Prediction vs Actual data comparison accuracy, both are linearly correlated.

**Chart, scatter chart

Description automatically generated**

Below graph shows Test vs predicted values by Linear Regression model.

**Chart, histogram

Description automatically generated**

**Sentiment Analysis Results:**

Below R plot shows correlation comparison between each word, example: asu highly correlated to tweet contains ssu, uwg, science words. Similarly Pitch with Dinner, Item with Auction..etc

Graphical user interface, diagram, table

Description automatically generated

I ran sentiment analysis with UNCG twitter data with R tdm, tm packages. It shows below positive vs negative comparison. Most of words related to soccer game.I observe UNCG tweets are more positiveness than negativity in Tweets.

Chart

Description automatically generated

I ran below Emotion type for hashtags #UNCG tweets using R NRC data dictionary words. It shows Positive with 501 count, anticipation & trust ~200 words. This shows that UNCG tweets are clearer & more perfect.

**Chart, bar chart

Description automatically generated**

With word cloud package in R, I have executed UNCG tweets word comparison with max 500 words. Below results shows that UNCG highly uses in tweets based on size of word.

**A map of a city

Description automatically generated with medium confidence**

**Conclusion:**

Github URL for code: <https://github.com/BalaMallampatiGIT/IAL621_TwitterProject>

I learnt very new subject social media network analysis using NodeXL & how we can utilize twitters tweets & connections contents data for organization (UNCG) growth prediction.

Massmine made my life very easy for twitter data extraction with simple commands, thanks to Dr Aaron.

UNCG tweets have more positivity and most of them related to games (Baseball, Soccer) & ssu(science, asu), uncg.

Though my models ultimately did not result in an effective way to classify and predict UNCG tweets data, the improvements I made to our ML model over time did significantly reduce the MSE loss. Given more time and resources to do further research and feature engineering, I believe I could retrain this model to better success.

I believe that I did not have enough of the contributing features that drive classification of UNCG Twitter dataset & not enough data. Additional data points year over year Tweet growths, seasonal trends of tweets, other modern bigdata techniques may have improved these model results.

**References:**

* Sklearn, Python library
* <https://www.massmine.org/docs/twitter.html>
* <https://developer.twitter.com/en/portal>
* <https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet>
* <https://towardsdatascience.com/3-super-simple-projects-to-learn-natural-language-processing-using-python-8ef74c757cd9>
* <https://nodexl.com/guides/how-to-install-nodexl-pro>
* <http://nodexlgraphgallery.org/Pages/Default.aspx>
* How to open NodeXL file : <https://alice.endicott.edu/examples/9>