Exploring, Discovering and Detecting News Headlines

Visualising News Headlines from 2020 with R Shiny

Gerald Chee [1]
gerald.chee.2020@mitb.smu.edu.sg

Jihun Nam [1]
nam.jihun.2020@mitb.smu.edu.sg

Atticus Foo [1]
atticusfoo.2020@mitb.smu.edu.sg

[1] *Singapore Management University, School of Computing and Information Systems*

Problem Statement

Being able to cut through the noise and quickly make sense of news in today's news landscape has become critically important because of the following reasons:

1. The volume and scale of information available to you today via digital channels;
2. Social media echo chambers that reflect and confirm one's biases in understanding news;
3. Online news can be manipulated by various methods such as social media engagement campaigns.

Having the ability to quickly capture snapshots of news headlines can thus have an impact not just on personal sense-making but also on society.

Motivation

Our project is thus focused on helping users to create snapshots to better explore topics news topics and events within a news corpus, discover word sequences in a news headlines and detect anomalous performing news articles and events.

This is enhanced by research studies that have identified the issue of a *lack of meaningful methods* to deal with news analysis; despite news media's *tangible impact on public perception and sentiment* and; have warned against dangers caused by *social media echo chambers* in perpetuating a narrow world view.

As identified by Data Journalism, there is *not enough being done to utilize visualization tools* to help readers make sense of the stories that are out there.

Approach

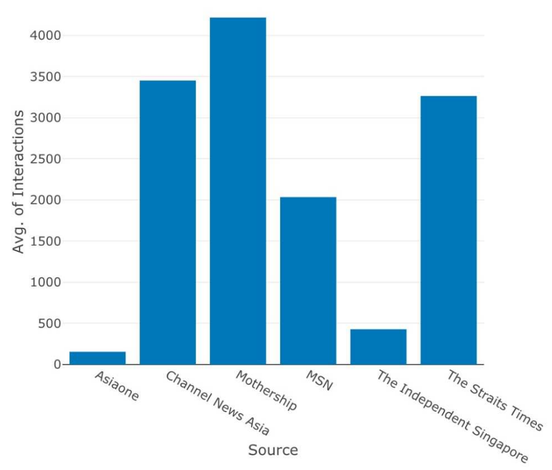
The project's goal is thus to blending both datatypes and utilize different statistical methods to build an interactive media dashboard that can help readers of local news sites quickly navigate and understand a large corpus of news. The intended result is different snapshots that would allow readers to:

1. `Explore` the news corpus using Latent Dirichlet Allocation to sort and identify key news events and topics by time period, topic and keyword;
2. `Discover` words used in news headlines using N-grams to analyse reporting styles;
3. `Detect` and be alerted to amomalous patterns in events and articles using Max-Anoms and social media engagement.

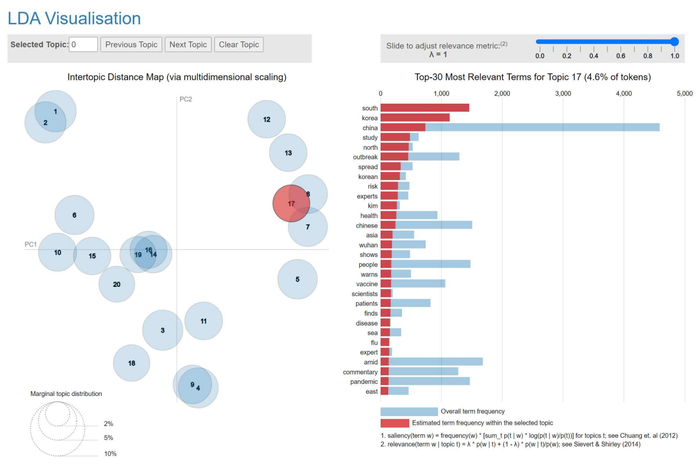
As such, while the presentation of the 'snapshot' may appear simple, the methods that would derive these insights are not.

Results

Overall, the six Singaporean news outlets published an average of 225 articles per day in 2020 (for our analysis). Each article received about an average of 353 Facebook engagement. While Straits Times and Channel News Asia may appear to dwarf other local news sites.



Results from the *explore* module in the dashboard actually demonstrate that Mothership is more effective at reaching Singaporeans with its articles receiving a higher engagement rate on average.

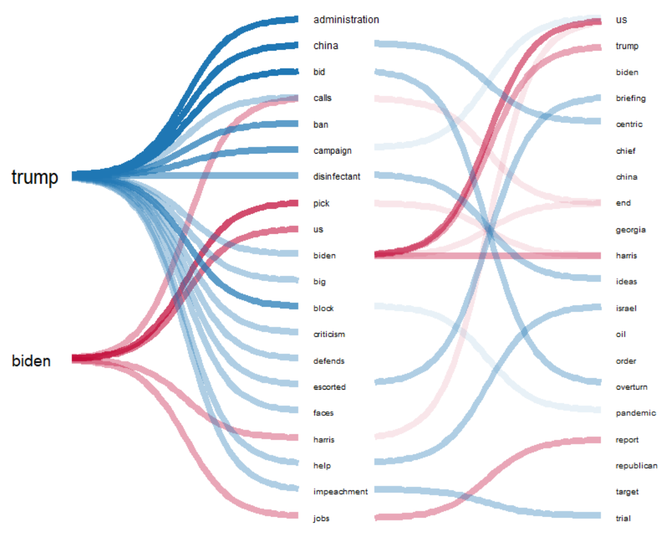


Using the `Explore` module in the dashboard, readers can interact with the top topics generated by LDA and visualised using LDAViz. We can easily observe the dominant topics are related to the COVID-19 pandemic. Reader can then use CoporaExplorer to *explore* how frequently related COVID-19 keywords occur in the news in 2020.

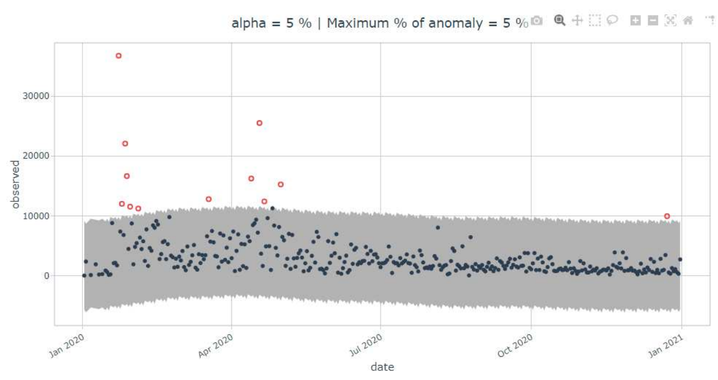
ï..Term	Hits	Proportion
Covid	15,807	16.98%
Coronavirus	10,667	11.46%
Wuhan	1,673	1.80%

From the results in Table 1, we observe that COVID dominated news coverage with nearly 1 in 5 articles related to the pandemic in 2020.

In `Discover` word sequences are extracted from the news corpus based on how it more frequently appears in news headlines.



Flow lines from the two keywords in the figure above, we able to *discover* that Trump was more likely to feature in local news reporting as compared to Biden. Compared to Trump, Biden was involved in less news topics - with coverage focusing on his choice of Vice President, Kamala Haris and Jobs.



Using the `Detect` module we find that an anomaly was detected on 18 April 2020. The cause of this was due to the news of a '*Daily high of 942 new Covid-19 cases reported in Singapore*' which gained 182,148 engagement. This cuts through expansive news corpus' to *detect* key events or articles that could be of potential importance.

Future Work

There are two areas for future work. The first extends the dataset to include full articles for analysis instead of just news headlines. This would enhance the usability of the system and the user wishes of a 'one-stop-shop' for media analysis.

The second area for future work is being able to identify and visualise how a topic evolves in a new corpus. Paired together with engagement scores, the dashboard can help readers to discern the inflection points involved in how news unfolds.