DATA301 Project Proposal

Due Midnight, Friday May 6 – Worth 6%

Lily Williams *42415299* *lfw25*

What is the similarity in average sentiment of health-related events from March 2019 to April 2022 between New Zealand and other countries?

Summary

Using a subset of the GDELT 2.0 Event Database filtered by the "HLH" type code, the average sentiment progression of 6 different countries will be calculated using fine-grained sentiment analysis. This sentiment will be determined by comparing the EventCode of each event to a manually categorized EventCode list, and matched to an integer value from -2 to 2 which describes a scale from "Very Negative" to "Very Positive", respectively. Summing the sentiments from each month for each country and rounding to the nearest integer on the scale, the numerical value will be stored in a vector. The time scale of interest is monthly, spanning 37 months from March 2019 to April 2022, and the six countries of interest will be New Zealand, USA, UK, China, Russia, and South Africa. This will result in 6 vectors of 37 elements, which can then be analyzed using the cosine similarity algorithm to determine how similar New Zealand's sentiment progression has been to the other 5 countries. This similarity index will provide insight on the overall tone interpretation in different countries over the course of the pandemic, and could perhaps revel interesting correlations between the behavior of the individuals in each country and the sentiment of the reported events.

Motivation

The COVID-19 pandemic continues to overwhelm countries, economies, healthcare systems, and the general psyche of human culture. In New Zealand, the pandemic response was initially received as bold but wise restrictions on travel, work, and general life, and I felt like the general feeling was one of unity and cooperation. However, now, two years later, it feels as if most people view it as a thing of the past, and continued restrictions feel like a cage preventing us from moving on. It seems like the belief in the advice of governments and healthcare professionals has gone from the obvious choice to feeling like an unnecessary burden. I believe that investigating the perception of health events and health-related news like COVID-19 over the last few years, both here in New Zealand and overseas, could reveal unique insights into the similarities and differences in how these events are being received amongst people of different countries.

Background

Each entry in the GDELT Project's event database has several CAMEO "codes" associated with each event; CAMEO is a taxonomy that categorizes different event or actor types into numerical codes. This truncates the amount of data that must be stored in the database while retaining the general intention or action of the event. A few relevant codes for this research question would be "Actor1CountryCode", which identifies the primary country involved in the event, and an "EventCode", which describes the nature of the event, among others. The Actor1CountryCode can be used to compare one country's response to another, and the EventCode can be used to determine which of the relevant events are positive and which are negative. Grouping the events by months, the general portrayal on a numerical scale for each country can be averaged. This is a concept known as sentiment analysis, and, over time, can show the progression of how a particular issue is perceived or portrayed. This sentiment can be compared between a list of countries using an algorithm known as cosine similarity. Cosine similarity is a mathematical similarity measurement between two vectors, where each vector represents the sentiment progression of a country, obeying the formula

$$\text{Similarity} = u \quad v||u|| * ||v||$$

Where u and v are sentiments of countries represented as vectors and ||u|| and ||v|| are the magnitudes of each sentiment vector.

Research Question

The research question I would like to answer in this project is this: **what is the similarity in average sentiment of health-related events from March 2019 to April 2022 between New Zealand and other countries?** The countries investigated will be New Zealand, USA, UK, China, Russia, and South Africa. By formatting the average sentiment of health-related events each month of different countries as a vector, the cosine similarity algorithm can be used to find how similar each country's sentiment is to New Zealand's. This similarity can show how each country is interpreting health-related events, particularly over the course of the pandemic, and how different that progression is to that of New Zealand.

Design and Methods

Initially, the GDELT 2.0 Event Database will have to be filtered to only download "HLH" type events between the time periods of March 2019 and April 2022 for each of 6 countries: New Zealand, USA, UK, China, Russia, and South Africa. These will be filtered using the Actor1CountryCodes NZL, USA, GBR, CHN, RUS, and ZAF, respectively. This amounts to 37 months, or 1156 days, spanning March 2019 until April 2022 for each of 6 countries, resulting in 222 lots of data sorted monthly. Assuming each day is approximately 1 MB of data, this is 1156 MB of data. The CAMEO Event Codes can be manually sorted into "Very Positive", "Positive", "Neutral", "Negative", and "Very Negative" sentiments, corresponding to numerical values 2, 1, 0, -1, and -2, respectively. Going through each month of health events for each country, each event code will be passed through a function that determines its sentiment and returns the associated numerical value. Summing these numerical sentiments will calculate an "average" sentiment for the country's month, which will then be rounded to the nearest integer and stored as a numerical value in a vector. Repeating this process for every month and every country will result in 6 vectors of 37 elements which detail the sentiment progression over time. Each country's sentiment will be plotted using Matplotlib to see the explicit differences in sentiment over time. Once these vectors have been calculated, the cosine similarity formula will be applied, where u is the New Zealand vector and v is each of the other countries. This will determine a similarity value between 0 and 1, where 1 is identical and 0 is completely different. This will answer the research question about the similarity between New Zealand's sentiment progression compared to other countries.

There are a few issues with the process of sentiment analysis; since it is a rule-based system, it can be naïve in the realm of natural language processing. Particularly, it can overlook language features like double negatives, irony, sarcasm, implications, missing context, and many other pitfalls that the human brain skims over. Because of this, there is a good chance that some of the events in the database have been categorized incorrectly, or that the proposed scale of very positive to very negative will miss out on important nuances. Without the use of advanced language-processing neural networks or extensive manual categorization, this, unfortunately, is unavoidable.

References

Atlas, J. (n.d.). *Google colaboratory*. Google Colab. Retrieved May 3, 2022, from https://colab.research.google.com/drive/11mF5gwgsoOPmwd5rDwDmDWMOUvoHoYsv?usp=sharing

Google BigQuery Sample Queries. The GDELT project. (n.d.). Retrieved May 3, 2022, from https://blog.gdeltproject.org/google-bigquery-gkg-2-0-sample-queries/

Gupta, S. (2018, January 19). *Sentiment analysis: Concept, analysis and applications*. Medium. Retrieved May 3, 2022, from https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

How to structure quantitative research questions: Lærd dissertation. How to structure quantitative research questions | Lærd Dissertation. (n.d.). Retrieved May 3, 2022, from https://dissertation.laerd.com/how-to-structure-quantitative-research-questions-p2.php

Leetaru, K. (2016, January 13). *Mapping world happiness and conflict through Global News and image mining*. Forbes. Retrieved May 3, 2022, from https://www.forbes.com/sites/kalevleetaru/2016/01/13/mapping-world-happiness-and-conflict-through-global-news-and-image-mining/?sh=78c05923e224

Sentiment Analysis Guide. MonkeyLearn. (n.d.). Retrieved May 3, 2022, from https://monkeylearn.com/sentiment-analysis/

Sentiment Analysis: Comprehensive Beginners Guide. Thematic. (n.d.). Retrieved May 3, 2022, from https://getthematic.com/sentiment-analysis/