A Social Media Mining Using Topic Modeling and Sentiment Analysis on Tourism in Malaysia During Covid19

N A Deraman1*, A G Buja1, K A F A Samah1, M N H H Jono1, M A M Isa1, and S Saad1

1 Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, Malaysia
Email: noora465@uitm.edu.my*

Abstract. Malaysia’s tourism is affected by the Covid19 pandemic and the MCO implementation, where borders are closed and non-essential activities are halted. Negative effects are also felt by Malaysians and are reflected in social media. This study examines two research questions, finding the issues that Twitter users have been addressing on tourism activities during the MCO period and analyze users’ sentiment regarding their ability to travel after MCO. 5000 data were extracted manually from 11357 data scraped from Twitter, of which 3243 were pre-processed keywords using RapidMiner. The results show that the topic of the debate focuses on three themes, namely the destination of tourism, future planning, and public emotions. In addition, 63% gave a positive view and 22% negative sentiment on domestic tourism. Overall, users of Twitter gave an optimistic outlook on domestic travel and hoped that Covid19 would soon be over.

1. Introduction
The Covid 19 pandemic that affected the world started at the end of 2019, leading to an unprecedented public health, social, and economic emergency. Travel and tourism are among the most affected industries, with on-site airlines, hotels closed, and travel restrictions imposed in most countries around the world [1]. The enforcement of the Movement Control Order (MCO) in Malaysia, beginning on 17th March 2020 to curb the spread of COVID-19, has listed domestic and international tourism activities as prohibited [2]. Also affecting the Malaysian tourism industry is the closing of borders. As a result, the tourism industry loses RM 53.5 billion in tourist receipts from March to September 2020 [3]. This condition continued until the Recovery MCO implementation on 10th June 2020, the government eased the restriction by giving public accountability for adhering to the set standard operating procedures (SOPs) and reopening the national economy in a regulated manner. Following this development, zone mapping has been carried out the government where the infection rate is measured within 14 days. Areas in the green zone category are areas that have not reported new Covid-19 positive cases in 14 days, while the yellow zone has recorded 1-20 cases. The orange zone has reported 40 cases and the red zone has more than 41 Covid-19 positive cases in 14 days [4]. As Malaysians are ready to initiate the new norm, domestic tourism activities have received a very encouraging response.

However, Malaysia was struck again by the third wave of Covid-19, in which some states shifted their status from green to red zone. This is resulting in the implementation of Targeted Enhanced MCO (TEMCO) in some states where TEMCO’s implementation applies to a much smaller area rather than to whole districts or cities, such as a residential complex or an office building, with tighter regulations. All social activities and movements inside and outside TEMCO’s premises are
forbidden. Furthermore, where screening and monitoring are compulsory, all residents must be subject to mandatory quarantine. Once again, this affects the tourism industry, and people are more vigilant in their everyday lives. This implementation affects not only the tourism industry but also the mental and physical aspects of Malaysians. Being confined to one’s home has a negative impact on the person, and vacationing is one way to destress oneself [5]-[6]. In addition, the tourism industry should also take into account the need to wear masks and 3C campaigns by the Ministry of Health (MoH): avoiding closed areas, enclosed spaces and closed contact.

Social media such as Facebook, Twitter, and Instagram are defined as applications of a website that allows users to share their ideas or views and participate in social networking. It has become a vital resource for coordinating different real-life events by instantly providing a broad audience forum [7]. In line with the advent of digital technology, various parties, such as hotels and travel agents, adopted social media platforms to provide information and promote tourism in Malaysia. Taking advantage of social media high market penetration, the industry relied primarily on the destination’s popularity, the visitors’ opinion, the distribution of information, and the positive word of mouth advertisement [8]. Ranked fourth in the world in the field of mobile social penetration, Malaysians have publicly discussed MCO issues and the closure of non-essential activities from the date of implementation of the MCO [9].

To understand Malaysia's perception on tourism during MCO and pandemic, further investigation needs to be conducted. This study uses Twitter as a data source and RapidMiner as a medium for implementing the opinion mining process to address the following research questions:

RQ1: What is the Malaysian topic of tourism-related discussion during the MCO?
RQ2: What are the Malaysian perceptions towards domestic travel during the Covid 19 Pandemic?

The rest of the paper is structured as follows. Section 2 outlines the relevant issues regarding topic modeling and sentiment analysis. Section 3 detail the process of opinion mining and operator used in RapidMiner. Section 4 addresses the overall summarization result of the opinion mining process and answering the research questions stated. Section 5 draws some conclusions and possible work.

2. Related works

2.1 Topic Modeling

Mining opinion on social media has become a large subject of study. One of the methods used in addition to sentiment analysis is topic modeling. Topic modeling is one of the text analytics tools whose primary purpose is to derive information from a large text data volume. It will reflect the information in the set using an unsupervised machine learning method [10]-[11]. Topic modeling has two main objectives: to discover hidden themes in text data by clustering similar words groups accordingly and classifying the document into the discovered theme. The effect of topic modeling is useful in improving classification by grouping related terms together in topics and identifying social media patterns [12]. It is often used in the recommender system with the use of similarity measures.

There are two basic topic models Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA). PLSI is a technique that learns the latent topic by performing matrix decomposition. Meanwhile, LDA is a generative probabilistic corpus model that uses Dirichlet over the latent topic. The basic concept is that documents are represented as random mixtures over latent topics, where each topic is defined by a distribution over words [13]. Although the PLSI is easier to train than the LDA, it has low accuracy.

A lot of research on the topic of modeling using the LDA technique has been done. A study by [14] used LDA to examine the topic discussed during Brexit on Twitter and find the connection between Brexit sentiment and the British pound sterling exchange rate. A similar idea is applied by [15], where Twitter is mined using the topic modeling approach to understand consumer attitudes towards vaccination. LDA technique is used to interpret public perception of the non-pharmaceutical intervention (NPI) method during Covid19 was explored [16]. This study’s findings defined issues relevant to keywords across six countries and assisted the government in the implementation of NPI decision-making and strategies. Apart from that, LDA-based topic modeling has been adopted by [17]
to identify product opportunities among Twitter users and quantify the importance of the topic identified. Similar work has been done by [18] where post related to climate change is clustered according to countries using geotagged tweet data.

2.2 Sentiment Analysis
Sentiment analysis is contextual text mining, which detects and extracts subjective information in the source and seeks to explain the product's social perception or service by monitoring online conversations. It is the most common text classification method that analyses the incoming message and decides whether the underlying sentiment is positive, negative, or neutral. A lot of research has been conducted using sentiment analysis to grasp the public understanding of certain topics.

Research [19] used a deep-learning approach to evaluate sentiments about the reviews that tourists publish online and study how new tourists plan their trip using data from the TripAdvisor website. A similar study [20] was carried out by performing a sentiment analysis to understand user perceptions of Chinese attractions based on Japanese tourism website comments: 4Travel. Research using Twitter as a data context has been performed by [21]–[22]. Their analysis aims to examine tourism perception at the travel site, halal tourism trends globally, and develop SmartSenti, which analyses tourism perception sentiment at tourist locations in Turkey, respectively.

3. Research methodology
A classical opinion mining process comprises of 4 stages: opinion retrieval, pre-processing data, topic modeling, and summarization, as shown in Figure 1.

![Figure 1. Architecture of Opinion Mining](image)

3.1 Opinion retrieval
Opinion retrieval is the mechanism by which social media views and feedback on an issue are made. Postings related to Malaysia tourism during pandemic on Twitter was selected for analysis. A collection of keywords has been established to scrap the related post. Some of the keywords used are “cuti-cuti malaysia”, “percutian domestik”, “traveling,” and the hashtag of states name in Malaysia. This project uses the RapidMiner “Search Twitter” operator to collect data. Attributes collected on
Twitter are the tweet itself, retweet count, and the number of likes received. The data collected is from 1st August 2020 to 30th October 2020. All collected data is stored in a CSV file to be processed.

3.2 Pre-processing
At this stage, raw data collected from Twitter are pre-processed using RapidMiner. All raw data will go through 2 processes: text preparation and text pre-processing. Text preparation is the process of readying the data for the machine learning process through various operators in Rapid Miner. Figure 2 shows the detailed process done using RapidMiner. First, the raw data in excel format will be added to the RapidMiner repository. Then process Extract, Transform, and Load (ETL) is done on the data using RapidMiner operator. Standard operators were used at this process: “Select Attribute”, “Nominal to Text”, and “Process Document to Data”.

After all the data has been prepared, it will be converted into documents for further processing. Four main text pre-processing components consist of tokenization, transform case, remove stop words, and stemming process, as shown in Figure 3. Tokenization is the process of extracting (also called tokenizing) words from the body of the text. Then all text will be converted to lowercase using the “Transform Case” operator of RapidMiner. After that, all terms with a high recurrence rate and little analysis value in the document are omitted using the text “Filter Stopwords” operator. The remaining words will be stemmed using the “Stem” operator to remove suffixes. Similarly, stemming turns noun plurals into singular forms. Lastly, the operator “Generate n-Gram” is set to two and will search for terms that often follow the others.

3.3 Topic Modeling
In RapidMiner, topic modeling is done using Operator Toolbox extension where the LDA method to find hidden theme in processed data. Since topic modeling is unsupervised learning, the data are set to iterate five times. Figure 4 illustrates the process of extracting topic from a document using LDA. All data is pre-processed before it clustered into a similar group. After that, the process of topic annotation is done to group the topic into a meaningful theme.
3.4 Sentiment Analysis

Figure 5 demonstrates the structure of the RapidMiner sentiment analysis implementation. The “Read Excel” operator is used to retrieve data, and the “Set Role” operator is used to tell the system what value to use. Other common operators used are “Nominal to Text” and “Process Document to Data” which correspond to pre-processing. Sentiment analysis preprocessing would be used by the operator “Tokenization,” “Transform Cases” and “Filter Stopwords.” To train the data, the “Cross Validation” operator is used “SVM,” “Apply Model” and “Performance.” The purpose of these models is, essentially, to apply the model to the dataset to get a prediction on the test data. The output of the model is then evaluated.

Figure 3. Text pre-processing activities

Figure 4. Topic modeling process

Figure 5. Sentiment analysis model
3.5 Summarization
Finally, the modeling result is visualized in the form of a table and word cloud format to comprehend better in answering RQ1 and RQ2.

4. Result and Discussion
A total of 11357 data was scrapped on Twitter that comprised English and non-English tweets. However, due to processing limitations, 5000 data are selected. After the data cleaning process has been completed, 3243 keywords are obtained. The reduction is attributable to the removal of URL link, repetition, comments in the form of gif file and emoji, and an unimportant statement by the reader that is not relevant to the subject being scrapped. Non-English words that are meaningless and irrelevant to the issues are omitted manually. Furthermore, words with a document frequency (DF) of less than 15 are removed, and tweets with less than 15 tokenized words are also omitted. This section will answer two research questions of this study.

RQ1: What is the Malaysian topic of tourism-related discussion during the MCO?
In answering RQ1, the detailed process of LDA is analyzed. Topic modeling using LDA required three inputs: the corpus of processed data, the word dictionary acquired at the pre-processing stage, and the number of topics, which is RapidMiner set to 10 by default.

Table 1 shows part of the LDA analysis result where the document is grouped into a relevant topic. The topic extracted from the document is calculated using the LDA analysis. For example, Document 1 has a high confidence value for Topic 6: 0.6144, while the rest are less than 0.100. The topic represented is manually defined by referring to Document 1 and the high-frequency keyword. Document 1 has been linked to domestic traveling, where Topic 6 is Melaka word. Upon completion of the text annotation process, the theme is listed, as shown in Table 2. The keywords that define the subject are described based on the value of confidence received. The outcome of this study shows that the subjects frequently discussed on Twitter among Malaysians are linked to emotion, their planning after MCO, and places they want to go.

| Document | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 0        | 0.0235  | 0.0438  | 0.0727  | 0.0361  | 0.0458  | 0.6068  | 0.0469  | 0.0413  | 0.0207  | 0.0623  |
| 1        | 0.0604  | 0.0392  | 0.0424  | 0.0392  | 0.0423  | 0.6144  | 0.0440  | 0.0436  | 0.0425  | 0.0319  |
| 2        | 0.0425  | 0.0452  | 0.0350  | 0.0494  | 0.0113  | 0.0192  | 0.0500  | 0.0399  | 0.0453  | 0.6623  |
| 3        | 0.0477  | 0.6419  | 0.0466  | 0.0410  | 0.0463  | 0.0324  | 0.0188  | 0.0415  | 0.0409  | 0.0428  |
| 4        | 0.0425  | 0.1011  | 0.0411  | 0.5321  | 0.0444  | 0.0593  | 0.0537  | 0.0419  | 0.0423  | 0.0415  |
Table 2. Topic discovered using LDA

| Topic                  | Major Keywords and Representative Data                                                                 |
|------------------------|--------------------------------------------------------------------------------------------------------|
| Point of Interest      | “Melaka”, “Penang”, “Redang”, “Langkawi”, “Kampung”, “Kelantan”                                       |
|                        | 1. “Pkpb sambung lg. Tepu dah rasa. Lama tak balik kg. Lama tak pergi bercuti.”                         |
|                        | 2. “Nasib sempat pergi bandung bulan 2. Time pkpp bercuti dekat selangor, PD and melaka. Tu je la.”   |
|                        | 3. “Terengganu pls end this. I want to go on holiday in Terengganu.”                                 |
| Emotion                | Stress, Pressure, “Geram”, “Rindu”                                                                    |
|                        | 1. “I think I can get depressed working and going home, I want to go on holiday to PKPB here and there. Haish mental sia .” |
|                        | 2. “Just want to go on holiday to kuantan. canceled plan because pkpb!”                                |
|                        | 3. “Nk meroyan bleh dak sebab nk g bercuti tpi kna pkpb! Geram??”                                     |
| Planning               | “Nikah”, Holiday, family, “Sekolah”                                                                    |
|                        | 1. “Majlis nikah tangguh… huwaaaaaa”                                                                    |
|                        | 2. “But really, i need my holiday. This sem is really challenging. Bru week 4, tpi keje dh bertimbun”   |
|                        | 3. “PKP cepat habis..nak plan balik kg & dah organize activity with family.”                           |

RQ2: What are the Malaysian perceptions towards domestic travel during the Covid 19 Pandemic? Sentiment analysis produces four types of attributes: polarity, confidence in polarity, subjectivity, and confidence in subjectivity. The polarity attribute indicates the mood of the text: positive, negative, or neutral. Polarity confidence will calculate the likelihood that the value set by the previous variable is valid. Subjectivity means writing style or expression, whether it is subjective or objective. The confidence of subjectivity is the probability value of the previous variable is true.

The sentiment analysis result shows that the polarity value for positive sentiment is 63%, negative 22%, and neutral 15%. This model accuracy is 85.23%, which means the model is reliable. From the result, it can be concluded that the majority of Malaysian Twitter users are looking forward to domestic traveling. Since it’s almost the end of the year, most of them are ready to start their holidays. The most-tweeted place with a frequency of 574 is Melaka, followed by Penang and the Islands in Malaysia: Sipadan, Pulau Redang, and Langkawi. A high percentage value for negative feelings due to canceled vacations and Conditional MCO implementation, which is predicted. Figure 7 shows the top 20 positive and negative words with the high total occurrence and document occurrence in the corpus.

![Figure 7](image-url)

Figure 7. Top 20 positive (left) and negative (right) words in sentiment analysis

With the Covid-19 pandemic not over anytime soon, as one of the drivers of the Malaysian economy, Malaysia's tourism industry is most affected. As people spend money on hotels, food, shopping, which benefits a wide range of industries, it is also the sector with the biggest multiplier effect. The government has to take proactive action in order to improve the economy of the country by taking into account the wellbeing of the citizens. Based on the results of this survey, a positive rate of 63% suggests that people are excited to continue living with new standards and eager to help move the economy of the country through domestic tourism. Furthermore, the key topics discussed were holiday
planning and family activities, tourist destinations in Malaysia and their feelings during quarantine, certainly helps the government, especially tourism operators, to make strategic planning by offering a package at affordable prices through attractive promotions. The provision of unique and attractive travel packages will boost domestic tourism during the MCO. Government initiatives will also enable people to take advantage of the current situation through individual income tax relief of up to RM1,000 for domestic tourism spending until 31 December 2021.

Future planning, which is a vital topic addressed by the public, should be taken into account. While maintaining a safe and healthy work environment, tourist operators can arrange all sorts of activities (indoor and outdoor) that comply with the health and safety guidelines laid down by the MoH and the National Security Council, such as guided nature walks or hiking trips with a limited number of participants, travelling food trail, attractive spa and wellness packages and virtual classes: cooking and baking, ceramic classes. One of the optimistic terms obtained is waterpark, where the discounted entrance ticket is definitely welcomed by those impacted by MCO: a decrease in salaries. All this planning also opens up job opportunities for the surrounding communities and helps them to grow their family economy.

5. Conclusion
The objective of this paper is to study Malaysia's perception of tourism activities during the MCO through two research questions. This study analyses 5000 clean tweets from Malaysia's Twitter user and measures their sentiment towards domestic tourism. The results of this study have shown that the main theme addressed in social media is future planning, Malaysia's destination of tourism and public emotion. These findings lead to a 63% positive acceptance of domestic travel and 22% negative sentiment. The outcome of this study allows the government and the tourism authorities to consider the public's concerns and viewpoints. The effect of this information and public voice enables the various parties in the tourism sector visualize their potential consumer needs and to be prepared to provide facilities and activities that comply with the Covid-19 SOP. Maintaining the sustainability of these activities and encouraging continuous development during the crisis are essential to pave the way for future success in the tourism sector. Despite the contributions made to this study, more work still needs to be done. A comparison of techniques is needed for the study of hidden subjects in the corpus. A practical comparison with the LDA solution can be made in terms of performance, as the LDA cannot handle implicit aspects of the solution.

6. References
[1] Henry-biabaud, O. (2020). COVID-19 - a global insight on travel and tourism impacts, UNWTO & Data Partners COVID-19. (April), 31.
[2] Shah, A. U. M., Safri, S. N. A., Thevadas, R., Noordin, N. K., Rahman, A. A., Sekawi, Z., ... Sultan, M. T. H. (2020). COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government. International Journal of Infectious Diseases, 97, 108–116. https://doi.org/10.1016/j.ijid.2020.05.093
[3] Tourism Malaysia. (2020), Malaysia’s Tourism Performance Records Negative Growth Of 78.6% From Jan – Sep 2020, Retrieved from: https://www.tourism.gov.my/media/view/malaysia-s-tourism-performance-records-negative-growth-of-78-6-from-jan-Sep-2020
[4] Aziz, N. A., Othman, J., Lugova, H., & Suleiman, A. (2020). Malaysia’s approach in handling COVID-19 onslaught: Report on the Movement Control Order (MCO) and targeted screening to reduce community infection rate and impact on public health and economy. Journal of Infection and Public Health, 13(12), 1823–1829. https://doi.org/10.1016/j.jiph.2020.08.007
[5] Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N., & Rubin, G. J. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. The Lancet, 395(10227), 912–920. https://doi.org/10.1016/S0140-6736(20)30460-8
[6] De Bloom, J. (2010). How do vacations affect workers’ health and well-being? Vacation (after-) effects and the role of vacation activities and experiences.
[7] J. Kawash (ed.), Online Social Media Analysis and Visualization, Lecture Notes in Social Networks, 2014 DOI 10.1007/978-3-319-13590-8_1
[8] Irfan, A., Rasli, A., Sami, A., & Liaquat, H. (2017). Role of social media in promoting education tourism. Advanced Science Letters, 23(9), 8728–8731. https://doi.org/10.1166/asl.2017.9959

[9] Social Media Marketing Statistic in Malaysia, Retrieved from: https://marketingsignallab.com/social-media-marketing-statistic-in-malaysia/ Dec 2019

[10] Tribhuvan, P. P., Bhirud, S. G., & R. Deshmukh, R. (2017). Product Features Extraction for Feature Based Opinion Mining using Latent Dirichlet Allocation. International Journal of Computer Sciences and Engineering, 5(10), 128–131. https://doi.org/10.26438/ijcse/v5i10.128131

[11] Stokes, D. C., Andy, A., Guntuku, S. C., Ungar, L. H., & Merchant, R. M. (2020). Public Priorities and Concerns Regarding COVID-19 in an Online Discussion Forum: Longitudinal Topic Modeling. Journal of General Internal Medicine, 35(7), 2244–2247. https://doi.org/10.1007/s11606-020-05889-w

[12] Lamba, Manika, & Madhusudhan, Margam. (2018). Application of Topic Mining and Prediction Modeling Tools for Library and Information Science Journals. In Library Practices, Digital Era: Festschrift in Honor of Prof. V Vishwa Mohan. Hyderabad: BS Publications. http://doi.org/10.5281/zenodo.1298739

[13] ANKARALI, E., & KÜLCÜ, Ö. (2020). Topic Modeling of Twitter Data. Bilgi Yönetimi, 3(1), 0–10. https://doi.org/10.33721/by.641878

[14] Ilyas, S. H. W., Soomro, Z. T., Anwar, A., Shahzad, H., & Yaqub, U. (2020). Analyzing brexit’s impact using sentiment analysis and topic modeling on twitter discussion. ACM International Conference Proceeding Series, (June), 1–6. https://doi.org/10.1145/3396956.3396973

[15] Zhang, H., Wheldon, C., Dunn, A. G., Tao, C., Huo, J., & Zhang, R. (n.d.). Mining Twitter to Assess the Determinants of Health Behavior towards Human Papillomavirus Vaccination in the United States Corresponding

[16] Doogan C, Buntine W, Linger H, Brunt S, Public Perceptions and Attitudes Toward COVID-19 Nonpharmaceutical Interventions Across Six Countries: A Topic Modeling Analysis of Twitter Data, J Med Internet Res 2020;22(9):e21419, DOI: 10.2196/21419

[17] Byeongki Jeong, Janghyeok Yoon, Jae-Min Lee, Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis, International Journal of Information Management, Volume 48, 2019, Pages 280-290, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2017.09.009.

[18] Dahal, B., Kumar, S.A.P. & Li, Z. Topic modeling and sentiment analysis of global climate change tweets. Soc. Netw. Anal. Min. 9, 24 (2019). https://doi.org/10.1007/s13278-019-0568-8

[19] Martin, C. A., Torres, J. M., Aguilar, R. M., & Diaz, S. (2018). Using deep learning to predict sentiments: Case study in tourism. Complexity, 2018. https://doi.org/10.1155/2018/7408431

[20] Yu, C., Zhu, X., Feng, B., Cai, L., & An, L. (2019). Sentiment analysis of Japanese tourism online reviews. Journal of Data and Information Science, 4(1), 89–113. https://doi.org/10.2478/jdis-2019-0005

[21] Gupta, G., & Gupta, P. (2019). Twitter mining for sentiment analysis in tourism industry. In Proceedings of the 3rd World Conference on Smart Trends in Systems, Security and Sustainability, WorldS4 2019 (pp. 302–306). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/WorldS4.2019.8903940

[22] Sulaiman Ainin, Ali Feizollah, Nor Badrul Anuar, Nor Aniza Abdullah, Sentiment analyses of multilingual tweets on halal tourism, Tourism Management Perspectives, Volume 34,2020, 100658,ISSN 2211-9736, https://doi.org/10.1016/j.tmp.2020.100658.

Acknowledgments
Sincere appreciation goes to Universiti Teknologi MARA Cawangan Melaka for the support given to this research endeavor, TEJA: Internal Grant (GDT2020-14).