Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Comprehensive analytics of COVID-19 vaccine research: From topic modeling to topic classification

Saeed Rouhani *, Fatemeh Mozaffari

Department of IT Management, College of Management, University of Tehran, Tehran, Iran

ARTICLE INFO

Keywords:
COVID-19
Vaccine research
Topic modeling
Sentiment analysis
Classification

ABSTRACT

COVID-19 vaccine research has played a vital role in successfully controlling the pandemic, and the research surrounding the coronavirus vaccine is ever-evolving and accruing. These enormous efforts in knowledge production necessitate a structured analysis as secondary research to extract useful insights. In this study, comprehensive analytics was performed to extract these insights, which has moved the boundaries of data analytics in secondary research in the vaccine field by utilizing topic modeling, sentiment analysis, and topic classification based on the abstracts of related publications indexed in Scopus and PubMed. By applying topic modeling to 4803 abstracts filtered by this study criterion, 8 research arenas were identified by merging related topics. The extracted research areas were entitled "Reporting," "Acceptance," "Reaction," "Surveyed Opinions," "Pregnancy," "Titer of Variants," "Categorized Surveys," and "International Approaches." Moreover, the investigation of topics sentiments variations over time led to identifying researchers' attitudes and focus in various years from 2020 to 2022. Finally, a CNN-LSTM classification model was developed to predict the dominant topics and sentiments of new documents based on the 25 pre-determined topics with 75 % accuracy. The findings of this study can be utilized for future research design in this area by quickly grasping the structure of the current research on the COVID-19 vaccine. Through the findings of current research, a classification model was developed to classify the topic of a new article as one of the identified topics. Also, vaccine manufacturing firms will achieve a niche market by having a schema to invest in the gap of fields that have yet to be concentrated in extracted topics.

1. Introduction

The outbreak of COVID-19 (Corona Virus Disease 2019) and its global spread have negatively affected the economic, political, and social aspects of countries worldwide [1]. It affected financial markets and the global economy, as well as communities, businesses, and organizations. For example, stock market indexes fell dramatically in many countries before governmental support. It also caused importation issues and staffing deficiencies as the key concerns for businesses because of disruption to supply chains and self-isolation policies. The pandemic had widespread socio-economic implications, such as social isolation and school dropout due to the nationwide closure of educational facilities imposed in over 100 countries. The travel industry, as well as the tourism sector, is one of the hardest hit by the outbreak of the Corona Virus, with an impact on both travel supply and demand [2]. The outbreak also had a detrimental effect on healthcare systems worldwide and faced them with unprecedented challenges, such as risk to healthcare workers and shortages of protective equipment [2].

Therefore, it has made researchers in various fields respond immediately to this pandemic due to advances in science and technology [1]. It created opportunities for companies involved in vaccine and medicine development in such a way that they announced collaborative plans to develop a viral vaccine [2]. Due to the great achievements that vaccines have had in controlling infectious diseases, the media and public placed hope at large on having a vaccine that protects against coronavirus as soon as possible [3]. The successful development of the COVID-19 vaccine gave the world a sense of optimism at the end of this crisis despite many challenges, from guaranteeing its safety and efficacy to the rising vaccine hesitancy in high-income nations [4]. Hence, many articles and research about this subject have been published, bringing various findings. The publication of these articles in scientific communities enables the researchers to quickly understand the development process of the pandemic [1], and the first articles in this field after the start of the pandemic, based on a search through PubMed, were released

* Corresponding author.
E-mail addresses: SRouhani@ut.ac.ir (S. Rouhani), famozaffari@ut.ac.ir (F. Mozaffari).

https://doi.org/10.1016/j.artmed.2024.102980
Received 19 February 2023; Received in revised form 10 June 2024; Accepted 9 September 2024
Available online 18 September 2024

as soon as March 2020 [5,6]. After the start of vaccination in many countries, the focus of research has shifted to the rapid and effective rollout of the vaccine, as well as public sentiment toward it and its effects [7].

In this regard, extensive efforts have been made to produce knowledge in this field, and an enormous amount of research and various studies have been presented in a short period of time [8]. Techniques would be needed to convert these findings into applicable insights, extract knowledge, and summarize the results and research trends through a comprehensive analysis of research topics and sentiment analysis of them. Text analytics techniques such as Topic Modeling and Sentiment Analysis, as a Natural Language Processing (NLP) technique, are among the most popular methods that researchers use to study themes, sentiments, viewpoints, etc., applying Machine Learning (ML) algorithms [9]. By utilizing these methods, the main topics and sentiments related to them can be automatically extracted from a large number of articles so that trends and important themes in the intended field can be achieved.

Several studies were conducted to extract the topics related to the COVID-19 vaccine and public sentiments toward it on social media, such as Twitter [10,11] and Reddit social media platforms [12]. For example, Marcec and Likic [13] used sentiment analysis on the Twitter social network to compare public attitudes toward different vaccines. Luo et al. [14] employed text-mining techniques to examine the differences between professionals' and laypeople's opinions in forum discussions by applying topic modeling to their comments. Abd-Alrazaq et al. [15] conducted a study to identify people's top concerns by extracting their tweets' topics during the pandemic. Ogbuokiri et al. [16] used sentiment analysis on geotagged Twitter posts to find out about vaccine hesitancy hotspots.

Despite extensive studies conducted to extract topics related to the COVID-19 vaccine in social media and sentiment analysis of public opinions, a comprehensive review of articles related to this subject and obtaining topics that express the concerns and findings of the scientific and academic communities and research databases, and also sentiment analysis of them have not been performed yet. It can be highlighted from the mentioned gap that there is a need for data analytics in this field, and due to a large number of published articles, the use of text mining in this research has been considered. Accordingly, this research is based on secondary analysis, which includes reusing existing data from previous studies and utilizing them to gain new insights about the intended subject. Secondary information includes sources of data collected by others that can be used to quickly and inexpensively answer many questions. In other words, secondary analysis is the analysis of most of the information that has already been obtained and can be in line with the purpose of collecting primary data or with completely different purposes [17].

In the current study, latent topics from a large volume of data are discovered by using Latent Dirichlet Allocation (LDA), a probabilistic topic modeling introduced by Blei et al. [18]. Therefore, this research has been designed to answer the following research questions regarding the COVID-19 vaccine:

RQ1: What are the main research topics in the “COVID-19 vaccine” field?

RQ2: What are the sentiments of the topics, and how have they been changed from 2020 to 2022?

RQ3: How can a new document about the COVID-19 vaccine be classified on provided topics and measure its sentiment?

So, in this study, the above questions will be answered using topic modeling, and then the topics and research areas are entitled. Besides, the sentiment of each topic and its variations through time are analyzed by a lexicon-based approach. Moreover, for new documents as input to this model, their topic and related sentiment can be specified.

By introducing a novel approach, the scientific contribution of current research includes extracting the main research topics in the “COVID-19 vaccine” field and presenting a roadmap for these research trends. Also, the current paper bounces the sentiment of each topic, which can direct future studies, and upon the provided model, classifying a new document to the most dominant topic would happen. The integration of topic modeling, sentiment analysis and topic classification for performing new Systematic Literature Review (SLR) approach in the healthcare and vaccine studies would be the novelty of this research. Moreover, from the practical perspective, budget-balancing of the COVID-19 vaccine can be benchmarked and adopted through extracted topics. Vaccine manufacturing companies will have a schema to invest in the gap of fields that have yet to be concentrated in extracted topics, so they will achieve a niche market to invest in.

The paper is organized as follows: Section 2 illustrates the background, and Section 3 describes the research method. This section describes the identification of search strategy and configuring filtering criteria, data exporting and storing, data preprocessing, LDA model configuration, sentiment analysis, classification model development, and evaluation. In the following, Section 4 reports findings for research questions. As a consequence, Section 5 discusses the contributions of the research. Finally, Section 6 summarizes the findings and implications as a conclusion.

2. Background

2.1. COVID-19 vaccine

Coronaviruses (CoVs) are generally regarded as nonfatal human pathogens, mostly causing the common cold, while two human pathogenic CoVs, namely Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle-East Respiratory Syndrome Coronavirus (MERS-CoV), have caused epidemics during the last 18 years [19]. The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), which is responsible for the Coronavirus Disease 2019 (COVID-19) pandemic that emerged in December 2019, is caused by a new CoV named SARS-CoV-2 due to the relevance of COVID-19 symptoms with those of Severe Acute Respiratory Distress (SARD) and the similarity of COVID-19 with the previous human CoV disease, SARS [20]. Various treatments, such as immunosuppressants, steroids, and antiviral drugs, have been used before for handling MERS-CoV and SARS-CoV infections or other viral diseases [20]. However, an effective vaccine was urgently needed to protect humans against COVID-19 and reduce the pandemic's economic and societal impacts [19]. Vaccine development is typically a lengthy, expensive process, and it takes multiple candidates and many years to produce a licensed vaccine. So, due to the cost and high failure rates, developers typically follow a linear sequence of stages, with multiple breaks for data analysis or manufacturing process checks [21]. For example, the US Food and Drug Administration (FDA) only approved the first vaccine against Ebola 43 years after the fatal virus was discovered [22].

Nevertheless, with the COVID-19 crisis, everyone hoped this time would be different. Rapidly developing a vaccine requires a novel pandemic paradigm, with a fast start and many steps performed in parallel before approving a successful outcome of another step, resulting in raised financial risk [21]. However, the catastrophic effects of COVID-19 catalyzed the unprecedented development of vaccines and vaccine technologies in the struggle against this pandemic [8]. The need to quickly develop a vaccine against SARS-CoV-2 comes during the eruption in basic scientific understanding in areas such as genomics and structural biology, supporting a new epoch in vaccine development [21]. In fact, a number of the most advanced vaccine candidates utilize emerging technology platforms [22]. Therefore, in less than six months, several COVID-19 vaccine candidates entered into clinical trials and were conditionally approved ten months after the start of the COVID-19 outbreak, which is a record-breaking speed in vaccine development history. This unprecedented speed was enabled by the availability of pioneering vaccine technologies, the timely issue of the viral genomic sequence, active cooperation among the global scientific community, sufficient funding from various sources, and the massive/urgent market demand [8].

Therefore, the free availability of basic science data has allowed the creation of vaccines based on state-of-the-art platforms [3]. It has been revealed that taking advantage of genetic engineering technology to design effective vaccines against emerging and re-emerging viral diseases with pandemic potential would be vital for controlling the COVID-19 pandemic and future pandemics. Generally, development strategies for antiviral vaccines can be divided into three main groups (i) the first-generation vaccines encompass live-attenuated and inactivated vaccines, (ii) the second-generation vaccines include vaccine platforms such as protein subunit and vector-based vaccines, and (iii) the third-generation vaccines with nucleic acid and nanomaterial-based vaccines [23]. Based on these strategies, COVID-19 vaccine platforms can be categorized as Live-attenuated virus vaccines, Inactivated virus vaccines, Protein subunit vaccines, Replication-deficient vectors, Genetic vaccines (Deoxyribonucleic Acid (DNA), and Ribonucleic Acid (RNA)) [19], and Virus-Like Particles (VLP) vaccines which represent an evolution of protein subunit vaccinology and may also be regarded as a specific category of protein subunit vaccines [8]. Among them, RNA vaccines have shown very hopeful results with considerable success against COVID-19 with high protection percentages that US FDA EUA approved vaccine candidates from Pfizer and Moderna. Moderna's mRNA-1273, which entered into clinical trials just 66 days after SARS-CoV-2 was first sequenced, indicates the potential for nucleotide-based vaccines [22]. Also, the creative and technological efforts that led to the development of COVID-19 vaccines have transformed the approach and way of designing new vaccines for other diseases [3].

After many COVID-19 candidate vaccines were tested and granted emergency use authorization, vaccine effectiveness became an important issue studied in numerous research. Phase III trials reported high vaccine effectiveness (VE) against SARS-CoV-2 infection with these vaccines, such as 70.4 % effectiveness of AZD1222 (Oxford-AstraZeneca), 95 % effectiveness of the BNT162b2 mRNA COVID-19 vaccine (Pfizer-BioNTech), 94.1 % effectiveness of the mRNA-1273 vaccine (Moderna), and 50.7 % effectiveness of an absorbed COVID-19 (inactivated) vaccine (CoronaVac) [24]. Since studies in real-world settings around the world showed that the approved vaccines are highly protective against SARS-CoV-2, the full vaccination according to the standard schedule to achieve maximum VE was considered a priority by many countries. While a number of vaccines use traditional approaches, several innovative technologies, such as mRNA vaccines and non-replicating adenovirus vaccines, have rapidly mounted to a prominent position by leading the race for mass production and distribution [8].

2.2. COVID-19 vaccine research

Through analyzing past research, it can be realized that an enormous amount of research in the field of COVID-19 vaccine has been presented in a short period of time. An analysis of the CAS content collection was performed by Li et al. [8] at the end of February 2021 to assess COVID-19 vaccine-related research. Reviewing over 4000 published journal articles related to COVID-19 vaccine development showed that the United States, China, the UK, India, and Italy are the top five countries, accounting for over 50 % of the total publications, and the University of California, University of Oxford, and the National Institutes of Health (USA) have published the highest number of documents on COVID-19 vaccine-related studies. Among the published journal articles, about 15 % have been devoted to the investigation of various vaccine platforms, design, and formulation, as publications about the protein subunit vaccine platform account for the largest number of articles, and studies related to mRNA vaccines represent the second largest group. There have also been remarkable efforts to discover the parameters of immunity/efficacy and epitope/mutations. Correlations between COVID-19 severity/morbidity and the status of previous vaccinations, as well as cross-protection by other vaccines, have also been explored in a substantial number of published articles. After the beginning of the vaccination process, a significant portion of the articles (~18 %) addressed vaccination policies, such as the vaccine administration program and strategy and its social and psychological outlooks. Mutations in the context of vaccine development and the effectiveness of approved COVID-19 vaccines against these mutations are other hot topics that have been studied in many publications. For example, studies have shown that the variant that emerged in early 2020 with the D614G mutation was detectable by the antibodies provoked by the mRNA-1273 vaccine developed by Moderna [8].

In the rapidly changing COVID-19 research environment, new studies continue to be conducted and published. Safety monitoring of additional COVID-19 vaccine doses and further studies on the side effects of vaccines in overall populations, as well as in immunocompromised (IC) populations, is ongoing worldwide [25]. Duration of protection, optimal time intervals between primary series and additional/booster vaccine doses, the effectiveness and safety of additional/booster vaccine doses, and prevention of the advent of highly mutated novel SARS-CoV-2 variants are some of the concerns among others in this regard [25]. Many case report articles consider emerging concerns about the side effects of vaccines and emphasize the importance of clinical vigilance. For example, cross-reactivity with human tissue may contribute to the development of vaccine-associated immune-mediated diseases (IMDs) such as vaccine-induced immune thrombotic thrombocytopenia, autoimmune liver disease, IgA nephropathy, rheumatoid arthritis, and systemic lupus erythematosus (SLE) has been examined by Saleh et al. [26]. Multiple cases of COVID-19 vaccine-induced vasculitis [27], COVID-19 mRNA vaccine-related interstitial lung disease (ILD) [28], several cases of thrombotic thrombocytopenic purpura (TTP) [29] have been reported following COVID-19 vaccination, and they have allocated a part of the literature in this field to themselves.

Moreover, since these modern vaccines have had a short documentation history and might prompt hypothetical side effects after a long time, along with the progressive development of COVID-19 vaccines, opinion movements against vaccination have also thrived [3]. Vaccine hesitancy, which is the term used to describe “delay in acceptance or refusal of vaccination despite the availability of vaccination services” [30], became a major hindrance to the approved and prospective COVID-19 vaccination though the vaccine acceptance among the general public and healthcare workers appears to have a crucial role in the successful control of the pandemic [31]. Therefore, many articles examine vaccine hesitancy and the factors influencing the attitude toward the acceptance of vaccination. Evaluating opinions and acceptance rates toward COVID-19 vaccines can help start communication campaigns to strengthen trust in health authorities [31]. So, despite many challenges and unanswered questions, such as uncertainty regarding its long-term efficacy, the remarkable advances in COVID-19 vaccine development have offered the world hope that this pandemic can be defeated [8].

2.3. Topic modeling

Topic modeling is one of the most popular methods utilized for modeling the evolution of events over time [32]. A topic model can be a probabilistic model that relates documents and words through variables called topics [33]. Topics are the main subject matter or dominant themes [34] of a given text extracted from social media platforms, news articles, purchase behavior, etc. [35]. Topic discovery can be conducted through various methods based on clustering [36] or machine-learning techniques [37], enabling tracking of emerging issues [38]. Two leading methods utilized for topic modeling are Probabilistic Latent Semantic Analysis (PLSA) [39] and Latent Dirichlet Allocation (LDA) [18]. There are also several extensions to the standard LDA, such as Dynamic Topic Models [40] and Hierarchical Dirichlet Processes [41].

Latent Dirichlet Allocation (LDA) is utilized for topic modeling most frequently [42]. It is a generative probabilistic model for analyzing discrete data collections such as text corpus. It is defined as a "three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics" [18]. In the literature, LDA has been used for topic modeling texts related to many applications in various fields, such as medical science [43,44], political science [45,46], analyzing social networks [47,48], etc. LDA can also be used to conduct a literature review over a large volume of documents and perform prediction of research trends about hot subjects [49,50]. Following the COVID-19 outbreak, several studies have been formed based on applying LDA to find the most common topics expressed by social media users regarding the vaccines and vaccination process [11,51–53].

2.4. Sentiment analysis

Sentiment Analysis (SA) is a research field that analyzes people's sentiments toward different topics, events, individuals, issues, products, services, organizations, and their attributes [54]. In the literature, SA is also called opinion mining, review mining, appraisal extraction, or attitude analysis [55]. Sentiment analysis is considered a subdivision of Natural Language Processing (NLP), Machine Learning (ML), and computational linguistics. Some elements are also borrowed from sociology and psychology. The history of NLP started in the 1950s; however, the SA came to the spotlight by the growth of social media in the past few years [56].

Early studies in the field of SA concentrated on sentiment or subjectivity identification at the document or sentence level of granularity. Document-level and sentence-level tasks include the classification of reviews as positive or negative and distinguishing objective from subjective sentences. Today, aspect-oriented opinion mining, which analyzes opinions toward individual attributes of an object, also comes into consideration. Regardless of the level of analysis, these tasks can be conducted through ML-based or non-ML-based approaches [57]. Hence, sentiment analysis methods can be classified into three groups: ML-based, lexicon-based, and hybrid methods. ML-based techniques can be supervised, unsupervised, or semi-supervised, while lexicon-based methods use sentiment lexicons containing words annotated with the sentiment orientation [58]. SA can have applications in various fields, such as politics [59,60], financial [61,62], medical science and healthcare [63–66]. Following the COVID-19 outbreak, many studies have been performed based on the sentiment analysis of public opinions toward various issues related to the pandemic and especially COVID-19 vaccines through different social networks such as Twitter [10,11,13,67,68] and Reddit [12], or news articles [69].

2.5. Text classification

Text classification involves creating models that can categorize new documents into pre-determined categories. This process is now quite complicated, encompassing not just the training of models but also several other steps, such as data preprocessing, transformation, and dimensionality reduction. It continues to be a significant research area, employing various techniques and their combinations in intricate systems [70]. Over the past few decades, text classification problems have been extensively studied and applied in numerous real-world scenarios. Recent advancements in NLP and text mining have sparked increased interest among researchers in creating applications utilizing text classification methods. Typically, text classification and document categorization systems can be broken down into four main phases: feature extraction, dimension reduction, classifier selection, and evaluation [71].

Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and Global Vectors for Word Representation (GloVe) are common techniques utilized for feature extraction. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Non-negative Matrix Factorization (NMF), and also novel techniques such as random projection, autoencoders, and t-distributed Stochastic Neighbor Embedding (t-SNE) can be used for dimensionality reduction. Moreover, there are various classification techniques in this regard, such as Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM), while neural network architectures such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) have achieved surpassing results compared to previous ML algorithms. The final step in this pipeline is the evaluation that can be conducted based on various metrics such as accuracy, precision, recall, Receiver Operating Characteristic (ROC), and Area Under ROC Curve (AUC). Among these metrics, precision, and recall are widely utilized to measure the effectiveness of text classifiers [71].

Text classification techniques have been used in several studies related to COVID-19 vaccine. Some studies utilized text classification techniques to classify the tweets on different vaccines into positive, neutral, or negative based on vaccine brands to find out about COVID-19 vaccine hesitancy and sentiments toward numerous vaccines [72,73]. In a similar study, Bidirectional Encoder Representations from Transformers (BERT) and bidirectional Long Short-Term Memory (LSTM) were utilized to identify anti-vaccination Tweets [74]. Topic classification was also subject of some studies in this field. For example, a stacking ensemble classifier was used for the multi-class classification of COVID-19 vaccines topic on Twitter [75].

Table 1 summarizes some of the related works in the field of COVID-19 vaccine topic modeling, sentiment analysis, and text classification.

3. Research method

This section explains the research method steps to answer the research questions. Fig. 1 presents the stages of the research process.

3.1. Identify search strategy and configure filtering criteria

To conduct the literature review in the field of the COVID-19 vaccine, Scopus, which is one of the most widely employed academic databases, and PubMed, a free resource supporting the search and retrieval of life sciences and biomedical topics, were chosen. Scopus, according to its developer Elsevier, "is the largest abstract and citation database of peer-reviewed literature: scientific journals, books, and conference proceedings." Hence, Scopus can provide good coverage of scientific literature. PubMed, according to the National Library of Medicine, "comprises more than 34 million citations for biomedical literature from MEDLINE, life science journals, and online books," which provides an appropriate coverage of the literature related to the subject of this research. In the search strategy, suitable keywords are defined to search the literature on the "COVID-19 vaccine" in the Title.

Moreover, only the journal or review articles with accessible abstracts and articles written in English were investigated. By applying these criteria, documents published from March 2020 to April 2022 can be retrieved. Finally, the results obtained by searching in Scopus and PubMed were merged, and the duplicates were removed. Fig. 2 briefly shows the search strategy, filtering criteria, and the final number of articles prepared for the following steps.

3.2. Data exporting and storing

As mentioned, Scopus abstract and citation database and PubMed were used by searching ("covid_19 vaccine" OR "covid vaccine" OR "corona virus vaccine" OR "coronavirus vaccine" OR "covid19 vaccine" OR "covid 19 vaccine" OR "covid-19 vaccine") in Title of the articles, and the journal articles and reviews written in English from 2020 were selected. The results were exported and stored as an Excel file containing citation information, abstracts, and keywords.

Table 1
Summary of some of the research in the field of COVID-19 vaccine-related text analytics.

Authors	Year	Title	Method
Praveen et al. [51]	2021	Analyzing the attitude of Indian citizens toward COVID-19 vaccine–A text analytics study	LDA for topic modeling and Textblob for sentiment analysis
Lyu et al. [11]	2021	COVID-19 vaccine–related discussion on Twitter: topic modeling and sentiment analysis	LDA for topic modeling and the National Research Council of Canada Emotion Lexicon for sentiment and emotion analysis
Liu & Liu [10]	2021	Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis	LDA for topic modeling and Valence Aware Dictionary and sEntiment Reasoner (VADER) tool for sentiment analysis
Quyen et al. [74]	2021	Applying Machine Learning to Identify Anti-Vaccination Tweets during the COVID-19 Pandemic	BERT and Bi-LSTM with pre-trained GloVe embeddings and classic machine learning algorithms including Support Vector Machine and Naïve Bayes for text classification.
Marcec & Likic [13]	2022	Using twitter for sentiment analysis toward AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines	AFINN lexicon for sentiment analysis
Zulfiker et al. [52]	2022	Analyzing the public sentiment on COVID-19 vaccination in social media: Bangladesh context	LDA for topic modeling and deep learning and machine learning algorithms for sentiment analysis
Xu et al. [53]	2022	COVID-19 vaccine sensing: Sentiment analysis and subject distillation from twitter data	LDA for topic modeling and VADER model for sentiment analysis
Ogbuokiri et al. [68]	2022	Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts	LDA for topic modeling and the VADER model for sentiment analysis. Machine learning classification algorithms for validation of the outputs.
Jayapermana et al. [75]	2022	Implementation of Stacking Ensemble Classifier for Multi-class Classification of COVID-19 Vaccines Topic on Twitter	Combining Logistic Regression, Random Forest, and Support Vector Machine algorithms as first-level learners and Logistic Regression as a meta-learner for text classification.
Qorib et al. [72]	2023	COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset	TextBlob, VADER, and Azure ML for sentiment analysis. LinearSVC, LR, MultinomialNB, Random Forest, and Decision Tree for the classification model.

3.3. Data pre-processing

After data collection, the next step is Data preprocessing, which is text preprocessing in our case. The Python language was used to perform text preprocessing and other steps. This step was conducted by using Python libraries "nltk" [76], "string," and "re." Lowercasing all letters, removing punctuation and numbers, tokenization, which means splitting the text into units, and removing stop-words, which are common English words that do not provide much insight into the sentence, have been applied to the abstracts of the selected documents from the previous stage. Then, stemming was performed as another step of text preprocessing in which the variant word forms are mapped to their base form [77]. Finally, the words with less than three letters were eliminated. Next, a Term-Document Matrix was built to feed LDA. Term-Document matrix, also known as Document-Term matrix, describes each term's frequency in each document. This was done using the corpora module of the "gensim" library [78], which provides the corpus of texts as input to the LDA model.

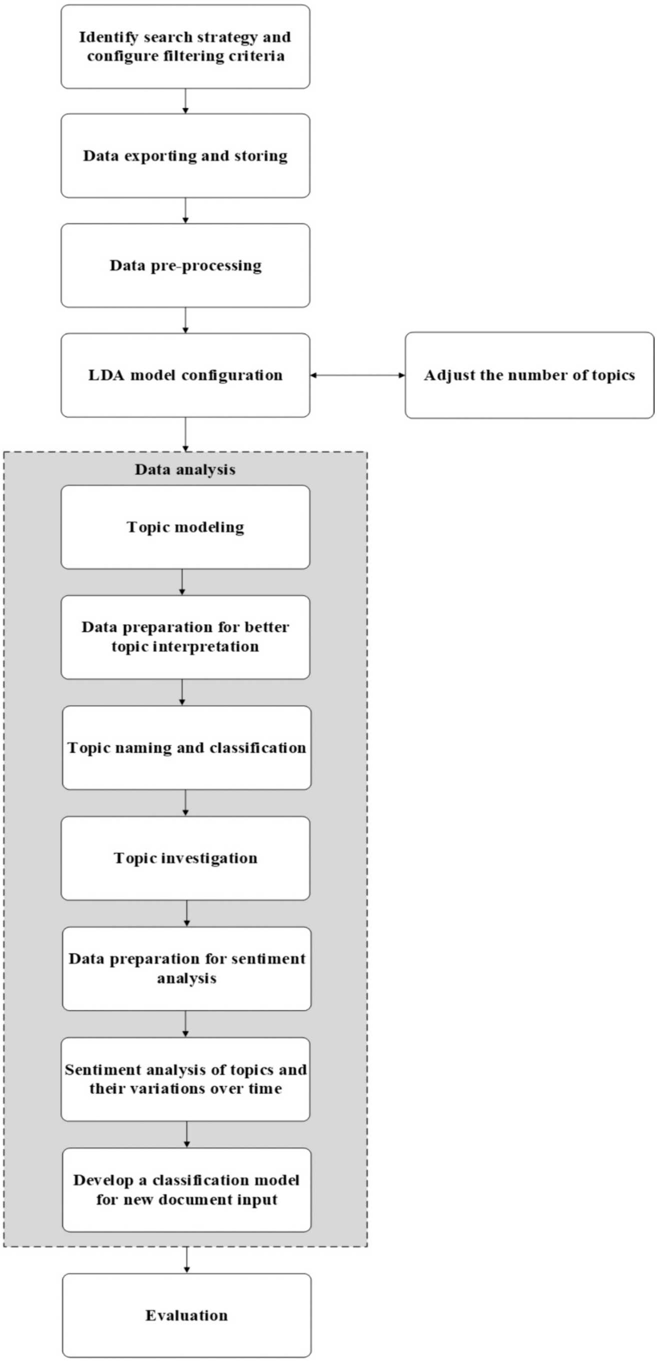

Identify search strategy and configure filtering criteria
Data exporting and storing
Data pre-processing
LDA model configuration
Adjust the number of topics
Data analysis
Topic modeling
Data preparation for better topic interpretation
Topic naming and classification
Topic investigation
Data preparation for sentiment analysis
Sentiment analysis of topics and their variations over time
Develop a classification model for new document input
Evaluation

Fig. 1. Research method steps.

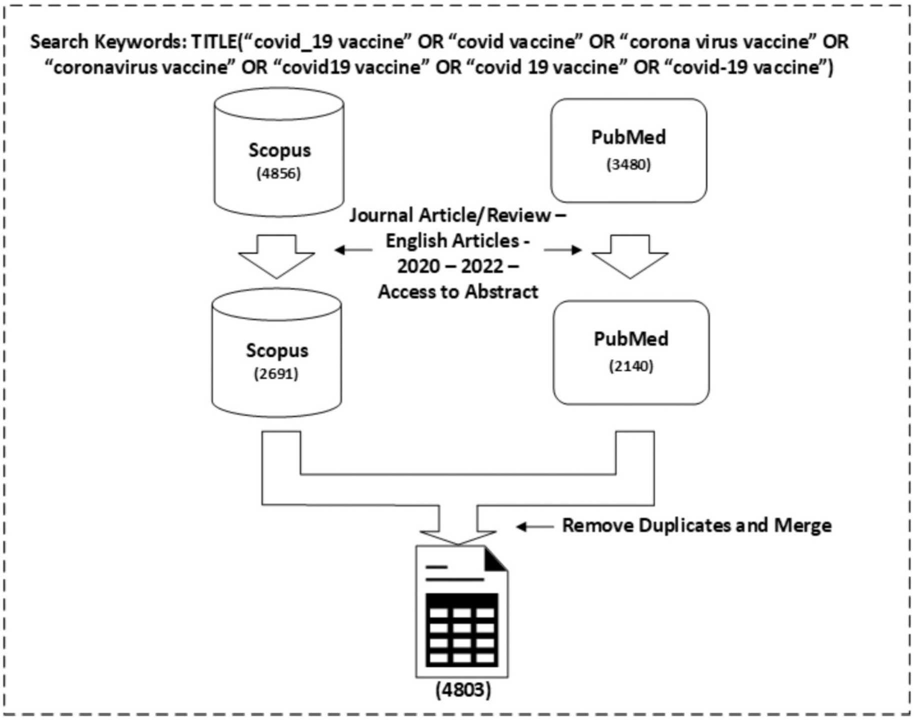

Search Keywords: TITLE("covid_19 vaccine" OR "covid vaccine" OR "corona virus vaccine" OR "coronavirus vaccine" OR "covid19 vaccine" OR "covid 19 vaccine" OR "covid-19 vaccine")
Scopus (4856)
PubMed (3480)
Journal Article/Review – English Articles - 2020 – 2022 – Access to Abstract
Scopus (2691)
PubMed (2140)
Remove Duplicates and Merge
(4803)

Fig. 2. Search strategy and filtering criteria.

3.4. LDA model configuration

The number of topics is the most important parameter of LDA that should be inferred from the corpus. Too few topics can result in an incomplete analysis, while too many topics would lead to several topics representing one cohesive subject [79]. To tackle this issue, the coherence score was used in this study. This measure is used to investigate the coherency of topics. Several studies that use LDA suggest that good topic modeling should lead to human-interpretable topics that are coherent but distinct from each other [79]. Therefore, the topic coherence score is an indicator of topic quality [80]. In this regard, to adjust the number of topics, the coherence score metric, which can be computed by CoherenceModel class in the "gensim" library, was used, and the number of topics resulted in a higher coherence score was chosen. In recent years, several topic coherence formulas have been proposed in the literature, and researchers have shown that these measures correlate with human judgment [81]. In order to calculate the coherence score in this research, the CV measure was used, which combines the indirect cosine measure with the Normalized Pointwise Mutual Information (NPMI) and the sliding window [82]. Hence, by changing the number of topics from 5 to 100, the coherence score of each model could be computed and compared to each other.

Topic distribution over documents and word distribution over topics are considered to have a prior probability of Dirichlet, so other hyperparameters that should be determined to configure the model are α and β, which are prior probability distribution for topics over documents and words over topics, respectively [83]. The value of β, a-priori belief on topic-word distribution affects the details of the model. The larger β is, the more words we will have in a topic and vice versa. On the other hand, α, a-priori belief on document-topic distribution relates to the number of topics that make up each document. The larger α is, the documents will be divided into more topics, and the smaller it is, the fewer topics we will have. In the current study, both hyperparameters were set to 'auto,' which means that the model would learn these parameters automatically.

3.5. Topic modeling

Data analysis consists of seven steps: Topic modeling, Data preparation for better topic interpretation, Topic naming and classification, Topic investigation, Data preparation for sentiment analysis, Sentiment analysis of topics and their variations over time, and Developing a model for new input to LDA clustering. After finding the optimal number of topics, the LDA function, which is LDAModel in "gensim" library, was applied. In other words, the LDA algorithm is implemented using gensim.models in python. As a result, the subset terms for each topic were derived. These terms were arranged based on the probability of their occurrence in a given topic, which is attained as weights. Therefore, each topic can be represented by its most probable terms, obtained from the posterior distribution over the assignments of words to topics.

3.6. Data preparation for topic interpretation

In the first round of LDA model training, it can be observed that some common and repeated words occur in several topics. Moreover, the first term in some topics is repeated in other topics with smaller weights, making interpreting the topics difficult. So, to refine the results and get more distinct topics, first, some common words such as "develop," "patient," "author," "among," etc., were eliminated besides the search keywords. Secondly, the first term in each topic, the word with the highest weight, was investigated in other topics, and if it had a smaller weight, it was eliminated from other topics. Thus, the words in each topic are more specific to that topic, and the interpretation and naming of the topics can be performed more effectively.

3.7. Topic naming

Latent topics were extracted as the clusters in which the terms were ordered by their probability of occurrence, which can be derived as coefficients multiplied by the terms. By labeling these clusters, the topics could be explored, and then by merging the topics, the research areas related to each group of topics were revealed. Hence, utilizing domain knowledge, each topic's semantic or representative label was inferred based on its most representative terms. Then again, based on expert knowledge, the labels merged to form the research areas. So, each research area consists of two or more labels that are related to each other.

3.8. Topic investigation

As mentioned, topics were aggregated to form the research areas, which can be utilized to discover the trend of research in the intended subject. In order to investigate the quality and interpretability of topics, different metrics, and tools such as coherence score, as described in Section 3.4, perplexity, and visualization of the topics were used. Perplexity evaluates how well a statistical model describes a dataset and is a commonly used measurement in information theory. A lower perplexity indicates a better probabilistic model [84]. Moreover, by utilizing the Inter-topic Distance Map implemented in the “pyLDAvis” library [85], a good topic model that consists of fairly big, non-overlapping bubbles scattered throughout the chart can be investigated.

3.9. Data preparation for sentiment analysis

For the purpose of sentiment analysis of each topic, ten most representative terms of the topic were considered. Moreover, since the frequency of each term in a specific topic can affect the emotional load of that topic, first, the number of occurrences of each term in a given topic was determined. This could be conducted through pyLDAvis.gensim.models.prepare function as it uses the frequency of terms to visualize the clusters. pyLDAvis provides information about each topic that can be extracted, including the frequency of each term in the topic [85]. By utilizing these frequencies, data preparation for sentiment analysis was performed by multiplying the frequency of each term by its sentiment score and finally dividing the whole score by the total number of terms. In data preparation for sentiment analysis steps, all the previous preparations were also taken into account; however, the representative terms were completed from their stemmed forms since they should be matched to lexicon terms to analyze their sentiment score.

3.10. Sentiment analysis of topics and their variations over time

Sentiment analysis of the topics was performed using SentimentIntensityAnalyzer of VADER from “nltk” package. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon-based and rule-based sentiment analysis tool. It is an efficient tool as it successfully deals with various types of texts [86]. VADER sentiment analysis relies on a dictionary that maps lexicon features to emotion intensities known as sentiment scores. The sentiment score of a given text can be attained by adding up the intensity of each word in the text. This model not only determines the polarity of a word but also specifies the strength of emotion. To determine the sentiment of each term, it generates a normalized score called a compound score, ranging from −1 (extremely negative) to +1 (extremely positive). Researchers typically classify texts as positive, neutral, and negative based on the following thresholds: positive (compound $\geq$ 0.05), neutral ($-0.05 <$ compound <0.05), and negative (compound≤ -0.05) [10]. However, the threshold is flexible and can be optimized based on domain knowledge. In this study, in addition to the thresholds mentioned above, which are typically utilized, other thresholds were used to see whether the results significantly changed.

The other important part of this step includes topic modeling of the articles based on the year of publication and comparing the sentiments of topics over time. This would lead to a better insight into the variations of sentiments in academic perspectives through different years and their probable causes. For performing this analysis, the articles were divided into three groups: 2020, 2021, and 2022. For each of the years, the previous steps were conducted separately, and final sentiment scores for the topics were derived to be compared over time.

3.11. Develop a classification model for new document input

In this step, a model was proposed to classify any new document as one of the pre-defined topics extracted in previous steps. In topic models, the words of a document are treated as resulting from a set of latent topics, a set of unknown distributions, over the lexicon. In other words, topic models are a reduction of classical document mixture models, which relate each document to a single topic [87]. The idea used in this research is to convert the unsupervised topic modeling to be used in a supervised classification problem to test whether the distribution per abstract of latent topics could predict the dominant topic of a new document and its positive or negative sentiment accordingly.

First, feature extraction, or more simply, vectorization, was used, which is a necessary step toward language-aware analysis. In order to vectorize a corpus, every document should be represented as a vector. Various vector encodings are used to vectorize a corpus, such as the bag-of-words approach, TF-IDF approach, word2vec or doc2vec algorithms, etc. One of the simplest while extremely effective models is the encoding of semantic space based on the bag-of-words model, whose main insight is that meaning and similarity are encoded in the corpus [88].

After feature extraction and defining the labels, the dominant topic of each document from 0 to 24 in our case, a classification model was trained. Building a classifier after topic analysis for the whole dataset is based on the correspondence between the words of the abstracts and the dominant topic of each document derived by posterior probability over the topics. To train the classifier, one of the classification algorithms, such as Artificial Neural Network (ANN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), etc., was chosen based on the accuracy metric. Finally, this model was evaluated on the test set so that the classifier could predict the topic of the new abstract and its corresponding sentiment. “keras” and “sklearn” libraries were used to implement the algorithms.

3.12. Evaluation

Two measures, as described in Section 3.8, were used to investigate the coherency and interpretability of the topics. CV measure that uses NPMI is presented in formula (1) [89,90], and perplexity is presented in formula (2) [18].

$$NPMI(w_i) = \sum_{j}^{N-1} \frac{log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-logP(w_i,w_j)} \tag{1}$$

where w_i represents the words. Based on formula (1), the word association features for the top-N topic words of a topic can be computed.

The perplexity is defined for a test set of M documents as follows:

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} logp(w_d)}{\sum_{d=1}^{M} \mathrm{N}_d}\right\} \tag{2}$$

where a document is a sequence of N words denoted by $\mathrm{w} = (\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_\mathrm{N})$ and w_n is the nth word in the sequence [18].

Moreover, to evaluate the classification model, accuracy, precision, recall, and F-Measure metrics were employed. One of the most common metrics in practice utilized by researchers is accuracy, which is used to evaluate the generalization capability of classifiers. By using accuracy, the trained classifier is evaluated based on the total number of instances of data that are correctly predicted by the trained classifier when tested with unseen or new data, so this metric measures the ratio of correct predictions over the total number of instances evaluated as provided in formula (3) [91].

$$Accuracy\ (acc) = \frac{t_p + t_n}{t_p + f_p + \mathrm{t}_n + f_n} \tag{3}$$

where t_p and t_n represent the number of positive and negative instances that are correctly classified. While f_p and f_n represent the number of misclassified negative and positive instances, respectively.

However, due to the imbalanced nature of the dataset, other metrics such as recall, precision, and F-Measure were also used to comprehensively evaluate the classification model performance. Recall is used to measure the fraction of positive samples that are correctly classified, and precision is used to measure the positive samples that are correctly predicted from the total predicted patterns in a positive class. Moreover, F-Measure represents the harmonic mean between recall and precision values. Formulas (4) to (6) provide these metrics, respectively [91].

$$Recall\ (r) = \frac{t_p}{t_p + f_n} \tag{4}$$

$$Precision\ (p) = \frac{t_p}{t_p + f_p} \tag{5}$$

$$F - Measure\ (FM) = \frac{2 \times p \times r}{p + r} \tag{6}$$

4. Results

This section provides the results based on the research method and data analysis explained in Section 3. Fig. 3 shows that the number of publications and citations on the COVID-19 vaccine has increased from 2020 to 2022. It can be seen that the COVID-19 vaccine is a hot research area, and a considerable amount of research has been conducted in this area.

Based on the search results, the articles that met our filtering criteria consist of 4803 abstracts of papers about the COVID-19 vaccine indexed in Scopus or can be retrieved through PubMed. Table 2 provides information about the collected dataset, including 4803 papers.

The number of distinct terms in the Term-Document Matrix after data preprocessing was 11,737. As described in Section 3.3, the number of topics was changed to find the optimal number of topics based on the coherence score. The result of the model selection is shown in Fig. 4.

The results show that the coherence score for 25 topics is higher than the other number of topics and is equal to 0.44. So, 25 was chosen as the LDA model's number of topics. Fig. 5 presents the inter-topic distance map for the configured LDA model. As mentioned before, a good topic model will have relatively big, non-overlapping bubbles dispersed all over the graph rather than being clustered in one quadrant. So, Fig. 5 indicates that the result of topic modeling is acceptable from this point of view since the clusters are dispersed in all four quadrants, and most are non-overlapping.

Table 2
Dataset used for literature review.

Number of abstracts	4803
Total number of words	1,083,041
Total number of characters (without space)	6,455,269
Total number of characters (with space)	7,533,507

Perplexity, as mentioned, is another measure used to evaluate the model and is a statistical method used for testing how efficiently a model can handle data it has never seen before. Generally, the lower the perplexity value, the higher the accuracy [92]. The perplexity for our model with 25 topics is −7.87.

The selected LDA model with 25 topics was applied to the Term-Document Matrix using the LdaModel function implemented in the Python library genism. As LDA provides soft clustering of the terms, there are overlaps between the terms in different clusters. Table 3 presents the five most probable terms in each cluster. Each topic's terms are ordered by the weight multiplied by each term which is a probability that each term is assigned to a given cluster. This weight indicates the probability that the term relates to the topic.

Therefore, each cluster can be labeled by investigating each topic's most probable (representative) terms. In the next step, clusters that relate to each other are grouped to develop the research areas of the dataset by merging the related topics. Fig. 6 depicts the graph of topics and research areas of the corpus. The research areas were extracted in a 2-stage process:

1. A label was assigned to each cluster by reviewing the semantics of each cluster's terms (T1 to T25).
2. Major research areas of the domain under consideration were obtained by experts' domain knowledge.

The stages of this process (Labeling the topics and extracting the research areas) were conducted carefully by expert opinions and domain

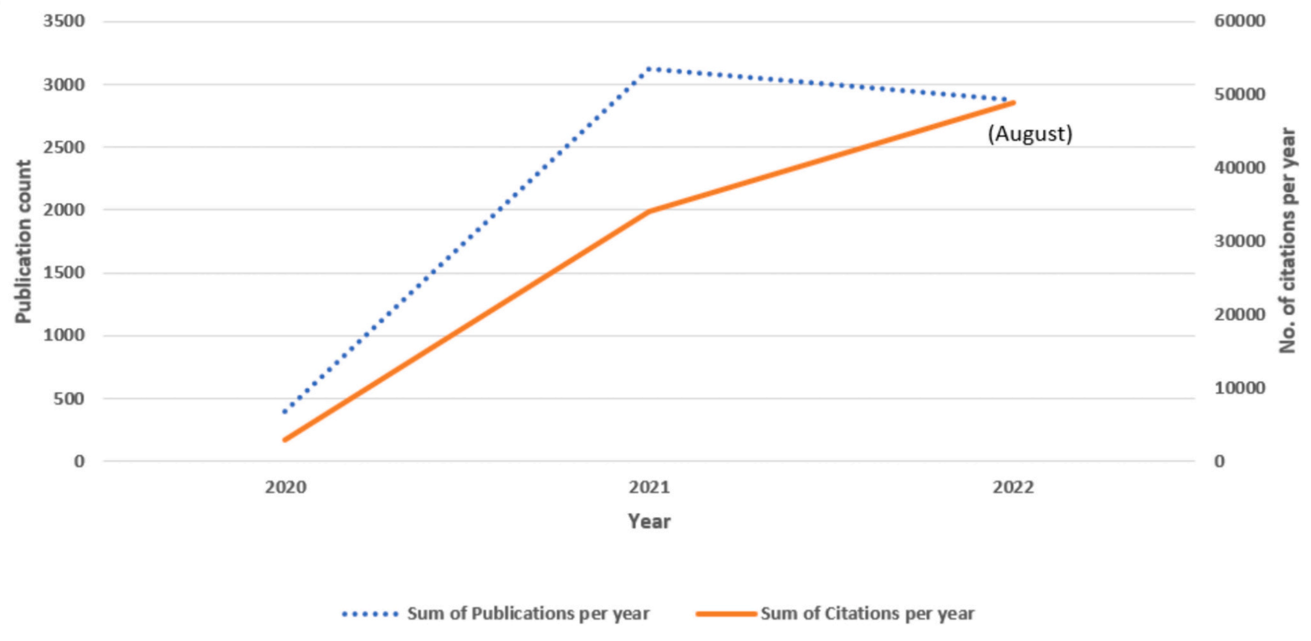


Fig. 3. Annual publications and citations based on the Scopus database.

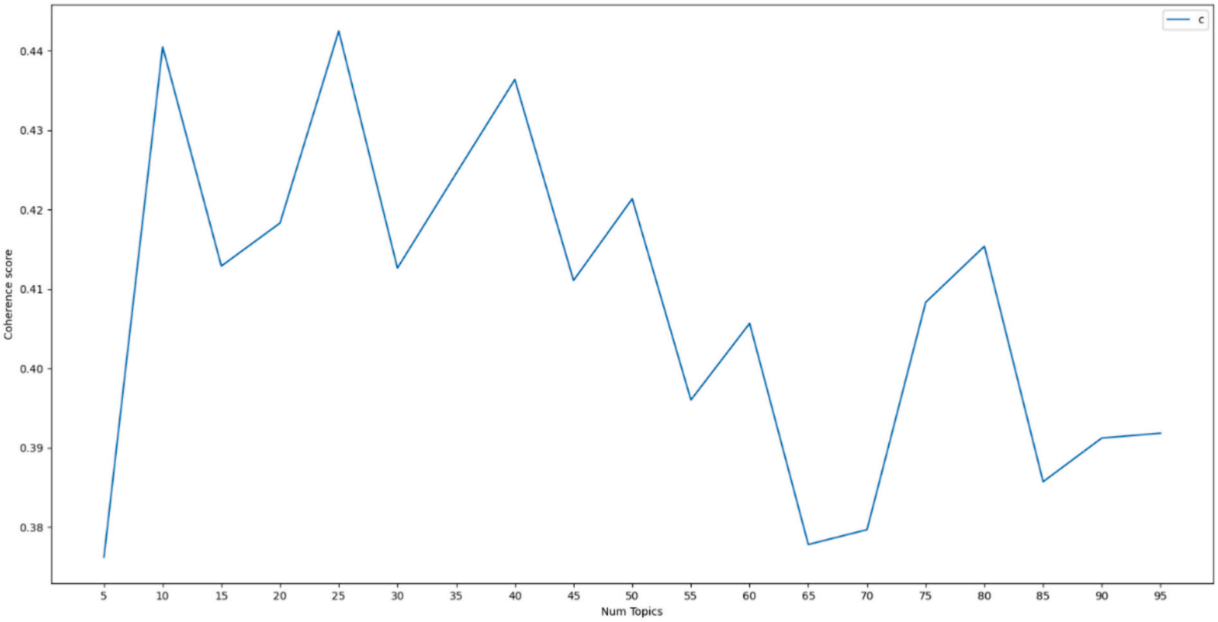
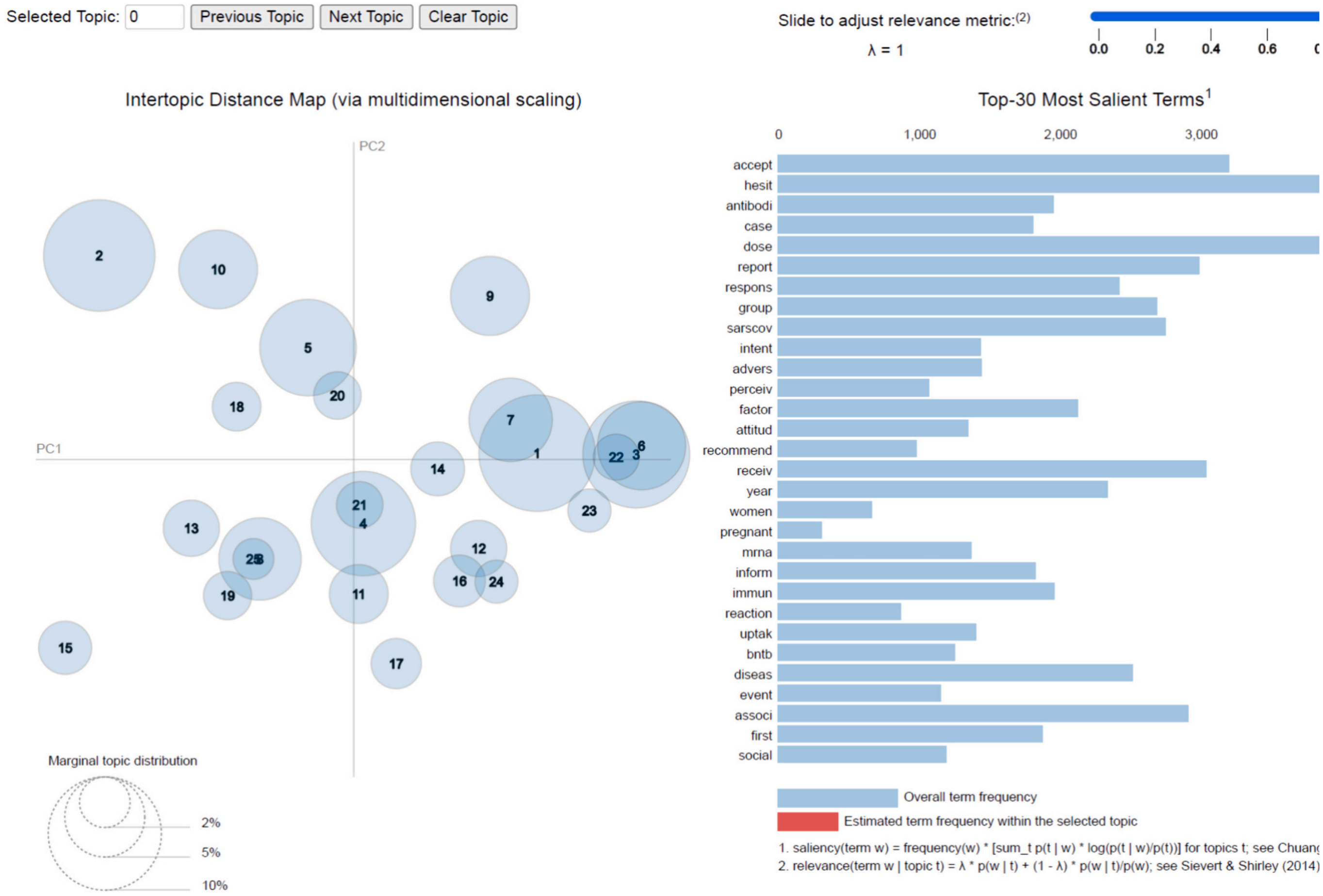


Fig. 4. The coherence score of each LDA model based on topic number.


Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 5. Inter-topic distance map for the LDA model with 25 topics.

knowledge, including specialist physician, pharmacist, and epidemiologist.

As depicted in Fig. 6, eight research areas were recognized through topics extracted from the literature. These research areas are "Reporting," "Acceptance," "Reaction," "Surveyed Opinions," "Pregnancy," "Titer of Variants," "Categorized Surveys," and "International Approaches." Each research area includes two or more topics. For example, "Reaction" relates to the topics that mention the side effects of

Table 3
Final topics and their five representative terms.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	dose	response	allergy	case	like	report	efficacy
2	report	immune	anaphylaxis	report	social	mrna	pandemic
3	adverse	control	pandemic	rate	disease	disease	disease
4	event	immunogen	public	present	inform	reaction	safety
5	receive	data	allergic	show	media	receive	review

	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
1	antibody	accept	group	hesitate	myocarditis	attitude	country
2	infect	perceive	year	inform	immune	intent	population
3	risk	factor	receive	receive	disease	accept	strategy
4	titer	receive	public	social	year	knowledge	death
5	level	imid	phase	factor	test	factor	model

	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21
1	factor	reside	individual	recommend	pregnant	thrombocytopenia	communicate
2	accept	staff	year	acip	pregnancy	report	result
3	pharmacist	population	associate	issue	women	symptom	cancer
4	uptake	compare	survey	inform	receive	thrombosis	access
5	increase	year	uptake	year	adverse	present	willing

	Topic 22	Topic 23	Topic 24	Topic 25
1	node	accept	accept	sarscov
2	inject	survey	population	infect
3	lymph	concern	associate	immune
4	site	associate	pandemic	bntb
5	first	healthcare	adult	mrna

vaccines or cross-reactivity with other diseases. The articles discussing these topics were mainly published after the start of the vaccination and comprised a part of the literature in this field.

In the next step, a sentiment analysis of the topics was conducted. After using SentimentIntensityAnalyzer() from the nltk package, as explained in Section 3.10, the amount of positive, negative, and neutrality of each word and their combination (compound) was calculated. This value was multiplied by the frequency of that word, and then the scores obtained from 10 words related to each topic were added together and divided by the sum of frequencies. First, the threshold for determining the sentiment of each topic is considered equal to 0.05 since researchers typically use it. Then, the 0-threshold was also considered. Table 4 presents the sentiment score and its corresponding sentiment based on 0.05 and 0 thresholds, respectively, for 25 topics.

It can be observed that if sentiment analysis is applied to all of the abstracts, two of the topics are recognized as positive, two as negative, and the other topics have a neutral sentiment. By considering the 0-threshold, 11 topics are identified as negative, 8 as positive, and 6 topics are also recognized as neutral. So, the number of negative topics is more than positive or neutral topics.

In order to investigate the sentiment of topics more deeply, the variation of the sentiments over time was also examined. As mentioned in Section 3.10, for this purpose, separate LDA models were developed for the articles of each year from 2020 to 2022. Table 5 presents the number of articles, the optimum number of topics computed based on the coherence score, and the percentage of every sentiment for the total topics of each year's LDA model. The threshold for determining the sentiment was considered 0.05.

Finally, a classification model was developed. Comparing the results of various algorithms, Long Short-Term Memory (LSTM) outperforms other algorithms significantly. LSTM is a kind of Recurrent neural network (RNN). RNNs show remarkable performance when dealing with serialized data since the network can memorize previous information and can use this information in current output calculations. So, it remains a connection between nodes in hidden layers, which leads to the integration of the information of the front and back positions in an effective way. However, these networks have deep memory for the last input signal and relatively shallow memory for the early input signal, which causes a problem called gradient disappearance. Therefore, RNN's LSTM model would solve this problem effectively by using the context's feature information, preserving the sequence information of the text [93]. Hence, LSTM networks are widely used for text classification [94].

Moreover, in order to improve the accuracy of the classification model, the combination of LSTM with Convolutional Neural Network (CNN) was utilized. CNN can be considered the main technique for extracting data features in deep learning. The convolutional layer is the core of CNN, which performs convolution operations on words to obtain a more sophisticated feature representation. It uses the convolution kernel to conduct feature extraction and mapping on the text data [95]. CNN-LSTM is a Recurrent CNN (RCNN) variant that consists of a recyclable convolutional neural network and a maximal pool layer [96].

So, in this research CNN-LSTM model was used since it outperforms other models based on the accuracy metric. Table 6 presents the accuracy metric of the prediction on the test dataset for various classification algorithms trained on the COVID-19 vaccine corpus.

The CNN-LSTM model uses CNN to extract high-level phrase expressions from sequences and then uses LSTM. This way, better performance can be achieved than a standalone CNN model and a standalone LSTM model [96]. First, by using word embedding, the text is converted into low-dimensional word vectors. Then, CNN is used for feature extraction and combines LSTM to preserve the feature of historical information in text sequences. Therefore, the inadequacy of CNN in extracting contextual association semantics would be compensated [95]. Lastly, a fully connected layer is used for classification output.

Table 7 and Table 8 summarize the CNN-LSTM model developed for the COVID-19 vaccine corpus and the hyperparameters of the model, respectively. The hyperparameters of the model were achieved through grid search and by comparing the accuracy metric for various sets of

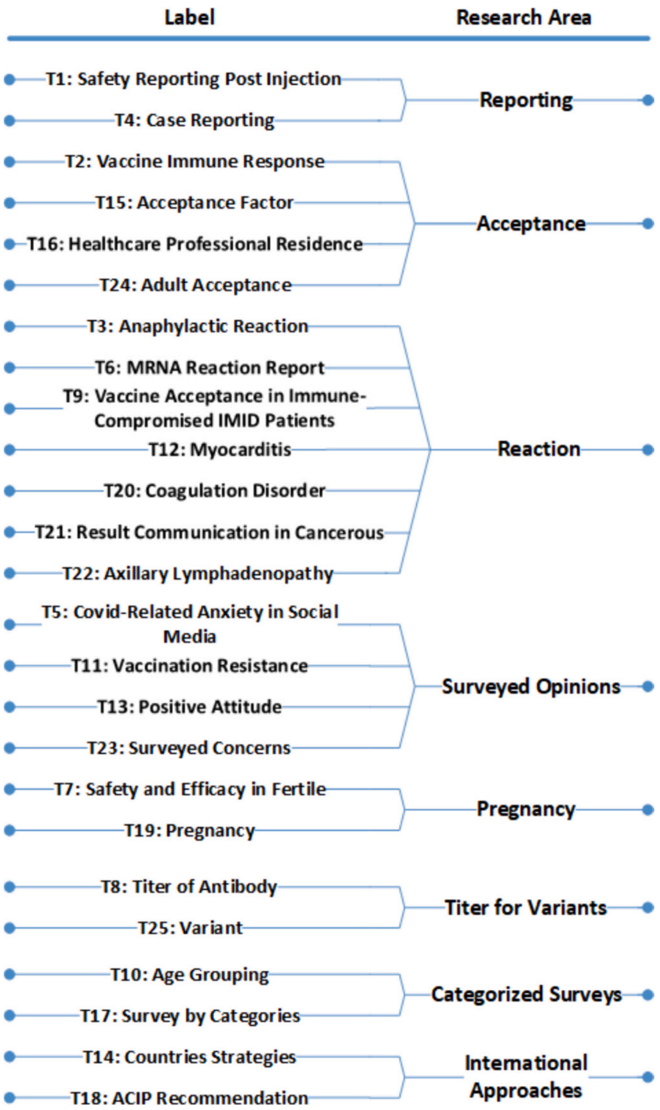

Label
Research Area
T1: Safety Reporting Post Injection
T4: Case Reporting
Reporting
T2: Vaccine Immune Response
T15: Acceptance Factor
T16: Healthcare Professional Residence
T24: Adult Acceptance
Acceptance
T3: Anaphylactic Reaction
T6: MRNA Reaction Report
T9: Vaccine Acceptance in Immune-Compromised IMID Patients
T12: Myocarditis
T20: Coagulation Disorder
T21: Result Communication in Cancerous
T22: Axillary Lymphadenopathy
Reaction
T5: Covid-Related Anxiety in Social Media
T11: Vaccination Resistance
T13: Positive Attitude
T23: Surveyed Concerns
Surveyed Opinions
T7: Safety and Efficacy in Fertile
T19: Pregnancy
Pregnancy
T8: Titer of Antibody
T25: Variant
Titer for Variants
T10: Age Grouping
T17: Survey by Categories
Categorized Surveys
T14: Countries Strategies
T18: ACIP Recommendation
International Approaches

Fig. 6. Graph of topics and research areas extracted by topic modeling.

hyperparameters. Other hyperparameters not mentioned in Table 8 were set to their default values.

Then, the model was trained by the training dataset, which is 95 % of the whole dataset, 4562 abstracts, to reduce the overfitting. The input data for training the model was derived by tokenizing preprocessed abstracts into a sequence of words and considering the maximum length of the abstract as 120. So, zero-padding was applied to the abstracts with a length of <120. This length was obtained by trying various values to get the maximum accuracy. The hyperparameters of the models were tuned based on the accuracy metric. The model was evaluated using hold-out validation and K-Fold cross-validation (K = 15). Fig. 7 shows the accuracy variations for the training set and test set per epoch, while the best model can predict the new document with 75 % accuracy.

Table 9 provides the accuracy, recall, precision, and F-Measure of the final model to predict the dominant topic of new documents. It should be noted that for calculating the metrics such as recall, precision, and F-Measure, since there are 25 classes, metrics for each label were computed, and then their unweighted mean were derived which can control the imbalanced multiclass problem.

Table 4
Sentiment scores of 25 topics related to COVID-19 vaccine literature.

Topic No.	Topic label	Sentiment_Score	Sentiment (threshold = 0.05)	Sentiment (threshold = 0)
1	Safety reporting post injection	−0.018733767	Neutral	Negative
2	Vaccine immune response	0.036211623	Neutral	Positive
3	Anaphylactic reaction	−0.025742017	Neutral	Negative
4	Case reporting	0	Neutral	Neutral
5	COVID-related anxiety in social media	−0.02079648	Neutral	Negative
6	MRNA reaction report	0	Neutral	Neutral
7	Safety and efficacy in fertile	0.012649654	Neutral	Positive
8	Titer of antibody	−0.027075869	Neutral	Negative
9	Vaccine acceptance in immune-compromised IMID patients	0.212039045	Positive	Positive
10	Age grouping	0	Neutral	Neutral
11	Vaccination resistance	−0.044599378	Neutral	Negative
12	Myocarditis	0.032301354	Neutral	Positive
13	Positive attitude	0.018661974	Neutral	Positive
14	Countries strategies	−0.092049477	Negative	Negative
15	Acceptance factor	0.027148545	Neutral	Positive
16	Healthcare professional residence	0	Neutral	Neutral
17	Survey by categories	−0.02722305	Neutral	Negative
18	ACIP recommendation	0.080810126	Positive	Positive
19	Pregnancy	−0.01759323	Neutral	Negative
20	Coagulation disorder	0	Neutral	Neutral
21	Result communication in cancerous	−0.065082583	Negative	Negative
22	Axillary lymphadenopathy	0	Neutral	Neutral
23	Surveyed concerns	−0.020840013	Neutral	Negative
24	Adult acceptance	−0.020399772	Neutral	Negative
25	Variant	0.022842863	Neutral	Positive

Table 5
The proportion of each topic sentiments per year.

Year	No. of articles	No. of topics	Positive	Negative	Neutral
2020	234	15	40 %	0	60 %
2021	2981	30	10 %	7 %	83 %
2022	1588	25	0	8 %	92 %

Table 6
The accuracy of various classification algorithms in comparison to CNN-LSTM.

Classification algorithm	Accuracy
RBF Kernel SVM	19 %
Random Forest	20 %
Linear Kernel SVM	16 %
Gaussian Naïve Bayes	13 %
Decision Tree	16 %
XGBoost	21 %
CNN-LSTM	75 %

5. Discussion

There are several studies that utilized topic modeling for literature analysis in the COVID-19 area [97–99]. For example, Cao et al. [100] used LDA and topic visualization methods to extract research trends and topic similarities of COVID-19 research. Analysis of trends and patterns in COVID-19 research was also conducted by Dornick et al. [101] by

Table 7
Model specifications for classification of new documents.

Layer (type)	Output Shape	Param #
embedding_24 (Embedding)	(None, 120, 100)	1,607,000
conv1d_22 (Conv1D)	(None, 120, 100)	30,100
max_pooling1d_23 (MaxPooling1D)	(None, 60, 100)	0
lstm_25 (LSTM)	(None, 20)	9680
dense_19 (Dense)	(None, 25)	525
Total params: 1,647,305		
Trainable params: 1,647,305		
Non-trainable params: 0		

Table 8
The hyperparameters of the CNN-LSTM text classification model.

Hyperparameter	Value
Embedding vector length	100
CNN kernel size	3
CNN activation function	'relu'
MaxPooling1D pool size	2
LSTM units	20
Dense layer units	25
Dense layer activation function	'softmax'
Optimizer	Adam
Learning rate	0.001
Loss function	'categorical crossentropy'
Metrics	'accuracy'
Epochs	30

utilizing LDA, Bidirectional Encoder Representations from Transformers (BERT), and sentiment analysis. Furthermore, some researchers performed the literature review and meta-analysis as their secondary research method to investigate the COVID-19 vaccine-related articles [24,102]. Zheng et al. [24] reviewed the articles about the effectiveness of COVID-19 vaccines using meta-analysis, and Noruzi et al. [103] applied a scientometric approach to review the vaccine-related articles. Topic modeling and sentiment analysis have been employed widely on social media and news outlets for analyzing public opinions and outlooks toward different vaccines such as COVID-19 [104], Influenza [105], and HPV [106]. These methods have also been applied to other rapidly growing publications in the medical field, such as cancer immunotherapy [107], to find emerging trends. However, the findings of this research make it possible to present a framework for COVID-19 vaccine research topics. The novel approach of combining topic modeling and sentiment analysis for the desired research topics and detecting the dominant topic of new abstracts leads to achieving sentiments of researchers toward the topics in any research field.

Based on the collected dataset, which contains studies from 2020 to 2022 about the COVID-19 vaccine, and data analysis by topic modeling, sentiment analysis, and developing a classification model, our findings are now presented to answer the research questions.

RQ1: What are the main research topics in the "COVID-19 vaccine" field?

The LDA model was applied to the dataset, and after adjusting the number of topics, generated 25 latent topics. After applying LDA, one of the key tasks is to label these latent topics based on the domain knowledge, representing the main research COVID-19 vaccine areas. After assigning labels to 25 extracted topics, these topics were categorized into 8 main research areas, as presented in Fig. 6. Table 10 presents these 8 research areas derived by merging the relevant labels.

Each research area includes several labels. For example, "Reaction" consists of recognized reactions studied in the COVID-19 vaccine literature such as "Anaphylactic Reaction," "MRNA Reaction Report,"

Table 9
CNN-LSTM text classification model evaluation results.

Dataset	Accuracy	Recall	Precision	F-Measure
Train	98 %	98 %	98 %	98 %
Test (hold-out validation)	75 %	65 %	70 %	67 %
Test (15-Fold cross-validation)	72 %	66 %	68 %	67 %

Table 10
Main research areas of the COVID-19 vaccine were obtained by applying LDA and merging related labels.

Main research areas			
Reporting	Acceptance	Reaction	Surveyed opinions
Pregnancy	Titer for variants	Categorized surveys	International approaches

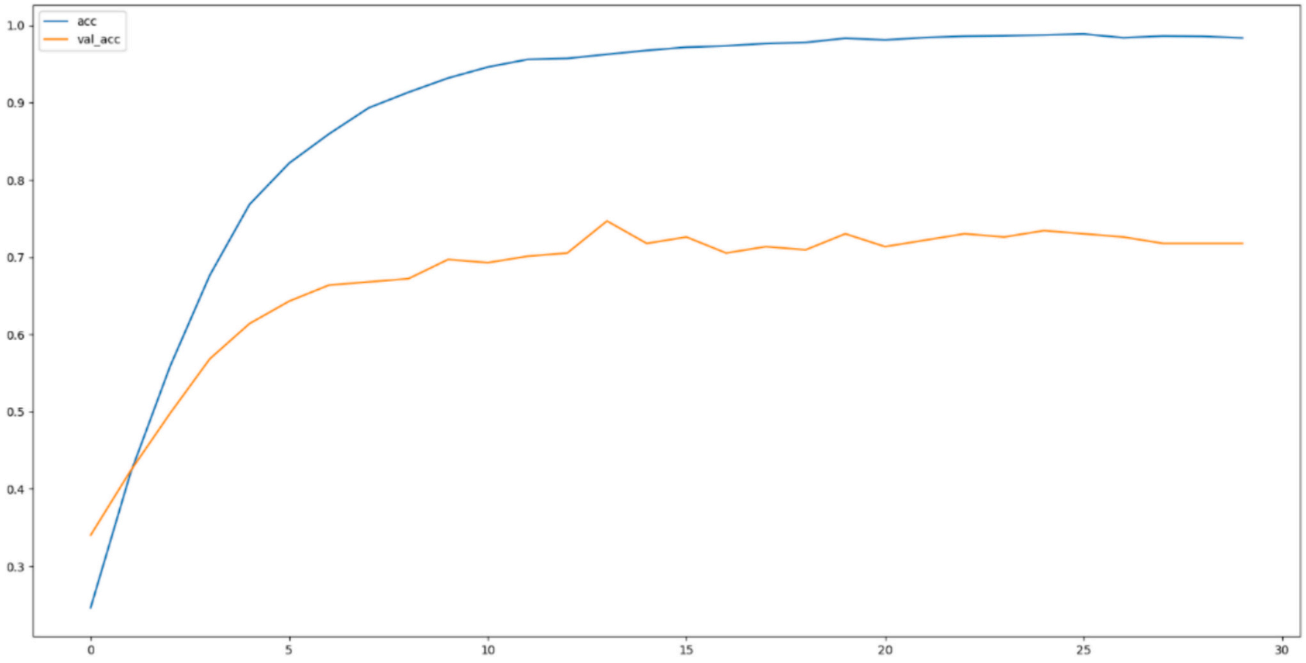


Fig. 7. Accuracy of the classification model for the training set and validation set in 30 epochs.

"Vaccine Acceptance in Immune Compromised IMID Patients," "Myocarditis," "Coagulation Disorder," "Result Communication in Cancerous" and "Axillary Lymphadenopathy." As another example, the "Surveyed Opinions" research area relates to public opinions in social media and their attitudes toward vaccination, which assigned many articles to itself. This research area consists of "Covid-Related Anxiety in Social Media," "Vaccination Resistance," "Positive Attitude," and "Surveyed Concerns."

A study which was conducted by Li et al. [8] in the field of the COVID-19 vaccine indicates that a significant portion of the articles addressed vaccination policies, vaccine social and psychological outlooks, and the effectiveness of approved vaccines against mutations, which are also three research areas derived by topic modeling as "International Approaches," "Surveyed Opinions," and "Titer for Variants" respectively.

RQ2: What are the sentiments of the topics, and how have they been changed from 2020 to 2022?

As presented in Section 4, the sentiments of 25 topics are analyzed by utilizing VADER and based on ten representative terms of each topic. The results, which are presented in Table 4, show that by considering 0.05 as the threshold, the numbers of negative and positive topics are equal. However, considering the 0-threshold, negative topics are more than positive ones. Topics such as "Anaphylactic Reaction," "Covid-Related Anxiety in Social Media," "Vaccination Resistance," and "Result Communication in Cancerous" are recognized as negative topics, while topics such as "Vaccine Immune Response," "Vaccine Acceptance in Immune-Compromised IMID Patients," and "ACIP Recommendation" are recognized as positive ones. Investigating the sentiments of topics over time can lead to a better understanding of the sentiments in various time epochs. As can be observed in Table 5, the related articles in 2020 are positive or neutral, and none of the topics are negative. It shows that in 2020, the scientific communities, similar to the media and public, placed hope at large on having a vaccine that protects against COVID-19. In 2021, though there are still more positive topics than negative ones, there are some negative topics in the articles, which shows that the focus of research has shifted to the rapid and effective rollout of the vaccine and its effectiveness. Finally, in 2022, more negative topics have been addressed in scientific articles, which mainly relate to the side effects of vaccines, case-report articles, and concerns about the effectiveness and safety of additional vaccine doses against highly mutated novel variants.

RQ3: How can a new document about the COVID-19 vaccine be classified on provided topics and measure its sentiment?

As demonstrated in Sections 3 and 4, a model based on a Convolutional Neural Network and a Recurrent Neural Network (LSTM) was developed to classify new documents to one of the pre-defined topics. This model was chosen from various classifying algorithms such as Random Forest and Support Vector Machine due to its significantly better performance on a text corpus. By using embedding, a mapping of words to a vector of continuous numbers was achieved, which allowed the words with similar meanings to have a similar representation. Utilizing the CNN led to feature extraction, and finally, LSTM preserved the feature of historical information in text sequences. After hyperparameter tuning, the results show that the model can predict the dominant topic and, consequently, the sentiment of a new document with 75 % accuracy, which outperforms remarkably compared to other algorithms that were tried on our corpus, as presented in Table 6. Moreover, this model can be utilized for future research design in this arena. Researchers can classify their research plan as one of the topics and research areas determined in this study so that they can pinpoint their position in the topic structure. Furthermore, if their research plan does not belong to any identified topics, it indicates that this can be a novel study. In this way, this study can be utilized as a guideline or strategy to find research areas that are facing more attention or, on the other hand, novel research areas with fewer studies performed on them if they are not categorized as any pre-determined topic. Also, the top representative terms of the dominant topic can help researchers with their search strategy and be utilized as their keywords.

6. Conclusion

The outbreak of COVID-19 in late 2019 and its overwhelming impacts catalyzed the development of vaccines and their technologies. So, due to the need for timely sharing of information about findings, collaboration among different research entities and scientists increased significantly [8]. Consequently, a considerable number of articles about the COVID-19 vaccine have been published since 2020. In this research, the applicable insights from this large number of articles have been extracted by utilizing text mining and machine learning tools. Applying the LDA model to the corpus of articles led to identifying 25 topics and 8 main research areas in this field. Moreover, sentiment analysis of the most representative terms of these topics and the variations of these sentiments over time can indicate the focus of research articles and researchers' attitudes toward this subject in various years. Also, in this study, a classification model was developed by using advanced machine learning algorithms such as CNN and LSTM. By employing this model, a new document can be classified as one of the pre-determined topics derived by LDA based on its abstract words. The findings of this article show that "Reporting," "Acceptance," "Reaction," "Surveyed Opinions," "Pregnancy," "Titer for Variants," "Categorized Surveys," and "International Approaches" are the research areas in the literature which can be derived by topic modeling and using expert knowledge to assign labels to the topics.

The current study has moved the boundary of data analytics in secondary research in the vaccine field by applying topic modeling and sentiment analysis to COVID-19 vaccine-related publications. In addition, for the first time, in this study, a classification model was developed to classify the topic of a new article as one of the identified topics.

It should be noted that this research has some limitations in terms of data source, filtering criteria, and method. Scopus and PubMed were used as the sources of collected literature, and only the journal articles and reviews were selected. Furthermore, several clustering methods can be used and compared to find the most coherent and interpretable topics, and other sentiment analysis approaches, such as ML-based methods, would lead to more precise sentiment analysis of the topics.

The implications of this research fall into two categories based on a novel approach utilized. In this approach, though some previously developed models and methods were employed, a specific application and the integration of topic modeling, sentiment analysis and topic classification for performing new kind of Systematic Literature Review (SLR) in the healthcare and vaccine studies, along with expert knowledge in the COVID-19 vaccine field, are the novelty and scientific contributions of this research. From the managerial point of view, identifying the scientific fields and the trends in this domain may help researchers find a roadmap for these research trends. In addition, sentiment analysis of the topics can direct future studies, and based on the developed classification model, the dominant topic and sentiment of new documents can be predicted. From the practical perspective, budget-balancing of the COVID-19 vaccine can be benchmarked and adopted through extracted topics, and vaccine manufacturing companies will plan to invest in the gap of areas that have not been focused on extracted topics so they will achieve a niche market to invest in. It brings up the practical implications of current research. Future research can extend our work in various ways, such as more coverage in the database and using other databases and other types of documents so that the generalization and accuracy of the classification model would also be improved. Other clustering and sentiment analysis techniques can also be applied and compared.

CRediT authorship contribution statement

Saeed Rouhani: Writing – review & editing, Writing – original draft, Supervision, Project administration, Formal analysis,

Conceptualization. **Fatemeh Mozaffari:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Data curation.

Declaration of competing interest

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

Acknowledgements

The authors would like to thank Dr. Keyvan Kohgir and Dr. Taha Mansouri for their assistance with topic labeling and developing the classification model, respectively.

References

[1] Zhou Q, Zhang C. Breaking community boundary: comparing academic and social communication preferences regarding global pandemics. J Informetr 2021;15(3):101162.
[2] Nicola M, Alsafi Z, Sohrabi C, Kerwan A, Al-Jabir A, Iosifidis C, et al. The socio-economic implications of the coronavirus pandemic (COVID-19): a review. Int J Surg 2020;78:185–93.
[3] Forni G, Mantovani A. COVID-19 vaccines: where we stand and challenges ahead. Cell Death Differ 2021;28(2):626–39.
[4] Badiani AA, Patel JA, Ziolkowski K, Nielsen FBH. Pfizer: the miracle vaccine for COVID-19? Public Heal Pract 2020;1:100061.
[5] Chen WH, Strych U, Hotez PJ, Bottazzi ME. The SARS-CoV-2 vaccine pipeline: an overview. Curr Trop Med reports 2020;7(2):61–4.
[6] Myint A, Jones T. Possible method for the production of a Covid-19 vaccine. Vet Rec 2020;186(12):388.
[7] Thelwall M, Kousha K, Thelwall S. Covid-19 vaccine hesitancy on English-language Twitter. Prof Inferm 2021;30(2).
[8] Li Y, Tenchov R, Smoot J, Liu C, Watkins S, Zhou Q. A comprehensive review of the global efforts on COVID-19 vaccine development. ACS Cent Sci 2021;7(4):512–33.
[9] Keikhosrokiani P, Pourya Asl M. Handbook of research on opinion mining and text analytics on literary works and social media. IGI Global; 2022.
[10] Liu S, Liu J. Public attitudes toward COVID-19 vaccines on English-language Twitter: a sentiment analysis. Vaccine 2021;39(39):5499–505.
[11] Lyu JC, Le Han E, Luli GK. COVID-19 vaccine–related discussion on Twitter: topic modeling and sentiment analysis. J Med Internet Res 2021;23(6):e24435.
[12] Melton CA, Olusanya OA, Ammar N, Shaban-Nejad A. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: a call to action for strengthening vaccine confidence. J Infect Public Health 2021;14(10):1505–12.
[13] Marcec R, Likic R. Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgrad Med J 2022;98(1161):544–50.
[14] Luo C, Ji K, Tang Y, Du Z. Exploring the expression differences between professionals and laypeople toward the COVID-19 vaccine: text mining approach. J Med Internet Res 2021;23(8):e30715.
[15] Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J Med Internet Res 2020;22(4):e19016.
[16] Ogbuokiri B, Ahmadi A, Nia ZM, Mellado B, Wu J, Orbinski J, et al. Vaccine hesitancy hotspots in Africa: an insight from geotagged Twitter posts. IEEE Trans Comput Soc Syst 2023;11(1):1325–38.
[17] Stewart DW, Kamins MA. Secondary research: information sources and methodsvol. 4. Sage; 1993.
[18] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3(Jan):993–1022.
[19] Nagy A, Alhatlani B. An overview of current COVID-19 vaccine platforms. Comput Struct Biotechnol J 2021;19:2508–17.
[20] Malik YS, Ansari MI, Ganesh B, Sircar S, Bhat S, Pande T, et al. BCG vaccine: a hope to control COVID-19 pandemic amid crisis. Hum Vaccin Immunother 2020;16(12):2954–62.
[21] Lurie N, Saville M, Hatchett R, Halton J. Developing Covid-19 vaccines at pandemic speed. N Engl J Med 2020;382(21):1969–73.
[22] Mullard A. COVID-19 vaccine development pipeline gears up. Lancet 2020;395(10239):1751–2.
[23] Soleimanpour S, Yaghoubi A. COVID-19 vaccine: where are we now and where should we go? Expert Rev Vaccines 2021;20(1):23–44.
[24] Zheng C, Shao W, Chen X, Zhang B, Wang G, Zhang W. Real-world effectiveness of COVID-19 vaccines: a literature review and meta-analysis. Int J Infect Dis 2022;114:252–60.
[25] Di Fusco M, Lin J, Vaghela S, Lingohr-Smith M, Nguyen JL, Scassellati Sforzolini T, et al. COVID-19 vaccine effectiveness among immunocompromised populations: a targeted literature review of real-world studies. Expert Rev Vaccines 2022;21(4):435–51.
[26] Saleh AM, Khalid A, Alshaya AK, Alanazi SMM. Systemic lupus erythematosus with acute pancreatitis and vasculitic rash following COVID-19 vaccine: a case report and literature review. Clin Rheumatol 2022;1–6.
[27] Hekmat M, Jafari Naeini S, Abbasi Z, Dadkhahfar S. Drug-induced vasculitis: thiazide or the COVID-19 vaccine, which one is guilty? A case report and literature review. Clin Case Rep 2022;10(6):e5978.
[28] So C, Izumi S, Ishida A, Hirakawa R, Kusaba Y, Hashimoto M, et al. COVID-19 mRNA vaccine-related interstitial lung disease: two case reports and literature review. Respirol Case Rep 2022;10(4):e0938.
[29] Ben Saida I, Maatouk I, Toumi R, Bouslama E, Ben Ismail H, Ben Salem C, et al. Acquired thrombotic thrombocytopenic Purpura following inactivated COVID-19 vaccines: two case reports and a short literature review. Vaccines 2022;10(7):1012.
[30] MacDonald NE. Vaccine hesitancy: definition, scope and determinants. Vaccine 2015;33(34):4161–4.
[31] Sallam M. COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. Vaccines 2021;9(2):160.
[32] Koller D, Friedman N. Probabilistic graphical models: Principles and techniques. MIT press; 2009.
[33] Dueñas-Fernández R, Velásquez JD, L'Huillier G. Detecting trends on the web: a multidisciplinary approach. Inf Fusion 2014;20:129–35.
[34] Koltsova O, Koltcov S. Mapping the public agenda with topic modeling: the case of the Russian LiveJournal. Policy Internet 2013;5(2):207–27.
[35] Akter S, Bhattacharyya M, Wamba SF, Aditya S. How does social media analytics create value? J Organ End User Comput 2016;28(3):1–9.
[36] Ogunleye B, Maswera T, Hirsch L, Gaudoin J, Brunsdon T. Comparison of topic modelling approaches in the banking context. Appl Sci 2023;13(2):797.
[37] Shah AM, Yan X, Qayyum A, Naqvi RA, Shah SJ. Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: machine learning approach. Int J Med Inform 2021;149:104434.
[38] Kim EHJ, Jeong YK, Kim Y, Kang KY, Song M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. J Inf Sci 2016;42(6):763–81.
[39] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 2001;42(1–2):177–96.
[40] Blei DM, Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning; 2006. p. 113–20.
[41] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical dirichlet processes. J Am Stat Assoc 2006;101(476):1566–81.
[42] Vayansky I, Kumar SAP. A review of topic modeling methods. Inf Syst 2020;94:101582.
[43] Joshi C, Attar VZ, Kalamkar SP. An unsupervised topic modeling approach for adverse drug reaction extraction and identification from natural language text. In: Advances in data and information sciences. Springer; 2022. p. 505–14.
[44] Wu Y, Liu M, Zheng WJ, Zhao Z, Xu H. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. In: Biocomputing 2012. World Scientific; 2012. p. 422–33.
[45] de Campos LM, Fernandez-Luna JM, Huete JF, Redondo-Expósito L. LDA-based term profiles for expert finding in a political setting. J Intell Inf Syst 2021;56(3):529–59.
[46] Zirn C, Stuckenschmidt H. Multidimensional topic analysis in political texts. Data Knowl Eng 2014;90:38–53.
[47] Debnath R, Bardhan R, Reiner DM, Miller JR. Political, economic, social, technological, legal and environmental dimensions of electric vehicle adoption in the United States: a social-media interaction analysis. Renew Sustain Energy Rev 2021;152:111707.
[48] Matsumoto K, Ren F, Matsuoka M, Yoshida M, Kita K. Slang feature extraction by analysing topic change on social media. CAAI Trans Intell Technol 2019;4(1):64–71.
[49] Gupta RK, Agarwalla R, Naik BH, Evuri JR, Thapa A, Singh TD. Prediction of research trends using LDA based topic modeling. Glob Transitions Proc 2022;3(1):298–304.
[50] Rouhani S, Mozaffari F. Sentiment analysis researches story narrated by topic modeling approach. Soc Sci Humanit Open 2022;6(1):100309.
[51] Praveen SV, Ittamalla R, Deepak G. Analyzing the attitude of Indian citizens towards COVID-19 vaccine– a text analytics study. Diabetes Metab Syndr Clin Res Rev 2021;15(2):595–9.
[52] Zulfiker MS, Kabir N, Biswas AA, Zulfiker S, Uddin MS. Analyzing the public sentiment on COVID-19 vaccination in social media: Bangladesh context. Array 2022;15:100204.
[53] Xu H, Liu R, Luo Z, Xu M. COVID-19 vaccine sensing: sentiment analysis and subject distillation from twitter data. Telemat Informatics Rep. 2022;8:100016.
[54] Liu B. Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 2012;5(1):1–167.
[55] Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Syst 2015;89:14–46.
[56] Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. Knowl Inf Syst 2019:1–47.
[57] Grljević O, Bošnjak Z, Kovačević A. Opinion mining in higher education: a corpus-based approach. Enterp Inf Syst 2022;16(5):1773542.
[58] Abo MEM, Raj RG, Qazi A. A review on Arabic sentiment analysis: state-of-the-art, taxonomy and open research challenges. IEEE Access 2019;7:162008–24.
[59] Diakopoulos NA, Shamma DA. Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the SIGCHI conference on human factors in computing systems; 2010. p. 1195–8.

[60] Jaidka K, Ahmed S, Skoric M, Hilbert M. Predicting elections from social media: a three-country, three-method comparative study. Asian J Commun 2019;29(3): 252–73.

[61] Rouhani S, Abedin E. Crypto-currencies narrated on tweets: a sentiment analysis approach. Int J Ethics Syst 2019;36(1):58–72.

[62] Smailović J, Grčar M, Lavrač N, Žnidaršič M. Predictive sentiment analysis of tweets: a stock market application. In: International workshop on human-computer interaction and knowledge discovery in complex, unstructured, big data. Springer; 2013. p. 77–88.

[63] Abualigah L, Alfar HE, Shehab M, Hussein AMA. Sentiment analysis in healthcare: a brief review. Recent Adv NLP Case Arab Lang 2020:129–41.

[64] Kumar S, Prabha R, Samuel S. Sentiment analysis and emotion detection with healthcare perspective. In: Augmented intelligence in healthcare: a pragmatic and integrated analysis. Springer; 2022. p. 189–204.

[65] Nguyen A, Pellerin R, Lamouri S, Lekens B. Managing demand volatility of pharmaceutical products in times of disruption through news sentiment analysis. Int J Prod Res 2022;1–12.

[66] Rodrigues RG, das Dores RM, Camilo-Junior CG, Rosa TC. SentiHealth-Cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. Int J Med Inform 2016;85(1):80–95.

[67] Xu H, Liu R, Luo Z, Xu M, Wang B. COVID-19 vaccine sensing: Sentiment analysis from Twitter data. In: In: 2021 IEEE international conference on systems, man, and cybernetics (SMC). IEEE; 2021. p. 3200–5.

[68] Ogbuokiri B, Ahmadi A, Bragazzi NL, Movahedi Nia Z, Mellado B, Wu J, et al. Public sentiments toward COVID-19 vaccines in South African cities: an analysis of Twitter posts. Front Public Heal 2022;10:987376.

[69] de Melo T, Figueiredo CMS. Comparing news articles and tweets about COVID-19 in Brazil: sentiment analysis and topic modeling approach. JMIR Public Heal Surveill 2021;7(2):e24585.

[70] Mirończuk MM, Protasiewicz J. A recent overview of the state-of-the-art elements of text classification. Expert Syst Appl 2018;106:36–54.

[71] Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D. Text classification algorithms: a survey. Information 2019;10(4):150.

[72] Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P. Covid-19 vaccine hesitancy: text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Syst Appl. 2023;212:118715.

[73] Shamrat F, Chakraborty S, Imran MM, Muna JN, Billah MM, Das P, et al. Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indones J Electr Eng Comput Sci 2021; 23(1):463–70.

[74] To QG, To KG, Huynh VAN, Nguyen NTQ, Ngo DTN, Alley SJ, et al. Applying machine learning to identify anti-vaccination tweets during the COVID-19 pandemic. Int J Environ Res Public Health 2021;18(8):4069.

[75] Jayapermana R, Aradea A, Kurniati NI. Implementation of stacking ensemble classifier for multi-class classification of COVID-19 vaccines topics on Twitter. Sci J Informatics 2022;9(1):8–15.

[76] Bird S, Klein E, Loper E. Natural language processing with Python: Analyzing text with the natural language toolkit. O'Reilly Media, Inc.; 2009.

[77] Singh J, Gupta V. Text stemming: approaches, applications, and challenges. ACM Comput Surv 2016;49(3):1–46.

[78] Srinivasa-Desikan B. Natural language processing and computational linguistics: a practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd; 2018.

[79] Dahal B, Kumar SAP, Li Z. Topic modeling and sentiment analysis of global climate change tweets. Soc Netw Anal Min 2019;9(1):24.

[80] Bakharia A, Bruza P, Watters J, Narayan B, Sitbon L. Interactive topic modeling for aiding qualitative content analysis. In: Proceedings of the 2016 ACM on conference on human information interaction and retrieval; 2016. p. 213–22.

[81] Omar M, On BW, Lee I, Choi GS. LDA topics: representation and evaluation. J Inf Sci 2015;41(5):662–75.

[82] Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining; 2015. p. 399–408.

[83] Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci 2004;101 (Suppl. 1):5228–35.

[84] Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. In: BMC bioinformatics. Springer; 2015. p. 1–10.

[85] Mabey B. pyLDAvis documentation. 2021.

[86] Bonta V, Janardhan NKN. A comprehensive study on lexicon based approaches for sentiment analysis. Asian J Comput Sci Technol 2019;8(S2):1–6.

[87] Mcauliffe J, Blei D. Supervised topic models. Adv Neural Inf Process Syst 2007;20.

[88] Bengfort B, Bilbro R, Ojeda T. Applied text analysis with python: Enabling language-aware data products with machine learning. O'Reilly Media, Inc.; 2018.

[89] Bouma G. Normalized (pointwise) mutual information in collocation extraction. Proc GSCL 2009:31–40.

[90] Lau JH, Newman D, Baldwin T. Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th conference of the European chapter of the association for computational linguistics; 2014. p. 530–9.

[91] Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. Int J data Min Knowl Manag Process 2015;5(2):1.

[92] Hasan M, Rahman A, Karim M, Khan M, Islam S, Islam M. Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA). In: Proceedings of international conference on trends in computational and cognitive engineering. Springer; 2021. p. 341–54.

[93] Xiao L, Wang G, Zuo Y. Research on patent text classification based on word2vec and LSTM. In: 2018 11th international symposium on computational intelligence and design (ISCID). IEEE; 2018. p. 71–4.

[94] Luan Y, Lin S. Research on text classification based on CNN and LSTM. In: 2019 IEEE international conference on artificial intelligence and computer applications (ICAICA). IEEE; 2019. p. 352–5.

[95] Wang K, Zhang P, Su J. A text classification method based on the merge-LSTM-CNN model. In: Journal of Physics: Conference Series. IOP Publishing; 2020. p. 12110.

[96] Liang S, Zhu B, Zhang Y, Cheng S, Jin J. A double channel CNN-LSTM model for text classification. In: 2020 IEEE 22nd international conference on high performance computing and communications; IEEE 18th international conference on Smart City; IEEE 6th international conference on data science and systems (HPCC/SmartCity/DSS). IEEE; 2020. p. 1316–21.

[97] Ebadi A, Xi P, Tremblay S, Spencer B, Pall R, Wong A. Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. Scientometrics 2021;126(1):725–39.

[98] Älgå A, Eriksson O, Nordberg M. Analysis of scientific publications during the early phase of the COVID-19 pandemic: topic modeling study. J Med Internet Res 2020;22(11):e21559.

[99] Tran BX, Ha GH, Nguyen LH, Vu GT, Hoang MT, Le HT, et al. Studies of novel coronavirus disease 19 (COVID-19) pandemic: a global analysis of literature. Int J Environ Res Public Health 2020;17(11):4095.

[100] Cao Q, Cheng X, Liao S. A comparison study of topic modeling based literature analysis by using full texts and abstracts of scientific articles: a case of COVID-19 research. Libr Hi Tech 2022;41(2):543–69.

[101] Dornick C, Kumar A, Seidenberger S, Seidle E, Mukherjee P. Analysis of patterns and trends in COVID-19 research. Procedia Comput Sci 2021;185:302–10.

[102] Marra AR, Kobayashi T, Suzuki H, Alsuhaibani M, Schweizer ML, Diekema DJ, et al. The long-term effectiveness of coronavirus disease 2019 (COVID-19) vaccines: a systematic literature review and meta-analysis. Antimicrob Steward Healthc Epidemiol 2022;2(1).

[103] Noruzi A, Gholampour B, Gholampour S, Jafari S, Farshid R, Stanek A, et al. Current and future perspectives on the COVID-19 vaccine: a scientometric review. J Clin Med 2022;11(3):750.

[104] Yin H, Song X, Yang S, Li J. Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web 2022;25(3):1067–83.

[105] Agarwal A, Romine WL, Banerjee T. Leveraging natural language processing to understand public outlook towards the influenza vaccination. In: 2020 IEEE international conference on big data (big data). IEEE; 2020. p. 4981–7.

[106] Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res 2016;18(8):e6045.

[107] Pouliliou S, Nikolaidis C, Drosatos G. Current trends in cancer immunotherapy: a literature-mining analysis. Cancer Immunol Immunother 2020;69(12):2425–39.