Student Name(s): **Samiul Islam**

Student Number(s): **500602494**

Course Title: **Social Media Analytics**

I hereby certify that I am the author of this document and all sources used in the preparation of this assignment have been cited in accordance with Ryerson's Code of Student Conduct directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by me (all group members if it is a group paper) for this purpose.

Patients' Sentiment Analysis Using Twitter Data Regarding the Health Services of Canada.

Table of Contents

Summary:

The goal of this project is to identify peoples' sentiment regarding 'Canada Health Care System' and the services they received using social media data. To do so, we have considered patients' or users' sentiments by analyzing their tweets. The main part of this project is 'text analysis' to reveal the patients' sentiments. Tweet data are collected from the selected sources for a period of ten days. Meta- data analysis has been performed on the collected data to understand the data pattern and to determine necessary steps those could be require in data preprocessing stage. The tweets of 12249 unique users are analyzed in the text analysis phase. Sentiments are extracted from those tweets and the "NRC" method is adopted to carry out this task. The value of count of those sentiments are grouped in ten categories. The final outcome of this sentiment scores depict that the ratio of negative and positive sentiments is approximately 1:1.38. Though people are talking more about their fear than the joy. This study also reveals that the patients' trust on the health care system overpowered their sadness, anger and disgust.

Problem statement:

In July 2017, a comparative survey report [1] was published by the Commonwealth Fund about the health care system performance among Australia, Canada, France, Germany, the Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, and the United States. That report considered only the top 11 high income countries of the world. The report reflected that Canada Health Care System ranked 9 out of 11. The ranking process was instituted by analyzing seventy-two indicators which were grouped in five domains: Care Process, Access, Administrative Efficiency, Equity, and Health Care Outcomes.

Figure 1: Country wise performance ranking of healthcare system

That result reflected the whole health care services performance of those countries, and not an impressive result for the Canadian health care system. In this report, we tried to measure what the Canadian people actually thinks about the health services they receive. To do so, we have fetched twitter data to observe what the people are talking about those services.

Dataset selection:

The dataset used in this work has been fetched from different sources those can be broadly categorized as government sources - who provides those healthcare services, public sources - who use those services and different organizations - who measure those services.

Literature review:

In 2014, review research was conducted by Francisco et al. [2] to present case studies that clarify "how social media being used in the medical and healthcare sectors". To conduct this research, they have used both traditional and nontraditional method. Where techniques included but not limited to peer review, case-studies, guidelines, and content of social media. In the same year, Coulter et al. [3] surveyed the patients' experiences regarding the health services they received to identify the level quality standards of those service providers. All of these results ending with an outcome that the information in social media has a contribution in healthcare quality improvement.

In 2015, Lin et al. [4] conducted research, which aimed to grade the quality of blogs and podcasts with the help of international instructors of health professionals. They used the data from various sources, such as patients' online blog, hospitals dashboards and complaints sections, etc. Based on these, they have categorized the quality standards in 13 domains and set multiple quality indicators for each domain. The researchers suggested that there is no pre-prescribed threshold point for each indicator to meet the minimum quality, rather it will depend on the organizations policy, stakeholders' requirements, and company's goal. To measure and improve the quality of cardiac care in Ontario, CA research was conducted by Tu et. al. [5] in 2015. The researcher put their effort to find out a way to boost up the primary and secondary anticipation of cardiac events in Ontario province which is well known for its multi-cultural population. In this work the researchers' used the multiple data sources from online blogs and customized websites developed to fetch the opinions of the patients. The researchers certainly did not carry out the evaluation of the service provided to the patients, instead they focused on the opinion of the patients regarding the effect of the medications those were provided to them. "The possibility of value-based health care for driving improvement and outcomes measurement must accelerate after conducting a critical research" was mentioned by Micheal et. al. [6] in 2015. The researchers' used social media platform as their main data sources. Based on their data, the researchers proposed a model where the services were categorized in to 5 domains and each domain composed of certain quality standards. Each standard was measured by different key indicators of those standards.

In 2017, Tursunbayeva et al. [7] investigated twitter data of last five years in medical journals and found that most companies used social media for "transparency/accountability, democratic participation, co-production, and evaluation". This well-structured review aimed to collect and present the evidence that the influence of social media regarding public health sector could help the government to understand the public sentiment on health services. The researchers identified that social media data reflect good picture of service standards and the government could take necessary steps to improve service quality in the sectors those were lagging in quality measurement.

Data collection:

Data have been fetched from several twitter accounts to review the opinion from different angels. Some are government sources (e.g. @GovCanHealth, #cdnhealth) where we can find the offerings of services, changes made by the government, and what people commented regarding those services and changes. Few are public sources (e.g. @PatientsCanada, @Patient_Safety) from where we can accumulate the statements of patients who are directly related to those services. Some of the sources are from different welfare organizations (e.g. @HQOntario, #ptsafety) who mainly concerned about the quality of different health services. We have also fetched sample data from the employee perspective (e..g @OntariosDoctors) regarding the services they are providing.

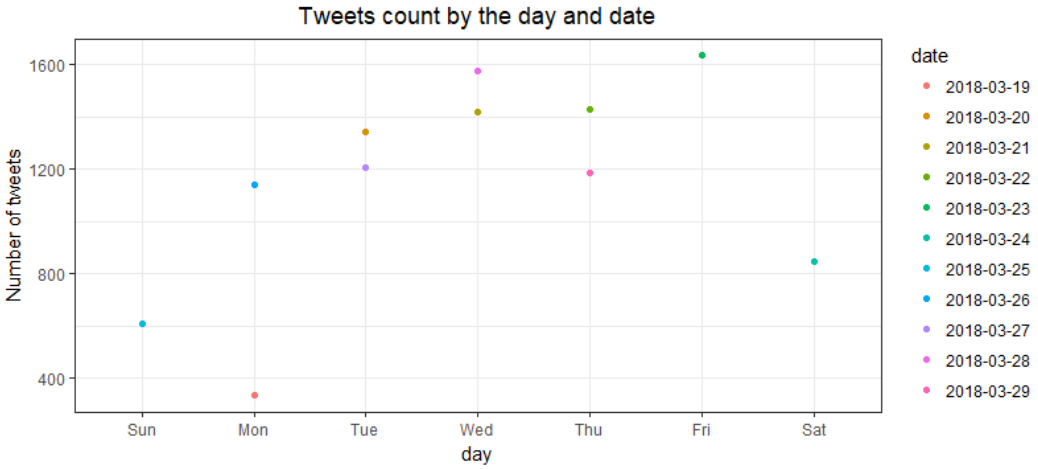


Figure 2: Distribution of the collection of data samples

Data preprocessing:

Collected data are converted in to data frame and are combined in a single .csv file, named as "CDN_health_tweet_data.csv". Tweets from total 12249 unique account has been considered for this analysis. The original data file has been reloaded to the file named ‘health_tweet’ as this data file need to be molded and twisted as required during the analysis phases. As example, to figure out the timing of the tweets data, data has been converted to the class “POSIXlt”. An extensive data cleaning is carried out in text analysis phase. A text scrubber has been declared and implemented on the dataset to remove url, punctuation, numbers, html-links and unnecessary spaces. Manual data processing is conducted to remove extra noise form the data file such as removing words less than 4 characters and grater than 11 characters form the tweets. To removing all "stopwords"(words that do not add meaning to the topic) and convert the text into a Term Document Matrix, a TDM function is used. Extra words are added in the TDM using ‘tmap’ function those need to be removed from the tweets, and this process carried out by multiple observations.

Analysis:

Meta data Analysis:

In this part we first find out the unique users in our data by looking up their IDs and total 12249 unique user have been found. Next, we would like to figure out the days on which the users tweeted most. The reason behind is to observe any abnormal tweeting tendency normally caused by 'tweet-bots'.

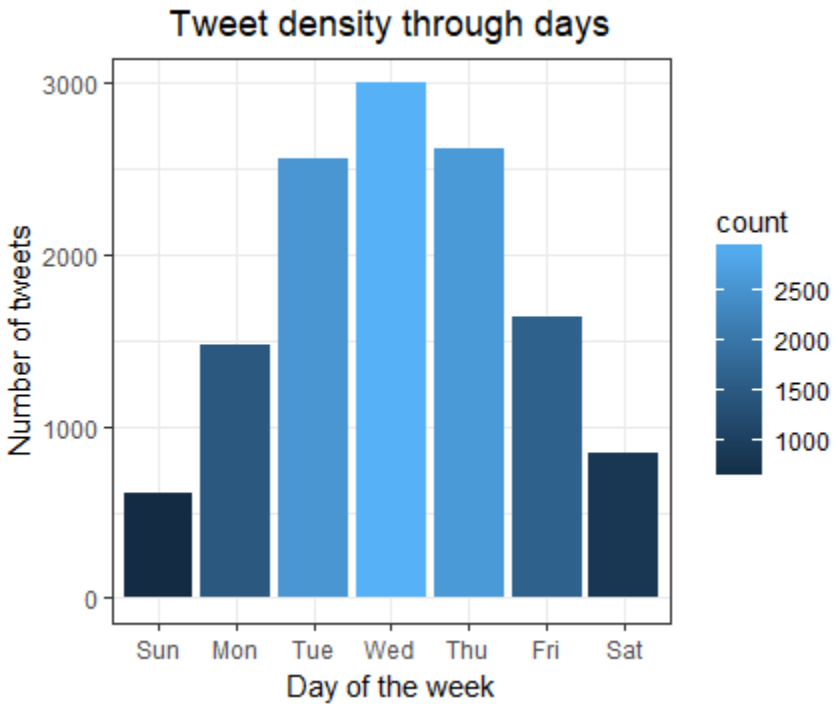


Figure 3: Users' tweeting tendency

Figure 3 depicts that tweets are lees occurred in weekends which is normal as people are less tends to visit the doctors in weekends if not emergency or maybe they do not want to spoil their day offs.

When we observed the density of tweeting time of our data in figure 4, we have spotted some unusual behavior. Some tweets are occurred during midnight to 4 in the morning. We have observed manually our data set and found some of the government account produced some tweets regarding new services or update of services in those time slots. We can assume that it is also normal as many of the government and news services made their announcements/ service updates using automated tweet services. The figure also shows that the lowest tweet density occurs during the time 4 am to 10 am which indicated that the users merely uses the offered services in that time slot. After 10 am the number of tweets increases and reaches at the peak at 8 pm. This tweeting behavior is also usual as people tends to tweet after receiving the services or visiting their physician/hospitals.

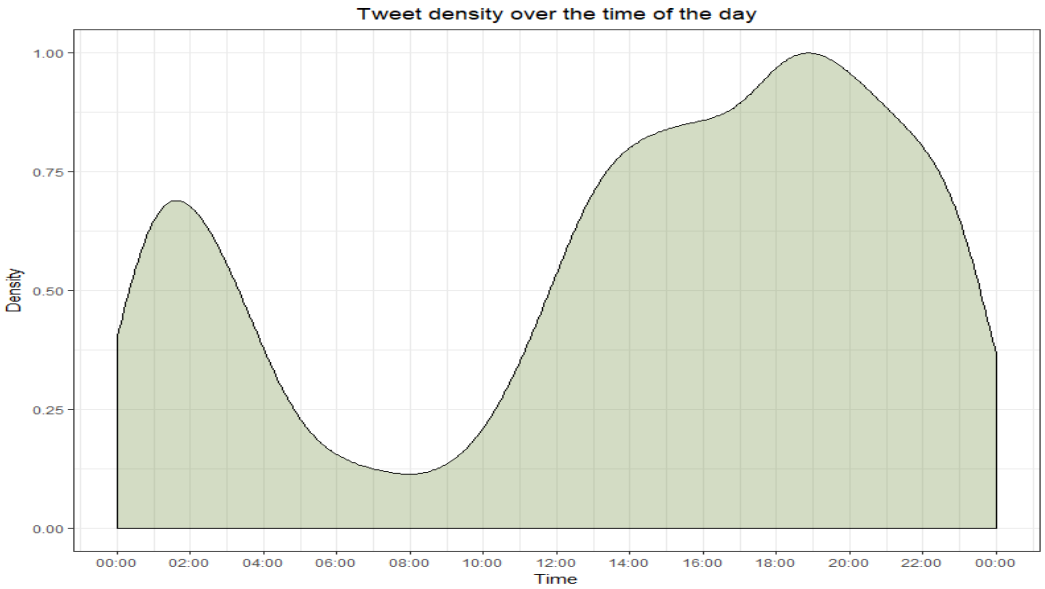


Figure 4: Users' tweeting tendency over time

This observation tells us that our dataset is free of unusual user accounts or harmful bots those can bias our final sentiment outcomes. We have also observed that the people merely replies on the post, but they tend to retweet on others' tweets as depicted in figure 5. This happens probably because users want to share same sort of experiences when users tweet to expresses their feelings.

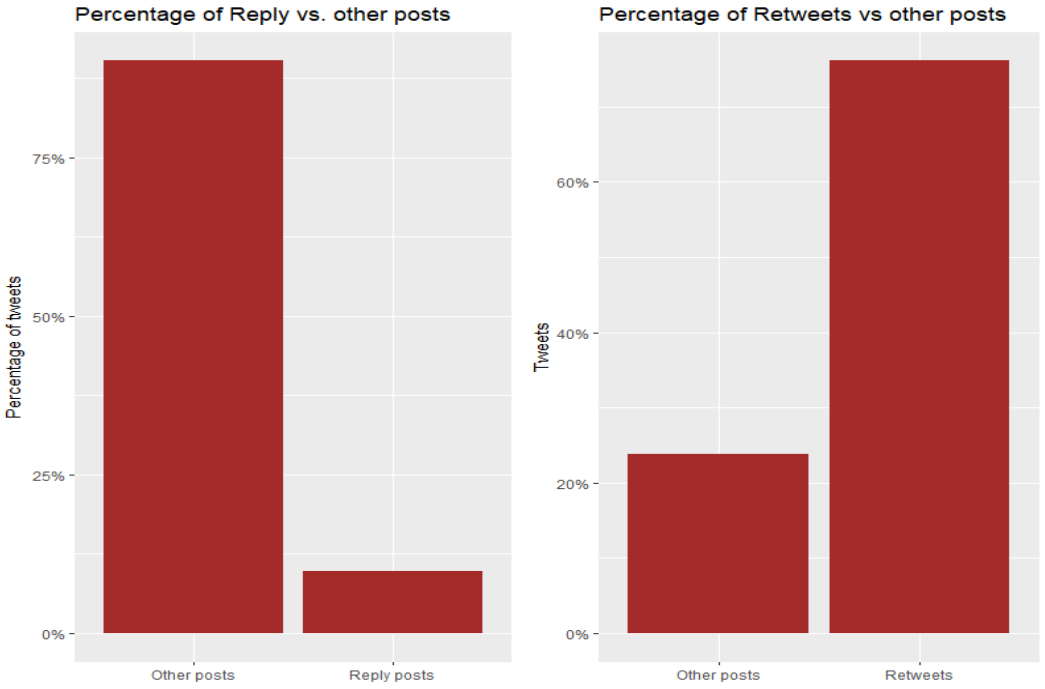


Figure 5: Percentage of reply retweet and other posts

Text Analysis:

In this part we have tried to find the words those can convey the feelings of the users in their tweets. To do so, NRC emotion lexicon [8] has been used which is a part of 'syuzhet' package. This package has very rich built in word dictionary and have different lexicons. Each lexicon has its own pros and cons. So, we have decided to use most recent NRC lexicon which tried to overcome most of the flaws of other lexicons and released in December 2017. This lexicon is a list of words and their associations with eight emotions classified as anger, fear, anticipation, trust, surprise, sadness, joy, and disgust, and two sentiments - negative and positive [9].

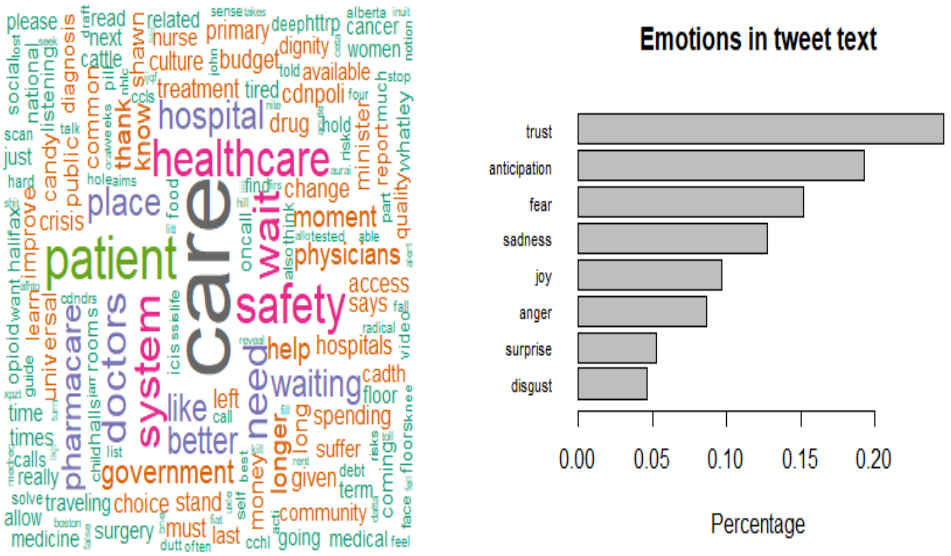


Figure 6: Word cloud and users emotions

Each tweet is examined and score of the category has been updated whenever a relevant word found inside the tweet as shown in figure 7. These emotions reveal that Canadian health care system and provided services are trustworthy among the users. Peoples expectancy meet their desire in most cases. Though many people expressed their anxiety relevant to the outcome or future goal of services. The comparison of sadness and expression also gives us different picture.

As figure 7 depicts that the expression of joy and sadness of users are in a ratio of approximately 1:1.6. This means that people tends to share their sad experiences most often than the good one. This probably the case that people anticipate high quality of services, and when it meets their expectancy, they tend to be neutral. When it reaches below their expectancy level, the expressed anger and disgust.

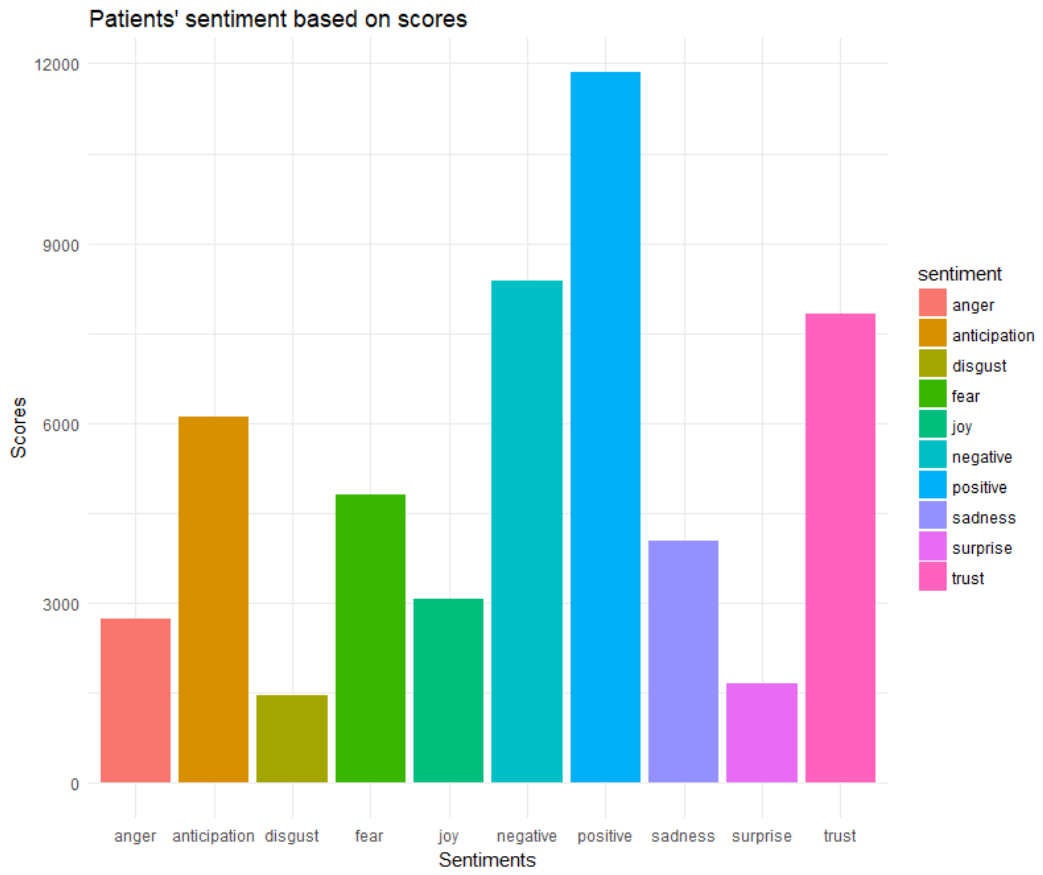


Figure 7: Users' emotions in health tweet data

The ratio of anger and joy is almost same but when we combine the ratio of anger and disgust together, it overwhelms the joy. But in total the proportion of positive sentiment is much higher than the negative sentiment, approximately 1.4:1.

Conclusion:

The analysis of this project reveals that people are satisfied about the health care services they receive. Yet the most satisfaction can be achieved by minimizing the fear factors among the users. The fear factor in our observation can be minimized if proper plan and goal projection has been publicly announced. The disgust and anger factors need to be investigated by the policy makers before making the decision.

The decision makers and the concerned authority may conduct such survey on a periodic interval to fetch the concern of the people regarding this service. The policy maker can also take advantages from this type of analysis for future goal projections.

References:

[1] “International Comparison Reflects Flaws and Opportunities for Better U.S. Health Care”, Eric C. Schneider, Dana O. Sarnak, David Squires, Arnav Shah, and Michelle M. Doty, July 2017. Available at, http://www.commonwealthfund.org/interactives/2017/july/mirror-mirror/assets/Schneider_mirror_mirror_2017.pdf

[2] “Social Media: A Review and Tutorial of Applications in Medicine and Health Care”, Francisco Jose Grajales, Samuel Sheps, Kendall Ho, Helen Novak-Lauscher, and Gunther Eysenbach, feb 2014, US National Library of Medicine and National Institutes of Health.

[3] “Collecting data on patient experience is not enough: they must be used to improve care”, Coulter, Angela; Locock, Louise; Ziebland, Sue; Calabrese, Joe. BMJ : British Medical Journal; London Vol. 348, (Mar 26, 2014).

[4] Lin M, Thoma B, Trueger NS, et al “Quality indicators for blogs and podcasts used in medical education: modified Delphi consensus recommendations by an international cohort of health professions educators” Postgraduate Medical Journal 2015;**91:**546-550.

[5] “Using Big Data to Measure and Improve Cardiovascular Health and Healthcare Services”, Jack V. Tu, Anna Chu, Linda R. Donovan, Dennis T. Ko, Gillian L. Booth, Karen Tu, Laura C. Maclagan, Helen Guo, Peter C. Austin, William Hogg, Moira K. Kapral, Harindra C. Wijeysundera, Clare L. Atzema, Andrea S. Gershon, David A. Alter, Douglas S. Lee, Cynthia A. Jackevicius, R. Sacha Bhatia, Jacob A. Udell, Mohammad R. Rezai, Thérèse A. Stukel, Circulation: Cardiovascular Quality and Outcomes. 2015;CIRCOUTCOMES.114.001416 Originally published February 3, 2015

[6] “Standardizing Patient Outcomes Measurement”, Michael E. Porter, Stefan Larsson, and Thomas H. Lee, Feb 2016, N Engl J Med 2016; 374:504-506 DOI: 10.1056/NEJMp1511701.

[7] Aizhan Tursunbayeva, Massimo Franco, Claudia Pagliari, Use of social media for e-Government in the public health sector: A systematic review of published studies, Government Information Quarterly, Volume 34, Issue 2, 2017, Pages 270-282, ISSN 0740-624X, https://doi.org/10.1016/j.giq.2017.04.001. Available at

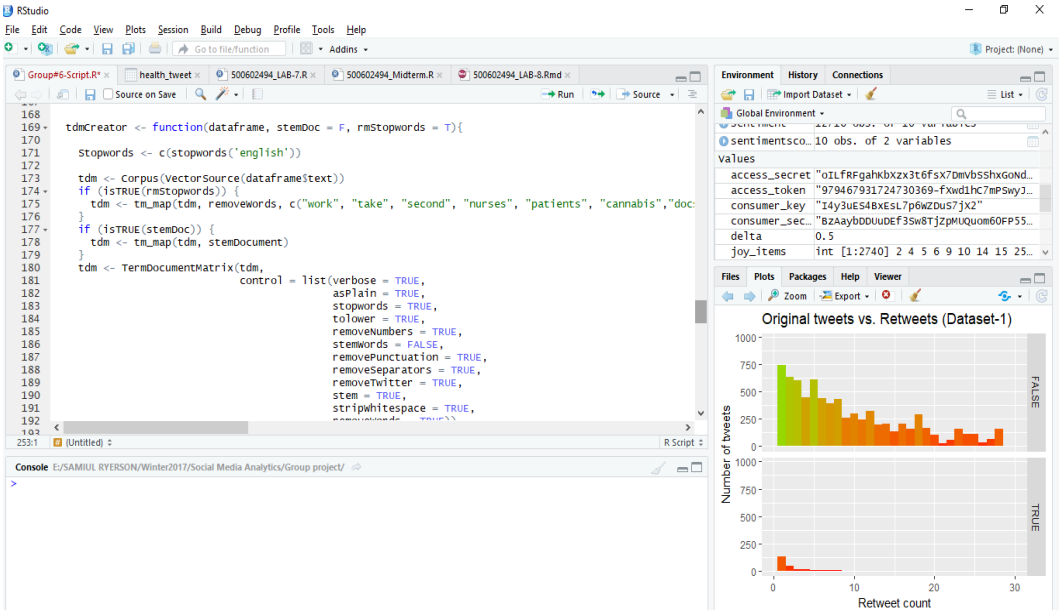
(http://www.sciencedirect.com/science/article/pii/S0740624X16302088)

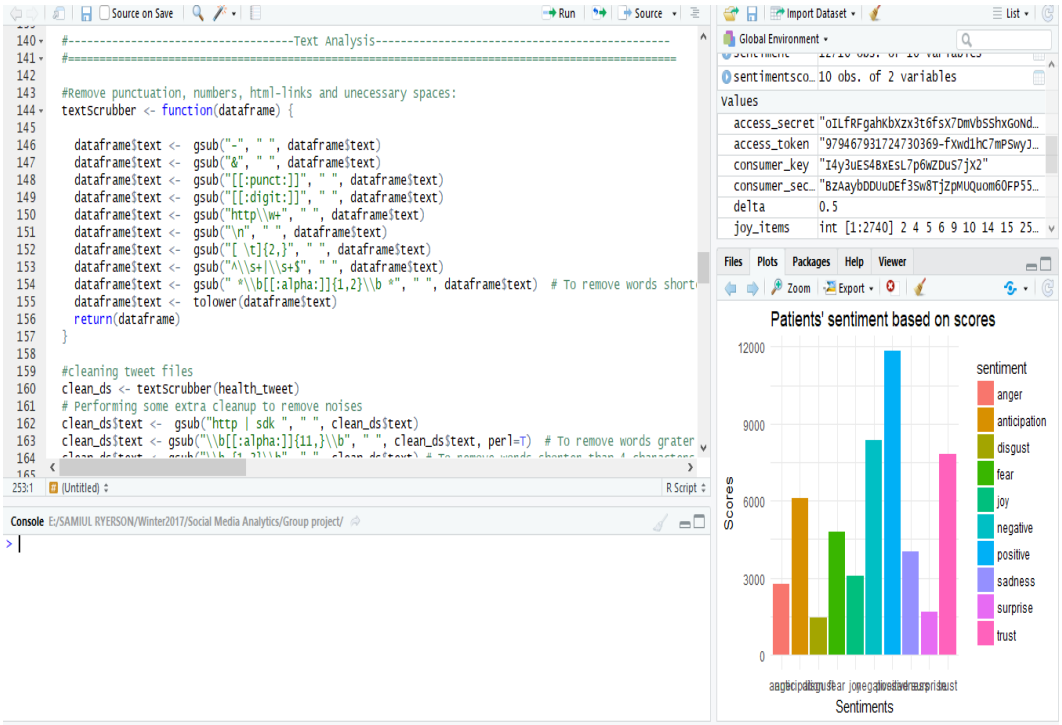
[8] Crowdsourcing a Word-Emotion Association Lexicon, Saif Mohammad and Peter Turney, Computational Intelligence, 29 (3), 436-465, 2013.

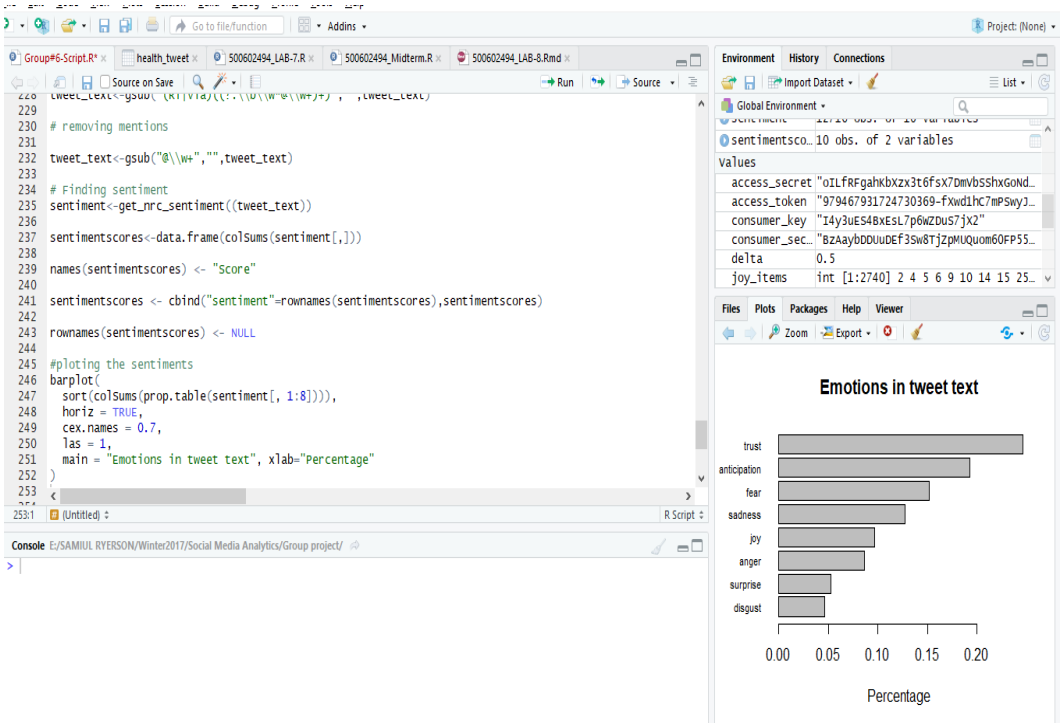
[9] Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon, Saif Mohammad and Peter Turney, In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, LA, California.

Appendix:

Screen shots:






Group#6-Script.R*
health_tweet
500602494_LAB-7.R
500602494_Midterm.R
500602494_LAB-8.Rmd
Source on Save
Run
Source
# removing mentions
tweet_text<-gsub("@\\w+","",tweet_text)
# Finding sentiment
sentiment<-get_nrc_sentiment((tweet_text))
sentimentscores<-data.frame(colSums(sentiment[,]))
names(sentimentscores) <- "Score"
sentimentscores <- cbind("sentiment"=rownames(sentimentscores),sentimentscores)
rownames(sentimentscores) <- NULL
#ploting the sentiments
barplot(
sort(colSums(prop.table(sentiment[, 1:8]))),
horiz = TRUE,
cex.names = 0.7,
las = 1,
main = "Emotions in tweet text", xlab="Percentage"
)
(Untitled)
R Script
Console E:/SAMIUL RYERSON/Winter2017/Social Media Analytics/Group project/
Environment
History
Connections
Import Dataset
List
Global Environment
sentimentsco... 10 obs. of 2 variables
Values
access_secret "oILfRFgahKbXzx3t6fsX7DmVbSShxGoNd...
access_token "979467931724730369-fXwd1hC7mPSwyJ...
consumer_key "I4y3uES4BxEsL7p6WZDuS7jX2"
consumer_sec... "BzAaybDDUuDEf3Sw8TjZpMUQuom6OFP55...
delta 0.5
joy_items int [1:2740] 2 4 5 6 9 10 14 15 25...
Files
Plots
Packages
Help
Viewer
Zoom
Export
Emotions in tweet text
trust
anticipation
fear
sadness
joy
anger
surprise
disgust
0.00
0.05
0.10
0.15
0.20
Percentage