Leveraging social media data using latent dirichlet allocation and naïve bayes for mental health sentiment analytics on Covid-19 pandemic

Nurzulaikha Khalid [b,1], Shuzlina Abdul-Rahman [a,b,2,*], Wahyu Wibowo [c,3], Nur Atiqah Sia Abdullah [b,4], Sofianita Mutalib [a,b,5]

[a] Research Initiative Group of Intelligent Systems, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[b] College of Computing Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[c] Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, East Java, Indonesia
[1] zulaikhakhalid.zk@gmail.com; [2] shuzlina@fskm.uitm.edu.my ; [3] wahyu_w@statistika.its.ac.id; [4] atiqah684@uitm.edu.my;
[5]sofi@fskm.uitm.edu.my
* corresponding author

ARTICLE INFO

Article history
Received June 30, 2022
Revised April 13, 2023
Accepted September 19, 2023
Available online October 15, 2023

Selected paper from The 2022 5th International Symposium on Advanced Intelligent Informatics (SAIN'22), Yogyakarta (Virtually), September 14, 2022, http://sain.ijain.org/2022/. Peer-reviewed by SAIN'22 Scientific Committee and Editorial Team of IJAIN journal.

Keywords
Social Media
COVID-19
Latent Dirichlet Allocation (LDA)
Lexicon-Based
Mental Health
Naïve Bayes

ABSTRACT

In Malaysia, during the early stages of the COVID-19 pandemic, the negative impact on mental health became noticeable. The public's psychological and behavioral responses have risen as the COVID-19 outbreak progresses. A high impression of severity, vulnerability, impact, and fear was the element that influenced higher anxiety. Social media data can be used to track Malaysian sentiments in the COVID-19 era. However, it is often found on the internet in text format with no labels, and manually decoding this data is usually complicated. Furthermore, traditional data-gathering approaches, such as filling out a survey form, may not completely capture the sentiments. This study uses a text mining technique called Latent Dirichlet Allocation (LDA) on social media to discover mental health topics during the COVID-19 pandemic. Then, a model is developed using a hybrid approach, combining both lexicon-based and Naïve Bayes classifier. The accuracy, precision, recall, and F-measures are used to evaluate the sentiment classification. The result shows that the best lexicon-based technique is VADER with 72% accuracy compared to TextBlob with 70% accuracy. These sentiments results allow for a better understanding and handling of the pandemic. The top three topics are identified and further classified into positive and negative comments. In conclusion, the developed model can assist healthcare workers and policymakers in making the right decisions in the upcoming pandemic outbreaks.

1. Introduction

The World Health Organization (WHO) has declared COVID-19 a public health pandemic [1]. The pandemic of COVID-19 has overtaken the world. Many countries wordwide, including Malaysia, are experiencing a crisis that may result in high mortality and morbidity rates [2]. The negative impact on mental health becomes noticeable in the early stages of the COVID-19 pandemic in Malaysia. The public's psychological and behavioural responses have risen as the COVID-19 outbreak progresses. A high impression of the elements of severity, vulnerability, impact, and fear has influenced a higher level of anxiety [3]. Social media data can be used to track Malaysian sentiments in the COVID-19 era. However, social media data is often found on the internet in text format with no labels, and manually decoding this data is usually a complicated process [4]. Despite the introduction of automated interpreting methods, the underlying technology still needs advancement. Furthermore, traditional data gathering approaches, such as filling out a survey form, may not completely capture the sentiments [5], [6]. Researchers can discover the topic being discussed, such as mental health during the COVID-19 outbreak, by utilizing a popular data mining technique called topic modelling on social media.

To date, there is still a lack of investigation on mental health in the context of COVID-19 in Malaysia. Wong et al. [7] claim that there are limited studies on the uncertainties over how the prolonged physical incarceration and containment measures imposed because of the COVID-19 epidemic have affected the mental health of Malaysians. There is also a lack of study on how the COVID-19 pandemic's resulting economic breakdown has affected the mental health of Malaysians because mental health is not given enough exposure and attention [8]. Furthermore, there may be cultural variations that necessitate the exploration of these issues in Malaysia [9]. Repeated tracking of social media data could provide a diachronic perspective on public morale and collective attitude alterations, as individuals actively contribute to narratives, providing unprompted and varied understandings of various circumstances [10]. On top of that, this study shows the possibility of utilising sentiment analysis on social media data to investigate the mental state among social media users, which can offer valuable proxies for mental health. This research was able to build methods that may assess, on a scale, whether components of the mental health issue posted on social media are more positive or negative in tone by including lexicon-based approaches sentiment analysis into the study. To evaluate the performance of the sentiment analysis, this study implements the Naïve Bayes classifier to both models.

Therefore, the main focus of this study is to determine the topics of conversation on social media related to the COVID-19 pandemic using Latent Dirichlet Allocation (LDA) topic modelling approach and to determine the varied levels of mental health on COVID-19 related topics using lexicon based sentiment analysis. In the next section of the paper, Section 2 reviews the related work, Section 3 explains the material and method used in this research, Section 4 summarizes the results, and Section 5 concludes this paper with suggestions for future research.

2. Related Work

This section reviews the related works that begin with the overview of mental health and the usage of social media in sharing knowledge and information. Then, it covers the deployment of Artificial Intelligence during COVID-19. This section ends with sentiment analysis and the application of machine learning in sentiment analysis.

2.1. Mental Health

Mental health is defined as "a condition of well-being in which each individual fulfils his or her own potential, is able to cope with regular life challenges, is able to work successfully and fruitfully, and is able to contribute to her or his community" [11]. In general, academics agree that mental health plays a critical role in determining an individual's overall health and well-being. Mental illness can impact psychological and physiological components of one's health [12]. Furthermore, pandemics have a long history of being linked to severe mental repercussions [13]. For example, stress, worry, symptoms of depression, insomnia, denial, rage, and dread are just a few of the significant mental health disorders linked to the COVID-19 pandemic that have been recorded globally [14]. Mental health illnesses are becoming more prevalent among diverse population groups. COVID-19, according to a recent report published in JAMA Psychiatry, may enhance the risk of suicide [15]. A recent Chinese study [16] finds that COVID-19 has led to 16.5% of moderate to severe depression, 28.8% of moderate to severe anxiety, and 8.1% of moderate to severe stress. Other countries, including Japan, Singapore, and Iran, have experienced similar effects of COVID-19 on mental health [17]. Stress, worry, and sadness are all linked to the COVID-19 pandemic, according to research conducted around the world, which reveal an increase in the prevalence of mental health issues among diverse population groups [18]. Individuals may resort to extreme methods due to sadness and depression caused by the loss of a loved one, fear and panic caused by an unclear future, and financial hardship [19].

2.2. Social Media

Social media refers to “Web-based services that enable individuals, groups, and organizations to cooperate, connect, interact, and develop a community by enabling them to produce, co-create, modify, share, and engage with easily accessible user-generated content” [20]. Social media is also recommended as a source of data to monitor social interactions on conservation-related events [21], [22], and understand worldwide patterns of trade in wildlife [23]. Data acquired through citizen science programmes, in which individuals voluntarily give data for research in a systematic manner, differs from data created spontaneously on social media [24]. Social media analysis methods are advancing rapidly in computer science and other related domains. In conservation science, these methodologies are often used with a delay [25].

Medical practitioners also utilize these platforms to provide patient care and education, as well as to increase personal knowledge of news and discoveries and also to distribute health information to the general public. Furthermore, as a way of discussing and debating scientific facts, these platforms are increasingly popular. Glowacki et al. [26], for example, studied tweets about electronic cigarettes posted by physicians from the United Kingdom and the United States, and discovered that physicians discussed critical subjects like the possibility of electronic cigarette usage among minors, food, and more. According to Wahbeh et al. [27], Physicians use Twitter primarily to communicate clinical news from scientific meetings, discuss treatment issues, market themselves, and give social commentary.

With the recent outbreak of the COVID-19 epidemic, many nations have enforced travel and movement restrictions, as well as "lockdowns" [28]. Social distancing campaigns, travel restrictions, self-quarantines, and company closures have increased globally. As individuals can no longer openly access public venues, most discussions about the COVID-19 outbreak take place in online forums and social networking sites [29]. Furthermore, medical experts across the world utilize social media, such as Twitter and Facebook, as they become key players in the COVID-19 epidemic. People use social media at this time to share thoughts based on their present state of mind and to convey their emotions to their loved ones.

Facebook, one of the social networking sites, has become among the most widely used platforms for social media [30], [31]. Many people around the globe use this medium to share their ideas, thoughts, emotions, joys, and poems. According to various research and observations, Facebook is the most adaptable social media platform as issues may be openly discussed. Facebook status posts are also more concise than reviews and are easier to categorise. This results in better writing and a more accurate description of emotions [32].

2.3. Artificial Intelligence (AI) for COVID-19

Artificial intelligence (AI) emulates human intellect in computers trained to think and act like humans. The term may also apply to any computer with human-like features such as learning and problem-solving ability. AI is a cutting-edge technology that can help combat the COVID-19 pandemic [33]. According to the World Health Organization (WHO), the most common outcome of a COVID-19 epidemic is severe pneumonia [34]. COVID-19 can be deadly for patients who develop pulmonary symptoms [35]. As a result, AI is being deployed to aid in the fight against the viral pandemic that has been sweeping the planet since 2020. The press and scientific community are optimistic that data science and artificial intelligence (AI) may be utilised to tackle the COVID-19 [36]. This AI technology helps radiologists and clinicians to make faster diagnoses by screening, monitoring, and predicting patients today and in the future. The primary use of this AI is early identification and diagnosis of illnesses. AI is being utilised to manufacture drugs and vaccines, as well as to reduce the burden of medical personnel [37].

Besides, AI applications have also been developed to gain a more in-depth understanding of methodologies that can rapidly classify novel viruses by identifying their intrinsic genomic signatures [38]. Deep Learning methods can be used to help control the illness after the basic processes of transmission have been understood. In particular, for human behaviour control, which is not directly related to a specific pandemic, pre-existing software for mask usage tracking and distance identification may be utilised. For example, as mentioned by [39], suggested a hybrid model for face mask identification that included deep transfer learning with traditional ML classifiers [40].

2.4. Sentiment Analysis

Sentiment analysis, also known as emotion AI or opinion mining, is an approach to analysing social media contents using natural language processing (NLP) and text analysis to systematically measure, extract, identify, and assess effective states and personal information [41]. The primary concept is to analyse textual data from multiple sources to determine the polarity of a sentence, paragraph, or entire document. Opinion mining refers to determining whether text has a positive, negative, or neutral viewpoint [42]. Text polarity represents the public's mood or an individual's viewpoint [43]. Sentiments can also be categorized into n-point scales: very good, good, satisfactory, bad, and very bad.

Sentiment analysis is commonly used in customer-facing content, such as responses to a survey and reviews. Millions of likes and retweets can be used to analyse people's sentiments, but this large involvement with a post does not necessarily indicate the significance of the sentiments toward that post [41]. This is due to several factors, including happiness, irony, satisfaction, sadness, and anger. The elements above can influence the content of a post. On the other hand, broad extractions of human sentiments from social media networks are important and have a considerable effect on global trends, market choices, and policy development [44]. This emphasizes the significance of sentiment analysis in interpreting human emotions.

According to Bose et al. [13], there is a glossary and law-dependent sentiment analysis tool named VADER, which is standardized solely for social media sentiments. VADER is a rule-based, open-source tool that identifies popular phrases, idioms, acronyms, and jargon while taking into account grammatical features, such as punctuation, negation, hedging, and magnification, frequently employed in the vernacular social networking platforms [45]. The VADER lexicon is one of the biggest of its kind, with over 7,500 frequently-used phrases rated for emotional valence by ten independent human raters. The term virus and its various derivatives (e.g., viruses and viral) are not included in the VADER lexicon; thus, changes in their frequency will not affect VADER ratings. VADER has been thoroughly validated for Twitter [46] with some of the highest accuracy and coverage for tweets in comparison to over 20 sentiment analysis tools [47].

2.5. Machine Learning in Sentiment Analysis

Machine learning, lexicon-based approaches, and hybrid techniques are used for sentiment analysis. The polarity of a sentence is already specified in the labelled dataset used in the machine learning methods. Once computers are exposed to fresh data, machine learning allows them to develop, alter, and learn independently [48]. Machine learning algorithms employ a variety of computational approaches to extract information from the data without depending on pre-programmed equations. As the number of samples available for learning grows, the algorithms strive to enhance their performance [49]. Training and testing are the two stages of processing. The model is trained using labelled data that includes both input and output in supervised learning. A dataset with labels is supplied to a classification algorithm during training, which generates a model. The results of the tests are put into a model that predicts the category [50]. On the other hand, unsupervised learning approaches do not require training data or labelled data. It discovers the unlabelled data's hidden structures or patterns [51].

In text mining, some documents, such as blog posts or news articles, must be gathered and then classified into topics so that people can comprehend them independently and clearly. Several researchers have recently concentrated on utilizing topic models to detect latent topics from text. Text mining and natural language processing are the foundations of topic modelling, a computational social science technique. It analyses text data to automatically find cluster terms for a collection of texts [52], [53]. In political science and rhetorical analysis, topic modelling has gained tremendous popularity [54]. Topic modelling is the most commonly used unsupervised learning approach for text classification in text mining [43], latent data exploration, and connection discovery between data and text to identify terms and phrases in a series of documents [55]. Topic modelling is a statistical text mining tool to identify possible (hidden) trends in a data corpus and to classify main words in a corpus as topics. It is a quick and straightforward technique to start examining data because it does not require any training [56]. The fundamental purpose of a topic model is to cluster documents in a text domain; each document has a topic probability distribution, and documents with a high probability for the same subject may be put together as a cluster [57], [58]. As a result, unlike traditional clustering, a topic model permits data from multiple clusters instead of just one.

Latent semantic indexing (LSI) is the beginning of topic modelling and serves as a foundation for its evolution [59]. However, because LSI is not a probabilistic model, it is not a probabilistic topic model. Hofmann [60] presents probabilistic latent semantic analysis (PLSA) as a genuine topic model based on LSI. Blei et al. [61] presents Latent Dirichlet allocation (LDA) as an even more thorough probabilistic generative model and an extension of PLSA. Nowadays, a rising number of probabilistic models based on LDA in association with specific goals are being developed. All the above topic models were first introduced in the text analysis community for unsupervised topic discovery in a corpus of documents. A study by Kherwa & Bansal [62] review topic modelling states that in LDA, the top words of all subjects indicate highly crisp topics, clearly separated and cohesive to tell the nature of distinct topics. Many previous researchers have used the LDA topic modelling method for clustering terms. Xue et al. [63] mentioned that their results show that tweet topic modelling effectively presents information about COVID-19 topics and concerns.

3. Method

This section presents the methodology of this study. It starts with the data collection on the secondary data from Facebook. The data is then pre-processed to obtain clean data. It continues with the topic modelling to select the topics and frequencies. Two sentiment analysis techniques are used to determine the mental health status. Lastly, the model is developed and evaluated.

3.1. Data Collection

This study used secondary data from the “Kementerian Kesihatan Malaysia” Facebook page regarding COVID-19 pandemic in Malaysia. A total of 74,266 comments were scrapped from “Kementerian Kesihatan Malaysia” Facebook page from 1 June 2021 to 31 August 2021 comprising those in English and Malay. The process of scraping the data used the library Facebook-scraper from Python and saved it in the form of CSV (comma separated values) format. The extracted dataset consists of 40 metadata, such as username, time, number of likes, number of comments, shares and reaction, text post, and the full comments. The full comment provides the user details name such as username, time and date when the comment was made, the comment text and also details of the replies to the comments. With all these metadata available only the comment text is concentrated upon to study the sentiment expressed by the commenter towards the COVID-19 pandemic. Researchers filter the dataset to consist of only the COVID-19 pandemic posts only by looking at the “COVID-19” keyword.

3.2. Data Pre-Processing

Researchers filter the dataset to consist of only posts on the COVID-19 pandemic by manually looking for the “COVID-19” keyword. Then, the dataset is processed. Researchers perform various data cleaning methods, such as removing irrelevant comments, translating, removing special characters, converting words to lowercase, tokenization, deleting stop words from the corpus, and lemmatization, all of which are critical tasks in text analytics. Firstly, comments in gif and emoji, repetition and statements irrelevant to the subject being scrapped are manually omitted. Then, Malay words are translated to English using google-translator library version 4.0 in Python. Special characters such as '! " # $ % & \ ' ( ) * are removed from the datasets because these characters provide no value to text-understanding and induce noise in the algorithm. After researchers convert the words into lowercases, the dataset is tokenized. Tokenization is the process of turning sensitive data into non-sensitive data. Researchers then use the NLTK package in Python to eliminate stop words. Stop words include articles (e.g., "a", "an", and "the") and prepositions, which have no inherent meaning. Stop words are not useful in text analytics, and they have to be eliminated from the corpus before the analysis begins. Researchers then use the NLTK package to lemmatize each term to determine the root or stem. Lemmatization is the process of combining different types of words to reduce dimensionality. There is a total of 8,890 data when the cleaning process finishes.

3.3. Topic Modelling

This study uses Latent Dirichlet Allocation (LDA) for topic modelling. LDA is a generative model that may be used to model how documents are formed given a set of subjects and their words. LDA begins by identifying the words in each document, then develops a topic mixture based on a predetermined set of subjects, with topic selection primarily based on the document's multinomial distribution, and word selection based on the document’s multinomial distribution. LDA is an unsupervised method that uses related indicators to detect the semantic relationship between words in a group. In line with previous research, the researchers use the Gensim (RARE Technologies Ltd) [64] coherence model to determine the most appropriate number of topics based on the data. Researchers choose the number of topics that generate the highest coherence value. The higher the coherence score, the easier it is to comprehend the topic's word distribution on which subjects belong to it [65]. Each model's coherent topics are calculated using the coherence score by importing Coherence Model from Gensim, a model library in Python. The number of topics selected ranges from 2 to 15, and the coherence score for each method with k topics is calculated. At the same time, researchers also display the line graph for easy reference of the coherence score versus the number of topics using Microsoft Excel.

Three topics are extracted for further analysis and discussion as this number of topics has the highest coherence score compared to others. The LDA is utilized from the LDA model library in Python to extract keywords form the 3-topic set. After extracting the keywords using LDA, the authors manually validate and label the topics by referring to the high-frequency keywords. Then, the 2D Plane of Intertopic Distance is displayed by using the pyLDAvis package in Python to visualize the distance between the topics selected and the frequency of the terms mentioned in each topic.

3.4. Text Polarity with VADER and TextBlob

For this study, two sentiment analysis approaches are used which are VADER (Valence Aware Dictionary for Sentiment Reasoning) and TextBlob. Sentiment analysis are used to determine the varied levels of mental health on COVID-19 related topics. Topic obtained through the LDA approach that related to mental health is evaluated on both approaches.

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a tool that analyses sentiments in social media and classifies posts based on a dictionary of terms. It is a lexicon-based method to detect sentiment polarity in comments. Unlike traditional dictionaries, VADER's dictionary includes emoticons, slang, contractions, negations, and acronyms often used in casual online conversations. VADER additionally considers degree modifiers that impact sentiment intensity, word order, and sensitive interactions between terms. The VADER sentiment analysis is based on a lexicon that maps lexical elements to emotional intensity using sentiment scores. The sentiment score is calculated by adding the intensity of each word in the text. VADER may categorise a sentiment as negative, neutral, or positive using a compound score that is calculated by summing up the valence ratings of each word in the lexicon and normalised in the range (-1,1), with "-1" being the most negative and "1" the most positive [66]. The text is considered positive if the compound score is greater than 0.05, neutral if the score is between 0.05 and -0.05, and negative if the score is less than -0.05 [67]. To perform binary classification, positive and neutral comments are combined, and neutral tweets are coded as positive, as practiced in previous validation studies [68]. The binary sentiment classification (positive/neutral vs. negative) is more accurate than trinary classifications (positive, neutral, negative) [68].

TextBlob is a Python library for processing textual data. It provides a standardized API for common NLP activities. TextBlob is similar to a Python string in terms of functionality. In this study, TEXT BLOB is applied by importing TextBlob from the text blob package. TextBlob’s sentiment function returns two characteristics, which are polarity and subjectivity. The float polarity is in the range [-1,1], with 1 denoting a positive statement and -1 a negative statement. The text is considered positive if the polarity score is greater than 0.00, neutral if the score equals 0.00, and negative if the score is less than 0.00 [67]. Similar to VADER, neutral comments are coded as positive in binary classification. This study imports Seaborn into Python using TextBlob to display the sentiment bar chart.

3.5. Model Development and Evaluation

This study uses a 70:30 split ratio where 70% of the data is used for the models' training while 30% is for testing. Researchers imported Naïve Bayes classifier in Python to classify the dataset. Naive Bayes is a group of supervised machine learning techniques used for classification to predict the sentiment of the comments. It predicts membership probabilities for each class in the dataset, such as the likelihood that a given data item belongs to a specific class. The performance of the Naive Bayes classifier is calculated using the confusion matrix. The confusion matrix illustrates the number of correctly and incorrectly predicted positive and negative comments by the classifier. [69]. Accuracy (ACC), precision (P), recall (R), and F-measure (F) values are employed as performance measures in this study. The experiment is modelled with the support of machine learning Naive Bayes, the most common text mining classifier which uses Bayes theorem to calculate the possibility of the given label related to a particular feature [70]. Moreover, Naive Bayes is used due to their high accuracy on textual data [71].

4. Results and Discussion

This section presents the topic modelling results using Latent Dirichlet Allocation (LDA). It continues with the text polarity analysis on the selected topic. Then this section highlights the discussion and insights on the model development and evaluation results.

4.1. Topic Modelling using Latent Dirichlet Allocation

Fig. 1 shows the findings of the coherence score for the LDA model. Three is the best number of themes in LDA, with a coherence score of 0.5606. A higher coherence score indicates that a topic's word distribution is easier to comprehend [65]. As can be seen, the best coherence score comes with three topics, then the score declines from four to 15. For this study, three topics are extracted for further analysis and discussion as this number receives the highest coherence score compared to others.

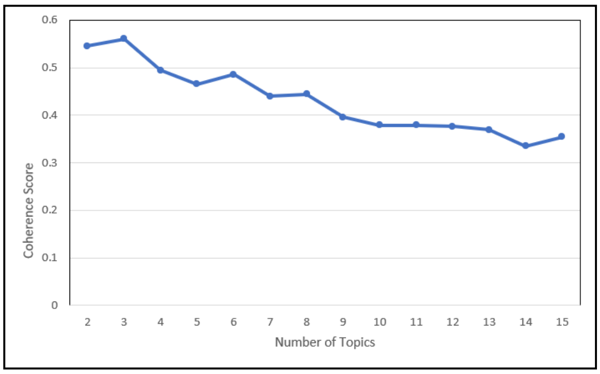

Fig. 1. Coherence Score for LDA

Based on the topic modelling result in Table 1, the three topics extracted from the keywords are "Covid-19 Cases with Hospital Quarantine", "Lockdown and Mental Health", and "Vaccination with Covid-19 Cases". In Topic 1, ten keywords are extracted that contribute to the topic. These are "case", "covid", "day", "hospital", "people", "new", "please", "state", "quarantine", and "time". These keywords are sorted by their respective weights. For example, the weight of "case" in Topic 1 is 0.031, and "covid" is 0.021.

To characterize the underlying content of the topics, it is easier to label them rather than present them as a combination of words. Unfortunately, automatic labelling of topics is not possible as discovering the topics is an unsupervised learning process. It requires human intervention to examine the coherence and meaningfulness of the topics and subsequently label them through their judgement [59]. After extracting the topics using LDA, researchers validate and label the topics, which can be seen in the third column of Table 1. The represented topic is manually defined by referring to the high-frequency keyword. The keywords that define the subject are described based on the value of confidence received. The total number of documents in each dominant topic in Topic 1, Topic 2, and Topic 3 are 2,561, 2,910, and 3,419.

Table 1. List of Extracted Topics

Topic	Keywords	Topic extracted from keywords
1	0.031*"case" + 0.021*"covid" + 0.017*"day" + 0.012*"hospital" + 0.011*"people" + 0.011*"new" + 0.011*"please" + 0.010*"state" + 0.010*"quarantine" + 0.010*"time"	Covid-19 Cases with Hospital Quarantine
2	0.054*"people" + 0.023*"case" + 0.023*"stay" + 0.023*"home" + 0.022*"government" + 0.017*"mental" + 0.016*"work" + 0.015*"covid" + 0.014*"sop" + 0.014*"factory"	Lockdown and Mental Health
3	'0.041*"vaccine" + 0.025*"case" + 0.024*"still" + 0.024*"covid" + 0.022*"people" + 0.019*"already" + 0.016*"day" + 0.015*"factory" + 0.014*"dose" + 0.014*"high"	Vaccination with Covid-19 Cases

Some representative comments on each topic are generated to explain the themes of these topics. The topic distance and a 2D plane of the intertopic distance are presented in Fig. 2. Each bubble on the left represents a topic: Topic 1 (Covid-19 Cases with Hospital Quarantine), Topic 2 (Lockdown and Mental Health), and Topic 3 (Vaccination with Covid-19 Case). The centres are determined by computing the distance between topics. All three bubbles show decent size, which means that all three topics are prevalent. Furthermore, the bubbles do not overlap and are scattered throughout the chart, meaning that the topic modelling has good cross-validation of the classification for the three themes. On the right is a list of the most frequently used terms for the topic and the frequency of occurrence. Based on the result, the word “people”, followed by “vaccine”, has the highest mention in the dataset. The words “home”, “stay”, “government”, and “factory” are also frequently mentioned.

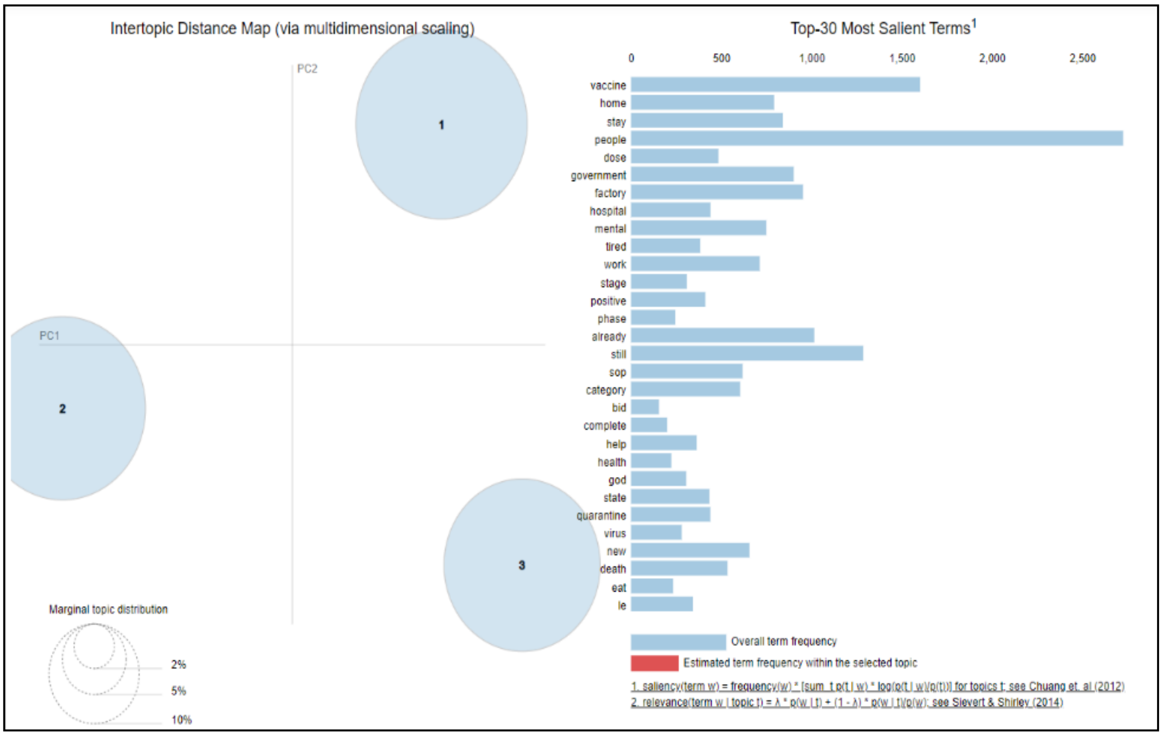

Fig. 2. 2D Plane of Intertopic Distance

Table 1 illustrates that the topic with an element of mental health is Topic 2 (Lockdown and Mental Health). This topic and the keywords show that most COVID-19 cases, and extended home stays or lockdowns, can affect a person's mental health because the words "stay at home" and "mental" are often mentioned in the same context. Some commenters express great displeasure with the outcome, believing that the outbreak cannot be managed when policymakers ignore the pandemic's severity. People believe that factories are to blame for the rise in COVID-19 instances, and cases increase due to the policymakers' incompetence in allowing factories to run resulting in a prolonged lockdown. On the other hand, lockdown requires an individual to stay home for an extended period while also practising SOP (Standard Operating Procedure). Spending extra time at home can be exceedingly stressful if the individual lives in a toxic home environment [2]. People are also becoming more stressed as the COVID-19 instances do not appear to recede despite the lockdown.

Apart from that, many lose their jobs because of the lockdown. Losing a job may impact one's emotions and cause instability and uncertainty, which can lead to mental health issues, such as anxiety and depression. Financial troubles rapidly set in and many groups among the general population, particularly those in the B40 and M40 categories, have either lost or are on the verge of losing their source of income [2]. Despite the government's stimulus packages designed to alleviate the financial hardships faced by many Malaysians, many small and medium enterprises (SMEs) in the country are forced to cut wages, reduce the number of employees, and enforce unpaid leave for an indefinite period due to the country's economic uncertainty. After the topic modelling process is complete, researchers choose a mental health-related topic for further experimentation on the issue. Topic 2 is more conducive to mental health based on the topic modelling results. The sentiment analysis is performed using open-source tools: TextBlob and VADER. These tools are integrated with the Natural Language Toolkit (NLTK) library in Python using Jupiter notebook.

4.2. Model Development

Table 2 shows the results of the number of comments in training and testing sets. A total of 2,210 and 700 comments are set as training and testing data, respectively. The training set with positive and negative classes based on VADER has 1,038 and 1,172 comments, respectively. The testing set has 320 and 380 comments with positive and negative classes, respectively, based on VADER. The training set with positive and negative classes based on TextBlob has 1,257 and 953 comments, respectively. The testing set with positive and negative classes based on TextBlob has 404 and 296 comments, respectively.

Table 2. Distribution of comments

	Number of Comments					
	VADER		***Total***	***TextBlob***		***Total***
	Pos	***Neg***		***Pos***	***Neg***	
Training Data	1038	1172	2210	1257	953	2210
Testing Data	320	380	700	404	296	700

Based on Table 3, the number of actual positive comments is 320, and the number of actual negative comments is 380. The algorithm predicts 297 positive comments and 403 negative ones based on the VADER lexicon-based approach. The total number of comments (N) is equal to 700. The total number of positive and negative comments correctly estimated is 211 and 294, respectively.

Table 3. Confusion matrix of VADER

	Predicted Positive	**Predicted Negative**	**Total**
Actual Positive	211	109	320
Actual Negative	86	294	380
Total	297	403	N = 700

Based on Table 4, the number of actual positive comments is 404, and the number of actual negative comments is 296. The system estimated 356 positive comments and 344 negative ones based on the

TextBlob lexicon-based approach. The total number of comments (N) is the same as VADER. The total number of positive and negative comments correctly predicted is 274 and 214, respectively.

Table 4. Confusion matrix of TexBlob

	Predicted Positive	Predicted Negative	Total
Actual Positive	274	130	404
Actual Negative	82	214	296
Total	356	344	N = 700

4.3. Model Evaluation

This section contains the results of the Naive Bayes model with each lexicon-based approach. The performance of the Naive Bayes model with each lexicon-based approach is evaluated by observing the scores of accuracy, precision, recall, and F-measure. The performance results are illustrated in Table 5 which shows the sentiment results of the Naive Bayes model, which reveals that machine learning performs better with VADER than with TextBlob.

Table 5. Performance measures for VADER and TextBlob

Lexicon Based	Accuracy (%)	Precision (%)		Recall (%)		F-Measure (%)	
		Pos	*Neg*	*Pos*	*Neg*	*Pos*	*Neg*
VADER	72.0	71.0	73.0	66.0	77.0	68.0	75.0
TextBlob	70.0	77.0	62.0	68.0	72.0	72.0	67.0

As seen in Table 5, Naive Bayes performs better with VADER sentiment analyzer for the lexicon-based approach. The accuracy value of the VADER lexicon-based approach is 72% which means it is regarded as a good model. The precision scores for both positive and negative sentiments using this approach are 71% and 73%, indicating that this model is good at predicting the sentiments on COVID-19 mental health topics. The recall value for negative sentiments is 11% higher than the positive ones, which means they perform better in classifying negative sentiments on COVID-19 mental health topics. The higher value of the F-measure for negative sentiments at 75% means that this model has a perfect balance of precision and recall for negative sentiments on COVID-19 mental health topics.

The accuracy of Naïve Bayes with TextBlob sentiment analyzer reaches 70%, as shown in Table 3, 2% lower than VADER. The precision score for positive sentiments is 15% higher than the negative ones indicating that when it predicts the sentiments on COVID-19 mental health topics, it is correct 77% of the time. The higher value of recall for negative sentiments at 72% indicates that this model correctly predicts negative sentiments on COVID-19 mental health topics. The F-measure value for positive sentiments is 72%, higher than the negative sentiment for this lexicon-based approach. It indicates that this model with TextBlob sentiment analyzer is a good model that can be used to classify negative sentiments on COVID-19 mental health topics.

5. Conclusion

This study sets out to (i) identify the topics related to the COVID-19 pandemic discussed on social media using the Latent Dirichlet Allocation (LDA) and (ii) classify the sentiment on COVID-19-related topics using lexicon-based approaches. Three topics are discovered as a result of the topic modelling technique; they are “Covid-19 Cases with Hospital Quarantine”, “Lockdown and Mental Health”, and “Vaccination with Covid-19 Cases”. After the topic modelling process is complete, researchers choose Topic 2, a mental health-related topic for future experiments on sentiments. In Topic 2, "Lockdown and Mental Health", most individuals disagree with the policymaker's decision to allow factories to function during the lockdown, even though factories are a primary source of the rising number of COVID-19 cases. As a result, the lockdown is extended, and many are concerned about their job security, being laid off, and losing their source of income.

This study performs sentiment classification on comments that fall under Topic 2. Neutral and positive sentiments are combined to perform binary classification. VADER lexicon-based classification sees negative sentiment with the highest score of 1,552 comments. TextBlob lexicon-based classification sees positive sentiment score higher than negative with 412 comments. This study uses a 70:30 split ratio where 70% of the data is used for the models' training while 30% is taken for testing supported by Naive Bayes. Among both the lexical-based methods, VADER shows the highest accuracy at 72% compared to TextBlob at 70%. It is also observed that the lexicon-based approach in classifying sentiments from social media text using VADER has a good impact on the Naive Bayes classifier, especially in classifying negative sentiments on COVID-19 mental health topics. Moving forward, the extension of the work can be carried out with other machine learning or deep learning methods. Future studies can include other social media platforms, like Twitter and YouTube. A future study can also extend the timeline of the extracted data period or consider conducting a systematic study through the period of interest explored through social-media analysis. This will enable more fine-grained spatio-temporal analysis, allowing more robust comparisons from reciprocal findings and deeper insights for policymakers.

Acknowledgment

The authors would like to express the gratitude to Research Management Center, Shah Alam, Selangor, Malaysia for the financial support from 600-IRMI/FRGS 5/3 (209/2019) through the Ministry of Higher Educationn grant (FRGS/1/2018/ICT04/UITM/02/9), and Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research support.

Declarations

Author contribution. Nurzulaikha Khalid is the main author who conducted the study and wrote the paper. Shuzlina Abdul-Rahman and Wahyu Wibowo are the supervisors who validate the methodology and results. Sofianita Mutalib and Nur Atiqah Sia Abdullah reviewed and improved the paper. All authors have read and agreed to the published version of the manuscript.

Funding statement. This research was supported by the Research Management Center, Shah Alam, Selangor, Malaysia (600-IRMI/FRGS 5/3 (209/2019)) through the Ministry of Higher Education grant (FRGS/1/2018/ICT04/UITM/02/9).

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] A. A. Zanke, R. R. Thenge, and V. S. Adhao, "COVID-19: A pandemic declare by world health organization," *IP Int. J. Compr. Adv. Pharmacol.*, vol. 5, no. 2, pp. 49–57, 2020, doi: 10.18231/j.ijcaap.2020.012.

[2] H. Shanmugam, J. A. Juhari, P. Nair, S. K. Chow, and C. G. Ng, "Impacts of COVID-19 Pandemic on Mental Health in Malaysia: A Single Thread of Hope | Shanmugam | Malaysian Journal of Psychiatry," *Malaysian J. Psychiatry Ejournal*, vol. 29, no. 1, pp. 78–84, 2020. Available at : https://www.mjpsychiatry.org/index.php/mjp/article/view/536/415

[3] L. Ping Wong and H. Alias, "Temporal changes in psychobehavioural responses during the early phase of the COVID-19 pandemic in Malaysia," *J. Behav. Med.*, vol. 44, pp. 18–28, 2021, doi: 10.1007/s10865-020-00172-z.

[4] P. E. Kummervold *et al.*, "Categorizing Vaccine Confidence With a Transformer-Based Machine Learning Model: Analysis of Nuances of Vaccine Sentiment in Twitter Discourse," *JMIR Med. informatics*, vol. 9, no. 10, p. e29584, 2021, doi: 10.2196/29584.

[5] Z. Wang, V. Joo, C. Tong, and D. Chan, "Issues of social data analytics with a new method for sentiment analysis of social media data," *Proc. Int. Conf. Cloud Comput. Technol. Sci. CloudCom*, vol. 2015-Febru, no. February, pp. 899–904, 2015, doi: 10.1109/CloudCom.2014.40.

[6] M. Ghiassi and S. Lee, "A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach," *Expert Syst. Appl.*, vol. 106, pp. 197–216, 2018, doi: 10.1016/j.eswa.2018.04.006.

[7] L. P. Wong *et al.*, "Escalating progression of mental health disorders during the COVID-19 pandemic: Evidence from a nationwide survey," *PLoS One*, vol. 16, no. 3 March, pp. 1–14, 2021, doi: 10.1371/journal.pone.0248916.

[8] M. F. bin Hassan, N. M. Hassan, E. S. Kassim, and M. I. Hamzah, "Issues and Challenges of Mental Health in Malaysia," Int. J. Acad. Res. Bus. Soc. Sci., vol. 8, no. 12, pp. 1685-1696, Dec. 2018, doi: 10.6007/IJARBSS/V8-I12/5288.

[9] W. D. Shoesmith *et al.*, "Reactions to symptoms of mental disorder and help seeking in Sabah, Malaysia," *Int. J. Soc. Psychiatry*, vol. 64, no. 1, pp. 49–55, 2018, doi: 10.1177/0020764017739643.

[10] S. E. Cho, K. Jung, and H. W. Park, "Social media use during Japan ' s 2011 earthquake : How Twitter transforms the locus of crisis communication," vol. 149, no. 1 pp. 28–40, Nov. 2013, doi: 10.1177/1329878X1314900105.

[11] WHO, "WHO urges more investments, services for mental health 2019," 2019. Available at : https://www.who.int/news/item/12-08-2010-who-urges-more-investments-services-for-mental-health

[12] C. Wang *et al.*, "A longitudinal study on the mental health of general population during the COVID-19 epidemic in China," *Brain. Behav. Immun.*, vol. 87, pp. 40–48, 2020, doi: 10.1016/j.bbi.2020.04.028.

[13] D. Bose, P. S. Aithal, and S. Roy, "Survey of Twitter Viewpoint on Application of Drugs by VADER Sentiment Analysis among Distinct Countries," *Int. J. Manag. Technol. Soc. Sci.*, vol. 6, no. 1, pp. 110–127, 2021, doi: 10.47992/IJMTS.2581.6012.0132.

[14] J. Torales, M. O'Higgins, J. M. Castaldelli-Maia, and A. Ventriglio, "The outbreak of COVID-19 coronavirus and its impact on global mental health," *Int. J. Soc. Psychiatry*, vol. 66, no. 4, pp. 317–320, 2020, doi: 10.1177/0020764020915212.

[15] Y.-T. Xiang, Y. Jin, and T. Cheung, "Joint international collaboration to combat mental health challenges during the coronavirus disease 2019 pandemic," *JAMA psychiatry*, vol. 77, no. 10, pp. 989–990, 2020, doi: 10.1001/jamapsychiatry.2020.1057.

[16] C. Wang *et al.*, "Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China," *Int. J. Environ. Res. Public Health*, vol. 17, no. 5, p. 1729, 2020, doi: 10.3390/ijerph17051729.

[17] Y. Song et al., "COVID-19 treatment: close to a cure? A rapid review of pharmacotherapies for the novel coronavirus (SARS-CoV-2)," Int. J. Antimicrob. Agents, vol. 56, no. 2, pp. 1-8, Aug. 2020, doi: 10.1016/J.IJANTIMICAG.2020.106080.

[18] R. Mohindra, R. Ravaki, V. Suri, A. Bhalla, and S. M. Singh, "Issues relevant to mental health promotion in frontline health care providers managing quarantined/isolated COVID19 patients," *Asian J Psychiatr*, vol. 51, no. 3, p. 102084, 2020, doi: 10.1016/j.ajp.2020.102084.

[19] A. Roy, A. K. Singh, S. Mishra, A. Chinnadurai, A. Mitra, and O. Bakshi, "Mental health implications of COVID-19 pandemic and its response in India," *Int. J. Soc. Psychiatry*, vol. 67, no. 5, pp. 587–600, 2021, doi: 10.1177/0020764020950769.

[20] L. McCay-Peet and A. Quan-Haase, "What is social media and what questions can social media research help us answer," *SAGE Handb. Soc. media Res. methods*, pp. 13–26, 2017, doi: 10.4135/9781473983847.n2

[21] E. Lunstrum, "Feed them to the lions: Conservation violence goes online," *Geoforum*, vol. 79, pp. 134–143, 2017, doi: 10.1016/j.geoforum.2016.04.009

[22] D. W. Macdonald, K. S. Jacobsen, D. Burnham, P. J. Johnson, and A. J. Loveridge, "Cecil: A moment or a movement? Analysis of media coverage of the death of a lion, Panthera leo," *Animals*, vol. 6, no. 5, pp. 26-38, 2016, doi: 10.3390/ani6050026.

[23] E. Di Minin, C. Fink, T. Hiippala, and H. Tenkanen, "A framework for investigating illegal wildlife trade on social media with machine learning," *Conserv. Biol.*, vol. 33, no. 1, p. 210-2014, 2019, doi: 10.1111/cobi.13104.

[24] L. See *et al.*, “Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information,” *ISPRS Int. J. Geo-Information*, vol. 5, no. 5, pp. 55-77, 2016, doi: 10.3390/ijgi5050055.

[25] K. Sherren, M. Smit, M. Holmlund, J. R. Parkins, and Y. Chen, “Conservation culturomics should include images and a wider range of scholars,” *Front. Ecol. Environ.*, vol. 15, no. 6, pp. 289–290, 201, doi: 10.1002/fee.1507.

[26] E. M. Glowacki, A. J. Lazard, and G. B. Wilcox, “E-cigarette topics shared by medical professionals: a comparison of tweets from the United States and United Kingdom,” *Cyberpsychology, Behav. Soc. Netw.*, vol. 20, no. 2, pp. 133–137, 2017, doi: 10.1089/cyber.2016.0409.

[27] A. Wahbeh, T. Nasralah, M. Al-Ramahi, and O. El-Gayar, “Mining physicians’ opinions on social media to obtain insights into COVID-19: Mixed methods analysis,” *JMIR Public Heal. Surveill.*, vol. 6, no. 2, pp.1-10, 2020, doi: 10.2196/19276.

[28] G. Kaplan and Z. Y. Avdan, “COVID-19: Spaceborne nitrogen dioxide over Turkey,” *Eskişehir Tech. Univ. J. Sci. Technol. A-Applied Sci. Eng.*, vol. 21, no. 2, pp. 251–255, 2020, doi: 10.18038/estubtda.724450.

[29] E. Chen, K. Lerman, and E. Ferrara, “Covid-19: The first public coronavirus twitter dataset,” *JMIR Public Heal. Surveill,* vol. 6, no. 2, pp. 1–9, May 2020, doi: 10.2196/19273.

[30] S. Kamal and M. S. Arefin, “Impact analysis of facebook in family bonding,” *Soc. Netw. Anal. Min.*, vol. 6, no. 1, p. 9, 2016, doi: 10.1007/s13278-015-0314-9.

[31] H. T. Vu, M. Blomberg, H. Seo, Y. Liu, F. Shayesteh, and H. V. Do, “Social media and environmental activism: Framing climate change on Facebook by global NGOs,” *Sci. Commun.*, vol. 43, no. 1, pp. 91–115, 2021, doi: 10.1177/1075547020971644

[32] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and As. Perera, “Opinion mining and sentiment analysis on a twitter data stream,” in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*, IEEE, 2012, pp. 182–188, doi: 10.1109/ICTer.2012.6423033.

[33] J. Frankenfield, “Artificial Intelligence (AI),” 2022. Available at : https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp

[34] A. Haleem, M. Javaid, and R. Vaishya, “Effects of COVID-19 pandemic in daily life,” Curr. Med. Res. Pract., vol. 10, no. 2, pp. 78–79, Mar. 2020, doi: 10.1016/J.CMRP.2020.03.011.

[35] X. Mei *et al.*, “Artificial intelligence–enabled rapid diagnosis of patients with COVID-19,” *Nat. Med.*, vol. 26, no. 8, pp. 1224–1228, 2020, doi: 10.1038/s41591-020-0931-3.

[36] L. Wynants *et al.*, “Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal,” *BMJ*, vol. 369, pp.1-22, 2020, doi: 10.1136/bmj.m1328.

[37] R. Saha, S. Aich, S. Tripathy, and H. C. Kim, “Artificial intelligence is reshaping healthcare amid covid-19: A review in the context of diagnosis & prognosis,” *Diagnostics*, vol. 11, no. 9, pp. 1–15, 2021, doi: 10.3390/diagnostics11091604.

[38] G. S. Randhawa, M. P. M. Soltysiak, H. El Roz, C. P. E. de Souza, K. A. Hill, and L. Kari, “Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study,” *PLoS One*, vol. 15, no. 4, pp. 1–24, 2020, doi: 10.1371/journal.pone.0232391.

[39] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, “A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic,” *Meas. J. Int. Meas. Confed.*, vol. 167, pp. 1-11, 2021, doi: 10.1016/j.measurement.2020.108288.

[40] F. Piccialli, V. S. di Cola, F. Giampaolo, and S. Cuomo, “The Role of Artificial Intelligence in Fighting the COVID-19 Pandemic,” *Inf. Syst. Front.*, vol. 23, no. 6, pp. 1467–1497, 2021, doi: 10.1007/s10796-021-10131-x.

[41] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, “Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media,” *Appl. Soft Comput.*, vol. 97, pp. 1-14, 2020, doi: 10.1016/j.asoc.2020.106754.

[42] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press, 2015, doi: 10.1017/CBO9781139084789.

[43] P. Chauhan, “Sentiment Analysis: A Comparative Study of Supervised Machine Learning Algorithms Using Rapid miner,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. V, no. XI, pp. 80–89, 2017, doi: 10.22214/ijraset.2017.11011.

[44] A. Alamoodi *et al.*, “Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review,” *Expert Syst. Appl.*, vol. 167, p. 1-13, 2020, doi: 10.1016/j.eswa.2020.114155.

[45] D. Valdez, M. ten Thij, K. Bathina, L. A. Rutter, and J. Bollen, “Social media insights into US mental health during the COVID-19 pandemic: Longitudinal analysis of twitter data,” *J. Med. Internet Res.*, vol. 22, no. 12, pp. 1–11, 2020, doi: 10.2196/21418.

[46] S. Elbagir and J. Yang, “Language Toolkit and VADER Sentiment,” *Proceedings of the International MultiConference of Engineers and Computer Scientists*, pp, 1-5, 2019. Available at : https://www.iaeng.org/publication/IMECS2019/IMECS2019_pp12-16.pdf

[47] M. T. Ribeiro and C. Guestrin, “Model-Agnostic Interpretability of Machine Learning,” *University of Washington Seattle*, pp. 91-95, 2016, doi: 10.48550/arXiv.1606.05386

[48] R. A. Priyadharshini, S. Arivazhagan, and M. Arun, “A deep learning approach for person identification using ear biometrics,” *Appl. Intell.*, vol. 51, no. 4, pp. 2161–2172, 2021, doi: 10.1007/s10489-020-01995-8.

[49] N. K. Chauhan and K. Singh, “A review on conventional machine learning vs deep learning,” in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, IEEE, 2018, pp. 347–352, doi: 10.1109/GUCON.2018.8675097.

[50] D. Sharma, M. Sabharwal, V. Goyal, and M. Vij, “Sentiment analysis techniques for social media data: A review,” *Adv. Intell. Syst. Comput.*, vol. 1045, no. September, pp. 75–90, 2020, doi: 10.1007/978-981-15-0029-9_7.

[51] K. B. Priya Iyer and S. Kumaresh, “Twitter sentiment analysis on coronavirus outbreak using machine learning algorithms,” *Eur. J. Mol. Clin. Med.*, vol. 7, no. 3, pp. 2663–2676, 2020. Available at : https://ejmcm.com/article_3797.

[52] M. E. Roberts, B. M. Stewart, and D. Tingley, “stm: R package for Structural Topic Models; 2017,” *R Packag. version 0.6*, vol. 21, pp. 1-40, 2016, doi: 10.18637/jss.v091.i02.

[53] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using topic modeling methods for short-text data: A comparative analysis,” *Front. Artif. Intell.*, vol. 3, pp. 1-14, 2020, doi: 10.3389/frai.2020.00042.

[54] R. Debnath and R. Bardhan, “India nudges to contain COVID-19 pandemic: A reactive public policy analysis using machine-learning based topic modelling,” *PLoS One*, vol. 15, no. 9, pp. 1-25, 2020, doi: 10.1371/journal.pone.0238972

[55] Y.-C. Wu, C.-S. Chen, and Y.-J. Chan, “The outbreak of COVID-19: An overview,” *J. Chinese Med. Assoc.*, vol. 83, no. 3, p. 217-220, 2020, doi: 10.1097/JCMA.0000000000000270.

[56] Y. Du, “A Deep Topical N-gram Model and Topic Discovery on COVID-19 News and Research Manuscripts,” *Electron. Thesis Diss. Repos.*, pp. 1-96, 2021. Available at: https://ir.lib.uwo.ca/etd/7797/

[57] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, “An overview of topic modeling and its current applications in bioinformatics,” *Springerplus*, vol.1608, pp. 1-22, 2016, doi: 10.1186/s40064-016-3252-8.

[58] H. Jiang, R. Zhou, L. Zhang, H. Wang, and Y. Zhang, “Sentence level topic models for associated topics extraction,” *World Wide Web*, vol. 22, no. 6, pp. 2545–2560, 2019, doi: 10.1007/s11280-018-0639-1.

[59] G. Narravula, “Text Embedding Based Topic Modeling on Noisy Historical Drilling Data,” *Dalhousie University,* no. Dec-2021, pp. 1-74, 2021. Available at: https://dalspace.library.dal.ca/handle/10222/81119.

[60] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, Association for Computing Machinery, Inc, Aug. 1999, pp. 50–57. doi: 10.1145/312624.312649.

[61] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, Mar. 2003, doi: 10.5555/944919.944937.

[62] P. Kherwa and P. Bansal, “Topic Modeling: A Comprehensive Review,” EAI Endorsed Trans. Scalable Inf. Syst., vol. “7,” no. 24, pp. 1–16, Jul. 2019, doi: 10.4108/EAI.13-7-2018.159623.

[63] J. Xue *et al.*, “Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach,” *J. Med. Internet Res.*, vol. 22, no. 11, pp. 1-14, 2020, doi: 10.2196/20550.

[64] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408, 2015, doi: 10.1145/2684822.2685324

[65] H. M. Alash and G. A. Al-Sultany, “Improve topic modeling algorithms based on Twitter hashtags,” *J. Phys. Conf. Ser.*, vol. 1660, no. 1, pp. 1-10, 2020, doi: 10.1088/1742-6596/1660/1/012100.

[66] H. Yin, S. Yang, and J. Li, “Detecting Topic and Sentiment Dynamics Due to COVID-19 Pandemic Using Social Media,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12447 LNAI, no. June, pp. 610–623, 2020, doi: 10.1007/978-3-030-65390-3_46.

[67] M. Ismail, “Sentiment Analysis of Patients’ Opinions in Healthcare using Lexicon-based Method,” *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 6977–6981, 2019, doi: 10.35940/ijeat.a2141.109119.

[68] C. Borchers, J. M. Rosenberg, B. Gibbons, and M. A. Burchfield, “To Scale or Not to Scale : Comparing Popular Sentiment Analysis Dictionaries on Educational Twitter Data,” *Fourteenth International Conference on Educational Data Mining (EDM 2021)*, pp. 2–7, 2021. Available at: https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_122.pdf

[69] V. D. Chaithra, “Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments,” *Int. J. Electr. Comput. Eng.*, vol. 9, no. 5, pp. 4452–4459, 2019, doi: 10.11591/ijece.v9i5.pp4452-4459.

[70] M. Umair and A. Hakim, “Sentiment Analysis of Students’ Feedback before and after COVID-19 Pandemic Sentiment analysis of Students Feedback before and after COVID-19 Pandemic View project,” *Int. Journal on Emerging Tech.,* vol.12, no.2, pp.177-182, July, 2021. Available at: https://www.researchgate.net/publication/353305417_Sentiment_Analysis_of_Students'_Feedback_before_and_after_COVID-19_Pandemic.

[71] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, “Sentiment analysis using SVM: A systematic literature review,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, pp. 182–188, 2018, doi: 10.14569/IJACSA.2018.090226.