Sentiment Analysis Of COVID-19 Related News Media Across Nations

Chhiring Yangjen Lama
May 10 · 7 min read

By Alejandro Garcia, Chhiring Lama, Stuti Goyal

Introduction

The coronavirus pandemic approaches 2 months and over that time the virus has garnered an abundance of attention in the news. Sites have covered the economic impact, governmental responses, conspiracies, and more as it pertains to COVID-19. However, the responses and impacts vary amongst different countries. The cases of Denmark and South Korea displayed a quick response to the spread of the virus, mitigating future costs and deaths. Furthermore, in the case of Denmark, employees maintained their jobs or received paid leave in the weeks of disruption. However, the

USA displayed a divided response and reaction amongst states. Mandates and enforcements differed depending on which state boundaries you found yourself in, and protests have broken out due to the experienced restriction of freedoms. News sites have covered the contrasting responses and developments, and the sentiments expressed towards different countries in the articles may have differed.

To understand the sentiments expressed towards different countries during the time of corona, we devised our project with the goals:

- To understand the sentiment of most relevant articles written by the UK or US-based news media towards the management of the COVID-19 crisis by select countries using News API.
- To understand the trajectory of change in the sentiment associated with countries over the past few months.
- To visualize the data using line graphs and WordClouds

Through our analysis, we expect to find a positive sentiment towards Taiwan, South Korea, and Denmark while countries such as the UK, India, and the USA leaning towards the negative side. We also expect to see a negative sentiment for Italy but with sympathetic undertones. While earlier sentiments towards China were mostly positive there has been a shift in the reporting with many allegations and controversial theories being attributed to the Chinese government. Hence, we predict a gradual downward trend in sentiment towards China over the past month.

The Data

To gather our data from a range of news sources, we used NewsAPI which sources data from over 50,000 news sources and blogs. NewsAPI offers a free developer key but with an imposed limit on the quantity of requested data, limit to articles up to a month old, and limit to the first 260 characters of an article's content. With the restrictions on our queries and data, we constructed a proof-of-concept in which someone with the business or enterprise API key could generate more thorough data and more accurate portrayals of the sentiment trends.

We obtained articles between 5th April 2020 and 4th May 2020. The articles were in English and were primarily published in the US- and UK-based news platforms such as:

- CNN
- Wall Street Journal
- The New York Times
- Fox News
- The Washington Post
- The Guardian
- BBC

We filtered the data to only include the articles that mentioned *COVID-19* and the specific country and analyzed the data from the hundred most relevant articles.

The Libraries

In our project, the following python libraries were used:

- Pandas
- Numpy
- Matplotlib
- Nltk
- Newsapi
- Wordcloud
- Sklearn
- Scipy
- Datetime
- Json

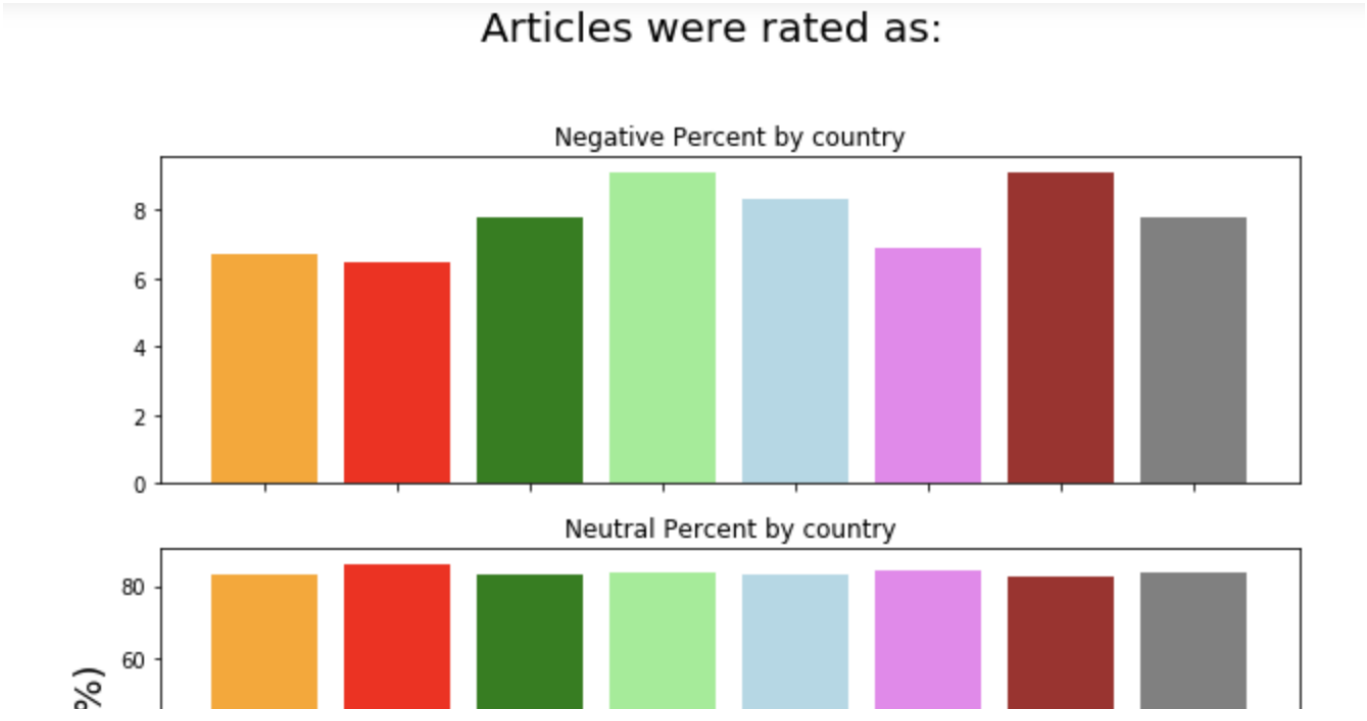
Data Pre-processing

In order to perform data analysis on the articles, we need to remove the unnecessary information. This includes cleaning and wrangling the data. Firstly, only essential information such as the news source, title, content, and date of publication was extracted. To get the correct format of the dataset for the word cloud, the content of the articles was used to generate a new flat list of all the words used in the articles about the specific country. Further data cleaning was done to remove integers, shorter words (with less than three letters as most of them are articles and prepositions), scheme from URLs (Http/ www), and ellipses. Similarly, for the VADER sentiment analysis, the content was converted into a list of sentences for each article. We chose to run sentiment analysis on articles using sentences as supposed to words because doing so gives the context to the usage of words and gives a more accurate sentiment score[1].

Sentiment Analysis

The method of sentiment analysis or opinion mining used for this project is VADER. "VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to the sentiment in social media"[2]. VADER uses a combination of lexicons such as words and sentences to not only identify whether the text is positive or negative but also how positive or negative it is.

The results for VADER for each of the country is as follows:



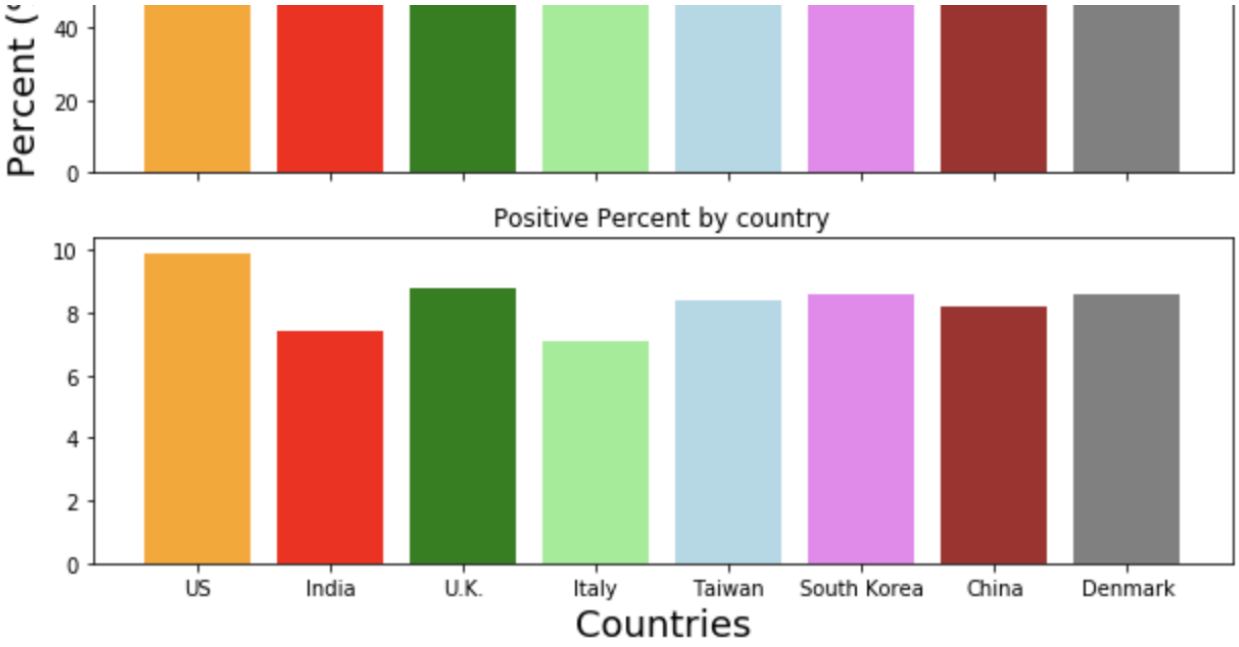
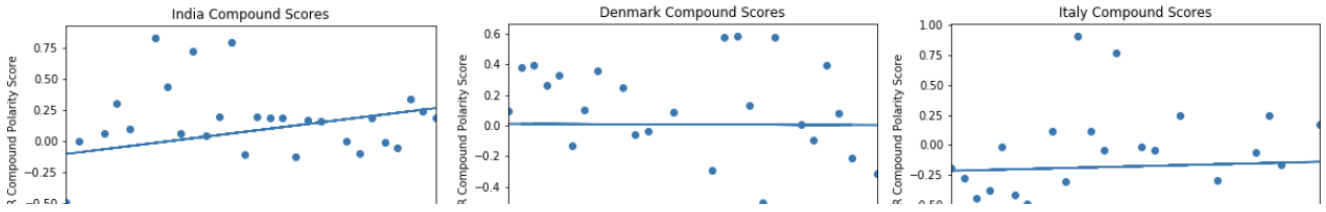


Fig 1: Overall Sentiment score the countries for all COVID-19 related articles about the respective countries

As shown in Figure 1, articles about most countries over the last month have been mostly neutral, with most of them rated as around 80% neutral. This indicates that the news articles are mostly relaying facts that could range from citing demographic trends, stating the country's institutional and individual responses to mapping out new information about the virus and disease prognosis. The articles seem to be mostly informative as opposed to argumentative. However, the positive and negative sentiment score suggests that there are some opinions expressed in the articles. Pieces on Italy and China are about 9% negative in general, followed by Taiwan (8%) and then the UK (>7%). While results for Italy, China, and the UK were not as surprising to us as we had personally read a number of articles questioning their responses to the pandemic, results for Taiwan were slightly surprising. International media had praised Taiwan for its rapid response towards the outbreak.

A higher percent of the articles about the US were categorized as positive (nearly 10%) compared to other countries, followed by the UK (about 8%). Articles about Italy are less than 7% positive. This trend could be due to a number of reasons, which we will be looking into further with the help of word clouds.



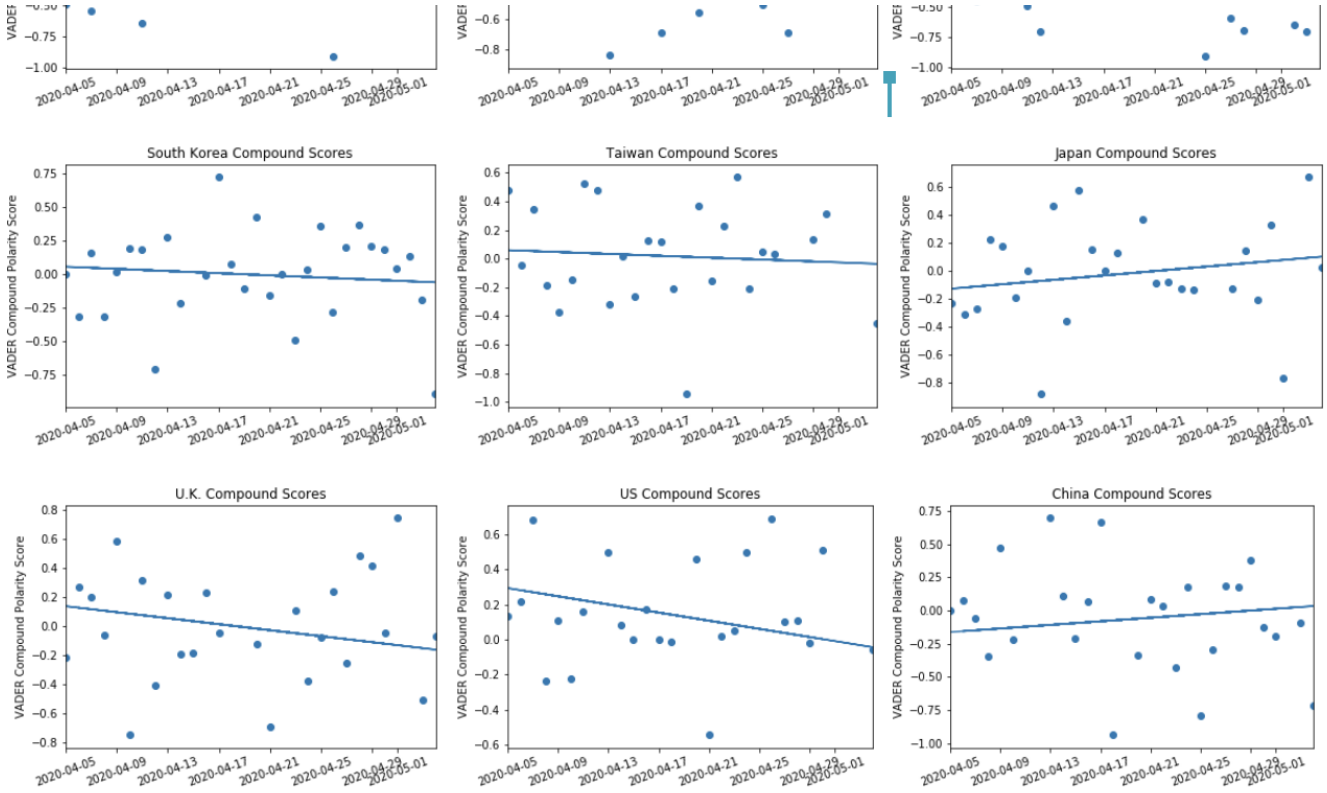


Fig 2: Trend of compound sentiment score for the eight countries over April

To see how the opinions of the news articles changes in April, we ran a linear regression for the VADER compound polarity score using the time (day) as the independent variable. The compound score for India has increased over the month the most (correlation(r) = 0.294. Similarly, the score for China and Japan is predicted to have increased by 0.150 and 0.229 respectively per day. The polarity score for Denmark has remained more or less the same, the articles about the country being mostly neutral in April. The sentiment towards Itay in the articles are negative (compound score of around -0.25), it has remained consistent throughout the month. The sentiments expressed in the articles about the US, the UK, South Korea, and Taiwan have become more negative over the month, particularly for the US (r: -0.333) and the UK (r: 0.251).

These trends show a clearer view of how the opinions towards the country changed within April. This might be correlated with the various relevant events and decisions taking place or affecting these countries. However, these results are not significant and thus might not be as effective in generalizing for articles not included in the dataset (due to our limitations with NewsAPI) or those which are not in English.

Word Frequency Data Visualizations

To further get a sense of the sentiment of the article we decided to make a word cloud to understand the usage of words. We made word clouds using the Word Cloud library.

Here is a cloud depicting the frequency of the words in articles written about China, Taiwan, and the UK in the news media about the COVID-19 virus (Figure 3). It sources words from articles from the 5th of April 2020 to the 4th of May 2020.

Fig 3: Word Cloud for China, UK, and Taiwan for COVID-19 related articles

It is interesting to see the high frequency for certain words concerning the countries above. For instance, for China apart from the terms related to the pandemic such as pandemic, patients, viruses, outbreak, etc, there are highly used words like Trump and apple. We can potentially relate them to recent news about Apple's stock price sinking in China and regular remarks that Trump has been making throughout this outbreak about China.

With regards to the UK, widespread news and updates about Prime Minister Borris Johnson after testing positive for the COVID-19 and about the need for more about overwhelming the NHS and availability of adequate ventilators have been in the news a lot recently. This shows in the word cloud also suggests why the compound polarity score of the UK dropped from 0.2 down to about -0.2 at the end of April.

In the case of Taiwan, what is surprising is 'racist' is one of the most used words in the articles about Taiwan. This could have contributed to articles pertaining to Taiwan to be rated higher than the expected negative. With some further research online, we came to know about the row between the WHO chief and Taiwanese President about racists remarks that originated from Taiwan and were directed towards the WHO chief.

• • •

To get more information about our analysis and check out word clouds for other countries, please follow this link to our Github repository.

[1]*Sentiment Analysis* https://www.nltk.org/howto/sentiment.html

[2] Parul Pandey (Sep 23, 2018). *Simplifying Sentiment Analysis using VADER in Python (on Social Media Text)* https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f