2022 IEEE 7th International conference for Convergence in Technology (I2CT)
Pune, India. Apr 07-09, 2022

Using Word2Vec-LDA-Word Mover Distance for Comparing the Patterns of Information Seeking and Sharing during the COVID-19 Pandemic

Wei Wei Chan
Department of Computing and Information Systems
Sunway University
Selangor, Malaysia
16052748@imail.sunway.edu.my

Hui Na Chua
Department of Computing and Information Systems
Sunway University
Selangor, Malaysia
huinac@sunway.edu.my

***Abstract*—During pandemics such as COVID-19, government announcements were sources to convey accurate and relevant information to the public in times of outbreak. Prior studies attempted to explore the public awareness and behavioral changes from various research disciplines in response to the COVID-19 pandemic. Literature has pointed out that the appropriate use of information sources significantly relates to public attitudes in battling the pandemic. Social media has been the widely used medium to express public interests in current events. Literature shows that social media use during a crisis effectively coordinates relevant information from different sources and promotes situational awareness. Therefore, it is crucial to investigate scalable approaches to promptly gather insights into the public's interests and how governments responded to the interests relevant to the COVID-19 pandemic. However, there is little empirical research found that tackles these needs. Therefore, we aim to close the research gap by examining the feasible approaches for (1) identifying if public information-seeking has similar patterns as information-sharing on social media during the COVID-19 pandemic, and (2) comparing the patterns with the government announcements to confirm if the announcements show aligned response to the public information-seeking and sharing during the COVID-19 pandemic. We applied text processing, LDA topic modeling, and Word Mover Distance techniques to realize our aim through a Malaysian case study. Our research work contributes to the application of the LDA-Word2Vec-Word Mover Distance architecture and algorithms that can be used for future investigation and comparison of information seeking and sharing patterns in different research subjects.**

***Keywords*—Latent Dirichlet Allocation (LDA), Word2Vec, Word Mover Distance (WMD), COVID-19 pandemic, data mining, text mining, Natural Language Processing (NLP).**

I. INTRODUCTION

Social media plays a crucial role before and after the outbreak of COVID-19. It has become a medium for the public to gain helpful information, share their thoughts, socialize and exchange information [1]. Hence, the public discusses and shares the issues with restrictions implementation and other matters related to the pandemic using social media platforms such as Twitter and Facebook. In addition, governments worldwide also disseminate their messages of policies, controls, and updates relevant to the pandemic to the public via social media platforms [2].

Literature has shown that content sharing on social media platforms can increase public awareness [3]. As a result, it has garnered immense interest over time due to its potential to spread awareness and, subsequently, better precautious and preventive practices among the public.

Numerous research works have deployed machine learning and Natural Language Processing (NLP) approaches for gaining insights into the publics' subjects of interest from the aspects of understanding awareness through information-seeking patterns via the internet search data [3] and topics discussed in the social media [2] [4]. More extensive studies have also been conducted to identify the associations between news disseminated and public awareness [3] and agenda-setting patterns of the news and social media to influence the public's opinions [5] and analyse news dissemination patterns in support of government-citizen engagement [2]. However, there is scarce research that examines the approaches to understanding the associations between public information seeking and sharing and how the government's announcements align with the public's interests via their information seeking and sharing patterns. Therefore, this study intends to answer the following research questions through a Malaysian case study:

1) Does public information-seeking has similar patterns as information-sharing on social media during the COVID-19 pandemic?

2) Do the government announcements show aligned response to the public's interests through information-seeking and sharing during the COVID-19 pandemic?

To realize our research objective in answering the questions, we employed text mining approaches, including crawling data from Google Trends, Twitter, and Facebook. For analyzing and modelling the data, we integrated the techniques of Latent Dirichlet (LDA) topic modelling and Word Mover Distance (WMD) with Word2Vec to derive insights into the patterns and comparisons of the public's information seeking, sharing, and government's announcements.

II. LITERATURE REVIEW

A. Text Mining Approaches

The process of text mining involves using Natural Language Processing (NLP) to extract valuable insights from unstructured text. Text mining involves transforming data into information that machines can understand. Combined with machine learning, data models can be built to automate the text classification tasks such as topic analysis for identifying the similarity between documents [6].

- ***Word Embedding with Word2Vec***. Word Embedding is one of the NLP techniques used for mapping words to vectors of real numbers [7]. The process of converting words into numbers is called vectorization. Word2Vec

plots the words in a multi-dimensional vector space, where similar words tend to be close to each other. Hence, word embedding via Word2Vec can make natural language computer-readable, then further implementation of mathematical operations on words can be used to detect their similarities. Essentially, the Word2Vec model is an unsupervised model which can take in massive textual corpora, create a vocabulary of possible words and generate dense word embeddings for each word in the vector space representing the vocabulary. As a result, the vector representation can extract semantic relationships and syntactic similarity based on the co-occurrence of words in the dataset.

- ***LDA***. LDA for topic modelling is used for latent data discovery and finding relationships among data and text documents [8]. The topic modelling concept is widely applied in NLP to discover hidden themes in collections, summarize and compare documents by topic [2] [6]. LDA ignores the occurrence of words and syntactic information. Meaning that LDA treats documents as a collection of words. The core concept of LDA is that it assumes a probabilistic model for documents. More specifically, the LDA model assumes that each document consists of a mixture of topics while each consists of a mixture of words. The main goal of the LDA algorithm is to learn the topic mixture in each document and word mixture in each topic. LDA represents topics by word probabilities. The word with the highest probabilities in each topic usually gives a good idea of the topic.
- ***WMD for identifying Text Similarity***. Finding the similarity between words is a primary stage for sentence, paragraph, and document similarities. Text similarity can be conducted in two ways which are lexical similarity and semantic similarity. The difference is that lexical similarity does not consider the actual contextual meaning behind the words or entire phrase in context, while semantic similarity does. Lexical similarity measures strings in terms of their common words and character sequence. For instance, the word pair of (book, cook) have high lexical similarity, but they are not semantically related. While the word pair of (car, wheel) does not have lexical similarity, they are semantically related as both are automotive-related terms [9]. WMD is a metric for the distance between text documents that leverages word vector relationships of word embeddings such as Word2Vec. WMD considers distances between embedded word vectors are to some degree, semantically meaningful. The text documents are represented as a weighted point cloud of embedded words. The WMD is the minimum total travel distance required to transport all word vectors from one document to another at a high level. The minimized travel distance is used to measure dissimilarity between the two documents [10]. WMD can target both semantic and syntactic approaches to get similarity between text documents has overcome the issues of identifying synonyms in other text similarity techniques [11].

B. Related Work on Text Mining for Analysing COVID-19 Social Media Data

Mangono et al. [12] conducted a study, which analysed Google Trends data to provide insights and potential indicators of essential changes in information-seeking patterns during the COVID-19 pandemic across the United States. The statistical methods of pairwise correlations and principal component analysis were used to extract search patterns across states. However, this analysis is limited by specific assumptions of Google Trends, including pulling the data from only a sample and not the whole database of searches and providing it in relative search volume instead of absolute search volumes per geographic region. Moreover, Lim et al. [13] conducted a study on Google Trends to understand whether the public's inquisitiveness towards COVID-19 and its recommended precautionary measures had increased during the initial duration of the pandemic in Malaysia. Spearman's rank correlation coefficient was used to gauge the correlation between search trends with the number of cases and deaths in Malaysia. Finally, Bento et al. [14] examined the public information-seeking behaviours in responding to the state government announcement of the first COVID-19 case. The data was collected using Google Health Trends API and implemented poison models and regression analysis in an event study framework. The result highlights search patterns that occur in the days leading up to and following the first case announcement in a state.

Besides, Boon-Itt et al. [15] studied the public perception towards the COVID-19 pandemic through sentiment analysis and topic modelling to identify the explore discussion topics over time through Twitter. The LDA algorithm identified the most shared tweets, categorized clusters, and identified themes based on keyword analysis. Furthermore, Lyu et al. [16] utilized twitter data to conduct LDA topic modelling to understand the public discussion about the Centers for Disease Control and Prevention. Furthermore, Amara et al. [17] exploited the Facebook data to perform LDA-based topic modelling methods and explored the evolution over time of the user's interest across different periods of the pandemic outbreak. Finally, Kok and Chua [2] deployed the Word2Vec-LDA-Cosine Similarity technique to discover how news coverage pattern aligns with government-disseminated information through a case study of Covid-19 in Malaysia during the pandemic.

There are various studies conducted surrounding the investigation of the public knowledge, attitudes, behaviors towards the COVID-19 pandemic [13] [14] [15] [16] [17] [18]. However, most of the studies were conducted using statistical approaches where limitations could exist due to the dissemination of the questionnaire. Also, we observed a lack of social media-based research studying the public's interests and the government's responses to the interests on COVID-19 pandemic-related matters.

In a nutshell, prior studies evidenced the ability of text mining to derive insights from a massive amount of data for performing surveillance analysis. Works of literature above also revealed that social media text mining is an effective way to track disease and assess public perception and give a global sight of the pandemic on different aspects. Besides, Google Trends can be a potential tool for risk communication as it effectively examines web-based information-seeking behaviour based on search queries. On the other hand, social media data enable researchers to obtain a large sample of user-generated content and represents an important source that gives real-time monitoring of public perception to inform early response strategies. The literature review has also proven that LDA can be a suitable technique

to discover hidden structures related to user behaviour on social media.

III. Research Methodology

A. Data Collection

The data were collected from Google Trends, Twitter, and Facebook. All the data ranging from 1 January 2020 to 30 September 2021 focusing on Malaysian data were collected from the three data sources. It is observed that different sources have different APIs that have different parameter settings when performing data crawling. For example, Twitter allows unlimited keywords and hashtags, while Google Trends limits only five keywords. Facebook only allows crawling on specified public pages instead of using keywords. Hence, the keyword selection relevant to the COVID-19 context based on domain knowledge was applied to each data source. The terms used are English and Malay to capture the COVID-19 related terms in the Malaysian context.

Google Trends allows searching for a particular topic on Google or a specific set of search terms. The Google Trends data was collected through the *PyTrends* open-source API, using keywords of {*kes COVID-19, mco, sop, bantuan, vaksin*}. Top searches are the most frequently searched terms with the term users entered in the same search session, within the chosen category and country. Tweets were extracted using *the snscrape* library. The specified COVID-19 related keywords used for scraping were {*coronavirus, COVID-19, COVID-1919, pkp, pkpp, pkpb, pkpd, mco, rmco, cmco, emco ,bantuan, PRN, vaccine, vaksin , #COVID-1919,#KerajaanGagal, #KitaJagaKita, #StayAtHome , #dudukrumah*}.

In Addition, Facebook posts were crawled from three COVID-19 related public pages through third-party software, Facepager [19]. The selected Facebook posts were mainly official public pages from Malaysia government namely Majlis Keselamatan Negara (MKN), Jawatankuasa Khas Jaminan Akses Bekalan Vaksin COVID-19 (JKJAV) and Noor Hisham Abdullah. These Facebook pages were selected as they are the official sources responsible for communicating information about COVID-19 in the country through Facebook posts. Furthermore, these Facebook pages are chosen to investigate whether the information posted as announcements by government authorities aligned with public interests via their information seeking and sharing during the COVID-19 pandemic.

B. Data Preparation

Since the context of the study is Malaysia, nearly 95% of the data collected were in the Malay Language. In this scenario, translation to the English Language might result in loss of information and context because tweets primarily consist of informal Malay language. Hence, data pre-processing were carried out using the Malaya Python library [20] to retain the originality of the contents. The data cleaning steps involved the conversion of texts to lower cases, removing stop-words and non-words characters such as punctuation, emoji, and numbers. Moreover, stemming and lemmatization were conducted to reduce the words into their respective root forms. Furthermore, the words were concatenated into n-grams and tokenized. This process has been applied to all three data sources.

C. Data Modelling

Two LDA models were built to extract topics from tweets and government posts. Both models produced a k-number of topics whereby each topic comprises top-15 descriptive words. Several model parameters and k-number of topics need to be adjusted to generate a more interpretable topic. During this process, an iteration of the k-value that ranges from 2 to 5 topics was tested. The rationale for choosing five topics as the maximum topics is because tweets usually are short texts; hence specifying too many topics to be extracted will increase the probability of topic overlapping. The best model was selected through the hyper-parameter tuning process using topic coherence score as an evaluation matrix. A higher topic coherence score indicates better human interpretability of the topic. This analysis was supplemented with the manual evaluation of topics' interpretability. We performed a quarterly topic modelling analysis from the year 2020 quarters Q1 to 2021 Q3 to gain more specific insights into changes in the interest of topic discussion.

Given the k number of topics from quarterly tweets and government posts, the topic similarity is measured through the WMD algorithm. Each topic will consist of its top 15 topic words extracted from LDA. Before fitting the model, each topic word needs to be vectorized using the Word2Vec model. The Word2Vec was trained using COVID-19 domain-specific corpus. Each topic word will be converted into its representative word embedding during this process. First, hyper-parameter tuning on window size and model architecture was conducted to identify the best trained Word2Vec model representing the COVID-19 context. Then, a pairwise WMD calculation was adopted to compute the similarities between Google Trends-tweets topic and Google Trends-government posts for each quarter from 2020 Q1 – 2021 Q3. After that, the similarity between two topics, T1 and T2, was calculated using the WMD algorithm.

D. Model Evaluation and Selection

Model selection on Word2Vec was carried out to select a more precise vector model in capturing topic similarities between two contents. Model selection on LDA is through the hyper-parameter tuning process to identify the optimum numbers of topics, denoted as the k value that produces the best results in its topic coherence score.

IV. Results

A. The Public's Information Seeking Patterns

Figure 1 describes the Google search trends for the specified five keywords over the study period. During the year 2020 Q1 (1 January 2020 – 31 Mac 2020), when the pandemic just began, all the search terms have reached a spike in search trends as people were curious about the unprecedented pandemic. Hence, it can be observed that the search queries were mainly about searching the general information about COVID-19 related terms such as the definition of Standard Operation Procedure (SOP), Movement Control Order (MCO), vaccination, government grant, and reported numbers of COVID-19 cases. Overall, the trends showed relatively higher search volume during the first few days of the event and a gradual decrease in search interest afterward.

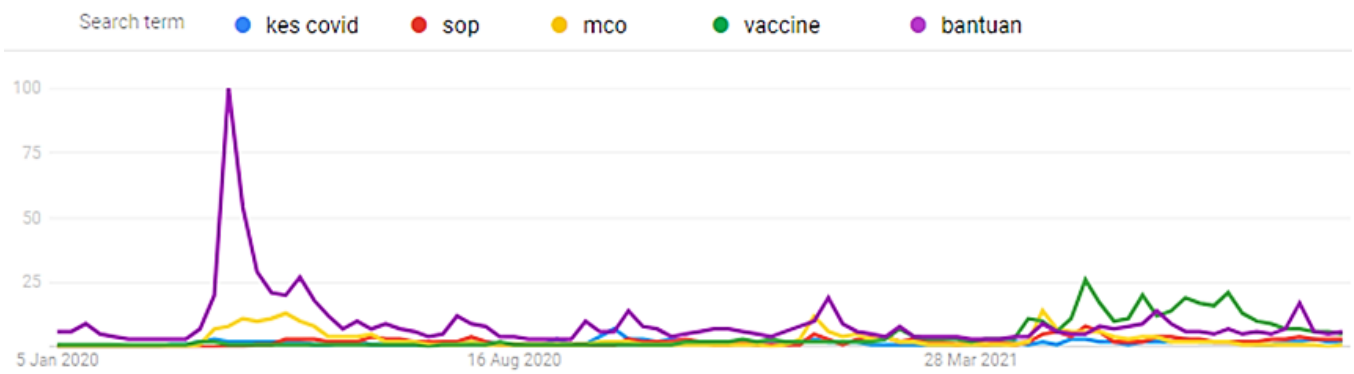

Fig. 1. The Public's Information Seeking Patterns Over Time based on Top 5 Search Terms from Google Trends

The public interest in the pandemic status of Malaysia was reflected through search queries such as *kes COVID-1919 terkini, pecahan kes COVID-19, jumlah kes COVID-19*. Moreover, search queries on cases specific to states in Malaysia can be observed throughout the study period.

The Malaysian government had imposed the MCO for the common good starting from 18 March 2020. It was replaced by the Conditional MCO (CMCO) on 4 May 2020 and subsequently replaced by the Recovery MCO (RMCO) on 10 June 2020. The MCO measure encompassed restrictions on movement, assembly, international travel and mandated closure of the business, industry, government, and educational institutions to curb the spread of COVID-19 infections in Malaysia. From time to time, the Malaysian government reviewed the implemented measure based on the current country's pandemic situation and made news announcements regarding the latest SOP of different sectors accordingly. As a result, the sought topics related to MCO and SOP were somehow similar, where people were seeking information about rules and regulations enforced during each phase of MCO. For instance, terms like *SOP terkini, SOP PKP, SOP PKPB, SOP PKPD, SOP MKN, MCO 1.0, MCO 2.0, MCO 3.0, MCO 4.0, CMCO, RMCO* appeared to be frequent search queries for each quarter throughout the study period.

In addition, the enforcement of the MCO by the government has put various sectors of the economy in jeopardy. The result of the prolonged lockdown imposed had a significant impact on the country's economy. Hence, It is noticeable that the search term *bantuan*, which means financial aids in English, has caught people's attention starting from 2020 Q2 (1 April 2020 -30 April 2021). This finding inferred that people were actively looking for information on financial aids through search queries. The top sought topics in Malay occurred were *bantuan COVID-1919 sabah, bantuan sara hidup, bantuan ihsan johor, bantuan mysalam, bantuan khas COVID-1919, bantuan i-Citra*.

When the COVID-19 vaccine has not been developed, people started to search for general knowledge about vaccines. However, its queries are relatively lower compared to other search terms. However, the public gradually shifted their interest to the vaccination topic from 2021 Q2 onwards. During that period, the National COVID-19 Immunisation program was implemented by the Malaysian government to curb the spread of COVID-19 infections in the country. Search queries such as *cara daftar vaksin, mysejahtera, vaksin sinovac, vaksin pfizer, vaksin astrazeneca, vaksin comirnaty, jenis vaksin COVID-1919 , temujanji vaksin, semakan vaksin COVID-19, pusat vaksin* implies that people were searching for information about vaccination and its registration procedure. The following 2021 Q3 was when people started to get vaccinated progressively. Public concern regarding the side effect of vaccination was revealed through search queries in Malay, such as *kesan vaksin and selepas vaksin*.

B. The Public's Information Sharing Patterns

The tweets topics comprised three topics as determined by the LDA model. Table I shows the top-30 terms for each topic. Topic ID 1 describes COVID-19 cases through descriptive terms includes *kes, negeri, positif, mati, turun*. Topic ID 2 revolved around topics of vaccination. This topic indicates that the people shared about their vaccination experience can be reflected through terms such as *ambil, terima, suntik, cucuk, daftar, uji, sakit, demam, phizer, sinovac*. Moreover, it can be observed that the most discussed topic (Topic ID 3) was regarding MCO. Descriptive terms such as *makan, kedai, kerja, rumah, beli, cuti* possibly indicate people's sharing about their conditions during the MCO period. The public also discussed Topic ID 4, the SOP to comply with during the enforcement of MCO, which is reflected through the descriptive terms *tatacara kendali piawai, mysejahtera*.

TABLE I. LDA TOPIC OF INFORMATION SHARING (VIA TWEETS) FROM 2020 Q1- 2021 Q3

Topic_ID	Topic Words
1	*kes, mco, moga, COVID-19, malaysia, selamat, selangor, vaksin, kuala_lumpur, doa, jun, gembira, sambung, hidup, mati, keluarga, negeri, mudah, sembuh, julai, positif, turun, urus, jaya, tarikh, pikir, pulih, sayang, baharu, johor*
2	*vaksin, COVID-19, dos, cucuk, ambil, fasa, indonesia, sakit, demam, positif, pfizer, vaccination, lindung, sinovac, doktor, vaccinate, lantik, anti, tarikh, hasil, swab, sebar, terima, vaksinasi, virus, takut, suntik, Singapore*
3	*kena, vaksin, kerja, makan, rumah, beli, COVID-19, raya, anak, esok, jumpa, buka, jalan, kedai, musim, duduk, masuk, pakai, sikit, ambil, penat, ramai, kerja, cuti, rindu, kawan*
4	*vaksin, mysejahtera, COVID-19, rakyat, raja, daftar, kena, kes, tatacara, kendali, piawai, ramai, menteri, ambil, malaysia, kerja, bodoh, bantu, buka, semak, negara, bohong, mati, kilang, salah, tutup, percaya, pilih, masuk, cakap, kedai, makan*

TABLE II. LDA TOPIC OF GOVENRMENT ANNOUNCEMENTS (VIA FACEBOOK) FROM 2020 Q1- 2021 Q3

Topic_ID	Topic Words
1	*berkuatkuasa, laksana, vaksin, vaccination, vaksinasi, daftar, terima, lindung, kampong, oktober, julai, walk,sunti, bahagian, do, mysejahtera, sabah, lantik, maklumat, janji_temu, COVID-19,mudah, selamat,vaccinate, nasional, jalan,kasih, ogos, lokaliti, awam*
2	*kes, kumulatif, ogos, september, lapor, julai, cluster, negeri, COVID-19, mati, warganegara, vaksinasi, do, baharu, pulih, takat, status, discaj, vaccination, Malaysia, nyata, tamat, peratus, positif, gerak, perintah, vaksin, aktif, kawal, sakit*
3	*menteri, malaysia, negara, jabat, lampir, kadar, kesihatan, tatacata, kendali, piawai, tahan, ismail_sabri, yakoob, perdana, negeri, pecah, siar_sidang, patuh, media, operasi, daftar, selamat, tugas, terang, komunikasi_multimedia, COVID-19, paksa, kawal, kena, hospital*

C. The Government Announcements Patterns

Throughout the study, the government posts topics comprised four topics determined by the LDA model. Table II shows the top-30 terms for each topic. Topic ID 1 described vaccination. This topic indicates that the government reported the daily figures of vaccination rate and encouraged the public to register for vaccination which can be examined through descriptive terms like *suntik, dos, vaksin, daftar, program_imunisasi*. Besides, Topic ID 2 is the second most posted topic, which revolved around the statistics of COVID-19 cases with terms such as *kes, positif, kumulatif, import, kluster, discaj, icu, sembuh*. Finally, topic ID 3 has a combination of MCO and SOP topics. One possible reason for this mixture of aspects is that the number of posts may contain more than one unique topic, leading the LDA model to identify a topic amalgamation of a few topics. However, this seems reasonable as these two topics are somehow associated with each other as SOP describes the rules and regulations of the MCO period.

D. Comparing Information Seeking and Sharing Patterns

The alignment of public information seeking and sharing patterns was investigated by examining its topic similarity across each quarter. A threshold to determine the topic similarity was established in this research. The minimum and maximum WMD score range fall between 0 to infinity value which is too broad and ambiguous. Score "0" represents complete similarity when documents contain the exact words. On the contrary, WMD generates infinity value (INF) when document pairs are very distant from each other in the semantic space of the COVID-19 pandemic context. In this study, the degree of similarity between Google search queries and social media posts was measured relatively. The topic pairs containing synonyms or similar meanings will have a smaller distance in semantic space computed by the WMD algorithm. In contrast, topic pairs consisting of closely unrelated words or different meanings will have a larger distance score than other topic pairs.

Table III shows the summarized topics of the public's information-seeking (via Google Trends) and sharing (via Tweets) for 2021 Q3. Finally, table IV shows the summarized WMD score results from 2021 Q3. As observed in Table IV, the topic pairs with a minimum WMD score indicate that the particular searched topic of information seeking is similar to one of the topics of information sharing during the quarter. Meanwhile, the topic pairs with a maximum WMD score indicate that the particular information seeking is not similar to information sharing during the quarter. In other words, it indicates that topic similarity exists between the following topic pairs between information seeking (GT) and sharing (TT), GT1 versus TT1; GT2 versus TT3; GT3 versus TT2; GT4 versus TT1; and GT5 versus TT2. The minimum and maximum distance scores generated from both topic pairs during 2021 Q3 are 2.404 and 4.307, respectively.

TABLE III. SUMMARISED TOPIC OF INFORMATION SEEKING AND SHARING FOR 2021 Q3

Information Seeking (Google Trend Search Queries)			**Information Sharing** (Tweets)		
ID	**Topic**	**Interpretation**	**ID**	**Topic**	**Interpretation**
GT1	*tatacara, kendali, piawai, mkn, fasa, pkp, pkpd, ppn, sabah, selangor, emco, miti*	Standard Operation Procedure (SOP)	TT1	*COVID-19, uji, kes, positif, malaysia, vaksin, kena, anak, tatacara, kendali, piawai, anti, habis, ambil, mati, sekolah*	Standard Operation Procedure (SOP) + Vaccination
GT2	*COVID-19, vaksin, semak, pfizer, sinovac, astrazeneca, daftar, status, pusat, jenis, temujanji, mysejahtera, sijil, tarikh, cucuk, kesan*	Vaccination	TT2	*COVID-19, kes, lindung, moga, indonesia, vaksin, kerja, hidup, kuala, lumpur, terima, keluarga, rakyat, bantu, vaksinasi, mati*	Financial Aids + COVID-19 cases
GT3	*bantu, khas, COVID-19, semak, sabah, nadma, prihatin, rakyat*	Financial Aids	TT3	*vaksin, dos, mysejahtera, cucuk, kena, ambil, tarikh, do, semak, sakit, kemas, fasa, lantik, masuk, semak*	Vaccination
GT4	*mco, malaysia, selangor, tatacara, kendali, piawai, fasa, tarikh*	Movement Control Order (MCO)	-	-	-
GT5	*kes, COVID-19, malaysia, kedah, dunia, indonesia, pahang, kelantan, terengganu, selangor, sarawak, perak, jumlah*	COVID-19 Cases	-	-	-

TABLE IV. SUMMARISED TABLE OF WMD SCORE BETWEEN INFORMATION SEEKING AND SHARING FOR 2021 Q3

Information Seeking Topic ID	Min_{wmd}	Max_{wmd}	Min<WMD<Max
GT1	TT1, 3.346	TT3, 4.307	TT2, 4.174
GT2	TT3, 2.404	TT2, 3.526	TT1, 3.487
GT3	TT2, 3.045	TT1, 3.572	TT3, 3.524
GT4	TT1, 3.040	TT2, 4.072	TT3, 3.563
GT5	TT2, 3.121	TT3, 3.999	TT1, 3.296
Average	2.991	3.895	TT3.609

TABLE V. SUPPLEMENTARY TABLE OUTLINES TOPIC WORDS FROM INFORMATION SEEKING AND SHARING DURING 2021 Q3

ID	Information Seeking (Goggle Trends Topic of 2021 Q3)	ID	Information Sharing (Tweet Topic of 2021 Q3)	WMD distance scores
GT1	*tatacara, kendali, piawai, mkn, fasa, pkp, pkpd, ppn, sabah, selangor, emco, miti*	TT3	*vaksin, dos, mysejahtera, cucuk, kena, ambil, tarikh, do, semak, sakit, kemas, fasa, lantik, masuk, sijil*	4.307 (max)
GT2	*COVID-19,* ***vaksin, semak, pfizer, sinovac, astrazeneca,*** *daftar, status, pusat, jenis, temujanji,* ***mysejahtera, sijil, tarikh, cucuk,*** *kesan*	TT3	***vaksin, dos, mysejahtera, cucuk,*** *kena, ambil,* ***tarikh,*** *do, semak, sakit, kemas, fasa, lantik, masuk,* ***sijil***	2.404 (min)

*Note: Bold words in a topic indicate semantic related terms

Table V shows the topic pairs (GT2 versus TT3) with minimum distance scores consisting of several overlapping words such as *cucuk, vaksin, mysejahtera, tarikh*. The rest of the topic words have semantic meaning and are associated with the vaccination based on domain knowledge on COVID-19. On the other hand, the topic pairs (GT1 versus TT3) with maximum distance scores are closely unrelated in which sought topic is about SOP and MCO while the shared topic is about vaccination. This implies that the constructed WMD model can capture the text similarities between topic pairs by assigning lower distance scores to topic pairs that are similar among others and vice versa.

The word frequency plot for each document has been plotted to visualize better the similarity and differences in terms of topic words across the topic pairs between information seeking and sharing, based Google Trends and Twitter data, as shown in Fig. 2. Through qualitative assessment conducted on Fig. 2, the topic words with lexical meaning appeared mutually in corpora such as *mco, vaksin, tatacara kendali piawai, kes, daftar, mysejahtera, bantu*, which are words associated with aspects of COVID-19.

E. Comparing Information Seeking and Government Announcements Patterns

Table VI shows the summarized information-seeking topics and government posts for 2021 Q3. Finally, Table VII shows the summarized WMD score results from 2021 Q3.

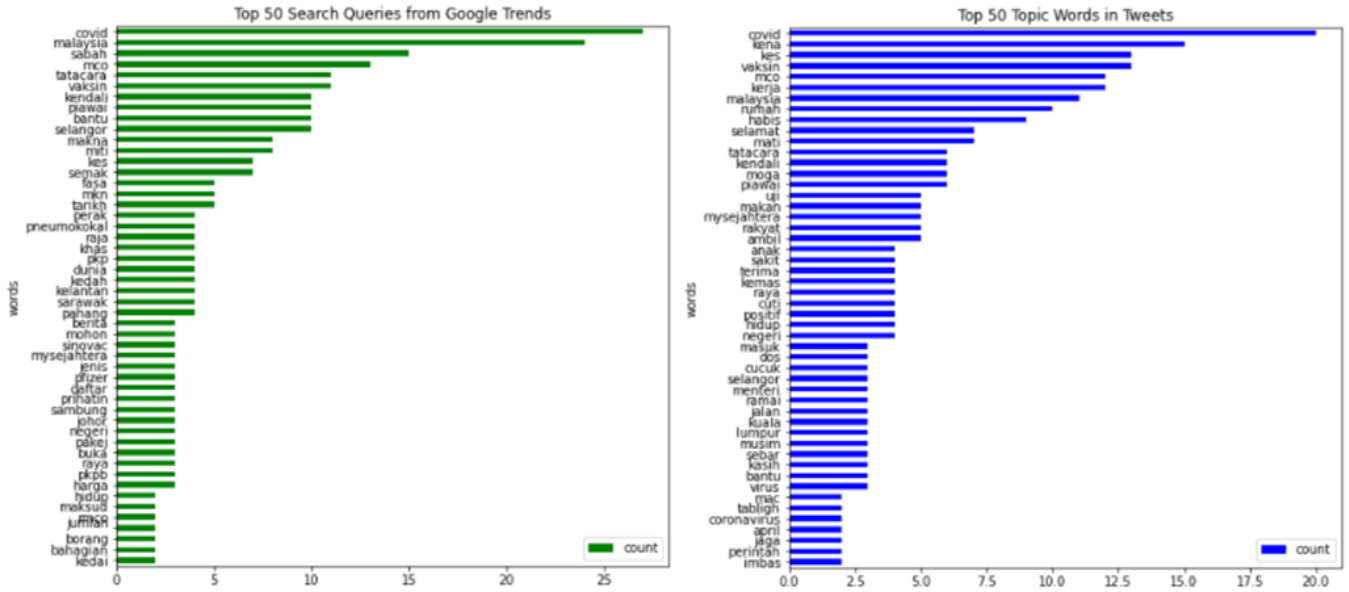

Fig. 2. Top 50 topic words from Google Trend search queries and Tweets

TABLE VI. THE SUMMARISED TOPIC OF INFORMATION SEEKING AND GOVERNMENT ANNOUNCEMENTS FOR 2021 Q3

Information Seeking (Google Trend Search Queries)			**Government Announcements** (Facebook posts)		
ID	**Topic**	**Interpretation**	**ID**	**Topic**	**Interpretation**
GT1	*tatacara, kendali, piawai, mkn, fasa, pkp, pkpd, ppn, sabah, selangor, emco, miti*	Standard Operation Procedure (SOP)	FT1	*vaksin, do, julai, tamat, dos, kategori, tadbir, vaksinasi, daftar, vaccination, lapor, negeri, ogos, COVID-19, terima*	Vaccination
GT2	*COVID-19, vaksin, semak, pfizer, sinovac, astrazeneca, daftar, status, pusat, jenis, temujanji, mysejahtera, sijil, tarikh, cucuk, kesan*	Vaccination	FT2	*negara, menteri, lampir, kadar, media, malaysia, pecah, kawal, gerak, perintah, selamat, negeri, tahan, kena, ketat*	Movement Control Order
GT3	*bantu, khas, COVID-19, semak, sabah, nadma, prihatin, rakyat*	Financial Aids	FT3	*pahang, kelantan, labuan_putrajaya, perlis, kes, johor, melaka, sabah, perak, negeri, pulau_pinang, sarawak, selangor, kedah, pecah*	COVID-19 cases
GT4	*mco, malaysia, selangor, tatacara, kendali, piawai, fasa, tarikh*	Movement Control Order (MCO)	FT4	*kes, september, kumulatif, lapor, kluster, ogos, vaksinasi, warganegara, status, mati, negeri, COVID-19, takat, discaj, malaysia*	COVID-19 cases
GT5	*kes, COVID-19, malaysia, kedah, dunia, indonesia, pahang, kelantan, terengganu, selangor, sarawak, perak, jumlah*	COVID-19 Cases	-	-	-

TABLE VII. SUMMARISED WMD SCORE BETWEEN INFORMATION SEEKING AND GOVERNMENT ANNOUNCEMENTS FOR 2021 Q3

Information Seeking Topic ID	Min_{wmd}	Max_{wmd}	Min<WMD<Max
GT1	FT3, 3.490	FT4, 4.594	FT1, 4.410; FT2, 4.108
GT2	FT1, 3.162	FT3, 4.203	FT2, 4.131; FT4, 3.987
GT3	FT3, 3.570	FT4, 4.122	FT1, 3.762; FT2, 3.733
GT4	FT2, 3.821	FT4, 4.249	FT1, 4.226; FT3, 3.864
GT5	FT3, 1.968	FT1, 3.732	FT2, 3.634; FT4, 3.459
Average	2.488	4.180	3.931

TABLE VIII. SUPPLEMENTARY TABLE OUTLINES TOPIC WORDS FROM INFORMATION SEEKING AND GOVERNMENT ANNOUNCEMENTS DURING 2021 Q3

ID	Information Seeking 2021 Q3	ID	Government Announcements 2021 Q3	WMD Scores
GT1	*tatacara, kendali, piawai, mkn, fasa, pkp, pkpd, ppn, sabah, selangor, emco, miti*	FT4	*kes, september, kumulatif, lapor, kluster, ogos, vaksinasi, warganegara, status, mati, negeri, COVID-19, takat, discaj, malaysia*	4.594 (max)
GT2	***COVID-19, vaksin, semak***, ***pfizer, sinovac, astrazeneca***, ***daftar***, *status, pusat, jenis, temujanji,* ***mysejahtera, sijil***, ***tarikh***, ***cucuk***, *kesan*	FT1	***vaksin***, *do, julai, tamat,* ***dos***, *kategori, tadbir,* ***vaksinasi***, ***daftar***, ***vaccination***, *lapor, negeri, ogos,* ***COVID-19***, ***terima***	3.162 (min)

*Note: Bold words in a topic indicate semantic related terms

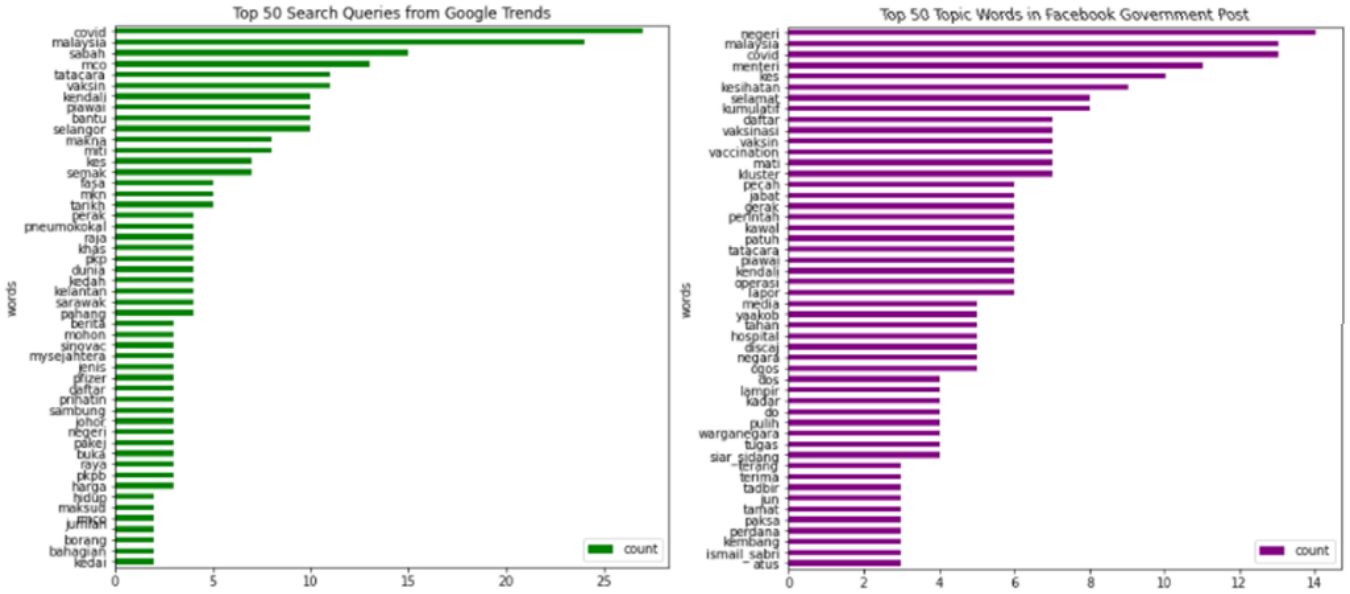

Fig. 3. Top 50 topic words from Google Trend search queries and Government Posts

Table VII shows the summarized WMD score results from 2021 Q3. As observed in Table VII, the topic pairs with a minimum WMD score indicate that the particular sought topic is similar to one of the topics announced on government posts during the quarter and vice versa. Table VII presents the topic similarity between these topic pairs, GT1 versus FT3; GT2 versus FT1; GT3 versus FT3; GT4 versus FT2; and GT5 versus FT3. The minimum and maximum distance scores generated from the Google Trends and government posts topic pairs during 2021 Q3 are 3.162 and 4.594, respectively.

Table VIII shows the topic pairs (GT2 versus FT1) with a minimum distance score consisting of several overlapping words such as *vaksin and daftar*. The rest of the descriptive terms have semantic meaning and are associated with the vaccination based on domain knowledge on COVID-19. On the other hand, the topic pairs with maximum distance scores (GT1 versus FT4) are closely unrelated in which sought topic about SOP and MCO. At the same time, government announcements are related to the statistics of the COVID-19 cases.

The threshold for differentiating the meaning of topic pairs will result in INF (infinity) value. However, based on the quarterly analysis, no topic pairs generated an infinity distance score. Hence, it can be concluded that similarities still exist between topic pairs from information seeking and sharing, information seeking, and government announcements, just that matters of similarity strength. In conclusion, the public information seeking on a specific topic through Google search engines shows similar patterns with the information sharing through social media through Twitter and government Facebook posts.

The word frequency plotted to visualize better the similarity and differences in terms of topic words across the topic pairs between information seeking and government announcements, based Google Trends and Facebook posts, as shown in Fig. 3. Through qualitative assessment conducted on Fig. 2, the topic words with lexical meaning appeared mutually in corpora such as *mco, vaksin, tatacara kendali piawai, kes, daftar, mysejahtera, bantu*, which are words associated with aspects of COVID-19.

V. DISCUSSION

Literature shows a lack of approaches that can promptly gather insights into the public's interests and how governments responded to the interests relevant to the COVID-19 pandemic. The text mining approaches presented in this paper prove the feasibility of tackling the research gap. Our proposed approaches successfully yield the results that display both sought topics, and the shared topic was related to aspects of COVID-19 such as SOP, MCO, vaccination, and daily statistics of death toll and cases. Our work presented in this paper contributes to the application of LDA-Word2Vec-Word Mover Distance architecture that can be used for future investigation and comparison of information seeking and sharing patterns in different research subjects.

Furthermore, the results showed distinguishing characteristics between information seeking and sharing. For instance, it is found that the topic of financial aids has a high search interest in information seeking but does not have a sign of high similarity to information sharing. This observation could be explained as financial matters are considered a more sensitive topic widely shared on social media. Our findings also imply that government announcements responded to public interests (via their information seeking and sharing) throughout the pandemic, indicating that the Malaysian government responded to the public interests timely.

VI. CONCLUSION

To the best of our knowledge, this is the first study that presented the feasible methods through text mining approaches for comparing information seeking and sharing related to the COVID-19 pandemic and how the government responded to the public interest via the information seeking-sharing patterns. The application of the LDA-Word2Vec-Word Mover Distance architecture presented in this paper can be used for future investigation and comparison of information seeking and sharing patterns in different research subjects. The experimented architecture can capture the topic similarity between both information seeking and sharing but raised the question of appropriate threshold value to be used in general topic comparison tasks. Hence, future research is required to determine the threshold to produce better empirical topic similarity comparison results.

This study poses several limitations. Firstly, the conversion of English to Malay during the data cleaning process may result in minimal information loss. Besides, the study adopted a self-trained Word2Vec model instead of using the pre-trained model to meet the context of the study. Hence, future research can further improve and build a more comprehensive Word2Vec on Malay corpus. Secondly, using Google Trends output to generalize the public's information-seeking pattern may not reflect a perfect search interest because only five keywords can be specified.

ACKNOWLEDGMENT

Funding: This research was supported by the Malaysian government FRGS grant [FRGS/1/2019/ICT04/SYUC/02/2].

REFERENCES

[1] S. Kemp, "Digital 2021: Malaysia," Datareportal, [Online]. Available: https://datareportal.com/digital-in-malaysia. [Accessed July 2021].

[2] Kok, K. S., & Chua, H. N. (2022). Using Word2Vec-LDA-Cosine Similarity for Discovering News Dissemination Pattern to Support Government–Citizen Engagement. In Proceedings of International Conference on Data Science and Applications (pp. 703-716). Springer, Singapore.

[3] N. W. J. Yan and H. N. Chua, "A Path Analysis Model to Identify the Effects of Social Media, News Media and Data Breach on Data Protection Regulation Awareness," 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 2020, pp. 1-6, doi: 10.1109/IICAIET49801.2020.9257846.

[4] Q. Khan and H. N. Chua, "Comparing Topic Modeling Techniques for Identifying Informative and Uninformative Content: A Case Study on COVID-19 Tweets," 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 2021, pp. 1-6, fdoi: 10.1109/IICAIET51634.2021.9573878.

[5] Wong, N. J. Y.; and Chua, H. N. (in press). A Framework Integrated with Time Series and Text Mining Models to Compare Agenda-Setting Patterns of News and Social Media. Journal of Engineering Science & Technology (JESTEC).

[6] Chaw, C. Y., & Chua, H. N. (2021). A Framework System Using Word Mover's Distance Text Similarity Algorithm for Assessing Privacy Policy Compliance. In IT Convergence and Security (pp. 79-89). Springer, Singapore.

[7] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," ICLR, 2013.

[8] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211.

[9] D. D. Prasetya, A. P. Wibawa and T. Hirashima, "The performance of text similarity algorithms," International Journal of Advances in Intelligent Informatics, vol. 4, no. 1, pp. 63-69, 2018.

[10] C. G. Gao Huang, M. J. Kusner, K. Q. W. Yu Sun and F. Sha, "Supervised Word Mover's Distance," NeurIPS Proceedings, 2016.

[11] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K., 2015, June. From word embeddings to document distances. In International conference on machine learning (pp. 957-966). PMLR.

[12] T. Mangono, P. Smittenaar, Y. Caplan, V. S. Huang, S. Sutermaster, H. Kempt and S. K. Sgaier, "Information-Seeking Patterns During the COVID-19 Pandemic Across the United States: Longitudinal Analysis of Google Trends Data," J Med Internet Res 2021, vol. 23, no. 5, 2021.

[13] J. L. Lim, C. Y. Ong, X. Beiqi and L. L. Low, "Estimating Information Seeking-Behaviour of Public in Malaysia During Covid-19 by Using Google Trends," Malays J Med Sci, vol. 27, no. 5, pp. 202-204, 27 October 2020.

[14] A. I.Bento, T. Nguyeh, C. Wing, F. Lozana-Rojas, Y.-Y. Ann and K. Simon, "Evidence from internet search data shows information-seeking responses to news of local COVID-19 cases," PNAS, 2020.

[15] S. Boon-Itt and Y. Skunkan, "Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study," JMIR Public Health And Surveillance, vol. 6, no. 4, 2020.

[16] J. C. Lyu and G. K. Luli, "Understanding the Public Discussion About the Centers for Disease Control and Prevention During the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis Study," Journal Of Medical Internet Research, vol. 23, no. 2, 2021.

[17] A. Amara, M. A. H. Taieb and M. B. Aouicha, "Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis," Applied Intelligence, vol. 51, pp. 3052-3073, 2021.

[18] T. Mangono, P. Smittenaar, Y. Caplan, V. S. Huang, S. Sutermaster, H. Kempt and S. K. Sgaier, "Information-Seeking Patterns During the COVID-19 Pandemic Across the United States: Longitudinal Analysis of Google Trends Data," J Med Internet Res 2021, vol. 23, no. 5, 2021

[19] J. Junger and T. Keyling, "Facepager. An application for automated data retrieval on the web," 2019. [Online]. Available: https://github.com/strohne/Facepager/.

[20] Z. Husein, "Malaya, Natural-Language-Toolkit library for bahasa Malaysia, powered by Deep Learning Tensorflow," 2018. [Online]. Available: https://github.com/huseinzol05/malaya.