|  |  |
| --- | --- |
| Sentiment Analysis and Visualization  ISMG 6470 H01 | ABSTRACT  Do women now have a general lack of interest in Science, Technology, Engineering and Mathematics (STEM) positions Or are we in for a BIG CHANGE? This report summarizes the analysis and findings of the twitter data regarding the presence of women in science and technology.  Neena Shereen  Final Assignment: OPTION 3 |

# ABSTRACT

This report summarizes the analysis and findings of the twitter data regarding the presence of women in science and technology. The data was scrapped from Twitter and cleaned in JMP and Excel. The data analysis was primarily done in RStudio using text data analysis packages like rtweet, qdap, tokenizer, tm etc. RStudio was extensively used for generating visualizations that summarizes the trend and pattern of the data.

Table of Contents

[ABSTRACT 1](#_heading=h.gjdgxs)

[ABOUT THE PROJECT 4](#_heading=h.30j0zll)

[OBJECTIVES OF THE PROJECT 4](#_heading=h.1fob9te)

[METHODOLOGY 4](#_heading=h.3znysh7)

[DATA FOR PROJECT 4](#_heading=h.2et92p0)

[Data cleaning and Recoding. 5](#_heading=h.tyjcwt)

[ANALYSIS 5](#_heading=h.3dy6vkm)

[Analysis Part 1: Basic Structure of The Data 5](#_heading=h.1t3h5sf)

[Analysis Part 2: Sentiment Polarity Analysis of Tweets 6](#_heading=h.4d34og8)

[Analysis Part 3: Sentiment Polarity Analysis of Retweet Description 6](#_heading=h.2s8eyo1)

[Analysis Part 4: Sentiment Word Cloud of the Tweets 6](#_heading=h.17dp8vu)

[Analysis Part 5: Emoji/Emoticon Analysis of the Tweets 6](#_heading=h.3rdcrjn)

[Analysis Part 6: Tracing the Tweet Flow 7](#_heading=h.26in1rg)

[Analysis Part 7: Who retweets whom: The Spread of tweets of prominent contributor 7](#_heading=h.lnxbz9)

[Analysis Part 8: Geospatial Visualization 7](#_heading=h.35nkun2)

[INSIGHTS FROM THE DATA 7](#_heading=h.1ksv4uv)

[APPENDIX 8](#_heading=h.44sinio)

[Dashboard 8](#_heading=h.2jxsxqh)

[Analysis Part 1: Basic Structure of The Data 8](#_heading=h.z337ya)

[Analysis Part 2: Sentiment Polarity Analysis of Tweets 9](#_heading=h.2xcytpi)

[Analysis Part 3: Sentiment Polarity Analysis of Retweet Description 10](#_heading=h.2bn6wsx)

[Analysis Part 4: Sentiment Word Cloud of the Tweets 10](#_heading=h.3as4poj)

[Analysis Part 5: Emoji/Emoticon Analysis of the Tweets 11](#_heading=h.49x2ik5)

[Analysis Part 6: Tracing the Tweet Flow 11](#_heading=h.3o7alnk)

[Analysis Part 7: Who retweets whom: The Spread of tweets of prominent contributor 12](#_heading=h.ihv636)

[Analysis Part 8: Geospatial Visualization 12](#_heading=h.32hioqz)

[R-packages used 13](#_heading=h.2grqrue)

[REFERENCE 13](#_heading=h.vx1227)

[Figure 1 No: of quotes in the given sample ,Figure 2 Proportion of retweets among the tweets 8](#_heading=h.3j2qqm3)

[Figure 3 Major source of the tweets 8](#_heading=h.1y810tw)

[Figure 4 Top contributors 9](#_heading=h.4i7ojhp)

[Figure 5 Polarity analysis of Tweets 9](#_heading=h.1ci93xb)

[Figure 6 Frequently appearing words in Tweets 9](#_heading=h.3whwml4)

[Figure 7 Polarity analysis of Retweets 10](#_heading=h.qsh70q)

[Figure 8 Sentiment word cloud 10](#_heading=h.1pxezwc)

[Figure 9 #womeninstem emojis , Figure 10 Common emojis used 11](#_heading=h.2p2csry)

[Figure 11 Common Emoticons used 11](#_heading=h.147n2zr)

[Figure 12 Popularity plot , Figure 13 Circulation Plot 11](#_heading=h.23ckvvd)

[Figure 14 User Locations 12](#_heading=h.1hmsyys)

[Figure 15 No of tweets per region 13](#_heading=h.41mghml)

# ABOUT THE PROJECT

Twitter has identified 6 fastest growing topics, *Well-being, Everyday wonders, Creator culture, Identity, Planet, Techlife,* that reflect the fundamental worldwide cultural shift. This helps brands to be more culturally relevant, green, and lean. Of all the themes, “Identity”, as per the twitter data, showed around 305% increase in conversations relating to gender equality, LGBQT, disability especially in the field of sports and technology, in the past 5 years. Twitter has released the popular hashtags in this theme: #WomenInBiz #LGBTQ #LGBT #WomenInStem #WomenInTech #GirlsWhoCode #film #Christian #PrideMonth #movies #Hollywood #misogynoir #TV #football #racism #reading #bookclub.

Intrigued by this insight, the study proposes to analyze the tweets with hashtag #WomenInStem #WomenInTech #GirlsWhoCode. Data Science is one of the trendiest jobs of this generation and women have already made their mark in this field. This study proposes to analyze the involvement of women in this field, through their presence in social media.

# OBJECTIVES OF THE PROJECT

The following are the major objectives of the study:

* Capture the overall sentiments expressed in tweets with hashtag: #WomenInBiz, #WomenInStem, #WomenInTech, #GirlsWhoCode.
* Do the emojis used in tweets corresponds with the general sentiment of the tweets?
* Are these tweets actively retweeted and shared to a wider group?
* Trace the origin of the tweeters under this topic

# METHODOLOGY

For the analysis, the Twitter data was scraped using rtweet package, using the API code and relevant keys. The raw data was saved in Excel. Small level data cleaning and recoding was done in JMP. The detailed data cleaning and analysis was done in Rstudio. Sentiment analysis was performed using Qdap Polarity analysis, emojis were categorized using data wrangling packages like dplyr, geo visualizations were used to show the origin of the tweets.

# DATA FOR PROJECT

The tweets with hashtag "#WomenInStem", "#WomenInTech","#GirlsWhoCode were parsed from the Twitter using retweet package and saved in the file “data4project.xslx”. ". The raw data has 8914 rows and 91 columns. From the raw data, columns that have relevant information are selected for the analysis, and subsequently saved as “sample data” in the same excel file.

## Data cleaning and Recoding.

Majority of the data cleaning was done in RStudio. However minor levels of cleaning and recording was done in Excel and JMP. The Raw data had tweets in different languages, only those in English was selected for analysis. Further the Raw data had information about location and Retweet location, signifying the location of the users. These columns are recoded in JMP in the format: Location, Country. In Excel they are split as Tweet.Location, Tweet.Country, Retweet.Location, Retweet.Country.

The following table summarizes the structure of the data

|  |  |
| --- | --- |
| File name | data4project.xlsx sheet: “sampledata” |
| Data dimensions | 8409,26 |
| **Column information** | |
| **user\_id, screen\_name, retweet\_user\_id, retweet\_screen\_name** | Information about the user |
| **created\_at, retweet\_created\_at** | Time stamp of tweet/retweet creation |
| **text** | The tweet |
| **Source, retweet\_source** | Whether the tweet was made from app or web |
| **is\_quote, is\_retweet** | Whether the given tweet is a quote or retweet |
| **favorite\_count** | No of times the tweet marked as favorite |
| **retweet\_count** | No of times the tweet is retweeted |
| **followers\_count,friends\_count, favourites\_count** | Friends and followers for the user who made the tweet. |
| **retweet\_text** | Usually same as the text |
| **retweet\_retweet\_count** | No of times the retweet is retweeted |
| **retweet\_favorite\_count, retweet\_followers\_count. retweet\_friends\_count** | Friends and followers for the user who made the retweet. |
| **retweet\_description** | Description on the Retweet |
| **Tweet.location, Tweet.Country** | Location of the tweeter |
| **retweet.location, retweet.Country** | Location of the retweeter |

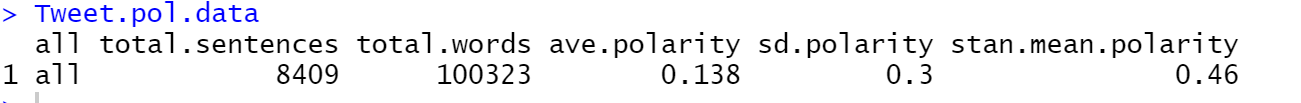
# ANALYSIS

## Analysis Part 1: Basic Structure of The Data

This section of analysis is devoted to examining the structure of the data. The visualizations in analysis helped to know whether the given tweets are retweet or not, who is the top contributor etc. This section helped to finalize the direction of further analysis.

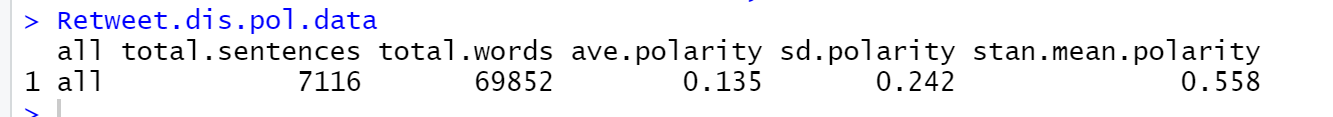
## Analysis Part 2: Sentiment Polarity Analysis of Tweets

In this section sentiment polarity of the tweets are examined using Polarity function of qdap package. The tweets were cleaned, stripped of the emoticons, and standardized before the analysis. The analysis shows a generally positive sentiment in the tweets. Even though the general sentiment is positive, some tweets have 0 polarity value: especially ones that are announcements. For visualization purpose such neutral tweets were excluded. In addition to this the study revealed that most common word used in these texts is “women”.



## Analysis Part 3: Sentiment Polarity Analysis of Retweet Description

The same type of analysis was performed on Retweet description. Though appear to be similar, the mean polarity of the retweets is slightly higher than tweets.



## Analysis Part 4: Sentiment Word Cloud of the Tweets

Polarity analysis revealed that there are 3 kinds of words present in the tweets: Positive words, Negative words, and neutral words. So, the next step was to create a word cloud depicting these categories of words. This step was largely and extension of previous analysis, so the same data frame was used. The tweets are categorized based on their polarity values and plotted as a comparison word cloud.

## Analysis Part 5: Emoji/Emoticon Analysis of the Tweets

This section was devoted to find out the common emojis and emoticon used in the tweets, and whether they really reflect the sentiment of the tweets. For this emoji analysis 2 datasets were used: “emojis” and “emDict”. The former shows a variety of emojis like man, man scientist, woman, girl etc. whereas the latter is crisper and more general. So “emDict” used to analyze the general range of emojis used, where as “emojis” dataset is used to find out the frequency of women specific emojis used. In the same way the built in “emoticon” dataset was used to find out the emoticons used in the tweets.

## Analysis Part 6: Tracing the Tweet Flow

This analysis concentrated on examining the popularity and spread of the tweets. For this purpose, two additional variables were created:

tweetpopularity=favorite\_count+(0.67\*retweet\_count) +(1.25\*retweet\_favorite\_count) +(1.74\*retweet\_retweet\_count),

tweetspread=scale((mean(followers\_count+friends\_count+favourites\_count)) + 2.5\*((retweet\_followers\_count+retweet\_favorite\_count+retweet\_friends\_count)/3)))

The first variable gauges the popularity of the tweet and the second gauges the rate at which the tweet is circulated. This analysis helped to find out the users with wider circulation and popular tweets.

## Analysis Part 7: Who retweets whom: The Spread of tweets of prominent contributor

This is a continuation of Part 6. The analysis showed the top contributor and the most popular one is the user with screen name “ femtech\_”. The objective of this analysis is to trace how the tweets from this user is shared and circulated. The graph was built using igraph package. Longer the edge, grater the circulation.

## Analysis Part 8: Geospatial Visualization

The geo spatial visualizations were used to find out the top location of the active users. Also, it was used to find out no of tweets generated in each country.

# INSIGHTS FROM THE DATA

* The overall sentiment of the tweets is positive.
* Twitter is becoming a platform for learning and inspiring especially for women in technology.
* Twitter offers an efficient way of marketing skill development courses and workshops.
* Emojis/emoticons largely correlates the sentiment of the tweets.
* Popular users have the propensity for larger spread.
* Majority of the tweets are coming from US and UK, but that does not mean other countries are lagging.
* Radical shift in the outlook: “programming was for boys”
* Throughout the world, women are seriously considering Data science /technology related jobs as a career option.

# APPENDIX

## Dashboard

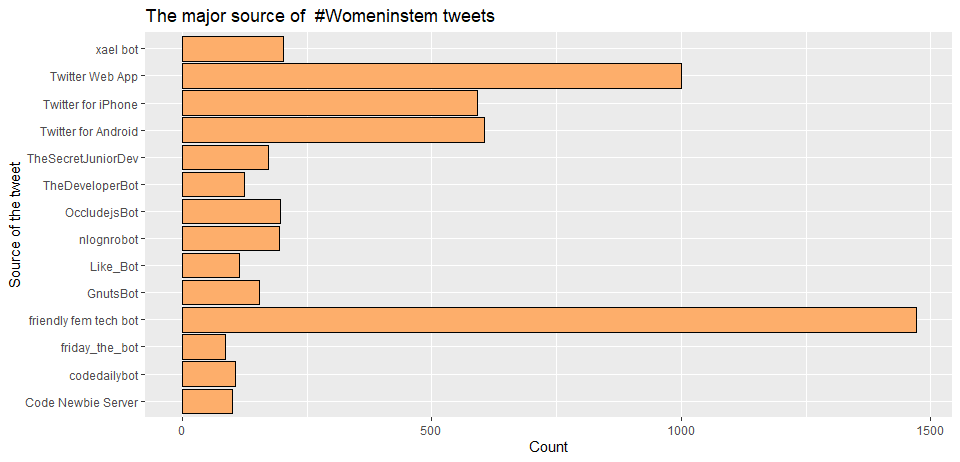
## Analysis Part 1: Basic Structure of The Data

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

*Figure 1 No: of quotes in the given sample Figure 2 Proportion of retweets among the tweets*



*Figure 3 Major source of the tweets*

Chart, bar chart

Description automatically generated

*Figure 4 Top contributors*

## Analysis Part 2: Sentiment Polarity Analysis of Tweets

Chart, histogram

Description automatically generated

*Figure 5 Polarity analysis of Tweets*



A picture containing timeline

Description automatically generated

Figure 6 Frequently appearing words in Tweets

## Analysis Part 3: Sentiment Polarity Analysis of Retweet Description

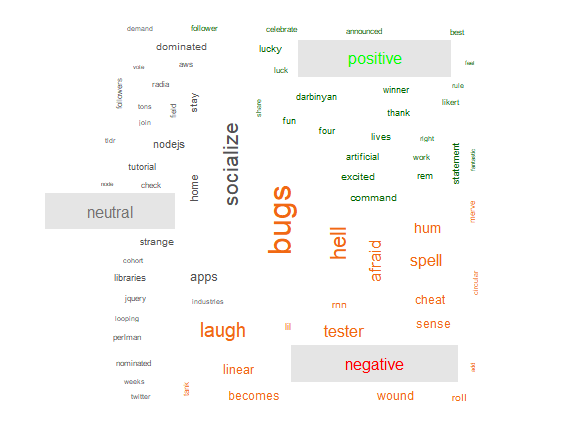
Chart, histogram

Description automatically generated

*Figure 7 Polarity analysis of Retweets*

## Analysis Part 4: Sentiment Word Cloud of the Tweets





*Figure 8 Sentiment word cloud*

## Analysis Part 5: Emoji/Emoticon Analysis of the Tweets

A picture containing chart

Description automatically generatedChart

Description automatically generated

*Figure 9 #womeninstem emojis Figure 10 Common emojis used*

Chart, bar chart

Description automatically generated

*Figure 11 Common Emoticons used*

## Analysis Part 6: Tracing the Tweet Flow

Graphical user interface, application, table

Description automatically generatedChart, background pattern

Description automatically generated

*Figure 12 Popularity plot Figure 13 Circulation Plot*

## Analysis Part 7: Who retweets whom: The Spread of tweets of prominent contributor

A picture containing chart

Description automatically generated

## Analysis Part 8: Geospatial Visualization

Map

Description automatically generated

*Figure 14 User Locations*

Map

Description automatically generated

*Figure 15 No of tweets per region*

## R-packages used

library("dplyr");library("rtweet");library("maps");library(RColorBrewer);library(tidyverse);library(ggrepel;library(openxlsx);library(plotrix);library(stringr);library(readr);library(wordcloud);library(readxl);library(qdap);library(tokenizers);library(tm);library(stopwords);library(RWeka);library(SnowballC);library(reshape2);library(igraph);library(ggthemes);library(rematch2);library(openxlsx);library(Unicode)

# REFERENCE

<https://www.searchenginejournal.com/twitter-highlights-18-of-the-hottest-trending-topics-going-into-2020/337625/#close>

[Twitter Developers](https://developer.twitter.com/en/portal/projects-and-apps): Twitter account created for data collection.

[emojis/emoji\_analysis.R at master · today-is-a-good-day/emojis · GitHub](https://github.com/today-is-a-good-day/emojis/blob/master/emoji_analysis.R)