IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 May 2024, accepted 4 June 2024, date of publication 10 June 2024, date of current version 17 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3411717

RESEARCH ARTICLE

Enhanced Sentiment Analysis and Topic Modeling During the Pandemic Using Automated Latent Dirichlet Allocation

AMREEN BATOOL[1] **AND YUNG-CHEOL BYUN**[2]

[1]Department of Electronic Engineering, Institute of Information Science Technology, Jeju National University, Jeju 63243, South Korea
[2]Department of Computer Engineering, Major of Electronic Engineering, Institute of Information Science Technology, Jeju National University, Jeju 63243, South Korea

Corresponding author: Yung-Cheol Byun (ycb@jejunu.ac.kr)

This work was supported by the ''Regional Innovation Strategy (RIS)'' through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE).

ABSTRACT The COVID-19 pandemic has profoundly impacted human societies, resulting in the loss of millions of lives and slowing economic growth worldwide. This devastating pandemic underscores the gravity of viral threats and led to multifaceted consequences, including loss of livelihoods, dynamic labor force migration, and significant ramifications on mental health. Furthermore, different scientific institutions and companies are attempting to accelerate research and innovation by analyzing large data corpus for fighting against the pandemic. In this research study, an advanced approach based on automated Latent Dirichlet Allocation (LDA) is suggested dealing with a large data corpus for efficiently providing visualization of sentiment analysis and discovered topics. This innovative approach seeks to interrogate a substantial pandemic corpus, delving into the intricacies of public sentiment and discerning evolving trends pertinent to the pandemic. A sophisticated 10-topic LDA model was implemented, revealing Topic 8 as the most prevalent, with a frequency peak of 22.29, eclipsing other enumerated topics. We employ text-mining techniques like WordCloud and Word2Vec to offer insights into specific terms relevant to the pandemic, such as ''Origin,'' ''Symptom,'' ''Diagnostic,'' and ''Transmission.'' Applying the t-SNE method enriches the analysis by visually unraveling semantic clusters within the corpus. The subsequent phase involves modeling strategic topics within the corpus through an unsupervised LDA-based approach, leveraging our suggested framework. This novel perspective contributes to a deeper understanding of the underlying dynamics by analyzing a large data corpus quickly and automatically for providing visualization of discovered topics aiming to aid front-line workers, healthcare practitioners, and community support to fight against the pandemic.

INDEX TERMS Topic modeling, LDA, sentiment analysis, machine learning, deep learning, feature extraction.

I. INTRODUCTION

The pandemic has impeded millions of lives. The majority of countries around the world were driven to order a temporary shutdown of their economy to stop the virus from spreading. This pandemic leads towards health crises as well as slowed down the economic growth [1], [2]. In addition, this fact provides evidence of the virus menace. Because of the long-term restriction to houses or residences, the consequences of this pandemic included loss of Coronaviruses (CoVs), which are significant RNA viruses enclosed in single positive strands that infect humans and animals, gastrointestinal tracts, or respiratory systems [3]. Several viruses affect human life, such as Human Coronaviruses (HCoVs), which are known to exist in seven different strains. Beta-CoVs HCoVs-OC43 and HCoVs-HKU1 and alpha-CoVs NL63 and 229E cause only moderate respiratory illnesses [4], [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhengmao Li.

Over the past 20 years, however, two new coronaviruses, CoV (MERS-CoV) and severe acute respiratory syndrome CoV (SARS-CoV) in the Middle East, have caused human infections in local nations and areas and have caused mortality rates around 10% and 35%, respectively [6], [7], [8]. The coronavirus is now a world wide SARS-CoV-2, 7th hcov, (2019-nCoV) formerly. Medical and social astrocytic 2022. tough efforts globally for the pandemic caused by the SARS-CoV-2 coronavirus which started infecting people in late December 2019 worldwide. It is an evolving medical issue, so the global monitoring of the disease and possible effective global research is currently being up-to-date to thwart and regroup the current pandemics of large concern to WHO. Here, is the seventh SARS-CoV-2. It is the seventh coronavirus that appeared in late December 2019, and the most recent one discovered. 14 September 2022, WHO called the COVID-19 pandemic amidst the disease (COVID-19 invoking the shape of the new virus). As of 14 September 2022, 607,083,820 people were afflicted with the disease, and 6,496,721 people died [9].

Furthermore, as the scientific research community strives to develop innovative solutions to mitigate the pandemic, it also paves the way for accelerating the pace of innovation and discovery [10]. In addition, it has become a serious health concern globally, increasing demand for health technology to enhance development for aiding health practitioners and policy makers to provide breakthrough against future pandemic [11]. In addition, research funding institutions have focused significant attention on supporting research and innovation at the utmost pace to address the pandemic. However, to accelerate the pace of the research community towards innovative solutions, timely processing of the large data corpus of the pandemic is required to facilitate growing research. According to scientific evidence, many researchers are writing on different issues, including text for communities with disease outbreaks, epidemic alarms, and other medical services, all of which are included in the discussion. Therefore, an efficient solution is required to timely process the large pandemic data corpus for providing a visualization of discovered topics to scientific researchers and other health practitioners for mitigating with the pandemic.

Several studies have applied Natural Language Processing (NLP) [12] on SM text to the work related to COVID-19. NLP approaches are gaining popularity in processing enormous natural language data. NLP is a multidisciplinary science combining artificial intelligence (AI) and linguistics, leveraging computers to interpret and understand human language. This convergence of knowledge makes machines capable of processing, analyzing, and generating text to enable communication between humans and computers [13]. The importance of NLP nowadays is further heightened by the fact that we produce large amounts of unstructured text data in our daily routine is called entity recognition sentiment analysis [14], machine translation [15], topic modeling [16], text filtering [17], reviews analysis [18], etc. Text summarizing [19] are a few of the most widely used and well-liked NLP approaches. In addition, in [20], the authors employed epistemic network analysis to extract sentiment score in blended learning data. Similarly, in [21], the authors used collaborative filtering approach to analyze users behaviour through social media data.

In the context of the COVID-19 crisis, large language models (LLMs) like BERT offer exceptional NLP capabilities but are computationally and memory-intensive [22], [23]. Binarization, which reduces model weights to 1 bit, significantly cuts these demands, yet often results in performance drops [24]. To address this, methods like BiLLM and BiBERT introduce innovative techniques such as binary residual approximation, optimal splitting search, Bi-Attention structures, and Direction-Matching Distillation (DMD) to enhance accuracy and efficiency [25]. These approaches achieve high-accuracy inference and substantial savings in FLOPs and model size, demonstrating their potential for real-world, resource-constrained scenarios [26]. The author used non-negative matrix factorization and probabilistic latent sentiment analysis to normalize the mutual and distance information [27]. In addition, in [28], the authors proposed a two stage adaptive distillation model to capture aesthetic and context information in crowd-sensing environment. By leveraging these advancements, binarized LLMs can maintain performance while being more computationally and memory efficient. In the context of the COVID-19 crisis, NLP techniques have been crucial for analyzing vast amounts of natural language data from various sources, enabling effective sentiment analysis, topic modeling, and information retrieval.

Automated LDA (Latent Dirichlet Allocation) discussion for the pandemic outbreak involves applying topic modeling techniques to extract key themes or topics from large-scale discussions related to the pandemic. This method helps researchers identify and categorize various aspects of the pandemic outbreak, such as medical issues, public responses, economic impacts, and policy discussions. Moreover, sentiment analysis can provide insights into prevailing sentiments surrounding the pandemic and its effects on different communities by gauging the emotional tone of these discussions.

This study has significant research topics and their connections and applied LDA modeling and NLP to evaluate the current status of literature on COVID-19 and COV infection. This study can also aid research on pandemic coordination by identifying high-priority scientific topics. This research is urgently needed in pathogens, treatments, virus diagnostics, vaccines, and viral genomes, while clinical characterization, epidemiology, and virus transmission are now priorities.

The distinctive contributions of our study are delineated as follows:

- Pioneering a novel approach, we introduce an automated LDA based topic modeling method to scrutinize an extensive pandemic corpus. This method goes beyond conventional analyses, offering an enhanced

understanding of public sentiment and emerging trends linked to the pandemic.
- Elevating the analysis, we employ word-to-vector as an embedding technique, delving into the intricate semantic relationships and similarities among words. Specifically, we explore terms such as origin, symptom, diagnostic, transmission, etc., providing a nuanced perspective on their interconnections with the pandemic.
- Employ statistical analysis to analyze statistical significance of the proposed research study.
- In addition, a detailed comparison is provided to highlight the empirical effectiveness of the proposed research study over the existing studies.

The rest of the paper organized as follows. The literature review on pandemic publications examined sentiment analysis and topic modeling in Section II. Our method and research design section III introduces data preprocessing and profoundly explores the dataset. We explain the precise methodology of this study. The subject distribution and topic representations are discussed in more detail in section IV and V. In section VI, we explain the results and over-generalizations.Finally, the conclusion section VII summarizes how this study fits into the research framework. We also describe the limitations of our research work and provide suggestions for future research.

II. RELATED WORK

In 2020, the pandemic of coronavirus disease (COVID), presented by the WHO (World Health Organization), will occur. The pandemic COVID-19 topic has a lot of research. Addressing issues like the COVID-19 transmission process, the virus symptoms, and psychological conditions of COVID-19 patients, boosting human immunity to prevent health consequences, prediction of COVID-19 data based on the technique of machine learning (ML), and the importance of online tools of technology in this context. Our literature evaluation will examine the trend of general research on COVID-19 and the application of ML algorithms for related research. In the current study, an unsupervised ML method is used to identify gaps in existing literature and suggest future research directions. Lots of searches are to be carried out to analyze this pandemic.

A. TOPIC MODELING

Topic Modeling is a technique that may be used to manage an extensive collection of documents by grouping them according to various subjects. Although topic modeling is often called a clustering option, it is more reliable and frequently provides more accurate results than a clustering technique like k-means. The clustering technique presupposes that each document is assigned a subject, and the distance between them is measured. TM assigns a document to a group of topics with different weights or probabilities without making any assumptions about how close or far apart the subjects are. Are several TMs available, with the latent Dirichlet allocation (LDA) model being the most often used [29]. Global researchers are working to comprehend the COVID-19 pandemic's many facets. Many researcher has also appeared in the literature since the outbreak of COVID-19, which was reported at the end of December 2019. For example, one metric of the growing body of COVID-19 research is a NLP-based analysis of social media posts, scholarly papers and the daily news relevant to the disease. A topic considered as the study of COVID-19 news stories from Canada is presented by Bai et al. in [30]. To examine the news media during the early stages of the COVID-19 epidemic in China [31] adopted a digital topic modeling technique. During the COVID-19 conference, [32] presented a system for identifying and following pertinent subjects from social media. Reference [33] examined how the local public responded to the new Coronavirus (COVID-19).

B. SENTIMENT ANALYSIS OF COVID-19

Some studies used sentiment analysis to examine how individuals responded to the epidemic through social media posts. The tweeter posts and Weibo postings made by China and America between January 2020 and May 2020 during the epidemic were examined by [49]. The results showed that most people were confident in controlling the pandemic, but sentiments of people like fear, sadness, and disgust also appeared worldwide. They compared the people's emotions, i.e., anger, hate, fear, happiness, sadness, and surprise. An existing study of the sentiment dynamics of residents of the Australian state of New South Wales (NSW) throughout the pandemic, [50] retrieved five months' worth of COVID-19-related tweets from Twitter. They grouped tweets into groups based on the worth of local government areas (LGAs) and tracked dynamic mood shifts over time. To dynamically assess the subject and mood of 13 million tweets about COVID-19, [51] devised a unique methodology. In addition, in [52], the authors carried out a cross-sectional study to investigate the impact of negative emotions and risk perception of health practitioners during COVID-19 pandemic. Despite several issues with social media data's biases, confounding, and representatives [53], social media platforms have an estimated 3.96 billion users worldwide. Several methodologies have been used based on characteristics, including Part of Speech (POS), uni-grams, bi-grams, statistical techniques, words, and sentence embedding [54]. In [55], the authors proposed a crowd-sourcing method to estimate moral elevation in medical data to facilitate well-being. Word embedding in Deep Learning (DL) models have received greater attention recently [54]. In [56], the authors developed a DL-assisted model to distinguish between positive and negative emotions in medical data. In [57], Doc2Vec and Word2Vec were used for the sentiment analysis of medical documents. When assessing unsupervised models for the medical domain, the study's authors also employed WordNet's Welsh statistic. In addition, in [58], the authors employed attentive multi-tasking ML model to recognize emotions for sustainable and livable environment

TABLE 1. Critical analysis of existing studies.

Year	Discussion of Topics	Advantages
2021 [34]	Using the COVID-19 dataset from February to March 2020 categorization of sentiment	CNN, Distil BERT, BiLSTM, XLNET, ware used
2021 [35]	No positive emotions of COVID-19 pandemic ware discussed	Remove the content related COVID-19
2021 [36]	Detection of fake new from COVID-19	For improving the accuracy used modified LSTM and modified GRU
2021 [37]	Automatic lung segmentation of CT-images of Coronavirus effected patients	a fully connected layer of (PQIS-Net) gives better results.
2021 [38]	Comparative study of continuous variable	SVM, bag of words and n-gram used
2021 [39]	Identify damaged assessment	The model made use of random forest, SVR, and linear regression technologies.
2021 [40]	Identify multi modal training disaster	Model based on Dense-Net and BERT
2021 [41]	Analyzing outliers in fin-tech data	Optimal RBF Neural Network, SOM and traditional RBF
2020 [42]	Quantum neural networks and quantum	Analysis of rule base fuzzy Gaussian membership
2020 [43]	Back-propagation multi-layer perceptrons	COVID-19 sentiment categorization, LSTM, Neural Network outperformed other machine-learning algorithms in terms of accuracy.
2020 [44]	Analyse the position of critical patients	consider of different analyser of Bayesian, Linear and SVM
2020 [45]	Analyze unstructured data	Multi ML learners and soft-voting assisted ensemble model
2020 [46]	Comparative study of continuous variable quantum neural networks and quantum back-propagation multi-layer perceptrons	Promising results with sporadic and convoluted data.
2020 [47]	Twitter depression dataset was analyzed COVID-19	Used are the pre-trained BERT, RoBERTa, and XLNet transformer classification models
2020 [48]	Sentences assigned automatically for COVID-19 briefings corpus	CNN paired with BERT performs better than CNN mixed with other embeddings (Word2Vec, Glove, ELMo)

development. It continues to be a great source of textually rich, semantic data with excellent chances to monitor various social interaction-related characteristics, particularly conversations on public health challenges.

C. SENTIMENT ANALYSIS AND TOPIC MODELING

From [59], [60], [61], [62], [63], and [64], of the research works cited either performed the topic of Using COVID-19 data for modeling or sentiment analysis. In contrast, there are extremely few studies that have integrated examine using topic modeling and sentiment analysis data from COVID-19 topics covered by Chandrasekaran et al. [65] included Using methods like LDA and VADER; we can analyze the trends and opinions expressed in tweets concerning the COVID-19 pandemic. Xue et al. [66] investigated tweets for assessing public sentiment and conversation during the COVID-19 epidemic. They employed the LDA approach for topic modeling. In addition, in [67], the authors suggested an hybrid NLP model based on multi-layered features to analyze the hidden insights of the data. Similarly, in [68], the authors suggested a neural network (NN) based approach to transform the representations into a unified latent vector space. In [69], the authors investigated emotion using traditional ML in interactive education systems. Similarly, in [70], the authors attempted to investigate difference in behavioural embedding between two entities. Our research fits into a category of studies that use topic modeling and sentiment analysis to evaluate COVID-19 data. Although this research is a positive feature to our understanding of cross-cultural COVID-19 news, the nature of our COVID-19 (research article) and the use of topic modeling and Textual Similarities for sentiment categorization make this research necessary.

III. MATERIALS AND METHODS

In this section, we presented a detailed methodology of the proposed architecture of COVID-19. The main sections of these studies are as follows in Fig 1.

- Data Collection
- Data Prepossessing
- Word Cloud Generation
- Topic Modeling

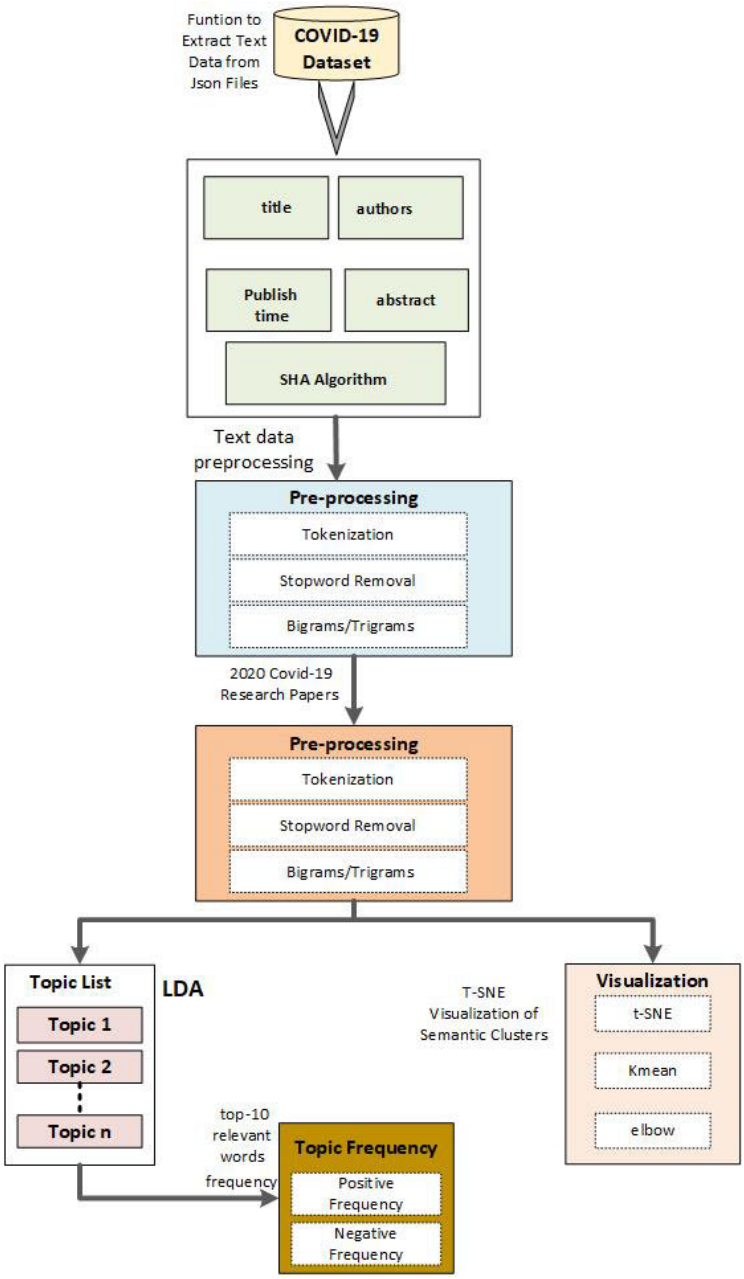


FIGURE 1. Overview model of the proposed methodology.

A. COLLECTION OF DATASET

All abstracts of papers in the scientific literature indexed in the SCOPUS database, and all the documents in the scientific literature on the outbreak of COVID-19, and after COVID-19, and after COVID-19 date, in the literature search of the COVID period paper, minimum of three keywords for searching paper in the scientific literature on COVID-19 are the keywords of COVID-19, coronavirus and SARS-CoV2. A minimum of three keywords for searching papers in the scientific literature on COVID-19, coronavirus, and SARS-CoV2. We found a lot of the papers during the mentioned period. We skipped those papers without abstracts. A total of 1994 papers were finally selected for analysis of COVID-19 academic published research paper data that were extracted from JSON files using push shift API. The histogram of COVID-19 research papers is shown in Fig 2.

B. DATA PREPROCESSING AND CLEANING

The raw data gathered for each step is pre-processed to conduct the analysis more effectively and efficiently. Unprocessed published research paper abstract data might impede analysis since they are filled with incorrect terms

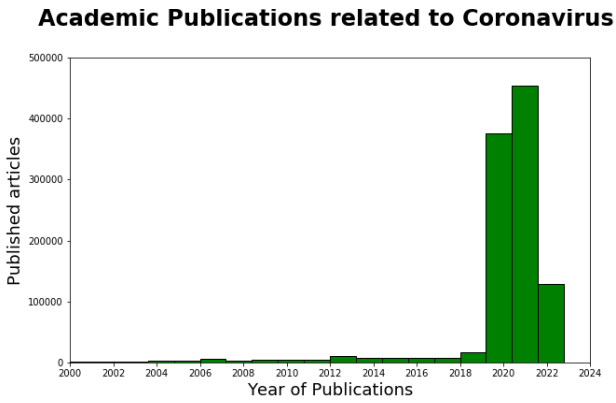


FIGURE 2. COVID-19 publications histogram.

and stop words that provide unclear results. The essential processes for cleaning and converting the data into something usable are all part of the pre-processing. This transforms the text input into a complete form, improving the functionality of ML algorithms. The pre-processing is carried out in the subsequent steps:

C. STOPWORDS REMOVING

A stop word is a term most often used in a language. Examples of these terms in English are "the," "a," "an," "in," etc. These words don't significantly modify the meaning of a statement or its topic, either. As a result, it is permissible to omit them without changing the sense of the statement. By removing these terms, the algorithm can focus more on words that help define the text.

D. TOKENIZATION

The second pre-processing stage involves breaking down abstract phrases and sentences into smaller parts, such as individual words. Tokenization turns each newly acquired smaller unit into a separate entity known as a token. The extracted tokens help create more accurate models and identify the context of the analyzed text.

E. BIGRAMS/TRIGRAMS

A connection between the first two words in each bi-gram in the abstract. In contrast to the co-word use, this bigram occurrence only considers a relation/edge if two words are placed one after the other in a sentence. Similar to the prior bigram occurrence, a relationship model is used instead. There is an additional edge between the first and third words in a trigram of three words.

IV. TOPIC MODELING OF COVID-19 PUBLICATIONS

Topic modeling is an unsupervised classification technique of documents. Management of project and engineering studies have gradually embraced the topic modeling technique known as LDA. An unsupervised ML method called LDA can identify the main topics from a collection of unlabeled texts. Each document in LDA is considered as a probabilistic

TABLE 2. Sentiment analysis and topic modeling of COVID-19.

Symbol	Feature	Description
N	$total-word$	total number of words in document topics
D	$total-doc$	total number of documents
K	$total-topic$	total number of topics
α, η	$dir-prmtr$	Dirichlet parameters
θ_d	$per-doc$	Per document topic probability
Z_d, n	$per-word$	Per word topic assignment
W_d, n	$ob-word$	Observed word
β_k	$topic-dis$	Topic distribution over the vocabulary

distribution over a set of K topics. Each topic $K \in \{1, \cdots, K\}$ is represented as a distribution ϕ k over vocabulary words [71]. Each word contributes in a specific way to each topic.

The Figure below clearly shows the mathematical annotations; for instance, it designates a matrix with rows defined by documents and columns defined by topics, and θ (d, k) indicates the probability that topic k will appear in the document d. Similarly, ϕ is a matrix where the columns are words, and the rows are subjects. Below is a simplified illustration of the LDA procedure.

In LDA, it is assumed that the topic distribution has a Dirichlet prior, resulting in a uniform topic distribution for each document model in Equation IV the probability for a corpus [72]. Fig 3 explains the LDA plate notation, and Table 2displays the significance of the notations.

LDA aims to find topic θ matrix and document ϕ topic that maximizes the following joint probability distribution across the hidden and observable variables.

$$\prod_{d=1}^{N_d} P\left(w_1, \cdots, w_{N_d} \mid \beta, \alpha\right) = \prod_{d=1}^{N_d} \int_{\theta_d} P\left(\theta_d \mid \alpha\right) \times \prod_{n=1}^{N_d} \left(\sum_k \theta_{dk}, \beta_{kw_n}\right) d\theta_d$$

LDA assumes the following generating process given a corpus D made up of M documents, each of length N_i

- Creat $\theta_i \sim \text{Dir}(\alpha)$, where $i \in \{1, 2, \cdots, D\}$. $Dir\alpha$ is a LDA distribution with symmetric parameter α where α is frequently sparse.
- Creat $\beta_k \sim \text{Dir}(\eta)$, where $k \in \{1, 2, \cdots, K\}$ andβ is often spares.
- For the n_{th} space in documents d where $n \in \{1, 2, \cdots, N_d\}$ and $d \in \{1, 2, \cdots, D\}$.
- select a topic Z_d, n for the position is generated the from of $Z_d, n \sim \text{Multinomial}(\theta_i)$.
- Word W_d, n which is produced from the word distribution of the subject chosen in the preceding step $W_i, j \sim \text{Multinomial}(\theta_{zd}, n)$., should be used to fill that position.

This work uses LDA to model themes and separately cover trending topics. In topic modeling, the number of subjects is a significant variable. We utilize the coherence score to calculate the determined number of topics to make

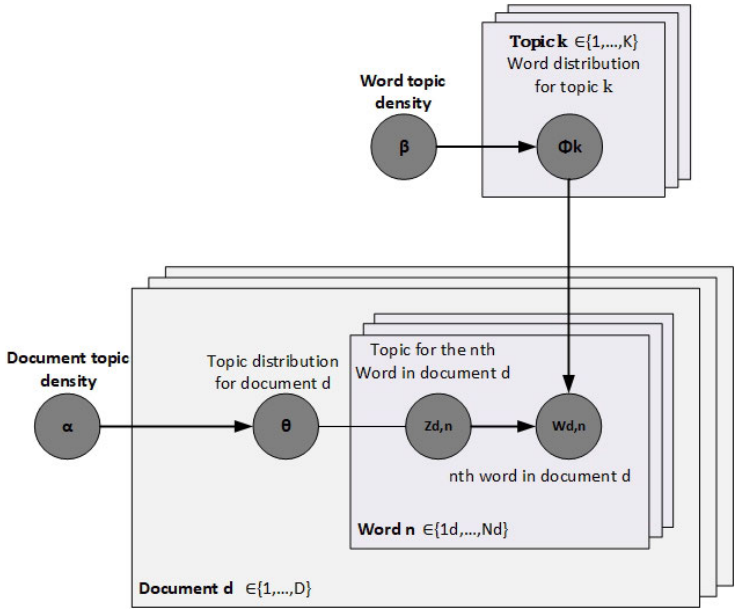


FIGURE 3. Explanation of LDA in topic modeling.

these subjects interpretable by humans. The coherence score in Equation 1 aids in separating themes with a human understanding from statistical inference.

The coherence chooses the top n words in each topic that appear often and averages all the scores pairwise for those topic n words $w_i, \ldots, w_n$ of the topic. Finally, we got the total coherence score for the current topics. The total number of topics across two validation sets, fixed $= 0.01$ and $= 0.1$. We selected the number of subjects to be between 1 and 100. We decide on 10 topics since the findings show that this number produces the maximum coherence score, and we use LDA topic modeling to analyze the abstracts.

$$\text{Coherence} = \sum_{i<j} \text{score}\left(w_i, w_j\right) \tag{1}$$

A. GRID BASED DETERMINATION

We used a grid-search optimization approach to determine the K number of subjects that results in the most compelling model. A detailed overview of documents and word probability is explained in Figure 4. To further explain, after training baseline models spanning the range of K, the C_v computed the coherence measure to estimate the ideal number of topics K for the corpus of abstracts. The topic model's coherence score C_UMass averages the coherence ratings for each subject included in the model. The log causes C_UMass to produce negative values, with values closer to 0 denoting more easily understood topics by humans.

$$C_{\text{UMass}}\left(k; W^{(k)}\right) = \frac{2}{N(N-1)} \sum_{i<j} \log \frac{D\left(w_i^{(k)}, w_f^{(k)}\right) + \varepsilon}{D\left(w_i^{(k)}\right)}$$

$$C_{\text{UMass}} = \frac{1}{K} \sum_{k=1}^{K} C_{\text{UMass}}\left(k; W^{(k)}\right) \tag{2}$$

In Equation 2 where $D(_i)$ is the count of the documents containing the word $w_(i)$ and $D(_i, w_j)$ the count of documents containing both word wi and wj, and $W^{(k)} = \left(w_1^{(k)\cdot}, \ldots, w_N^{(k)}\right)$ is the list of N most probable words of the topic k [73].

V. NON-NEGATIVE MATRIX FACTORIZATION (NMF)

Non-negative Matrix Factorization (NMF) was utilized to extract and identify the underlying topics from a vast collection of research articles. In the vector space model, the non-negative matrix is represented by d x n, where d represents the size of the words in the topic, and n represents the total number of documents.

A. NMF FOR COVID-19 PUBLICATIONS

In Non-negative Matrix Factorization (NMF), the corpus matrix $Z \in \mathbb{R}_{\geq 0}^{d \times n}$ is factorization into two low-rank non-negative matrices: $W \in \mathbb{R}^{d \times x}$, known as the dictionary matrix, and $H \in \mathbb{R}^{x \times n}$, known as the coding matrix. This factorization is accomplished by solving the optimization problem as described in Equation 3:

$$\inf_{W \in \mathbb{R}_{\geq 0}^{d \times x}, H \in \mathbb{R}_{\geq 0}^{x \times n}} \|Z - WH\|_F^2 \tag{3}$$

where $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$ denotes the Frobenius norm of matrix A. NMF is essentially an iterative optimization algorithm. However, it has a significant drawback: the objective function is usually non-convex and possesses multiple local minima. As a result, different random initializations of the NMF procedure can lead to different matrix factorizations. The variability impacts how the results are interpreted., including the topic vector representations in W and the relevance between articles and topics in H.

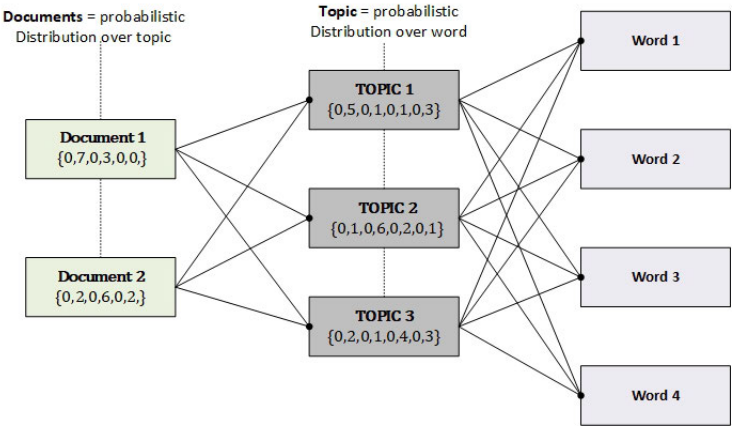


FIGURE 4. Grid-based determination of the optimal number of topics.

Algorithm 1 The Process NMF Algorithm

Step 1:Input Corpus matrix X
Apply Non-negative Matrix Factorization (NMF) to decompose X
into matrices W and H with x topics.
Select the optimal number of topics x^*
by using a threshold value in matrix H,
categorizing the articles into topics $Z_1, \cdots, Z_{x^*}$.
Any articles that do not meet the threshold
are placed in an "Extra Document" matrix Z_e.
The value of $x*$ is chosen to allocate articles
to the relevant topics
Z_1, Z_2 Â· Â· Â·, z_x* based on a specified threshold in matrix H,
and any remaining articles are assigned to an
"Extra Document" matrix x_e;
while No of the articles assigned to a topic
i > m do
Apply NMF to the sub-matrix
Z_i to obtain W_i and H_i with x_i^* sub-topics.
assign the documents to the topics by
the threshold Î± in matrix H.
assign the rest to Z_e;
end
For each article z_i in Z_e **do**
Calculate the cosine similarity between z_i
and each of the topic of leaf
Assign x_i to the most similar topic.
end
repeat the loop process of each topic in
articles until every topic has less than m articles.

Moreover, the choice of the number of topics, k, introduces another source of variability. Different combinations of initial values for W and H, along with varying values of k, produce different topics, thereby leading to different clustering results for the articles.

B. NMF TOPIC VISUALIZATION

We used NMF topic visualization with the algorithm implementation. In Algorithm 1 the data is visualized into 10 topics.

After the topic visualization, topic 8 was assigned the highest accuracy depending on the health and vaccine, and the subtopics were waste, supply, environment, and supply. The public health topic led to articles related to public sentiment about the COVID-19 outbreak.

VI. RESULTS AND DISCUSSION

This section presents experiment results and analysis to evaluate the proposed topic modeling approach. Topic coherence is considered the most frequent word in each generated topic and measures the sentiment simulated between the words of topics. Using either UCI or Umass to perform the pairwise calculations and calculate the mean coherence score across all the topics for the model.

A. SENTIMENT ANALYSIS

This study used 32314 research publications and 428,265 words for sentimental analysis. Among the data, 22.1% had positive sentiments, 12.3% had negative sentiments, and the majority had neutral opinions. It accounted for 64.1%. Table 1 shows how many words were related to each month. Based on the results COVID has a primarily neutral sentiment. Topic 1 records the highest number of positive words, and from there, the tone of the public toward the COVID crisis seems less optimistic. While positive sentiment decreased by 4.5%, 4.2%, and 3% in topic 2, topic 3, and topic 4, neutral sentiment increased by 2.7%, 1.7%, and 1.3% during those topics. In topic 2, there is also a high number of negative sentiments. Negative sentiments increase in the latter topic when compared to topic 1 and topic 3. Except for topic 2, the percentage of Neutral tweets almost remains the same throughout the year. As the COVID crisis piled up, there was a drop in positive sentiments and a significant increase in neutral sentiments as shown in Table 3.

B. TOPIC MODELING USING PROPOSED APPROACH

We extract topics from the COVID-19 papers that have been published using the LDA model in genism. The number of

TABLE 3. Positive and negative sentiment analysis.

Topics	Total Words	Neutral Words	Neutral W%	Positive Words	Positive W %	Negative Words	Negative W%
Topic 1	85,653	60, 086	20.5	16, 841	16	8, 856	17.8
Topic 2	85,653	50, 775	19.2	23, 703	23.9	13, 305	21.3
Topic 3	85,653	58, 202	20.9	18, 955	19.1	9, 626	17.4
Topic 4	85,653	55, 666	18.9	19, 006	19.1	12, 111	22.8
Topic 5	85,653	54, 720	20.5	20, 697	20.9	11, 366	20.7
Total	428,265	278,449	100	98,202	100	51,264	100

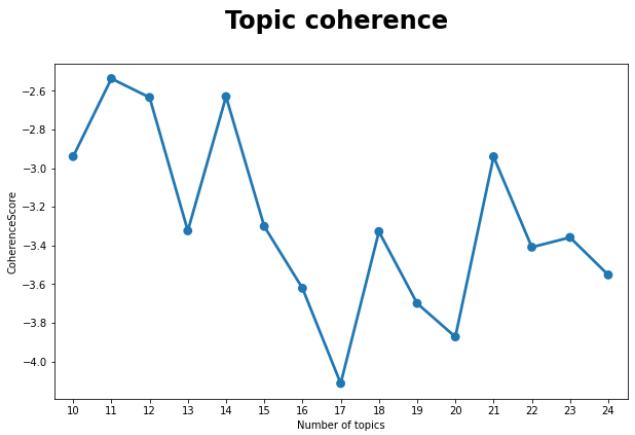


FIGURE 5. Coherence scores of the different number of topics.

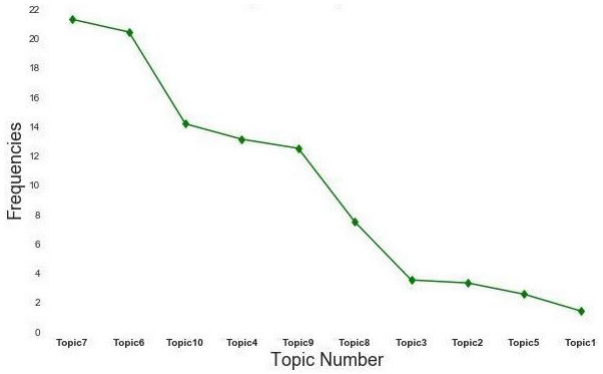


FIGURE 6. Analysis of the topic frequency of optimal topics.

LDA topics depends on topic coherence. Regarding topic coherence, the distributional hypothesis is that the start of words with similar meanings is grouped together [74]. According to this hypothesis, similar words are more likely to occur in similar situations. In the tmtoolkit package, we use the ''coherence gensim c npmi'' function to calculate the coherence value for each topic number k in the abstract collection. Fig 5 illustrates the coherence value between topics. Topic number 8 achieves the highest coherence value.

As a result, 8 is the ideal topic number. The results of the LDA model are interpreted to illuminate the significance of the subjects. Visualization of 10 topics can be found. The full-text collection's subject number is also set to 10 to maintain consistency with the numbering. In Equation 4, according to topic coherence, topic V coherence value is the sum of pairwise distributional similarity scores over topic words [75], [76], [77].

$$\text{Score}\left(w_i, w_j, \epsilon\right) = \left(\frac{\log \frac{P(w_i, w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log\left(P\left(w_i, w_j\right)+\epsilon\right)}\right)^{\gamma}$$
$$\text{Coherence}(V) = \sum_{(w_i, w_j)\epsilon V} \text{Score}\left(w_i, w_j, \epsilon\right) \quad (4)$$

Furthermore, Fig 6 shows the frequency of the optimal topics. Topic 8 has the highest frequency, 22.97 percent, compared to the other listed topics. The average frequency percentage of topic 9 is 21.80, topic 6 is 21.79, topic 10 is 13.26, and topic 1 is 13.16. Similarly, Topic 5 achieved the lowest frequency of 1.167 compared to all listed topics.

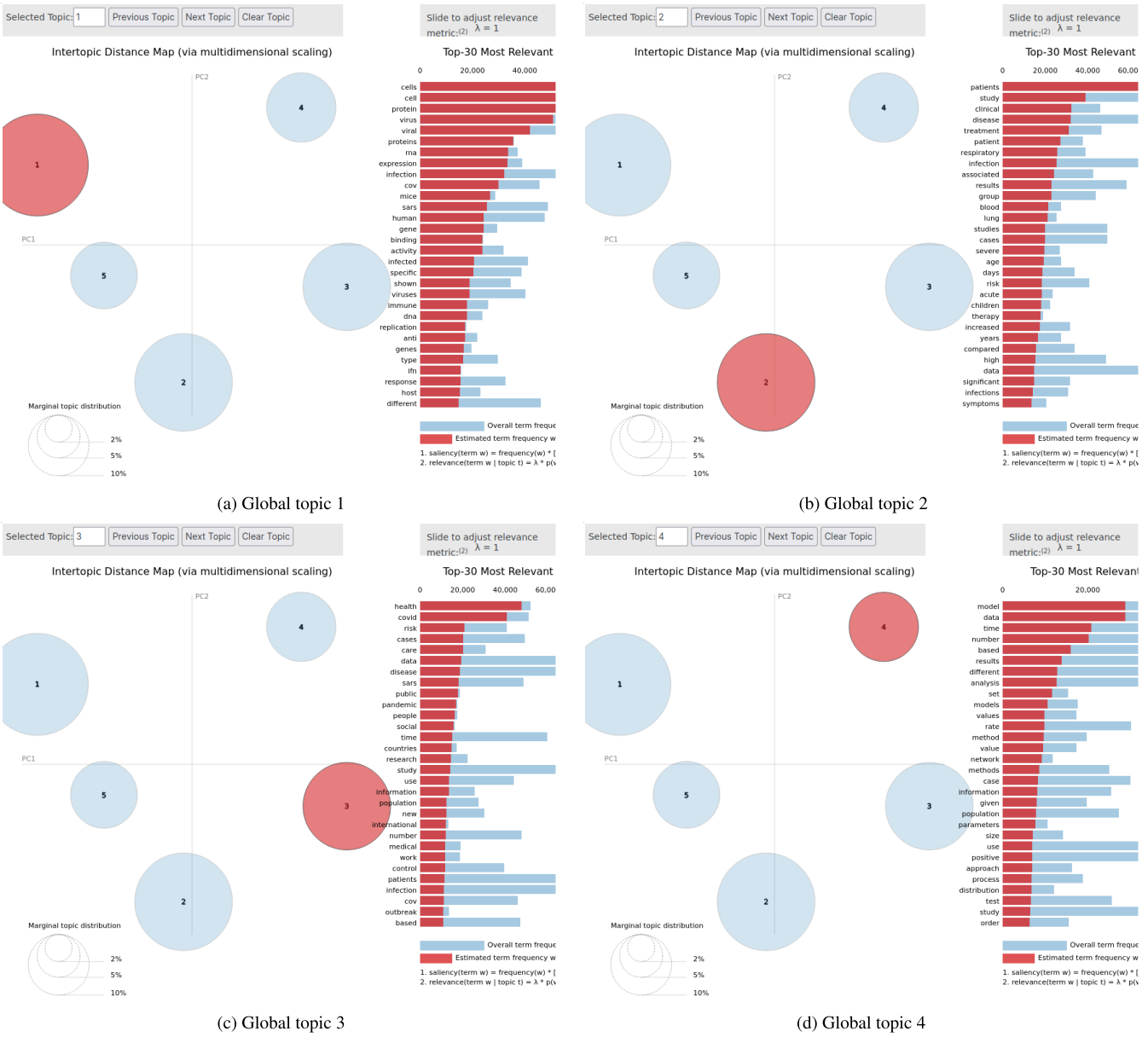
In addition, Fig. 7 illustrates an overview of the global topics and associated terms to investigate the topic-term relationship. The analysis provides two types of visualization to interpret a selected topic by determining useful terms. On the left side, it provides a visual view of the selected global topics on a two-dimensional plane. On the right side, a bar chart is given to illustrate the frequency of the terms associated with the selected global topic. The ranking frequency of the associated terms is given in decreasing order for topic interpretation. In this way, each global topic is analyzed compactly according to the frequency of the terms. The proposed analysis aims to facilitate health practitioners for interpreting the relationship between global topics and associated terms in a large corpus to mitigate the pandemic.

C. WORLDCOULD OF COVID-19 RESEARCH PAPERS

A wordcloud item is font size depending on the importance of the word in the data, here words obtained by analysis of the abstract's text belonging to the COVID-19 corpus. For the creation of the wordcloud, you must first perform various preprocessing (tokenization, lemmatization) of the text before creating the items of the wordcloud. Clearly, the words ''covid'', ''patient'', and ''Study'' do stand out from the rest of the abstract text analyzed. Other words such as ''Wuhan'' and ''protein'' have also been extensively shown in Fig. 8 shows WorlColud.

D. WORD2VEC MODEL AND TEXTUAL COSINE SIMILARITIES

Word2Vec models train on a corpus of text to see which words tend to be used in a similar context. We built the



(a) Global topic 1

(b) Global topic 2

(c) Global topic 3

(d) Global topic 4

FIGURE 7. **Overview of the global topics and associated terms for analyzing relationship between topic and terms.**

word embedding using the Python library Gensim for word embedding. In Word2Vec models, large corpora of text are used as inputs. As a result, each unique word in the corpus is represented by a vector. word2Vec model shown in Fig 9 Cosine similarity is used to measure vector similarity shown in Table 4 and Table 5 Cosine similarity in diagnostic, transmission to a similar word vector.

Similar vectors are used to represent semantically related words to the origin during the analysis of the original corpus. Coronaviruses are illnesses that can be passed from animals to humans. This type of transmission, known as zoonotic origin, occurs when a pathogen jumps from non-human animals to humans.

Similar to the words that have been used to describe COVID-19 symptoms, terms like "fever", "Patient", "Transmission," and "Vaccine."

E. T-SNE VISUALIZATION OF SEMANTIC CLUSTERS

This technique was finally used to reduce each word's dimensionality, allowing the 2D position to be projected along with its label. A ML algorithm such as K-mean was also implemented using Scikit-learn Python Library to partition n-words into semantic clusters. In the elbow method, K was optimally determined by summing the squared distances between clusters [1, 30]. If the plot looks like an arm, the elbow on the arm is the optimal K. Here, K = 7.

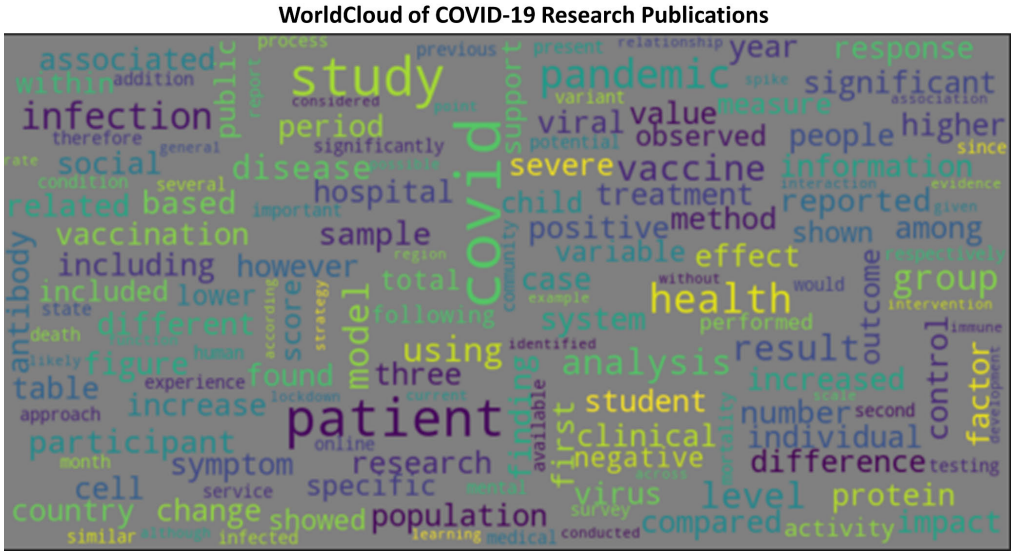


FIGURE 8. Worldcloud of COVID-19 research publications.

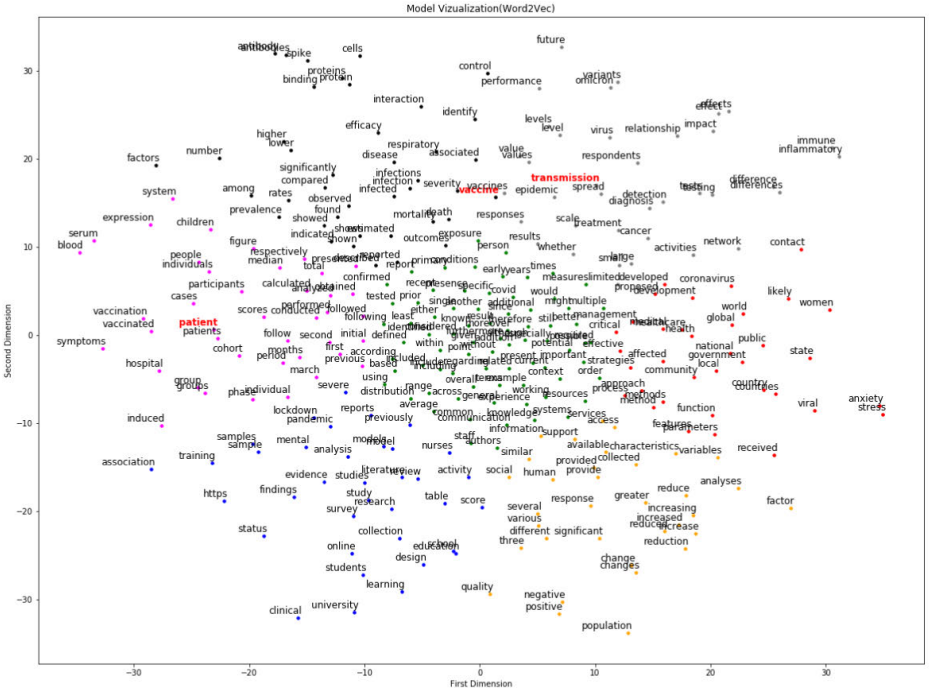


FIGURE 9. Word2vec model vizualization.

From the above TSNE visualization of Word2Vec embeddings, we can distinguish several clusters among which we can recognize semantic similarities, including medical treatment, government policies and measures, vaccine research, epidemiological research, and COVID-19 detection, transmission, causes, and consequences of the disease. Inter-word distance in the 2D plane is an indication of inter-word similarity.

F. WORD MOVER'S DISTANCE (WMD)

WMD is a tool for measuring the distance between a document and a word. The topic similarity between the topic and subtopics is where the most related results in word2vec embedding represent each topic. Analyzing the similarities between topics indicates a lower score with other correlated topics. In Table 5, the similarity metrics in Topic 3 and Topic 4, Dyspnea and fever also, and cough and myalgia in

TABLE 4. Cosine similarity in origin, symptom to similar word vector.

Cosine similarity in origin to similar word vector		Cosine similarity in symptom to similar word vector	
Origin	**Similar Words**	**Symptom**	**Similar Words**
Origins	0.7476460933685303	Symptoms	0.7272068858146667
Originated	0.7162984609603882	Dyspnoea	0.5813004970550537
Natural_reservoir	0.6514419317245483	Fever	0.5673059821128845
Zoonotic_origin	0.6216771006584167	Dyspnea	0.5655109882354736
Intermediate_hosts	0.6125218272209167	Cough	0.5579974055290222
Animal_markets	0.5840376615524292	Myalgia	0.5559396743774414
Horseshoe	0.5832125544548035	Fatigue	0.5482243299484253
Animal_reservoir	0.5780599117279053	Fever_cough	0.5202878713607788
Pangolins	0.5735911130905151	Illness	0.5187296867370605
Seafood_market	0.5658592581748962	Productive_cough	0.5070851445198059

TABLE 5. Cosine similarity in diagnostic, transmission to similar word vector.

Cosine similarity in diagnostic to similar word vector		Cosine similarity in transmission to similar word vector	
Diagnostic	**Similar Words**	**Transmission**	**Similar Words**
Diagnosing	0.6566814184188843	Diagnostics	0.6272414922714233
Transmissions	0.8198723793029785	Diagnosis	0.6242400407791138
Spread	0.6515756249427795	Testing	0.5751957893371582
Transmissibility	0.6484392881393433	Spreading	0.5507039427757263
Detecting	0.5729243755340576	Serological	0.5473078489303589
Transmitted	0.5454849600791931	Detection	0.5434328317642212
Superspreading_events	0.5300137996673584	Diagnose	0.5244864821434021
Transmitting	0.5024526715278625	Serological_assays	0.5178400874137878
Transmissible	0.4996853768825531	Molecular_diagnostics	0.5066964030265808

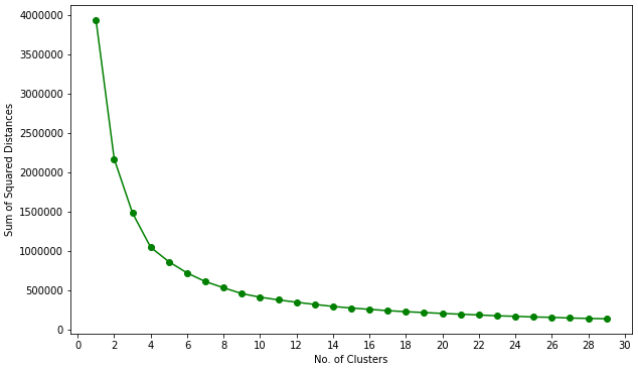


FIGURE 10. Optimal no of clusters K(Elbow method).

Topic 5 and 6 within Topic 7, the similarity metrics among its subtopic the words are notably diminished compared to those between the similar words.

G. DISCUSSION

1) NOVELTY AND SCOPE

This study stands out from existing research by focusing on the public sentiment analysis on COVID-19 through research publications. While previous studies by Ahammad [78] use a smaller dataset of 10,254 news headlines and combine sentiment analysis with topic modeling. Xie et al. [79] offer insights into public sentiment on Weibo during the COVID-19 outbreak. Gyftopoulos et al. [80] collected data from Twitter posts and analyzed public sentiments based on the content of the posts circulated during the COVID-19 period. By Thakur [81] focusing on COVID-19 and MPox simultaneously, revealing that nearly half of the tweets had a negative sentiment, followed by positive and neutral sentiments. Costola et al. [82] investigate the impact of COVID-19 news on financial markets, analyzing a large corpus of online articles from major news platforms and employing ML techniques for sentiment analysis. It also identifies top hashtags and most frequently used words, offering insights into prevalent topics in Twitter conversations. Our research takes a more comprehensive approach by encompassing a wide range of COVID-related analyses. Our approach is more extensive, encompassing a diverse range of COVID-related analyses. Leveraging data from the National Library of Medicine PubMed, we employed automated LDA to extract key themes from extensive discussions on the pandemic. This method aids researchers in identifying and categorizing various facets of the pandemic outbreak, ranging from medical issues to public responses, economic impacts, and policy. Additionally, sentiment analysis allows for insights into prevailing sentiments surrounding the pandemic and its implications for different communities by assessing the emotional tone of these discussions. Our study makes a significant contribution by examining research topics and their connections, employing LDA modeling and NLP (Natural Language Processing) to assess the current literature on COVID-19 and COV infection. Importantly, our research is academic and serves to aid in pandemic coordination efforts by identifying high-priority scientific topics. This is particularly crucial in areas such as pathogens, treatments, virus diagnostics, vaccines, and viral genomes, which are now deemed priorities alongside clinical characterization, epidemiology, and virus transmission research. Our study

TABLE 6. Comparison of the proposed study with similar approaches.

Ref	Dataset Size	Analysis Methods	Sentiment Trends	Sentiment Distribution	Findings
[78]	10,254 headlines	Sentiment Analysis	Not specified	Not specified	Specific misinformation themes identified
[79]	719,570 Weibo posts	Sentiment Analysis Text Mining	Not specified	Positive, negative neutral sentiments	Public responses on Weibo
[80]	Twitter posts	Sentiment Analysis	Not specified	Dynamics public response	Increased activity in COVID-19 channel, joy predominant
[81]	61,862 tweets	Sentiment Analysis	Not specified	Positive: 20.9%, Negative: 46.8%, Neutral: 21.1%	Identifies top 50 hashtags and top 100 most frequently used words
[82]	203,886 online articles	Sentiment extraction	Not specified	Not specified	The positive relationship between news sentiment and S&P 500 market
Proposed Study	32,314 publications	Sentiment Analysis	Positive decrease, Neutral increase over time	Positive: 22.1%, Negative: 12.3%, Neutral: 64.1%	Sentiment trends in research, primarily neutral

aims to provide a detailed analysis of public sentiment surrounding COVID-19. We use a new approach that involves research publications and advanced techniques such as LDA modeling and sentiment analysis. By building on existing studies, we aim to improve our understanding of the pandemic's impact and provide valuable insights into the scientific community's efforts in combating COVID-19. Table 6 presents a comprehensive comparison between our study and existing research using a similar approach, highlighting key differences and similarities in methodology.

2) LIMITATIONS AND FUTURE WORKS

It's important to note the limitations of this study. We have identified the issue to understand the medical treatment, governmental rules and regulations, vaccination research, epidemiological research, and the detection, transmission, causes, and effects of the illness COVID-19. However, our study is not exhaustive and there may be other aspects that could be explored in future research. Future research can consider other methods. The analysis exclusively refers to the COVID-19 pandemic literature. We may contend that our approach performs admirably on the sizable COVID-19 data set. Additionally, we concentrate on the literature analysis, which includes themes for study, research trends, and topic similarity networks. All of these are general information and not specific medical knowledge. Consequently, our study framework might be applied to various types of literature.

VII. CONCLUSION

The study aims to provide a framework for topic modeling to aid in analyzing the research themes and patterns surrounding the newly developing research topic. We compared subject modeling on full-text papers and matching abstracts using COVID-19 as a case to determine the impact of various document formats used as topic modeling input. With the help of the topic modeling approach, the presented work shows the common research themes, trends, and similarity networks for the COVID-19 study. This research makes several contributions. First, we summarize the COVID-19 Publications using topic modeling, including the most pertinent terminology, major research themes, and emerging trends. Many articles have been published about the virus's gene analysis. Government regulations and their effect have been discussed in particular articles. In the interim, the vaccination by the end of 2020, although not yet in the complete discussion. Furthermore, the proposed research not only adds to the technique by using literature analysis but also provides practical insights. The comparative study of topic extraction from full paper texts against their related abstracts can assist us in comprehending the impact of the various texts based on topic modeling analysis findings. This research shows that extracting ideas from abstracts could be more effective than full text because they might convey the same information with fewer words. Third, for librarians or documentalists to effectively manage the literature on a particular subject, the current study offers a practical methodological framework that may be used in any field. Understanding the results may be aided by our LDA-based topic modeling, word-cloud subject visualization, and essential terms' trends.

REFERENCES

[1] F. Hu, L. Qiu, X. Xi, H. Zhou, T. Hu, N. Su, H. Zhou, X. Li, S. Yang, Z. Duan, Z. Dong, Z. Wu, H. Zhou, M. Zeng, T. Wan, and S. Wei, "Has COVID-19 changed China's digital trade?—Implications for health economics," *Frontiers Public Health*, vol. 10, Mar. 2022, Art. no. 831549.

[2] F. Hu, Q. Ma, H. Hu, K. H. Zhou, and S. Wei, "A study of the spatial network structure of ethnic regions in Northwest China based on multiple factor flows in the context of COVID-19: Evidence from Ningxia," *Heliyon*, vol. 10, no. 2, Jan. 2024, Art. no. e24653.

[3] S. R. Weiss and J. L. Leibowitz, "Coronavirus pathogenesis," *Adv. Virus Res.*, vol. 81, pp. 85–164, Jan. 2011.

[4] A. Zumla, J. F. W. Chan, E. I. Azhar, D. S. C. Hui, and K.-Y. Yuen, "Coronaviruses—Drug discovery and therapeutic options," *Nature Rev. Drug Discovery*, vol. 15, no. 5, pp. 327–347, May 2016.

[5] J. Cui, F. Li, and Z.-L. Shi, "Origin and evolution of pathogenic coronaviruses," *Nature Rev. Microbiol.*, vol. 17, no. 3, pp. 181–192, Mar. 2019.

[6] V. C. C. Cheng, S. K. P. Lau, P. C. Y. Woo, and K. Y. Yuen, "Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection," *Clin. Microbiol. Rev.*, vol. 20, no. 4, pp. 660–694, Oct. 2007.

[7] J. F. W. Chan, S. K. P. Lau, K. K. W. To, V. C. C. Cheng, P. C. Y. Woo, and K.-Y. Yuen, "Middle east respiratory syndrome coronavirus: Another zoonotic betacoronavirus causing SARS-like disease," *Clin. Microbiol. Rev.*, vol. 28, no. 2, pp. 465–522, Apr. 2015.

[8] L. E. Gralinski and R. S. Baric, "Molecular pathology of emerging coronavirus infections," *J. Pathol.*, vol. 235, no. 2, pp. 185–195, Jan. 2015.

[9] (2001). *World Health Organization—COVID-19*. Accessed: Aug. 2022. [Online]. Available: https://covid19.who.int/

[10] S. Ahamed and M. Samad, "Information mining for COVID-19 research from a large volume of scientific literature," 2020, *arXiv:2004.02085*.

[11] F. Hu, L. Qiu, and H. Zhou, "Medical device product innovation choices in Asia: An empirical analysis based on product space," *Frontiers Public Health*, vol. 10, Apr. 2022, Art. no. 871575.

[12] S. Malla and P. J. A. Alphonse, "COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets," *Appl. Soft Comput.*, vol. 107, Aug. 2021, Art. no. 107495.

[13] Q. Chen, R. Leaman, A. Allot, L. Luo, C.-H. Wei, S. Yan, and Z. Lu, "Artificial intelligence in action: Addressing the COVID-19 pandemic with natural language processing," *Annu. Rev. Biomed. Data Sci.*, vol. 4, no. 1, pp. 313–339, Jul. 2021.

[14] S. Praveen and R. Ittamalla, "An analysis of attitude of general public toward COVID-19 crises—Sentimental analysis and a topic modeling study," *Inf. Discovery Del.*, vol. 49, no. 3, pp. 240–249, Sep. 2021.

[15] T. Tayir and L. Li, "Unsupervised multimodal machine translation for low-resource distant language pairs," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 4, pp. 1–22, Apr. 2024.

[16] K. A. R. Issam, S. Patel, and C. N. Subalalitha, "Topic modeling based extractive text summarization," 2021, *arXiv:2106.15313*.

[17] W. Dang, L. Cai, M. Liu, X. Li, Z. Yin, X. Liu, L. Yin, and W. Zheng, "Increasing text filtering accuracy with improved LSTM," *Comput. Informat.*, vol. 42, no. 6, pp. 1491–1517, 2023.

[18] S. Pan, G. J. Xu, K. Guo, S. H. Park, and H. Ding, "Cultural insights in souls-like games: Analyzing player behaviors," *IEEE Trans. Games*, 2024.

[19] R. Sandhiya, A. Boopika, M. Akshatha, S. Swetha, and N. Hariharan, "A review of topic modeling and its application," in *Handbook of Intelligent Computing and Optimization for Sustainable Development*, 2022, pp. 305–322.

[20] C. Huang, Z. Han, M. Li, X. Wang, and W. Zhao, "Sentiment evolution with interaction levels in blended learning environments: Using learning analytics and epistemic network analysis," *Australas. J. Educ. Technol.*, vol. 37, no. 2, pp. 81–95, May 2021.

[21] Y. Ban, Y. Liu, Z. Yin, X. Liu, M. Liu, L. Yin, X. Li, and W. Zheng, "Micro-directional propagation method based on user clustering," *Comput. Informat.*, vol. 42, no. 6, pp. 1445–1470, 2023.

[22] H. Chen, C. Lv, L. Ding, H. Qin, X. Zhou, Y. Ding, X. Liu, M. Zhang, J. Guo, X. Liu, and D. Tao, "DB-LLM: Accurate dual-binarization for efficient LLMs," 2024, *arXiv:2402.11960*.

[23] S. Pan, G. J. W. Xu, K. Guo, S. H. Park, and H. Ding, "Video-based engagement estimation of game streamers: An interpretable multimodal neural network approach," *IEEE Trans. Games*, 2023.

[24] H. Qin, M. Zhang, Y. Ding, A. Li, Z. Cai, Z. Liu, F. Yu, and X. Liu, "BiBench: Benchmarking and analyzing network binarization," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28351–28388.

[25] W. Huang, Y. Liu, H. Qin, Y. Li, S. Zhang, X. Liu, M. Magno, and X. Qi, "BiLLM: Pushing the limit of post-training quantization for LLMs," 2024, *arXiv:2402.04291*.

[26] H. Qin, Y. Ding, M. Zhang, Q. Yan, A. Liu, Q. Dang, Z. Liu, and X. Liu, "BiBERT: Accurate fully binarized BERT," 2022, *arXiv:2203.06390*.

[27] P. S. A. Babu, C. S. Rao Annavarapu, and A. Mohapatra, "A novel method for next-generation sequence data analysis using PLSA topic modeling technique," in *Proc. 2nd Int. Conf. Adv. Comput. Commun. Paradigms (ICACCP)*, Feb. 2019, pp. 1–6.

[28] T. Zhou, Z. Cai, F. Liu, and J. Su, "In pursuit of beauty: Aesthetic-aware and context-adaptive photo selection in crowdsensing," *IEEE Trans. Knowl. Data Eng.*, 2023.

[29] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, Dec. 2020, Art. no. 101582.

[30] Y. Bai, S. Jia, and L. Chen, "Topic evolution analysis of COVID-19 news articles," *J. Phys., Conf. Ser.*, vol. 1601, no. 5, Aug. 2020, Art. no. 052009.

[31] Q. Liu, Z. Zheng, J. Zheng, Q. Chen, G. Liu, S. Chen, B. Chu, H. Zhu, B. Akinwunmi, J. Huang, C. J. P. Zhang, and W.-K. Ming, "Health communication through news media during the early stage of the COVID-19 outbreak in China: Digital topic modeling approach," *J. Med. Internet Res.*, vol. 22, no. 4, Apr. 2020, Art. no. e19118.

[32] E. De Santis, A. Martino, and A. Rizzi, "An infoveillance system for detecting and tracking relevant topics from Italian tweets during the COVID-19 event," *IEEE Access*, vol. 8, pp. 132527–132538, 2020.

[33] S. Noor, Y. Guo, S. H. H. Shah, P. Fournier-Viger, and M. S. Nawaz, "Analysis of public reactions to the novel coronavirus (COVID-19) outbreak on Twitter," *Kybernetes*, vol. 50, no. 5, pp. 1633–1653, May 2021.

[34] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 4, pp. 1003–1015, Aug. 2021.

[35] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107057.

[36] D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein, and A. Nabil, "CoAID-DEEP: An optimized intelligent framework for automated detecting COVID-19 misleading information on Twitter," *IEEE Access*, vol. 9, pp. 27840–27867, 2021.

[37] D. Konar, B. K. Panigrahi, S. Bhattacharyya, N. Dey, and R. Jiang, "Auto-diagnosis of COVID-19 using lung CT images with semi-supervised shallow learning network," *IEEE Access*, vol. 9, pp. 28716–28728, 2021.

[38] L. L. Wang and K. Lo, "Text mining approaches for dealing with the rapidly expanding literature on COVID-19," *Briefings Bioinf.*, vol. 22, no. 2, pp. 781–799, Mar. 2021.

[39] S. Madichetty and M. Sridev, "A novel method for identifying the damage assessment tweets during disaster," *Future Gener. Comput. Syst.*, vol. 116, pp. 440–454, Mar. 2021.

[40] S. Madichetty, S. Muthukumarasamy, and P. Jayadev, "Multi-modal classification of Twitter data during disasters for humanitarian response," *J. Ambient Intell. Hum. Comput.*, vol. 12, no. 11, pp. 10223–10237, Nov. 2021.

[41] X. Li and Y. Sun, "Application of RBF neural network optimal segmentation algorithm in credit rating," *Neural Comput. Appl.*, vol. 33, no. 14, pp. 8227–8235, Jul. 2021.

[42] 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, Dec. 2020, Art. no. 106754.

[43] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2733–2742, Oct. 2020.

[44] L. Carnevale, A. Celesti, G. Fiumara, A. Galletta, and M. Villari, "Investigating classification supervised learning approaches for the identification of critical patients' posts in a healthcare social network," *Appl. Soft Comput.*, vol. 90, May 2020, Art. no. 106155.

[45] H. Qi, Z. Zhou, J. Irizarry, D. Lin, H. Zhang, N. Li, and J. Cui, "Automatic identification of causal factors from fall-related accident investigation reports using machine learning and ensemble learning approaches," *J. Manage. Eng.*, vol. 40, no. 1, Jan. 2024, Art. no. 04023050.

[46] P. Kairon and S. Bhattacharyya, "COVID-19 outbreak prediction using quantum neural networks," in *Intelligence Enabled Research*. Springer, 2021, pp. 113–123.

[47] Y. Zhang, H. Lyu, Y. Liu, X. Zhang, Y. Wang, and J. Luo, "Monitoring depression trend on Twitter during the COVID-19 pandemic," 2020, *arXiv:2007.00228*.

[48] K. Chatsiou, "Text classification of COVID-19 press briefings using BERT and convolutional neural networks," Tech. Rep., 2020.

[49] X. Li, M. Zhou, J. Wu, A. Yuan, F. Wu, and J. Li, "Analyzing COVID-19 on online social media: Trends, sentiments and emotions," 2020, *arXiv:2005.14464*.

[50] J. Zhou, H. Zogan, S. Yang, S. Jameel, G. Xu, and F. Chen, "Detecting community depression dynamics due to COVID-19 pandemic in Australia," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 4, pp. 982–991, Aug. 2021.

[51] H. Yin, S. Yang, and J. Li, "Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media," in *Proc. Int. Conf. Adv. Data Mining Appl.* Springer, 2020, pp. 610–623.

[52] J. Li, C. Huang, Y. Yang, J. Liu, X. Lin, and J. Pan, "How nursing students' risk perception affected their professional commitment during the COVID-19 pandemic: The mediating effects of negative emotions and moderating effects of psychological capital," *Humanities Social Sci. Commun.*, vol. 10, no. 1, pp. 1–9, May 2023.

[53] A. E. Aiello, A. Renson, and P. N. Zivich, "Social media- and internet-based disease surveillance for public health," *Annu. Rev. Public Health*, vol. 41, no. 1, pp. 101–118, Apr. 2020.

[54] S. A. Waheeb, N. A. Khan, B. Chen, and X. Shang, "Machine learning based sentiment text classification for evaluating treatment quality of discharge summary," *Information*, vol. 11, no. 5, p. 281, May 2020.

[55] C. Bao, X. Hu, D. Zhang, Z. Lv, and J. Chen, "Predicting moral elevation conveyed in Danmaku comments using EEGs," *Cyborg Bionic Syst.*, vol. 4, p. 28, Jan. 2023.

[56] X. Si, H. He, J. Yu, and D. Ming, "Cross-subject emotion recognition brain–computer interface based on fNIRS and DBJNet," *Cyborg Bionic Syst.*, vol. 4, p. 45, Jan. 2023.

[57] Q. Chen and M. Sokolova, "Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in analysis of scientific and medical texts," *Social Netw. Comput. Sci.*, vol. 2, no. 5, pp. 1–11, 2021.

[58] H. Zhang, H. Liu, and C. Kim, "Semantic and instance segmentation in coastal urban spatial perception: A multi-task learning framework with an attention mechanism," *Sustainability*, vol. 16, no. 2, p. 833, Jan. 2024.

[59] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, Jun. 2020.

[60] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets," *IEEE Access*, vol. 8, pp. 181074–181090, 2020.

[61] S. Siddiqui, M. S. Faisal, S. Khurram, A. Irshad, M. Baz, H. Hamam, N. Iqbal, and M. Shafiq, "Quality prediction of wearable apps in the Google play store," *Intell. Autom. Soft Comput.*, vol. 32, no. 2, pp. 877–892, 2022.

[62] S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study," *JMIR Public Health Surveill.*, vol. 6, no. 4, Nov. 2020, Art. no. e21978.

[63] S. Das and A. Dutta, "Characterizing public emotions and sentiments in COVID-19 environment: A case study of India," *J. Hum. Behav. Social Environ.*, vol. 31, nos. 1–4, pp. 154–167, May 2021.

[64] G. Barkur, Vibha, and G. B. Kamath, "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India," *Asian J. Psychiatry*, vol. 51, Jun. 2020, Art. no. 102089.

[65] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, "Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study," *J. Med. Internet Res.*, vol. 22, no. 10, Oct. 2020, Art. no. e22624.

[66] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic: Using latent Dirichlet allocation for topic modeling on Twitter," *PLoS ONE*, vol. 15, no. 9, Sep. 2020, Art. no. e0239441.

[67] D. Yang, T. Zhu, S. Wang, S. Wang, and Z. Xiong, "LFRSNet: A robust light field semantic segmentation network combining contextual and geometric features," *Frontiers Environ. Sci.*, vol. 10, Oct. 2022, Art. no. 996513.

[68] Y. Xu, E. Wang, Y. Yang, and Y. Chang, "A unified collaborative representation learning for neural-network based recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5126–5139, Nov. 2022.

[69] D. Li, "An interactive teaching evaluation system for preschool education in universities based on machine learning algorithm," *Comput. Hum. Behav.*, vol. 157, Aug. 2024, Art. no. 108211.

[70] F. Huang, Z. Wang, X. Huang, Y. Qian, Z. Li, and H. Chen, "Aligning distillation for cold-start item recommendation," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2023, pp. 1147–1157.

[71] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[72] H. Yin, X. Song, S. Yang, and J. Li, "Sentiment analysis and topic modeling for COVID-19 vaccine discussions," *World Wide Web*, vol. 25, no. 3, pp. 1067–1083, May 2022.

[73] D. M. Mimno, H. M. Wallach, E. M. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 262–272.

[74] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2017, pp. 165–174.

[75] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proc. 10th Int. Conf. Comput. Semantics (IWCS)*, 2013, pp. 13–22.

[76] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, Feb. 2015, pp. 399–408.

[77] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 952–961.

[78] T. Ahammad, "Identifying hidden patterns of fake COVID-19 news: An in-depth sentiment analysis and topic modeling approach," *Natural Lang. Process. J.*, vol. 6, Mar. 2024, Art. no. 100053.

[79] R. Xie, S. K. W. Chu, D. K. W. Chiu, and Y. Wang, "Exploring public response to COVID-19 on Weibo with LDA topic modeling and sentiment analysis," *Data Inf. Manage.*, vol. 5, no. 1, pp. 86–99, Jan. 2021.

[80] S. Gyftopoulos, G. Drosatos, G. Fico, L. Pecchia, and E. Kaldoudi, "Analysis of pharmaceutical Companies' social media activity during the COVID-19 pandemic and its impact on the public," *Behav. Sci.*, vol. 14, no. 2, p. 128, Feb. 2024.

[81] N. Thakur, "Sentiment analysis and text analysis of the public discourse on Twitter about COVID-19 and MPox," *Big Data Cognit. Comput.*, vol. 7, no. 2, p. 116, Jun. 2023.

[82] M. Costola, O. Hinz, M. Nofer, and L. Pelizzon, "Machine learning sentiment analysis, COVID-19 news and stock market reactions," *Res. Int. Bus. Finance*, vol. 64, Jan. 2023, Art. no. 101881.

AMREEN BATOOL received the bachelor's degree from GC University, Pakistan, the M.C.S. degree from Virtual University of Pakistan, and the M.S. degree in computer science and technology from Tiangong University, Tianjin, China, in 2021. She is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Jeju National University, Republic of Korea. She is a Project Coordinator with EUT Global Ltd. Her main role is to coordinate with clients and field engineers to plan project delivery. Her research interests include machine learning, deep learning, and blockchain technology.

YUNG-CHEOL BYUN received the B.S. degree from Jeju National University, in 1993, and the M.S. and Ph.D. degrees from Yonsei University, in 1995 and 2001, respectively. He was a Special Lecturer with SAMSUNG Electronics, in 2000 and 2001. From 2001 to 2003, he was a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI). He was promoted to join Jeju National University as an Assistant Professor, in 2003. He is currently an Associate Professor with the Computer Engineering Department, Jeju National University. His research interests include the areas of AI machine learning, pattern recognition, blockchain and deep learning-based applications, big data and knowledge discovery, time series data analysis and prediction, image processing and medical applications, and recommendation systems.