Topic Modeling of COVID-19 Research

Patrico Tyrell, MPH

Capstone Project 3

Springboard Data Science career track

February 2024

Mentor: Upon Malik

1. Introduction

The COVID-19 pandemic has generated an immense volume of scholarly articles, presenting a challenge for researchers and policymakers to stay abreast of evolving trends. This project applies Natural Language Processing (NLP) techniques for topic modeling and trend analysis to extract valuable insights from the COVID-19 Open Research Dataset (CORD-19). The purpose of the project was to use NLP to clearly define and interpret topics derived from scientific literature.

2. Problem

The problem addressed in this project lies within the vast and intricate landscape of scientific literature surrounding COVID-19. The objective was to distill key themes and topics from this extensive corpus, leveraging natural language processing techniques. By employing Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), the study aimed to uncover latent structures within the text data, providing a structured representation of the diverse research dimensions related to COVID-19.

3. Data

The project utilized the COVID-19 Open Research Dataset (CORD-19) published on Kaggle. This was prepared by the White House and a coalition of leading research groups. The CORD-19 is a resource of over 1,000,000 scholarly articles, including over 400,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses.

3. Data Wrangling

In the process of preparing the dataset for analysis, a series of data wrangling steps were performed on a collection of research papers related to the COVID-19 pandemic. The key steps in the data wrangling process include:

a. *Merging Metadata and JSON Files*:
    i. Metadata from CSV files was merged with the content of JSON files.
    ii. A random sample of 10,000 papers was chosen from the merged dataset.

b. *Exclusion of Invalid and Missing Papers*:
    i. Papers with invalid formats and those lacking metadata or having empty abstract/body text were excluded from the sample.
    ii. The resulting dataset comprised 8,656 papers.

c. *Language Identification and Filtering*:
   i. A language column was created to identify the language of each paper.
   ii. Of the 8,656 papers, 8,465 were identified as being in English, while non-English papers were dropped.

d. *Text Cleaning*:

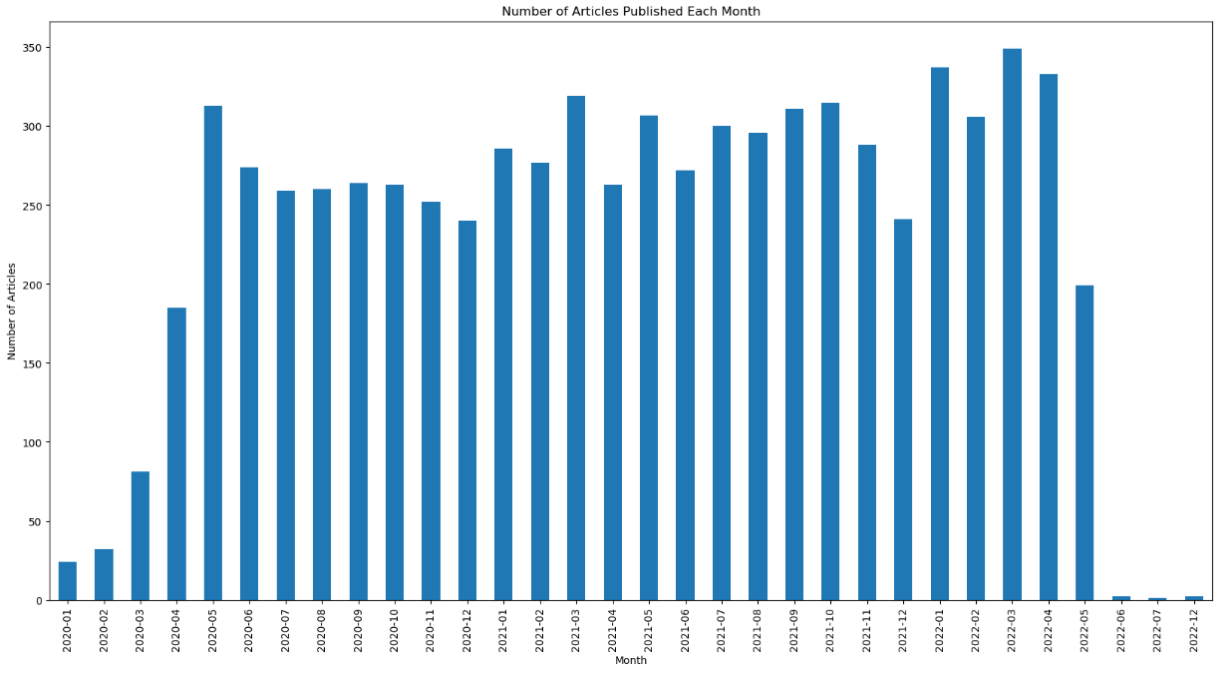
Both abstract and body text underwent a comprehensive cleaning process, including:

i. Removal of special characters, numbers, and punctuation using regular expressions.
ii. Conversion of text to lowercase.
iii. Tokenization of text into words.
iv. Removal of stop words using NLTK's stop words.
v. Application of stemming using the Porter Stemmer.

3. Exploratory Data Analysis

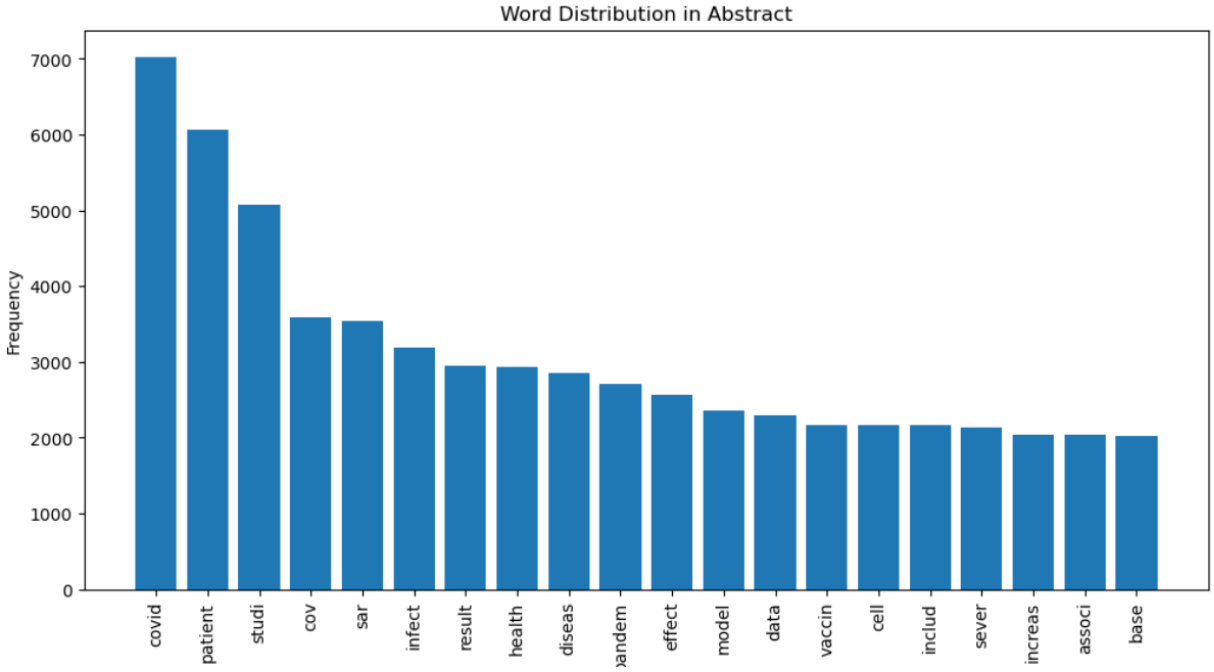
In the early months of 2020, the publication trend was relatively modest, with less than 50 papers published monthly from January to February. However, I observed a notable increase in March, reaching nearly 200 publications, and a substantial surge in April, coinciding with the global lockdown, with over 300 publications. The mean number of publications per month was 232, peaking at 349 in March 2022, reflecting the dynamic nature of COVID-19 research over time (Figure 1).

Figure 1. Number of Publications Monthly 2020-2022



Analyzing the word frequency within the abstract highlights that the terms most utilized are "covid," "patient," "study," "SARS," "infection," "results," "health," "disease," "pandemic," "effect," "model," "data," "vaccine," "cell," "severe," "increase," "association,". These terms suggest a wide-ranging research focus that span from the virus itself (SARS-CoV-2) to clinical and patient-oriented research, impact on Public Health Systems, and the role of data and modeling (Figure 2).

Figure 2. Word Distribution in Abstracts



Analyzing the word clouds from COVID-19 research publications between 2020 and 2022 reveals evolving focal points in the scientific community's response to the pandemic. In 2020, the emphasis was on immediate concerns related to the disease, with terms like "patient," "SARS," "COVID," and "study" being prominently featured, highlighting the initial efforts to understand and manage the virus's impact (Figure 3). The year 2021 marked a shift towards vaccine development and deployment, as evidenced by the emergence and rising prominence of terms such as "vaccine" (Figure 4). This trend continued into 2022, with "vaccine" becoming even more central, reflecting the global emphasis on vaccination as a key strategy in combating the pandemic (Figure 5).

Figure 3. Word Cloud for Abstracts in 2020

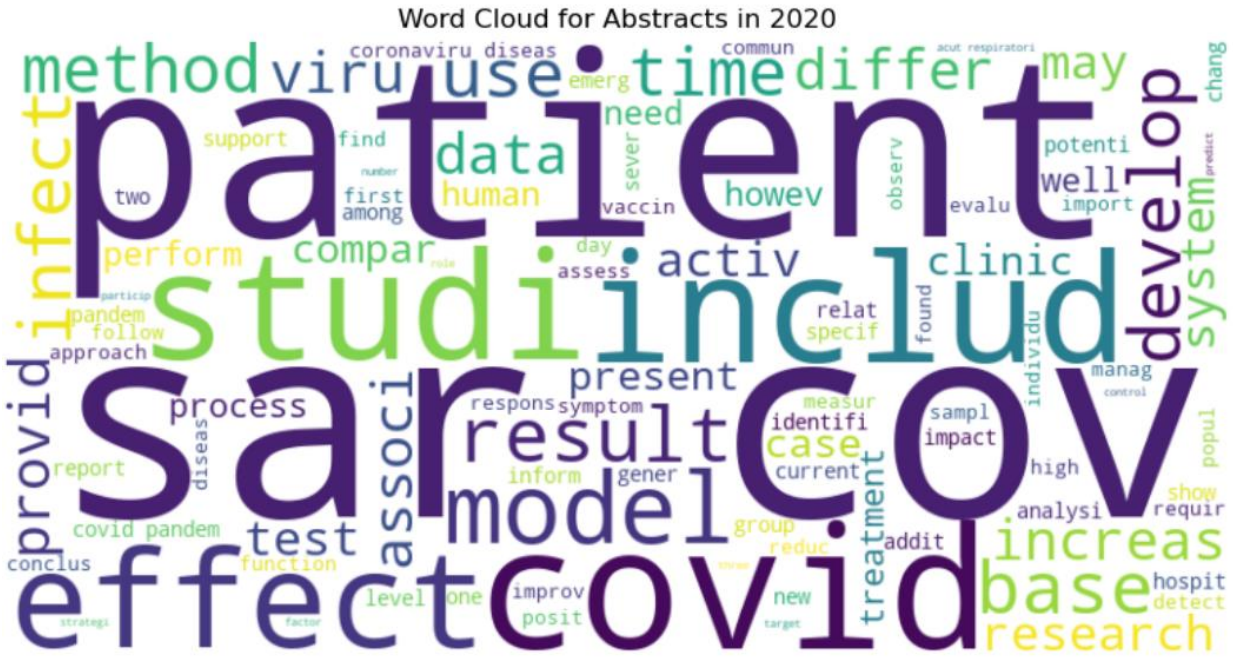


Figure 4. Word Cloud for Abstracts in 2021

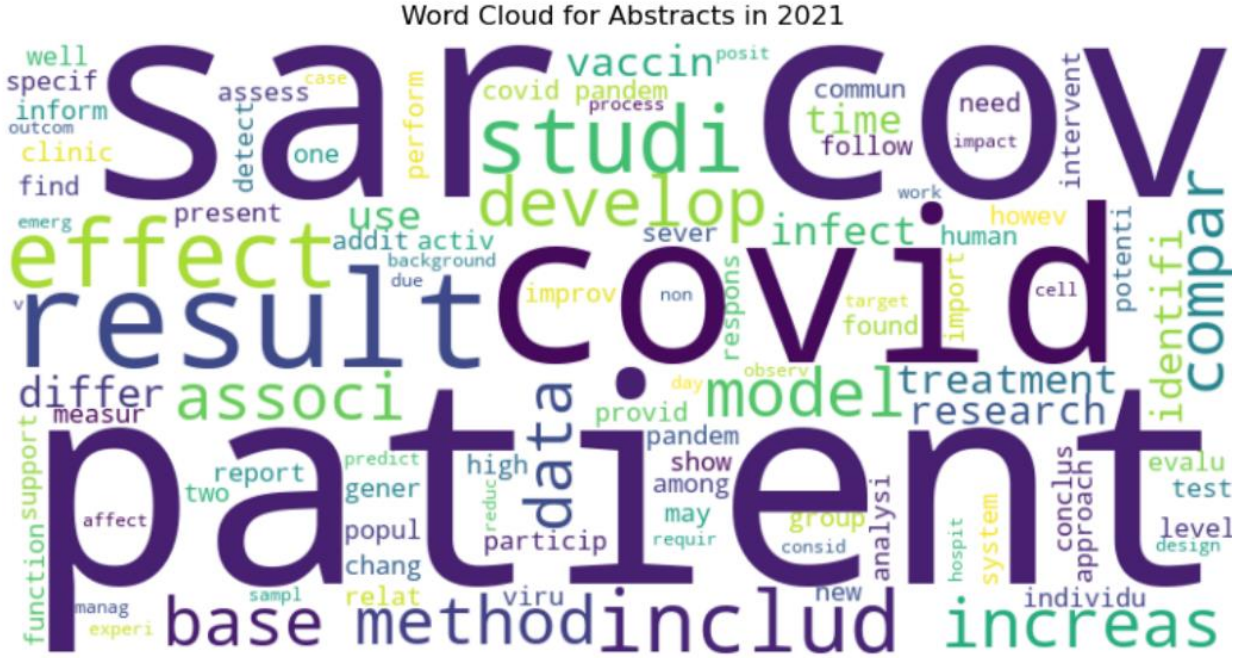
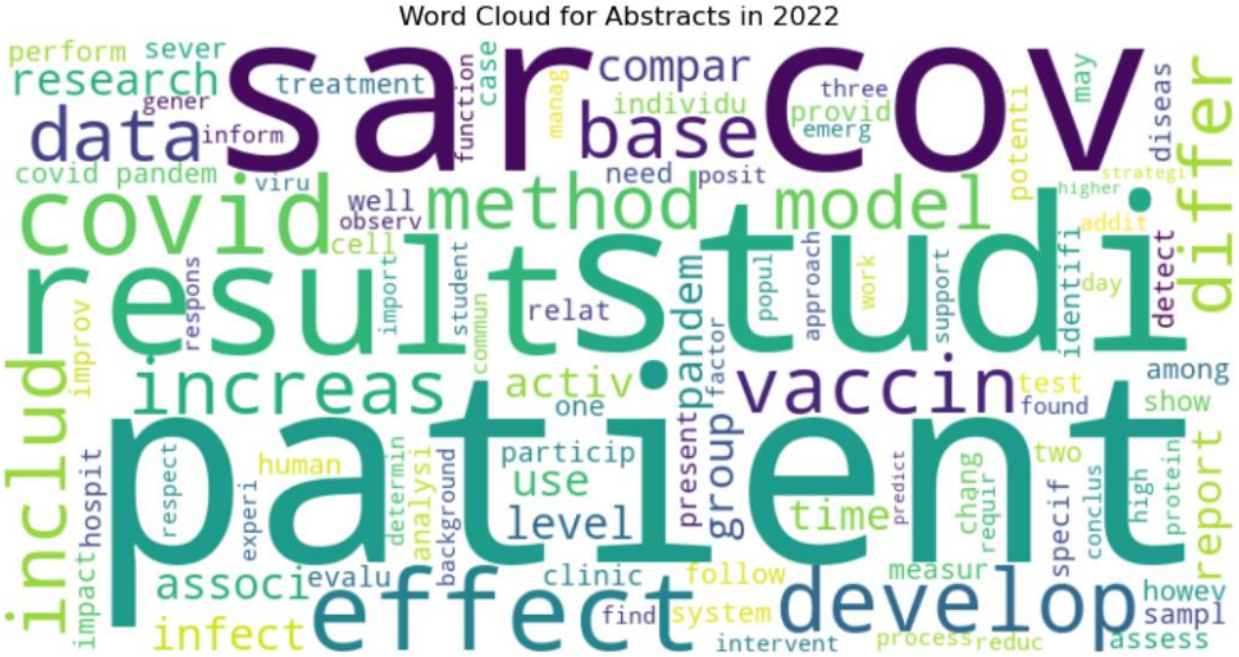


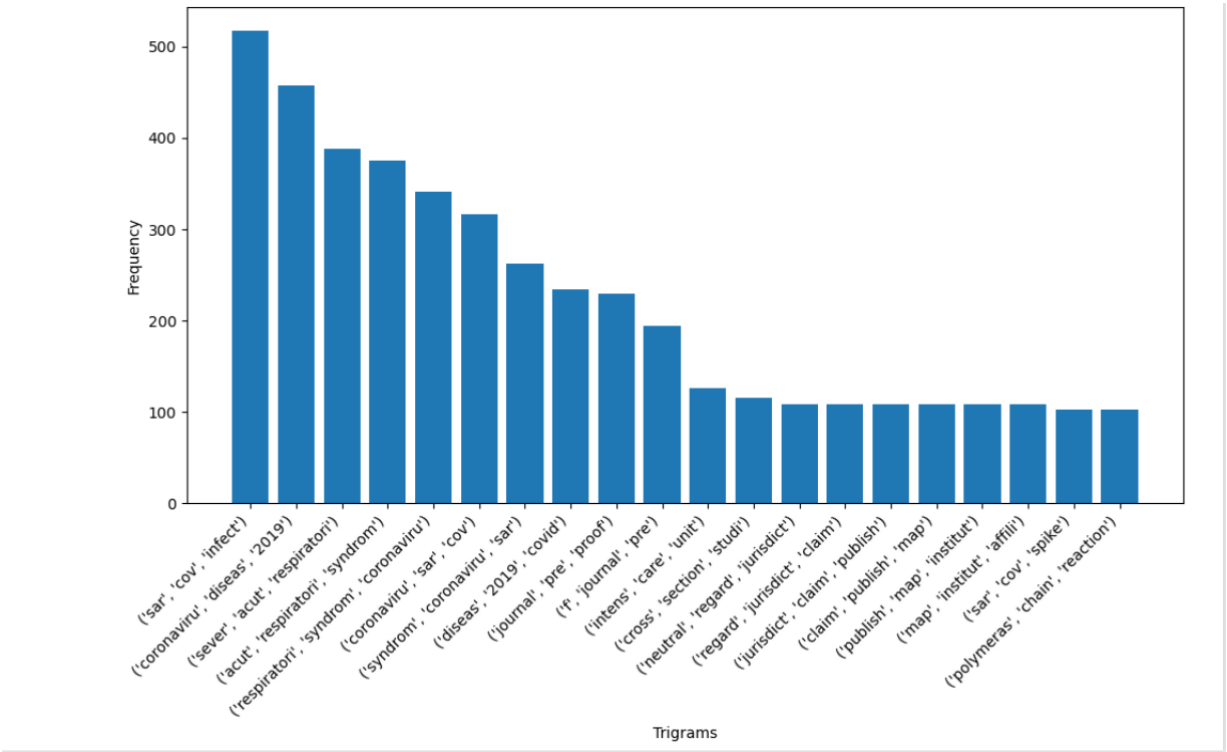
Figure 5. Word Cloud for Abstracts in 2022



My exploration of bigrams (*bigram is a sequence of two adjacent elements from a string of tokens, which are typically letters, syllables, or words*) and trigrams (*trigram is a sequence of three adjacent elements from a string of tokens, which are typically letters, syllables, or words*) within the COVID-19 literature has provided valuable insights into the diverse research themes that are present in this data.

In the abstracts, trigrams such as ('sar', 'cov', 'infect') and ('coronaviru', 'diseas', '2019') shed light on the virology and temporal context of the SARS-CoV virus. Notably, ('sever', 'acut', 'respiratori') and ('acut', 'respiratori', 'syndrom') highlight the severity and acute respiratory conditions associated with the virus. These findings, in the abstracts, serve as a snapshot, capturing key aspects of COVID-19 research and providing a high-level understanding of the priorities.

Figure 6. Top Trigrams in Abstracts



Interpreting the prevalence of certain trigrams and bigrams, I observed that their frequency reflects the evolving COVID-19 research landscape. The prominence of terms like 'immun', 'respons' suggests a keen interest in understanding immune responses, potentially indicating a focus on vaccine development or therapeutic strategies. The recurrent mention of 'long', 'term' underscores a commitment to studying the lasting effects of COVID-19 on individuals. Additionally, the consistent appearance of 'data', 'collect' emphasizes the importance of robust methodologies in the research process, reflecting a commitment to scientific rigor.

I also examined the distribution of various keywords related to the covid-19 pandemic. The wordclouds reveals distinct themes within documents containing specific keywords related to the covid-19 pandemic. For instance, documents featuring the "Symptomatic" keyword primarily focus on infection-related aspects, patients' clinical conditions, and the severity of cases. On the other hand, content associated with "Coronavirus" spans a broad spectrum, encompassing discussions on vaccines, patient care, respiratory issues, and the overall pandemic scenario. The "Spread" keyword is linked to discussions on infection, patients, and considerations such as lockdowns, risks, and outbreaks. The "Incubation" keyword is associated with topics like assay methods, cellular processes, animal control, viral dynamics, and vaccine development. Documents containing the "Ventilator" keyword delve into aspects such as hospitalization, mortality rates, respiratory health, mechanical ventilation, and medical considerations. The "Monoclonal" keyword centers on cellular processes,

proteins, antibodies, and genetic variations. In addition, documents featuring "Social Distancing" discuss social interactions, group behavior, lockdown strategies, and intervention measures.

3. Preprocessing

The preprocessing pipeline for the abstracts and body texts involved several steps to prepare them for topic modeling. Empty documents were discarded, eliminating those lacking valuable content. Next, infrequently appearing words were removed to focus on the most prominent vocabulary. The abstracts and body text then underwent thorough cleaning, which included removing special characters, numbers, punctuation, and converting text to lowercase for consistency. Each abstract was then broken down into individual words and filtered further by removing common stop words and domain-specific ones related to COVID-19 research. Finally, stemming reduced words to their base forms, ensuring better matching and capturing semantic similarities. The cleaned and processed abstracts and body texts were transformed into a numerical representation called a Document-Term Matrix (DTM) using CountVectorizer. This matrix captures the frequency of each word in each document, providing a foundation for quantitative analysis and topic modeling of the research content. The final files were exported to be used for modeling.

4. Modeling and Evaluation

In this research project, I utilized two topic modeling algorithms, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) techniques to extract and interpret topics from the extensive scientific literature on COVID-19.

Non-Negative Matrix Factorization (NMF): Non-Negative Matrix Factorization is a statistical method that helps us to reduce the dimension of the input corpora or corpora. Internally, it uses the factor analysis method to give comparatively less weightage to the words that are having less coherence.

latent Dirichlet allocation (LDA): a generative probabilistic model for collections of discrete data such as text corpora. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

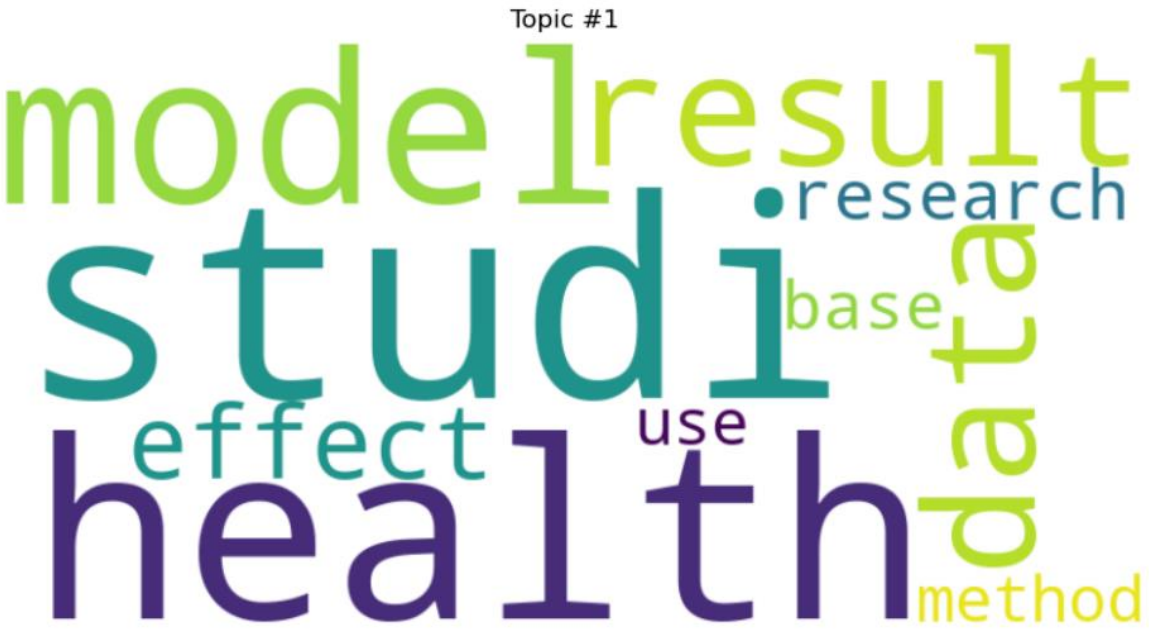
4a. Non-Negative Matrix Factorization (NMF)

In exploring thematic structures within the corpus of COVID-19 pandemic-related publications, I employed Non-negative Matrix Factorization (NMF). This technique is known for its efficacy in identifying latent topics in large text datasets. Following the initial data cleaning, I utilized a CountVectorizer to transform the filtered abstracts into a structured document-term matrix, subsequently normalized to accommodate the NMF algorithm's requirements. With the stage set, the NMF model was deployed, Initially, I examined the top 10 words across the top 5 topics

generated by the NMF model. My goal was to uncover the predominant themes that have shaped scholarly conversation during the pandemic. The results were as follow:

Topic 1: Focusing on health, research, and data-driven approaches to address the pandemic. Includes studies, modeling, and development of methods related to healthcare provision and analysis.

Figure 7. NMF Topic 1



Topic 2: Deeply delving into the biological aspects of the virus, particularly cellular, protein, and viral mechanisms. Covers infection, severe respiratory disease, and potential causes or targets for interventions.

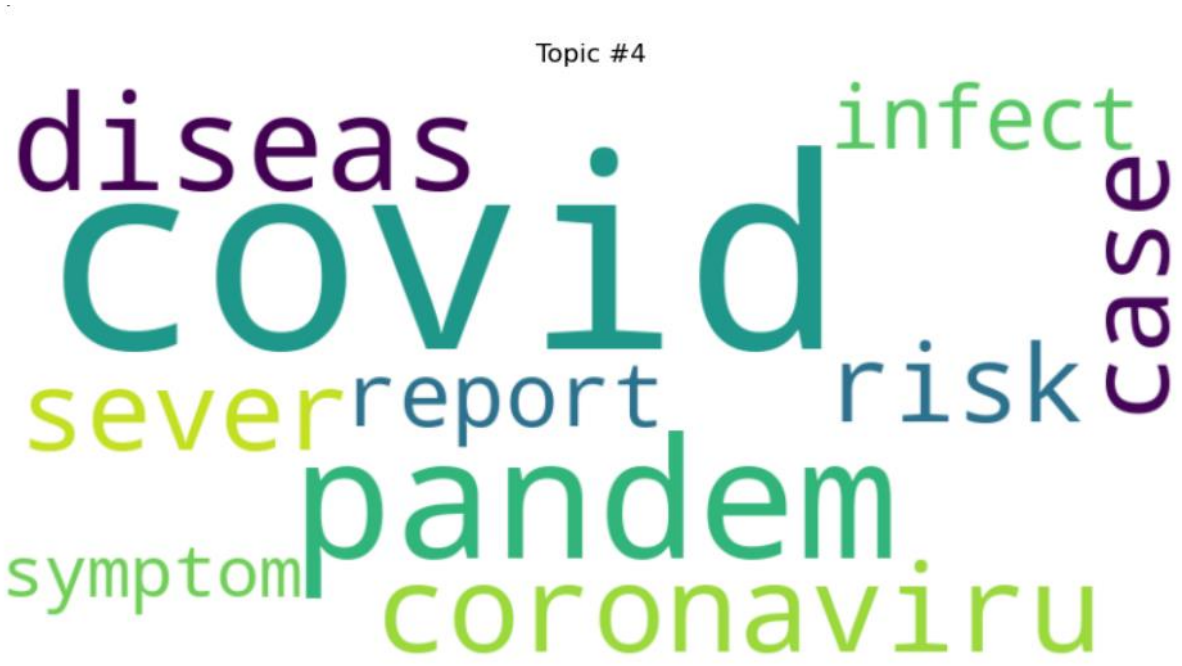
Figure 8. NMF Topic 2

Topic 3: Zooming in on patient experiences and clinical aspects. This theme explores treatment, outcomes, risk factors, and comparisons between groups, highlighting patient perspectives and healthcare settings.

Figure 9. NMF Topic 3

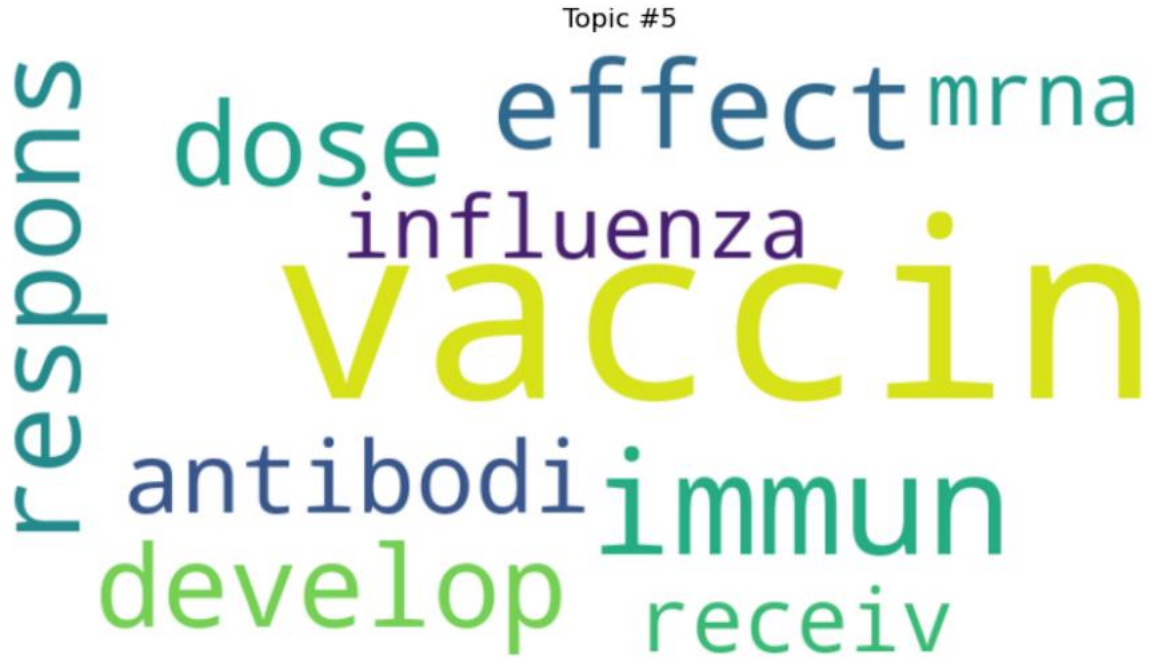
Topic 4: Discussing the broader impacts of the pandemic, encompassing case reports, disease spread, societal effects, and health consequences. Addresses concerns like severity, risk, symptoms, and global implications.

Figure 10. NMF Topic 4



Topic 5: Primarily investigating vaccination, immune responses, and protective measures. Covers vaccine development, effectiveness, dosage, safety concerns, and societal factors like hesitancy and accessibility.

Figure 11. NMF Topic 5



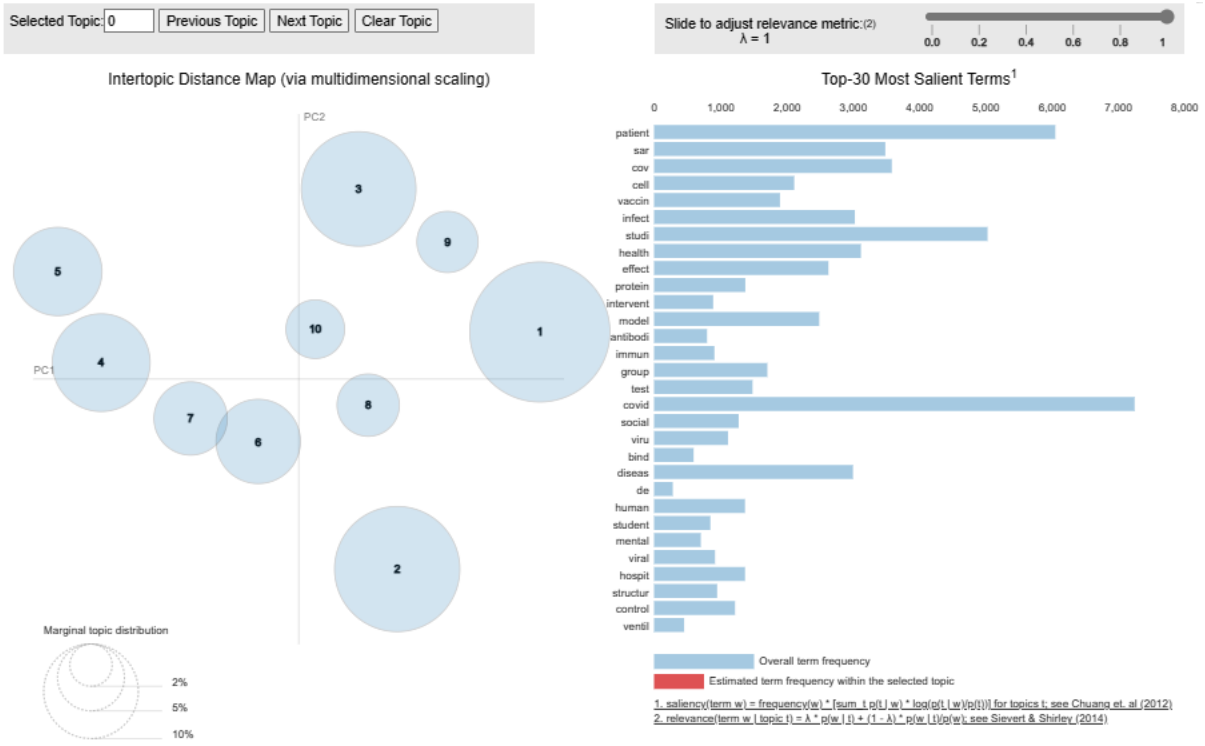
4b. latent Dirichlet allocation (LDA)

The other model that I tested on this data was LDA. This probabilistic modeling approach allowed me to systematically categorize the corpus into a predefined number of topics, aiming for a comprehensive overview of the dominant themes. Using the Gensim's implementation of the LDA model, the model was fitted to pre-processed corpus, which was represented as a document-term matrix and accompanied by a bespoke dictionary mapping of terms to their respective ids.

Following the model training, I sought to enhance interpretability and accessibility of the derived topics through the innovative use of pyLDAvis. This visualization tool facilitated an interactive exploration of the topics, providing an intuitive means to discern the distribution and relevance of terms within each topic.

Based on visual observations: There appears to be well-separated circles which indicate distinct topics with unique word distributions. Topics 7 and 6, being less separated, might share some common themes or vocabulary. Dominant Topics: Larger circles (topics 1, 2, and 3) suggest a higher prevalence of documents that heavily feature those topics' words. These topics might represent central themes or research areas within your dataset (Figure 12).

Figure 12. LDA Topics Visualization



4c. Evaluation

In the evaluative phase of the study, I assessed the efficacy and applicability of the two topic modeling algorithms: Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). This comparative analysis was foundational in my methodological decision-making process, with a particular focus on the interpretability of the models and the semantic coherence of the topics they generated. Preliminary evaluations, as depicted in Figure 13, revealed that LDA not only demonstrated superior interpretability but also yielded higher coherence scores compared to its NMF counterpart. These scores are indicative of the degree to which topic words are semantically related and thus, are critical in determining the quality and meaningfulness of the topics extracted. NMF, despite its strengths in matrix factorization and pattern discovery, consistently produced topics with lower coherence scores, suggesting the emergence of less semantically cohesive themes. This observation led to the decision to prioritize LDA for in-depth analysis, motivated by its capacity to uncover more interpretable and semantically meaningful topics within the corpus of COVID-19 research literature.

Figure 13: Evaluation Metrics

	Number of Topics	Silhouette Score (NMF)	Average Coherence Score (NMF)	Perplexity (LDA)	Coherence (LDA)
0	5	-0.003732	-1.403546	-7.639905	-1.349885
1	8	-0.033213	-1.673822	-7.716630	-1.471810
2	10	-0.036386	-1.901098	-7.758841	-1.527735
3	12	-0.049032	-1.703759	-7.821914	-1.613772
4	15	-0.037094	-1.766529	-7.905755	-1.717944

5. Hyperparameter tuning

Hyperparameter tuning was done to refine and maximize the efficacy of the LDA model. My primary objective was to augment the model's coherence and ensure the distinctiveness of the topics generated, thereby aligning more closely with the study's analytical goals. This process involved systematically varying the hyperparameters, specifically the document-topic density (alpha) and word-topic density (beta), across a predetermined range of values to identify the optimal configuration.

Despite minimal differences in coherence scores before and after hyperparameter tuning, LDA was deemed more suitable for its ability to generate topics that align with the study's objectives and maintain consistency in theme representation. Since the coherence scores before and after hyperparameter tuning is very small. I decided to assess the topics generated by the model qualitatively. Examining the top words for each topic to see if they make sense and are interpretable. I determined that 5 topics were optical for the LDA model based on coherence and perplexity scores.

The topics were as follows:

Topic 1: General COVID-19 and Model-related Keywords

Keywords: covid, model, method, data, study, base, result, time, patient, perform

Topic 2: Pandemic and Social Impact

Keywords: covid, study, pandemic, result, social, health, student, effect, model, care

Topic 3: Patient and Disease Characteristics

Keywords: patient, covid, cov, sar, study, infect, disease, severe, risk, hospital

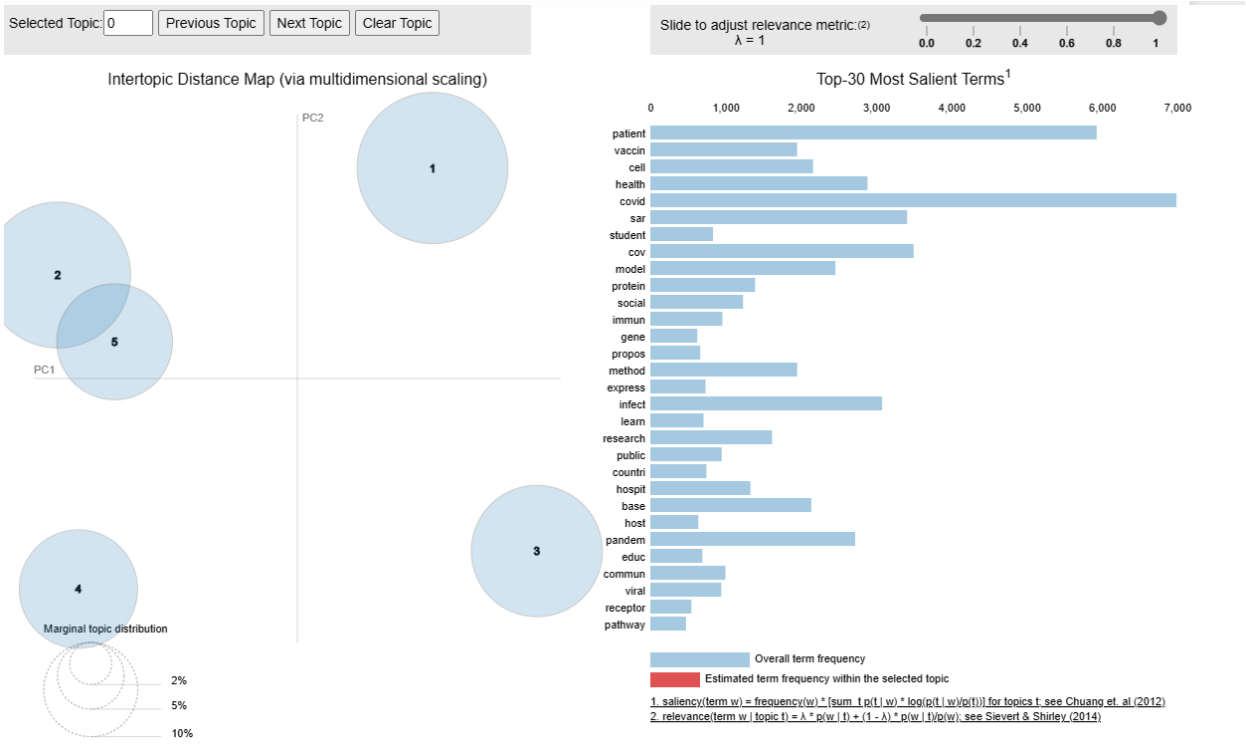
Topic 4: Vaccination and Health Research

Keywords: vaccine, health, study, covid, effect, research, pandemic, infect, result, cov

Topic 5: Cellular and Protein Aspects

Keywords: cell, sar, patient, study, infect, disease, protein, cov, effect, covid

Figure 17. Final Topic Visualization Using LDA Model



5. Summary of Findings

The topics generated through Latent Dirichlet Allocation (LDA) modeling of the scientific literature on COVID-19 offer insight into the extensive body of work produced during the pandemic. These identified topics span a diverse range of themes, including general discussions on COVID-19, the societal impact of the pandemic, patient and disease characteristics, vaccination efforts, and cellular aspects of the virus. The consistency in themes across models, both before and after hyperparameter tuning, underscores the robustness of the topic identification process. These topics provide a structured representation in line with the study's goal of clearly defining and interpreting themes from the literature.

The following general conclusions and implications emerged from the findings:

A. **Diverse Research Themes**: The identified topics encompass a wide array of research themes related to COVID-19, reflecting the diversity within the scientific literature on the subject.

B. **Consistency in Themes:** Despite subtle changes after hyperparameter tuning, the overall themes and keywords within topics remain consistent, indicating the robustness of the identified topics.

C. **Interpretable Topics:** The topics are interpretable and align with the broader context of COVID-19 research, providing valuable insights for researchers and practitioners into different aspects of the pandemic.

D. **Relevance to Study Goals:** Aligned with the study's goal, the identified topics offer a structured representation of key themes in the literature, enhancing understanding of the research landscape.

E. **Implications for Further Analysis:** The topics serve as a foundation for in-depth analysis, including exploration of subtopics, trends over time, and correlations between topics. Researchers can prioritize areas for further investigation based on the prevalence and significance of specific themes.

F. **Communication and Decision-Making:** Insights from the topics are valuable for communicating key findings to diverse audiences, including policymakers, healthcare professionals, and the general public. Understanding prevalent themes is crucial for informed decision-making during the ongoing pandemic and future ones.

G. **Validation and Expert Input:** While LDA models provide automated topic identification, validation by domain experts is essential. Collaboration with subject matter experts ensures that identified topics align with the latest developments and nuances in COVID-19 research.

6. Future Research Ideas

1. **Temporal Evolution of Research Themes:**
    - Investigate the temporal evolution of topics within the COVID-19 literature. Analyze how research themes have shifted over time, identifying emerging topics and tracking the prevalence of key themes during different phases of the pandemic. This longitudinal analysis could provide insights into the dynamic nature of scientific discourse and evolving priorities in COVID-19 research.
2. **Cross-Disciplinary Collaboration Analysis:**
    - Explore the potential for cross-disciplinary collaboration in COVID-19 research by examining the co-occurrence of topics across different domains. Identify interdisciplinary intersections and assess the extent to which research themes from fields such as medicine, public health, social sciences, and technology converge. This could inform strategies for fostering collaboration and knowledge integration across diverse scientific communities.
3. **Sentiment Analysis of Research Themes:**
    - Conduct sentiment analysis on the identified topics to gauge the emotional tone and sentiment prevalent in COVID-19 literature. Analyzing sentiments associated with specific themes could provide insights into the emotional context of research findings. Understanding the emotional undertones may contribute to a more nuanced understanding of how scientific discourse reflects the challenges, concerns, or optimism within the research community.

7. Recommendations

1. **Enhanced Validation with Expert Input:**
    - Collaborate with domain experts, including researchers, clinicians, and public health professionals, to validate and refine the identified topics. Expert input can offer additional context, ensuring that the topics align with the latest developments, emerging trends, and nuances in COVID-19 research. This validation process enhances the reliability and accuracy of the identified topics.
2. **Integration of External Data Sources:**
    - Integrate external data sources, such as citation networks, funding information, or global health data, to enrich the context around identified topics. Linking topics to citation patterns or funding sources can provide a broader understanding of the impact and influence of specific research themes. This integrated approach contributes to a more comprehensive analysis of the research landscape.
3. **Interactive Visualization for Knowledge Exploration:**
    - Develop interactive visualization tools to facilitate knowledge exploration for researchers, policymakers, and the general public. Create user-friendly interfaces that allow stakeholders to navigate and explore the identified topics, enabling a deeper understanding of the interconnectedness of themes. Visualization tools can enhance the accessibility of complex research insights and promote data-driven decision-making.