Course Project Report

Exploring COVID Scientometrics Data: Topic Modeling and Query-Driven Recommendations

Submitted By

Prathipati Jayanth - 211AI027
A D Mahit Nandan - 211AI001

as part of the requirements of the course

Social Computing (IT480) [Dec 2023 - Apr 2024]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Artificial Intelligence

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

DEC 2023 - APR 2024

DEPARTMENT OF INFORMATION TECHNOLOGY

National Institute of Technology Karnataka, Surathkal

CERTIFICATE

This is to certify that the Course project Work Report entitled **‘Exploring COVID Scientometrics Data: Topic Modeling and Query-Driven Recommendations ”** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
Prathipati Jayanth	2110504	Jayanth P 15/04/2024
A D Mahit Nandan	2110160	Mahit 15-04-2024

this report is a record of the work carried out by them as part of the course **Social Computing (IT480)** during the semester **Jan - Apr 2024**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Artificial Intelligence.**

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S

DECLARATION

We hereby declare that the project report entitled **"Exploring COVID Scientometrics Data: Topic Modeling and Query-Driven Recommendations System"** submitted by us for the course **Social Computing (IT480)** during the semester **Jan - Apr 2024**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Prathipati Jayanth	2110504	Jayanth P 15/04/2024
2. A D Mahit Nandan	2110160	Mahit 15-04-2024

Place: NITK, Surathkal
Date: 15 *th April* 2024

Exploring COVID Scientometrics Data: Topic Modeling and Query-Driven Recommendations

Prathipati Jayanth [1], A D Mahit Nandan[2],

Abstract—**This project delves into scientometrics data using the COVID-19 Open Research Dataset (CORD-19). Employing a diverse array of topic modeling techniques including Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), BERTopic, Non-negative Matrix Factorization (NMF), Probabilistic Latent Semantic Analysis (pLSA), the Contextualized Topic Model (CTM), and Top2Vec. Additionally, a SciBERT-based recommendation system is developed to provide query-driven access to relevant information within the dataset. Furthermore, extensive network analysis is conducted on author-author networks, examining properties such as average degree, degree distribution, clustering coefficient, connected components, and the giant component. The study integrates hierarchical clustering, K-means clustering, and DBSCAN clustering algorithms to cluster papers based on their abstracts, offering further insights into the thematic composition of the COVID-19 literature. This comprehensive approach contributes to a nuanced understanding and exploration of COVID-19 research within the scientometrics domain.**

*Keywords:***Author-Author Relation, Article Recommendation, Clustering, Network Analysis, Scientometrics, , Topic Modelling**

https://www.overleaf.com/7252243825phgthwhtnqrnc99538 to Overleaf project: **paste link here**

I. Introduction

The COVID-19 pandemic has spurred an unprecedented surge in scientific research, resulting in a vast corpus of scholarly literature aimed at understanding various facets of the disease, its transmission, treatments, and societal impacts. The volume and diversity of academic literature continue to expand exponentially, exemplified by the fact that as of March 22, 2021, the Open Research Dataset 1 contained a staggering 497,906 research articles focused on COVID-19 [2]. This surge has underscored the importance of scientometrics—the quantitative study of scientific publications and citations—in comprehensively analyzing and understanding scientific research. Studying these research papers reveals evolving scientific trends and helps identify key contributors.

Various analytical techniques and methodologies can be leveraged to extract actionable insights from COVID-19 scientometrics data. Topic modeling techniques, such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and BERTopic, offer effective means to identify latent themes and clusters within the COVID-19 literature. Additionally, advanced natural language processing (NLP) techniques, including Named Entity Recognition (NER) and sentiment analysis, can provide deeper insights into the content and context of COVID-19 research papers.

To address these challenges, we aim to accomplish the following objectives:

- Conduct a thorough Exploratory Data Analysis (EDA) by identifying languages, author-author relation graphs, Title text analysis and word cloud showing important words.
- Clustering documents using existing models based on abstract and title.
- Experimenting on different topic modelling models and recent large language models like bert-topic.
- Developing a recommendation system using sci-bert LLM which gives a relevant text from scientific documents based on given query.

The paper is structured as follows: Section II gives a overview of past existing research. Section III describes the dataset used. In Section IV, we analyze scientometric documents. Section V describes our proposed methodology. Section VI discusses our experiments and results. Finally, Section VII wraps up the paper and suggests future research directions.

II. Literature Survey

Liu et.al in [3] explores the growing field of early scientific research on COVID-19, particularly focusing on advancements propelled by artificial intelligence (AI). It employs the Latent Dirichlet Allocation (LDA) model to categorize research articles into 50 key topics pertinent to COVID-19, as discerned from their abstracts. Through this analysis, the paper provides a comprehensive overview of various dimensions of early COVID-19 studies, including aspects such as referencing and citation behavior, the diversity of research topics, and their interrelations.

Pivetta et.al in [4] introduces a novel tool designed to assist researchers in navigating the extensive collection of COVID-19 literature. Leveraging Information Retrieval (IR) and Information Extraction (IE) techniques applied to the Open Research Dataset, the tool enhances search capabilities for COVID-19-related papers. By employing Latent Dirichlet Allocation (LDA) to model research topics and extracting relevant named entities, the tool enables automated, topic-based searches of scientific papers. While showing promise, the authors note the need for further refinement to improve accuracy and reliability.

Li et.al in [5] conducts a bibliometric analysis of articles published between 2000 and 2017 in prestigious academic journals indexed in SCIE, SSCI, and A&HCI to assess the development and trends of topic modeling studies. The analysis focuses on productive authors, countries, and institutions, while also examining thematic changes within the field of topic modeling. The results highlight China's prominent role in topic modeling research and demonstrate the technique's significance across various disciplines, including natural and formal sciences, as well as social sciences. The study identifies LDA, social networks, and text analysis as topics experiencing increasing popularity, while certain models like pLSA and applications such as topic detection are declining in popularity.

Li et.al in [6] explored topic modeling, a statistical method used in machine learning to identify latent ”topics” within a collection of documents. Specifically, the authors focus on latent Dirichlet allocation (LDA), a popular approach in topic modeling. They investigate methods, including LDA and its extensions, for clustering scientific publications into distinct groups based on their content. The study evaluates the effectiveness of these methods by analyzing whether papers from the same field are clustered together. Additionally, the paper discusses potential applications of text analysis in scientometrics, the quantitative study of scientific research output.

Yang et.al in [7] explores the recognition of scientific data citation, vital for promoting data sharing and facilitating scientific analysis. Comparing classical machine learning and deep learning models, the study demonstrates the effectiveness of BERT-based models, such as BioBERT and SciBERT, in this task. Using full-text scientific papers, annotated citation classifications form a dataset for analysis. The study concludes that the proposed methods enable automated identification and extraction of data citations, with full-text information significantly influencing recognition accuracy.

Gündoğan et.al in [8] introduces a novel journal recommendation system aimed at assisting researchers in selecting suitable journals for their articles. Unlike existing systems, this approach utilizes article content, including title, abstract, keywords, and references, without requiring user-specific information. By analyzing publications from the last three years, the system determines the scope of journals, providing comprehensive recommendations across multiple publishers. Using SBERT for sentence-level similarity, the system outperforms traditional methods like Word2vec, Glove, and FastText. Experimental results highlight the effectiveness of this approach in guiding researchers towards appropriate journal selections.

III. Dataset Description

The dataset used for this research [1] contains cord_id, sha, source_x, title, pmcid, pubmed, abstract, authors, licence and json file link to respective report pdf. It contains 970836 articles out of which 850367 articles has mentioned title. A custom file reader function is added to convert the dataset into useful dataframe containing doi, abstract, body text, authors, title, source. Language distribution for the reports is given in Fig.1 shows many articles published are in english.

IV. Exploratory Data Analysis

A. Author-Author Graph Analysis

In our exploration of COVID Scientometrics data, we've constructed a collaboration network graph to

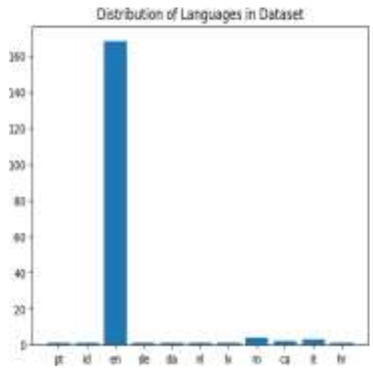


Fig. 1. Author-Author Graph

visualize scientific collaboration among researchers studying the pandemic. This graph maps authors as nodes and their collaborations as edges, offering a snapshot of the inter-connectedness within the COVID-19 research community. Through this visualization, we aim to uncover collaboration patterns, identify influential authors and communities, and understand how scientific collaboration evolves over time. This analysis contributes to our understanding of the collaborative dynamics shaping research efforts.

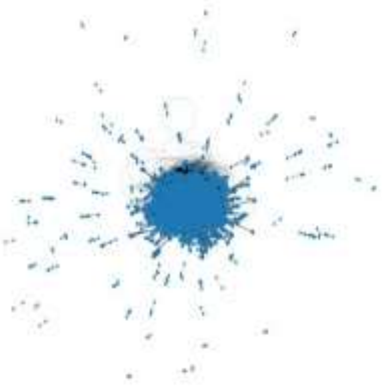


Fig. 2. Author-Author Graph

- Authors demonstrate an average collaboration degree of 2.0, indicating an average of two collaborations per author. This points to a moderate level of collaboration within the research community, underscoring the inter-connectedness among researchers
- In our analysis, we observed a total of 9083 nodes in the collaboration network, with the giant component size being 1. This suggests that the network lacks a significant interconnected structure. The coverage, calculated at approximately 0.00011, indicates a minimal proportion of the overall network, potentially reflecting isolated clusters or limited collaboration among authors.

B. Distribution of title length

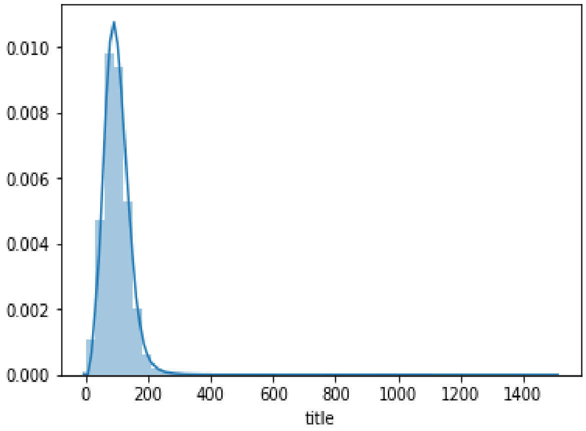


Fig. 3. Distribution of Title Length

The analysis of title length distribution revealed a clear trend. Most titles are relatively short, under 600 characters, with a few outliers being significantly longer. This suggests potential for categorizing titles by length for further analysis. However, it's important to note that length alone might not be a strong indicator of content. To gain a more complete picture, examining the distribution of word frequencies within the titles is recommended, which could reveal commonly used keywords and phrases.

C. Most Common Words in Title and Abstract

As part of exploratory data analysis (EDA), we conducted an investigation of the most prevalent words in both titles and abstracts within our scientometric dataset. To show these findings, a histogram where the y-axis represents the words and the x-axis denotes their correspondning frequencies of occurrence is plotted.

D. Most Common Bi-Grams and Tri-Grams in Title

In the project, the existence of bigrams and trigrams within the titles of the scientometric dataset is studied. This study tries find recurring phrases

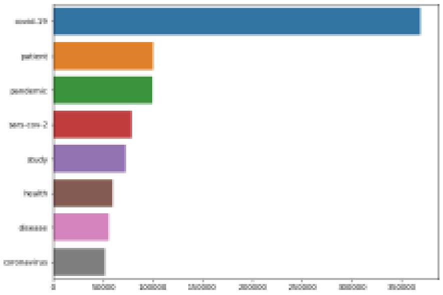


Fig. 4. Most Common words in Title

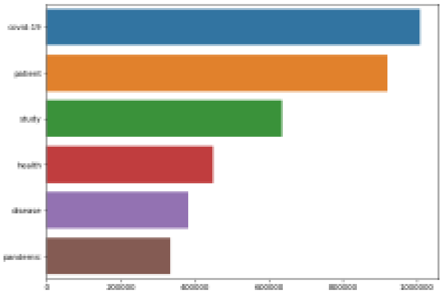


Fig. 5. Most Common words in Abstract

and associations indicative of important research topics and themes. Histograms were constructed to represent these findings, with the y-axis denoting the bigrams/trigrams and the x-axis illustrating their frequencies of occurrence. This exploration shows the nuanced language patterns in research titles, providing insights into key subject areas and trends within the literature.

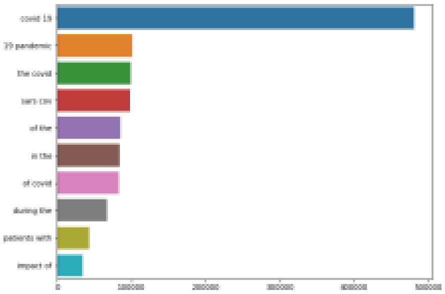


Fig. 6. Most Common bi-grams in Title

E. Word Cloud

A word cloud was generated to further illustrate the prevalence of terms within the titles and abstracts of the scientometric dataset. This graphical representation visually highlights the most frequently occurring words, with larger font sizes indicating higher frequencies as shown in 8.

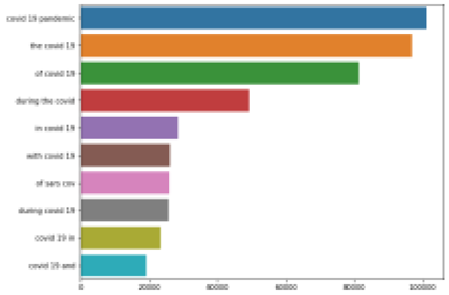


Fig. 7. Most Common tri-grams in Title

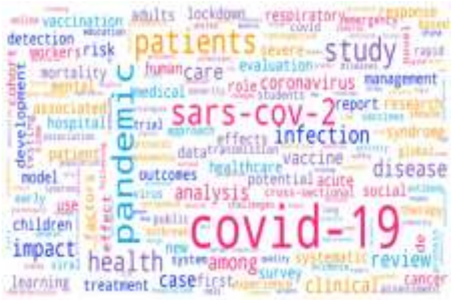


Fig. 8. Word Cloud

V. Methodology

The methodology employed in this project encompasses a comprehensive approach to analyzing the scientometric dataset. Our methodology consists of four components, each designed to find valuable information from the dataset. These components include author-author graph analysis, clustering papers according to concept, topic modeling, and a recommendation system as shown in 9

A. Clustering Papers according to Concept

1) K Means Clustering: Documents are grouped into K clusters by K-means based on how similar they are. Documents are represented as high-dimensional vectors (TF-IDF scores, for example), and K-means minimizes the distance to cluster centroids in order to iteratively assign documents to clusters. It works well with big datasets that have distinct clusters.

2) DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Clustering: Documents are grouped by DBSCAN according to their density distribution. It locates areas that are closely spaced and enlarges clusters by joining nearby points. DBSCAN is robust against noise and works well with datasets that have irregular clusters and varying densities.

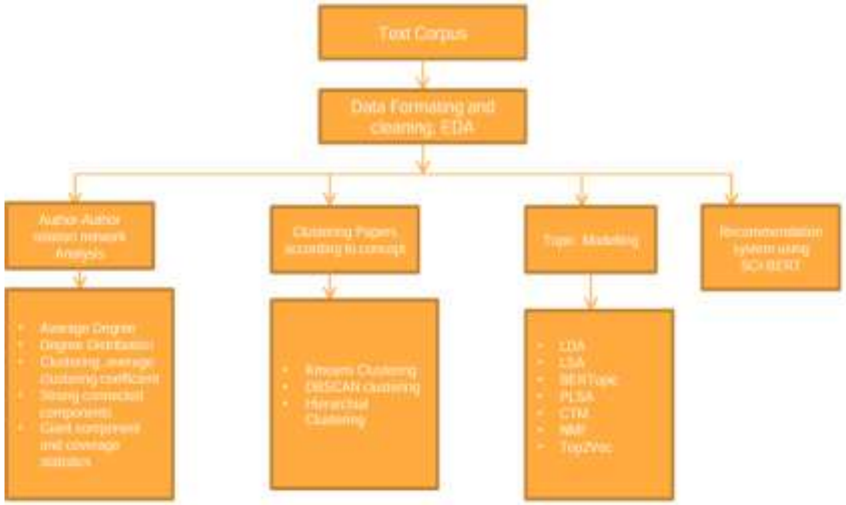

Text Corpus
Data Formating and cleaning, EDA
Author-Author citation network Analysis
Average Degree
Degree Distribution
Clustering, average clustering coefficient
Strong connected components
Giant component and coverage statistics
Clustering Papers according to concept
Kmeans Clustering
DBSCAN clustering
Hierarchial Clustering
Topic Modeling
LDA
LSA
BERTopic
PLSA
CTM
NMF
Top2Vec
Recommendation system using SCI BERT

Fig. 9. Proposed Methodology

3) Hierarchial Clustering: By recursively joining or dividing clusters, hierarchical clustering creates a hierarchy of clusters. Researchers can examine clusters at various granularities depicting the hierarchical structure of document similarities. Datasets with nested or hierarchical relationships between clusters are a good fit for hierarchical clustering.

B. Topic Modeling

1) LDA: Assuming that every document is a mixture of topics and that each topic is a distribution over words, LDA is a generative probabilistic model. Determining the underlying topic distribution and word distribution for every document is the aim of latent distribution analysis (LDA). Latent domain analysis (LDA) models the co-occurrence of words in documents to uncover latent topics and their frequency throughout the corpus. It is frequently used to find patterns and themes in textual data.

2) pLSA: The chance of observing words in documents given the latent topics is modeled by pLSA, a probabilistic variation of latent semantic analysis (LSA). The likelihood of word occurrences in documents is directly modeled by pLSA based on the topic distribution, in contrast to LDA, which introduces a prior distribution over subjects. By breaking down the document-term matrix into low-dimensional latent variables, pLSA is able to extract the corpus's semantic structure.

3) BertTopic: BertTopic extracts topic representations from text input by utilizing pre-trained BERT (Bidirectional Encoder Representations from Transformers) models. BertTopic is capable of learning dense representations of documents that capture semantic information by optimizing BERT embeddings on topic classification tasks. The next step is to cluster these representations in order to find latent topics in the corpus. BertTopic is able to capture complex links between words and their context in documents because of BERT's contextualized word embeddings.

4) CTM: Correlations between themes are incorporated into CTM, which is an extension of LDA. CTM uses a Gaussian distribution to model topic dependencies, in contrast to LDA, which makes the assumption that topics are distributed independently. By allowing subjects to share a similar correlation matrix, CTM is able to identify more logical and understandable topic structures and capture relationships. When it comes to discovering theme clusters within the data and capturing intricate links between topics, CTM is especially helpful.

5) NMF: In order to describe themes and their distributions over terms, a document-term matrix is broken down into non-negative matrices using the matrix factorization technique known as NMF. It makes themes and document representations interpretable by immediately identifying low-rank components, especially for high-dimensional and sparse text data.

6) Top2Vec: Expanding on Word2Vec, Topic2Vec is a neural network-based topic modeling technique that can capture semantic representations of themes. Topic2Vec builds dense vector representations of topics by training on word embeddings, which allows for more sophisticated topic modeling in text data.

C. Recommendation System

Our recommendation system utilizes SciBERT, a variant of BERT trained on scientific text, to extract dense representations of user queries and COVID-19 papers. By computing similarity scores between these embeddings using cosine similarity, the system efficiently ranks and recommends papers most relevant to the user query. This approach leverages SciBERT’s contextual understanding to provide accurate recommendations tailored to the user’s needs.

VI. Experiments and Results

A. Clustering Papers according to Concept

1) K Means Clustering: First the documents were converted to vectors using Tf-Idf. Optimal number of clusters were found using the elbow method. Here were are taking K as 19 as shown in Fig.10. To visualize these high dimensional vectors we used T-SNE to reduce the dimensions to 2D as shown in Fig.11

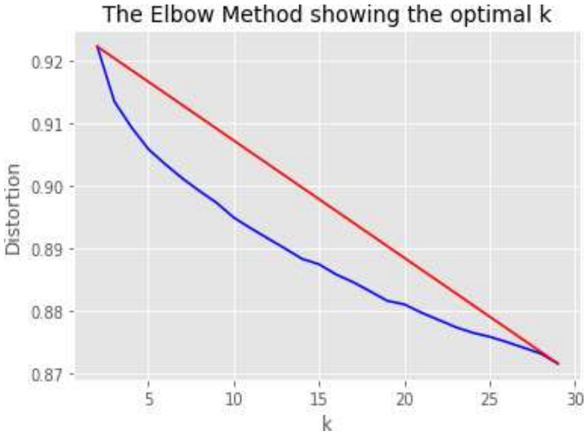


Fig. 10. Elbow Method

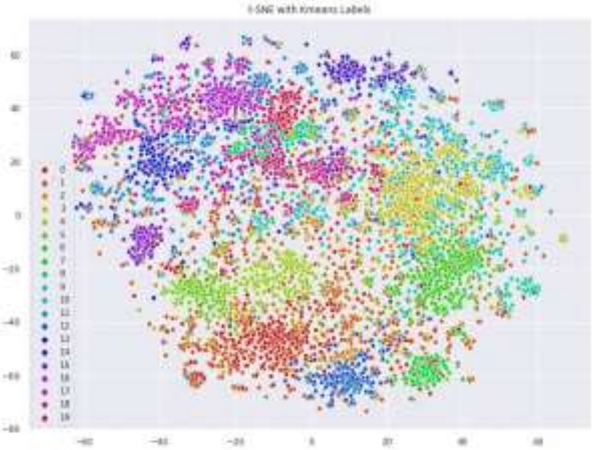


Fig. 11. T-SNE visualization after K-Means clustering

2) DBSCAN clustering: Documents are clustered based on combined text of abstract, title. English language model is loaded using ’en_core_web_sm’ module. Sentences are converted into vectors using this model which are passed into DBSCAN module with parameters as epsilon=0.08, min_samples=2 and cosine metric. 3 clusters are formed with labels as 0,-1,1. TSNE function is called with n_components=2, random_state=42 parameters for dimensionality reduction and 2D plotting as shown in Fig.12

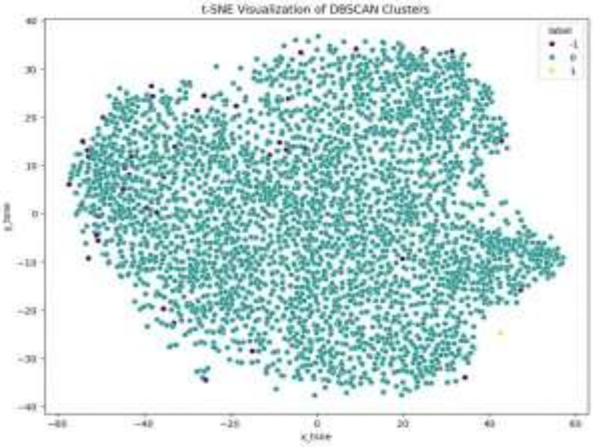


Fig. 12. T-SNE visualization after DBSCAN clustering

3) Hierarchical Clustering: Documents are clustered based on combined text of abstract, title. English language model is loaded using ’en_core_web_sm’ module. Sentences are converted into vectors using this model which are passed into AgglomerativeClustering module with distance_threshold=0.5 and ward linkage. Number of clusters need not be specified. Dendogram is plotted for hierarchial understanding with sentence vs euclidean distance in plots as shown in Fig.13

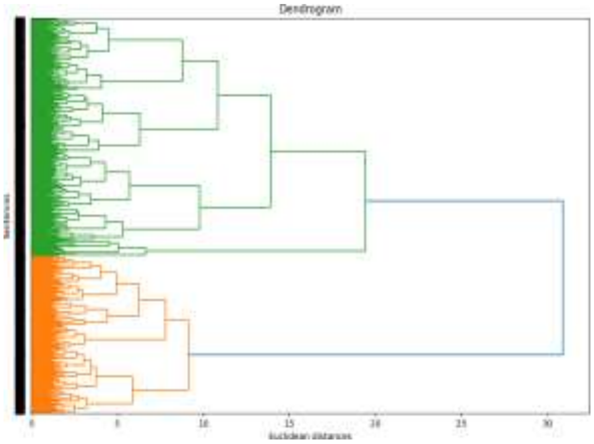
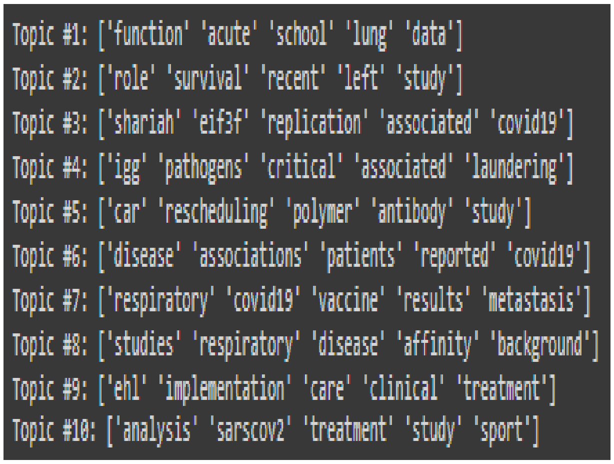


Fig. 13. Dendogram plot

B. Topic Modeling

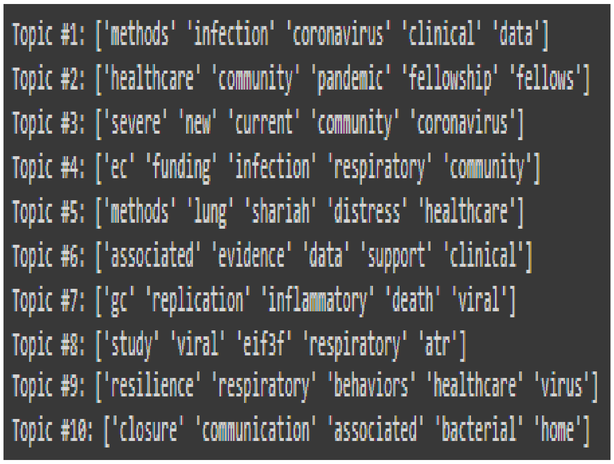
1) LDA: Preprocessed body text sentences are converted into text using Countvectorizer function after removing english and additional stopwords. LatentDirichletAllocation module is called with 10 components required and random state of 42. 10 topics are extracted as shown in Fig.14.



```
Topic #1: ['function' 'acute' 'school' 'lung' 'data']
Topic #2: ['role' 'survival' 'recent' 'left' 'study']
Topic #3: ['shariah' 'eif3f' 'replication' 'associated' 'covid19']
Topic #4: ['igg' 'pathogens' 'critical' 'associated' 'laundering']
Topic #5: ['car' 'rescheduling' 'polymer' 'antibody' 'study']
Topic #6: ['disease' 'associations' 'patients' 'reported' 'covid19']
Topic #7: ['respiratory' 'covid19' 'vaccine' 'results' 'metastasis']
Topic #8: ['studies' 'respiratory' 'disease' 'affinity' 'background']
Topic #9: ['ehl' 'implementation' 'care' 'clinical' 'treatment']
Topic #10: ['analysis' 'sarscov2' 'treatment' 'study' 'sport']
```

Fig. 14. LDA topics

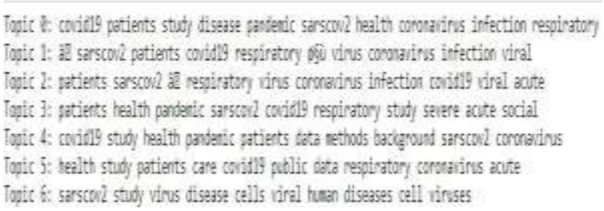
2) LSA: Preprocessed body text sentences are converted into text using Countvectorizer function after removing english and additional stopwords. TruncatedSVD module is called with 10 components required and random state of 42. 10 topics are extracted as shown in Fig.15.



```
Topic #1: ['methods' 'infection' 'coronavirus' 'clinical' 'data']
Topic #2: ['healthcare' 'community' 'pandemic' 'fellowship' 'fellows']
Topic #3: ['severe' 'new' 'current' 'community' 'coronavirus']
Topic #4: ['ec' 'funding' 'infection' 'respiratory' 'community']
Topic #5: ['methods' 'lung' 'shariah' 'distress' 'healthcare']
Topic #6: ['associated' 'evidence' 'data' 'support' 'clinical']
Topic #7: ['gc' 'replication' 'inflammatory' 'death' 'viral']
Topic #8: ['study' 'viral' 'eif3f' 'respiratory' 'atr']
Topic #9: ['resilience' 'respiratory' 'behaviors' 'healthcare' 'virus']
Topic #10: ['closure' 'communication' 'associated' 'bacterial' 'home']
```

Fig. 15. LSA topics

3) pLSA: Using Probabilistic Latent Semantic Analysis (pLSA), we were able to narrow down our dataset to seven user-defined categories. To aid in the rapid understanding of the thematic material of each topic, we visualized term distributions using bar charts. These charts provide visitors with a quick overview of the subject core of each topic as shown in Fig.17 and Fig.16.



```
Topic 0: covid19 patients study disease pandemic sarscov2 health coronavirus infection respiratory
Topic 1: â sarscov2 patients covid19 respiratory ðŸ virus coronavirus infection viral
Topic 2: patients sarscov2 â respiratory virus coronavirus infection covid19 viral acute
Topic 3: patients health pandemic sarscov2 covid19 respiratory study severe acute social
Topic 4: covid19 study health pandemic patients data methods background sarscov2 coronavirus
Topic 5: health study patients care covid19 public data respiratory coronavirus acute
Topic 6: sarscov2 study virus disease cells viral human diseases cell viruses
```

Fig. 16. Topics found using pLSA

4) BertTopic: After applying the pre-trained Bert-Topic Model on the dataset we cluster the entire scientometric dataset into 140 topics. We then plot the Intertopic Distance plot where we can see the keywords in each topic and how closely related and far apart they from each other as shown in Fig.18.

5) CTM: Our dataset was reduced to 7 themes after applying CTM, while users are free to choose how many topics to include. By capturing connections between subjects, CTM provides a detailed depiction of thematic structures. Similar to our pLSA technique, we used bar charts to visualize keyword distributions in each topic. Thematic substance of any topic can be quickly understood by comparing the keyword prevalence throughout these charts as shown in Fig.19 and Fig.20.

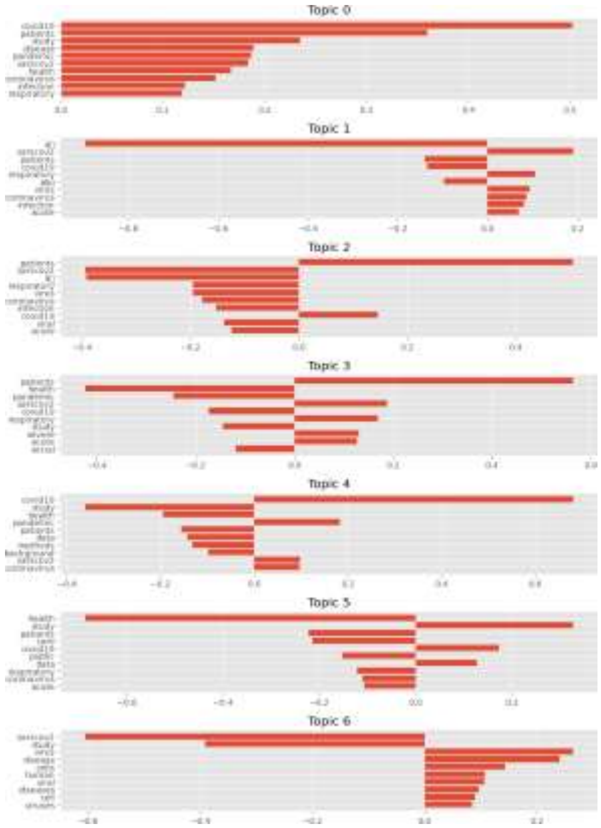


Fig. 17. Frequency Chart for each Topic found using pLSA

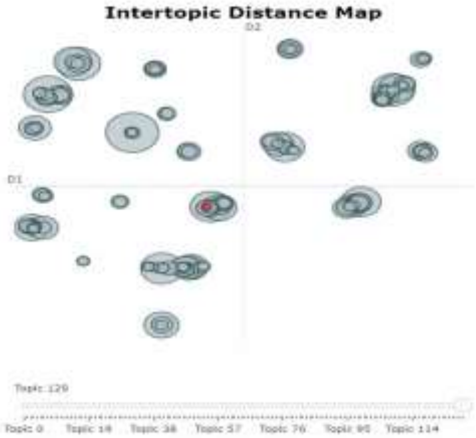
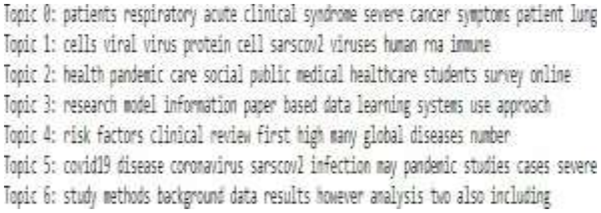


Fig. 18. Intertopic Distance Map using BertTopic

6) *NMF:* NMF is a topic modeling technique that uncovers latent topics in text data by decomposing non-negative matrices. Sentences are converted into vector form using Tf-idf embeddings. Using NMF module with num_components=5, we generate 5 relevant topics as shown in Fig.21.

7) *Top2Vec:* A list of body text is passed into top2vec module installed along with tensor-flow. get_topic function is called with the model which list out key words occured in different topics as shown in Fig.22 and Fig.23.



```
Topic 0: patients respiratory acute clinical syndrome severe cancer symptoms patient lung
Topic 1: cells viral virus protein cell sarscov2 viruses human rna immune
Topic 2: health pandemic care social public medical healthcare students survey online
Topic 3: research model information paper based data learning systems use approach
Topic 4: risk factors clinical review first high many global diseases number
Topic 5: covid19 disease coronavirus sarscov2 infection may pandemic studies cases severe
Topic 6: study methods background data results however analysis two also including
```

Fig. 19. Topics found using CTM

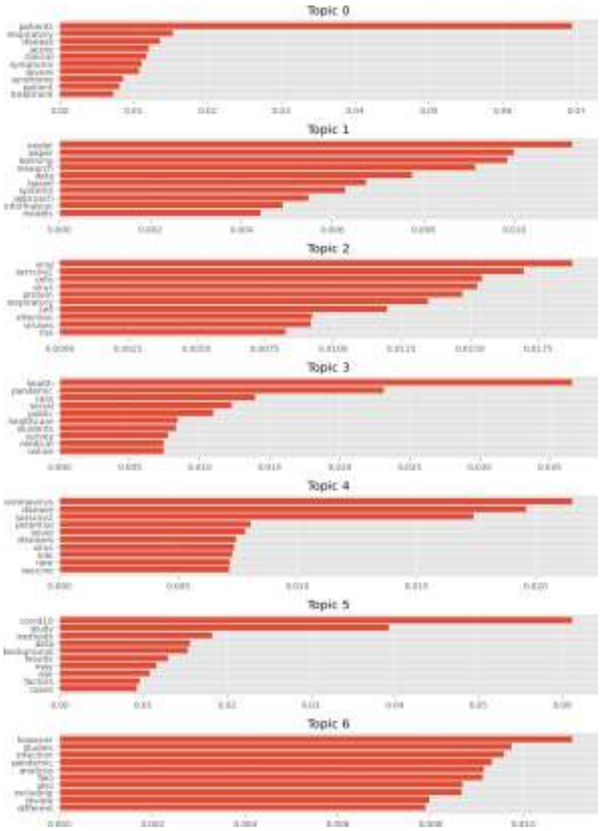
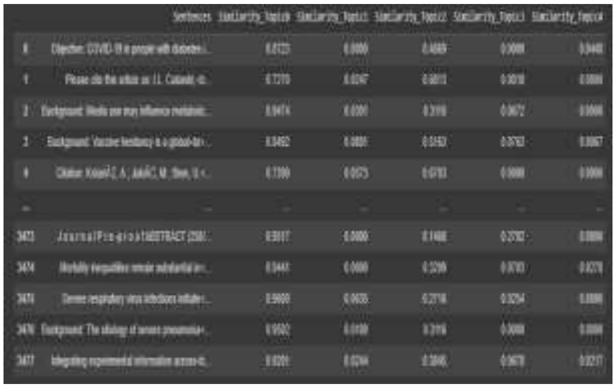


Fig. 20. Frequency Chart for each Topic found using CTM

TABLE I
METRICS FOR PROBABILISTIC MODELS

Model Name/Metric	Coherence Score	Perplexity
LDA	0.484	-9.25
pLSA	0.493	-9.36
CTM	0.623	-8.64

8) *Metrics for comparing models:* In probabilistic topic modeling like LDA, CTM and pLSA, coherence score and perplexity are used to assess topic interpretability and predictive accuracy, respectively. Coherence score measures semantic similarity within



	Sentences	Similarity_Topic0	Similarity_Topic1	Similarity_Topic2	Similarity_Topic3	Similarity_Topic4
0	Objective: COVID-19 in people with diabetes i...	0.2723	0.0000	0.4669	0.0000	0.0440
1	Please cite this article as: J.L. Cabanit, et...	0.7270	0.0247	0.6011	0.0018	0.0000
2	Background: Media use may influence metabolic...	0.9474	0.0391	0.3119	0.0672	0.0000
3	Background: Vaccine hesitancy is a global-tin...	0.3462	0.0801	0.0353	0.0763	0.0067
4	Citation: Kovač, A., Jukić, M., Stee, U....	0.7399	0.0575	0.0783	0.0000	0.0000
...	...	...	...	...	...	...
3473	Journal Pre-proof ABSTRACT (250...	0.3017	0.0000	0.1468	0.2782	0.0000
3474	Mortality inequalities remain substantial in ...	0.3441	0.0000	0.3299	0.0783	0.0278
3475	Severe respiratory virus infections initiate ...	0.5600	0.0635	0.2718	0.0254	0.0000
3476	Background: The etiology of severe pneumonia ...	0.9502	0.0100	0.3116	0.0000	0.0000
3477	Integrating experimental information across d...	0.0201	0.0264	0.3846	0.0678	0.0217

Fig. 21. NMF based similarity score from each topic

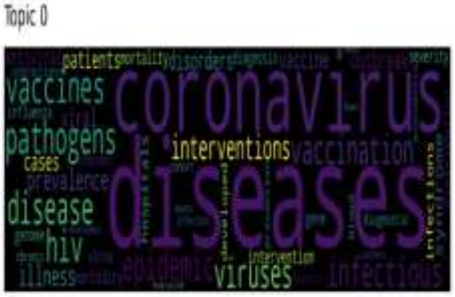


Fig. 22. Topic 0 Word-cloud in top2vec

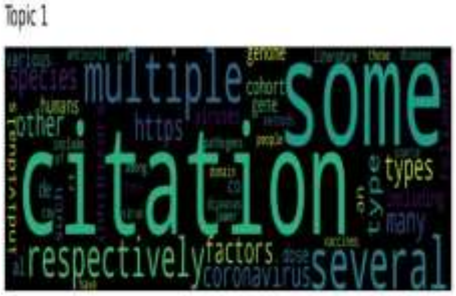


Fig. 23. Topic 1 Word-cloud in top2vec

topics, while lower perplexity indicates better predictive performance on held-out data as given in I. In non-probabilistic models such as LSA, NMF, BERTopic, Top2Vec, silhouette score evaluates clustering quality based on cluster compactness and separation as given in II. For LSA, CTM reconstruction error quantifies how well matrix factorization approximates the original data matrix as given in III, indicating model effectiveness in capturing semantic relationships. These metrics are employed to evaluate and refine topic models and dimensionality reduction techniques for text analysis tasks.

TABLE II
METRICS FOR NON-PROBABILISTIC MODELS

Model Name/Metric	Silhouette Score	Topic Coherence Score
LSA	0.356	-9.25
BERTopic	0.402	-8.55
Top2Vec	0.427	-7.57
NMF	0.410	-8.76

TABLE III
RECONSTRUCTION ERROR METRIC FOR MATRIX FACTORISATION MODELS

Model Name/Metric	Coherence Score
LSA	0.0049
NMF	0.0028

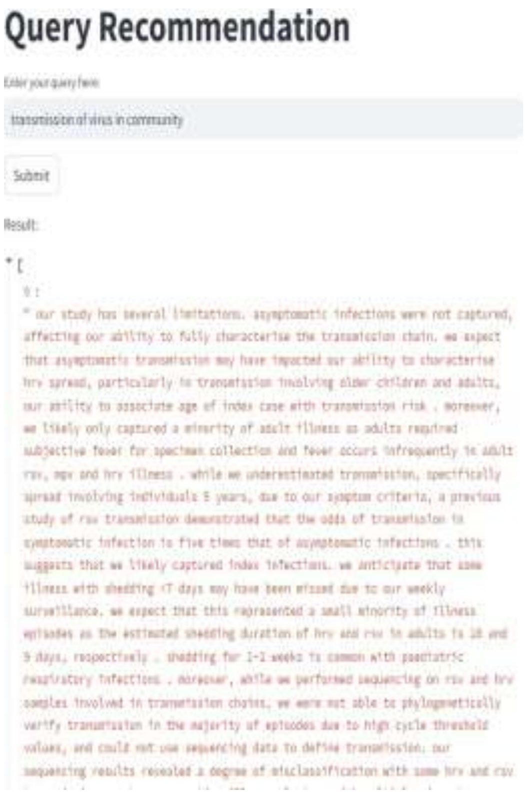
C. Recommendation System

To get sentence embeddings, We use scibert_scivocab_uncased model, a version of bert focuses on Scientific documents processing and training the sentences with the BertTokenizer with the same version. Cosine similarity is calculated between sentence and query embeddings. After sorting the list according to cosine similarity, the respective text in top is given as result as shown in Fig.24.

VII. CONCLUSION AND FUTURE SCOPE

In summary, this project has employed a variety of analytical tools, including topic modeling, network analysis, and recommendation systems, to offer insightful analysis of the COVID-19 research literature. We have improved knowledge of the academic landscape by examining thematic trends, author relations, and language distributions. The created recommendation engine, driven by SciBERT, makes it easier to find pertinent material quickly.

As a future this research can be extended by adding Incorporation of Multi-modal Data such as



Query Recommendation

```
▾ [
  0 :
  " our study has several limitations. asymptomatic infections were not captured,
  affecting our ability to fully characterise the transmission chain. we expect
  that asymptomatic transmission may have impacted our ability to characterise
  hrv spread, particularly in transmission involving older children and adults,
  our ability to associate age of index case with transmission risk . moreover,
  we likely only captured a minority of adult illness as adults required
  subjective fever for specimen collection and fever occurs infrequently in adult
  rsv, mpv and hrv illness . while we underestimated transmission, specifically
  spread involving individuals 5 years, due to our symptom criteria, a previous
  study of rsv transmission demonstrated that the odds of transmission in
  symptomatic infection is five times that of asymptomatic infections . this
  suggests that we likely captured index infections. we anticipate that some
  illness with shedding <7 days may have been missed due to our weekly
  surveillance. we expect that this represented a small minority of illness
  episodes as the estimated shedding duration of hrv and rsv in adults is 10 and
  9 days, respectively . shedding for 1-2 weeks is common with paediatric
  respiratory infections . moreover, while we performed sequencing on rsv and hrv
  samples involved in transmission chains, we were not able to phylogenetically
  verify transmission in the majority of episodes due to high cycle threshold
  values, and could not use sequencing data to define transmission. our
  sequencing results revealed a degree of misclassification with some hrv and rsv
```

Fig. 24. Recommendation system using SciBert

textual data provides insights into content, citation networks reveal research relationships, methodology graphs, author affiliations identify collaborations, publication metadata offers contextual information, and altimetrics gauge online impact. Semantic Search and Query Expansion, Interactive Visualization Tools, Integration with Knowledge Graphs, Evaluation of Impact and Influence.

References

[1] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge?select=metadata.csv

[2] Aletras, N. and Stevenson, M. (2013), “Evaluating topic coherence using distributional semantics”, Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers, pp. 13-22.

[3] Liu, Jiaying, Hansong Nie, Shihao Li, Xiangtai Chen, Huazhu Cao, Jing Ren, Ivan Lee, and Feng Xia. ”Tracing the pace of COVID-19 research: topic modeling and evolution.” Big Data Research 25 (2021): 100236.

[4] Pivetta, Marcos VL. ”An Information Retrieval and Extraction Tool for Covid-19 Related Papers.” arXiv preprint arXiv:2401.16430 (2024).

[5] Li, Xin, and Lei Lei. ”A bibliometric analysis of topic modelling studies (2000–2017).” Journal of Information Science 47, no. 2 (2021): 161-175.

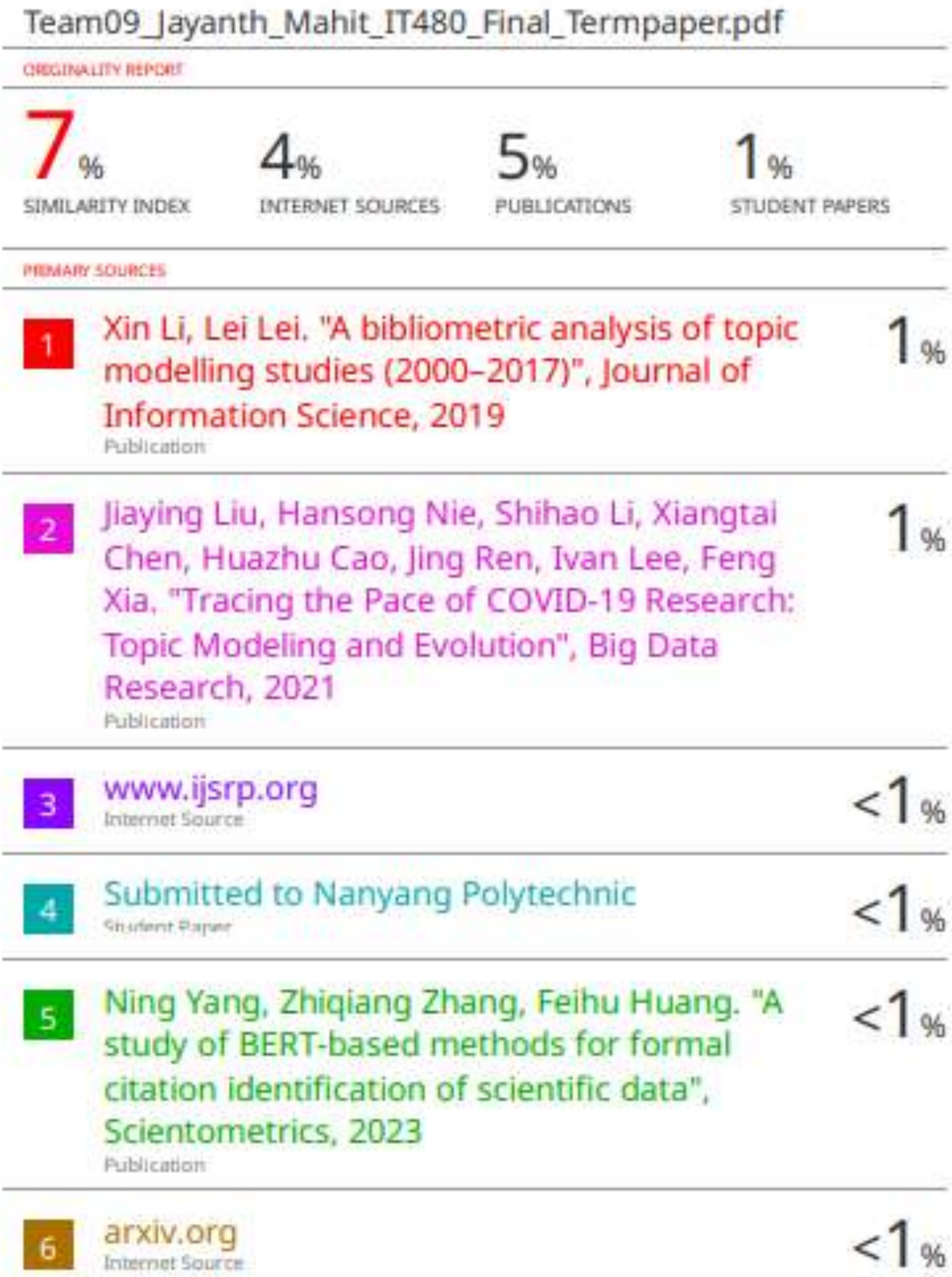
[6] Yau, Chyi-Kwei, Alan Porter, Nils Newman, and Arho Suominen. ”Clustering scientific documents with topic modeling.” Scientometrics 100 (2014): 767-786.

[7] Yang, Ning, Zhiqiang Zhang, and Feihu Huang. ”A study of BERT-based methods for formal citation identification of scientific data.” Scientometrics 128, no. 11 (2023): 5865-5881.

[8]G ündoğan, Esra, Mehmet Kaya, and Ali Daud. ”Deep learning for journal recommendation system of research papers.” Scientometrics 128, no. 1 (2023): 461-481.

APPENDIX

PLAGARISM REPORT



Team09_Jayanth_Mahit_IT480_Final_Termpaper.pdf

ORIGINALITY REPORT

7%	4%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Xin Li, Lei Lei. "A bibliometric analysis of topic modelling studies (2000–2017)", Journal of Information Science, 2019 Publication	1%
2	Jiaying Liu, Hansong Nie, Shihao Li, Xiangtai Chen, Huazhu Cao, Jing Ren, Ivan Lee, Feng Xia. "Tracing the Pace of COVID-19 Research: Topic Modeling and Evolution", Big Data Research, 2021 Publication	1%
3	www.ijsrp.org Internet Source	<1%
4	Submitted to Nanyang Polytechnic Student Paper	<1%
5	Ning Yang, Zhiqiang Zhang, Feihu Huang. "A study of BERT-based methods for formal citation identification of scientific data", Scientometrics, 2023 Publication	<1%
6	arxiv.org Internet Source	<1%