Exploring and prototyping visualisation methods for understanding the news corpus using news headlines collected from six Singaporean news outlets in 2020

Nam Jihun
Singapore Management University, School of Computing and Information Systems
jihun.nam2020@mitb.smu.edu.sg

Gerald Chee
Singapore Management University, School of Computing and Information Systems
gerald.chee.2020@mitb.smu.edu.sg

Atticus Foo
Singapore Management University, School of Computing and Information Systems
atticusfoo.2020@mitb.smu.edu.sg

ABSTRACT

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:

• The volume and scale of information available to you today via digital channels;
• Social media echo chambers that reflect and confirm one's biases in understanding news;
• 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 the wider social fabric. This can be seen from how media reporting can contribute to the public's understanding of safe distancing measures during COVID-19 to how it can influence share prices on the stock market.

Our project is thus focused on helping users to create snapshots to better explore, discover and detect events in the news corpus. This will be performed in this project, using an R Shiny dashboard to ingest unstructured text data from news headlines and social media engagement data to quantify reach. Using identified R Packages, we want to be able to adopt techniques to organize and analyse these two data types to help users to create snapshots to better understand coverage in a news corpus through features that allow for exploration of topics, discovery of context relating to topics and detecting anomalies in news coverage.

1. MOTIVATIONS

Our interest in this project is enhanced by research studies that have:

• Identified the issue of a lack of meaningful methods (Koivunen-Niemi & Masoodian, 2020) to deal with news analysis;
• Despite news media's tangible impact [(Bhargava , Bishop, & Zuckerman, 2020)] on public perception and sentiment and;
• Have listed increasing dangerous caused by social media echo chambers (Grimes, 2017) in perpetuating a narrow scope of news – which can lead to an inaccurate world-view (Seneca, 2020).

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

Kouivunen Niemi and Masoodian echo this sentiment and argue that the news media and its reports can have "widespread repercussions in the public perception of past and present phenomena." They posited that a tool or method could be designed to simply to report findings in an "easy to understand but effective way." They also go further to lament how these types of studies were "almost invariably presented in tables or using simple graphs such as bar charts or line charts" that were not meaningful.

1.1 Existing vendors focus solely on either structured or unstructured data, but not both

Meanwhile, commercially available options (both paid and free) offer limited scope. They either provide services in text analytics or social media analytics, but not both (Toros, 2019).

An interesting example for text analytics was a company called Primer who specializes in the use of ensemble machine learning models for analysis of unstructured data like news media articles. Some key features include information extraction, multi-document pairwise relationships and abstractive summarization.

On the other hand, NewsWhip's approach utilizes social media engagement data to quantify how quickly news spreads online. We are thus interested in using text analytics solutions in Primer and engagement metrics from NewsWhip to provide an intelligent media sensing dashboard that can offer more by blending sets of data and methods together.

When we spoke to users of these tools at communication agencies, we found that they had preferred for a 'one-stop-solution' for analysing the news corpus instead of having to 'piece' together their analysis separately to develop their reports for media sensing.

1.2 Goal: Blending both datatypes to provide readers with useful snapshots of news headlines

The eventual goal of blending both datatypes and methods are so that the team can build an interactive media dashboard that can help readers of local news sites (both mainstream and non-mainstream) quickly navigate and understand a large corpus of news. This will be done by applying statistical methods to sort large volumes of unstructured data. The intended result is different snapshots that would allow readers to Explore the news corpus by sorting and identifying key events (by time period, topic and keyword). Discover in greater detail the context around each theme and news source (by word sequence) and Detect and be alerted to unsual patterns based on engagement metrics.

As such, while the presentation of the 'snapshot' may appear simple, the methods that would derive these insights are not. The approach taken for this project is based on a mixed of deep research of current statistical methods, domain knowledge of the local news reporting landscape and rigorous data wrangling. We will describe the purpose, methods used and design intent of each subsequent module in the subsequent sections of the paper.

2. REVIEW OF PAST WORKS

2.1 Analysing unsctured news data

The main challenge of the project is that of sorting and organizing unstructured data (224,351 news headlines and 2,692,224 observations) into topics that were useful for further analysis when quantified with engagement data. To do this, we needed robust method that could cluster all the news headlines into topics.

We are also motivated in the project, to avoid simple word cloud visualisations which was a common method for visualising word frequency in a corpus as they tended to have limited utility (Temple, 2019) and could easily mis-present data (Keatext, 2021). Often, they include common verbs that do very little to explain word in the context of the corpus of data and can harm meaningful insights. As mentioned earlier.

As our subject of analysis is English news headlines, we benefitted from research and corresponding methods that were available for analysing western media. News media also tended to adopt standardized styles of reporting which also helped our text pre-processing. For instance, text cleaning and sentiment analysis could easily adopt the English packages as they were unlikely to contain colloquial lingua franca that was most used on local social networking sites. Furthermore, the use of standardised stylebooks for news reporting would help the Latent Dirichlet Model which the team intends to use for analysing unstructured text data.

2.2 Why Latent Dirichlet Allocation (LDA) for analysing unstructured data?

Our intention to utilize the LDA model is mainly because it is a generative statistical model "that allows for sets of observations to be explained by unobserved groups that explain why some parts of the data are similar." It has been adopted in other parts of the industry such as banking.

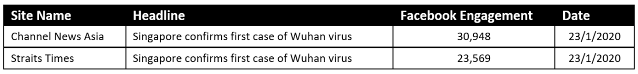
Although LDA has been criticized for issues with overfitting and lacks an intrinsic method for choosing several topics for clustering, it is a method that is extremely advantageous for the purposes of our study. This is because the method itself is entirely unsupervised and allows for the dashboard to assist readers in identifying topics that are 'unknown-unknowns' even though some of the topic models derived may not be immediately clear. This approach complements our intent for news analysis well because we are unable to 'predict' and therefore pre-define news topics for analysis in the eventual snapshot. Using an unsupervised approach with LDA, will also, in theory, enable the dashboard to be scalable and reproducible as it will be able to handle new datasets containing new or future news articles.

2.3 Analysing Structured Social Media Engagement Data

Social media engagement data has been frequently used as proxy for actual web visitorship data since the former is more publicly available. This has prompted many organizations and technology vendors to track online 'reach' and 'virality' by using social media engagement. Being able to rank topics identified using social media engagement can thus provide insight to the reach of different news topics. Further analysis can also be done to identify variances of engagement across topics and media sources.

Researchers have utilized social media engagement using CrowdTangle and NewsWhip to track virality of fake news articles. They have also proposed that readers pay attention the news source to ensure authenticity and the corresponding engagement metrics to determine if an article has been manipulated to reach a wider audience (e.g., when the engagement ratios for a content does not appear to be natural). This typically occurs when 'likes' or 'shares' are mobilized. As such, we intend to provide users with engagement metrics sot that they can explore the context surrounding a topic (i.e., sentiment and co-occurring words) and use engagement metrics to identify if there are anomalous articles.

Using social media metrics, NewsWhip and CrowdTangle can predict how viral an article is likely to be based on an algorithm that calculates how much an article is 'overperforming' when compared to the publisher's historical average. This allows for 'outliers' to be quickly identified. Using this method, they are also able to predict the trajectory of an article's social media engagement and predict its performance over the next hours. This can be useful for us to



Site Name	Headline	Facebook Engagement	Date
Channel News Asia	Singapore confirms first case of Wuhan virus	30,948	23/1/2020
Straits Times	Singapore confirms first case of Wuhan virus	23,569	23/1/2020

Figure 1: Table1

visualize anomalies in time series using packages in R.

Sorting engagement data by by topics identified by LDA can allow readers to identify if they are key topics that are trending and allow them to better understand the sources contributing to their report and the keywords used in their respective headlines.

2.4 Dataset used and Preparation Involved in Processing

The corpus of text articles to be analysed is obtained from the Prrt database, retrieved on February 22, 2021. The data selected for analysis is from 1 January 2020 to 31 December 2020. These local news sites are a mix of mainstream (Straits Times, Channel News Asia, AsiaOne) and non-mainstream (Mothership, MustShareNews, The Independent) sites. A preliminary query pulls 224,351 articles. Data extracted from Prrt is extremely useful for the project as it provides an easy tool for the extraction of the article headline, corresponding URL and Facebook Engagement metrics (Likes, Shares and Comments).

For this exercise, we will be extracting and performing analysis solely on article headlines. The reason for doing so as opposed to parsing the full article is due to a unique feature of media articles: in most cases, article headlines capture the key message of the article. We further filtered the dataset to observe only news articles with at least 10 Facebook engagement. Doing this reduces the number of articles in the dataset by 131,235 (or close to 58%) and removes articles that were (a) removed quickly after A/B testing or (b) removed by editors after corrections in the headlines or (b) republishe from international news wire agencies and therefore had low to no engagement.

Following our research into data pre-processing, we found that it was important to avoid using TF-IDF, a popular method used in text cleaning. TF-IDF was a method of focusing on 'interesting' or 'rare' terms in a text corpus by reducing the importance. of commonly used words that appear frequently in the corpus. However, as identified earlier, the nature of standardized style of news reports around common events will inevitably engender the occurrence of common terms. These 'common terms are therefore important and should not be reduced in 'weightage' when analyzing a corpus of text. Because of this, we will not be adopting the TF-IDF method for text cleaning for the purposes of this media dashboard.

3. DESIGN FRAMEWORK

As discussed in the above section, the goal of the dashboard is to allow users to explore, discover and detect anomalies in the news corpus. This section will discuss our approach towards achieving each in their respective modules within the Shiny dashboard. Our aim is to provide users the interactive tools to create snapshots that help their personal sensemaking of the news corpus based on the pre-loaded data - SG news headlines coverage from six mainstream and non-mainstream media outlets in 2020.

3.1 Explore – The news corpus to understand the topics contained within

3.1.1 Purpose

The purpose of this module is for readers to explore the entire news corpus (Gjerde, 2019) based on dominant topics. This will allow them to better understand the scale of coverage, the size of engagement and the keywords involved in each dominant topic.

3.1.2 Method Involved: LDA

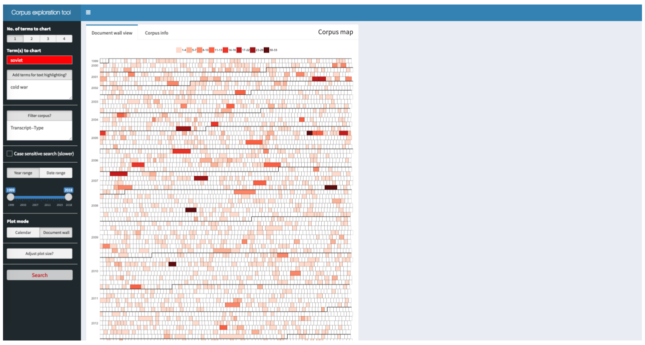
To organize unstructured data into topics for our explore module, we will be using the TM/Textminer package for creating the LDA model. The idea behind LDA is that (1) every document is a mixture of topics (2) every topic is a mixture of words. A simplified explanation of point (1) would be that a document may consists of 30% economics, 60% politics, and 10% social topic and a higher probability means that the document can be represented by the topic. According to the point (2), a topic is a mixture of words, where certain words have a strong association with a certain topic.

To determine the number of topics for analysis, we rely on the work done by David Meza to determine k number of topics using the harmonic mean. Harmonic mean is often used as an aggregated performance score for the evaluation of algorithms and systems: The F-score (or F-measure). In our case, we will calculate the harmonic mean based on the log-likelihood function calculated by the LDA algorithm of the 'tm' package. Since logarithm equations are increasing functions, is like maximizing log-likelihood values as recorded by the LDA function. Thus, this would be applicable to finding out the number of topics our model should have.

3.1.3 Design Choice:

For visualising the LDA results, we will be using LDAvis which provides interactive features that allows users to interact with the topic bubbles by hovering over them. This in turn provides additional data that would allow readers to discover word distribution within the topic and the corresponding headlines. They will also be able to change the word / lambda threshold using the dashboard feature, so as to make greater sense of the different topics if necessary. This gives users a sense of control in terms of customizing the snapshot that they want to see. LDAvis' visualisations are also clear because they provide a clean interface involving only 3 different colours. The subtle use of transparency allows for readers to see with clarity, overlaps in topics and keywords without too much visual clutter.

For this Explore module module, we have included a link to CorporaExplorer for users to further explore the news corpus. It is an excellent R package that allows for readers to quickly search a corpus of documents by up to 5 different keywords. The selected documents within the corpus will then be easily highlighted in the corpus wall and made available for retrieval.



Term(s) to chart: soviet

Add terms for text highlighting? cold war

Filter corpus? Transcript--Type

☐ Case sensitive search (slower)

Figure 2: Corpus Wall in Corpora Explorer

As seen in the figure above, the corpus wall in Corpora-Explorer is very effective for visualising how a keyword has appeared over the duration of the year. By clicking on the tile scrolls, users can also quickly retrieve engagement information and access the full article for further sense-making by clicking the embedded link.

3.2 Discover – In detail the context surrounding each topic in the media

3.2.1 Purpose

The purpose of the Discover module is to allow readers to discover in greater detail the context of the reported news headlines. To do this, we will allow users to input two words that they are interested in and generate the sequences of words that were likely to follow based on the news reports.

3.2.2 Method Involved: How we will generate tri-grams

To do this, we will utilize the r packages, tidyverse, tidytext and purrrlyr. Using TidyText's unnest_tokens function, we can tokenize the words in all the media headlines into consecutive sequence of words call n-grams. For this module, we will be using tri-grams, which are consecutive sequences of 3 words. This is about 50% of words in the length of the average recommended news headline (Santiago, 2019). By seeing how often each subsequent word follows the chosen word, we can build a model of the relationship between them for visualisation.

This provides readers with greater detail as it captures "structure that isn't present when one is just counting single words (Silge & Robinson , 2021)" thereby allowing readers to gain insight to the relationship and connections between words chosen by media in their reports.

3.2.3 Design Choice

Using trigrams identified in news headlines, readers should should be able to visualise how any two words in a selected corpus is used, where they are similar and where they diverge in the visualisation. Research into empirical linguistics have identified how the frequency of words occurring together can give insight and meaning to the treatment of a topic and the 'closeness' of association between entities. This can help readers understand the context to how words in media coverage was used when compared to each other.

As seen in the visualization created in the dashboard, the word tri-gram network can be elegantly displayed and compared with each other to see how they unfold more often in media coverage. This is more focussed than word networks as it allows for direct comparison between two keywords. Aesthetically, this involves a limited colour palette that allows for contrast (and thereby clarity) to identify the differences clearly, between the selected keywords.

3.3 Detect – Anomalies in news coverage

3.3.1 Purpose

Readers should be able to use the Detect module in order to identify potential anomalies in the news corpus based on social media engagement data. This will allow them to cut through an extensive corpus to quickly focus on the point of anomaly for further investigation.

3.3.2 Method Involved: How we will use social media interaction metrics for anomaly detection

For this, we will utilize the alpha and max anoms from the anomalize packages in R. These are parameters that are used to control the threshold of engagement data that can be then considered an anomaly. The alpha parameter determines the width of the critical values by default as 0.05. On the other hand, the max_anoms paramters control the maximum percentage of data that can be detected as anomalies with the default set at 5%.

In order to deal with seasonal trends that could affect engagement data, we decomposed the entire dataset as a time-series, in order to separate seasonal variations in data from observed values, leaving the remainder for anomaly detection. Once the seasonality data is is removed, anomaly detection is performed on the "remainder." Anomalies are identified, and boundaries (recomposed_l1 and recomposed_l2) are determined. The anomaly detection method uses an inner quartile range (IQR) of the +/-25 of the median.

3.3.3 Design Choice

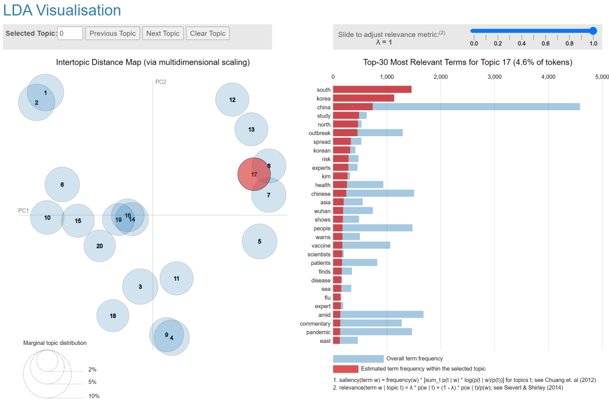
Using the interactive features for the visualization is key here in allowing readers detect anomalies in an extensive corpus. This is because they will be able to visually identify the anomaly, and interact with the datapoint to discover the potential articles that may have caused the anomaly.

Applying this visualisation techniques will also allow users to quickly select filter variables so that they can focus on specific periods of analysis as well as group by media type. In terms of clarity and aesthetics, the anomalous data point is also clearly segregated from the 'normal' band by distance and colour.

4. DEMONSTRATION

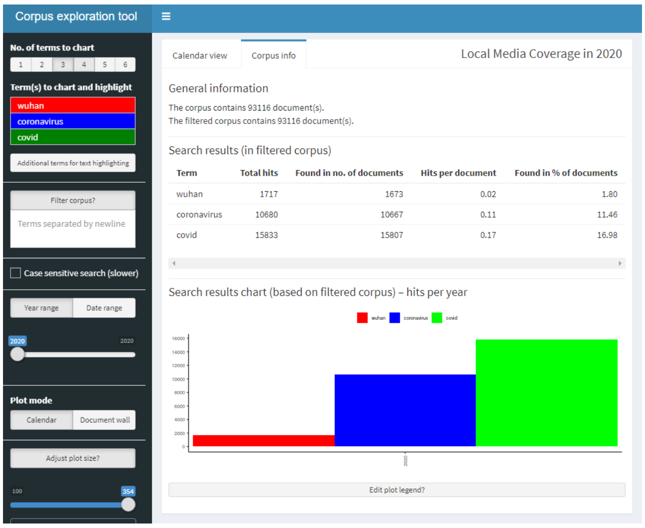
In the overview panel, users will be presented with an overview of all the news headslines in 2020 categorized by the topics and sorted by article count and engagement data. They will be able to then select from three modules that allow them to explore, discover and detect the news corpus.

4.1 Module 1: Explore



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3: Explore Topics with LDAViz



General information

The corpus contains 93116 document(s).
The filtered corpus contains 93116 document(s).

Search results (in filtered corpus)

Term	Total hits	Found in no. of documents	Hits per document	Found in % of documents
wuhan	1717	1673	0.02	1.80
coronavirus	10680	10667	0.11	11.46
covid	15833	15807	0.17	16.98

Search results chart (based on filtered corpus) – hits per year

Figure 4: Coverage of Covid-19

Using the R Package, LDAViz, readers will be able to explore the visualisation of all the 20 topics that were generated from the entire news corpus. This is visualised on the bubble chart on the left. Topics are close together (e.g. 8, 7 and 17) have overlaps because they share topic distribution based on the news headlines.

From the panel on the right, users should be able to tell from the bar charts, the word probabilities within the topic. Users will also be able to adjust the lambda in order to callibrate the generation of topics and associated keywords.

4.1.1 Insight

From the coporaexplorer, we can quickly derive 2 insights that can help us make sense of news coverage of COVID-19 in 2020.

First, CorporaExplorer is able to identify that there were 1673 news articles with the word Wuhan in the headline. This was much smaller than articles with the word COVID in the headline at 15,807 articles or nearly 17% of the new headlines analysed in 2020.

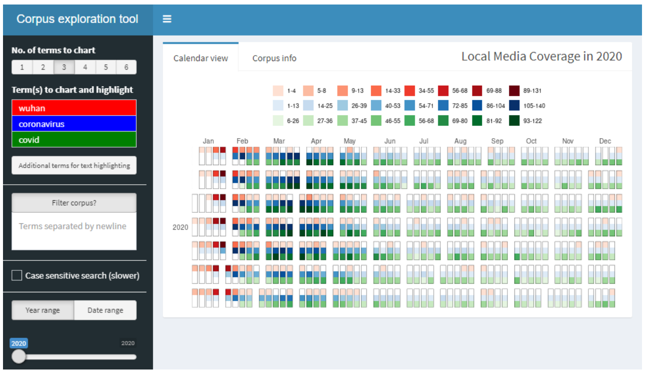


Figure 5: Coverage of Coronavirus Over 2020

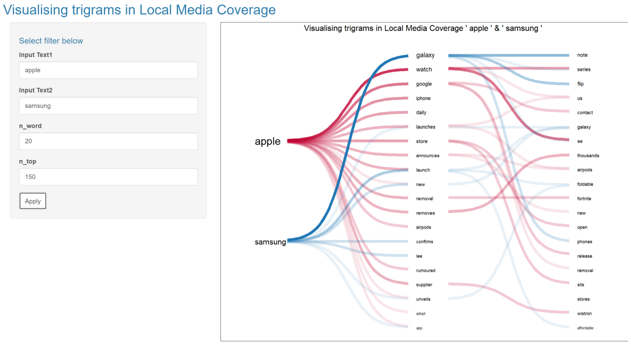


Figure 6: Discover Word Sequences in news reporting

Second, looking at the corpus wall using the three keywords, 'Wuhan,' 'Coronavirus' and 'Covid,' we can identify that local media was very quickly and effectively able to adapt and change their news reporting on the pandemic as seen in the change from using the term Wuhan virus (primarily in the month of January) to that of COVID. This is a contrast with Western media which continued to sporadically report the coronavirus as the Wuhan virus in late 2020.

4.2 Module 2: Discover

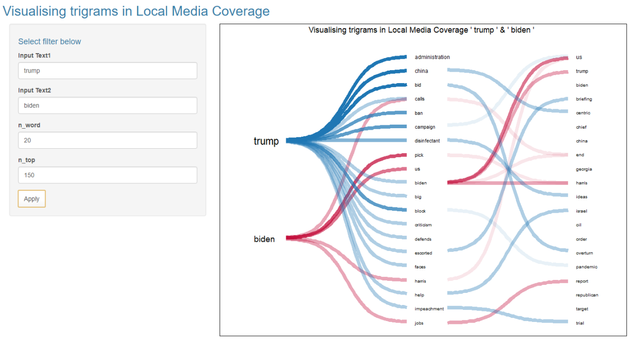
The second module allows users to input free text in the panel on the left in order to visualise the word sequences that follow the selected word.

The word sequences are extracted from the news corpus based on how it more frequently appears in news headlines. Flow lines are also visualised in contrasting colours and its intensity is based on frequency, allowing for a quick and elegant way of comparing two selected keywords.

4.2.1 Insight

Based on the sequence of words used in local media coverage, we are also able to quickly see that Trump was more likely to feature in local news reporting (from the thicker blue lines) 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 his focus on Jobs and the Economy.

4.3 Module 3: Detect



Visualising trigrams in Local Media Coverage

Select filter below

Input Text1: trump

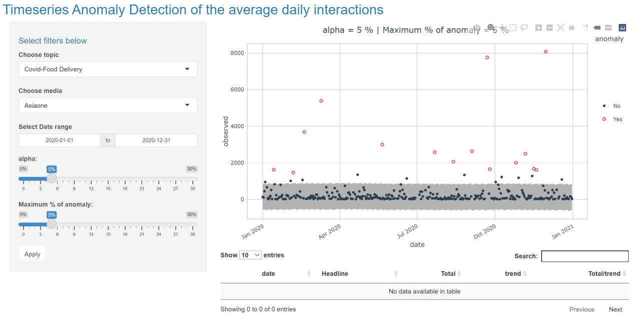
Input Text2: biden

n_word: 20

n_top: 150

Apply

Figure 7: Trump VS Biden



Timeseries Anomaly Detection of the average daily interactions

Select filters below

Choose topic: Covid-Food Delivery

Choose media: Asiaone

Select Date range: 2020-01-01 to 2020-12-31

alpha: 5%

Maximum % of anomaly: 5%

Apply

Show 10 entries

Search:

date	Headline	Total	trend	Total/trend
No data available in table				

Showing 0 to 0 of 0 entries

Previous Next

Figure 8: Detection Anomalies

Lastly, for anomaly detection, users of the module, Detect, will be able to customise their search terms based on a series of features that can be narrowed to their specific areas of interest.

This allows them to apply meaningful and relevant features such as topic, media and date. They will be able to then adjust the threshold for the anomaly detection to widen or narrow the points that would be reflected as a anomaly on the time series plot. Lastly, in order to get more detail into the potential causes of the anomalies, users will be able to click on the plots to view the articles with 'anomalous' engagement metrics.

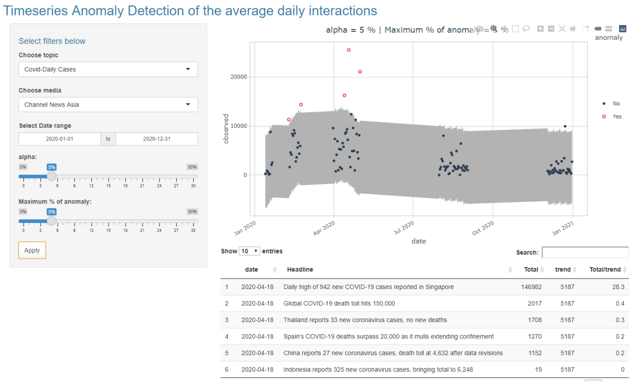
4.3.1 *Insight*

As seen in the figure above, and anomaly detected on 18 April 2020 was directly related to the article 'Daily high of 942 new Covid-19 cases reported in Singapore.' The anomaly detection feature was useful here in identifying the article and event that caused the anomaly. Such anomaly detection features can easily help to cut through an expansive corpus of news to identify key events or articles that could be of potential importance.

5. DISCUSSION

Based on the initial feedback gathered from technological vendors, communication practitioners and regular users, there were three key areas for further discussion.

5.1 Usability of the Dashboard



Timeseries Anomaly Detection of the average daily interactions

Select filters below

Choose topic: Covid-Daily Cases

Choose media: Channel News Asia

Select Date range: 2020-01-01 to 2020-12-31

alpha: 5%

Maximum % of anomaly: 5%

Apply

Show 10 entries

Search:

	date	Headline	Total	trend	Total/trend
1	2020-04-18	Daily high of 942 new COVID-19 cases reported in Singapore	146982	5187	28.3
2	2020-04-18	Global COVID-19 death toll hits 150,000	2017	5187	0.4
3	2020-04-18	Thailand reports 33 new coronavirus cases, no new deaths	1708	5187	0.3
4	2020-04-18	Spain's COVID-19 deaths surpass 20,000 as it mulls extending confinement	1270	5187	0.2
5	2020-04-18	China reports 27 new coronavirus cases, death toll at 4,632 after data revisions	1152	5187	0.2
6	2020-04-18	Indonesia reports 325 new coronavirus cases, bringing total to 6,248	19	5187	0

Figure 9: Anomaly in Covid Daily Cases

While the technology vendors and communication practitioners were able to easily use the dashboard to derive insights, users with little or no background in media analytics did not. This group also tended to ask questions on the different functions in the dashboard and the terminology used within the dashboard. To improve usability for the third group of users who had little to no background in media analytics, we updated the dashboard with more visual cues and replaced confusing terminology (e.g. n-gram) with easier to understand words.

5.2 Precision limitations in unsupervised topic modeling

While the LDA model was great at deriving topics without human intervention, it did not always generate topics that were immediately coherent or easily interpretable. Different variations of the LDA method will also most likely produce different results. As such, this poses questions in relation to the amount of precision required by users who intend to adopt such a system for their media analysis.

5.3 Scalability of eventual system

A system such as this will need to be scalable, to deal with the everchanging news landscape. This means that the system will need to be able to process new news articles from the same sources, or different local mainstream and non-mainstream news sources. As the dashboard was intentionally designed to rely on unsupervised methods, users will be able to load a different corpus for analysis. This ensures the scalability and reproducibility of the system.

6. FUTURE WORK

While there are many aspects to explore to help make sense of the news corpus, we have selected packages and methods that would best suit a large volume of unstructured data and contain interactive user functions to allow them to quickly navigate the corpus and create snapshots around topics they are interested in.

There are three major areas that can be worked on for the dashboard. The first being an extension of 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 and thus benefit their sense-making workflows. As such, they will benefit from being able to gain insights from a larger corpus of connected data, whilst making gains in time-savings (The Economist, 2020).

The second area for future work is being able to identify and visualise how a topic evolves in a new corpus (Sheidin & Lanir, 2017). This can provide greater context on a topic's evolution (e.g., based on engagement) as topics do not remain static across the year. Understanding how news media treats the headlines, together with engagement scores, can help readers to discern the inflection points (Whitney, 2014) involved in how a news topic develops.

Lastly, if we had more expertise and knowledge in branching off packages, we would be most interested in working off Kristian Lundby Gjerde's Corpora Explorer package for R. Corpora Explorer is a very neat and quick package for exploring all the 'books' within a package - but is unfortunately not customised for news articles. The main difference being the volume of a corpus of news articles is generally greater than a corpus of books. Being able to customise the Corpora Explorer package could thus greatly enhance the usability of the dashboard.

7. REFERENCES

1. Bhargava , R., Bishop, C., & Zuckerman, E. (March, 2020). Mapping and Visualizing News Images for Media Research. Retrieved from Northeastern EDU: https://cpb-us-w2.wpmucdn.com/sites.northeastern.edu/dist/d/53/files/2020/02/CJ_2020_paper_39.pdf

2. Gjerde, K. L. (2019). corporaexplorer: An R package for dynamic exploration of text collections. Journal of Open Source Software 4, 1342.

3. Grimes, D. R. (4 December , 2017). Echo chambers are dangerous - we must try to break free of our online bubbles. Retrieved from Guardian: https://www.theguardian.com/science/blog/2017/dec/04/echo-chambers-are-dangerous-we-must-try-to-break-free-of-our-online-bubbles

4. Keatext. (15 April, 2021). When word clouds are not enough: Using AI text analytics to reveal insights within data. Retrieved from keatext: https://www.keatext.ai/en/blog/artificial-intelligence/3-strengths-and-3-weaknesses-of-word-clouds

5. Koivunen-Niemi, L., & Masoodian, M. (2020). Visualizing narrative patterns in online news media. Multimedia Tools and Applications, 919–946.

6. Patel, N. (23 March, 2021). The Evolution of headlines at the New York Times. Retrieved from Columbia Journalism Review: http://experiment.cjr.org/experiment/features/the-evolution-of-headlines-at-the-new-york-times

7. Santiago, A. (5 August, 2019). How Long Should a Press Release's Headline Be. Retrieved from Newswire: https://www.newswire.com/blog/how-long-should-a-press-releases-headline-be

8. Seneca, C. (17 September, 2020). How to break out of your social media echo chamber. Retrieved from Wired: https://www.wired.com/story/facebook-twitter-echo-chamber-confirmation-bias

9. Sheidin, J., & Lanir, J. (2017). Visualising Spatial-Temporal Evaluation of News Stories. IUI, 65-68.

10. Silge, J., & Robinson , D. (6 April, 2021). Relationships between words: n-grams and correlations. Retrieved from Text Mining with R: A Tidy Approach: https://www.tidytextmining.com/ngrams.html

11. Temple, S. (13 May, 2019). Word Clouds are Lame. Retrieved from Towards Data Science : https://towardsdatascien
com/word-clouds-are-lame-263d9cbc49b7

12. The Economist. (10 September, 2020). Facebook offers a distorted view of American news. Retrieved from The Economist: https://www.economist.com/graphic-detail/2020/09/10/facebook-offers-a-distorted-view-of-american-news

13. Toros, E. (2019). Text Analysis of Newspaper News on Electoral Integrity and Electoral Violence in Turkey. Retrieved from Rpubs: https://rpubs.com/emretoros/dievt

14. Whitney, H. (30 September, 2014). It's About Time Visualizing temporal data to reveal patterns and stories. Retrieved from UX Mag: https://uxmag.com/articles/its-about-time

...