**NEWS COMPARISON ANALYSIS REPORT**

**Introduction**

In the past, news information was not easily accessible to a lot of people. This may have been due to the medium of propagation. But over time, the world evolved, and so has media to disseminate information. Information can now be easily gotten from social media, print media, television and news websites which is going to be our focus.

There is a plethora of information to be accessed on the web which makes it difficult for users to access their news of interest and that is why it has become pertinent to categorize news for the users. It means grouping news items relating to a particular sector together. This gives the user real-time access to interesting news without stress. It also helps interested parties gain insights into the state of affairs.

For this mini-project, I would be trying to compare news from the first 2weeks in December 2019 and the first 2 weeks in December 2020 by categorizing the news articles based on their title. These topic categories are decided arbitrarily.

**Data Collection**

The news article titles were gotten from the archives of “World News” at <http://archive.wn.com/>. World News is an international news agency that compiles, and supplies news obtained from various other international news agencies like “BBC News”, “The Independent”, “CNN”, “Al Jazeera”, etc. I scraped details of the news article like the title, publisher (other agencies), date, and link to the article. The scrapping was done using the “Beautiful Soup” library. I chose this method because of its ease of use. The dataset for this project is not large and the site is static so other more complex methods might be an over-kill.

**Data Analysis**

The visualizations were done using matplotlib. The environment used was Google Colab and the .ipynb file to view them can be accessed from [here](https://colab.research.google.com/drive/1oEi9pakREr9fAcLTT41NyCIq_2_Frt6R#scrollTo=Okq0hfgc3zbR).

From the visualizations, “The Independent” with about 11 articles has the highest number of news articles published in the first 2 weeks of December. Coming after it is “Yahoo Daily News”, which is then followed by “Russia Today”. Examining yearly, “The Independent” again topped the charts for 2019, followed by “Reuters”. However, it is noticed that Reuters was dropped totally in 2020. “Russia Today” has the highest, followed by a tie between “Yahoo Daily News” and “The Independent” in 2020.

In 2020, about 110 news articles in total were published as opposed to the 95 articles published in the previous year. I attempted to check if more articles were released at the beginning of the month compared to the middle of the month, and no meaningful correlation was found. An average of about 7 news articles is released daily.

As mentioned earlier in this report, the categories were chosen arbitrarily and include conflict, politics, health, climate, legal, sports, accidents, business and general. Over the 2 years, news related to conflict is the most published with over 65 stories in total. Politics is also well talked about, and the least published topics are sports, accidents, and business.

The topics related to conflict dropped from 37 to about 29, while politics-related topics increased by about 60% from 23 to 37. I believe the increase in political news articles may have been due to the US presidential elections that took place in November 2020.

In health, a very sharp increase of about 500% is seen in the number of health-related news articles between 2019 and 2020. This sharp increase may have been because of the COVID19 pandemic because the words “covid19” and “vaccine” were found among the top words in the titles.

**Conclusion**

I have been able to successfully categorize news titles into various topics for easy access to the topic of choice and for drawing more insights. I have also been able to make it interactive for users to pick a topic, receive a list of titles that fall in that topic and select a particular title to get the link.

While scraping data from the archives of World News, I discovered that the page structure changed from 3rd December 2020. To solve that problem, I had to write another function to handle the scraping of the new page structure.

In future, if we decide to scrape more news titles that have not been seen before, new categories may emerge. The algorithm used to categorize the titles could be further modified to do so more accurately.