BIG DATA TO APPROXIMATE MENTAL HEALTH INDICATORS
[by RINJANI]

General idea: Official statistics are usually produced based on surveys and/or census, which are held almost every year. Between surveys and/or censuses there is an information gap. Along with the enhancement of technology, this gap could be filled with real-time/big data to give early warning to stakeholders about certain issues of interest.

Problem statement: Suicide is one of the big issues in the mental health field - according to WHO, on average close to 800,000 people die from suicide every year. Referring to the CDC/WPR web page, suicide rates by state are supposed to be produced every year, although statistics for the year 2019 and 2020 are not yet available. Thus, we have identified the following problems that might arise.

1. We are lack of information on the suicide rate in 2019 and 2020 (although the data may be available later at the end of the year).
2. Even if we had yearly suicide rates, we still experience information gaps between the end of years' statistics, when the rates are likely to be available, e.g. we do not really know what had happened in the mid-year.
3. If we wanted to improve the disease surveillance and help the department of health to develop public health interventions in real-time (or at least near real-time), we need to fill this gap, which will be useful during the time of crisis such as in the current COVID-19 pandemic.

Objective: The aim of this project is to propose an approach for filling the information gap of mental health indicators, especially suicide rates, by making an approximation of such indicators using different sources of big data. Why big data? Because big data is generated in a real-time manner, i.e. every second there is always people (i) updating their social media, (ii) using search engines to find information online, as well as (iii) moving from one place to another of which the movement or mobility is recorded in their devices. It is important to note that this approximation is not to replace surveys/censuses but to complement them and fill the information gap between them.

Data:

As dependent or response variable, we used suicide rates as of March 2020, recorded in 2018, at the state-level gathered from the WPR web page. As independent variables or features, we used big data from the following sources.

1. Tweets from Twitter (#covid19, #covid_19, #coronavirus #coronavirusoutbreak #coronavirusPandemic) during September 10th-17th
    - We considered tweets contain 121 words/lexicons related to suicide/depression: 112 from https://github.com/gamallo/depression_classification + 9 additional based on researches
    - We also calculated the negative-positive sentiment score (ranging from -1 to 1) for each tweet using the "sentimentR" package.
    - We also did the emotion analysis based on eight emotion axes, for each tweet using the "syuzhet" package.
2. Google Mobility

    We used Google mobility data available on XPRIZE which contained the % of increase/decrease average mobility of people to certain points of interests by state: Retail_recreation, Grocery_pharmacy, Parks, Transit_stations, Workplaces, Residential.
3. Google Trend

    We used Google Trend to track the number of the following words/phrases searched in each state during September 10th-17th: Antidepressant, Psychiatrist, Psychiatric hospital, Drugs, Suicide, I hate my life, Life sucks, I wanna die.

Methods:

To deal with collinearity and high-dimensional data, we fit the data with L1 and L2 regularization, i.e. adding penalty term in the SSE minimization. We split the data containing 50 states with 122 final features into training data (80%, 40 states) and testing data (20%, 10 states).

- We use the training data to optimize the parameter of L1 and L2 regularization (λ) using 5-fold cross-validation, then fit the model with the best λ
- Using the testing data, we validate L1 and L2 regularization models and assess their performances by calculating MSE and R^2

Results:

Both L1 and L2 regularization models are actually promising, accounting for 60 – 70% variance in the data with pretty small errors, 13.71 and 14.58, respectively. We recommend using the L1 regularization (Lasso regression) model because it has a smaller error and is more interpretable with only 30 features included in the model. Thus, with this model suicide rate can be approximated with 30 features extracted from tweets, Google mobility, and Google trend.

Limitations: In general, the limitations of this project is that the available data are not in the same time frame.

1. The latest suicide rates available on WPR web page, as of March 2020, were reported in 2018.
2. We did not use tweets data available on the XPRIZE web page since this dataset does not have longitude and latitude information for the tweets. We needed longitude and latitude information to map each tweet to the states. Thus, we pulled tweets data on our own. We were only able to pull last week's tweets (pulled on September 17th, 2020, so we had September 10th - 17th 2020 tweets) as we used the free Twitter API.
3. The Google mobility data available on XPRIZE were last updated on March 30th, 2020.
4. The implications of (2) is that we pulled Google trend data during the period of September 10th - 17th 2020 to at least match the time frame of Twitter data.

Potential future works:

1. We could scale-up this work with two different scenarios.
    a. Perform similar cross-section analysis but using one-year tweets (paid API is needed) and Google mobility for more representativeness, since suicide rates data are also available yearly.
    b. Perform longitudinal/time-series analysis to look at the change/trend overtime – this one would need at least 5 years of suicide rates, tweets, Google mobility, and Google trend.
2. It is also worth trying experimenting with other modeling approaches.