Analysis of Regional Tweeting Patterns During The COVID Pandemic

Anuraag Kumar

***Abstract*— The paper uses tweets made during the pandemic**

I. INTRODUCTION

Even though it seems so long ago, we were trapped at home due to a pandemic that ravaged the world and killed millions. Those conditions are ripe for a large amount of emotion. We certainly saw that. Twitter was a hotbed for communication during the pandemic, and there were a lot of opinions. One key conflict was the narrative around medicine, vaccination, isolation and disinfectants. It felt like half the nation didn't believe in the vaccine while the other would swear and die by it. We plan analyze that conflict through the avenue of twitter. We took a sample of 500,000 tweets and separated them by state of origin and a region of origin. We also analyzed them using a cosine similarity function to a dictionary of terms relating to medicine, vaccination, isolation and disinfectants. We used clustering algorithms on our data set to figure out what states and regions expressed themselves in a similar manner to other states and regions.

I expected the Northeast and the West to talk more about isolation and medicine than their South and Midwestern counterparts as that was a greater concern for the liberal political base and those regions are more liberal than the other two. This fact was almost true. I found that the West and Northeast talked about the vaccine more and talked about medicine and isolation about as much as their South and Midwestern neighbors.

I also expected the states in each region to have similar scores when relating to medicine, vaccination, isolation and disinfectants. However, the results defied my expectations here yet again. I will expand on what that means in the analysis sections but the clusters that were formed were not strictly region based.

II. DATA

We used data scraped directly from Twitter (now X) during the year of 2020. We used a sample of 500,000 such tweets. Our total dataset was 1.6GB which is a lot of data contained in just text. This understandably took a while to process. Many of our methods and analysis during this lab took longer than usual to process.

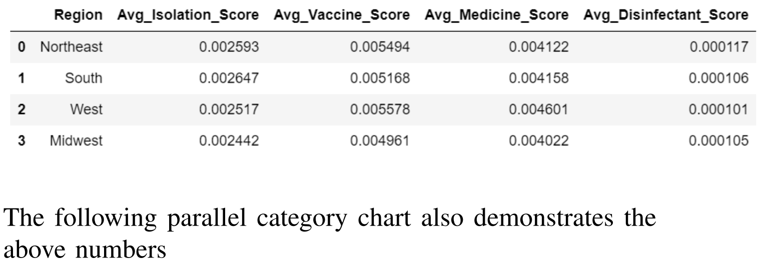
We have the following important features in our tweet dataset: time and day created at, tweet text, id, screen name, and location. I extracted the date, state and region from the "time and day created at" and "location" column and created new columns. I dropped any row that I couldn't find state for. However, I ended up dropping no rows (This probably is a mistake on my end). Therefore, each and every column was valid when dealing with our analysis. This was nice as we could a full and complete picture of the happenings.

We also had lists of words relating to each of vaccines, isolation, disinfectant, and medicine in separate excel files to do our cosine similarity analysis.

III. RESULTS

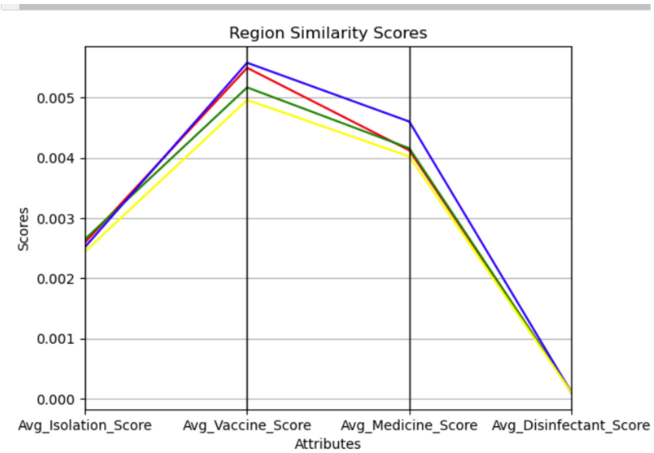
We start our analysis by first considering the text of the tweet. We need to get it to a place where we can actually compare tweets properly to the dictionary list of words for each topic discussed above. Many tweets contain links, hashtags, and filler words that interfere with this process. This is why we lemmatize our tweets. Lemmatizing is the process of turning words into their root words or most basic form (i.e running to run). However, we need to clean our tweet before hand. As an example, after cleaning (to remove unimportant words) and lemmatizing "Hello World! I love being alive and eating cake.", we get "hello world cake". You can see that we focus purely on the world. We ran this process over all 500,000 tweets to obtain "clean" versions of every tweet. I created a new column to contain this information called "text_clean".

I then did the analysis by taking the cosine similarity our clean texts and our dictionaries of different words relating to medicine, vaccination, isolation and disinfectants. We normalized this scores by dividing everything by the largest similarity value (which was about 0.57). We then took the mean of those scores for each category for each region and state. The below dataframe is that for similarity scores for region.



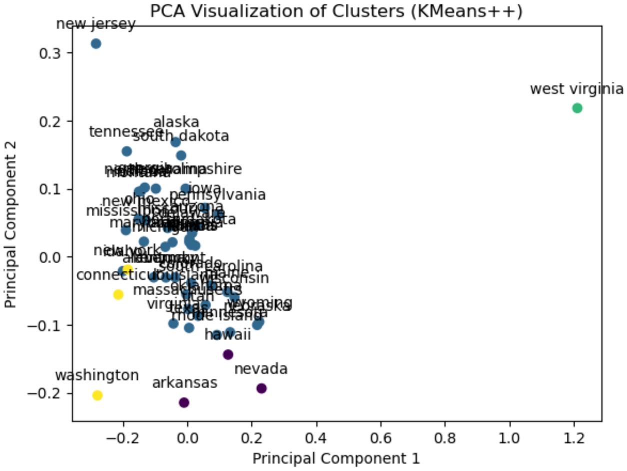
	Region	Avg_Isolation_Score	Avg_Vaccine_Score	Avg_Medicine_Score	Avg_Disinfectant_Score
0	Northeast	0.002593	0.005494	0.004122	0.000117
1	South	0.002647	0.005168	0.004158	0.000106
2	West	0.002517	0.005578	0.004601	0.000101
3	Midwest	0.002442	0.004961	0.004022	0.000105

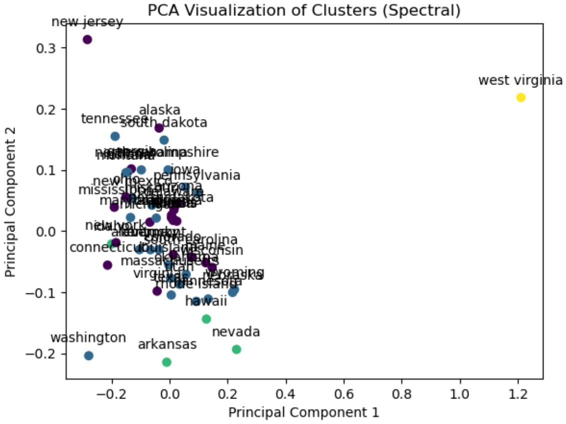
The following parallel category chart also demonstrates the above numbers



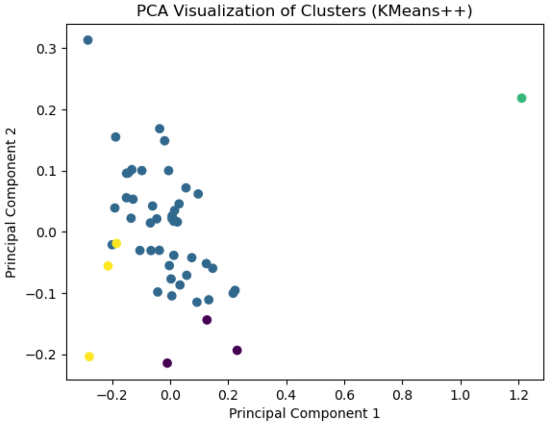
With the blue line representing the West region, the red line representing the northeast, the green line representing the South and the yellow line representing the midwest. We can see above that the scores follow a similar distribution and there doesn't seem like there is a large difference between the regions. This goes against our hypothesis. However, it is also clear that there does seem to be a difference between the west (the blue) and the rest of them. It seemed that the residents of that region were (slightly) more concerned with the pandemic.

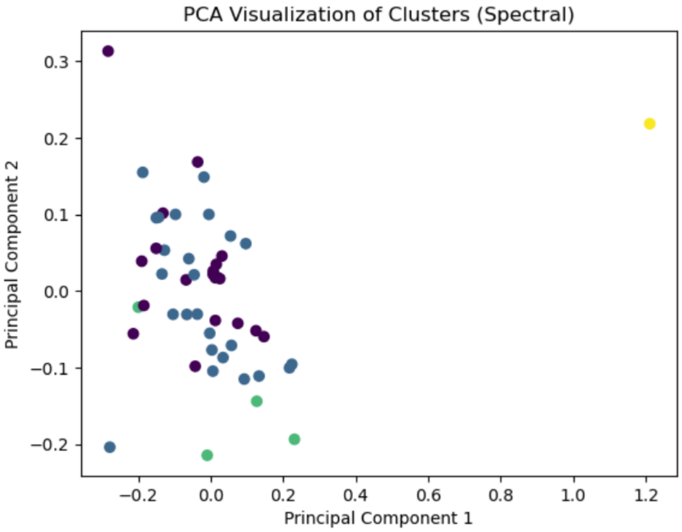
We then turned our focus to the mean scores amongst the states. Ideally, if our results are split well amongst regions then our rows would form 4 clusters. As we have four kind of scores, it follows that our points are in 4 dimensions. To visualize our data, we did PCA reduction and ran kmeans and spectral clustering. Here are the results.





Without labels, our results look like

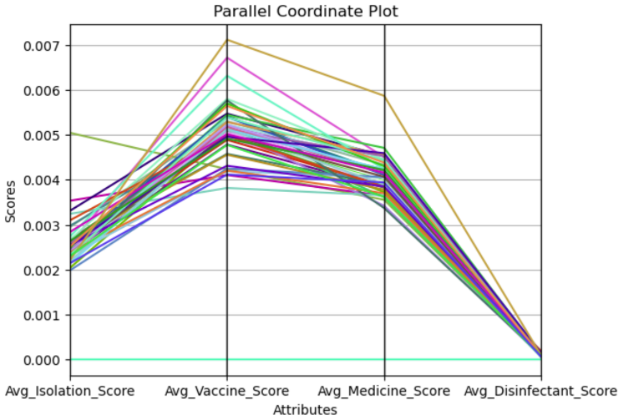




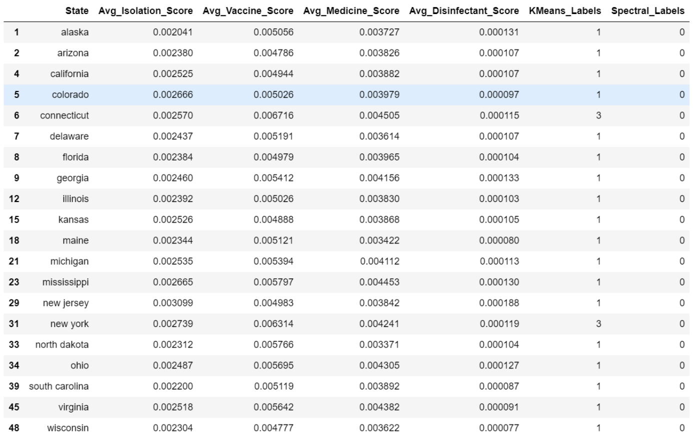
We can see a couple of interesting things. The first is that our data is not very clusterable. This is illustrated that that our optimal clusters for Spectral Clustering and Kmeans++ clustering was 20 and 2 respectively. We found this by taking the largest Calinski Harabsz score over cluster values from 2 to 20. This is an interesting find as it was clear that there was a partisan divide on how one tweeted about the pandemic and states tend to have a very partisan divide. This was the motivation behind my hypothesis. It turns out it was wrong.

One other thing to mention is that Spectral Clutering made less sense as the clusters interesect fold inside one another in many places. This should make sense as Kmeans is completely based on distance from prechosen centroid points and spectral tuning is based on fine tuning parameters of an affinity matrix. The parameters of the spectral affinity matrix might be heavily affected by the dimensionality reduction. However, by our analysis on the Calinski Harabsz scores, it is clear that neither Spectral Clustering nor Kmeans++ clustering was an effective way of spliting our data.

The following parallel coordinate plot makes it clear that the state similarity scores were too similar to cluster effectively.



(I'm not going to bother with the key as it would take too much space on the page and it mostly around the point). It also is worth mentioning that our clusters were not by region as well. For KMeans, one cluster contained nearly every single piece of data, and in spectral data wasn't much better. The following data represents the data in one spectral cluster. You will notice there a states from everywhere in the United States.



	State	Avg_Isolation_Score	Avg_Vaccine_Score	Avg_Medicine_Score	Avg_Disinfectant_Score	KMeans_Labels	Spectral_Labels
1	alaska	0.002041	0.005056	0.003727	0.000131	1	0
2	arizona	0.002380	0.004786	0.003826	0.000107	1	0
4	california	0.002525	0.004944	0.003882	0.000107	1	0
5	colorado	0.002666	0.005026	0.003979	0.000097	1	0
6	connecticut	0.002570	0.006716	0.004505	0.000115	3	0
7	delaware	0.002437	0.005191	0.003614	0.000107	1	0
8	florida	0.002384	0.004979	0.003965	0.000104	1	0
9	georgia	0.002460	0.005412	0.004156	0.000133	1	0
12	illinois	0.002392	0.005026	0.003830	0.000103	1	0
15	kansas	0.002526	0.004888	0.003868	0.000105	1	0
18	maine	0.002344	0.005121	0.003422	0.000080	1	0
21	michigan	0.002535	0.005394	0.004112	0.000113	1	0
23	mississippi	0.002665	0.005797	0.004453	0.000130	1	0
29	new jersey	0.003099	0.004983	0.003842	0.000188	1	0
31	new york	0.002739	0.006314	0.004241	0.000119	3	0
33	north dakota	0.002312	0.005766	0.003371	0.000104	1	0
34	ohio	0.002487	0.005695	0.004305	0.000127	1	0
39	south carolina	0.002200	0.005119	0.003892	0.000087	1	0
45	virginia	0.002518	0.005642	0.004382	0.000091	1	0
48	wisconsin	0.002304	0.004777	0.003622	0.000077	1	0

What can we get from this data? Both of our assumptions turned out to be wrong and it turns out many of our methods didn't tell us. Why did that happen? I think it is because using dictionaries for our analysis was not a good idea. I think it might have been a better idea to use sentiment analysis and natural language processing together to figure out the meaning. It might also have been worth it to use Machine Learning techniques to determine the meaning of each tweet.

It also might be that many people could have talked about the same topics in different ways. For example, "I hate staying inside, and I blame government." and "I like being inside!" both mention the word inside once even though the meaning is very different. We could see this pattern with how both sides of the political spectrum approached social distancing and vaccine mandates.

It didn't seem that we had many good predictors of what state would be related to other states. However, out of the bad predictors, our isolation score was the best. It was the topic talked about the most and therefore had the highest variation, so it follows that clusters were mostly determined by the isolation score as opposed to, for example, the disninfectant score.