Data-driven Insights into Pandemic Fatigue

Introduction

Almost a year since the first reported case of COVID-19, people are feeling less motivated to fight the virus. In the United States, case counts continue to rise, and it's critical that we address this fatigue and reinvigorate people to follow recommendations. With the Biden administration transitioning into the white house, it's critical to understand what policies and factors might aggravate the issue further.

Microsoft news recently published an article titled: "'there is no playbook' on how we handle what we all have: pandemic fatigue". Unsatisfied with this pessimistic outlook, we built a playbook.

Problem Description

In October of 2020 the World Health Organization (WHO) released a 25 page document titled "Pandemic fatigue: Reinvigorating the public to prevent COVID-19". It brought up the demotivation of the public, and a decrease in inclination to follow guidelines. The document outlined potential solutions to better understand and solve this issue, including suggested actions for decision-makers, and several considerations to take into account when approaching the issue. Strategies mentioned included: understanding and engaging people, allowing people to live their lives while also reducing risk, and acknowledging and addressing hardships.

Broad approach

Our goal was to develop a pipeline that would utilize data and statistics to better predict how the recommended strategies should be implemented. Through survey data and parsing social media hashtags, fatigue experienced by populations can be quantified. The survey data that we used contained information about public opinions towards the pandemic in general, as well as specific recommendations. However, since the dataset only contained around 30,000 responses, we decided to widen the scope. This meant investigating the ~3 million daily tweets related to COVID-19.

Data was extracted from the C3.ai API, processed, and normalized to obtain the subsets we were interested in investigating. A feature set was then compiled, containing information about demographics, state case counts, and policies implemented in these states. Machine learning algorithms were applied to draw insights into which factors had the largest impact on the sentiment scores of each tweet, and the survey responses. The pipeline developed contains a package with helpful functions we developed to help process datasets, run analysis, obtain twitter data and create visualisations.

Technical Details of Approach

To obtain tweets from the past week, Twitter's public API can be used. However, since we were interested in a longer timeframe, we utilized an IEEE dataset containing tweets related to COVID-19 and their sentiment scores. These scores were computed by a researcher using a Long Short-Term Memory (LSTM) deep network. Appendix A contains a list of the hashtags and keywords used to parse the tweets. Since we were interested in other attributes associated with the tweets, we used an application called Hydrator to rehydrate each tweet ID. We created a tutorial that outlines the specific steps to follow, and it can be found on the submitted GitHub repository. Note that this is a time-consuming task, and each day of tweets will take multiple hours to rehydrate. We were only able to run our pipeline on a few datasets.

The feature sets we created contained information from many of the datasets available from the C3.ai data lake. We extracted a wide range of information, including census data and case counts. From the policy data we extracted policies implemented in each state, and matched each survey and tweet entry to the most recently implemented policies. Once all the data was cleaned and normalized, we implemented machine learning algorithms to draw insights from the data.

One of the algorithms we implemented was a multivariate regression framework. Since our feature set contained many dichotomous variables, one instance of each category was dropped and treated as the reference level. We created the training and test datasets, and fit the model using the linear regression algorithm from sklearn-learn.

In addition to functions useful for data manipulation, we created several functions to help with data visualization. To map data across states, we utilized the

Matplotlib-base map tool and created a function to easily plot color maps of the US. To plot variables over time and compare them, we created a function to easily create subplots for each state. A few examples of plots we generated are attached in Appendix B.

Results

The pipeline identifies factors that relate to pandemic fatigue. For example, with the survey response “coronavirus_intent_mask”, we saw that the policy “mandatory quarantine lifted for travelers” was related to an increase in mask intent. The feature most negatively related to mask intent was “TrumpApproval”.

The data we used to create and demonstrate our pipeline was limited by the amount of resources and time we had access to. However, if implemented on a larger scale more accurate models could be generated, and a more continuous timeframe could be analyzed. This would allow organizations to act upon meaningful insights and re-motivate people in continuing to fight the virus effectively, whilst also taking into account the issues communities are facing.

Impact

We’ve provided a data-driven approach to studying pandemic fatigue and determining effective recommendations to uphold public morale. Using our pipeline, data can be used to draw insights into what barriers these communities are facing, and what initiatives and policies are helping to reduce pandemic fatigue.

Due to our finite timeframe, resources, and restrictions, we focused on the United States and several publicly available datasets, in addition to the data provided by C3. With more time the entirety of the Twitter dataset could be easily analyzed. In addition to the submission requirements, we have compiled and attached a library of useful resources which we hope will be of use to others.

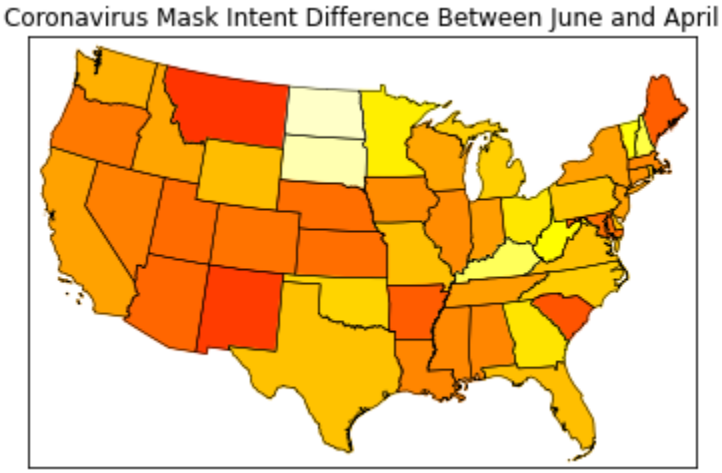
Appendix

Appendix A: List of hashtags and keywords

"corona", "#corona", "coronavirus", "#coronavirus", "covid", "#covid", "covid19", "#covid19", "covid-19", "#covid-19", "sarscov2", "#sarscov2", "sars cov2", "sars cov 2", "covid_19", "#covid_19", "#ncov", "ncov", "#ncov2019", "ncov2019", "2019-ncov", "#2019-ncov", "pandemic", "#pandemic" "#2019ncov", "2019ncov", "quarantine",

"#quarantine", "flatten the curve", "flattening the curve", "#flatteningthecurve", "#flattenthecurve", "hand sanitizer", "#handsanitizer", "#lockdown", "lockdown", "social distancing", "#socialdistancing", "work from home", "#workfromhome", "working from home", "#workingfromhome", "ppe", "n95", "#ppe", "#n95", "#covidiots", "covidiots", "herd immunity", "#herdimmunity", "pneumonia", "#pneumonia", "chinese virus", "#chinesevirus", "wuhan virus", "#wuhanvirus", "kung flu", "#kungflu", "wearamask", "#wearamask", "wear a mask", "vaccine", "vaccines", "#vaccine", "#vaccines", "corona vaccine", "corona vaccines", "#coronavaccine", "#coronavaccines", "face shield", "#faceshield", "face shields", "#faceshields", "health worker", "#healthworker", "health workers", "#healthworkers", "#stayhomestaysafe", "#coronaupdate", "#frontlineheroes", "#coronawarriors", "#homeschool", "#homeschooling", "#hometasking", "#masks4all", "#wfh", "wash ur hands", "wash your hands", "#washurhands", "#washyourhands", "#stayathome", "#stayhome", "#selfisolating", "self isolating"

Appendix B: Plots and figures




Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
DistrictofColumbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
NewHampshire
NewJersey
NewMexico
NewYork
NorthCarolina
NorthDakota
Ohio
Oklahoma
Oregon
Pennsylvania
PuertoRico
RhodeIsland
SouthCarolina
SouthDakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
WestVirginia
Wisconsin
Wyoming
Case data
coronavirusIntent_SixFeet
cases_per_pop