Communication on social media during 2020 coronavirus outbreak

Tang Chunmin claudiatang95@gmail.com

2/23/2020

Contents

word count: 1204

Overview of the project

Since the middle January 2020, the outbreak of coronavirus in China has aroused the world's attention. People have shared all kinds of information about this epidemic on social media. The main goals of this project is to analyze :

1. Is there any association between the number of followers and popularity of tweets?
2. What are the sources and contents of the most popular tweets?
3. Sentiment analysis.
4. The basic text analysis (e.g. the most frequent words, the most frequent hashtags).

Besides, it's also worth noting that during this outbreak, social media has become a hotbed for fake news, conspiracy theories and racism. These misleading information is circulating on social media and penetrating into people's daily life, aggravating public's panic. This is a good topic for future study and in order to prepare some materials, I collected some tweets that contain rumors about coronavirus outbreak in China.

The tasks list

1. Data collection
   To collect English tweets with hashtag #coronavirus posted from 25 Jan to 12 Feb.

2. Data combination and sampling
   To put all the tweets collected into one dataset and take a sample for text analysis.

3. Rumor tweets collection
   To collect tweets that contain two rumors about coronavirus:

- A Chinese nurse claimed in a video that there were already 90000 people infected.
- The coronavirus was invented by Chinese government for biowarfare and may have leaked from the Wuhan Institute of Virology

4. Analyze the users that received more "likes" and retweets.
5. Analyze the tweets that received more "likes" and retweets.
6. Sentiment analysis.
7. Basic text analysis.

Task review and difficulties encountered

1. Data collection

I collected tweets posted every two days during this period, i.e the tweets posted on 25 Jan, 27 Jan, 29 Jan... And in order to get the latest favorite counts and retweet counts, I collected the tweets seven days after the date when they were posted. Thus I got a dataset of 261745 tweets.

2. Data combination and sampling

I put all the data together and drew a sample of 10% of total tweets, which is a sample of 26174 tweets. This sample is used to do the text analysis.

3. Rumor tweets collection

I searched and collected rumor tweets with some keywords. Obviously, in this way my dataset contains some tweets that are not related to my topics. I need to go through them and manually delete the unrelated one when I use these data in the future. This is time-consuming and I hope in the future I can figure out a better way to collect and filter rumor tweets.

4. Analyze popular users and popular tweets

I added a new variable "twitter engagement", which is the sum of favorite count and retweet count, to indicate the popularity of the tweets. First I extracted top 100 accounts whose tweets have higher average twitter engagement values. Then I analyzed the percentages of the verified accounts and unverified accounts. I used ANOVA to test if the difference of mean twitter engagement of verified and unverified users is significant. I also used linear regression to test if there is any correlation between the number of followers and popularity of tweets. Second I extracted top 1000 tweets that have higher twitter engagement values. I manually analyzed the top 10 most popular tweets.

5. Sentiment analysis and text analysis

I used quanteda package to do the text analysis. I analyzed the most frequent words, hashtags and mentioned users. For sentiment analysis, I used NRC dictionary in quanteda.dictionaries package to analyze the emotions of sample tweets.

Findings

1. Findings of statistical analysis

Among the top 100 accounts whose tweets have higher average twitter engagement values, 73% are unverified accounts and only 27% are verified. It seems that the average twitter engagement value of unverified accounts is higher than that of verified accounts, but ANOVA shows that the difference is not significant at 0.05 significance level.

```
## # A tibble: 2 x 3
##   verified followers twitter_engagement
##   <lgl>        <dbl>              <dbl>
## 1 FALSE      211057.              8101.
## 2 TRUE      2363868.              6583.
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## verified     1 4.541e+07  45414039   0.239  0.626
## Residuals   98 1.863e+10 190053963
```

Linear regression also shows that there is no correlation between the number of followers and the popularity of the tweets.

```
##
## Call:
## lm(formula = average_engagement ~ followers, data = topusers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9905  -4944  -3928  -1383  86527
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.196e+03  1.460e+03   4.930 3.35e-06 ***
## followers   6.250e-04  6.245e-04   1.001    0.319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13730 on 98 degrees of freedom
## Multiple R-squared:  0.01012,    Adjusted R-squared:  1.628e-05
## F-statistic: 1.002 on 1 and 98 DF,  p-value: 0.3194
```

I used the top 1000 most popular tweets to repeat the above ANOVA and regression analysis. The results are consistent.

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## verified      1 1.085e+07 10848400   0.143  0.705
## Residuals   998 7.552e+10 75673534
```

```
##
## Call:
## lm(formula = twitter_engagement ~ followers_count, data = toptweets)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2844  -2124  -1683   -552 185419
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.134e+03  2.988e+02  10.487   <2e-16 ***
## followers_count 2.356e-05  9.481e-05   0.248    0.804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8699 on 998 degrees of freedom
## Multiple R-squared:  6.185e-05,  Adjusted R-squared:  -0.0009401
## F-statistic: 0.06173 on 1 and 998 DF,  p-value: 0.8038
```

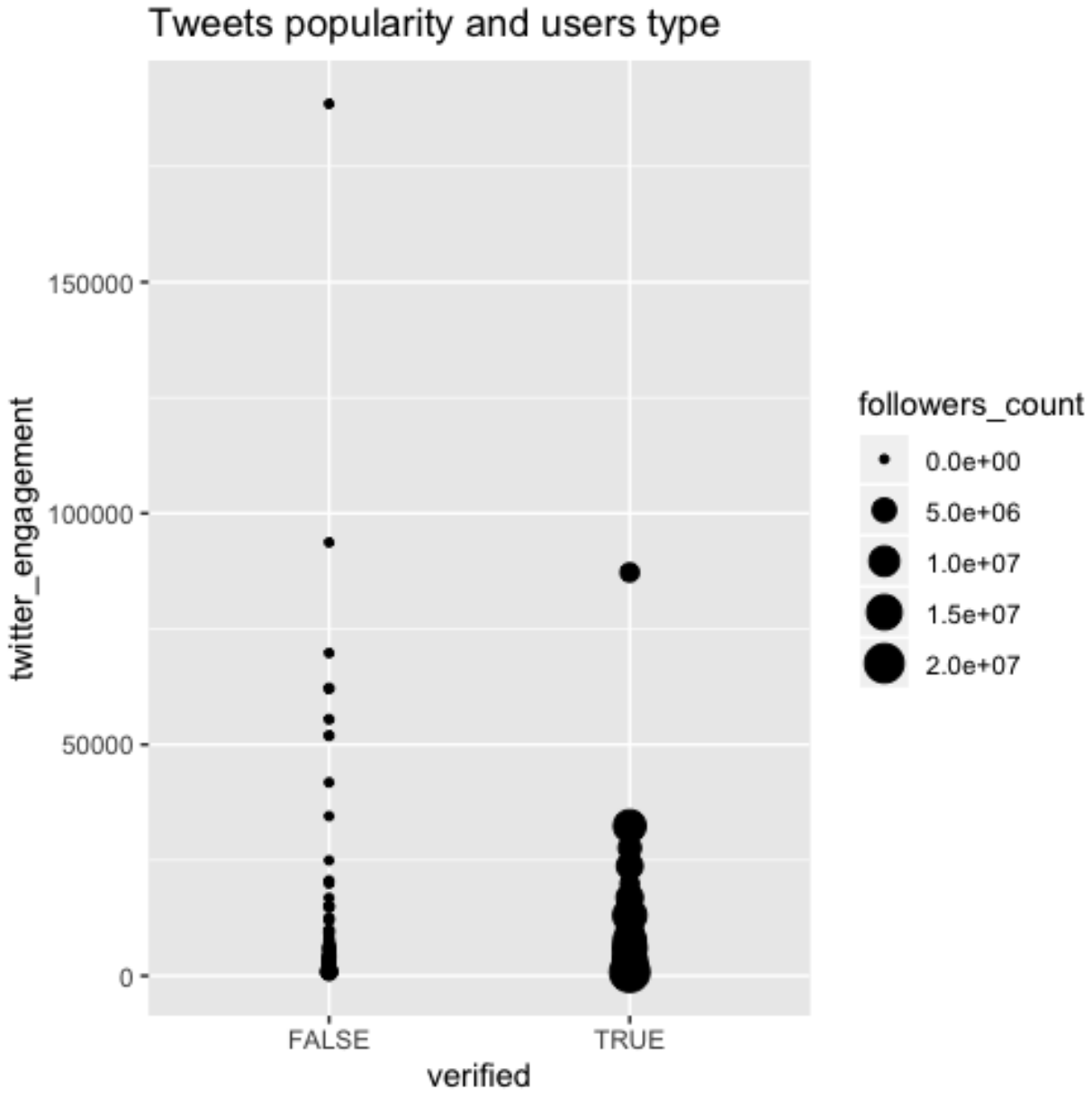
However, as to the percentages of unverified and verified accounts in toptweets dataset, the gap gets smaller. The unverified users account for 53.4% and verified users 46.6%.
Probing into the topusers dataset, I found that the first three accounts actually don't have a large number of followers, all below 200.

```
##      screen_name count average_favorite average_retweet average_engagement
## 1      new_prykm     1         48999.00        44724.00           93723.00
## 2       elmnzhri     1         33510.00        36293.00           69803.00
## 3    menumpahkan     3         35866.33        27986.33           63852.67
## 4 AngelinaShen36     1         19780.00        22038.00           41818.00
## 5    RahulGandhi     1         26726.00         5660.00           32386.00
## 6 RealJamesWoods     4         25137.75         6238.25           31376.00
##   followers verified
## 1        49    FALSE
## 2       172    FALSE
## 3       188    FALSE
## 4       448    FALSE
## 5  12128438     TRUE
## 6   2231395    FALSE
```

And there are also accounts that even have one follower or 0 follower, but still get thousands of favorite counts or retweet counts.

```
##       screen_name count average_favorite average_retweet
## 1 ClarenceBeeks14     1             2267             476
## 2 khongbengofkala     1              330            2141
##   average_engagement followers verified
## 1               2743         1    FALSE
## 2               2471         0    FALSE
```

Therefore, I believe that under a popular hashtag, a large number of followers is not the prerequisite for a tweet to go viral. Moreover, the verified users are not at an advantage to receive more favorites and retweets.
The following scatter plot of top 1000 most popular tweets further confirms my point of view.



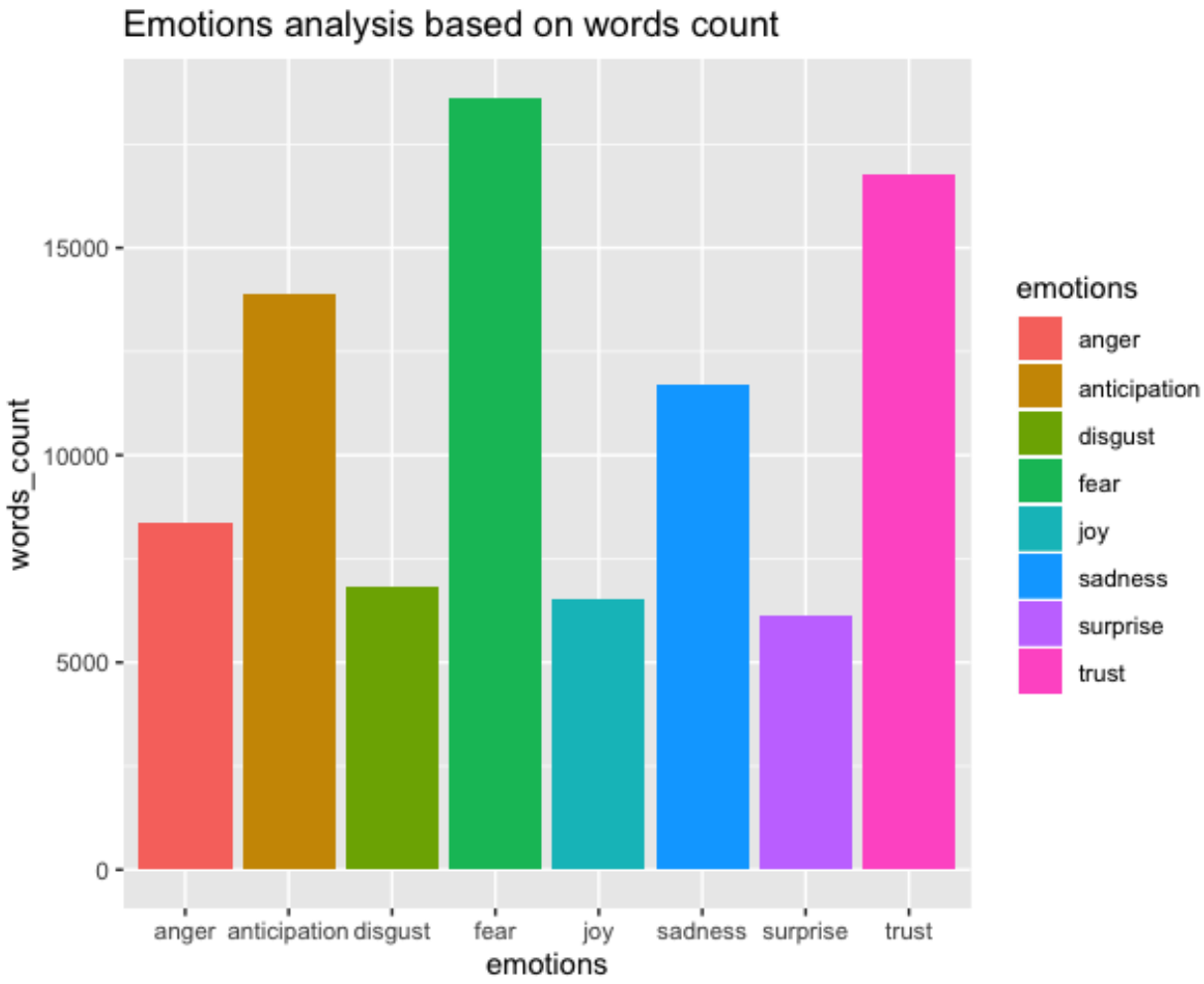
2. What are the sources and contents of the most popular tweets?

Rank	Name	Who is he/she	Topic	Attached media	Verified	Category	Followers	Favorite count	Retweet count
1	menumpahkan	Fan account of BTS (a South Korean boy band)	How to wear a mask	Non-original video	F	Information	188	105665	82888
2	new_prykm	Unknown	How south korea airpot is dealing with coronavirus	Original video (First time posted on Twitter)	F	Information	49	48999	44724
3	RealJamesWoods	James Wood (American actor, supporter of Trump)	The actual death toll is significantly more than what is reported by mainstream media	No media	T	Personal opinion	2231395	73168	14054
4	elmnzhri	Unknown	The government is doing nothing in his country to deal with coronavirus	Non-original video	F	Personal opinion	172	33510	36293
5	DarrenPlymouth	Chemist, Atheist	The patient infected was quivering on the hospital bed.	Non-original video	F	Information	30996	36784	18920
6	juliojiangwei	Spokesman of Chinese embassy in Panama	Two elderly patients of #coronavirus in their 80s said goodbye in ICU	Deleted	F	Information	5887	36554	18920
7	manyapan	Sinologist, China social trend & online media watcher	A video made by Chinese expressing people's boredom when staying indoors during the lockdown.	Original video (First time posted on Twitter)	F	Humor	12871	42353	9571
8	AngelinaShen36	A girl lives in Wuhan	The Wuhan local talks about what she saw and heard	Original picture	F	Personal experience	448	19780	22038
9	purplelovehime	Korean	English translation of the viral nurse video	Original picture	F	Misinformation	95	17795	16732
10	RahulGa	Member of the Indian	Coronavirus is a significant threat and the government is	No media	T	Personal opinion	12128438	26726	5660

The contents of top 10 popular tweets are diverse, including information, personal opinion, indirect personal experience, misinformation and humor. Most of them contain videos or pictures. Here original video means it was posted on Twitter for the first time although it may not be shot by the user himself/herself. What is unexpected is that the personal opinion tweets all expressed negative point of views towards government: distrust in mainstream media, disappointment in their governments for not taking effective actions to deal with coronavirus. Among 10 tweets, 8 are from unverified users. All 10 tweets were posted by individual accounts.

Due to limited time and skills, I didn't classify more popular tweets. Based on top 10 tweets, I didn't see any particular pattern that can lead the tweets to go viral. So I assume that the success of those tweets is random and unpredictable.

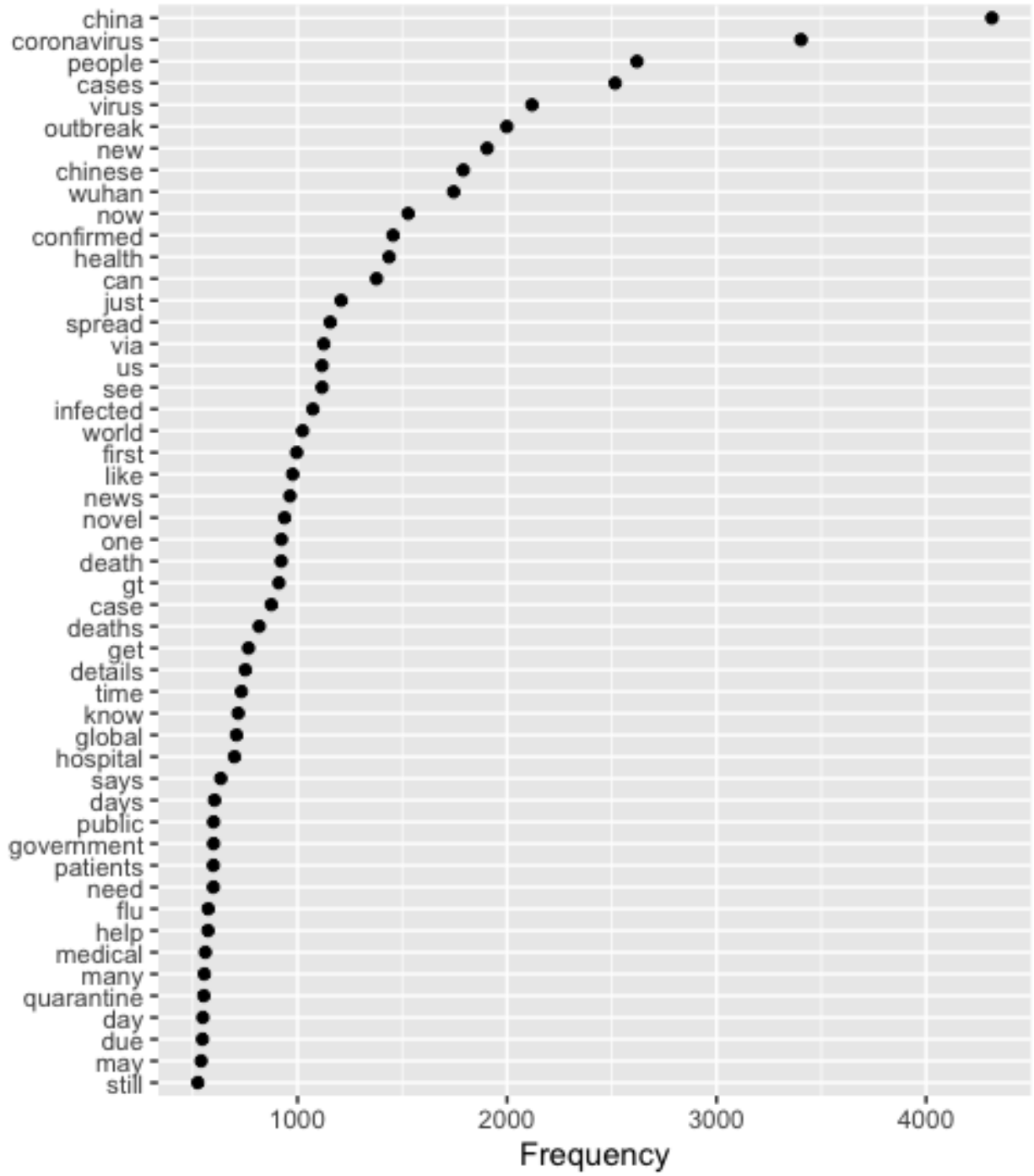
3. Findings of sentiment analysis



As expected, the most intense emotion is fear. The second one is trust. This may be attributed to the numerous tweets that contain the words like "confirm", "advice", "authority" etc. However, it is in question whether these tweets really expressed trust or actually it's the opposite.

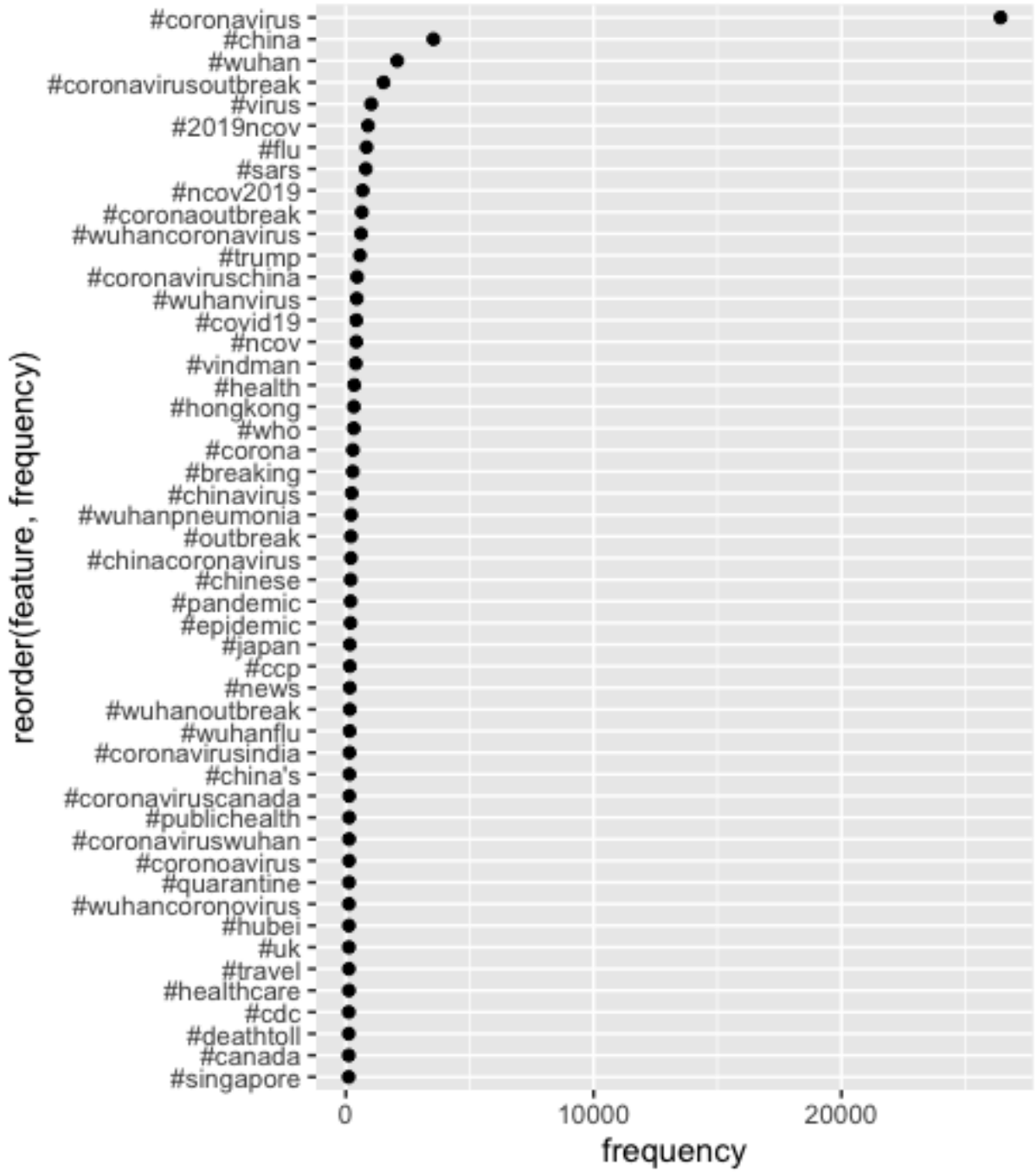
4. Findings of text analysis

1.The 50 most frequent words



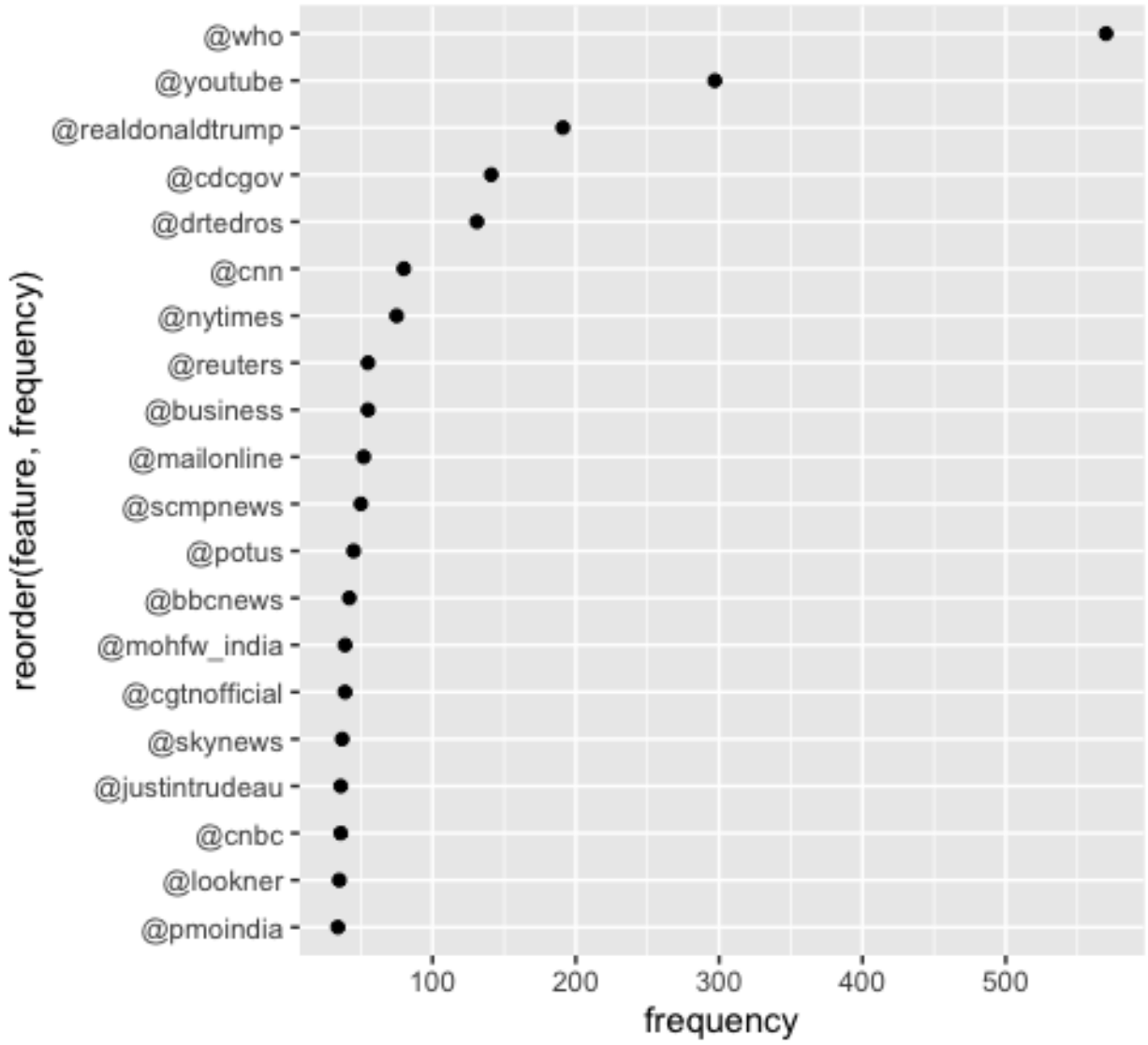
It shows that people are concerned about the outbreak in US and worldwide.

2. The 50 most frequent hashtags (#coronavirus excluded)



Among these hashtags, most of them are directly related to the coronavirus outbreak event. The rest of them are countries, health, government or politicians (#ccp , #trump and # vindman).

3. The 20 most frequently mentioned users



What could not be achieved

I planned to do topic modeling, but I found that it's difficult to categorize the topics because tweets are short and all the tweets are already under the same hashtag # coronavirus. For now I don't have time and skills to classify the contents. I hope in the future I can realize this task. And I'm also not very convinced by the results of the sentiment analysis due to the irony or other rhetoric in the tweeets.

Conclusion

Doing this project I learned how to collect tweets with R and how to do the basic text analysis. I'm happy that I got all these findings, but in the meantime I also felt limited, by the lack of both theoretical knowledge and technical skills.

Finally, let's pray that this coronavirus outbreak can be controlled and pass soon. Don't panic, don't blame, but be careful and cooperative!