Face off: Polarized Public Opinions on Personal Use of Face Masks during the COVID-19 Pandemic

Jonathan Lai
jlai11@u.rochester.edu

Neil Yeung
nyeung@u.rochester.edu

Abstract

The COVID-19 crisis has reached pandemic proportions. One controversial aspect of the U.S. response to the COVID-19 pandemic response has been the policy towards personal face mask usage. Confusion surrounding the effectiveness of masks, as well as shifting public policy espoused by the CDC and the Surgeon General, have ingrained many Americans with polarized opinions towards face masks. In this study, we seek to track and analyze public opinion in the U.S. towards personal use of face masks in the COVID-19 Pandemic from the dates of March 23 to April 13 by analyzing a COVID-19 Twitter chatter dataset. We utilize sentiment analysis to analyze the polarity of the public opinions. Furthermore, numerous tools are utilized to infer demographics, such as age and gender, from the Twitter data.

Introduction

On March 11, 2020, the COVID-19 disease was officially declared a pandemic by the World Health Organization (WHO, 2020). At the time of this writing, the disease has killed over 3.6 million people with over 250,000 people dead (WHO, 2020). According to the United Nations, COVID-19 is projected to cost 1 trillion dollars to the global economy (UN, 2020).Within the United States, one controversial aspect of the response towards the virus has been whether to allow face masks for personal use. There are two types of masks. The first kind is the N-95 masks. The N-95 mask, fits the NIOSH certification, filters at least 95 percent of airborne particles, allowing particles of only X size (CDC, 2020). The N-95 masks are the only masks that healthcare workers can use to minimize the change of contracting COVID-19 when interacting with patients within close quarters. The second kind of masks include surgical grade masks, reusable masks, and homemade masks, among others. These masks are not NIOSH-Approved; hence, they are not sufficient for healthcare workers who are in close contact with patients on a daily basis. In Asia, face mask usage has a different cultural connotation than within the United States. Face mask usage is considered common courtesy when one is sick, as to prevent others from being sick(Friedman, 2020). Because personal face masks usage is a common practice during times that are not a pandemic, personal face mask usage is largely uncontroversial in Asian countries. During February, the Center for Disease Control and Prevention's (CDC) official policy towards face masks was not to wear face masks for personal use. The CDC was concerned that recommending face mask usage as a policy for non-healthcare works would encourage hoarding behavior —behavior that has been observed towards other essential supplies such as hand sanitizer and toilet paper [Vigdor, 2020] — preventing healthcare workers from acquiring enough N-95 masks. On February 29, the surgeon general tweeted: "Seriously people- STOP BUYING MASKS! They are NOT effective in preventing general public from catching Coronavirus, but if healthcare providers can't get them to care for sick patients, it puts them and our communities at risk!

However, on April 03, the CDC reversed its earlier stance towards personal face mask usage, citing studies that showed the high amount of asymptomatic carriers of COVID-19(Goodnough and Sheikh, 2020). The rationale behind the policy change was that although there is no conclusive evidence that wearing a non-NIOSH-approved face masks prevents one from getting sick, asymptomatic carriers who wear face masks would prevent others from getting sick. The confusion towards the kinds of face masks coupled with shifting governmental policies has led face masks to become a political statement and an increasingly controversial topic (Rojas, 2020)). There have been reported instances of racism towards Asians wearing face masks (Rojas, 2020)). The president has espoused varying perspectives towards the mask(Rojas, 2020). Given the confusion surrounding the public policy, the efficacy of facemasks, and the distinction between N-95 masks and other masks, it is important to measure the perception of the public towards face masks. Furthermore, we would like to examine the change in public perception of facemasks against key events. Our study utilizes Twitter data and sentiment analysis to measure public perception of face masks. From Twitter data, we

infer key demographics such as age and whether the tweeter is a college student. Ultimately, we aim to analyze and track public sentiment towards face masks over time.

Related Work

Sentiment analysis of social media data is a widely-written topic (Rout et al, 2017). For Twitter data specifically, there has been multiple related works on analysis of tweets during other pandemics, such as H1N1 (Chew and Eysenbach, 2010) or Ebola (Kim et al, 2015).In terms of tracking twitter sentiment specifically during the COVID-19 crisis, there has been work on analyzing specific demographics (Duong et al, 2020) and analysis of controversial word usage (Lyn et al, 2020). Our contribution is twofold: we analyze a novel subtopic of COVID-19, face masks, and we implement state-of-the-art deep learning tools such as m3inference to infer demographics from the acquired Twitter data.

Data collection

Our first approach was to attempt to gather tweets using twitter's historical data and continuously streaming the data. We quickly realized that this would not be a viable option due to twitter rate limits and the slow speed at which we were gathering data. As such, we then chose to sample from the Panacea Lab's Covid-19 Twitter chatter dataset for scientific use. Due to Twitter's terms and conditions, full tweet information is not able to be shared, Panacea is only able to release the tweet ids. As such we had to use our hydration scripts to download the full JSON object of each tweet. The data relevant to COVID-19 was gathered starting from March 11th; each day yielded over 4 million tweets. We restricted our analysis to the tweets from March 23 to April 12. In this time period, there are 19 relevant days. For each of these days, we sampled 25 percent of the total tweets for each day due to the size of the JSON object for just one tweet would make it impossible for us to store all the data due to the weak nature of our computers. Then, we performed a regex search within the text section of the data, drawing upon only the texts that mention mask-related terms such as 'N-95', 'mask', or face maskk' and if they were english. After the sampling, we are left with around 250,000 tweets that mention mask-related terms and are dated from March 23 to April 12.

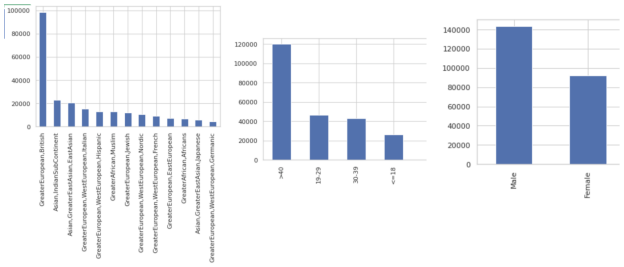


Figure 1: Demographic

Extracting Age and Gender Attributes from Twitter User Profiles

Age and Gender are two demographics that we believed would have different sentiment toward facemasks during the Covid-19 epidemic due to the nature of the disease. Covid-19 is especially dangerous amongst the older population and should not be taken lightly as it has a higher fatality rate amongst weakened immune systems. As such we attempted to infer both Age and Gender for each Twitter User. This was done these attributes are extracted using the M3 (Multilingual, Multimodal, Multi-attribute) deep learning system for inferring the demographics of users from four sources of information from Twitter profiles: user's name (first and last name in natural language), screen name (Twitter username), biography (short self-descriptive text), and profile image [Wang,2019]. M3's accuracy rate is greatly improved if we are able to use the full model involving both image and natural language identifiers. As such during the process of retrieving the profile images of any individual, if they had a faulty profile image link ,invalid of picture format, or had deleted their profile. They were then excluded from our analysis. The full M3 Model is shown below in a picture with a distribution of resulting demographics. A large majority of our population drawn came from the Male, ¿40 category when looking at users who tweeted about Coronavirus. From our data set with 243994 users. We were able to identify 143525 users as male and 92481 as female or a total of 236006 users. With 62.2 percent male and 37.8 female. This statistic is highly consistent with the website statista.com results of 61 percent of users were male and 39 were female. Updated as of April 2020. The age demographics drawn from M3 implies 59.8 percent of users in our data set was over the age of 40. This is against our natural intuition as we expect most twitter users to be at a younger age. As such we move forward with the assumption this is because our topic is a highly specific and controversial topic that poses a greater risk toward older members of society. As such they are more vocal on their opinions.

Sentiment Analysis

Text preprocessing steps include tokenizing the tweets in order to remove hashtags, @, Usernames they were referencing, and attempting to fix incorrectly spelled words using work based on Peter Novrig's in the form of py spellchecker. This uses a Levenshtein algorithm to find permutations within an edit distance of 2 from the original word. It then compares all permutations (insertions, deletions, replacements, and transpositions) to known words in a word frequency list. Those words that are found more often in the frequency list are more likely the correct results. But this does not ALWAYS produce the correct spelling for a word. This is however, very important and accurate when correcting cases where words such as "happy" are exaggerated to the form of "happyyyyy" and other cases similar to this. As not only are most sentiment analyzer models unable to differentiate the second case as being a variation of the first but also the overall compound score of the sentence goes down as it contributes nothing to the sentiment score. For sentiment analy-

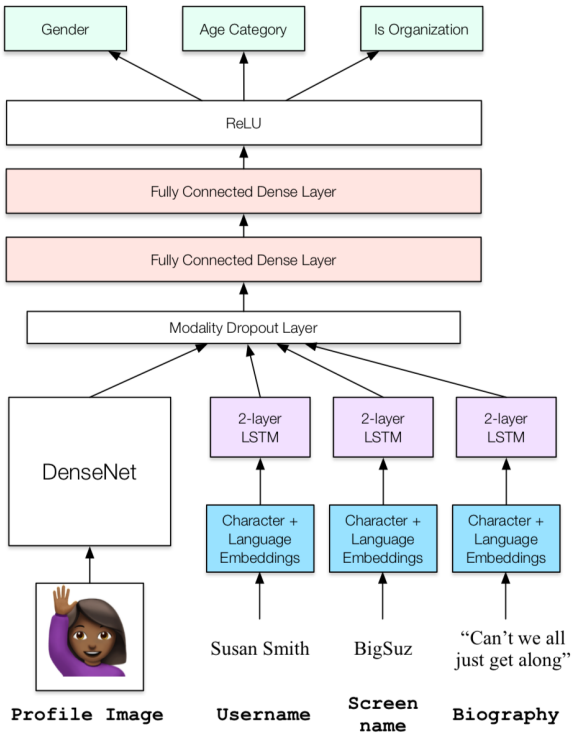


Figure 2: Example Pipeline of M3 Model

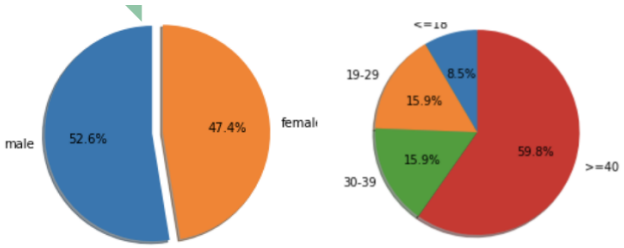


Figure 3: Demographics predicted with n=236006 predicted

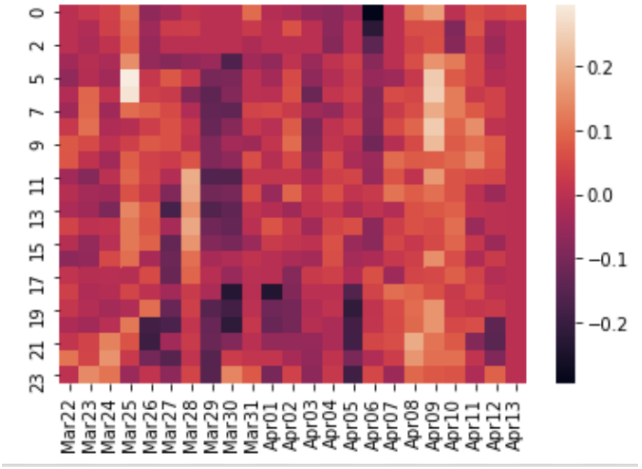


Figure 4: Heatmap of average sentiment by hour of day

sis of the twitter tweets. We use the VADER model for processing text and predicting valence scores. (Valence Aware Dictionary and Sentiment Reasoner. This is a model that is specifically attuned for analyzing the sentiment of social media data (Hutto, C.J. Gilbert, E.E. (2014) using a gold standard of tweets rated by 20+ raters to train the model in using LSTM architecture. Many other different sentiment analysis models that also use NLTK recommend removing emojis and removing uppercase and lowercase. However, VADER sentiment is attuned to these effects and adjusts its rule based compound score accordingly to these features. (See corresponding presentation slides). Vader sentiment produces a compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive)((Hutto, C.J. Gilbert, E.E. (2014). As such, it is appropriate to refer to this as a normalized weighted composite score. Typically for classification of positive and negative. Any text that has a compound score of ¿.05 is considered positive. ¡-.05 is considered to be negative and anything else can be considered to be neutral.

College Student Classification

College students and younger people are a demographic that often does not display the best intentions when preventing the spread of the coronavirus. This is evidenced by the fact that even with the spread of coronavirus. Many students still chose to travel to popular spring break locations. Florida beaches. New York city, and other attractions. A perfect place for the coronavirus to spread. As well as many choose to display poor judgment with the justification that even if they get the coronavirus they most likely won't suffer fatally from it. As such, we wanted to see if College students had any noticeable difference in their sentiment toward the term of face masks. To conclude if students were a college student or not we first attempted to gather a gold standard balanced set of training data. 250 students that were identified as college. And 250 of non college students. This was done

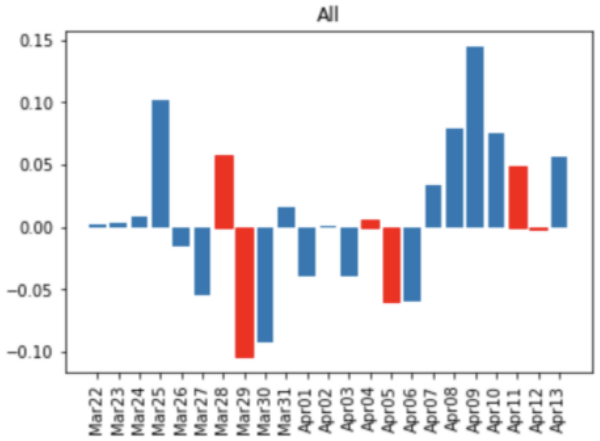


Figure 5: Average Sentiment by Day (Easier Visability

using hand labeling between two different raters. For each twitter user that they observed, if both raters reported the same score for that user. That rating was kept, else it would be discarded. This gold standard of labeled tweets could be improved by in the future outsourcing the labeling to popular sites such as qualtrics and only recording the label for a specific user if they had say a 80 percent success rate. After our set of training data had been identified. We discovered from Bergsma and Van Durme paper on characterizing social media users. They discovered a user most often reveals information not only in their profile picture, bio graph, and other twitter user object fields. But also to their tweets and the things that they reference. Specifically in the use of possessive objects. Such as commonly using phrases using my x where x refers to an item. Viet Duong's paper expanded on this idea and found that certain attributes in college students do in fact reveal themselves as being college students. (Duong, 2020). Using his list of attributes and their pointwise mutual information[Bergsma and Durme, 2013] to obtain our own list of distinguishable attributes we moved on to our next step. As such the next step was to download the full timelines for each user from twitter API. (JSON object of at most 200 tweets). Due to the nature of twitter rate limits and the fact that twitter has blocked the authors of this paper from applying for any new accounts. As well as the fact that obtaining 20k user-timelines already took up 30 GB of space. We chose to sample a subset of our original data and produce college predictions. Due to our initial data set being heavily unbalanced with more users being in the ¿40 group. We took a random sample of 6.25 tweets from each of the age bins. This was where our training data was drawn from. We then ran a 80/20 train,test split on our training data and used a tf-idf vectorizer to extract relevant terms of the timelines for the training data. Then this was run through a random forest classifier using our training labels to determine and predict class labels for if they were a college student or not. Using our list of x attributes. Users that displayed a high frequency of these x terms in the corresponding spot. Such

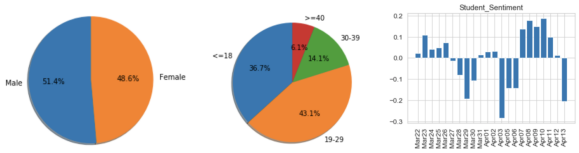


Figure 6: The following graphs are the male and female distribution as well as the student sentiment over time. Note the negative sentiment from April 03 to April 8 and the subsequent positive spike from April 08 to April 11. n = 4397

as "professor" and "textbook". My Final, my test. Etc had their labels manually overridden to be college students.

Race

Race was attempted to be inferred from the user specified name category of the twitter user. Many of the users had names such as John Smith. In this case, it is reasonable to infer that the name of the user is John Smith with first name John and last name Smith. We used the package human names from the parent package name parser to do our best guess at splitting up the name field into a first and last name in the case that a name was not so simple in the case as John Smith and to take care of excess words such as John "The Goat" Smith. Again, an example is referenced in the presentation, After a first and last name prediction was obtained, we used ethnicolor's tensorflow's implementation of the full wikipedia model database of name to predict race.(Laohaprapanon and Sood) based on the first and last name of the user. Or just the last name. This method's accuracy is highly dependent on the assumption that users did not pull troll categories for their name field such as 'sdfsdfdsfdsf'. Our method would try to predict name from this nonsensical string. However, a cursory research from our results seems to confirm that there were not a drastic amount of names like this. Most of our users used their real names or some sort of nickname variation of their names.

Location

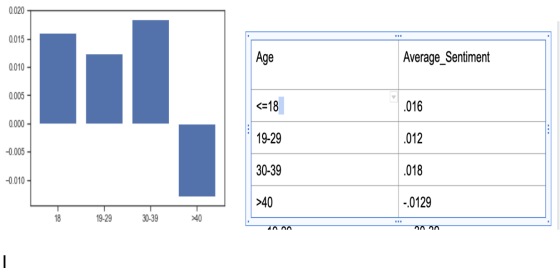
Around 6 percent of the gathered twitter data had a non-empty location that was a city within the United States. We searched for the zip code associated with this location through using the package us zip code. From this package, we were able to glean other relevant data, such as median household income. Then, we classified each county according to a 2013 Rural-Urban-County Classification scheme (RUCC, 2013). The RUCC classification scheme assigns each county in the country a number from 1 - 9. Numbers 1 - 3 are considered metro areas and Numbers 4- 9 are non-metro areas. Our data distribution skewed towards urban areas, so conclusions drawn from the data, must be contextualized with this fact.

Figure 7: We plotted the average sentiment of the West Coast vs. East Coast, Metro vs. Non-Metro zip codes, and median household income average sentiment.

Experiment

Age

It has been shown that Twitter users skew young (Omnicore, 2020). However, in the majority of the tweets we sampled, the 40 and over age demographic was the most prevalent. We can see this in how 59.8 percent of users collected were estimated to be over ¿40. This was also the only age group that had a sentiment lower than 0. As such we can speculate that the 40 and over age demographic disproportionately tweets about face masks, and the coronavirus in general at a more negative level than other age groups. These results are further supported when completing an ANOVA test. An analysis of variance test is used to measure within group differences for a population. In this case age, with a test statistic value of 2.27e-39, we can conclude that there is a statistical difference in populations of the age groups.



Age	Average_Sentiment
<=18	.016
19-29	.012
30-39	.018
>40	-.0129

Figure 8: Age

Gender

Gender Previous surveys of Twitter demographics have found that around 60 percent of Twitter users are male (Omnicore 2020); the gender distribution in our data is similar. Females were found to have a more negative average sentiment when compared to males. Females had a sentiment of -.01377. While Males had a sentiment of .0105 Female average sentiment patterns mirrored male average sentiment patterns. This suggests that although females had higher range, the fluctuations of various policies had the same effect (i.e. positive or negative) on both gender's average sentiment.

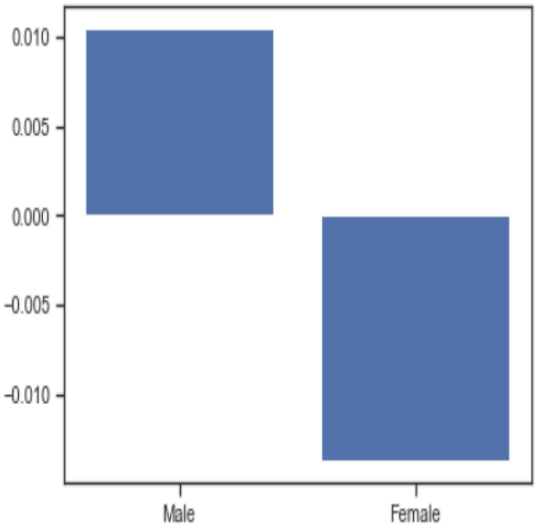


Figure 9: Visual Representation of Gender

Location

When comparing the top 3 West Coast states and the top 3 East Coast states in terms of yield, it can be shown that West Coast states had a higher average sentiment while East Coast. More rural counties had a more negative opinion towards face masks, while urban environments had a higher positive sentiment. We can speculate that face masks usage was more of a concern in metro environments due to high population density in these areas. The higher the median household income, the more negative opinion one has towards face masks. In terms of an explanation for this phenom, there does not seem to be a sound explanation for this pattern that we observed.

College

College students were affected greatly by the shift in CDC policy on April 03. College students showed a significant increase in average sentiment for the few days after Apr 03 when compared to non-college individuals. It can be speculated that college students react differently based on the data for each demographic. For example, in the over 40 subdivision of college students, sentiment was the most positive. However, within the wider population, the over 40 demographic was the most negative. This suggests that college students have reactions to governmental policies that may deviate with the broader American population.

Demographic	Avg sentiment
Female	0.009533003442340862
Male	-0.0032300315173345432
18	-0.0031331783846600767
19-29	-0.009029599121361876
30-30	0.06739068047337277
40	0.049461556603773674

Figure 10: A table of Average College Sentiment

Latent Dirichlet allocation

We ran latent dirichlet allocation (LDA) to observe topic similarities within the data set. It was found that running LDA with 6 topics yielded the highest coherence. The following table, word clouds, and graphs show the rough topic groupings.

Topic	Description
0	China, Masks Supplies from China
1	Work-related issues
2	March, Government, Shipment of Masks
3	Trump, 1 million masks, PPE
4	Home vs. Public
5	Positivity

Figure 11: Topics

Race

In terms of race, all Greater European races displayed lower average sentiment than every Asian race. This suggests that the cultural difference in face mask usage may apply to Asian Americans. However, we observe that after April 03, average sentiment each day of Asian and non-Asian races converge, suggesting that the April 03 policy had a significant effect in potentially alleviating racially-motivated tensions related to face mask usage.

Future Works

As pandemics become more common in an increasingly connected world(Bar-Yam, 2020), consistent public policies are essential for managing pandemics. Social media data mining and sentiment analysis have the potential to be important tools in measuring public response to policy changes during pandemics. More governments and health organizations need to leverage social media data mining technologies to improve policy outcomes.

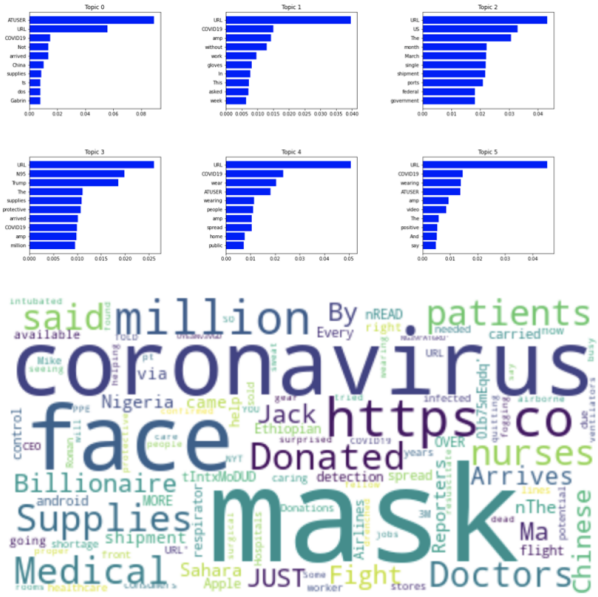


Figure 12: LDA and Word Map

References

Wang, Zijian and Hale, Scott A. and Adelani, David and Grabowicz, Przemyslaw A. and Hartmann, Timo and Fabian and Jurgens, David. "Demographic Inference and Representative Population Estimates from Multilingual Social Media Data". WWW '19: The World Wide Web Conference (2056 - 2067). May 2019. https://doi.org/10.1145/3308558.3313684.

World Health Organization. Coronavirus Situation Report. April 23 2020. Accessed May 5, 2020. Friedman, Uri. "Face Masks Are In". The Atlantic. April 2020. Accessed May 5, 2020.

Vigdor, Neil. "Tennessee Brothers Who Hoarded Hand Sanitizer Settle to Avoid Price-Gouging Fine". New York Times. April 22 2020. Accessed May 5, 2020.

Goodnough, Abby and Sheikh, Knuvul. "C.D.C Weighs Advising Everyone Wear a Mask". New York Times. March 31, 2020. Accessed May 7, 2020.

Bergsma, Shane, and Benjamin Van Durme. "Using conceptual class attributes to characterize social media users." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013

Duong, Viet, et al. "The Ivory Tower Lost: How College Students Respond Differently than the General Public to the COVID-19 Pandemic." arXiv preprint arXiv:2004.09968(2020).

Centers for Disease Control and Prevention. "NIOSH-Approved N95 Particulate Filtering Facepiece Respirators". Updated April 22, 2020. Accessed May 5, 2020.

Chew, Cynthia, and Gunther Eysenbach. "Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak." PloS one vol. 5,11 e14118. 29 Nov.

```
{0: 'GreaterEuropean,WestEuropean,Germanic',
 1: 'GreaterAfrican,Africans',
 2: 'GreaterEuropean,British',
 3: 'GreaterEuropean,WestEuropean,Hispanic',
 4: 'GreaterEuropean,EastEuropean',
 5: 'Asian,IndianSubContinent',
 6: 'GreaterEuropean,Jewish',
 7: 'GreaterEuropean,WestEuropean,French',
 8: 'GreaterEuropean,WestEuropean,Nordic',
 9: 'Asian,GreaterEastAsian,EastAsian',
 10: 'Asian,GreaterEastAsian,Japanese',
 11: 'GreaterEuropean,WestEuropean,Italian',
 12: 'GreaterAfrican,Muslim'}
```

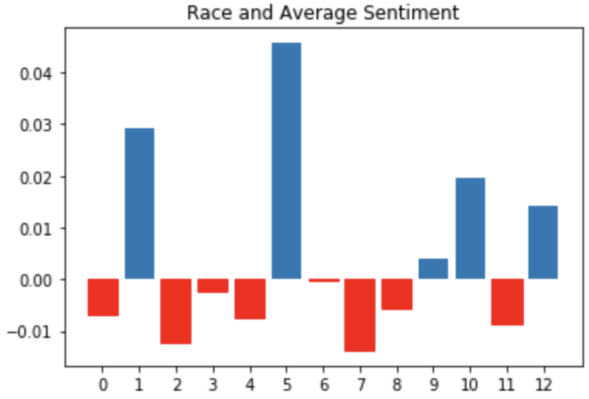


Figure 13: European users have been labeled red

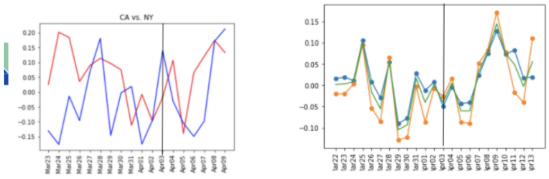


Figure 14: Red = CA, Blue = NY — Yellow=Female,Blue = Male , Green= All

2010, doi:10.1371/journal.pone.0014118

Rout, J.K., Choo, K.R., Dash, A.K. et al. A model for sentiment and emotion analysis of unstructured social media text. Electron Commer Res 18, 181–199 (2018). https://doi.org/10.1007/s10660-017-9257-8

Kim, Hea-Jin Jeong, Yoo Kyung Kim, Yuyoung Kang, Keun Song, Min. (2015). Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. Journal of Information Science. 42. 10.1177/0165551515608733.

Lyn et al. “The Ivory Tower Lost: How College Students Respond Differently than the General Public to the COVID-19 Pandemic”. Arxiv preprint. https://arxiv.org/pdf/2004.09968.pdf. 21 April 2020.

Hutto, C.J and Gilbert, Eric. “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”. AAI Conference. 2014.

Cromartie, John. “Rural-Urban Continuum Codes”. United States Department of Agriculture Economic Research Service. 2013. Accessed May 1, 2020.

Yaneer Bar-Yam, Transition to extinction: Pandemics in a connected world, NECSI (July 3, 2016).