CSC 240 Lab Report: Airbnb

Jiahang Wu and Qihao Yun

***Abstract*— This report is the product of a lab assignment from CSC 240, Fall 2021. The work prepared and analyzed a Covid Twitter dataset.**

I. Introduction

This report analyzes and explores a US twitter dataset from April to December 2020, focusing on Covid-related analysis. The report hypothesizes twitter discussion of covid related topics are connected to geographical differences, as it is the case with the US political differences. The report finds substantial evidence against this hypothesis and discusses possible alternativer insights from the dataset.

II. Data

A. Dataset

The Twitter sample contains 200,000 twitter posts during 2020. This report randomly selects 50,000 samples and explores the dataset. Only a small part of the features directly describes the tweet post; many others are information about the poster which is convenient for this report to explore the tweets.

B. Preprocessing

In preparing the dataset, we have selected 50,000 random tweets from the larger dataset. Because the focus of this report relates to locality, we have removed tweets that have unclear poster location (multiple locations, non-US locations) attached to the tweet, losing about 0.03% of the 50,000 tweets. This process took 8.718 second, excluding loading the original dataset. We also added three columns specifying US state, US region, and the calendar day of the post's creation by extracting from the exisitng features of the tweet (location and created_at). This process took 12.851 second.

To prepare to data for NLP exploration, we removed all non-lexicon tokens and stop words from the tweets and lemmatized the content. Punctuation, emoji, url links, and stop words which are words we are unable to analyze (contains little information) such as proposition, grammatical articles, etc. Here, we use stop words as defined in nltk.corpus. Lemmatization converts all words to their base form, without grammatical inflexions or as they would appear in dictionary, to help our analysis. This process took 1553.2397 second.

From here, we are able to explore the relevance of these tweets with covid-related topics such as vaccine, isolation, medicine, and disinfection. Using lists of topic keywords (lemmatized in 2.64 second), we are able to find the cosine similarity between our processed tweet text and the topics' words. The calculation and the subsequent 0-1 normalization took 318.69 second.

III. Results

The tweets are analyzed by their relevence to selected topics, and the results are compiled in tubular form grouping by day, state, and region (took 118.7140 second). There seems to be minimal variations in all four topic by state and similarly (naturally) by region. A clearer evaluation of the geographical variations can be seen in the following two parallel plots of the topic relevance score in the table.

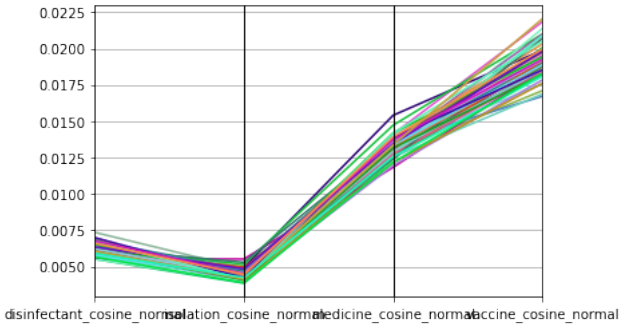

Fig. 1. Parallel Plot of State Scores

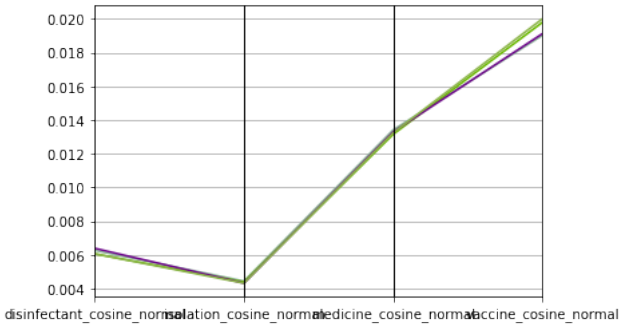

Fig. 2. Parallel Plot of Region Scores

It is almost immediately clear that region variation is quite minimal as said. There are some level of state variation, but overall, all states' values fall closely in a small range. An explanation to the state-wise variation in some of the topics could be population: less populous state may have less overall tweets and are thus more sensitive to outlier tweets. Similarly, populous states naturally have more cases overall which could contribute to more discussion and attentions on twitter.

There is, however, substantial variations when the topic is grouped by day. Among them, isolation and medicine are relatively stable across the 200 days, while disinfectant and vaccine vary much more. Vaccine topic remained low around

0.01 but jumped to around 0.05 by the end of the year. This in particular seems to have an obvious explanation in that covid-19 vaccines were only developed to a stage that garner media attention towards the end of 2020.

To confirm the observation that geographical distinctions seems to have little impact on the topics, we performed clustering using Kmeans and spectral clustering algorithms as defined in the sklearn package. We tested the observation with n=4, following the belief of 4 geographical regions that separates the US states.

The clusterings' immediate results confirms this hypothesis. There are some differences across the clusters generated by either methods but they are hardly relevant to the assumed geographical lines. Clusters from both methods produces seemingly random group of states if we are to consider their geographical locations.

To evaluate further the performance of the methods and the confounding similarity between the results, we performed PCA on the original data and plotted the two results on 2D space below.

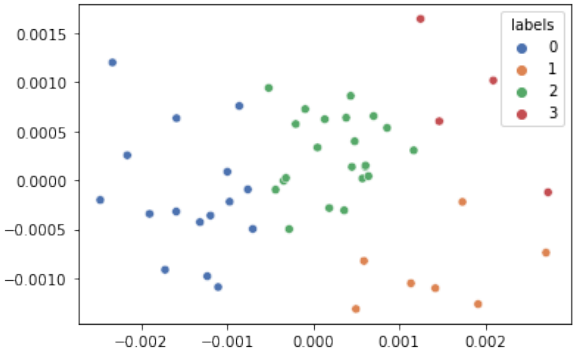

Fig. 3. Scatter plot of Kmeans Clustering on post-PCA dataset

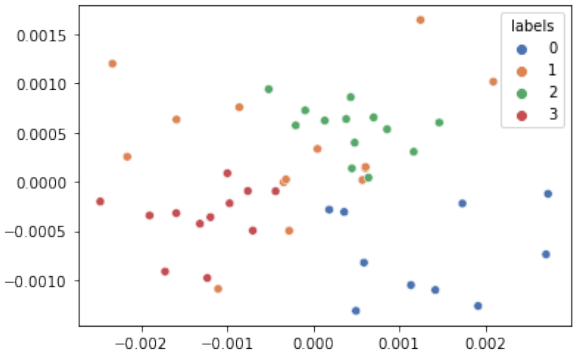

Fig. 4. Scatter plot of Spectral Clustering on post-PCA dataset

From the two plots, we see that Kmeans produces more defined boundaries in their clustering. This is not a result of its algorithm since the data here is casted on 2D plane via PCA. Spectral clustering, on the other hand, show no apparent grouping in our post-PCA dataset. The relative success of Kmeans method is confounded by the relatively loose scattering of the data; without the color code, we as humans are unable to see very apparent clusters within the dataset.

Consequently, we tested alternative n parameter for the clustering using Calinski-Harabasz score.

TABLE I
CALINSKI-HARABASZ SCORE OF DIFFERENT N

N	Kmeans	Spectral
2	40.28	40.41
3	33.88	17.48
4	28.77	15.56
5	26.04	11.3
6	25.12	11.88
7	23.96	6.97
8	24.04	8.06
9	22.76	4.95
10	23.69	4.34
11	21.28	5.53
12	22.08	6.14
13	22.37	4.84
14	21.28	3.25
15	22.22	5.14
16	21.19	2.88
17	22.06	5.04
18	22.12	2.81
19	22.28	3.25
20	22.17	2.39

The results of both methods show clear preference on n=2. This evidence against the geographical hypothesis. Contrary to the inconclusive finding earlier, we may also argue that Kmeans is more robust (better) for this particular dataset as it performs reasonably well at alternative (non-2) N values. While, spectral clustering begins to fail n>4.

A. *Discussion*

We could draw the conclusion that the discourse (at least on twitter) relating to covid in the US is very much a national sentiment, which may be somewhat surprising considering the perceived political polarization in the US. There are many possible factors that are bound to affect the applicability of our conclusions here to the larger world. The dataset we are using, in particular, can come under many doubts. We do not know whether the sample is representative to the larger sentiment of twitter during this timeframe, and we are certainly assuming a sort of independence of the US tweets with tweets of other English speaking countries. In fact, we might argue that Twitter may be a terrible form of representation of our general sentiment in the society, considering a number of assumptions or impression we have about twitter being a very charged and volatile space.

There is a big limitation to the design of this project as we intended - originally - only to confirm or refute the hypothesis regarding geography. Here, we offer a small graph plotting the changes of the topics compared to calendar day.

We can draw from this that the topics are perhaps more of a time-series data than we would have assumed. While somewhat of a better fit that the geographical hypothesis, under closer inspection, we would notice relatively weak variation with isolation and disinfection compared with

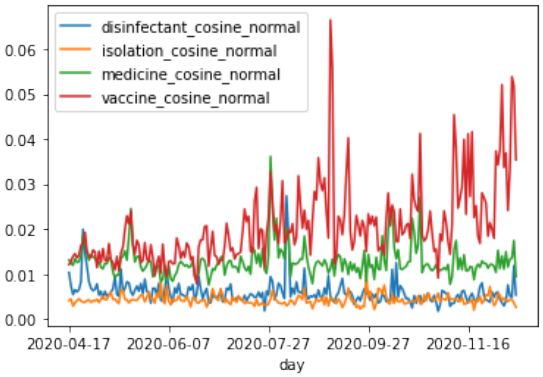

Fig. 5. Changes of Topics by Time

strong variation in vaccine and slightly less so medicine. This is not at all a conclusion but only serve to point out the complexity of "what contributes to the sentiments." It is likely that the actual contributing feature (if we are doing causal inference) is absent from the twitter dataset itself, and it is something that we need to consider for a larger project.