**Spring 2021 CS 505 Final Project Write-Up**

Project Description : Covid-19 News Classification

Team Fourteen: Ding Ma, Gordon Ng, Tegveer Ghura, Yining Wang

Code url: https://drive.google.com/drive/folders/1gu\_D\_q5Rm4zFO3HrNXEM53g4yHU-DNWv?usp=sharing

1. **Problem Introduction**

We aim to comprehensively analyze the correlation between the impression of the US public on pandemic from April 2020 to August 2020 captured by GALLUP poll together with the proportion of news reports in the United States, China and South Korea, utilizing tools such as the correlation coefficient and Granger causality. We use the rationality of these results to compare the reality people are being faced with within the real world.

1. **Approaches and Models**
2. *Dimension Reduction*: transform data from a high-dimensional space into a low-dimensional space.
3. *Z-score Normalization*: a strategy of normalizing data that avoids this outlier issue.
4. *Spearman Correlation*: a non-parametric test to measure the degree of association between two variables.
5. *Granger Causality Test*: a statistical hypothesis test for determining whether one-time series is helpful for forecasting another. The F-test value indicates the degree of impacts, and the P-value indicates reliability, a p-value <0.05 indicates significance;
6. **Experimental Setup and Findings**

**3.1 Data preparation:**

Since the proportion of news topics has a lot of missing and non -consecutive data, we decided to focus on the complete data gathered by certain countries, including China, Korea and the US’s news topics from April 4th to August 8th and public impressions from April 4th to August 8th. Fill the blanks in the data with 0s. Then, to reduce the dimension of different impressions, we first label scores to them:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Impression** | **A lot better** | **A little better** | **Staying about the same** | **A little worse** | **A lot worse** |
| **Score** | **5** | **4** | **3** | **2** | **1** |

We sum the label score by multiplying the proportion of its corresponding impression to calculate a single score for impression. This compresses the original five columns of data into one column while retaining the characteristics of different impressions.

Then we apply z-score normalization on both data sets to meet the requirement of using Granger Causality and avoid the influence of dimension, variable variation, and numerical value when computing the correlation coefficient.

**3.2 Findings:**

***USA :*** The result of correlation between impression on pandemic and different news topics.

**Spearman Correlation With Time Lags**

|  |  |
| --- | --- |
| correlartion | Explaination ：  According to the result of the Spearman correlation coefficient, the category most correlated with impressions on pandemic news topics are:   * Entertainment & sports (negative). * Domestic economic consequences (positive). * Racism & race relations (negative). * Domestic COVID-19 outbreak (negative). |

**Granger Causality**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1-Lag | | 2-Lag | | 3-Lag | | 4-Lag | |
|  | F-test | P value | F-test | P value | F-test | P value | F-test | P value |
| 1 | 12.968 | 0.0029 | 2.3006 | 0.1463 | 2.0010 | 0.1925 | 1.4428 | 0.3435 |
| 2 | 0.7356 | 0.4055 | 4.8710 | 0.0305 | 2.4654 | 0.1368 | 4.1375 | 0.0757 |
| 3 | 3.7116 | 0.0746 | 3.8085 | 0.0454 | 3.3322 | 0.0770 | 1.7686 | 0.2720 |
| 4 | 0.5213 | 0.4822 | 1.2166 | 0.3332 | 0.5938 | 0.6365 | 6.8458 | 0.0292 |

1: impact on entertainment & sports. 2: Domestic economic consequences. 3: Racism & race relation. 4:Domestic COVID-19 outbreak

We find the 1-lag impact on entertainment & sports Granger-cause publics’ impression on the pandemic. The Spearman correlation coefficients between them are negative, which indicates that an increase in coverage on the impact of entertainment & sports is followed by a decline in general impressions on the pandemic. The change of proportions of four news topics with high Spearman correlation coefficients Granger causes the transformation of the impression of the pandemic. Similarly, we could see that Racism has a negative impact on correlation, which is seen because the “Black Lives Matter” movement started at the end of May. Further, we could see the Domestic Covid-19 outbreak has negative impacts but an increase in time lag. The US’s public attitude did not change until the disease affected thousands and became a severe issue. Furthermore, from this data, we could see that Trump’s economic stimulation positively impacted people’s attitudes, which is reflected by the granger results of Domestic economic consequences.

***China:*** The correlation between impression on the pandemic and different news topics.

**Spearman Correlation With Time Lags**

|  |  |
| --- | --- |
|  | Explaination ：  According to the Spearman correlation coefficient, the category most correlated with impression on pandemic news topics is：   * Frontline healthcare. (positive) * Responses from international organizations (positive) * International COVID-19 outbreak. (negative). |

**Granger Causality**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1-Lag | | 2-Lag | | 3-Lag | | 4-Lag | |
|  | F-test | P value | F-test | P value | F-test | P value | F-test | P value |
| 1 | 3.1907 | 0.0957 | 2.3971 | 0.1367 | 1.1116 | 0.3996 | 1.0926 | 0.4504 |
| 2 | 1.9759 | 0.1816 | 0.9279 | 0.4242 | 1.6213 | 0.2596 | 1.7378 | 0.2778 |
| 3 | 1.1861 | 0.2944 | 0.8575 | 0.4507 | 0.5050 | 0.6894 | 1.0316 | 0.4732 |
| 4 | 0.0024 | 0.9609 | 0.5196 | 0.6086 | 0.1706 | 0.9132 | 1.3223 | 0.3761 |

1: Frontline healthcare. 2: Responses from international organizations.3: Impact on domestic politics. 4: International COVID-19 outbreak.

Though Spearman correlation coefficients show that they correlate with each other, the Granger causality result says otherwise. It shows that the news coverage in China is not followed by the US general impressions on the pandemic, so they are unrelated to one another based on this data. Most of the news topics related to the US pandemic impressions are about international pandemic or health care. However, there is no precise depiction/sequence of this relationship, and the correlated factors are not helpful for forecasting.

***Korea:*** The result of correlation between impression on pandemic and different news topics.

**Spearman Correlation With Time Lags**

|  |  |
| --- | --- |
| korea | Explanation：  According to the result of Spearman correlation coefficient, the category most correlated with impression on pandemic news topic are:   * Impacts on education sector (positive) * Global economic consequences (positive) * US-China relationship (positive) * Domestic economic consequences (negative), * Impact on entertainment & sports (negative). |

**Granger Causality**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1-Lag | | 2-Lag | | 3-Lag | | 4-Lag | |
|  | F-test | P value | F-test | P value | F-test | P value | F-test | P value |
| 1 | 5.1942 | 0.0388 | 1.0595 | 0.3794 | 0.5145 | 0.6836 | 0.8185 | 0.5647 |
| 2 | 1.2215 | 0.2876 | 2.1270 | 0.1655 | 2.0227 | 0.1893 | 2.7296 | 0.1501 |
| 3 | 1.8819 | 0.1916 | 1.7130 | 0.2250 | 2.4961 | 0.1338 | 2.3165 | 0.1908 |
| 4 | 1.2067 | 0.2905 | 0.8253 | 0.4634 | 0.5055 | 0.6891 | 1.1583 | 0.4274 |
| 5 | 5.8402 | 0.0298 | 3.1422 | 0.0832 | 1.2353 | 0.3588 | 4.5482 | 0.0638 |

1: impact on the education sector. 2: global economic consequences. 3: US-China relationship.

4. domestic economic consequences 5.Impact on entertainment & sports

The result shows that most news topics are not a direct cause of the impressions on the pandemic. The impact on the education sector and the impact on entertainment & sports affected the public opinion only in the first k-lag time frame or (1-lag).

This result probably shows the inconsistency of the peak of the epidemic situation between South Korea and the United States. Before the epidemic situation in the United States expanded, the epidemic situation in South Korea reached its peak, with the most pessimistic reports on the epidemic situation. From July to August, the epidemic situation in South Korea retreated. After the outbreak in the United States became severe, the negative reports on the epidemic situation in South Korea decreased. At the same time, the impression of pandemics in the United States also reduced. So they may have high Spearman correlation coefficients. But there is no precise sequence of this relationship.

**4. Conclusion:**

Our results restore the pandemic environment from April 2020 to August 2020, showing local and international news effects on the American people's attitude towards Covid-19. According to our expectations, most of our results show that local news about economics, pandemics, and racism influences public attitude. However, some of the outcomes are out of our expectations, like how Korean news on education and entertainment impacts sentiment. To further analyze these outliers in the results, we have to rely on more data sources. Overall, from our analysis, we can conclude that news reports closely correlate with the public's attitude on specific events during the pandemic.

**5. Cited References:**

* Amherst, UMass. “Machine Learning for Language Toolkit: Topic Modeling.” *Topic Modeling*, 2007, mallet.cs.umass.edu/topics.php.
* Field, Anjalie, et al. “Framing and Agenda-Setting in Russian News: a Computational Analysis of Intricate Political Strategies.” *Aclweb.org/Anthology*, Carnegie Mellon University, University of Haifa, Stanford University, 2017, [www.aclweb.org/anthology/D18-1393.pdf](http://www.aclweb.org/anthology/D18-1393.pdf).
* News, Gallup. “Coronavirus Pandemic.” *Gallup.com*, Gallup, 6 Apr. 2021, news.gallup.com/poll/308222/coronavirus-pandemic.aspx.
* University AIEM Research Group, Boston. “Communicating COVID-19 – Weekly Updates of Main Topics in International News Coverage of COVID.” *Philemerge.com*, Dr. Lei Guo, 2021, covid19.philemerge.com/.
* Maitra, Sarit. “Time Series Analysis Using Granger's Causality and VAR Model: an Example with Python Code.” *Medium*, Towards Data Science, 28 June 2020, towardsdatascience.com/granger-causality-and-vector-auto-regressive-model-for-time-series-forecasting-3226a64889a6.