# Exploring Malaysia's COVID-19 News Using N-gram and Concordances Methods

Tang Wen Shuen
School of Computer Science
Universiti Sains Malaysia
Penang, Malaysia
wenshuen_tang@student.usm.my

Chan Siang Sheng
School of Computer Science
Universiti Sains Malaysia
Penang, Malaysia
siangsheng.chan@student.usm.my

***Abstract*—This project is to explore and study the text exploration methods such as n-gram and concordances in Malaysia's COVID-19 Domain. We have chosen this domain due to the uprising and popularity of COVID-19 news in Malaysia and our sample data will be largely articles that is specifically related to Malaysia's COVID-19. We prepared 21 sets of sample data through internet for the use in text exploration. In this paper, we will be explaining the overview, describing the dataset, analyze and discuss the text exploration and make a final conclusion at the end of paper. At the end, we had found out how useful is n-gram in detecting important phrases and how concordances can be used to understand the relative context of a term, find similar terms, and understand about the acronyms.**

***Keywords—n-gram, concordances, Covid-19, Malaysia***

## I. OVERVIEW

There are some NLP problems exhibited in Malaysia's COVID-19 news domain. One of them would be the coreference resolution. For example, in the sentence "Minister Khairy Jamaluddin is discussing with the health ministry on the issue. He said that he will announce it after it is discussed", the NLP model cannot easily identify whether the words "he" and "it" are referring to "Minister Khairy Jamaluddin", "health ministry", or "issue". Besides, another NLP problem in the domain above would be understanding the abbreviations and acronyms. For example, it might be difficult for the NLP model to understand that "MoH" is the acronym of "Ministry of Health" and "Dr" is the abbreviation of "Doctor". Apart from that, Malaysia's COVID-19 news domain also suffers from NLP problems like detecting the phrases such as "Ministry of Health", "critical cases", etc.

Coreference resolution and understanding of abbreviations and acronyms in Malaysia's COVID-19 news domain could be solved through analyzing the Concordances of the selected terms. For example, by analyzing concordances of the term "he" in the sentence "Minister Khairy Jamaluddin is discussing with the Ministry of Health (MoH) on the issue. He said that he will announce it after it is discussed", we are able to see "he" is referring to "Minister Khairy Jamaluddin" and "MoH" is the acronym of "Ministry of Health".

Phrase detection in Malaysia's COVID-19 news domain could be solved through analyzing n-grams of the selected terms. For example, if we talk a look at the trigrams of the term "Ministry", it is more likely that we are able to catch the phrase "Ministry of Health".

## II. SCENARIO

### A. Dataset Background

The dataset of this project is obtained through online news articles and there are several factors supporting this method. One of the factors is due to COVID-19 being a relatively new popular phenomenon and it has grabbed everyone's attention. As such, a lot of news reporter taking advantage of the internet to cover COVID-19 topic on social media or online platform. This has also caused abundance of online news articles with COVID-19 coverage, and it is the easier method to obtain dataset compared to other methods such as looking through newspapers. Besides that, online news articles also easier for us to process the text for text exploration with related methods such as n-gram and concordances. With that said, this project will be focusing only on Malaysia's COVID-19 news, we will be able to get the insight and understand the convey of Malaysia's COVID-19 news from news reporters with the usage of n-gram and concordances in text exploration.

### B. Dataset Methodology

First of all, we search for dataset by inserting "Malaysia's COVID-19" as query in Google search engine. We looked for 21 news articles that have coverage of Malaysia's COVID-19 and determine if the articles are relevant to the topic and have substantial sentences for text exploration use. Next, we extract the text from the news articles by copying and pasting the text into 21 notepads respectively and saved as text file format. We didn't do any data preprocessing and this issue will be discussed in Section C Comparison of Text Exploration. Last but not least, we will import the dataset into AntConc which is a useful tool to analyze corpus especially in concordances and n-gram.

### C. Dataset Description

Before moving towards text exploration with n-gram and concordances method, we would like to further describe the dataset with some basic information such as word type. Our dataset is composed of 21 text files and the content of each text file is extracted from one news article respectively. With a total of 21 text files as input, we have 7316 word tokens and 1543 word type.

## III. RELATED METHODS

### A. N-gram based language processing using Twitter dataset to identify COVID-19 patients

Peoples often report themselves on the social media if they were diagnosed with COVID-19. So, N-gram is applied in this paper to analyse the Twitter dataset and identify the potential COVID-19 patients [1].

Since the Twitter dataset contains multilingual data, character n-gram is used to identify the language by looking at the first character of the words as non-Latin script languages such as Cyrillic script does not use Latin alphabets.

After the language of the tweet is identified, word n-gram is used to extract the important feature which is the groups of word that can identify the COVID-19 patients such as "COVID-19 Positive".

### *B. N -gram Density based Malware Detection*

The advancement of malware forced the researchers to attempt to detect the malware on the host environment's native opcodes at run-time. N-gram is applied in this paper to build the opcode density features based on native opcodes at run-time [2]. SVM classification model is used to classify whether the features are a malware or not. The illustration of the n-gram structure of the opcode density features is shown in Fig. 1.

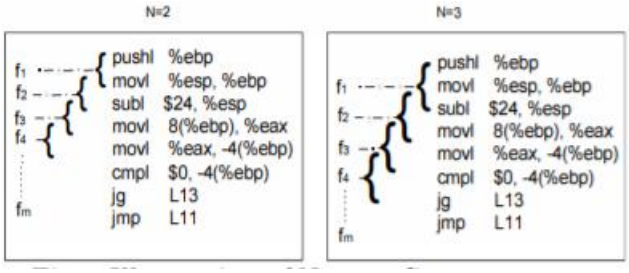


Fig. 1 Illustration of N-gram structure.

### *C. Using Google Ngram Viewer for Scientific Referencing and History of Science*

Google Books contains a lot of digitalized texts which spans across a lot of subjects. Google Books provides a useful tool named as Ngram Viewer which can show yearly count of n-grams. This paper attempts to prove Ngram Viewer as a useful tool for science research [3], not only for history of science, but also obtaining references which are less familiar to researchers.

### *D. Computational Concordance Analysis of Fictional Literary Work*

This paper attempts to apply the computational analysis of concordances in the text of George Orwell's novel 1984, a dystopian literature to examine the possibilities to measure the true, totalitarian nature of the concepts and catch phrases such as "Big Brother" or "Newspeak" which are used affirmatively by the characters in the text. [4]

### *E. Using Concordance to Decode the Ideological Weight of Lexis in Learning Narrative Literature: A Computational Approach*

Narrative literature contains a lot of texts, which causes the difficulty in understanding the core ideological messages of the literature. This paper attempts to apply the computational concordances to understand the themes and ideological meanings of the selected terms in the novel [5] in a more accurate, credible and faster way based on Frequency Distribution (FD) and Key Word In Context (KWIC).

### *F. Automatic Concordance Creation for Texts in Any Language*

The Bible Societies had translated and published scripture in many different languages, but they are lacking the general knowledge. It is also quite challenging to identify the key periscopes of scripture which consists of over 1000 pages. Thus, this paper attempts to build the concordance of the Bible in the original language which lists the important narratives and other key areas of the text. The concordance is then used as a model for a similar concordance in another language [6].

## IV. TEXT EXPLORATION

### *A. N-grams*

#### *1) Top 10 Most Frequently Occurred Bigrams*

We had attempted to look for the most frequently appeared bigrams in our dataset. Table 1 shows the top 10 most frequently occurred bigrams in our dataset.

TABLE I
Top 10 Most Frequently Occurred Bigrams

| Bigram | Frequency |
|---|---|
| of the | 45 |
| in the | 27 |
| he said | 24 |
| per cent | 23 |
| kuala lumpur | 21 |
| fully vaccinated | 20 |
| covid vaccine | 19 |
| to the | 18 |
| from the | 17 |
| number of | 17 |

From the result of Table 1, we can see that there are 6 out of 10 most frequently occurred bigrams such as "of the", "in the", "he said", "to the", "from the", and "number of" are mainly consists of the stopwords. This shows that our dataset should be properly cleaned so that these stopwords can be removed and make the result more viable.

Nevertheless, we are still able to obtain some useful insights from the obtained result. Firstly, we can see that the term "kuala lumpur" appears at the 5$^{th}$ position. This shows that most of the news sources are obtained at Kuala Lumpur, which is the capital of Malaysia. Since most of the COVID-19 news sources are provided by the Ministry of Health, we could also infer that the main location of the live announcement of COVID-19 news by Ministry of Health is mostly likely be held at Kuala Lumpur.

Secondly, we have "fully vaccinated" at the 6$^{th}$ position. From here, we can see that Malaysia government are working hard towards building a fully vaccinated community by encouraging all citizens to be fully vaccinated so that the infection rate drops, and the economy can return to a much stable state. On another insight, we could also infer that the fully vaccination progress among the citizens is on the right track as it is often stated in the articles.

Thirdly, we have the bigram "covid vaccine" appears at the 7$^{th}$ position. This shows that how important COVID-19 vaccine are currently, especially in Malaysia. The constant supply of COVID-19 vaccines can ensure the steady increase in the vaccination rate among Malaysia citizens.

#### *2) Top 5 Most Common Bigrams Started with "covid-19" word*

We would also like to know about some common phrases related to COVID-19. To achieve this, we had attempted to investigate the top 5 most common bigrams that started with the word "covid-19" and the result is displayed in Table 2.

TABLE II

Top 5 Most Common Bigrams Started with “covid-19”

| Bigram* | Probability | Frequency |
|---|---|---|
| covid-19 vaccine | 0.279 | 26 |
| covid-19 cases | 0.129 | 12 |
| covid-19 infection | 0.065 | 6 |
| covid-19 immunisation | 0.054 | 5 |
| covid-19 vaccination | 0.054 | 5 |

* Results are processed to treat singular and plural forms the same

From the result in Table 2, we can see there are a lot of 2-words phrases related to COVID-19. Among 5 phrases, we can see that the phrase “covid-19 vaccine” stands out of the crowd. This is most likely due to there are some issues related to covid-19 vaccines such as the post-vaccination deaths, efficiency rate of the vaccines, post-vaccination symptoms, etc.

COVID-19 related phrases such as “covid-19 cases” and “covid-19 infection” are also quite common. This is due to COVID-19 is still a thing until the date of writing this and news often reported about daily number of new COVID-19 cases and the infection areas.

Phrases such as “covid-19 immunisation” and “covid-19 vaccination” are also relatively common. This might be due to the facts that the Malaysia government often stress about the importance for the citizens to get themselves fully vaccinated under the COVID-19 Immunisation Programme so the infection rate can be decreased.

## B. *Concordances*

Since the domain of this project is related to Malaysia’s COVID-19 news, we will be using some keywords such as “vaccine”, “virus”.

### 1) *Find Semantic Or Context*

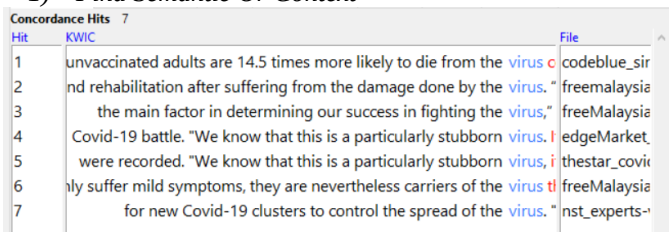


Fig. 2 Concordances of “virus”.

Based on the result of concordances of “virus”, with a good amount of window size we will be able to find out that “Covid-19” is one of the frequent words appear with “virus”. As such, we can understand that “Covid-19” is related to “virus”. This proves that using concordances, we can find semantic or context of keywords.

### 2) *Find Similarity Or Related Keywords*

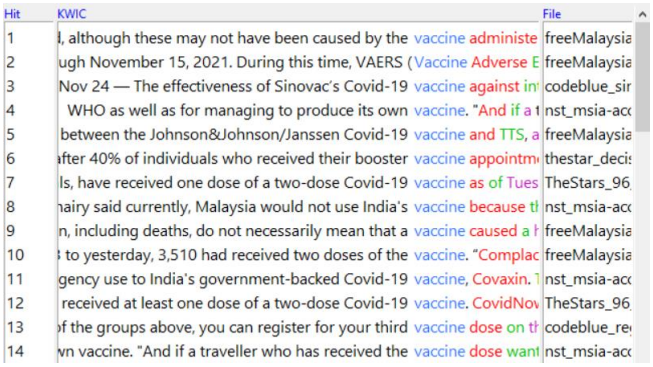


Fig. 3 Concordances of “vaccine”.

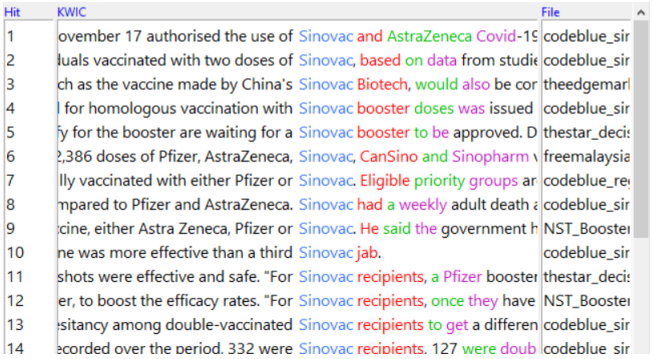


Fig. 4 Concordances of “Sinovac”.

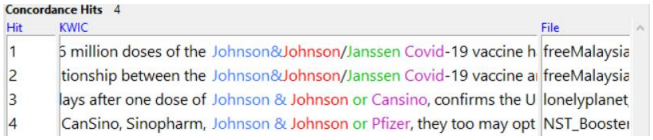


Fig. 5 Concordances of “Johnson&Johnson”.

Besides than finding the context of a keyword, from the Fig 3, we are able to find numerous mentions of “COVID-19" followed by “Sinovac”, “Johnson&Johnson”. Thus, “Sinovac” and “Johnson&Johnson” most likely related to each other. From Fig 4 and 5, we found out that both keywords have several frequent words again such as “booster”, “vaccine”, “effect”. Thus, we can conclude that “Sinovac” and “Johnson&Johnson” are correlated and probably serves the same purpose but still different from each other. Hence, concordances is useful in finding things or keywords that have similarity or related but probably have different semantic. For example, “vaccine” and “virus” can have same frequent words such as “Sinovac”, “Covid-19”, “infection” but “vaccine” and “virus” are differently things.

### 3) *Understanding Acronyms*

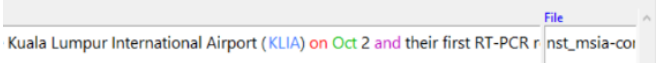


Fig. 6 Concordances of “KLIA”

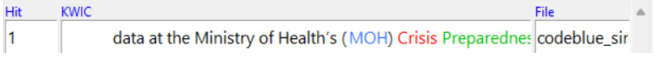


Fig. 7 Concordances of “MOH”

Using concordances on acronyms, we are also able to find the full form of acronyms. In Fig 6, we can find the full form of “KLIA” which is “Kuala Lumpur International Airport”. Apart from that, we can also find the full form of “MOH” which is “Ministry of Health”. In

modern days, we tend to use abbreviation and most of the time the full form of words is written before the acronym or mentioned at least once before the usage of acronym in the article.

### C. Comparison

First of all, n-gram method requires data preprocessing to be able to find a more meaningful of frequently occurred table. Based on Table 1, we learnt that stop words such as "of the", "in the" need to be removed from the dataset then "kuala lumpur" will be the most frequent occurred bigram. If that's the case, one of the findings from the Malaysia's COVID-19 article news would be the article news are most likely written from Kuala Lumpur or Covid-19 cases most likely reported from Kuala Lumpur. On the other hand, concordances method does not require any data preprocessing.

Secondly, probability play an important role in n-gram method while concordances method does not. N-gram method will make use of the probability to predict or rank the next keyword after the first n-gram. Meanwhile, concordances method does not make use of probability, it will only show the next keyword in the order of alphabetical and there are no calculation involves at all.

Additionally, n-gram does not require different genre of dataset to work around while concordances method do need multiple genres of dataset. Dataset of n-gram is not heavily dependent on genre but the quantity of occurrence n-gram. Even with less amount of dataset, n-gram still be able to predict the next keyword. On the other hand, concordances method requires multiple genres of dataset to really provide a complete context or semantic of keyword. For example, with a good amount of dataset in Psychology and Artificial Intelligence genre, concordances method will be able to understand "NLP" as "Neruo-linguistic Programming" and "Natural Language Processing". Without dataset in Psychology genre, "NLP" will always be understood as "Natural Language Processing" even if the context is regarding Psychology.

Last but not least, concept and application of n-gram are varying from concordances method. N-gram heavily involves with probability to predict the next keyword while concordances are used to understand the context or semantic of the keyword. N-gram is applied in querying or search engine and most of the time it is used with Least Recently Used concept as well. However, concordances method is more frequently applied in grammar check.

## V. CONCLUSION

We had attempted to perform some text exploration techniques such as n-gram and concordance on our dataset and successfully obtain some insights from the analyzed result. Both n-gram and concordance are able to help us to gain different insights on the data.

N-gram are useful in the way to help us extract the key phrases of the documents. For instance, bigram had been applied to our dataset to identify some important phrases related to the Malaysia COVID-19 news such as "covid vaccine" and "fully vaccinated". Apart from that, n-gram also useful in giving suggestions for the next word when searching for information through a document. For example, if the user would like to search for information related to COVID-19 among the Malaysia COVID-19 news, our result would suggest them to complete their search terms with word such as "vaccine", "cases", "infection", "immunisation", or "vaccination" after entering "covid-19".

On the other hand, concordances method is useful in finding semantic or context of a keyword in a specific genre. When both keywords have similar quantity of frequent words, concordances are able to conclude that the keywords are correlated and very similar but both keywords are still individually different item at the same time. This is very helpful especially when we are searching the first keyword in a search engine and another keyword can be used as a suggestion for the result. Besides that, concordance can be used to find acronyms. With sufficient amount of corpus with different genres, concordance will be useful in computing a list of acronyms and their full form respectively. This application is very useful in our modern days due to frequent usage of acronyms.

Conclusively, n-gram and concordances method are different, and they have their own uniqueness and specialty in terms of concept, application and methodology. Both are useful in text exploration and should be implemented in situation where they can perform their best benefits.

## REFERENCES


[1] N. Nasser, L. Karim, A. El Ouadrhiri, A. Ali, and N. Khan, "N-gram based language processing using Twitter dataset to identify COVID-19 patients," Sustainable Cities and Society, vol. 72, May 2021, p. 103048, doi: 10.1016/j.scs.2021.103048

[2] P. O'Kane, S. Sezer and K. McLaughlin, "N-gram density based malware detection," 2014 World Symposium on Computer Applications & Research (WSCAR), 2014, pp. 1-6, doi: 10.1109/WSCAR.2014.6916806.

[3] A. C. Sparavigna and R. Marazzato, "Using Google Ngram Viewer for Scientific Referencing and History of Science," ArXiv.org e-print archive, Dec-2015. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1512/1512.01364.pdf. [Accessed: 24-Nov-2021].

[4] I. Dunder and M. Pavlovski, "Computational concordance analysis of fictional literary work," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0644-0648, doi: 10.23919/MIPRO.2018.8400121.

[5] A. F. Khafaga and I. El-Nabawi, "Using concordance to decode the ideological weight of lexis in learning narrative literature: A computational approach," International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, Jan. 2020, pp. 1–7, doi: 10.14569/IJACSA.2020.0110433.

[6] N. W. Rees and J. D. Riding, "Automatic Concordance Creation for Texts in Any Language," *Proceedings of Translating and the Computer*, vol. 31, Nov. 2009, pp. 1–11.