ANTH 511 | Assignment 2 Twitter Analysis White Paper | Yasmin Stoss

Title:

A sentiment analysis of disinformation in scientific topics on Twitter

Brief explanation of the research questions and goals of your analysis:

A sentiment analysis of scientific topics was examined in the context of Twitter. Keywords used to identified tweets about scientific topics were identified by Suzie Allard, PhD and Natalie Rice, PhD as part of another project on the same topic. These keywords were chosen based on a literature review on the topics of misinformation, disinformation, and fake news and based on current news trends for scientific topics. This paper is examining scientific keywords that are commonly associated with disinformation, but it doesn not identify the tweets used in this project as disinformation. The aim of this analysis is to view tweets on these scientific topics through the lens of sentiment analysis. Of particular interest is the comparison between words identifying negative and positive sentiments generally and, specifically, the comparison between words identifying sentiments of fear and trust. Fear has been indicated as a powerful motivator in Fear Appeal theory, prominent in advertising. In addition, this project seeks to establish the power of the most prominent twitter accounts through predictive analysis. The 3 accounts with the most tweets will be compared to the entire set using a Granger Causality test. Events around the 2016 election are also examined in consideration of their impact on results. The research questions are as follows:

1. Are tweets on science disinformation topics more positive or negative?
2. Do tweets on science disinformation topics display more fear or trust?
3. Who are the top 10 tweeters of “science disinformation” topics (identified by greatest number of tweets) and do their sentiment patterns predict the general sentiment patterns across all tweets on these topics?

The keywords used in these analyses are: Clean coal, Dakota Access Pipeline, Ebola, Fracking, Paris Climate Accord, Pasteurization, Vaccinations, and Zika.

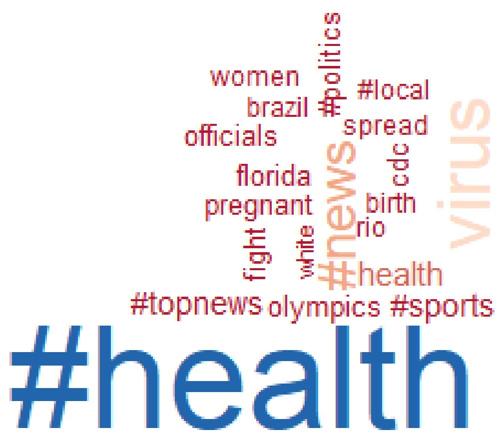
Short description of the data:

The data used are from the Clemson twitter dataset, created by Clemson University professors Darren Linvill and Patrick Warren. The dataset contains tweets from 2848 Twitter accounts identified as Russian bots. The dataset contains tweets from between February 2012 and May 2018 (Five Thirty Eight, 2018). I used the NRC Word-Emotion Association Lexicon created by Saif Mohammad to identify positive, negative, trust, and fear words for the sentiment analysis (2010).

Results and discussion:

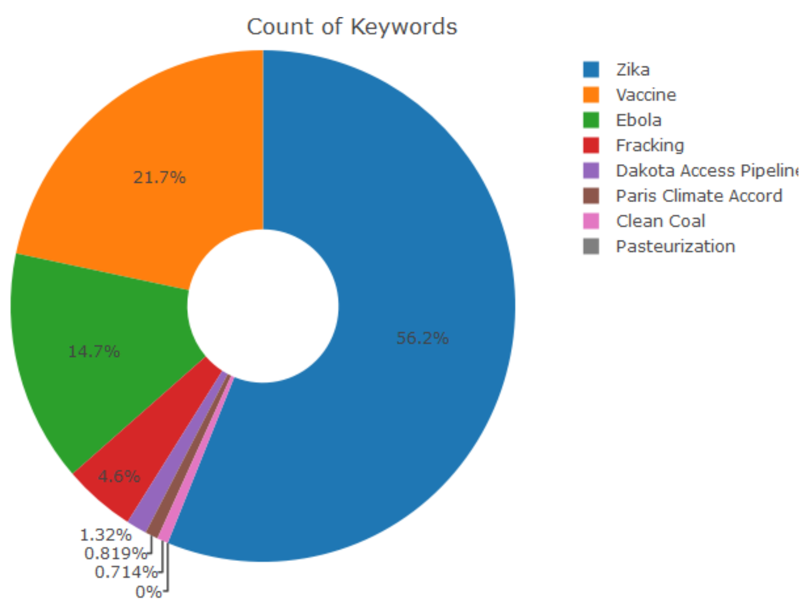
Tweets containing science disinformation keywords were isolated from the Clemson data set. The top 20 words in the new data set are presented in the word cloud in figure 1. The number one term was “health” suggesting that the bots were most interested in scientific topics surrounding health concerns. The list of keywords chosen for this project also shows that many of the scientific terms associated with disinformation are also associated with health, such as vaccination, zika, and ebola.

Figure 1:



These terms also make up a combined 92.6% of instances of the keywords in our data set (see figure 2).

Figure 2:



The reason for this focus on health issues is likely because it is a major concern for the public, with many lines of division identifying differing opinions or interests. Russia may consider this topic an easy entry point for spreading disinformation for this reason.

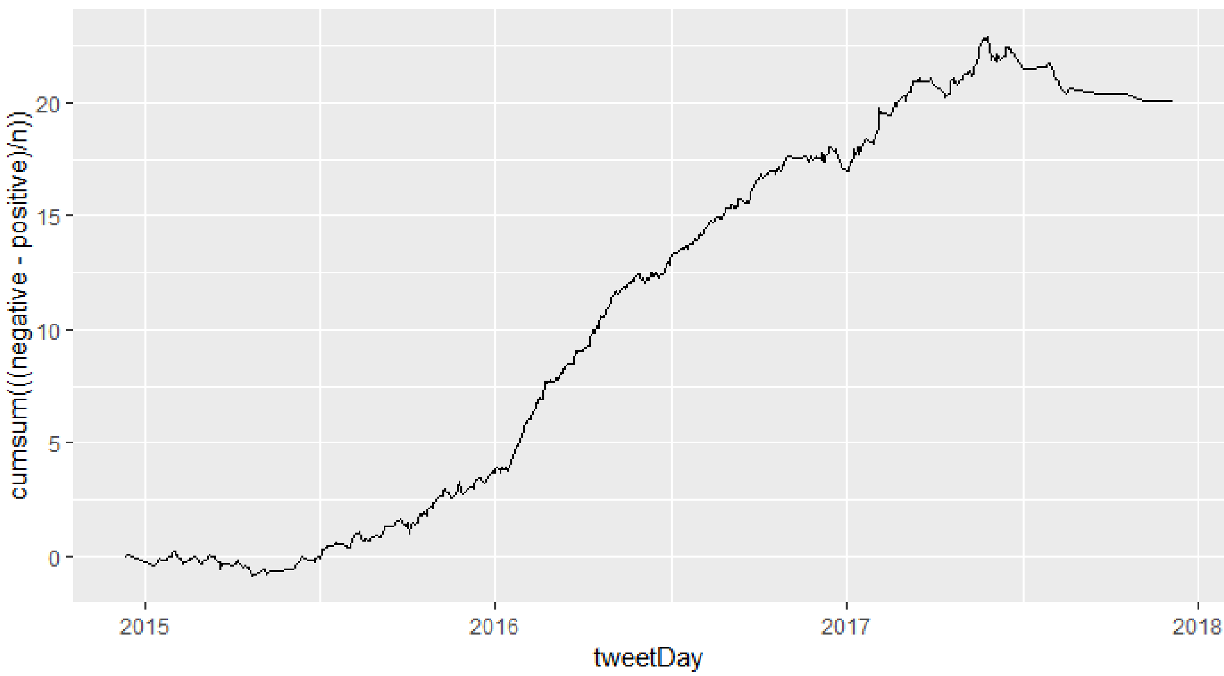
The data set was split into individual words and each words was identified as having either a positive or negative sentiment. A word cloud was created for both sentiments (see figures 3 and 4).

Figure 3:

Figure 4:

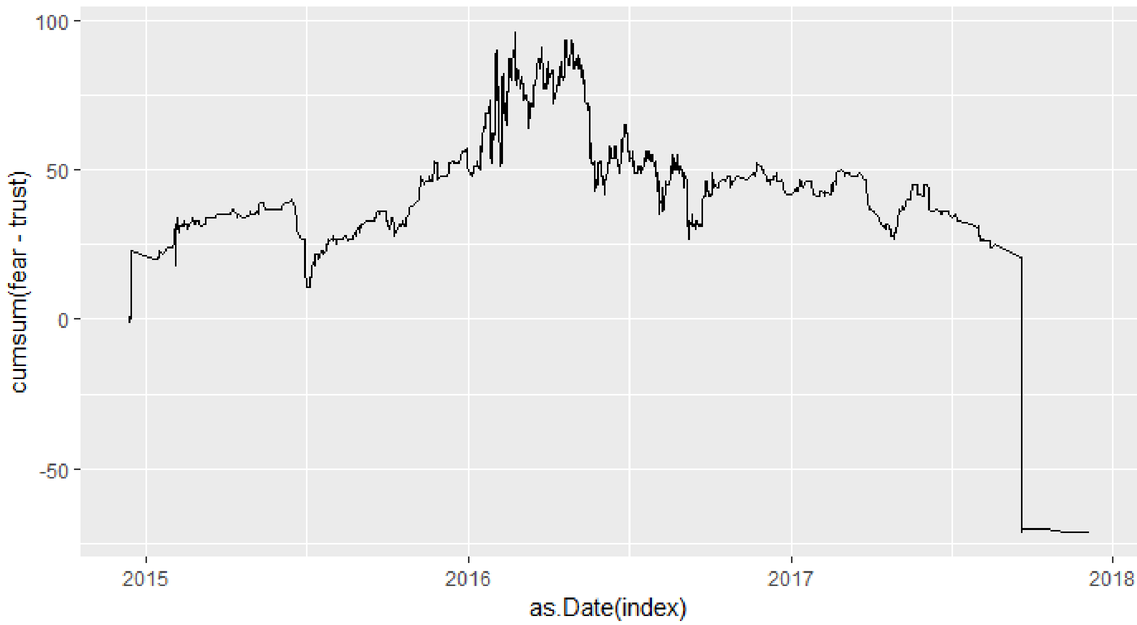
The keywords were filtered out of the data set to prevent them from confounding further analyses. During initial examination of the data the term “vaccine”, and variations of, substantially boosted the number of words in the positive sentiment grouping, skewing results. After removing the keywords, the relationship between positive and negative words was re-examined. This relationship was determined using the cumulative sum of the negative terms minus the positive terms to establish the trend between these sentiments over time. We are presented with a graph showing that a negative sentiment increased over time (see figure 5). In particular we see the negative sentiment increase drastically at the start of 2016 and then level back out toward the end of 2016. The line drops below zero in 2015 indicating a greater positive sentiment. These results may be related to the 2016 election. The positive sentiment in 2015 may be a result of excitement about the impending political campaigns and the promise of change while the great increase in negative sentiment around the election may indicate that Russia was trying to instill fear in potential voters to arouse action and create division.

Figure 5:



The same relationship was examined between fear and trust using the cumulative sum (see figure 6). Overall fear is greater than trust, particularly around the 2016 election. Interestingly we see a complete reversal of this trend toward the end of 2017. The reason for this enormous drop may be that many of the Russian bots left twitter at this point. This idea would be interesting to examine in further research.

Figure 6:



The cumulative sum for both the negative to positive and fear to trust relationship was re-examined with a data set normalized by the number of tweets per day. This normalization was done to account for an increase in general number of tweets over time. This graph (see figure 7) shows a more gradual increase in negative vs positive sentiment over time though, again, there is a greater increase in this difference at the beginning of 2016, potentially related to the election.

Figure 7:

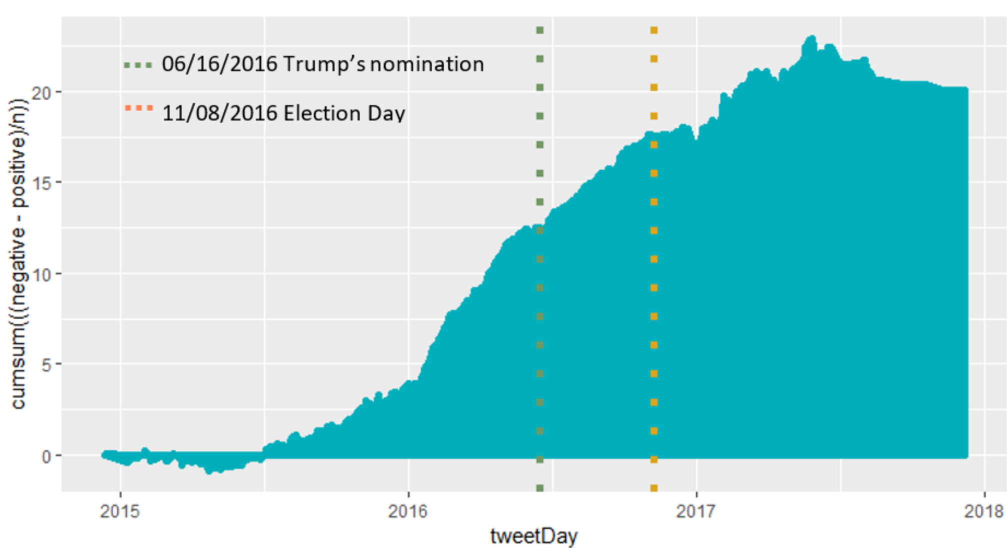
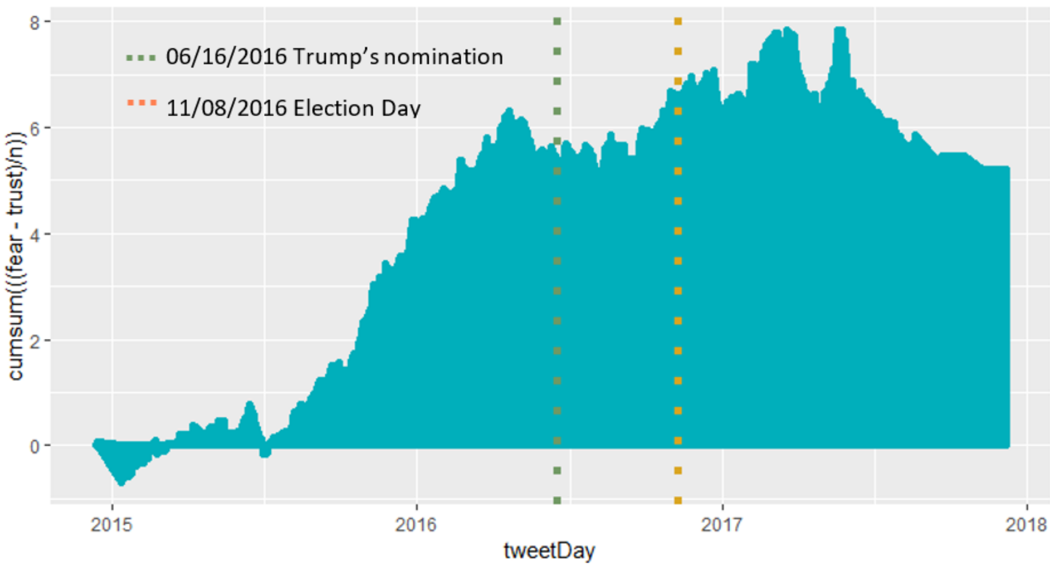


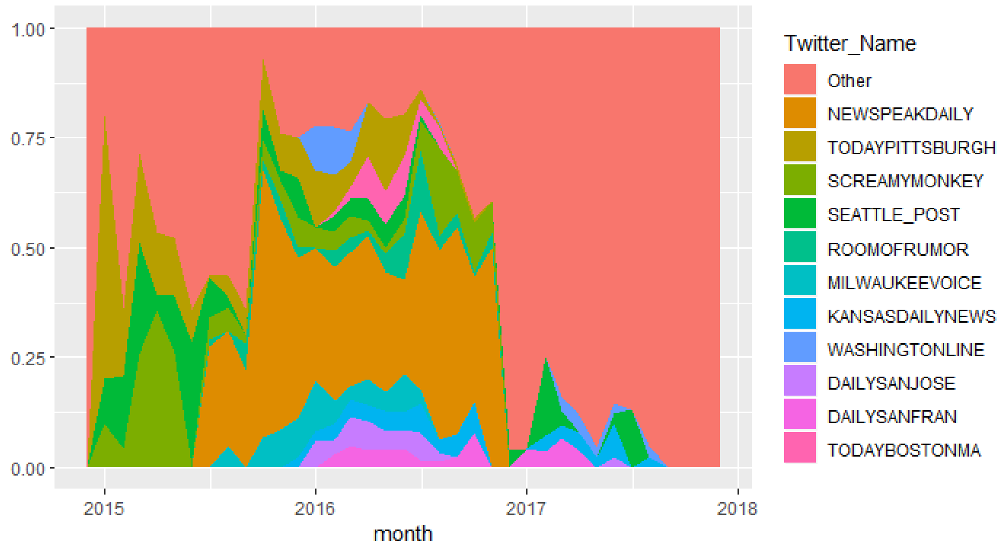
Figure 8 shows the cumulative sum for the fear vs trust relationship on the normalized data set. The results is, again, more negative over time, serving to confirm the result from the general test comparison. The reason for this increase is unclear however. The election and the announcement of Trump's running do not seem to have much of an impact. The difference between fear and trust sentiments does grow substantially starting in the summer of 2015. This could be the result of an increase in bot usage in social media beginning in the summer of 2014 (Roth, 2017). Another potential cause is the occurances of the Zika virus in South America starting early in 2015 with many reports following throughout that year and into 2016 (Center For Disease Control and Prevention, n.d.). Tweets associated with the keyword "zika" makeup 52.6% of the dataset so it is likely that a relationship exists.

Figure 8:



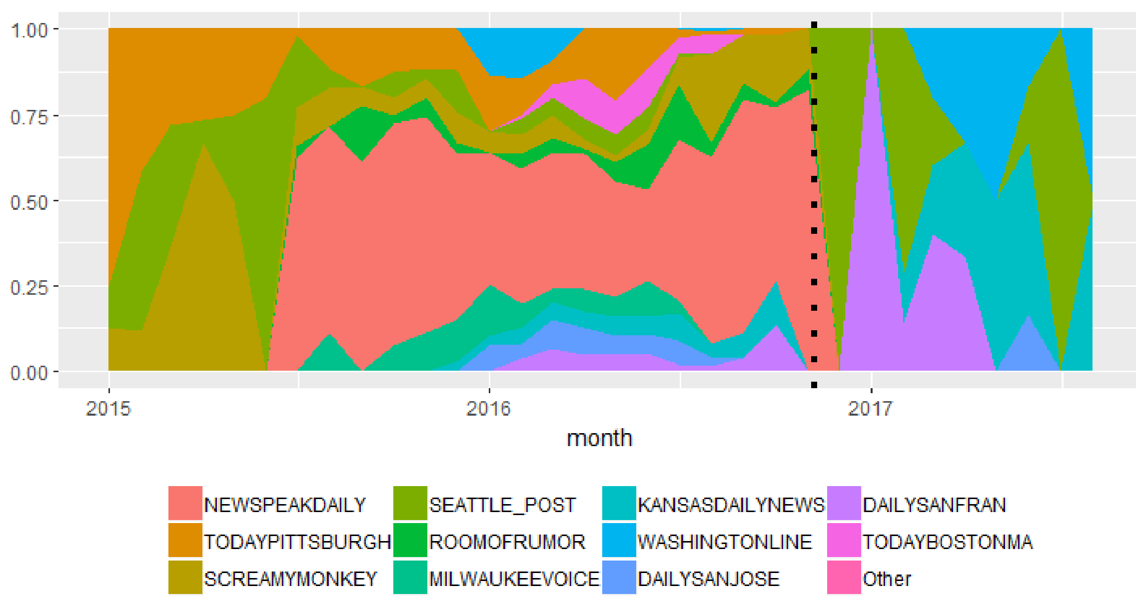
The ribbon plot in figure 9 depicts the activity of the top 10 tweeters compared to that of the 2838 remaining twitter accounts in this data set, labelled as "other". Surrounding the campaign trail and 2016 election these 10 twitter handles account for most of the bot activity.

Figure 9:



What is interesting, is the behavior of these twitter accounts after the 2016 election (see figure 10). Several disappear and the remaining accounts decrease their activity.

Figure 10:



To determine whether these top tweeting handles had an effect on the rest of the data set a granger causality test was conducted examining the top 3 accounts. Figures 11, 12, and 13 show that only one of the accounts, SCREAMYMONKEY, predicted the activity of the rest of the data set.

Figure 11:

```
Granger causality test

Model 1: all_timeSeries$percent ~ Lags(all_timeSeries$percent, 1:1) + Lags
(newspeakdaily_timeSeries$percent, 1:1)
Model 2: all_timeSeries$percent ~ Lags(all_timeSeries$percent, 1:1)
  Res.Df Df      F Pr(>F)
1    139
2    140 -1 2.3586 0.1269
```

Figure 12:

```
Granger causality test

Model 1: all_timeSeries$percent ~ Lags(all_timeSeries$percent, 1:1) + Lags
(todaypittsburgh_timeSeries$percent, 1:1)
Model 2: all_timeSeries$percent ~ Lags(all_timeSeries$percent, 1:1)
  Res.Df Df      F Pr(>F)
1    139
2    140 -1 0.0458 0.8308
```

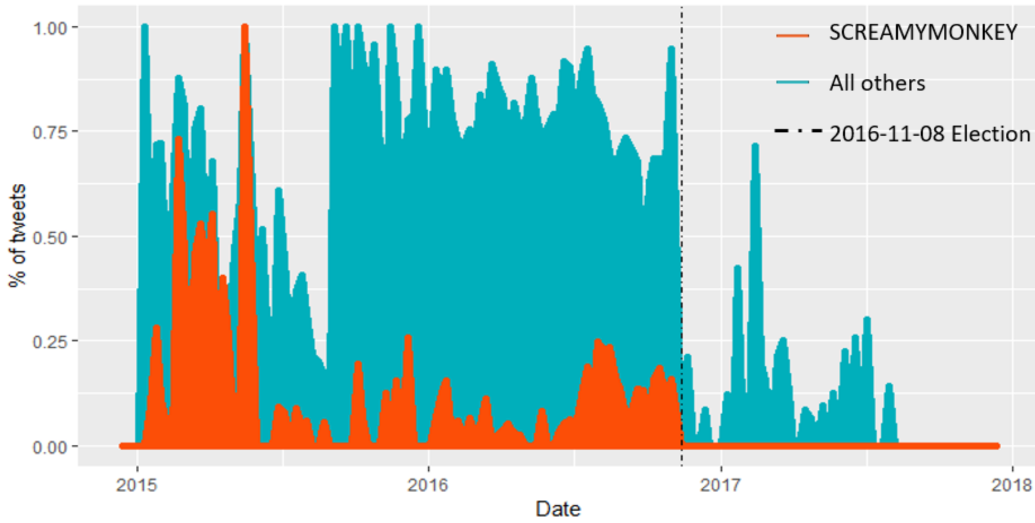
Figure 13:

```
Granger causality test

Model 1: all_timeSeries$percent ~ Lags(all_timeSeries$percent, 1:1) + Lags
(screamymonkey_timeSeries$percent, 1:1)
Model 2: all_timeSeries$percent ~ Lags(all_timeSeries$percent, 1:1)
  Res.Df Df      F  Pr(>F)
1    139
2    140 -1 3.6936 0.05667 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14 depicts the time series of SCREAMYMONKEY overlaid with the time series for the data set as a whole. The activity for SCREAMYMONKEY disappears right after the election and the activity for the entire data set dies down before picking back up just before Trump’s inauguration on January 20th, 2017.

Figure 14:



Short Conclusion:

In conclusion, much of the dataset is focused on health terms and on the Zika virus in particular. A sentiment analysis shows us that both negative and fearful sentiments have trended upward since 2015. The likely reasons for this are an increase in Russian bot activity focusing on dividing people, the Zika outbreak in South and North America in 2015-2016, and the 2016 U.S. election. Russian bots may be using more negative and fearful words in their tweets about health related issues in an effort to divide people and decrease trust in institutions. Russian bots have probably chosen to focus on health issues either because these issues are of greater concern to the general public compared to other scientific topics, especially in light of the Zika and Ebola outbreaks.

Additionally, there are several topics that could be examined in greater detail in future research. It would be interesting to determine when each of these keywords become more prominent in the data set, to see if that change is associated with a world event, and determine if this change is related to a change in sentiment.

Another potential examination of this data set would be to combine it with a general twitter dataset to analyze if and how the Russian bot accounts affect other twitter accounts. For example, a granger causality test could be done to determine if the top 10 tweeting accounts from this dataset predict activity in the general twitter dataset from the same time frame.

Appendix:

Packages used:
Plotly, lubridate, tidytext, tidyverse, lmtest, RColorBrewer, wordcloud

1. Read in csv file
    a. tweets <-read.csv("Clemson.csv")
2. Format date
    a. tweets_dated <-tweets %>% mutate(Date = mdy_hm(publish_date))
    b. colnames(tweets_dated)[colnames(tweets_dated)=="content"] <-"Tweet"
3. Clean data by removing URLs and other unwanted characters
    a. replace_reg <-"https://t.co/[A-Za-z\\d]+|http://[A-Za-z\\d]+|&|<|>|https"
    b. unnest_reg <-"([^A-Za-z_\\d#@']|'(?![A-Za-z_\\d#@]))"
    c. clean_tweets<-tweets_dated%>%mutate(Tweet=str_replace_all(Tweet,replace_reg,""))
4. Sort out all tweets containing the keyword(s)
    a. https://stackoverflow.com/questions/22850026/filtering-row-which-contains-a-certain-string-using-dplyr
    b. https://stackoverflow.com/questions/35962426/multiple-strings-with-str-detect-r
    c. keywords<-c("vacc", “Vacc", “vax”, “Vax”, “clean coal", “Clean Coal”, "paris climate accord", “Paris Climate Accord”, "ebola", “Ebola, "zika", “Zika, "dakota access pipeline", “Dakota Access Pipeline”,"pasteuriz", “Pasteuriz”, “frack”, “Frack”)
    d. tweets_sci_words<-clean_tweets%>%filter(str_detect(Tweet,paste(keywords, collapse="|")))
5. Count instance of each keyword
    a. https://stackoverflow.com/questions/7782113/count-word-occurrences-in-r
    b. Vaccines: sum(str_count(tweets_sci_words$Tweet, "vacc")) AND sum(str_count(tweets_sci_words$Tweet, "Vacc")) AND sum(str_count(tweets_sci_words$Tweet, "Vax")) AND sum(str_count(tweets_sci_words$Tweet, "vax"))
        i. Result = 821
    c. Ebola: Ebola = 556
    d. Zika: Zika = 2125
    e. Paris Climate Accord: Paris climate accord = 31
    f. Clean Coal: Clean coal = 27
    g. Dakota Access Pipeline: Dakota access pipeline = 50
    h. Pasteurization: Pasteuriz = 0
    i. Fracking: Frack = 174
6. Plot instance of each keyword in a donut plot
    a. Keywords_count <- read.csv(“Keywords_count.csv”)
    b. Keywords_count %>% group_by(keyword) %>% plot_ly(labels=~keyword, values=~count) %>% add_pie(hole=0.3)
7. Create a data frame with the tweets separated out into single words
    a. sci_words<-tweets_sci_words%>%unnest_tokens(word,Tweet,token="regex",pattern=unnest_reg)%>%filter(!word %in% stop_words$word,str_detect(word, "[a-z]"))
8. Filter out the keywords.

a. keywords2<-c("vacc", "Vacc", "vax", "Vax", "clean", "coal", "Clean", "Coal", "paris", "climate", "accord", "Paris", "Climate", "Accord", "ebola", "Ebola", "zika", "Zika", "dakota", "access", "pipeline", "Dakota", "Access", "Pipeline","pasteuriz", "Pasteuriz", "frack", "Frack")
b. sci_words2<-sci_words%>%filter(!str_detect(word,paste(keywords2, collapse="|")))

9. Create a new data frame including the sentiments for each word
    a. nrc<-get_sentiments("nrc")
    b. sci_tweet_sentiment <-sci_words2 %>% inner_join(nrc) %>% count(index = Date, sentiment) %>% spread(sentiment, n, fill = 0)
10. Create a word cloud for the top 20 words for the entire data set, the positive words, the negative words
    a. pal <- brewer.pal(9,"RdBu")
    b. wordcloud(words = count_sci_word$word, freq = count_sci_word$n, min.freq = 1, max.words=20, color=pal)
    c. wordcloud(words = positive_all$word, freq = positive_all$n, min.freq = 1, max.words=20, color=pal)
    d. wordcloud(words = Negative_all$word, freq = Negative_all$n, min.freq = 1, max.words=20, color=pal)
11. Plot the negative-positive cumulative sum
    a. sci_tweet_sentiment %>% ggplot(aes(x=as.Date(index), y=cumsum(negative-positive))) + geom_line()
12. Plot the fear-trust cumulative sum
    a. sci_tweet_sentiment %>% ggplot(aes(x=as.Date(index), y=cumsum(fear-trust))) + geom_line()
13. Normalize the data by daily tweets
    a. daily_sci <-sci_words2 %>% mutate(tweetDay= floor_date(Date, "day"))
    b. daily_sci_sentiment <-daily_sci %>% inner_join(nrc) %>% count(index = tweetDay, sentiment) %>% spread(sentiment, n, fill = 0)
    c. colnames(daily_sci_sentiment)[colnames(daily_sci_sentiment)=="index"] <-"tweetDay"
    d. wordsPerDay <-count(daily_sci, tweetDay)
    e. SentimentsPerDay <-inner_join(daily_sci_sentiment, wordsPerDay)
14. Plot the negative-positive cumulative sum with normalized data (include election day and date of Trump's running announcement)
    a. ggplot(data = SentimentsPerDay) + geom_area(aes(x=tweetDay, y=cumsum(((negative-positive)/n))),color="#00AFBB", size=2, fill="#00AFBB")+geom_vline(xintercept = as.POSIXct(as.Date(c("2016-11-08"))), linetype="dotted",color = "#daa520", size=2)+geom_vline(xintercept = as.POSIXct(as.Date(c("2016-06-16"))), linetype="dotted",color = "#709963", size=2)

15. Plot the fear-trust cumulative sum with normalized data (include election day and date of Trump's running announcement)
    a. ggplot(data = SentimentsPerDay) + geom_area(aes(x=tweetDay, y=cumsum(((fear-trust)/n))),color="#00AFBB", size=2, fill="#00AFBB")+geom_vline(xintercept = as.POSIXct(as.Date(c("2016-11-08"))), linetype="dotted",color = "#daa520", size=2)+geom_vline(xintercept = as.POSIXct(as.Date(c("2016-06-16"))), linetype="dotted",color = "#709963", size=2)
16. Determine the top 10 tweeters by count of tweets
    a. count(sci_words2, Twitter_Name, sort=TRUE)
17. Do a ribbon plot of the top 10 tweeters
    a. sci_words2 %>%   mutate(Twitter_Name = fct_lump(Twitter_Name, 10)) %>%   count(month = round_date(Date, "month"), Twitter_Name) %>% complete(month, Twitter_Name, fill = list(n = 0)) %>% mutate(Twitter_Name = reorder(Twitter_Name, -n, sum)) %>% group_by(month) %>%mutate(percent = n /sum(n),   maximum = cumsum(percent), minimum = lag(maximum, 1, 0)) %>% ggplot(aes(month, ymin = minimum, ymax = maximum, fill = Twitter_Name)) + geom_ribbon()+theme(legend.position="bottom", legend.title = element_blank())+geom_vline(xintercept = as.POSIXct(as.Date(c("2016-11-08"))), linetype="dotted",color = "black", size=1.5)
18. Filter out the top 10 tweeters and do a ribbon plot for just those 10 (add a line for the election day)
    a. top10_tweeters<-c("NEWSPEAKDAILY","TODAYPITTSBURGH","SCREAMYMONKEY","SEATTLE_POST","ROOMOFRUMOR","MILWAUKEEVOICE","KANSASDAILYNEWS","WASHINGTONLINE","DAILYSANJOSE","DAILYSANFRAN","TODAYBOSTONMA")
    b. Top10_sci_words<-tweets_sci_words%>%filter(str_detect(Twitter_Name,paste(top10_tweeters, collapse="|")))
    c. Top10_sci_words %>%   mutate(Twitter_Name = fct_lump(Twitter_Name, 10)) %>%   count(month = round_date(Date, "month"), Twitter_Name) %>% complete(month, Twitter_Name, fill = list(n = 0)) %>% mutate(Twitter_Name = reorder(Twitter_Name, -n, sum)) %>% group_by(month) %>%mutate(percent = n /sum(n),   maximum = cumsum(percent), minimum = lag(maximum, 1, 0)) %>% ggplot(aes(month, ymin = minimum, ymax = maximum, fill = Twitter_Name)) + geom_ribbon()+theme(legend.position="bottom",

legend.title = element_blank())+geom_vline(xintercept = as.POSIXct(as.Date(c("2016-11-08"))), linetype="dotted",color = "black", size=1.5)

19. Determine the top 3 tweeters
    a. count(sci_words2, Twitter_Name, sort=TRUE)
    b. Top3_tweeters<-c("NEWSPEAKDAILY", "SCREAMYMONKEY","TODAYPITTSBURGH")
20. Do a Granger Causality test between the top 3 tweeters and the rest of the data set
    a. all_timeSeries<-sci_words2%>%group_by(week=round_date(Date,"week"))%>%summarize(tweets=n(), account=sum(str_detect(Twitter_Name, paste(top10_tweeters, collapse="|"))), percent=account/tweets)
    b. newspeakdaily_timeSeries<-sci_words2%>%group_by(week=round_date(Date,"week"))%>%summarize(tweets=n(), account=sum(str_detect(Twitter_Name, "NEWSPEAKDAILY")), percent=account/tweets)
    c. screamymonkey_timeSeries<-sci_words2%>%group_by(week=round_date(Date,"week"))%>%summarize(tweets=n(), account=sum(str_detect(Twitter_Name, "SCREAMYMONKEY")), percent=account/tweets)
    d. todaypittsburgh_timeSeries<-sci_words2%>%group_by(week=round_date(Date,"week"))%>%summarize(tweets=n(), account=sum(str_detect(Twitter_Name, "TODAYPITTSBURGH")), percent=account/tweets)
    e. grangertest(newspeakdaily_timeSeries$percent,all_timeSeries$percent, order=1)
    f. grangertest(todaypittsburgh_timeSeries$percent,all_timeSeries$percent, order=1)
    g. grangertest(screamymonkey_timeSeries$percent,all_timeSeries$percent, order=1)
21. Visualize the time series for the granger causality tests with significant p values
    a. ggplot()+geom_area(data=all_timeSeries, aes(x=as.Date(week), y=percent), color="#00AFBB", size=2, fill = "#00AFBB")+geom_area(data=screamymonkey_timeSeries, aes(x=as.Date(week), y=percent), color="#FC4E07", size=2, fill="#FC4E07")+xlab("Date")+ylab("% of tweets")+geom_vline(aes(xintercept=as.numeric(as.Date("2016-11-13"))), linetype=4, color="black")

References:

Center For Disease Control and Prevention (n.d.). Zika Virus. Retrieved from https://www.cdc.gov/zika/reporting/2016-case-counts.html

Five Thirty Eight (2018). Russian Troll Tweets. GitHub. https://github.com/fivethirtyeight/russian-troll-tweets

Mohammad, S.M. (2010). NRC Word-Emotion Association Lexicon. Retrieved from http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Roth, A. (2017). Pro-Putin bots are dominating Russian political talk on Twitter. The Washington Post. Retrieved from https://www.washingtonpost.com/world/europe/pro-putin-politics-bots-are-flooding-russian-twitter-oxford-based-studysays/2017/06/20/19c35d6e-5474-11e7-840b-512026319da7_story.html?noredirect=on&utm_term=.364167fd9c43