UnwRappler

Revealing Rappler's underlying focus

Proponents

MSDS 2020 - Learning Team 01
Carmelita Esclanda
George Allan Esleta
Sandro Luis Silva
Elmer Robles

Executive Summary

Rappler, one of the leading online news publishers in the Philippines, seeks to inspire community engagement and create action for social change through citizen journalism. However, it has recently been under scrutiny by the Philippine government for "twisted" reporting. This study aims to uncover the underlying themes of Rappler news articles via unsupervised clustering techniques and see if there is an inherent concentration of news in a specific theme. A total of 11,079 articles were extracted from Rappler's national news section from January 2018 to May 2019. The articles were vectorized with a term frequency-inverse document frequency (TF-IDF) weighting and dimensions were reduced by implementing Latent Semantic Analysis (LSA). Lastly, unsupervised clustering via k-means algorithm was applied to group the articles and internal validation metrics were utilized to determine the optimal number of clusters. Ten clusters were uncovered, with Philippine president Rodrigo Duterte as the dominant cluster. The remaining themes touch on different branches of government, police and weather updates, and trending national issues. The insights gained from this research can aid Rappler in balancing its reporting by lessening bias towards specific topics.

Introduction

Rappler is one of the leading online news publishers in the Philippines. It was started in 2011 by former CNN journalist Maria Ressa. However, since 2018, Rappler has been under heavy scrutiny by the Philippine government for "twisted" and "biased reporting". Philippine President Rodrigo Duterte even went so far as calling Rappler a "fake news" outlet that publishes articles that are "pregnant with falsity." [1] Rappler has been on the receiving end of criminal cases from the Philippine governments. BIR filed several counts of tax evasion charges against Rappler CEO Maria Ressa. Cyberlibel cases were also filed against Ressa.

This begs the question “Why is Rappler being targeted by the Duterte Administration?” Is Rappler https://www.rappler.com/nation/197230-duterte-rappler-ban-twisted-reporting really “twisted” and “biased” in its reporting? To answer this question, we uncovered the underlying themes of Rappler’s news articles via unsupervised clustering, to see if there is indeed an inherent concentration of news in a specific topic or theme.

Business Value

Extracting themes from a set of articles programmatically has a widespread application in the digital publishing industry and can be used to deliver business value to readers, journalists, advertisers, aggregators, and researchers.

- **Readers** in the digital age have no patience. They are more likely to use a search engine to retrieve a set of ranked articles from multiple news sources and it becomes a challenge for online news platforms to engage a reader continuously. Automatic theme extraction and clustering of articles provide the ability to structure a set of articles in order to make it easier for readers to navigate to related articles. If articles can be presented in a logical hierarchy driven by patterns in the data, human intervention is minimized (saving on labor costs) and the reader is kept on the news platform, increasing engagement time on the site.

- **News Publishers** and **Journalists** seeking to address topics of broad interest to their readership can benefit from a thematic clustering articles to determine the balance of news coverage. Publishing organizations are always looking for gaps in news coverage in order to offer readers some perspective and insight into topics that are not heavily covered. By examining existing themes holistically, they can better gauge what is missing and what their next article should be about.

- **Advertisers** that are considering online news platforms to reach specific customer segments can tailor their messages so that they are congruent with the themes that readers are interested in. For instance, themes dealing with legislative or judicial proceedings may appeal to advertisers whose clients are lawyers or law schools. Advertisers of tourism and tourist agencies may be interested in weather and specific vacation destinations, such as Boracay.

- **Aggregators** in the licensed publishing industry create packages of content for redistribution. Their business model, which can be subscription or advertiser-based, is to target specific customers based on their geographic, demographic, and psychographic profile. For instance, if an investor wanted to understand market conditions for a particular industry, they would subscribe to a real-time news delivery service that was able to provide the relevant articles.

- **Researchers** make money by selling analytical reports. By studying the themes presented on news platforms and combining the study with other metrics such as readership engagement, sentiment analysis, and political stance, they can make in-depth comparisons with other news delivery platforms to better inform advertisers and aggregators.
    1.

Methodology

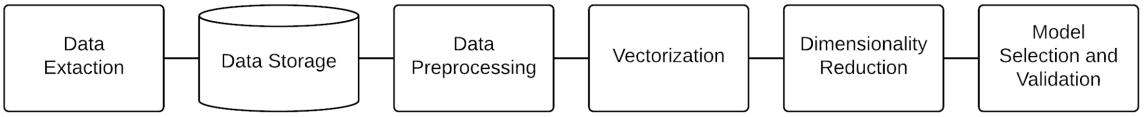

Figure 1. High-level pipeline for clustering Rappler news articles

Data Description

Data Processing

a. **Data Extraction** Web scraping tools, specifically Python's Requests and Beautifulsoup modules were implemented on the Rappler's nation subsection website. The process involved extracting the article text and selected meta-data. Initial cleaning was also implemented to remove image captions, author and photographer details, location headers, hyperlinks, social media texts, and general sign offs.

b. **Data Storage:** The extracted data was stored in a local database via Python's SQLite3 module. An estimated 11,079 articles were stored in the database. The data description is as follows:

Data	Data type	Description
article_id	VARCHAR	Unique ID identifier
url	VARCHAR	Relative URL
headline	VARCHAR	Headline title
metadesc	VARCHAR	Quick synopsis of the article
label	VARCHAR	Absolute URL
author	VARCHAR	Author name(s)
published_date	VARCHAR	Published date
updated_date	VARCHAR	Date updated
article	VARCHAR	Article text
metakey	VARCHAR	List of metatags associated with the article

c. **Data Preprocessing:** Data preprocessing was implemented on the acquired article text. The preprocessing:
   - Neutralizing text case-sensitivity by converting text to lowercase
   - Omitting unnecessary whitespaces by removing leading and trailing whitespace
   - Converting words to root form by performing stemming, via NLTK's porter stemmer.

d. **Feature Extraction/Vectorization:** The term frequency-inverse document frequency (TF-IDF) vectorizer was implemented, via scikit-learn's Tfidfvectorizer, to vectorize the text. As opposed to an equal weighting vectorizer, the TF-IDF statistic measures how important a word is to the document and the corpus. Additional parameters were taken into consideration:
   - Removal of English stopwords
   - N-gram range to include unigram, bigrams, trigrams
   - Words that appeared less than 0.1% or more than 70% were discarded

e. **Dimensionality Reduction:** The Latent Semantic Analysis(LSA) was implemented to reduce the number of components. LSA was chosen over Principle Component Analysis(PCA) since LSA decomposes the term-document matrix, as opposed to PCA which performs decomposition through the covariance matrix. Sensitivity analysis on the number of components were performed. The sensitivity analysis started with 50 components up until 1000 components, with steps of 50. The analysis observed using 300 components extracted the broader underlying themes.

f. **Model Selection and Validation:** K-means clustering was implemented on the resulting design matrix. The analysis performed sensitivity analysis for different cluster counts, ranging from 2 to 20. Selecting the appropriate number of clusters relied on examining the internal validation criteria: (1) minimizing the Intracluster to Intercluster distance ratio, (2) maximizing the Calinski-Harabasz score, (3) maximizing Silhouette coefficient, and (4) minimizing the sum of square distance to centroids coefficient. Selecting 10 clusters satisfied the four criteria.

Exploratory Data Analysis

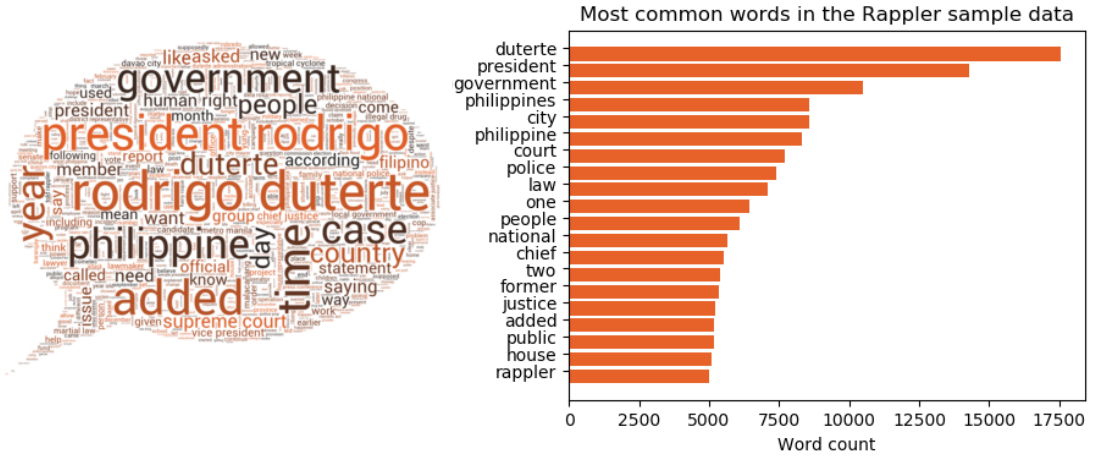

Figure 2. Word cloud of Rappler news articles from January 2018 to May 2019.

Results

To determine the optimal number of clusters, we perform k-means clustering for different values of the cluster count k from k=2 to
k=20. This range of values was chosen to keep the clustering parsimonious.

Several internal validation measures were computed for each value of
k. The internal validation measures that were used are (a) intracluster to intercluster distance ratio, (b) Calinski-Harabasz score (c) Silhouette coefficient, and (d) sum of square distances to centroids coefficient. The optimal value of k is then chosen based on the following criteria:

1. Sum-of-square distance to centroid is minimized
2. Calinski-Harabasz index is maximized
3. Intracluster to intercluster distance ratio is minimized
4. Silhouette coefficient is maximized

Plotted below are the values of the internal validation criteria for values of k between 2 and 20. Using the elbow method, k=10 is a good candidate for the number of clusters.

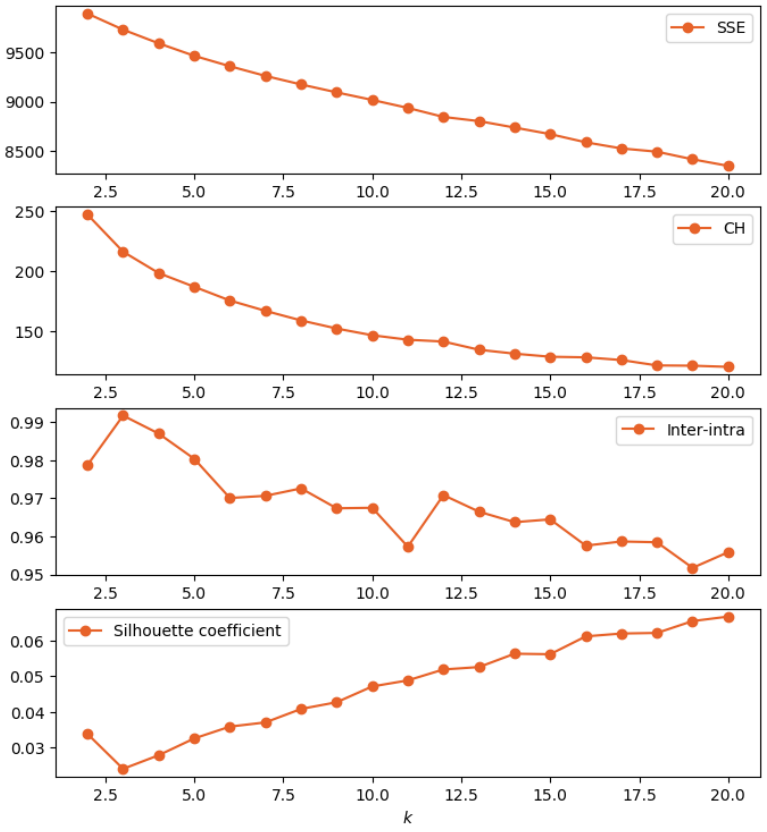

Figure ## Internal validation criteria

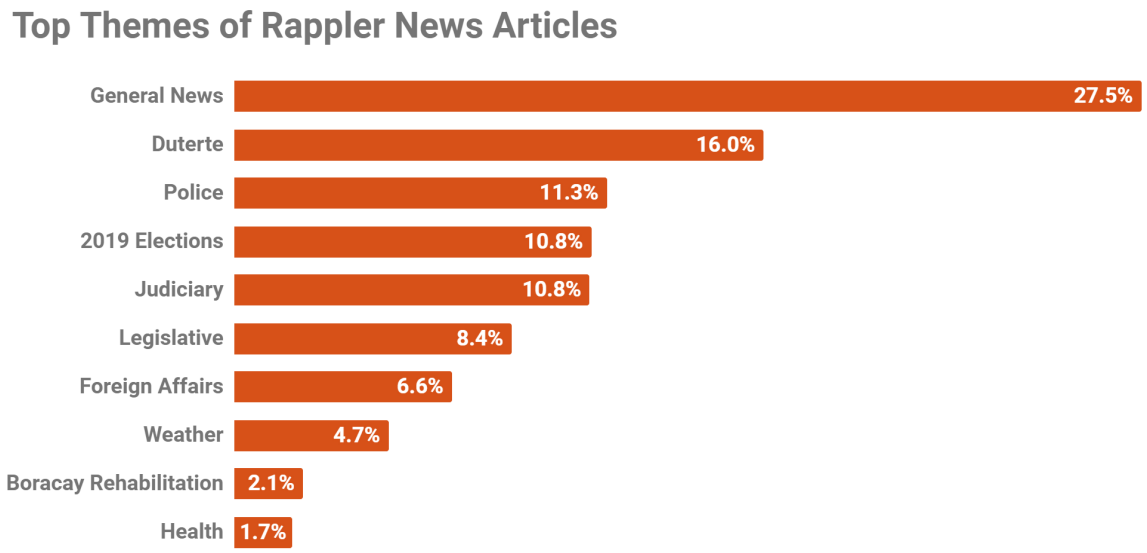

Figure 3. Underlying themes of Rappler news articles.

Insights

The application of unsupervised clustering technique on a corpus of 11, 079 news articles shows consistency of topic clusters across different resolutions that include:

A. Duterte article cluster

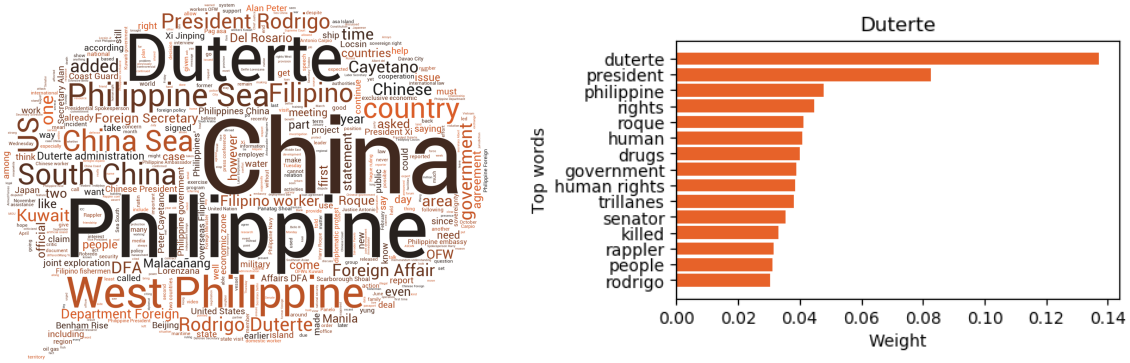

Figure 4. (a) Wordcloud and (b) Top words in the Duterte article cluster

The **dominant theme of Rappler news articles is Philippine President Rodrigo Duterte**, with nearly 16% of news articles in this cluster. Subthemes in this cluster include the war on drugs, human rights, Senator Antonio Trillanes, and Duterte's ongoing war versus Rappler. This can imply that Rappler is particular with the President's undertakings and that while this may be reasonable, Rappler has to review its focus given their strife with the President.

B. Philippine Politics article cluster (Legislative and Judiciary)

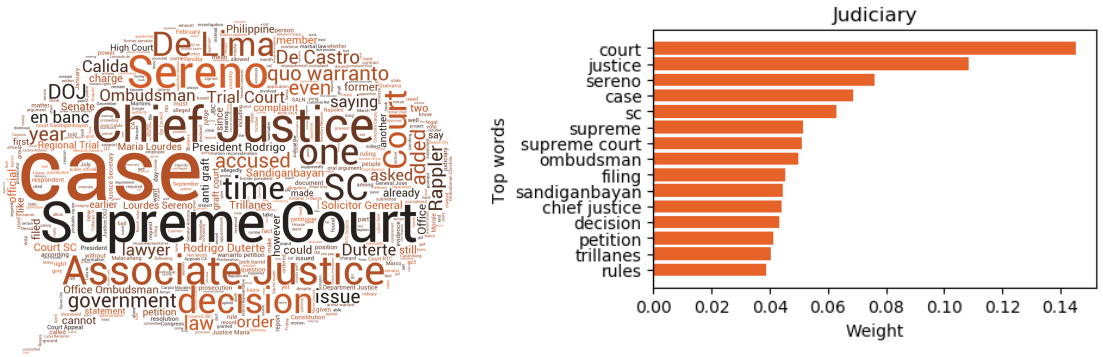

Figure 5. (a) Wordcloud and (b) Top words in the Judiciary article cluster

The other major themes are related to the other branches of the Philippine government, particularly the **Legislative branch** (House of Representatives, Senate) and the **Judicial branch (Supreme Court).** Subthemes in the Judicial cluster are the impeachment of former Chief Justice Maria Lourdes Sereno and Leila de Lima. Impeachment of former Chief Justice Lourdes Sereno because of failure to disclose assets in her SALN, and misuse of public funds among other things. Drug charges filed by the DOJ against Senator Leila de Lima. Subthemes in the Legislative article cluster are budget Proposed changes in the constitution to shift to a federal type of government. Like the dominant clusters, this cluster focus on politics. Hence, it is apparent that Rappler is focused on the political landscape in the Philippines with particular focus on hot and relevant issues.

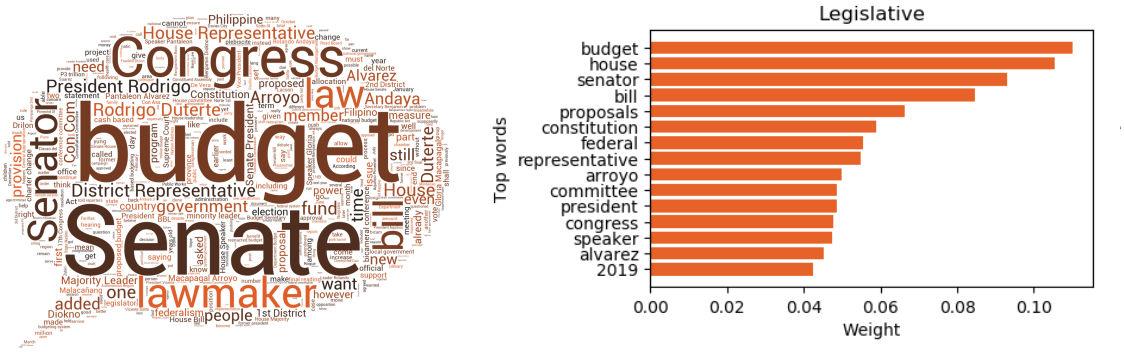

Figure 6. (a) Wordcloud and (b) Top words in the Legislative article cluster

C. Police and Weather clusters

Two recurring themes regardless of the number of clusters selected are police reports and weather reports. In line with the duterte cluster, it is sensible that the cluster about the police can be generated given the “oplan tokhang” implemented by the administration. Meanwhile, the weather cluster is sound considering the country’s geographic circumstances.

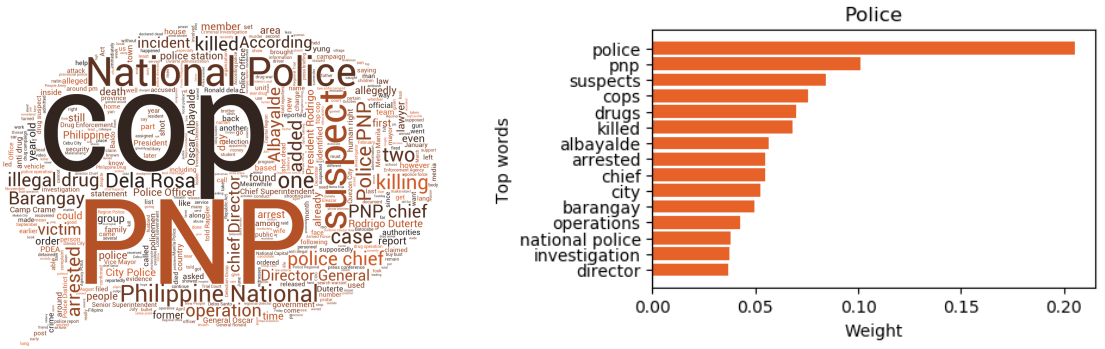

Figure 7. (a) Wordcloud and (b) Top words in the Police article cluster

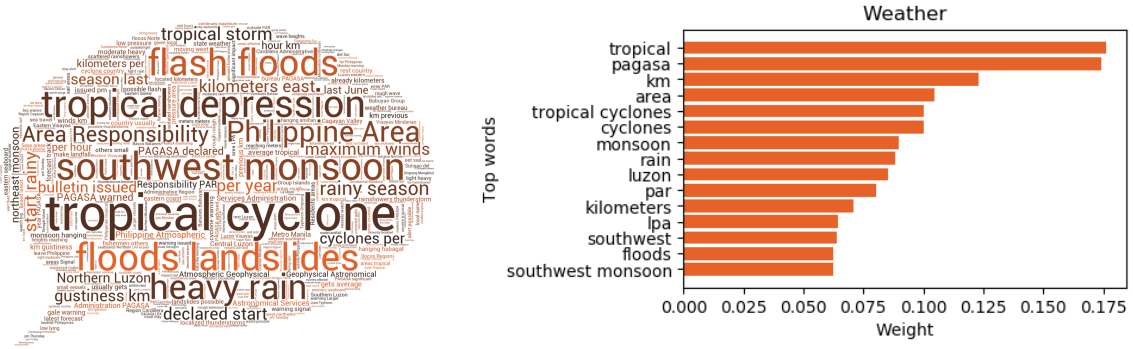

Figure 8. (a) Wordcloud and (b) Top words in the Weather article cluster

D. Time-specific article clusters

The remaining themes are time-related and reflect major events during the timeframe selected (January 2019 to May 2019), such as the **2019 Philippine Midterm Elections**, **Boracay Island Rehabilitation** (April 2018-October 2018), and the **Dengvaxia (dengue vaccine) controversy.**

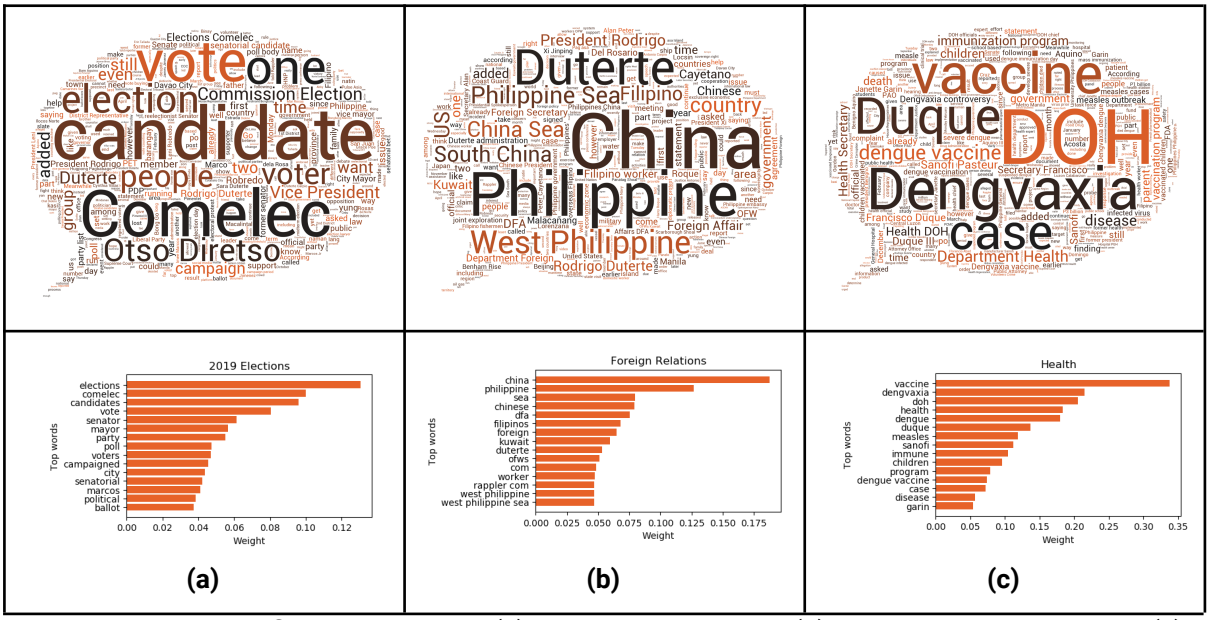

Figure 9. Time specific article clusters (a) 2019 Election cluster (b) Philippine-China relations (c) Dengue vaccine controversy

Sensitivity analysis

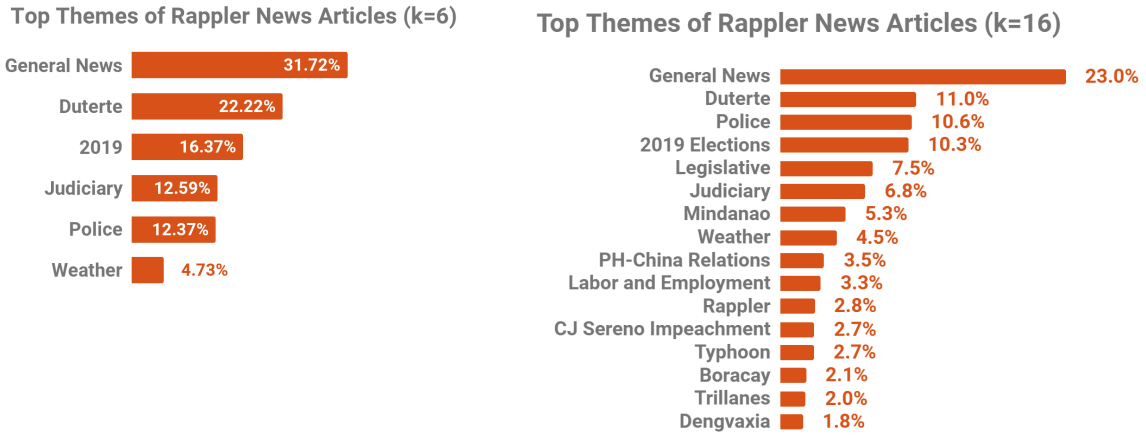

Figure 10. Uncovered article themes for (a) k=6, and (b) k=16

While it was concluded that 10 is the optimal number of groupings for the subject dataset, sensitivity analysis has been performed to test the robustness of the results gathered. From figures presented above, it is noted that as the number of clusters formed changes, the most affected clusters are bigger ones, especially the *Duterte* cluster. As such, it can be deduced that said cluster has many subclusters which are differentiated when clustering is increased. Nevertheless, the core clusters (e.g., General News, Duterte, 2019 Election, Police) created regardless of the number of clusters are constant.

Summary & Conclusions

Grouping of articles into clusters of related topics by unsupervised clustering method has significant value to communication researchers and media practitioners in studying news output at scale and its repercussions. This study was able to *unwrap 10 major themes* ranging from General News, Politics, Weather, Health and relevant events. As such, the result can be used as a starting point for a generalized theme extraction project from a national corpus in order to learn the general interest and sentiments of the people.

Recommendations

The following points can be considered in future research related to this work:

A. **Comparative cluster analysis with other Philippine news outfits** (Inquirer, ABS-CBN News, GMA News) can be explored to validate Rappler's focus on particular topics;
B. **Sentiment analysis can explored to look into the subjective information or emotional states of the article.** The value of combining clustering with sentiment analysis could be implemented in making better sense of Rappler's public opinion on trending issues or the current administration; and,

C. **Historical analysis can be explored to compare the rappler data during previous administrations with the current to recognize difference in rappler's focus per administration.** Administration changes can affect the political landscape and prioritization of complex issues. For instance, if an issue from one administration was solved in the next administration.

References:

[1] https://www.rappler.com/nation/197230-duterte-rappler-ban-twisted-reporting
[2] https://www.rappler.com/nation/193806-duterte-fake-news-outlet
[3] https://towardsdatascience.com/all-the-news-17fa34b52b9d