Geopoltiks indicator:

Step 1: Make bag of words as per page 1199, flag them as positive and negative as per impact.

Step 2: In google trends api, search for top 25 queries related to each word. The same step can be done from website also.

Context of doing this step -> In a newspaper the number of articles mentioning war, conflicts, etc is occasional and can tell the tone of the article. In the google search these words can be searched any no of times in a day by users without any context. Hence using specific keywords list so that data becomes less noisy. Using words like war, peace, conflicts etc was making data a lot noisy as these can be general words in terms of google search.

Step 3: the top 25 queries for each word could be related or unrelated to geopoltiks, so one has to manually go thru each list to pick up related words.

Step 4: Go to google trends api or website, search for top 25 queries for the words which are selected from step 3. After data download, repeat the process of manual selection as in step 3.

Step 5: As top 25 queries in google trends is a never unending loop, but after repeating step 3 and step 4 several times, the loop will start giving same answers or keywords. That’s when one has to stop the above step as we have reached end of loop for the specific word.

Step 6: Classify the selected words as per bag of words

Step 7: Download data for the words from trends api or website. For trends api, please don’t use free proxies from internet, google algorithms majorly blocks all of them after 2 tries. Try testing and buying proxies, not all proxies work especially data center proxies are instantly blocked by google.

Step 8: After downloading data, clean and organize data.

Step 9: Delete all rows where SVI >= 10. Some datapoints especially which are common keywords like war, peace, etc have lot of noise. Removing entries less than 10 will give less noise to data.

Step 10: Find delta SVI, i.e. current\_SVI – previous\_SVI = delta\_SVI

Step 11: Winsorize delta SVI at 2.5% level at each tail. Tried doing it without winsorization, but then data again became noisy.

Step 12: Remove seasonality from the delta\_SVI by using monthly dummies.

Step 13: keep the residual obtained from step 12.

Step 14: Normalize the residual for each keyword, by year, month and day. (actual residual – mean residual) / (std dev of residual)

Step 15: Sum up all the residuals for year, month and day to get sentiment index. The inspiration for the same is taken from gpr index paper slides. Screenshot attached. Google don’t give total no of searches. Also google trends algorithm have no of searches accounted and normalized in the SVI calculations. The intuition was to sum up the normalized residual for all search parameters for the month to get frequency of searches. The normalized residual parameter is a change in search parameters which is de-seasonalized and normalized. The reason for taking delta SVI instead of SVI is, every word in google is randomly searched by ppl all over the world. Taking SVI for calculations, will not be the exact parameter. But change in SVI tells, how much a search parameter is searched relative to its previous searches. General assumption, an event will increase the searches in google for any keyword relative to its previous values. The measure of increase is a better metric than taking SVI. In gpr parer, frequency of searches is taken, but as the data source have changed hence the calculations of frequency also changed.

