Exploring Topic Trends in CORD-19 Literature using Non-negative Matrix Factorization[⋆,⋆⋆]

ARTICLE INFO

Keywords:
Non-Negative Matrix Factorization
Topic Modelling
CORD-19
Natural Language Processing
Text Mining
Time Series Analysis

ABSTRACT

Objective: To demonstrate the effectiveness of Non-negative Matrix Factorization (NMF) in extracting latent topics and uncovering trends in literature related to COVID-19 using the CORD-19 dataset, and to explore the potential implications of these findings for researchers and policy makers.

Materials and Methods: The study employs a sophisticated approach to data preprocessing and analysis using the CORD-19 dataset, a comprehensive collection of over 800,000 scholarly articles related to COVID-19 and other coronaviruses. NMF is applied to identify latent topics and their distribution over time is analyzed. Correlations between topics are investigated to uncover hidden patterns and relationships. To ensure that the model produces meaningful results, we perform a topic stability analysis to determine the optimal number of topics for the data set so that number of topics will be more robust to perturbations in the data set.

Results: Our results show that NMF is a powerful technique for topic modeling on the CORD-19 dataset, providing valuable insights into trends in scientific literature related to COVID-19. Our findings demonstrate that certain topics have increased in popularity over time, while others have decreased. These results provide a nuanced understanding of the direction of COVID-19 research and can help guide future efforts.

Conclusion: This study highlights the potential of NMF as a tool for uncovering hidden patterns in large datasets and its applicability in the field of text mining. The use of topic stability analysis ensures that the model produces meaningful results by selecting an optimal number of topics for the data set. The insights gained from this analysis can inform future research efforts by highlighting areas of interest and potential gaps in the existing literature. For policy makers, the results of this analysis can provide valuable information for decision-making by providing a deeper understanding of trends and patterns in scientific literature related to COVID-19. Overall, this study demonstrates the value of NMF as a tool for researchers and policy makers seeking to understand and respond to the ongoing COVID-19 pandemic.

Exploring Topic Trends in CORD-19 Literature using Non-negative Matrix Factorization

1. Introduction and Motivation

The COVID-19 pandemic has had far-reaching and profound impacts on global health, economies, societies, and politics. These impacts have manifested in the extensive discourse surrounding the virus across media outlets, social media platforms, and among experts in various fields. To comprehensively understand the pandemic and its effects, it is crucial to systematically analyze this discourse.

Topic modeling is a powerful tool for uncovering latent structures within large text corpora. By applying topic modeling techniques to articles related to the pandemic, we can gain insight into how the conversation around COVID-19 has evolved over time. This research aims to use topic modeling to identify key themes and trends in the discourse surrounding COVID-19.

There are several motivations for conducting this research. Firstly, by identifying key topics discussed in articles related to COVID-19, we can better understand public perceptions of and responses to the virus. This understanding can inform public health communication strategies and policy decisions. Additionally, our research can help identify areas where information or consensus is lacking; this knowledge can guide future research and data collection efforts.

Furthermore, topic modeling can be used to identify key influencers within the discourse and track changes in their perspectives on the pandemic over time. Moreover, our research will provide a holistic overview of discourse on COVID-19 that could benefit researchers, policymakers, healthcare professionals, and members of the public. By presenting a comprehensive summary of main topics of discussion related to COVID-19 our research can aid in understanding the pandemic and its impact. This understanding can provide a foundation for further research on specific topics related to COVID-19.

This research has the potential to be used for various purposes. The insights gained from this analysis can inform future research efforts by highlighting areas of interest and potential gaps in the existing literature. Additionally, our findings can be used to develop predictive models that anticipate future trends in discourse related to COVID-19. Furthermore, our methodology can be applied to other large text corpora to uncover latent structures and trends in discourse on other topics.

2. Previous Work

There have been several attempts at topic modeling and topic trend analysis on scientific literature related to COVID-19. For instance, a study by **?** applied Non-negative Matrix Factorization (NMF) temporal topic models to clinical text

This note has no numbers. In this work we demonstrate a_b the formation Y_1 of a new type of polariton on the interface between a cuprous oxide slab and a polystyrene micro-sphere placed on the slab.

ORCID(s):

data to identify the effects of the COVID-19 pandemic on primary healthcare and community health in Toronto, Canada. While their topic modeling approach yielded some highly similar topics, the limited size of their dataset and lack of rigorous proof regarding topic model stability may have contributed to this result.

Another study by **?** conducted a comparative analysis between Latent Dirichlet Allocation (LDA) and NMF models using the COVID-19 corpus. However, their topic model was not as effective as it could have been due to the lack of blacklisting and whitelisting of words for topic modeling and the small number of topics (10) used in their analysis.

? explored the non-medical impacts of COVID-19 using Natural Language Processing. While they did an excellent job handling and preprocessing the dataset, their NMF model was not able to cluster topics properly and some topics contained words from different domains, making it difficult to label or classify the topics.

? proposed a novel methodology to identify the primary topics contained within the COVID-19 research corpus. However, they limited their model to only 10 topics and focused solely on the abstracts of articles in the CORD-19 dataset.

Finally, **?** conducted a topic trend analysis on COVID-19 literature. While they did an excellent job collecting and preprocessing data, they used Structural Topic Modeling (STM) instead of NMF.

In summary, while there have been several attempts at topic modeling and trend analysis on scientific literature related to COVID-19, there is still room for improvement in terms of dataset size, topic model stability, and clustering effectiveness.

3. Dataset

3.1. CORD-19 Dataset Description

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). 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. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

The dataset is provided in JSON format, with each article having several fields such as "paper_id", "doi", "abstract", "body_text", "authors", "title", "journal", "abstract_summary", "language", and "publish_time" that contain various information about the article such as the title, authors, journal, publication date, and the full text of the article. The dataset also includes meta-data information such as the license, citation count, and the source of the data.

The CORD-19 dataset is particularly useful for researchers interested in performing topic modelling on COVID-19 related research as it provides a vast amount of data for analysis. The dataset is also useful for researchers interested in natural language processing, text mining, and information retrieval, as it provides a large corpus of text data that can be used to train and test various machine learning models. The dataset is regularly updated to include the most recent research available, making it a valuable resource for researchers. It is worth noting that the dataset is also provided in a structured format making it easier to use and analyze, it also have a specific section for COVID-19 and non-COVID-19 related articles which makes it more specific for the research purposes.

Column	Description
paper_id	Unique identifier for each article
doi	Digital Object Identifier
abstract	Abstract of the article
body_text	Full text of the article
authors	Authors of the article
title	Title of the article
journal	Journal of the article
abstract_summary	Summary of the abstract
language	Language of the article
publish_time	Date of publication

Table 1: CORD-19 Dataset Columns

3.2. Data Preprocessing

The CORD-19 dataset contains a large amount of data, which makes it difficult to analyze. In order to make the dataset more manageable, we will perform some preprocessing steps to remove unnecessary information and reduce the size of the dataset. The preprocessing steps include removing duplicate articles and removing articles that are not in English. The preprocessing steps are described in more detail below.

- **Removal of articles (< 50 words):** Removing articles containing less than ten words is a common preprocessing step in text analysis because it helps to ensure that the resulting corpus is of high quality and contains only meaningful text data. Some of the reasons for removing articles containing less than ten words are:

    1. Such types of articles often contain a large number of stop words (common words such as "the", "and", "is", etc.), which do not carry much meaning and can skew the results of topic modelling.
    2. Articles containing less than 50 words may not have enough context or information to provide meaningful insights when performing topic modelling. They may be incomplete sentences or phrases that do not convey any meaning on their own.

3. They may also be errors in the data collection, such as typos, truncated sentences, or incomplete data.

Removing such articles helps to improve the computational efficiency of the topic modelling process, as it reduces the corpus's size and the number of irrelevant articles that need to be processed.

- **Tokenization**: Tokenization is the process of breaking down a sentence or a text into individual words or tokens. It is a crucial step in natural language processing (NLP) as it helps in identifying the syntactic structure of the text and aids in further analysis and text classification.
- **Converting words to lowercase**: Converting words to lowercase is an important step in preprocessing as it helps standardise the text and reduce the data's dimensionality. It also helps reduce the number of unique words in the text, which can improve the model's efficiency. For example, the words "coronavirus" and "Coronavirus" are the same word but are represented differently. By converting all words to lowercase, it helps in reducing the number of unique words in the text and improves the efficiency of the model.
- **Removing Numbers**: Removing numbers from the text helps in simplifying the data for further analysis. Numbers do not add any value to the text and can be removed without affecting the overall meaning of the text.
- **Contraction Expansion**: Contraction Expansion is the process of expanding contractions such as "didn't" to "did not" in the text. This step is important as it helps better understand the text's meaning. This is because contractions are informal and may be difficult to understand for NLP models. Expanding contractions makes it easier for NLP models to understand the text and improve the model's overall performance.
- **Lemmatization**: Lemmatization is the process of reducing words to their base or root form (e.g. "running" becomes "run"). This process is important as it helps standardize the text and reduce the dimensionality of the data. It also helps in understanding the meaning of the text, as different forms of a word may have different meanings. For example, "running" and "ran" have different meanings, but they are derived from the same root word "run".
- **Removing Punctuations**: Removing punctuation marks (e.g. ",", ".", "!", etc.) from the text is an important step in preprocessing as it helps in simplifying the data for further analysis. Punctuation marks do not carry any meaning and can add noise to the data. Removing them can improve the analysis's efficiency and the model's overall performance.
- **Stop-Words Removal**: Stop-words removal is the process of removing commonly used words that do not carry much meaning (e.g. "is", "and", "the", etc.) from the text. This step is important as it helps reduce the data's dimensionality and improve the analysis's efficiency. Stop-words do not carry any meaning and can add noise to the data. Removing them can improve the performance of the model and make it easier to understand the text.
- **Handling N-grams**: N-grams are a sequence of n words that occur together in a text. N-grams can be used to improve the performance of the model by providing more context to the model. For example, mental health is a phrase that is used to describe the state of a person's mental well-being. If we use the word "mental" in isolation, it may not be clear what the word means. However, if we use the phrase "mental health", it is clear that the word "mental" is used to describe the state of a person's mental well-being. N-grams can be used to improve the performance of the model by providing more context to the model. Therefore it is important to identify the most important n-grams in the text and use them to improve the model's performance. Also, it is necessary to process n-grams in descending order. For example, we must process tri-grams before bi-grams. This is because tri-grams contain information about bi-grams, and bi-grams contain information about unigrams. For example, we must merge 'acute respiratory distress syndrome ards' into a 5-gram before merging 'acute respiratory distress syndrome' into a 4-gram.
- **Creating Dictionary:** The goal of creating a dictionary is to remove rare and common words that are unlikely to provide useful information for the topic model. The Dictionary class has parameters such as no_below and no_above that can be used to remove these extremes. The no_below parameter removes words that appear in less than a certain number of documents (specified as an integer), while the no_above parameter removes words that appear in more than a certain percentage of documents (specified as a value between 0 and 1). Setting these parameters allows for the removal of rare and common words from the dictionary.
- **Vectorization:** Vectorization is a technique used in natural language processing (NLP) to convert text data into numerical form so that it can be used as input for machine learning algorithms. One common method of vectorization is term frequency-inverse document frequency (tf-idf), which assigns a weight to each word in a document based on how frequently it appears in the document and how rare it is across all documents in the corpus.

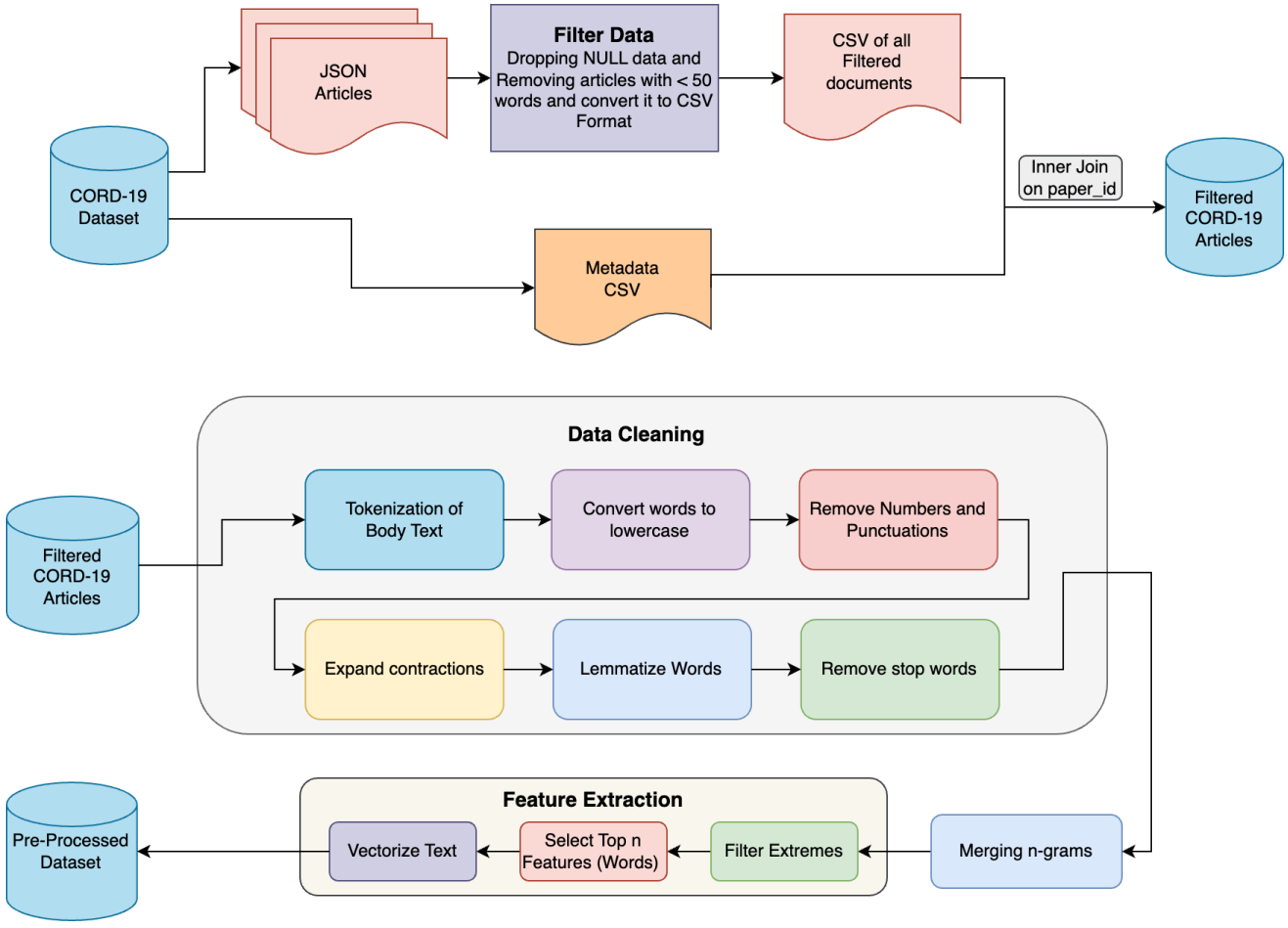

Figure 1: Methodology of Data Pre-Processing Approach which is used for Topic Modelling using NMF

4. Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) is a technique used in machine learning and data analysis to extract meaningful patterns and structures from high-dimensional data. In the context of topic modeling research, NMF can be used to factorize a document-term matrix into two matrices, one representing the topics and the other representing the distribution of these topics across the documents.

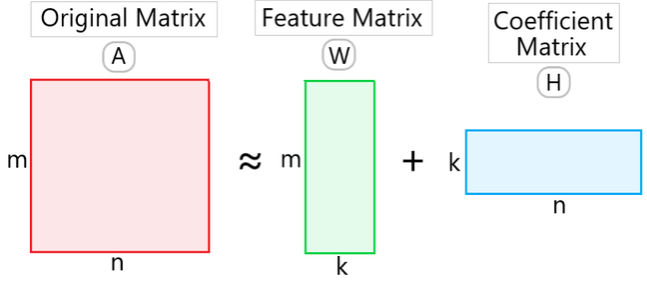

Figure 2: Illustration of Non-negative Matrix Factorization

5. The NMF Model

Given a non-negative matrix $X \in \mathbb{R}_{\geq} 0^{m \times n}$, where m is the number of documents and n is the number of terms in the vocabulary, the goal of NMF is to factorize X into two non-negative matrices $W \in \mathbb{R}_{\geq} 0^{m \times k}$ and $H \in \mathbb{R}_{\geq 0}^{k \times n}$, where k is the number of topics.

The factorization of X can be written as:

$$X \approx W H \tag{1}$$

where $\approx$ denotes the approximate equality.

5.1. The NMF Procedure

The NMF procedure consists of the following steps:

5.1.1. *Initialization*

The first step in NMF is to initialize the matrices W and H. Random initialization is commonly used, where the elements of W and H are set to random values in the range $[0, 1]$. However, other initialization methods, such as singular value decomposition (SVD), can also be used.

5.1.2. *Update Rules*

The update rules for NMF are derived from minimizing the Frobenius norm of the difference between X and WH:

$$\min_{W,H} ||X - WH||_F^2 \tag{2}$$

where $||\cdot||_F$ denotes the Frobenius norm.

The update rules for W and H are as follows:

$$W_{ia} \leftarrow W_{ia} \frac{(XH^T)ia}{(WHH^T)ia} \tag{3}$$

$$H_{aj} \leftarrow H_{aj} \frac{(W^T X)aj}{(W^T WH)aj} \tag{4}$$

where $i \in 1, 2, \ldots, m$, $a \in 1, 2, \ldots, k$, and $j \in 1, 2, \ldots, n$.

These update rules iteratively refine the factorization of X until convergence.

5.1.3. Stopping Criterion

The NMF procedure stops when a stopping criterion is met. Common stopping criteria include a maximum number of iterations or a threshold for the change in the Frobenius norm between iterations.

5.2. Interpretation

Each row of W represents a document and each column represents a topic. The entries in each row represent the contribution of each topic to reconstructing the corresponding document in V. Similarly, each column of H represents a word and each row represents a topic. The entries in each column represent the contribution of each word to defining the corresponding topic.

5.3. Stability Analysis

Topic models are widely used to discover latent topics in large collections of text documents. However, due to the probabilistic nature of these models, the topics identified by a topic model can vary between different runs of the model with different random initializations. This variability can make it difficult to assess the reliability and interpretability of the topics identified by a topic model.

To address this issue, we conducted a stability analysis of our topic model following the methodology proposed by Greene et al. in their paper "How Many Topics? Stability Analysis for Topic Models" **?**. This analysis measures the similarity between multiple runs of the same topic model with different random initializations.

Our results show that the stability scores for our topic model remain high even when the number of topics is increased to 40 and beyond. This suggests that our topic model is able to consistently identify a stable set of topics even when the number of topics is relatively large.

One possible explanation for this result is that our dataset contains a large number of distinct and well-separated topics. In this case, increasing the number of topics in the model allows it to better capture the underlying structure of the data, leading to more stable and interpretable topics.

These results provide strong evidence for the robustness and reliability of our topic model. By demonstrating that our model is able to consistently identify a stable set of topics across multiple runs, we can have greater confidence in the interpretability and generalizability of the topics identified by our model.

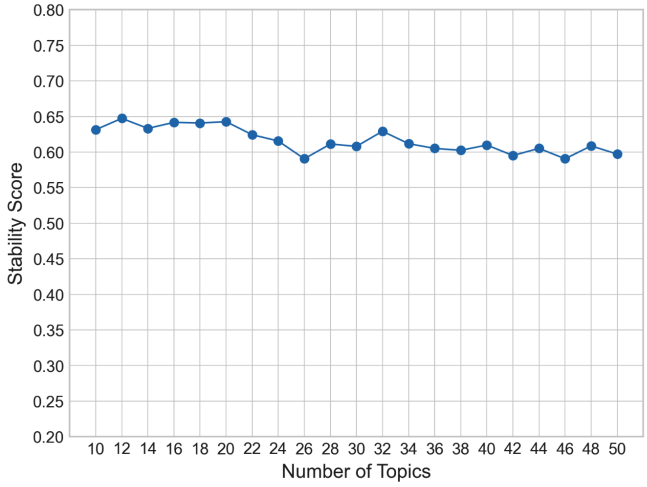

Figure 3: Stability scores of NMF topic model for different number of topics

Based on our stability analysis, we chose to use a topic model with 20 topics for our analysis. While our results show that the stability scores remain high even when the number of topics is increased beyond 20, we found that increasing the number of topics beyond this point resulted in the identification of additional topics with relatively small contributions to the overall structure of the data. In other words, while these additional topics were distinct and well-separated, their presence did not significantly improve the interpretability or generalizability of our model. As such, we chose to use a topic model with 20 topics as a balance between model complexity and interpretability.

For instance, when we increased the number of topics to 40, our topic model identified additional topics such as a separate topic for sleep-related problems, a separate topic for stroke, and a separate topic for stem cells. While these topics were distinct and well-separated, their contributions to the overall structure of the data were relatively small. In other words, the presence of these additional topics did not significantly improve the interpretability or generalizability of our model. As such, we chose to use a topic model with 20 topics as a balance between model complexity and interpretability.

6. Finding Topic Trends using NMF

W and H. The matrix W represents the relationship between articles and topics, while the matrix H represents the relationship between topics and terms.

To identify topic trends over time, we first divide our collection of articles and sort them based on their publication dates. For instance, we might group articles by month. Let n_t denote the number of articles published in month t, and let m represent the number of topics. For each month t, we compute the average topic distribution for all articles published in that month as follows:

Let n_i denote the total number of articles published in month i, and let W represent the matrix encoding the relationship between articles and topics.

Let $I_k = i | i \in [\sum_{j=1}^{k-1} n_j + 1, \sum_{j=1}^{k} n_j]$ represent the set of indices of articles belonging to month k. Then, the sum of all $W_i / \sum_{l=1}^{n} W_{il}$ for articles belonging to month k can be expressed as:

$$\textbf{Topic Trend Distribution for } month_k = \sum_{i \in I_k} \frac{W_i}{\sum_{l=1}^{n} W_{il}}$$

Where W_i denotes the i^{th} row of W, and W_{il} represents the l^{th} element in the i^{th} row of W.

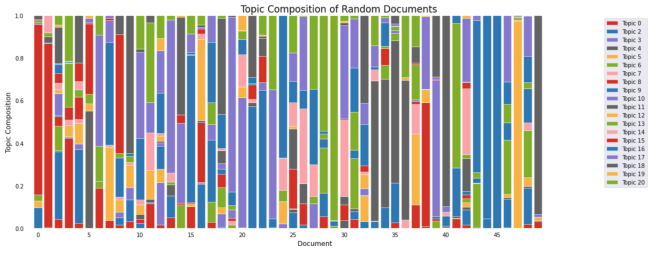

Figure 4: Distribution of Topics topics in a random sample of documents

This equation calculates a normalized topic distribution for each article by dividing each element in row i by the sum of all elements in that row. These normalized values are then summed for all articles published in a given month and divided by the total number of articles to obtain an average topic distribution for that month.

By repeating this process for all months, we can identify trends in how topics change over time. By analyzing these trends, we can gain insights into how different topics evolve and interact with one another.

6.1. Using Relevance to Find Relevant Terms for Topic Modeling on CORD-19 Dataset

Non-negative Matrix Factorization (NMF) is a popular technique for topic modeling that can be applied to the CORD-19 dataset. One challenge in using NMF for topic modeling is identifying relevant terms for each topic. In this section, we discuss how the relevance formula introduced by Sievert and Shirley (2014) can be used to address this challenge.

The relevance of a term w to a topic t is defined as:

$$\text{relevance}(w \mid t) = \lambda p(w \mid t) + (1 - \lambda) \frac{p(w \mid t)}{p(w)} \qquad (5)$$

where λ is a weight parameter that determines the weight given to the probability of term w under topic t relative to its lift. By adjusting the value of λ, we can control how much weight is given to the probability of a term under a specific topic relative to its lift.

When $\lambda = 1$, terms are ranked solely based on their probability under a specific topic. When $\lambda = 0$, terms are ranked solely based on their lift. For values of λ between 0 and 1, both probability and lift are taken into account when ranking terms.

This formula can be helpful in finding relevant terms for NMF related to topic modeling on CORD-19 dataset because it allows us to rank terms within topics based on their relevance. By selecting terms with high relevance scores, we can identify terms that are most likely to be informative and useful for understanding each topic.

7. Topic Modelling using NMF Results

Setting λ to 0.5 in the relevance formula introduced by Sievert and Shirley (2014) can help identify the best words for topic modeling on the CORD-19 dataset. When λ is set to 0.5, both the probability of a term under a specific topic and its lift are given equal weight when ranking terms.

The probability of a term under a specific topic measures how often that term appears in documents associated with that topic. The lift of a term measures how much more often that term appears in documents associated with a specific topic compared to its overall frequency in the entire corpus.

By setting λ to 0.5, we balance these two factors when ranking terms within topics. This means that terms with high probability under a specific topic and high lift will be ranked highly. These terms are likely to be informative and useful for understanding each topic because they appear frequently in documents associated with that topic and are more characteristic of that topic compared to other topics.

No.	Topic Label	Words
1	Hospitalization and Comorbidities	**Frequent:** icu, admission, diabetes, hospitalization, comorbidities, hospitalized, admitted, ci, cardiac, oxygen **Relevant:**icu, admission, diabetes, aki, mechanical_ventilation, comorbidities, hospitalized, admitted, anticoagulation
2	Economic Policy and Crisis	**Frequent:** food, policy, economic, government, market, crisis, sector, income, business, economy **Relevant:** food, market, economic, sector, business, policy, economy, firm, political, price, investment
3	Vaccines and Vaccination	**Frequent:** vaccine, vaccination, vaccinated, dose, vaccine_hesitancy, bnt, dos, influenza, efficacy, trial **Relevant:** vaccine, vaccination, vaccinated, vaccine_hesitancy, dose, bnt, pfizer, dos, mrna_vaccine, biontech
4	Education and Online Learning	**Frequent:** student, learning, teacher, teaching, school, education, academic, faculty, skill, online_learning **Relevant:** student, learning, teaching, teacher, online_learning, instructor, classroom, lecture, medical_student, semester
5	Immune Response and Inflammation	**Frequent:**il, cytokine, inflammatory, mouse, immune, expression, gene, in, lung, macrophage **Relevant:** il, cytokine, inflammatory, macrophage, ifn, tnf, activation, mouse, monocyte, inflammation, signaling
6	Mental Health and Well-being	**Frequent:** anxiety, depression, mental_health, stress, psychological, item, sleep, disorder, questionnaire, survey **Relevant:** anxiety, depression, mental_health, psychological, stress, sleep, depressive, coping, loneliness, mental
7	Antibody Response and Testing	**Frequent:** antibody, igg, rbd, assay, serum, igm, anti, titer, elisa, iga **Relevant:** antibody, igg, igm, iga, elisa, serum, rbd, titer, neutralization, antibody_response
8	Children	**Frequent:** child, parent, school, adolescent, pediatric, children, adult, mother, household, caregiver **Relevant:** child, parent, children, school, adolescent, pediatric, parental, parenting, mis, childhood
9	Machine Learning	**Frequent:** image, algorithm, layer, dataset, accuracy, training, network, classification, performance, learning **Relevant:** image, dataset, algorithm, layer, classification, accuracy, neural_network, classifier, deep_learning, architecture
10	Telemedicine and Healthcare	**Frequent:**telemedicine, nurse, staff, visit, telehealth, healthcare, team, provider, training, program **Relevant:** telemedicine, telehealth, visit, provider, team, staff, consultation, nurse, resident
11	Virus Variants and Mutations	**Frequent:** mutation, variant, sequence, genome, gene, lineage, omicron, orf, strain, sequencing **Relevant:** mutation, variant, genome, sequence, lineage, clade, omicron, gisaid, phylogenetic
12	PCR Testing and RNA Detection	**Frequent:** pcr, rt, rna, assay, swab, detection, specimen, sensitivity, saliva, gene **Relevant:** rt, pcr, specimen, swab, saliva, apcr, rna, ct value, lamp, assay
13	Pregnancy and Maternal Health	**Frequent:** woman, pregnant, pregnancy, maternal, mother, birth, delivery, infant, neonatal, fetal **Relevant:** woman, pregnant, pregnancy, maternal, birth, trimester, preterm, mother, postpartum, neonatal, gestational
14	Airborne Transmission and Mask Usage	**Frequent:** mask, particle, air, aerosol, droplet, pm, concentration, surface, temperature, dental **Relevant:** mask, aerosol, air, particle, droplet, pm, airborne, respirator, temperature, filtration, face_mask
15	Lung Imaging and Pneumonia	**Frequent:**ct, lung, chest, pneumonia, scan, imaging, pulmonary, lesion, consolidation, opacity **Relevant:**ct, chest, consolidation, lung, scan, opacity, ggo, pneumonia, lesion, lobe
16	Epidemic Control and Transmission	**Frequent:**epidemic, lockdown, transmission, outbreak, infectious, wave, reproduction_number, quarantine, susceptible, equation **Relevant:** epidemic, lockdown, reproduction_number, sir, susceptible, mobility, seir, equation, confirmed_case, cumulative
17	Cancer	**Frequent:** cancer, surgery, surgical, tumor, chemotherapy, oncology, breast cancer, surgeon, postoperative, screening **Relevant:** cancer, surgery, chemotherapy, tumor, surgical, oncology, breast_cancer, radiotherapy, colorectal
18	Drug Design and Antiviral Compounds	**Frequent:** compound, binding, ligand, residue, nsp, energy, docking, inhibitor, molecule, antiviral **Relevant:** compound, ligand, docking, residue, kcal_mol, molecule, nsp, atom, binding, hydrogen_bond
19	ACE2 Receptor and Virus Binding	**Frequent:** ace, bd, binding, receptor, arb, ang, angiotensin, tmprss, ace expression, lung **Relevant:** ace, ang, arb, ang_ii, rbd, ace_expression, angiotensin, tmprss, ace_protein, raas
20	Healthcare Workers	**Frequent:** hcws, hw, nurse, ppe, staff, worker, healthcare_worker, seroprevalence, burnout, working **Relevant:** hcws, hcw, ppe, nurse, ipc, occupational, worker, frontline, seroprevalence, healthcare_worker

Table 2: Topic Modeling Results

7.1. Word Cloud for Topics

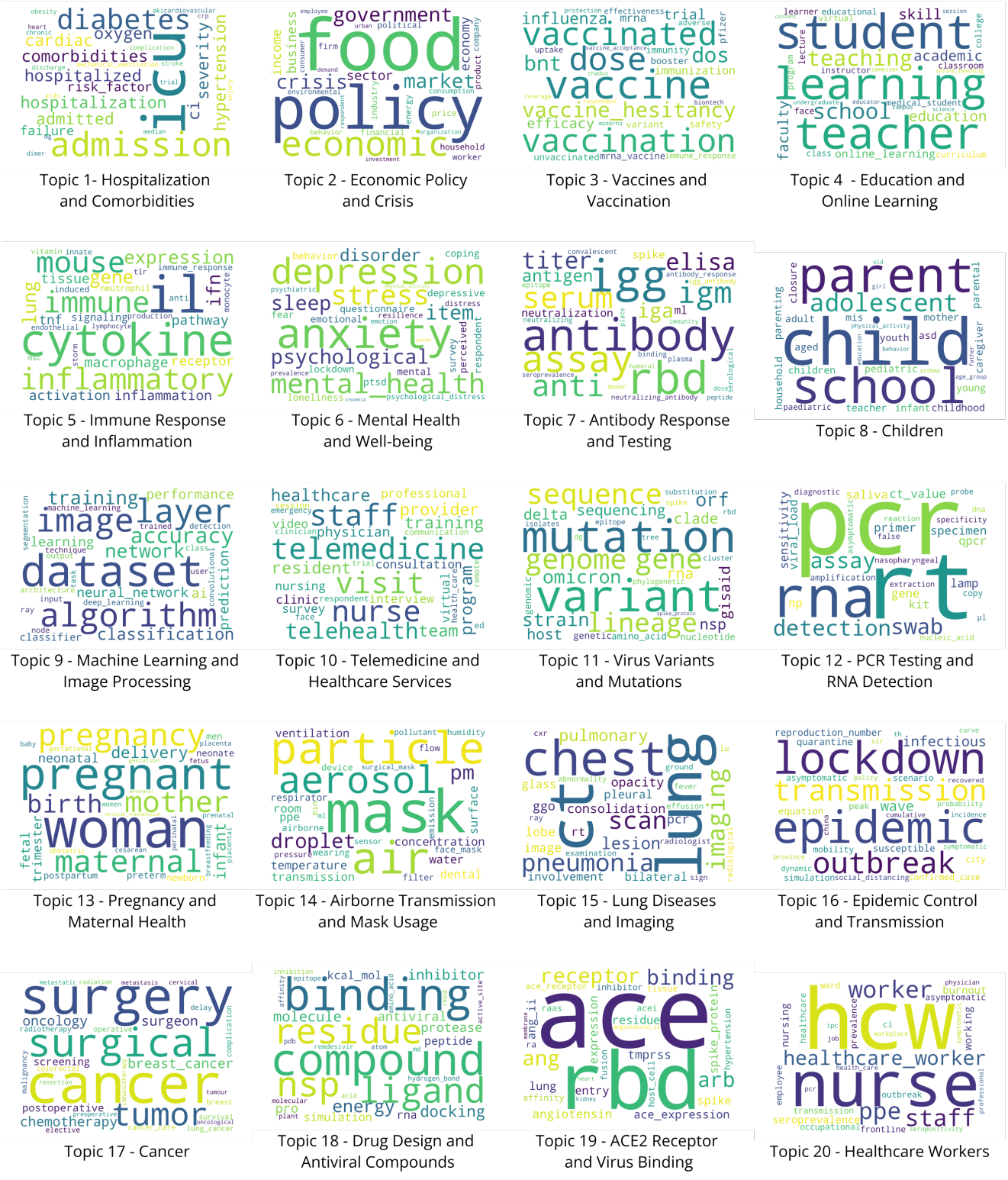

Figure 5: Top words for each topic for 20-Topics where the heat-map colour-scale denotes the percentage contribution of the topic in the corpus for that month.

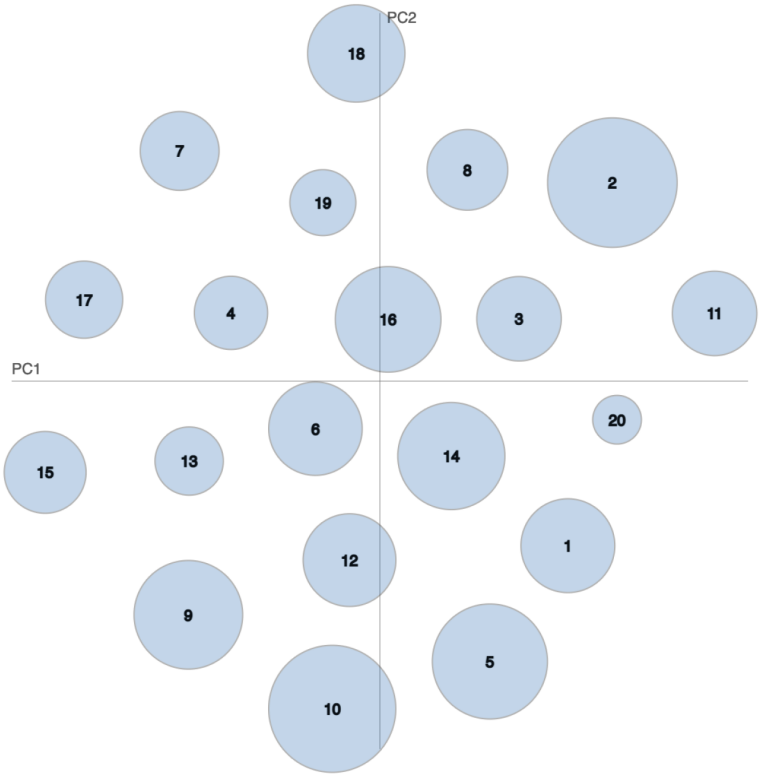

Figure 6: t-SNE visualization of NMF topic modeling results, where the size of each circle represents the relative prevalence of the corresponding topic in the corpus. The figure also visualizes the intertopic distance.

8. Topic Trends

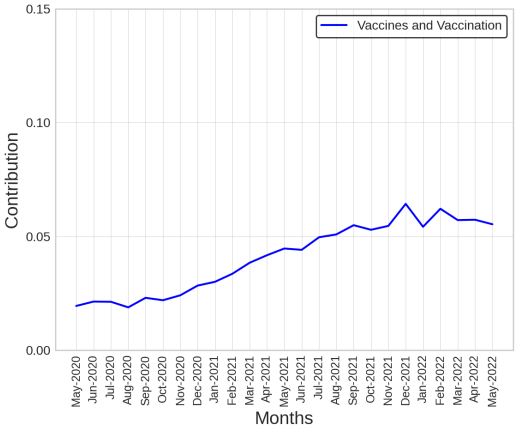

Figure 7: Topic Trend for Vaccines and Vaccination

Figure 7 shows that how the contribution of vaccine topic has significantly increased over the timeline of COVID-19 pandemic, it also shows how much important the vaccine was to the people. The topic started lifting up around from Sep-2020. This increasing trend can be said to be due to various factors including global media coverage, public interest, vaccine development and rollout, updates on vaccine efficacy and safety and more importantly the vaccine's importance in preventing the spread of disease and deaths in the world, thus creating a compelling environment over Pharma and research sector to develop the vaccines at the earliest.

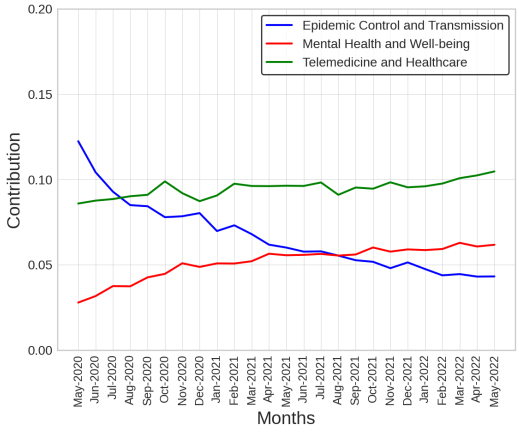

Figure 8: Topic Trend for Epidemic Control and Transmission, Mental Health and Well-being, Telemedicine and Healthcare

Figure 8 shows how the trends of Mental Health and Well-being, Telemedicine and Healthcare, and Epidemic Control and Transmission during the pandemic timeline are

interrelated.The increasing trend in Mental Health and Well-being and Telemedicine and Healthcare can be attributed to the pandemic's impact on people's health and the need to adopt new ways of delivering healthcare services remotely. This has led to a decrease in the trend of Epidemic Control and Transmission, as people have become accustomed to the measures put in place to control the spread of the virus, such as social distancing and wearing masks. On the other hand, the decreasing trend in Epidemic Control and Transmission can be attributed to the rollout of vaccines and a shift in focus from controlling the spread of the virus to managing its impact. This has led to an increase in the trend of Mental Health and Well-being and Telemedicine and Healthcare, as people seek information and resources on how to cope with the pandemic's emotional toll and adopt new ways of receiving healthcare services remotely.

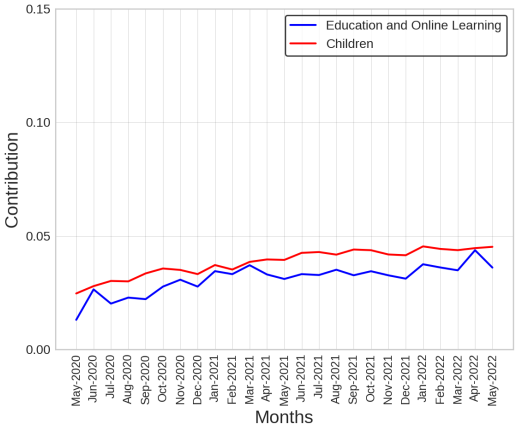

Figure 9: Topic Trend for Education and Online Learning, Children

Figure 9 shows how the pandemic has brought about a significant shift towards online learning, leading to an increased focus on child development and the impact of online learning on children's development. The increasing trend of child development and online learning during the pandemic can be attributed to several factors. Firstly, the pandemic has caused widespread school closures, leading to a shift towards online learning. This shift has raised concerns among parents, educators, and policymakers about how it may impact children's cognitive, social, and emotional development. Secondly, the pandemic has accelerated the adoption of digital technology in education and the major discussions have centered around the effectiveness of online learning, digital literacy skills, and the digital divide that exists among students. The pandemic has led to an increase in parental involvement in their children's education due to the shift towards online learning.

Figure 10 shows how trends of RT-PCR Testing, Airborne Transmission and Mask Usage, Lung Imaging, and Pneumonia are decreasing along the pandemic timeline. These trends can be attributed to several factors but one of the major factor is increasing vaccination over the time.

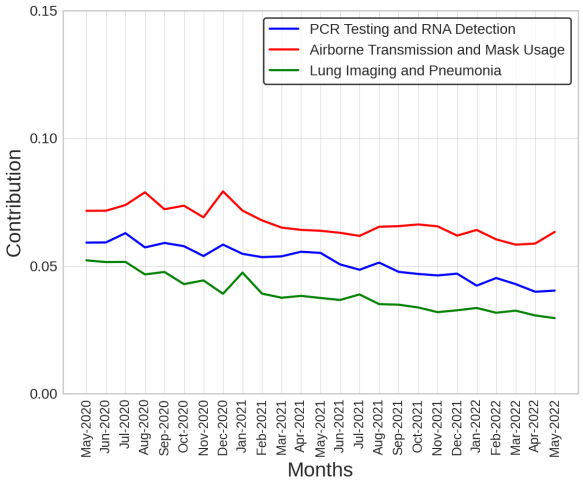

Figure 10: Topic Trend for RT-PCR Testing and RNA Detection, Airborne Transmission and Mask Usage, Lung Imaging and Pneumonia

Firstly, the widespread adoption of vaccines has led to a reduction in the need for RT-PCR Testing, as individuals who have received the vaccine may not need to get tested as frequently as those who have not. As more people receive the vaccine, the need for RT-PCR Testing decreases, leading to a decrease in discussions around this topic. Secondly, vaccines provide protection against severe illness and hospitalization from COVID-19, reducing the need for discussions around lung imaging and pneumonia. As more individuals receive the vaccine, the risk of severe illness and hospitalization decreases. Thirdly, the increasing trend of vaccine adoption has led to a decrease in the trend of discussions around airborne transmission and mask usage. This is because vaccines provide protection against the virus and reduce the likelihood of transmission, making it less necessary to discuss these measures.

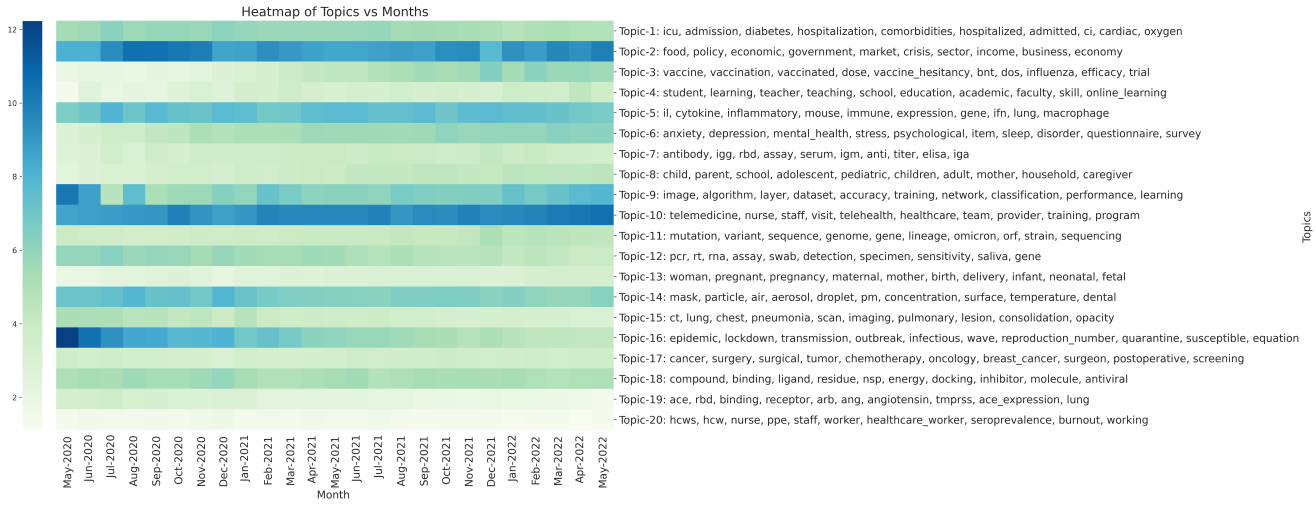

Figure 11: Heatmap of Topic Trends of all topics.

9. Correlation & P-Score Analysis

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. The Pearson coefficient will help us to get an analysis of topic trends. This will help us to get which topics are strongly/weakly connected with each other (increasing with each other, decreasing with each other, opposite trends).

Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. When we want to show support for our hypothesis through statistical results, our statistical results should reject the null hypothesis. We can reject the null hypothesis through the p-score value. Here we want to prove that there is a correlation between topics found through topic modeling. So this will become our alternative hypothesis. So, in this case, our null hypothesis will be that there is no correlation between any of the topics.

Most of the statistical tests are based on the comparison of within-group variance versus between-group variance.

If the between-group variance is large enough that there is little or no overlap between groups, our statistical test will reflect that by showing a low p-value. This means it is unlikely that the difference between these groups came about by chance. In our case, this will mean that when we see any correlation between two topics, it has influence, and we can reject our null hypothesis, that is, stating that there is no correlation between any topics.

Alternatively, if there is high within-group variance and low between-group variance, then your statistical test will reflect that with a high p-value. This means it is likely that any difference we measure between groups is due to chance. In our case, this means that when we see any correlation between two topics, it could come by chance. So, we can not reject the null hypothesis based on our statistical results.

So, the lower the p-value, the more confidently we can reject the null hypothesis and can strongly stand by our alternative hypothesis. In our case, when we calculated the correlation between topics through the residual value, we got a lower p-value in most of the cases but went for finding topic contribution by summing W[i]/sum(W[i]) for all documents in the time-slice which gives fraction of topic contribution for each month.

The fraction of residual in W matrix represents the proportion of each topic in the document, rather than just a binary classification of belonging to one topic or not. Additionally, summing the fractions of residuals for each document and normalizing by dividing by the sum of residuals allows for comparison and aggregation of topic distributions across multiple documents. On the other hand, assigning one topic to each document or taking a fraction of topics for each document can lead to oversimplification and loss of information about the multiple topics that a document may cover.

For our analysis, we selected Spearman correlation as the preferred method for measuring the strength and direction of the relationship between two variables. Spearman correlation is a non-parametric statistical method that is widely used when the relationship between variables is non-linear or the data are non-normally distributed.

Spearman correlation measures the monotonic relationship between two variables, which is a more general measure of association that can detect non-linear relationships. In contrast, Pearson correlation measures the linear relationship between variables, which may not accurately represent the strength of the relationship if it is non-linear. Spearman correlation can therefore be used to detect more complex relationships between variables.

Additionally, Spearman correlation is based on ranks and is therefore more robust to outliers compared to Pearson correlation. Outliers can heavily influence the results of correlation analysis, especially when the sample size is small.

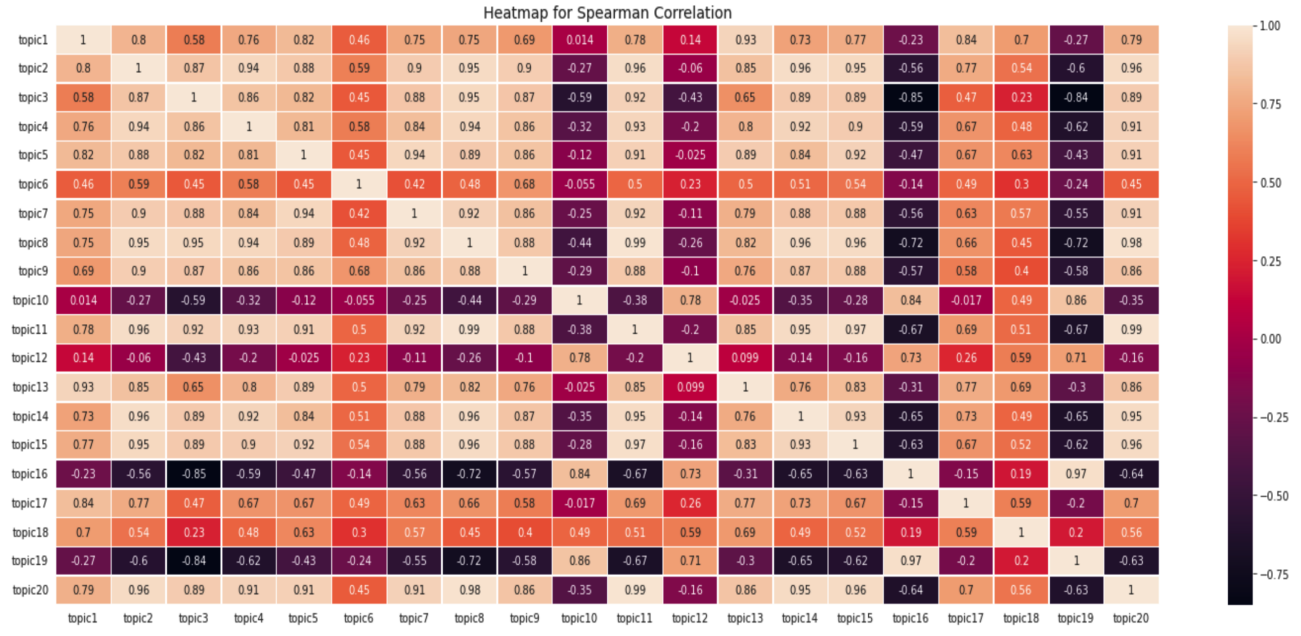

Figure 12: Correlation using Spearman Correlation

Ranks are a more stable and reliable way of measuring the strength of the relationship between variables, even in the presence of outliers.

Finally, Spearman correlation can be used with categorical or ordinal data, which can be converted to ranks. Pearson correlation, on the other hand, is only appropriate for continuous data. The ability to use Spearman correlation with categorical or ordinal data makes it a more versatile tool for correlation analysis.

In summary, we chose to use Spearman correlation for our analysis because it is a non-parametric method that is more suitable for non-linear or non-normally distributed data. Additionally, its reliance on ranks makes it more robust to outliers and suitable for categorical or ordinal data.

A. My Appendix

Appendix sections are coded under `\appendix`.

`\printcredits` command is used after appendix sections to list author credit taxonomy contribution roles tagged using `\credit` in frontmatter.

Figure 13: p-score