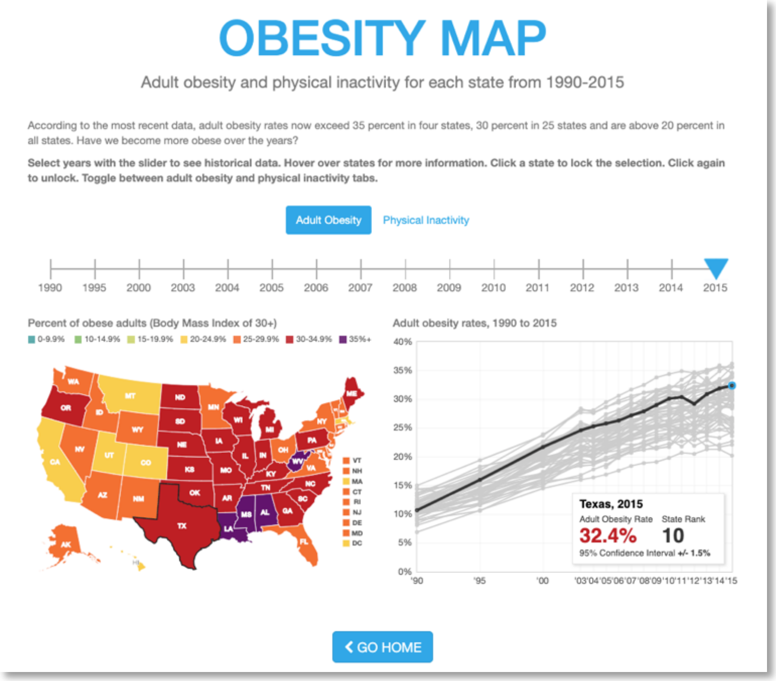
INFO 6010 Final Project
Trust or fear: Interactive Data Visualization in COVID-19 News Articles

Angel Hsing-Chi Hwang
hh695@cornell.edu

“A picture is worth a thousand words.” As cliché as it is, the idiom perhaps best describes the reason why data visualization gains its popularity across the fields. Other than research, business, and other professional uses, we can see across mass media how data visualization has been used to demonstrate complex statistics and information. With the current worldwide pandemic, the ubiquitous use of data visualization has surged to a whole new level, where the form of presentation explains a wide range of topic – from how the virus spread to how our near future may look like.

In this current project, I am interested in investigating how the public react to news articles with data visualization. In particular, I examine a specific form of data visualization – **interactive data visualization**, through which visualized information is combined with on-screen user actions, such as scrolling, clicking, dragging, mouse hovering, and more (Sander et al., 2015). In one of my former projects, we conducted an experiment to see how participants reacted to a website about the issue of obesity in the United States, where we manipulated the level of interactivity in data visualization on the website. We found even with embedding a minimal number of interactive features, participants tended to demonstrate a stronger intent to take actions for resolving or preventing obesity. However, they also perceived the issue of obesity much more severe and reported a higher degree of risks and threats regarding how obesity may affect their life.


OBESITY MAP

Adult obesity and physical inactivity for each state from 1990-2015

According to the most recent data, adult obesity rates now exceed 35 percent in four states, 30 percent in 25 states and are above 20 percent in all states. Have we become more obese over the years?

Select years with the slider to see historical data. Hover over states for more information. Click a state to lock the selection. Click again to unlock. Toggle between adult obesity and physical inactivity tabs.

Texas, 2015
Adult Obesity Rate 32.4%
State Rank 10
95% Confidence Interval +/- 1.5%

Figure 1. A screenshot of the stimulus website in our former project. The website can also be accessed here.

Specifically, interactive data visualization works as an effective way of communication and story-telling due to its higher degree of on-screen **vividness** and a higher demand for **user actions**. In particular, vividness implies how rich, sensory stimulating, and visually appealing a piece of media content is (Ward, 2010). Besides, a higher level of user actions indicates an interactive data visualization inviting the viewers to exploit one or more of the above-mentioned on-screen features to view the visualized data (Murray, 2017). While vivid, rich media content captures individuals' attention during content consumption, their on-screen actions also call for greater degree of user engagement.

Research Questions

Despite a powerful way to illustrate information, data visualization can also serve as a double-edged sword. Considering the unprecedented circumstances, watching numbers tolling up or data points spreading all across our screens can, to some degree, be intimidating. Therefore, in the current project, I look into how the audience reacted to news articles with interactive data visualization by monitoring their content sharing behaviors on Twitter and performing a series of text and sentiment analyses accordingly. Specifically, there are a few questions I am particularly interested in:

- What are the audience's emotional reactions to these news articles with interactive data visualization?
- How are these news and information shared among the audience? Has the information been spread appropriately?
- Does the trends of the audience's sentiment reactions and their engagement correlate with the degree of vividness and user actions in interactive data visualization?

Method

To collect data for the current project, I searched across major news media in the United States for news articles with interactive data visualization. To narrow down the scope for this course project, I only looked into news articles released from March 1st to April 30th, 2020. Table 1 is a list of articles I found. Please note that though there are a much larger number of news posts with data visualization, the use of interaction data visualization is relatively limited, where articles from The New York Times contributed to a dominating number of sources for this project.

Based on each news article, I conducted tweets scrapping based on the following criteria to collect public tweets:

- Because each news article was released on a different date, instead of scrapping by fixed dates, the time range of tweets scrapped is within 7 days after the news article is released.
- Language of the tweets is English.
- The tweets must include the URL link to the news article.

The features I scrapped for each tweet include:

- **handle:** Twitter handle

- **name:** Name of twitter account
- **content:** Text content of the tweet
- **replies:** Number of replies
- **retweets:** Number of retweets
- **favorite:** Number of likes
- **date:** the date when the tweet was posted
- **url:** link to the tweet
- **search_url:** link to filter search and find the tweet
- **hastags:** hashtags used in the tweet.

Table 1. The list of news articles with interactive data visualization used for current study.

News article headline	Link to news article	Remarks on interactive features
This 3-D Simulation Shows Why Social Distancing Is So Important	link	High vividness
How Severe Are Coronavirus Outbreaks Across the U.S.? Look Up Any Metro Area	link	
Where America Didn't Stay Home Even as the Virus Spread	link	
How the Virus Got Out	link	High vividness
Trump Wants to 'Reopen America.' Here's What Happens if We Do.	link	High user actions
Location Data Says It All: Staying at Home During Coronavirus Is a Luxury	link	
How Much Worse the Coronavirus Could Get, in Charts	link	High user actions
Does My County Have an Epidemic? Estimates Show Hidden Transmission	link	
America Will Struggle After Coronavirus. These Charts Show Why.	link	
The Workers Who Face the Greatest Coronavirus Risk	link	
Four Ways to Measure Coronavirus Outbreaks in U.S. Metro Areas	link	
Where the U.S. Stands Now on Coronavirus Testing	link	
Watch How the Coronavirus Spread Across the United States	link	
How Has Your State Reacted to Social Distancing?	link	
Could Coronavirus Cause as Many Deaths as Cancer in the U.S.? Putting Estimates in Context	link	High user actions
These Places Could Run Out of Hospital Beds as Coronavirus Spreads	link	
Why outbreaks like coronavirus spread exponentially, and how to "flatten the curve"	link	High vividness
From 'It's going to disappear' to 'WE WILL WIN THIS WAR' How the president's response to the coronavirus has changed since January	link	High vividness

Italy’s Virus Shutdown Came Too Late. What Happens Now?	link	High vividness
How Long Will a Vaccine Really Take?	link	High user actions
What 5 Coronavirus Models Say the Next Month Will Look Like	link	
260,000 Words, Full of Self-Praise, From Trump on the Virus	link	High vividness
See How the Coronavirus Death Toll Grew Across the U.S.	link	
The social distancing of America	link	High vividness
What 5 Coronavirus Models Say the Next Month Will Look Like	link	
Ventilators: a bridge between life and death?	link	High vividness

Analysis and Findings

In total, I conducted twitter scrapping for 25 news articles with interactive data visualization, collecting 41,732 tweets. Figure 2 shows the total tweets by each article.

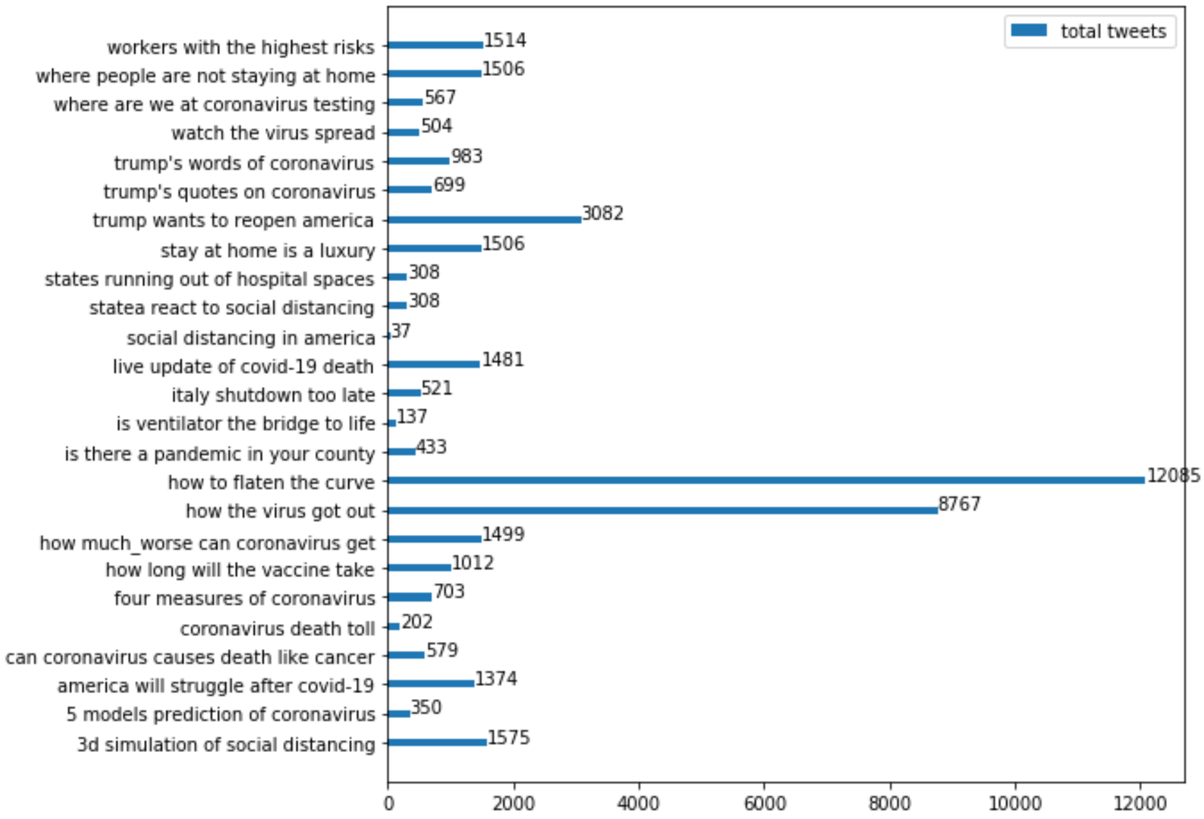


Figure 2. *Total tweets by each news article.*

Data exploration

Taking a quick look at the data, we can clearly see how tweets and their engagement surged during March, when the pandemic outbreak exploded across the country. In comparison, the audience have been much less active in April, even though there were still a considerable number of news articles with interactive data visualization being released in April.

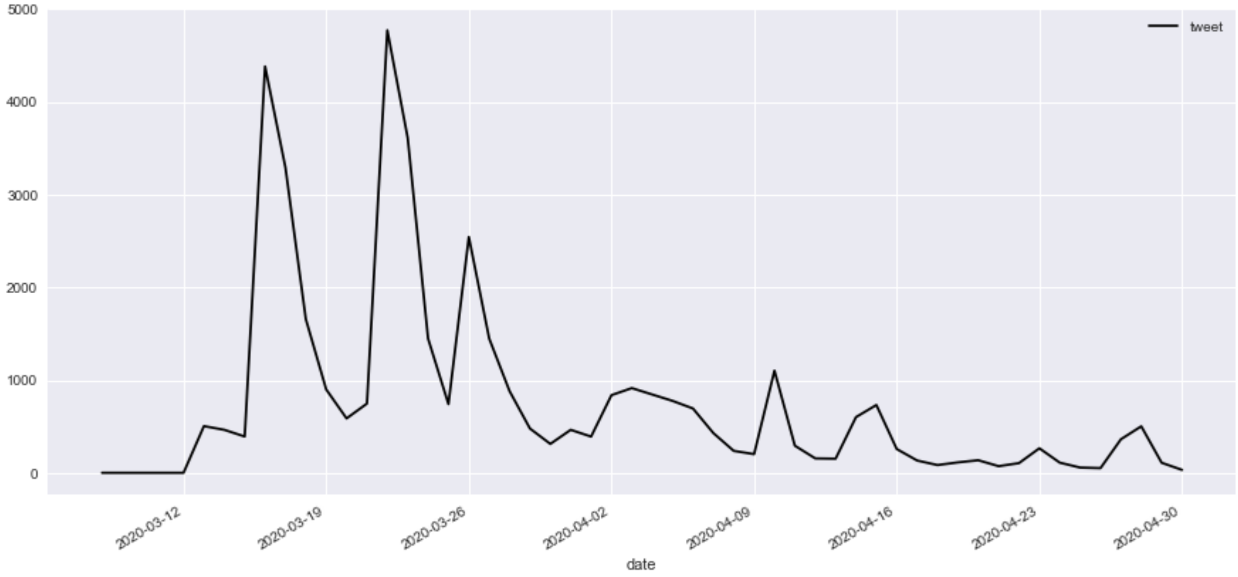


Figure 3. *Total tweets sharing news articles with interactive data visualization from March to April 2020.*

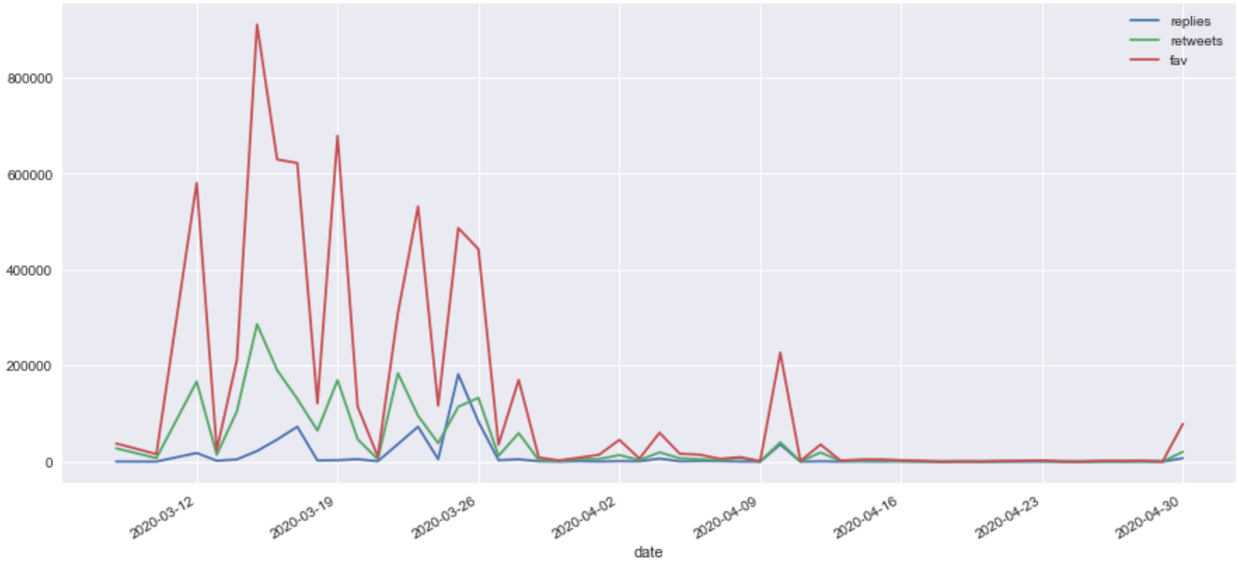


Figure 4. *Total engagement (including retweets, replies, and likes) reacting to tweets sharing news articles with interactive data visualization from March to April 2020.*

Sentiment analysis

The first sentiment analysis I experimented with was nltk vader sentiment intensity analyzer. After analyzing the positive, negative, neutral, and overall compound sentiment intensity of each tweet, I made a plot to examine all tweets' mean sentiments across time. As shown in Figure 5, most of the tweets seem to be composing relatively neutral sentiments, but we do see the positive tones diminished quickly from early March to mid-March, which is when the severity of the pandemic surged in the country (see also Appendix A for the sentiment of tweets of each article).

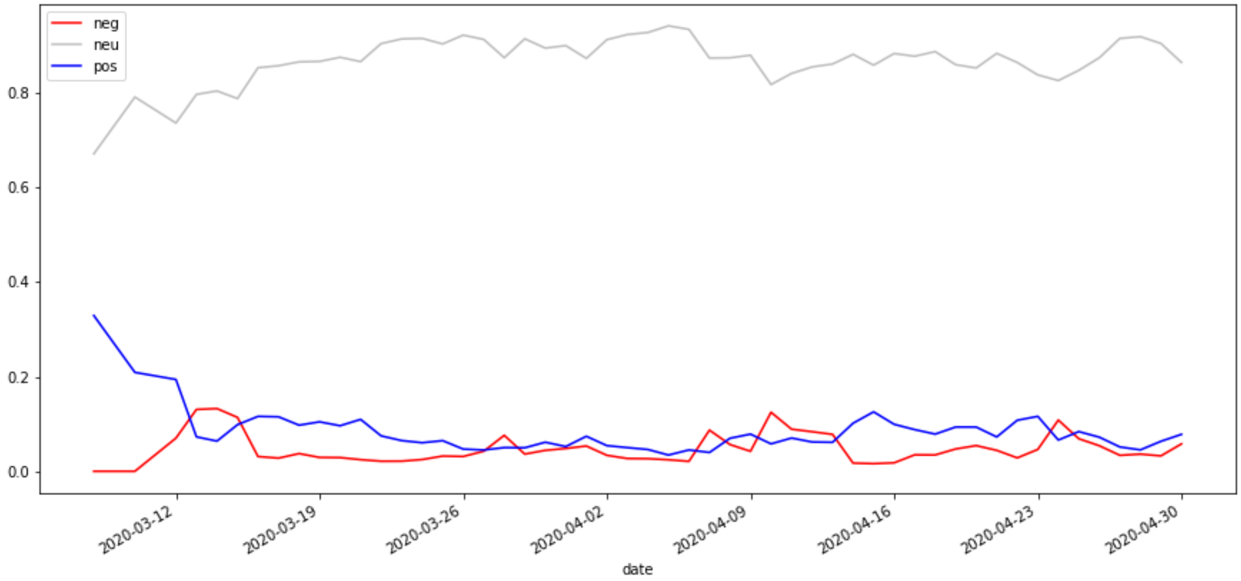


Figure 5. *Tweet sentiment over time.*

On top of that, I looked at whether there is any linear relationship between the amount of engagement versus sentiment intensity. That is, whether more positive or more negative tweets got retweeted, replied, or liked more often, or otherwise. To test this hypothesis, I plotted each tweet by its amount of engagement and its sentiment intensity. First, the results showed that the engagement patterns of positive and negative tweets seem to be fairly similar. Moreover, the linear fit lines (as shown in Figure 6, where orange lines represent fit lines of negative tweets, and yellow lines represent fit lines for positive tweets) and their parameters showed there is no significant relationship between engagement and sentiment intensity.

(Continue on the next page)

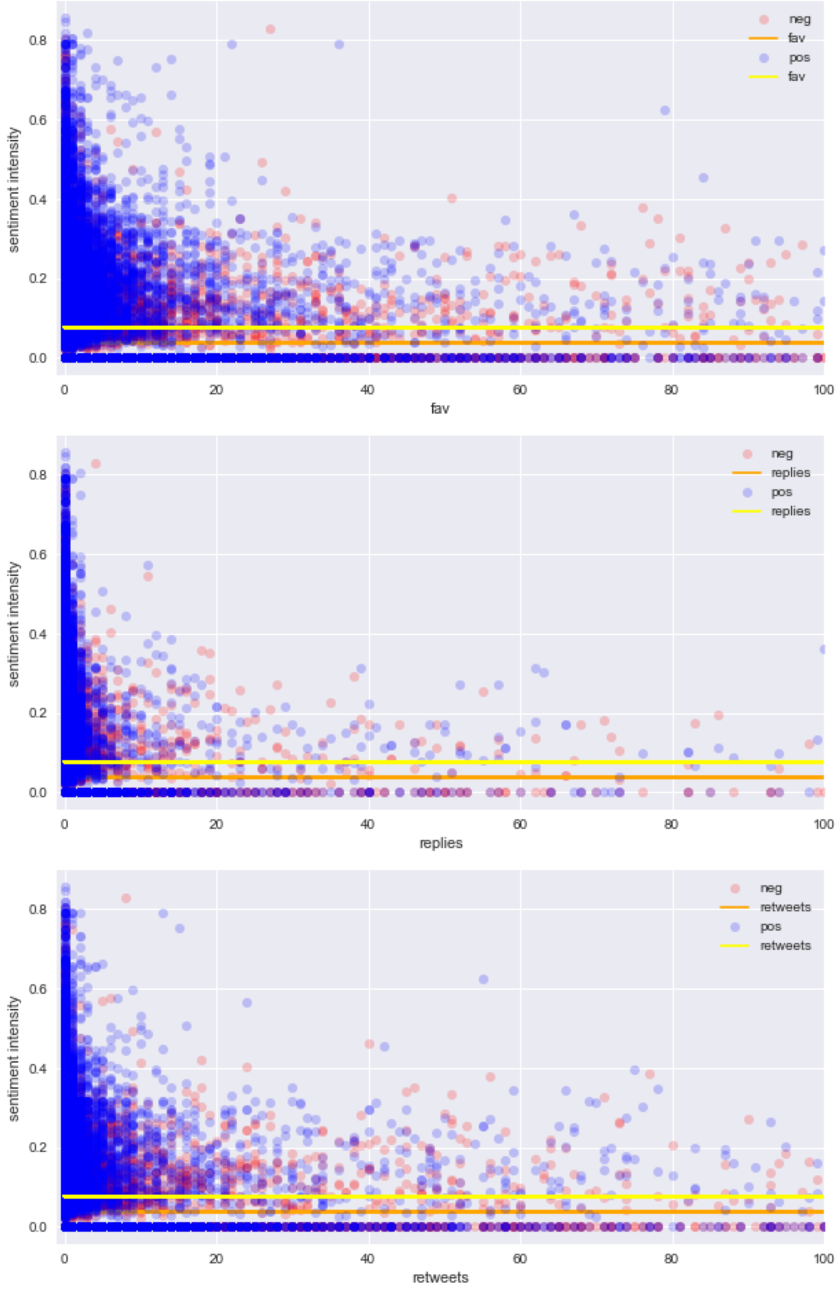


Figure 6. *Scatter plots of tweet engagement (likes, replies, and retweets) versus sentiment intensity of tweets.*

The next, I took a look at the mean sentiments of each article. I noticed a few articles are gaining much more negative sentiments than others. Just by a glance at the headers, we can see almost all these articles addressed deaths caused by coronavirus. By taking a closer look into the actual content of interaction data visualization, I noticed most of these articles are also the ones embedding with a higher degree of vividness (e.g., "is ventilator the bridge to life", "coronavirus death toll") or demanding greater user actions (e.g., "how much worse can coronavirus get", "can coronavirus cause deaths like cancer").

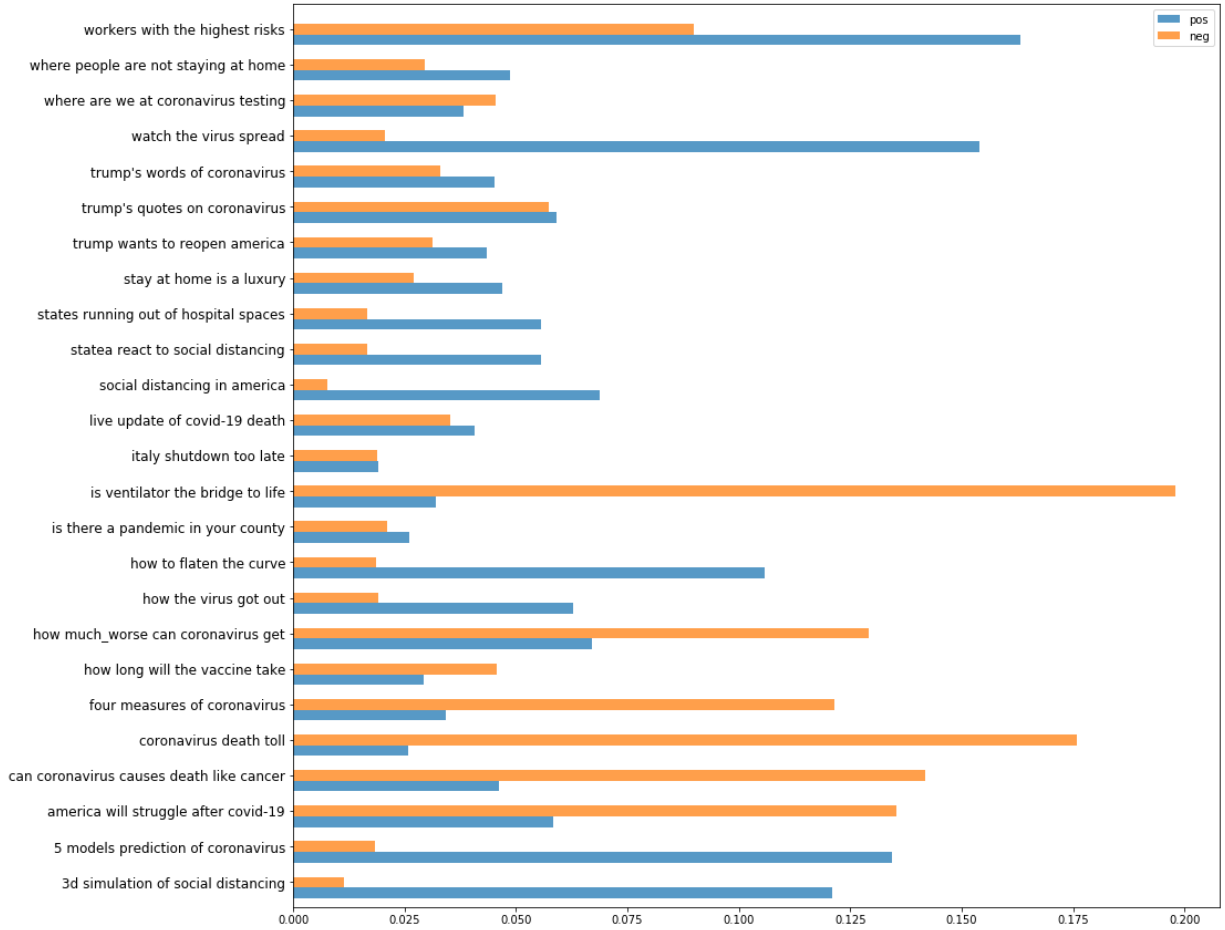


Figure 7. *Average tweet sentiments by news articles.*

Identify tweets with misinformation

In this part of the analysis, I am interested in how the audience shared these news articles – Through their tweets, were they sharing the information accurately? Based on the sentiment analysis above, it seems like overall, the tone used in individuals' posts seemed neutral, but can the language per se be off or misleading? To answer these questions, I refer to a project conducted collectively by the Communication and Computer Science Department at University of Southern California (USC), where they monitor all the tweets about coronavirus from March to early May of 2020[1]. On their project website, they collected an extensive list of tweets which contain words of misinformation. The researchers categorized them into four categories as follows:

- **Unreliable.** This category is defined to include false, questionable, rumorous and misleading news. In addition, we include satire, based on the consideration that satire has the potential to perpetuate misinformation or be used as a cover for misinformation publication (Sharma et al., 2019).
- **Conspiracy**. This category is defined to include conspiracy theories and scientifically dubious news.

[1] For more details about the project by USC Melady Lab, please refer to their preprint on ArXiv: https://arxiv.org/pdf/2003.12309.pdf

- **Clickbait**. This category includes clickbait news i.e. exaggerated or misleading headlines and/or body purposed to attract attention, for reliable and/or unreliable information.
- **Political**. This category includes political and biased news, written in support of a particular point of view or political orientation, for reliable and/or unreliable information such as propaganda.

Since the authors did not release their code for classification, I crawled the entire table (containing all tweets with misinformation) from their project website and attempted to train my own Naïve Bayes classifiers. In other words, using the tweets they collected on their website, I trained four classifiers (i.e., unreliable, conspiracy, clickbait, and political), each can perform binary classification to identify whether a tweet is unreliable, conspiracy, clickbait, political, or not. I choose to do so, instead of training a single multi-class classifier because a single tweet can contain text of more than one category of misinformation (i.e., a tweet can be misleading because it contains text that is unreliable and political at the same time). After removing 50 most common stop words, I trained each classifier by with up to 10,000 most discriminating words. Overall, I was able to bring the training accuracy up to about 90% and testing accuracy to around 80%. Below, Figure 8 shows how the test accuracy rates change along with including more features into training the classifiers.

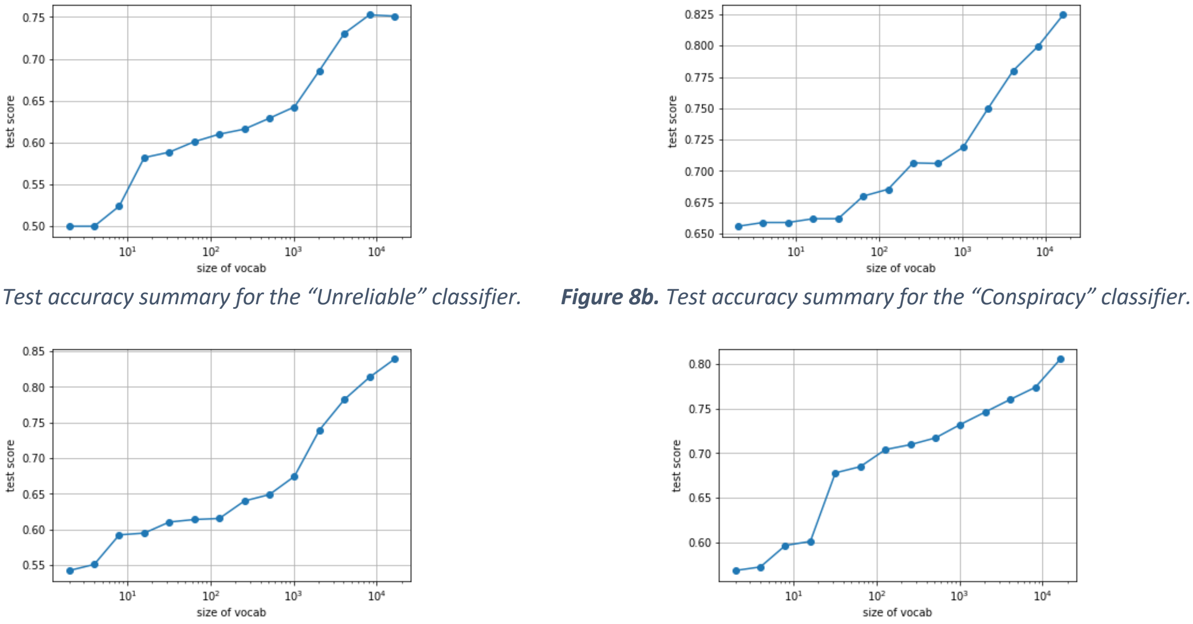


Figure 8a. *Test accuracy summary for the "Unreliable" classifier.*

Figure 8b. *Test accuracy summary for the "Conspiracy" classifier.*

Figure 8c. *Test accuracy summary for the "Clickbait" classifier.*

Figure 8d. *Test accuracy summary for the "Political" classifier.*

Again, I took a look at the overall likelihood of tweets being classified into one of the four misinformation categories for each of the 25 news articles. While articles that tend to be political per se also led the audience use more biased wordings in their tweets (e.g., trump's words of coronavirus, trump wants to reopen America), I noticed among those with the highest likelihood of being labeled as misinformation are also the ones that used interactive data visualization with the greatest **vividness**, such as workers with the highest risks, Italy's shutdown is too late, is ventilator the bridge to life, can coronavirus cause death like cancer, and the 3d simulation of social distancing.

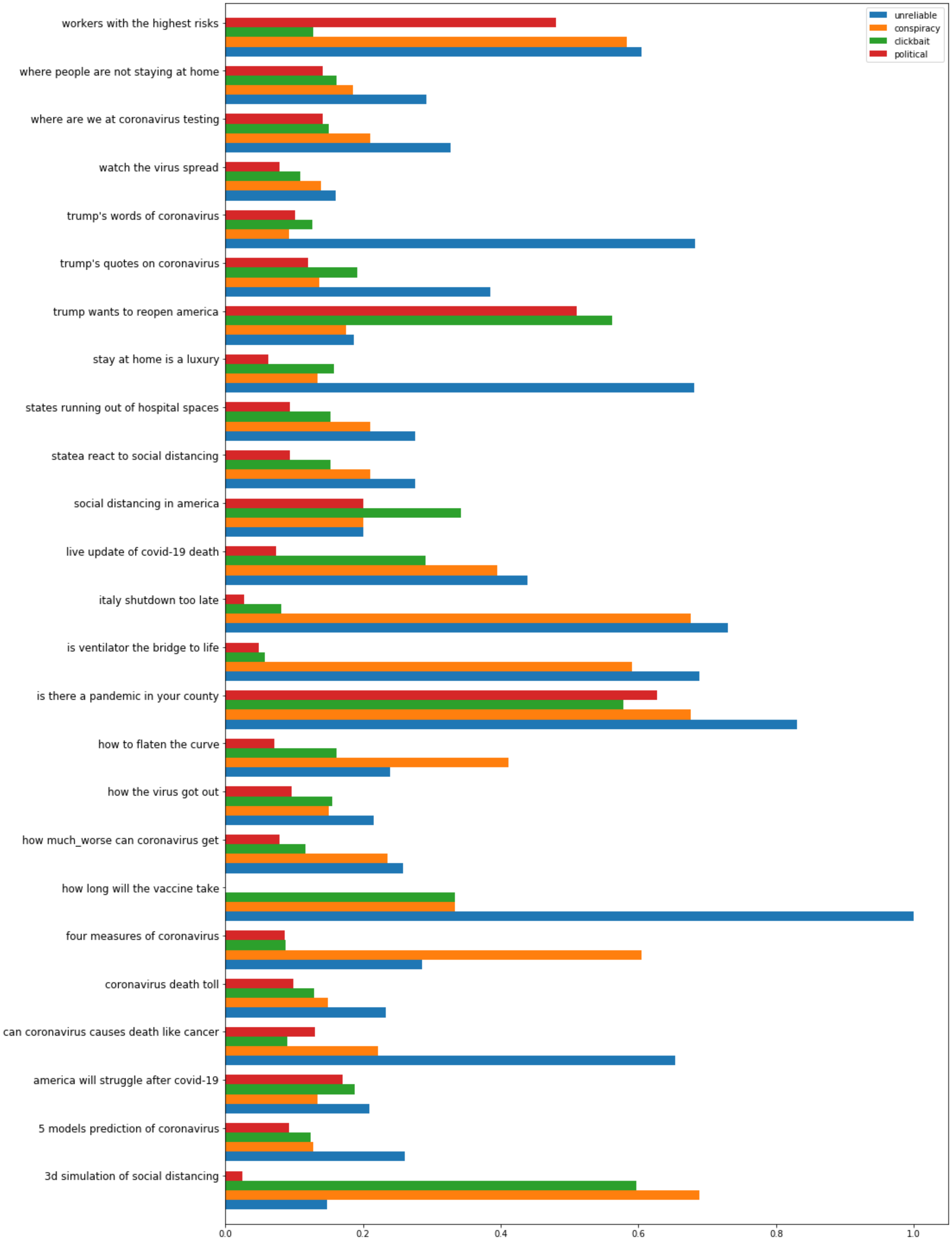


Figure 9. *Likelihood of tweets containing misinformation, by each news article.*

The next, I took a deeper dive into news articles using interactive data visualization with high vividness or high user actions. To accomplish this, I ran another sentiment analysis on the tweets by implementing the NRC emotion lexicon analysis, which gives a more granular view to the sentiments embedded in text. Other than positive and negative sentiments, I am also able to examine anger, anticipation, disgust, fear, joy, sadness, surprise, and trust in the audience's tweets. Figure 10 compares the granular sentiments among tweets sharing news articles with interactive data visualization versus all tweets collected in the social media sentiment project done at USC. Again, since the authors haven't released their code and data, I scrapped all the tweets from their project website. This time, instead of scrapping tweets with misinformation, I did web scraping to get all tweets collected for the project. These tweets are any tweets about the worldwide pandemic. As demonstrated on the project website, the authors categorized them into 21 topics. I did a html web scraping to get all tweets from all 21 categories. As shown in Figure 10, the average sentiment scores of tweets sharing interactive data visualization (in blue bars) are, in general, much lower than the mean of all tweets, with the exception for "anticipation" and "trust." Conversely, when I compare the likelihood of tweets containing misinformation text, tweets sharing news articles with data visualization seem to be more misleading and biased than all tweets about the pandemic in general (see Figure 11).

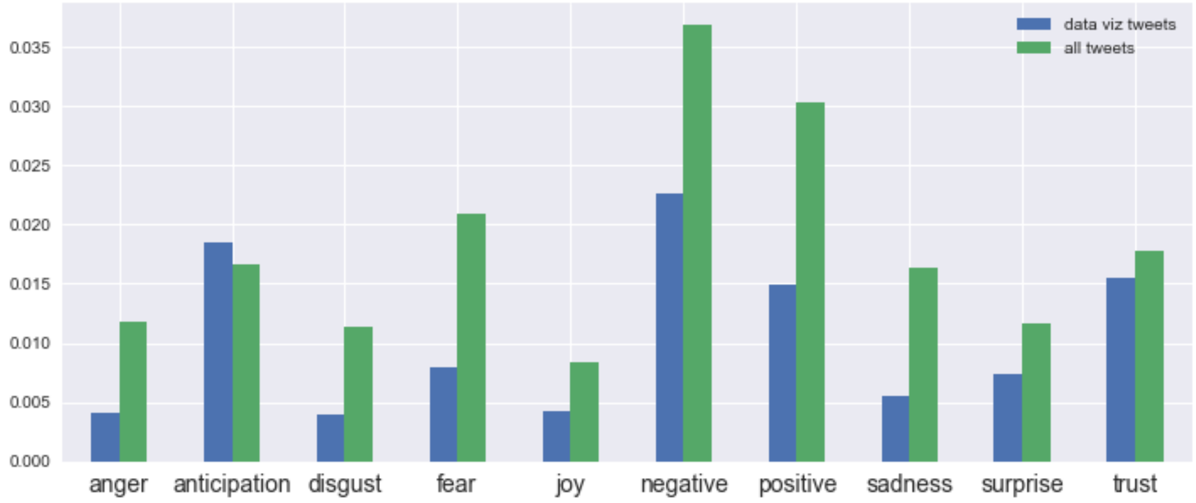


Figure 10. *Emotion lexicon analysis.*

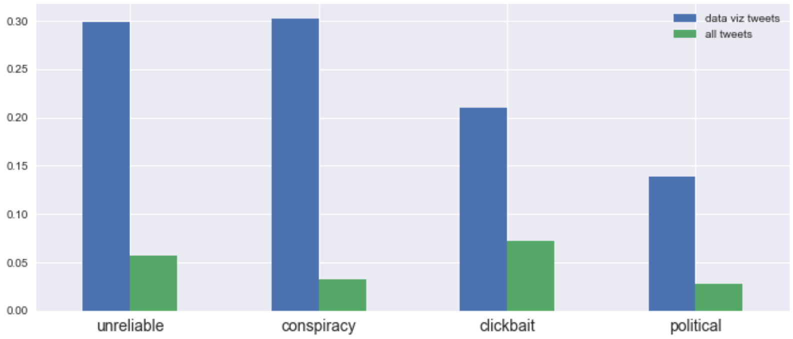


Figure 11. *Likelihood of tweet containing misinformation text.*

With a more granular set of sentiments, I created a confusion matrix for each article, using misinformation categories and sentiments as x- and y-axis respectively, and looked at where most of the tweets fall into. For instance, Figure 12 is a confusion matrix of misinformation categories by all sentiments, using all tweets. By looking at the confusion matrix, I wonder where the anticipatory sentiment came from. In addition, it seems a bit concerning when the bottom row reveals that people who tend to show a lot of trust in these articles are spreading the information with rather biased, unreliable words.

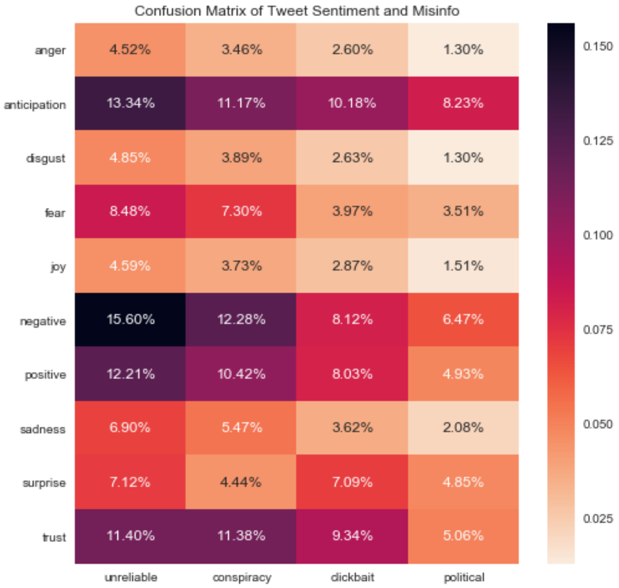


Figure 12. *Confusion Matrix of Tweet Sentiment and Misinformation Categories, using all tweets.*

Looking specifically into articles with high vividness or user actions, I created a similar confusion matrix for each of these articles. Similar to what I found above, when the audience tweet about articles that are more political-oriented, they tend to use more emotional and more unreliable phrasing.

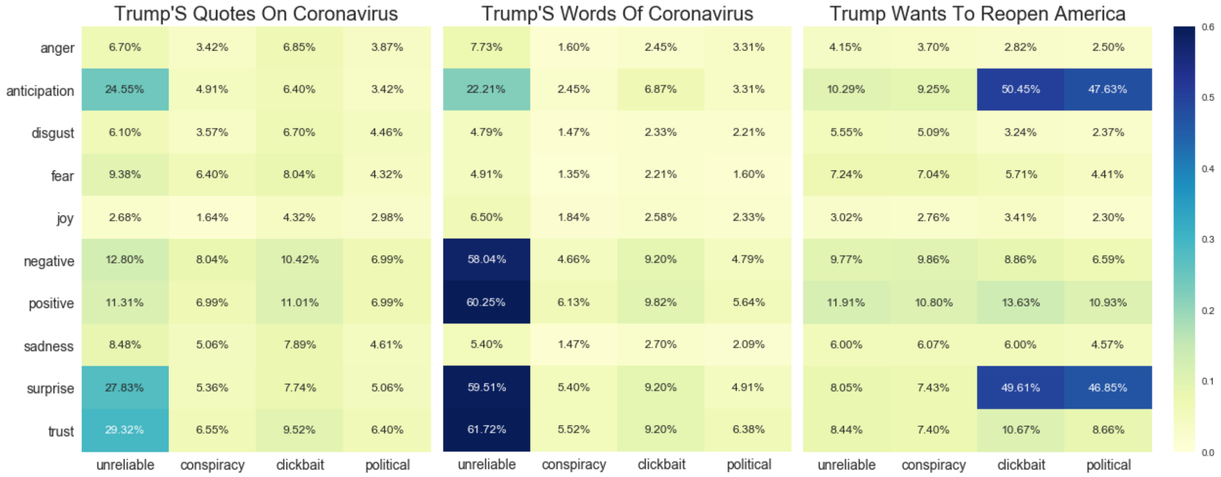


Figure 13. Confusion matrix of tweet sentiments and misinformation categories, using tweets sharing political-oriented news articles with interactive data visualization.

Again, it is concerning to find that people who trust these news articles with interactive data visualization are spreading the information in somewhat misleading tones. Furthermore, I find this pattern arises more apparently when the audience are tweeting about news articles that discuss scientific facts of coronavirus. I wonder if this indicates the general public trust visualized data “too much.” Or in other words, according to former literature (Sundar et al., 2015), when individuals experience highly vivid, appealing media content, the high extent of sensory arousal consumes much of their cognitive bandwidth, leaving them limited cognitive resource to think critically and analytically about the information at hand.

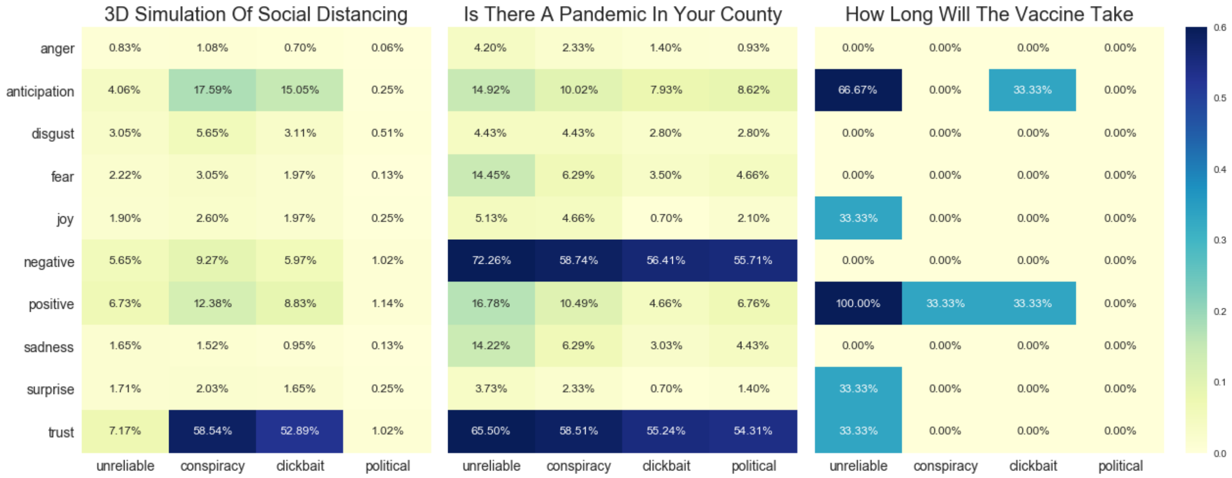


Figure 14. *Confusion matrix of tweet sentiment and misinformation categories, using tweets sharing news articles discussing scientific facts of coronavirus.*

Network analysis: How does misinformation spread?

In the last part of the analysis, I looked at how tweets with misinformation got spread. For each user who shared a news article with interactive data visualization, I used Twitter Developer API to scrape a list of his/her followers on twitter who retweeted the post. Due to time constraints (standard Twitter Developer API only allows scraping 200 followers per 15 minutes), I randomly sampled 400 twitter posts that include misinformation texts (i.e., 100 tweets for each of the four misinformation categories). Using NetworkX (using the user who originally tweeted an interactive data visualization news article as “source” and followers who retweeted the post as “targets”), I generated visualizations of these news articles sharing network and highlighting nodes spreading misinformation posts in Figure 15. I used four different network visualization layouts (i.e., random, circular, forced-directed, and spectral) to better understand the data. Though the results may not have provided a complete view to the data, since I was only sampling a relatively small number of accounts among the entire data set, the “forced-directed” graphs seem to offer the most information. Specifically, they demonstrated that users who shared misinformation tweets tended to spread within their own communities (i.e., larger bubbles in the graphs).

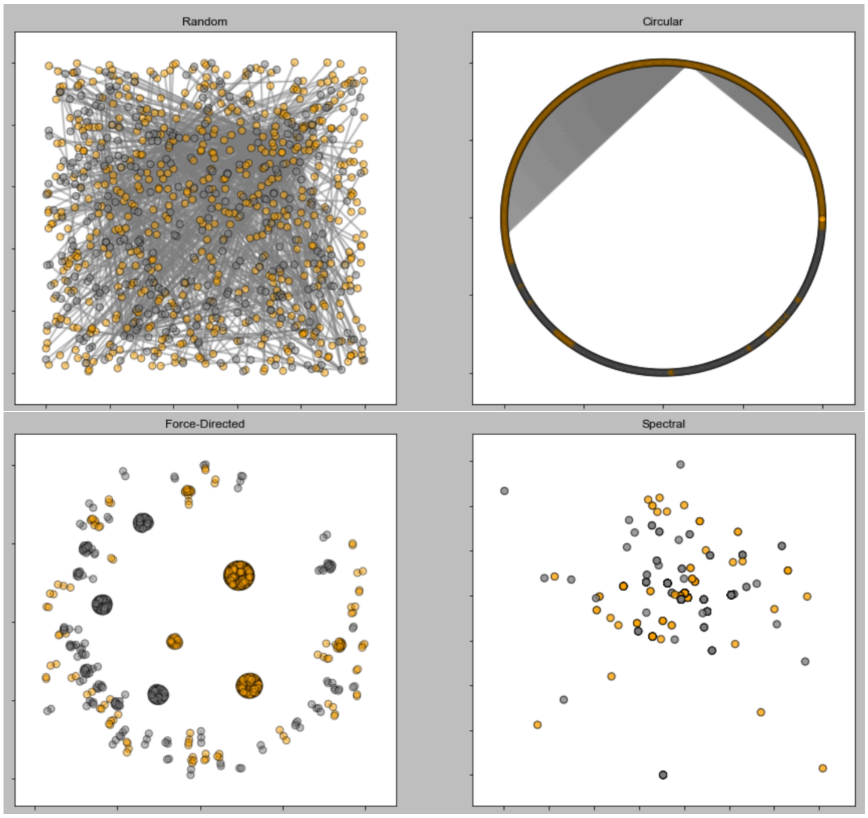


__Figure 15a.__ Network visualization of how tweets with "unreliable" texts were spread.

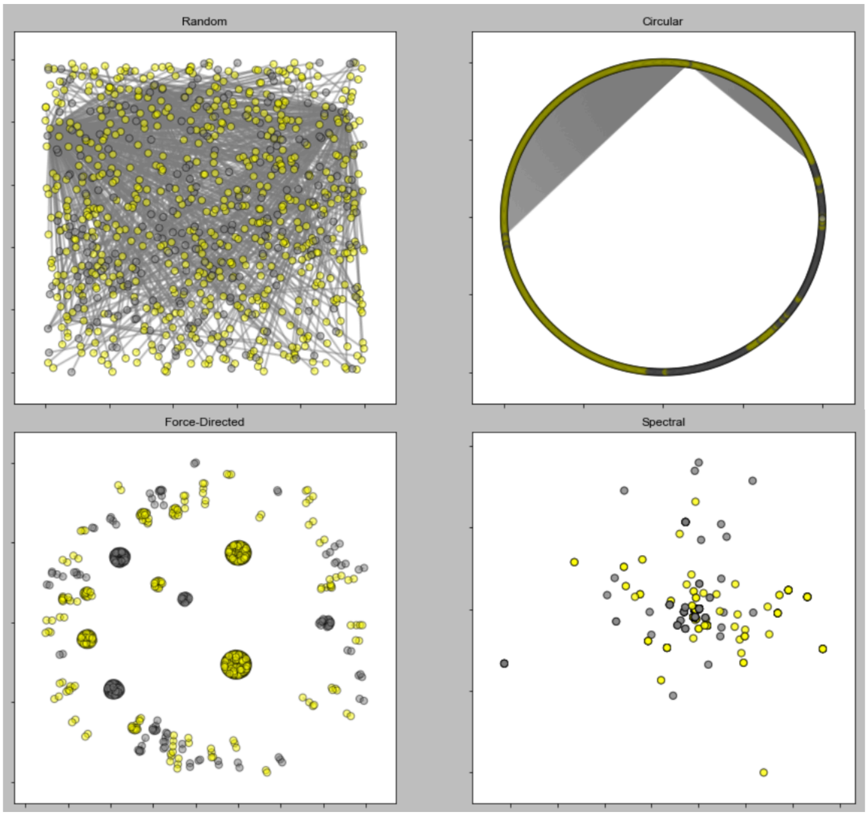


__Figure 16b.__ Network visualization of how tweets with "conspiracy" texts were spread.

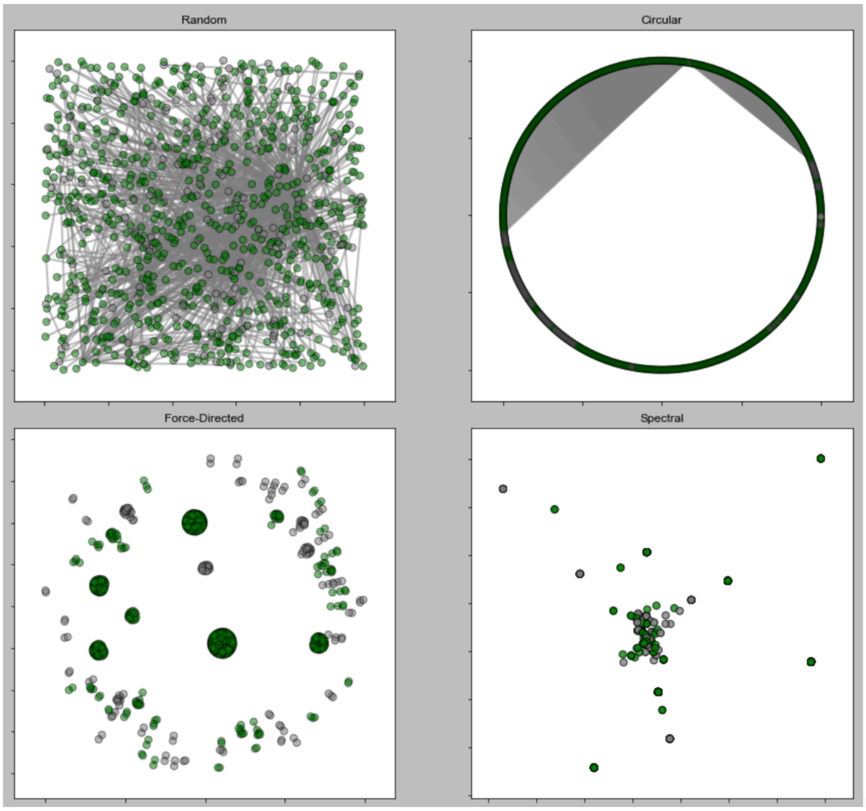


Figure 17c. *Network visualization of how tweets with "clickbait" texts were spread.*

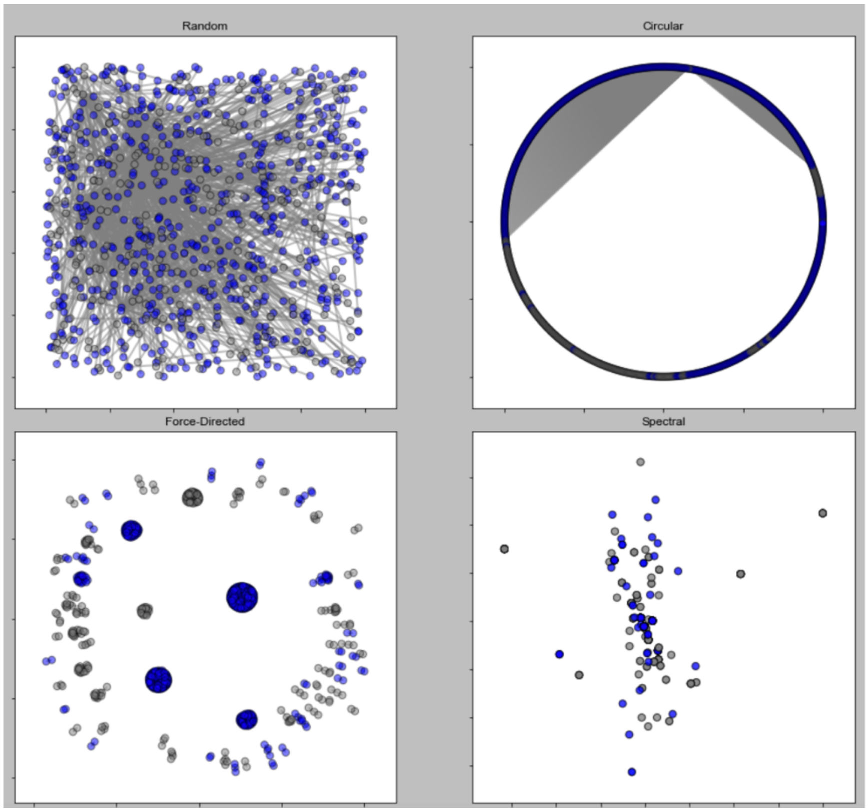


Figure 18d. *Network visualization of how tweets with "political" texts were spread.*

Future Research

The current data exploratory project reveals some trends among the audience's emotional reactions and how they shared these news articles with interactive data visualization. However, it also leaves a lot of open-ended questions to address. Moreover, it is challenging to tell whether some of the results from the current data exploration have more to do with the topics of the news content per se, or rather, the visual effect of the content also made an impact. To further understand how various features of interactive data visualization affect the viewers, I plan to design an online experiment comparing how static versus interactive data visualization illustrating the same topic and same news content may influence how participants react emotionally to information shared in the news. In addition, the current project discovered some concerns in terms of how the general public shared information of visualized data, which will also be further investigated in future research.

Reference

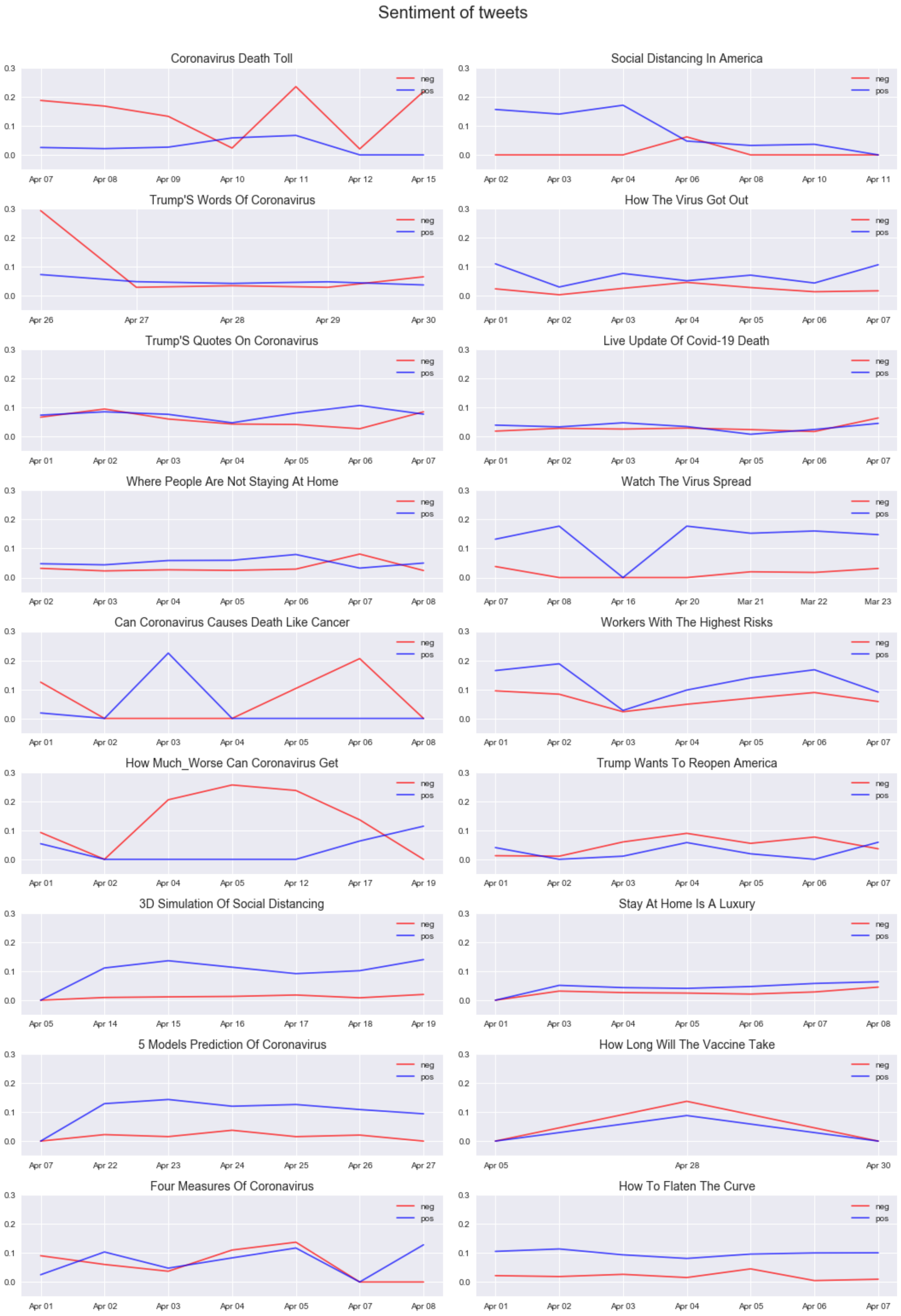
Murray, S. (2017). *Interactive data visualization for the web: an introduction to designing with*. " O'Reilly Media, Inc.".

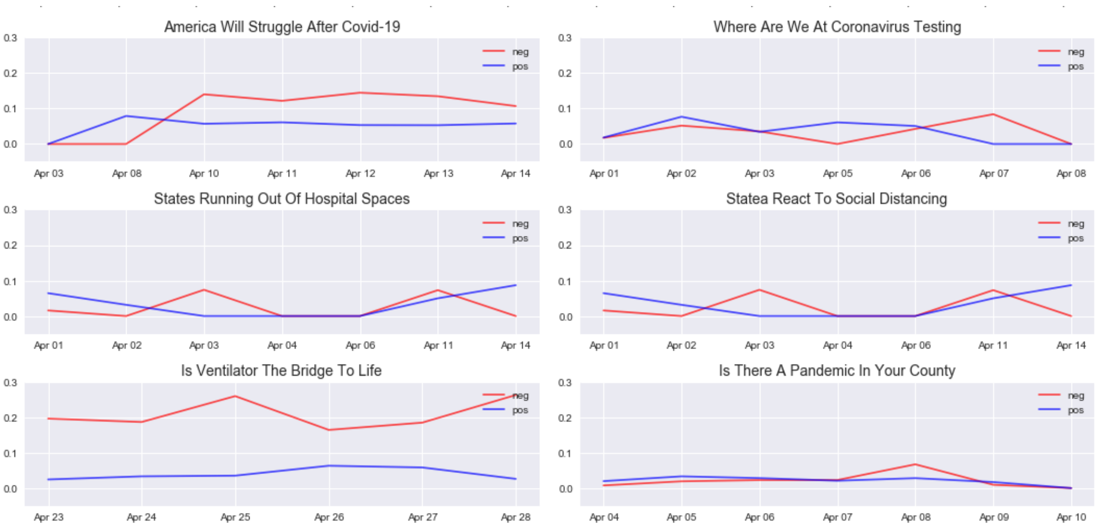
Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1-42.

Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a theory of interactive media effects (TIME): Four models for explaining how interface features affect user psychology.

Ward, M. O., Grinstein, G., & Keim, D. (2010). *Interactive data visualization: foundations, techniques, and applications*. CRC Press.

Appendix A:
Sentiment of tweets sharing each of the articles




America Will Struggle After Covid-19
0.3
0.2
0.1
0.0
Apr 03
Apr 08
Apr 10
Apr 11
Apr 12
Apr 13
Apr 14
neg
pos
Where Are We At Coronavirus Testing
0.3
0.2
0.1
0.0
Apr 01
Apr 02
Apr 03
Apr 05
Apr 06
Apr 07
Apr 08
neg
pos
States Running Out Of Hospital Spaces
0.3
0.2
0.1
0.0
Apr 01
Apr 02
Apr 03
Apr 04
Apr 06
Apr 11
Apr 14
neg
pos
Statea React To Social Distancing
0.3
0.2
0.1
0.0
Apr 01
Apr 02
Apr 03
Apr 04
Apr 06
Apr 11
Apr 14
neg
pos
Is Ventilator The Bridge To Life
0.3
0.2
0.1
0.0
Apr 23
Apr 24
Apr 25
Apr 26
Apr 27
Apr 28
neg
pos
Is There A Pandemic In Your County
0.3
0.2
0.1
0.0
Apr 04
Apr 05
Apr 06
Apr 07
Apr 08
Apr 09
Apr 10
neg
pos