Greta Thunberg on Social Media: The Fickle Public

Jeremy Chu, Yadi He, Jungeun Lim, Yanqi Li
Feb. 04, 2020

Abstract

This report is divided into three sections about network analysis, text analysis and sentiment analysis respectively. Network analysis was conducted over Twitter and Youtube, each including two research questions related towards the discussion of Greta Thunberg, a 17-year-old environmental activist on climate change, on the respective platform. For the Twitter section, a summary of the demographic breakdown of the top influencers is also included. Data for this report was gathered by Netlytics and YouTube Data Tools. Data analysis was performed in R Studio by computing Pearson's correlation coefficients and conducting tests such as ANOVA, regression, Pearson's correlation, and t-test. For text analysis and sentiment analysis, each section has one main research question. The analysis was conducted based on data gathered over Twitter two weeks after network analysis.

The Twitter Network

Network Properties:

Diameter: 17

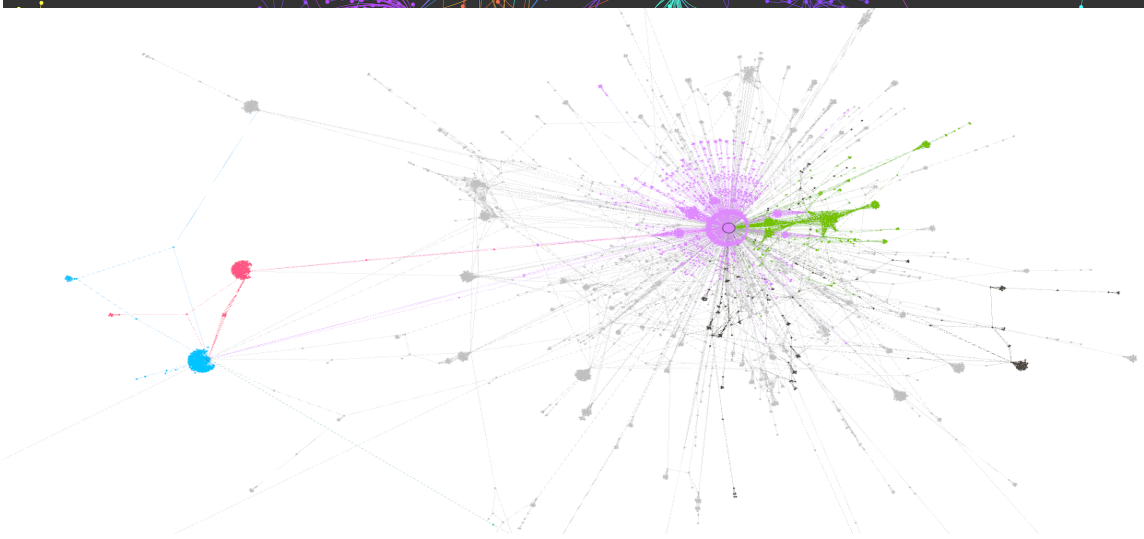
Density: 0.000206

Reciprocity: 0.017230

Centralization: 0.088240

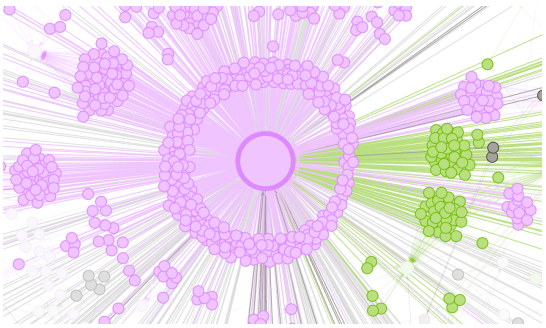
Modularity: 0.861200

Average Degree: 1.28



Insights:

At the centre lies @GretaThunberg with 1086 in-degree. This is unsurprising considering the dataset was gathered with @GretaThunberg as one of the search criteria, and it would come as a shock if she was not at the centre.

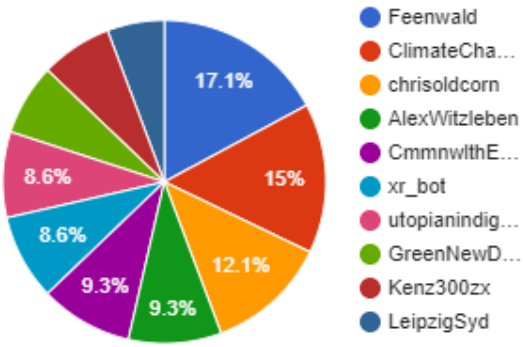


Diameter being 17 most likely is owing to the many divisive issues Greta Thunberg has been recently part of. The scandal that found out her father was posting on her Facebook, the Australian Fire, Trump, Greta is part of quite a number to trending topics.

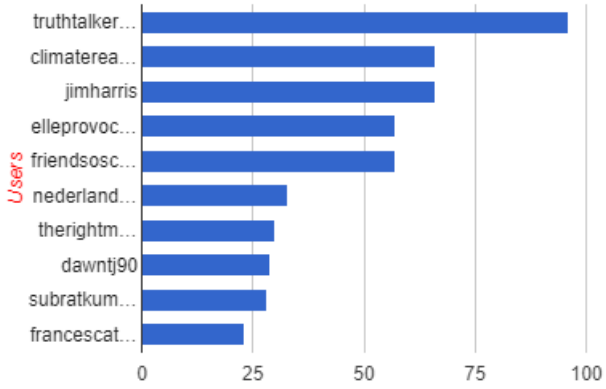
Reciprocity being at 0.017 would possibly indicate that people are merely tweeting @GretaThunberg instead of discussing issues. Directing outrage at, or simply tagging Greta in tweets, for example.

Modularity at 0.86 is exceedingly high, which suggests that the users are divided into groups and the groups do not communicate with each other very much. Part of the reason should be that there are non-English speakers who speak with each other; however, with a closer observation, we could identify two ideologically polarized groups (namely, pro-Greta who agree on climate change and anti-Greta who are skeptical about it).

Top 10 Posters



Top 10 Posters Mentioned in Messages



Interestingly, there is no overlap between the top 10 posters and the top 10 posters mentioned in messages.

Twitter Influencers (80 top in-degree accounts) Breakdown

Among the top (30 minimum-100 maximum) in-degree accounts from our Twitter network, there were 8 activists (including Greta Thurnberg herself), 10 politicians, 7 media accounts, 4 journalists, 14 non-profit organizations, 2 companies, and 19 influencers (we defined influencers as individual users with more than 10K followers). There are a few overlaps among these categories; e.g., Al Gore, who is a (former) politician and activist. In addition, there were 14 laypeople most of whom showed their interests in politics in their profiles.

Among the 80 twitter accounts we identified, 30 users were apparently pro-Greta who believe and fight climate change and 16 people users were apparently anti-Greta who are skeptical about climate change.

Research Questions - Twitter

Question 1. Is there any correlation between the number of followers and the indegrees/ or degrees?

This analysis tries to find out if the number of followers of a user is correlated with how many people mention and talk to him/her when tweeting about Greta Thunberg.

Hypothesis 1: If a person has more followers, she will have a higher indegree value.
Hypothesis 2: If a person has more followers, she will have a lower outdegree value.

The first hypothesis is based on the intuition that more popular users tend to be mentioned more and talked to. The second hypothesis is based on the intuition that more popular users tend to tweet toward a broader audience group rather than mentioning or talking to a specific user.

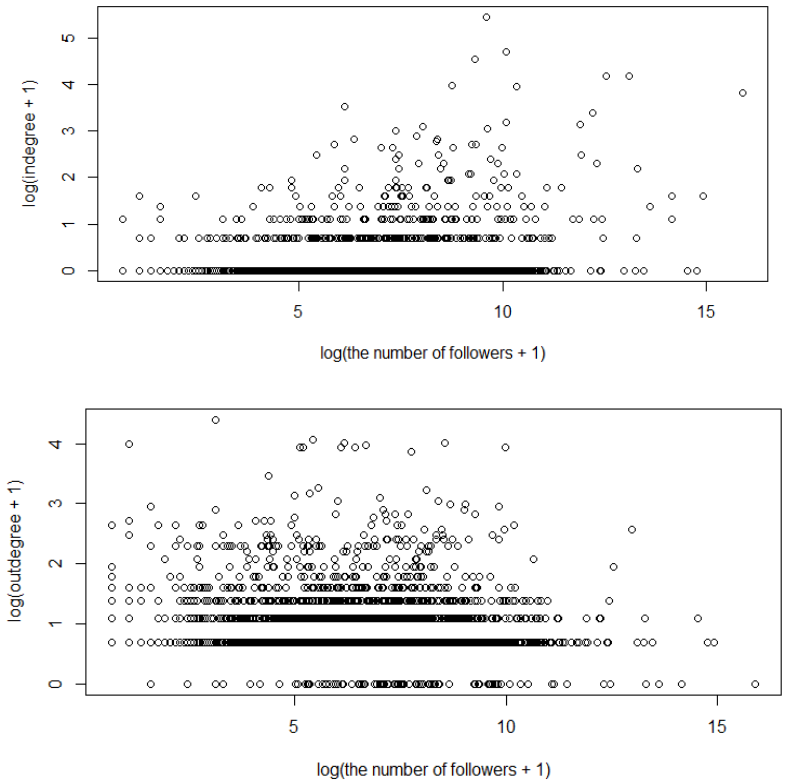


Figure 1. Scatterplots based on 4,184 users. Since the data were highly skewed, logarithm of each variable + 1 (1 was added to avoid the log of 0) was taken. Since most of the users have very low indegree and outdegree values, the relationships between the variables are hard to identify; however, a slightly positive relationship on the chart below and a slightly negative relationship in the chart above were observed.

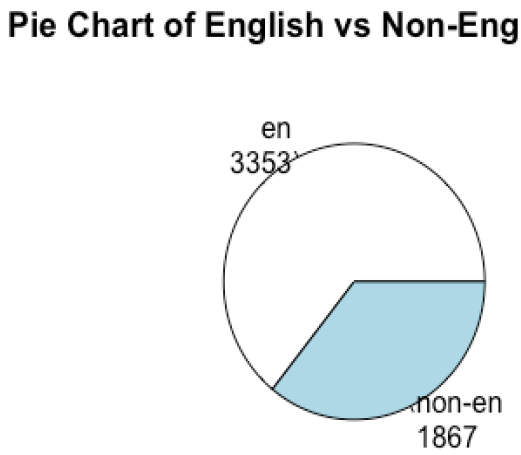
Pearson's correlation between the number of followers and indegree	Pearson's correlation between the number of followers and outdegree
Pearson's product-moment correlation data: greta_filtered$Followers and greta_filtered$indegree t = 9.8983, df = 4182, p-value < 2.2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.1215561 0.1807745 sample estimates: cor 0.1513011	Pearson's product-moment correlation data: greta_filtered$Followers and greta_filtered$outdegree t = -0.78593, df = 4182, p-value = 0.432 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: -0.04243899 0.01815659 sample estimates: cor -0.01215236

Table 1. Pearson's correlation coefficients show a positive correlation between the number of followers and indegree, and a negative correlation between the number of followers and outdegree. However, only the correlation between the number of followers and independence has a reasonably low p-value that is enough to reject the null hypothesis (p=2.2e-16).

There indeed exists a positive correlation between the number of followers and indegree value, and a negative correlation between the number of followers and outdegree value. However, the correlation coefficients are not very big and the p-value is low enough to reject the null hypothesis only for the former.

Question 2. People in which language group are more inclined to discuss about #GretaThunberg? Do users attract more followers based on the language they use?

The following pie chart reflects the distribution of language groups among the top Twitter influencers related to the subject Greta Thunberg. To better illustrate the distribution, the language groups are combined into English vs. non-English :



As shown above, English is the major language group that leads the discussion about Greta Thunberg on Twitter, which is assumed that people in the English-speaking countries are more interested in this topic. However, since Twitter is an American company, it might also be that the

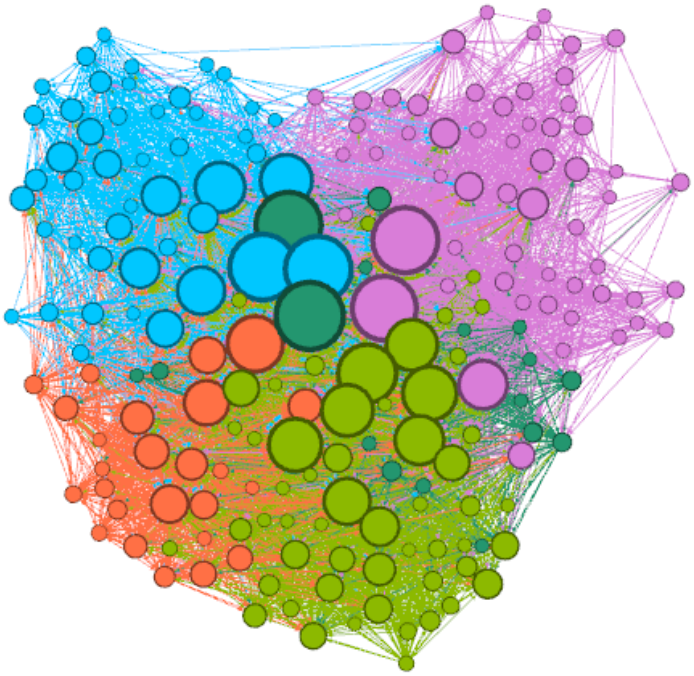
majority of Twitter users speak English. Also, Tweets written in the languages that do not use latin alphabets were not included in the dataset. With these two factors in mind, the assumption might not be true.

Based on the ANOVA test results shown below, the p value is 1, which is significantly greater than 0.05. With the null hypothesis being language does not impact the number of followers for users, the p value indicates that it fails to reject the hypothesis. As a result, the number of followers each account has does not have a significant relationship to the language of the account.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lang	26	8.22e+10	3.162e+09	0.164	1
Residuals	5193	1.00e+14	1.926e+10		

The YouTube Network

Network Properties:
Graph Density: 0.000206
Modularity: 0.203
Average Degree: 29.095



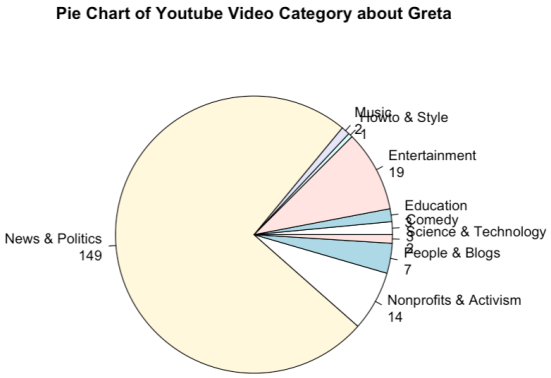
Insights:

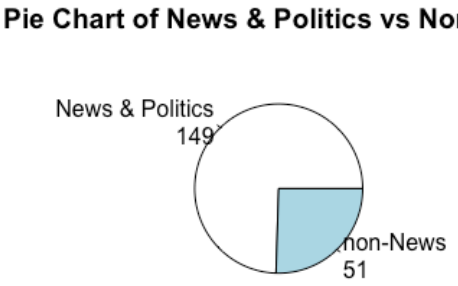
A quick glance would indicate that videos are connected through a core group of nodes. Rather than clear divisions between communities, the videos are largely interconnected. The nature of YouTube video recommendations most likely attest to this. Rather than random highly viewed videos, the videos are more often than not related as well. The communities would represent video categorizations, and therefore overlap is unsurprising as they are all about Greta Thunberg.

Research Questions - YouTube

Question 3. In YouTube video, what category of videos related to Greta Thunberg the most? Use pie chart to represent. Does view count and like count differ between video categories?

1. Use pie-chart to show what category of videos related to Greta Thunberg the most:





As shown above, most of the Great Thumberg videos are under the News & Politics category.

2. Use ANOVA to analyze the relationship between view counts and video categories:
   Null Hypothesis: there is no significant difference in the effect of the different video categories on the number of view counts.

```
                    Df    Sum Sq   Mean Sq F value Pr(>F)
videocategorylabel   8 9.444e+12 1.181e+12   1.705 0.0994 .
Residuals          191 1.323e+14 6.924e+11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As shown above, F(8,191)=1.705, P-value = 0.0994 >0.05, which fails to reject the null hypothesis. It means different video category labels have no impact on the number of view counts.

3. Use ANOVA to analyze the relationship between like counts and video categories:
   Null Hypothesis: there is no significant difference in the effect of the different video categories on the number of like counts.

```
                    Df    Sum Sq   Mean Sq F value Pr(>F)
videocategorylabel   8 8.279e+09 1.035e+09   2.441 0.0156 *
Residuals          187 7.928e+10 4.239e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
4 observations deleted due to missingness
```

As shown above, F(8,187)=2.441, P-value = 0.0156 < 0.05, which rejects the null hypothesis. It means at least one of the like counts at one video category label is different than at other video category labels.

Question 4. What factors are related to high in-degree videos, aka what is the relationship between highly recommended videos and views/comments/likes and dislikes?

Based on the attributes available, the hypothesized factors possibly responsible for a high in-degree value are: **dislike like ratio, commentcount, and viewcount**.

- In addition, calculations were made and 2 extra attributes were created: **likeratio** and **ratingratio**.
- **likeratio** represents the ratio of likes to dislikes;
- **ratingratio** represents the % of viewers who participated in liking/disliking a video.

Multiple regression was chosen to evaluate the relationship between the dependent variable **indegree** against selected attributes.

Results are as follows:

```
#Dependent: Indegree. Is there a relation between indegree and likeratios, comments, views?
my.lm <- lm(indegree ~ likeratio + dislikelikeratio + ratingratio + commentcount + viewcount, data=Youtube2)
summary(my.lm)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.484e+01  1.650e+01   0.899  0.36976
likeratio         1.123e+01  2.059e+01   0.545  0.58624
dislikelikeratio -3.882e+00  4.661e+00  -0.833  0.40612
ratingratio      -1.079e+02  1.516e+02  -0.712  0.47745
commentcount      2.326e-03  8.669e-04   2.684  0.00800 **
viewcount         1.905e-05  6.344e-06   3.003  0.00307 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.17 on 172 degrees of freedom
  (22 observations deleted due to missingness)
```

- The extraordinary high P-value indicates that likes and dislikes have little to no significance in determining a video's recommendation.
- Comments and views, on the other hand, were considerably more significant. With views more so than comments.
- One thing to note before any hypothesis, Greta Thunberg videos generally receive little to no likes or dislikes. The ratio of rating activity to viewers is only at a meager 3%.

The results seem to support the fact that YouTube algorithms decide which video to recommend based on views. Comments are believed to be significant simply because highly viewed videos are more likely to generate discussion.

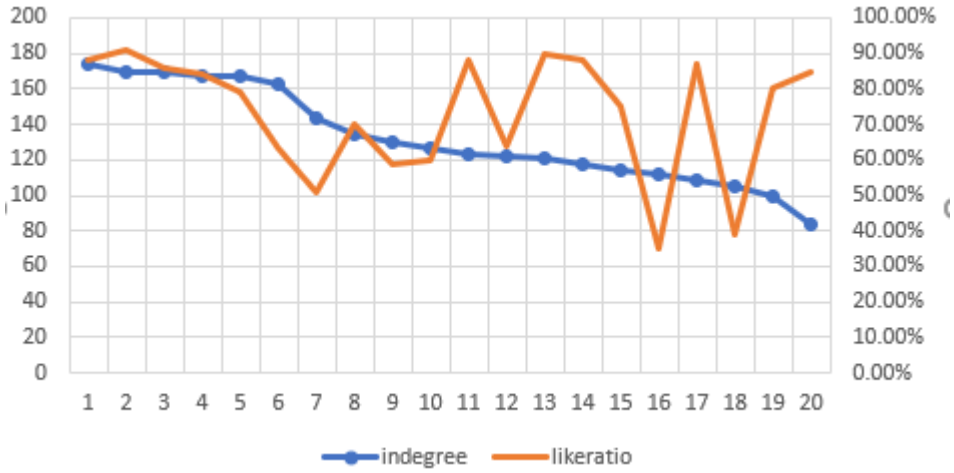
Hypothesis: Like/dislike ratios are hard to measure for Greta Thunberg videos because of how polarizing the topic is and so the ratio bounces when viewers from opposite political spectrums flood dislikes.

```
> mean(Youtube2$likeratio, na.rm = T)
[1] 0.6383673
```

Calculations reveal videos to have a mean of 64% (0.6383673) likes, supporting that hypothesis.

However, it could be that the YouTube algorithm disregards the whole like/dislike system when making recommendations.

But!



It could also be that Greta Thunberg videos firstly attract an audience that does not partake in the system, and secondly Greta and climate change are such topics of contention that we see the top 20 in-degree videos range from 35% to 91% likes, making this attribute hard to consider.

Assignment 2 - #GretaThunberg after 2 weeks

Q1. What are the most frequently used words in tweets about Greta Thunberg and profiles of the users who talk about Greta Thunberg?

1-1. Let's compare the most frequently used words two weeks ago and now. Is there any change? If there is any change, What is likely to be the reason behind it?

Figure 1. Left: most frequent words in tweets about Greta Thunberg scraped about 2 weeks ago, excluding Greta's name (i.e., Greta, GretaThunberg, Thunberg). Right: most frequent words in tweets about Greta Thunberg scraped February 3.

Why is this change happening? As we can see above, two weeks ago, when people talk about Greta on Twitter, they were more focused on climate change and how it affected the whole world. And now, when people talk about Greta on Twitter, they also mention the Nobel Peace Prize and she is nominated except climate change.

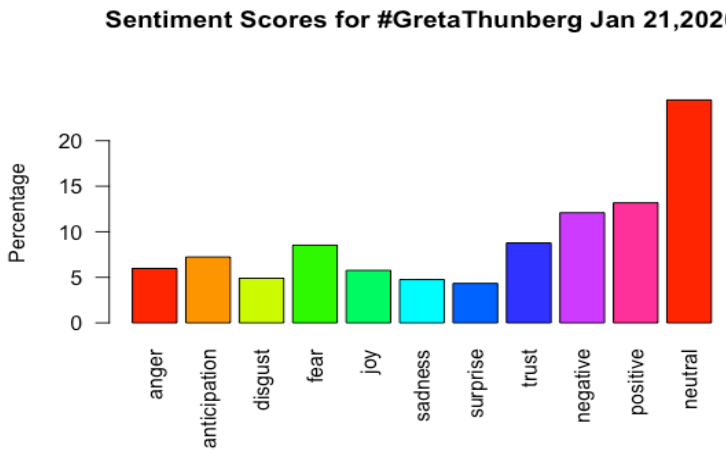
1-2. Can we get insights about who talks about Greta Thunberg based on the words used in their profiles?

Figure 2. Most frequent words in profiles of the users who tweeted about Greta Thunberg.

MAGA stands for "Make America Great Again", KAG stands for "Keep America Great", and WW1WGA stands for "Where We Go One We Go All", all of which are phrases frequently used by right-wing Americans who support Trump. The anti-Greta Twitter users we identified in Assignment 1 are likely to be these people. Words such as "Trump", "NRA(National Rifle Association)", "conservative", "Christian", "patriot", "family", "wife", "American", "country" also show values that right-wing Americans consider important.
"Politics" confirms our analysis in assignment 1 that many people talking about Greta Thunberg are interested in politics, and words like "married" and "retired" impy the older ages of the Twitter users talking about Greta Thunberg. "Supporter" and "fan" show that many users talking about Greta Thunberg like to identify them as somebody's supporters or fans.

Q2. Sentiment analysis and comparison on #GretaThunberg between the week of Jan 21, 2020 and the week of Feb 3, 2020.

Sentiment analysis of dataset from Jan 21, 2020:

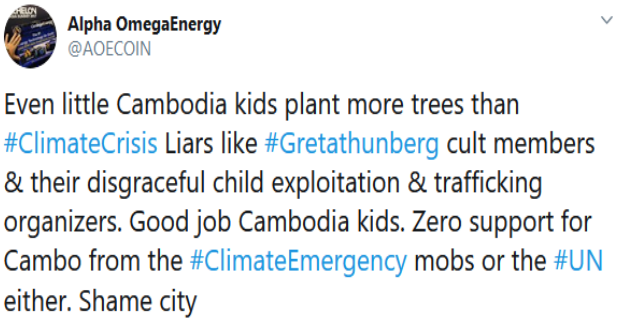


First glance: Sentiment towards topics relating to #GretaThunberg during the end of January looks generally neutral, and evenly polarized between fear and trust; positivity and negativity.

Highly ranked positive tweet about Greta Thunberg

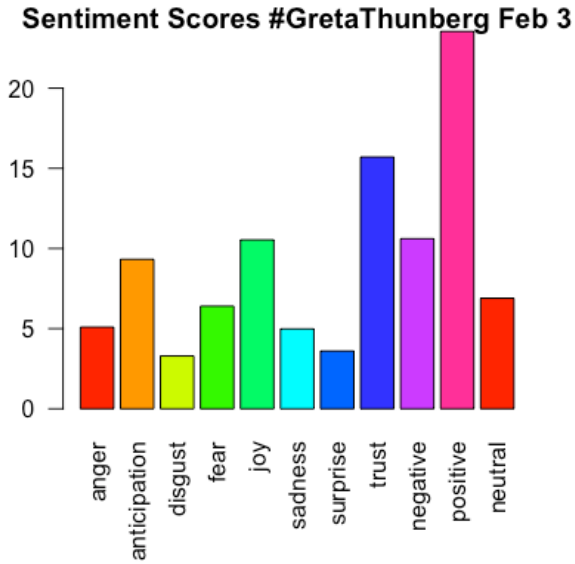
Highly ranked negative tweet about Greta Thunberg

However, as shown above, a deeper dive into opinions about Greta Thunberg and climate change show that they reach both ends of the extreme spectrum. The twitter community is dichotomized between climate change believers and deniers, and also between fans and critics about Greta Thunberg herself. When looking at the results of the sentiment analysis then, it is imperative to be aware that a tweet can be anti-Greta but pro-climate change. Below is a tweet categorized as neutral:



As a result, it can only be concluded that sentiment during the second half of January towards climate change and Greta Thunberg are split, and that sentiment against #GretaThunberg can come up even in pro-climate change tweets.

<u>Sentiment analysis of dataset from Feb. 3, 2020:</u>



Compared to the sentiment analysis from January, one can see that there is a significant increase in positive tweets and trust. The main reason behind this increase is that Greta was nominated for 2020 Nobel Peace Prize between January 21 and February 3, which marks her second nomination after 2019. By diving deeper into the tweets that scored high in positivity, most of them were talking about her nomination.

Highly ranked positive tweet about Greta Thunberg (Feb.3)

However, the following figure shows that some of the tweets that ranked high in positivity and trust don't seem like a real positive tweet. Instead, one can easily tell the irony associated with it. This illustrates the downside of conducting sentiment analysis using codes because it can't irony and sarcasm within the lines. As a result, the increase in positive and trust is not as significant as it shows in the bar chart.

A tweet that scored 5 in trust and 7 in positive