SMU Data Science Review

Volume 3 | Number 2 Article 1

2020

A Novel Methodology to Identify the Primary Topics Contained Within the COVID-19 Research Corpus

Allen Crane
Southern Methodist University, acrane@mail.smu.edu

Brock Freidrich
bfriedrich@mail.smu.edu

William Fehlman
wfehlman@gmail.com

Igor Frolow
ifrolow@austin.rr.com

Daniel W. Engels
dwe@alum.mit.edu

Recommended Citation

Crane, Allen; Freidrich, Brock; Fehlman, William; Frolow, Igor; and Engels, Daniel W. (2020) "A Novel Methodology to Identify the Primary Topics Contained Within the COVID-19 Research Corpus," *SMU Data Science Review*: Vol. 3: No. 2, Article 1.
Available at: https://scholar.smu.edu/datasciencereview/vol3/iss2/1

A Novel Methodology to Identify the Primary Topics Contained Within the COVID-19 Research Corpus

Allen Crane, Brock Friedrich,
William Fehlman, Igor Frolow, and Daniel W. Engels

Master of Science in Data Science
Southern Methodist University, Dallas TX 75275 USA
allen.s.crane@gmail.com, brocklfriedrich@gmail.com,
wfehlman@gmail.com, ifrolow@austin.rr.com, dwe@alum.mit.edu

Abstract. In this paper, we present a novel framework and system for the identification of primary research topics from within a corpus of related publications, the classification of individual publications according to these topics, and the results of the application of our framework and system to the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a corpus of published peer reviewed and pre-peer reviewed articles related to the coronavirus that causes COVID-19. Using machine learning techniques, such as Non-negative Matrix Factorization for Natural Language Processing and a Bayesian classifier, we developed a novel framework and system that automatically extracts sparse and meaningful features from the abstracts of the articles in CORD-19 allowing for primary topic identification and the classification of articles based upon these primary topics. The system uses an adaptive topic model classifier that allows for the identification of new primary topics as papers are added to CORD-19. New primary topics are added only when sufficiently many papers cover that topic. Using our system, we identified ten primary topics for the CORD-19 articles existing as of June 2020. The COVID-19 pandemic began in or around December 2019; therefore, the June 2020 CORD-19 dataset reflects the early research that has been performed related to COVID-19, as well as earlier coronaviruses related to previous epidemics such as SARS and MERS. The ten identified primary topics cover the breadth of the essential research questions that need to be answered in order to understand and find a cure or vaccine for COVID-19. This breadth and coverage demonstrates that beginning early in the pandemic, the research community began the investigation into all aspects of COVID-19 and the coronavirus that causes COVID-19, providing a broad foundation for the ending of the pandemic.

Keywords: coronavirus · COVID-19 · CORD-19 · Latent Dirichlet Allocation · Non-negative Matrix Factorization · topic modeling · adaptive topic modeling · visual analytics · interactive clustering · text analytics.

1 Introduction

We are facing an unprecedented public health crisis with the coronavirus pandemic and the lasting health and financial impacts of COVID-19. It is more important than ever to have the resources to answer critical questions that matter to both organizations and people. This includes having access to timely, detailed, and trustworthy data to think quickly and move fast in the identification, treatment, and prevention of COVID-19 [20].

In response to this global health crisis, a call to action has been issued to the world's artificial intelligence experts to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions[1]. Our research utilizes the COVID-19 Open Research Dataset (CORD-19), which represents the most extensive machine-readable coronavirus literature collection available for data mining to-date. As of June 9, 2020, CORD-19 contained 138,794 peer reviewed and pre-peer reviewed articles, and continues to grow. This open dataset is for the global research community to apply recent advances in natural language processing and other machine learning techniques to generate new insights in support of the ongoing fight against this highly infectious disease. There is a growing urgency for these automated and advanced analyses because of the rapid increase in both COVID-19 cases and the growing coronavirus literature that makes it more difficult for researchers to identify the advances and the papers in which they are being published.

In this paper, we present a novel framework and system for the identification of primary research topics from within a corpus of related publications, the classification of individual publications according to these topics, and the results of the application of our framework and system to the CORD-19 dataset. Unlike other indexed or searchable databases, such as the Allen Institute for AI CORD-19 explorer[2] or the National Institutes of Health National Library of Medicine[3], our method does not rely on keywords to be tagged for simple searches, but uses an innovative unsupervised topic modeling approach to classify new and emerging topics from the entire corpus, without the need for tagging or other supervised methods.

Using machine learning techniques, such as Non-negative Matrix Factorization for Natural Language Processing, we developed a novel framework and system that automatically extracts sparse and meaningful features from the abstracts of the articles in CORD-19. This feature extraction allows for primary topic identification and the classification of articles based upon these primary topics and the identification of new primary topics. Our framework and system is designed to adapt to the growing size of CORD-19 and the addition of new

[1] More information on this call to action and the CORD-19 Open Research Dataset Challenge can be found at https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[2] More info on the AI2 CORD-19 explorer can be found at https://cord-19.apps.allenai.org/

[3] To learn more about the NIH NLM, please refer to https://www.nlm.nih.gov/services/databases_subject.html

primary topics introduced by these yet to be added articles. We use an adaptive topic model Bayesian classifier that allows for the identification of new primary topics as papers are added to CORD-19.

Our Bayesian classifier is meant to be applied to recently published (or added) articles. As new research enters into the CORD-19 dataset, we apply a posterior probability to assign that observation into a topic probability or include the observation into a new/unknown bucket. As this new/unknown bucket grows, we refresh our model to include these new primary topics.

Using our system, we identified that the CORD-19 articles existing as of June 2020 contain ten primary topics. The COVID-19 pandemic began in or around December 2019; therefore, the June 2020 CORD-19 dataset reflects the early research related to COVID-19 that has been performed, as well as earlier coronavirus articles related to previous epidemics such as SARS and MERS.

The ten identified primary topics cover the breadth of the essential research questions that need to be answered in order to understand and find a cure or vaccine for COVID-19. This breadth and coverage demonstrates that beginning early in the pandemic, the research community began the investigation into all aspects of COVID-19 and the coronavirus that causes COVID-19, providing a broad foundation for the ending of the pandemic.

Our novel framework and system is applicable to a broad range of research corpora, particularly when the corpus is large, growing over time, and researchers need a means to quickly distill down the corpus into their specific area(s) of interest.

The remainder of this paper is organized as follows: In Section 2, we present the history of COVID-19. In Section 3, we provide an overview of topic modeling and its context in machine learning. In Section 4, we introduce the CORD-19 dataset and how it is created. In Section 5, we present our topic model design. We present our results and analysis in Section 6. In Section 7, we discuss ethical considerations for machine learning models applied to a large corpus. We draw the relevant conclusions in Section 8.

2 History of COVID-19

We begin with a brief history of COVID-19 and illustrate how researchers' understanding of the virus–in both its origin and spread into a global pandemic–is of great interest on many topics. Through this history, we connect the breadth of the background to the many types of research topics, and from there, our topic model itself.

COVID-19 is just one of the coronavirus epidemics of the past century. In addition, it reprsesents one of the increasing number of zoonoses (diseases which are transmitted to humans from animals) that impact global health. SARS CoV-2 (the virus causing the COVID-19 epidemic) is distinct from, but closely resembles, SARS CoV-1, which was responsible for the severe acute respiratory syndrome (SARS) outbreak in 2002. SARS CoV-1 and 2 share almost 80 percent of genetic sequences and use the same host cell receptor to initiate viral infection. However, SARS predominantly affected individuals in close contact with infected animals and health care workers. In contrast, CoV-2 exhibits robust person-to-person spread, most likely by means of asymptomatic carriers, which has resulted in greater spread of disease, overall morbidity and mortality, despite its lesser virulence [17].

In early December 2019, Li Wenliang, a physician from Wuhan, a large metropolitan area in China's Hubei province, reported in a group chat that he noticed a series of patients showing signs of a severe acute respiratory syndrome or SARS-like illness which was subsequently reported to the World Health Organization (WHO) Country Office in China on December 31, 2019. On January 12, Chinese scientists published the genome of the virus, and the WHO asked a team in Berlin to use that information to develop a diagnostic test to identify active infection, which was developed and shared four days later. On January 30, 2020, the outbreak was declared by the WHO a Public Health Emergency of International Concern (PHEIC). The first case of the disease due to local person-to-person spread in the United States was confirmed in mid-February 2020. On March 11, the WHO declared COVID-19 a pandemic [17].

SARS CoV-2 is similar to the SARS-CoV-1 virus, which caused the SARS epidemic originating in China in 2002 to 2003. Full-length genome sequences obtained from infected Chinese patients in 2019 showed 79.6 percent sequence homology to SARS-CoV-1. Because of the genome's homology with SARS-CoV-1, this virus was renamed SARS-CoV-2. Coronaviruses are known to be the causative agent for the "common cold," accounting for up to 30 percent of upper respiratory tract infections in adults. Coronaviruses, like other RNA viruses, mutate frequently and evolve in vast animal reservoirs. The overwhelming majority of coronaviruses pose no threat to humans, but recombination events, natural selection and genetic drift permit particularly virulent coronaviruses to jump to human hosts and to subsequently acquire the capacity for efficient person-to-person spread [10].

For reasons that are not well understood, zoonoses from wildlife have been increasing over the last half-century, and represent the most significant, growing threat to global health of all emerging infectious diseases. Geographic hotspots, or maps reflecting zoonotic infectious disease risk have been identified in South America, Africa, and South Asia. Both SARS-CoV-1 and SARS-CoV-2 arose in one of these hot-spots. Future outbreaks are believed to be all but inevitable [10] [18].

The SARS-CoV-1 epidemic in 2002 to 2003 was the first coronavirus pandemic in modern times, which spread to two dozen countries with approximately 8000 cases and 800 deaths before it was contained [6]. In 2012, another outbreak, referred to as MERS for the Middle Eastern Respiratory Syndrome, and also caused by a coronavirus, resulted in over 1000 infections and 400 deaths through 2015. Since SARS and MERS coronaviruses exhibited reduced person-to-person spread, the global impact of each was ultimately limited. Health care settings were the most frequent sites of person-to-person disease transmission [19] [6].

Healthcare workers and those in close contact with infected individuals were at greatest risk of contracting and succumbing to disease, while the general public was relatively spared during these outbreaks. In particular, otolaryngologists (ear, nose, and throat doctors) were at greater risk of infection due to shedding of virus from nasal and pharyngeal mucosa [1]. The implementation of infection control methods, aggressive contact tracing and isolation limited the spread of disease in 2003 and 2012. Antiviral treatments and vaccines were never developed [24].

SARS-CoV-2 is more closely related to two bat-derived severe acute respiratory syndrome (SARS)-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21, collected in 2018 in Zhoushan, eastern China (with 88 percent identity), and a strain isolated from pangolins (99 percent identity) than to SARS-CoV-1, suggesting SARS-CoV-2 should be considered a new human-infecting coronavirus, rather than re-emergence of SARS-CoV-1 [23]. Transmission is believed to have occurred from a bat-CoV to an intermediate host after which the virus jumped to human hosts at live animal markets [14]. Suggestions that SARS-CoV-2 was constructed in a laboratory have been widely discredited [9] [2].

The infectivity of SARS-CoV-2, relative to CoV-1, is likely accounted for by relatively high rates of asymptomatic carriers which promote the spread of infection to susceptible populations [3], rather than greater stability of virus in aerosolized droplets or on surfaces [7]. Despite variations and limitations in testing, CoV-2 appears to be less virulent than CoV-1 [23].

Finding relevant articles for a given focus area of COVID-19 research is challenging in this sea of information. The research varies widely, due to global authorship, language, localized and generalized issues, origin of the virus, virus pathology, transmission patterns, patient symptoms, patient treatments, physician and caregiver health, public safety, government regulations and policy, and many other themes. Given the large size of the CORD-19 data, researchers face the additional challenge of making sure they have access to the most relevant papers for their area of interest, while focusing less on the research that is pertinent to another topic. The use of topic modeling allows for faster search on issues, symptoms, and treatments of COVID-19, provides the "human-assist" machine learning insights to the researcher, and allows them to more quickly consume and act on the most relevant corpus of research for their area, with maximum efficiency in terms of article consumption. Table 1 lists example topics for the CORD-19 dataset and areas of research.

Table 1. Example Topics for the Growing CORD-19 Dataset.

Treatments	Risk Factors
Infection	History/Origin
Symptoms	Research Studies
Transmission	High Severity Patients
Public Health/Caregiver Risk	Genetics/Virology

Figure 1 shows the distribution of COVID-19 research papers by date of publication. Originating in 2000, coronavirus publications increased during and following the SARS and MERS epidemics, but the number of papers published in the early months of 2020 exploded in response to the COVID-19 epidemic. [12].

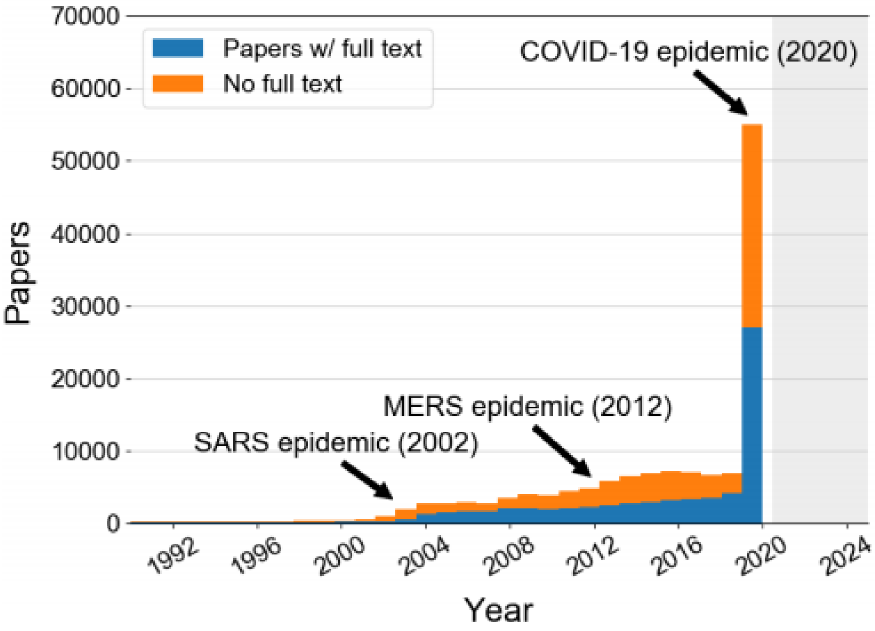


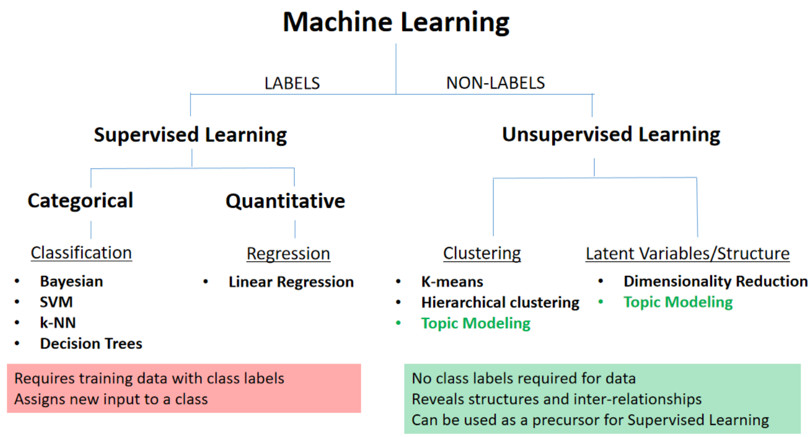
Fig. 1. COVID-19 Research Papers per Year (2000 - Present).

Two strategies are emerging with respect to making the large number of articles more searchable: creating easily accessible paper collections, including a few carefully curated collections designed to highlight strong papers; and building automated search tools that use artificial intelligence (AI) technologies to cut through the noise.[4]

3 Topic Modeling and Machine Learning

Reading hundreds of thousands of documents (as in the CORD-19 dataset) is a costly activity and an impossible task for humans to uncover all of the hidden patterns that could lead to actionable insights [15].

Topic modeling applies machine learning methods to uncover hidden patterns in unstructured data, at a speed and scale that is beyond humans. Machine learning can be broken down into two primary disciplines: **Supervised Learning**, in which a labeler (usually a person) affixes a label to the data to create a "known" dataset; and **Unsupervised Learning**–which does not use human labels, but rather, mathematical techniques to get to a similar outcome automatically. Supervised Learning lives in the world of regression, decision trees, and statistical methods that are easy to understand. By contrast, Unsupervised Learning requires us to "trust the machine", to tell us what it "sees", and report that as our new truth. Supervised learning is transparent, but new context is often overlooked. Unsupervised learning is more scalable, but it can be harder to explain, because categories are derived mathematically [8]. See Figure 2 for an illustration of topic modeling and its context in machine learning.



Requires training data with class labels
Assigns new input to a class

No class labels required for data
Reveals structures and inter-relationships
Can be used as a precursor for Supervised Learning

Fig. 2. Hierarchy of Supervised/Unsupervised Machine Learning and Topic Modeling.

[4] Read more about how the volume of COVID-19 articles is creating challenges for scientists here https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat

Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse, and summarize large archives of texts. [5] Topic modeling is a form of Unsupervised Learning for discovering the abstract "topics" that occur in a collection of documents and providing a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together.[6] For example, historians use topic modeling to identify important events in history by analyzing text, based on year. Web-based libraries use topic modeling to recommend books based on a person's past readings. News providers use topic modeling to understand articles quickly or cluster similar articles. Another interesting application is the unsupervised clustering of images, where each image is treated similar to a document [4]. [7]

It is important that we become familiar with certain terms in topic modeling. Table 2 contains common terms and their definitions for topic modeling that show up multiple times in the course of this paper and other research literature.

Table 2. Common Topic Modeling Terms and Definitions.

Term	Description
Document	A unit of text under analysis (e.g., typical document, paragraph, sentence, email, tweet, etc.)
Corpus	A collection of documents
Topic	A theme (or subject) discussed in one or more documents that is inferred based on a group of words that frequently occur together
Topic Model	An unsupervised machine learning algorithm that reveals the topics across a corpus and used to generate labeled topic categories
Adaptive Topic	A supervised machine learning model that assigns a document
Classification Model	(observation) to a topic category based on a degree of probability and provides the ability to update the classification model based on learning new emerging topics

[5] Additional discussion on topic modeling from Princeton can be found at http://www.cs.princeton.edu/ blei/topicmodeling.html

[6] See this link for University of Massachusetts discussion on topic modeling http://mallet.cs.umass.edu/topics.php

[7] More information on topic modeling use in image analysis can be found at http://en.wikipedia.org/wiki/topic_model

3.1 A Topic Modeling Example

We illustrate topic modeling using a methodology known as Latent Dirichlet Allocation (LDA).[8] LDA is an NLP technique that automatically discovers topics that documents contain. The concept behind it is very intuitive: Imagine a fixed set of topics, with each topic representing a set of words. The goal is to map all the documents to the topics in such a way, that the words in each document are mostly captured by those imaginary topics [5].

LDA is a probabilistic model, and to obtain cluster assignments, it uses two probability values: $P(\text{word} \mid \text{topics})$ and $P(\text{topics} \mid \text{documents})$. These values are calculated based on an initial random assignment, after which they are repeated for each word in each document, to decide their topic assignment. In an iterative procedure, these probabilities are calculated multiple times, until the convergence of the algorithm.

Consider the set of sentences shown in Table 3 for our example.

Table 3. The sentences for our LDA example.

Sentence
"I study **zoonotic diseases**."
"*Coronavirus* affects the *respiratory* system."
"The *patient* has **coronavirus**."

Given these sentences in Table 3, our algorithm might classify the **bolded** words under Topic O, which we label as “origin". Similarly, *italicized* words might be classified under a separate Topic S, which we label as "symptoms“. Each topic is essentially a bag-of-words, and we have to label the topics manually since LDA does not know the context or meaning of the words that it is classifying as belonging to the same topic. There are 2 primary benefits from defining topics on a word-level: 1) We can infer the content spread of each sentence by a word count as shown in Table 4, and 2) We can derive the proportions that each word constitutes in given topics. For example, Topic O might comprise words in the proportions shown in Table 5.

[8] More information on the LDA topic modeling tutorial can be found at https://algobeans.com/2015/06/21/laymans-explanation-of-topic-modeling-with-lda-2/

Table 4. Content Spread of Each Sentence by Word Count

Sentence	Topic O	Topic S
"I study **zoonotic diseases**."	100%	0%
"*Coronavirus* affects the *respiratory* system."	33%	67%
"The *patient* has **coronavirus**."	0%	100%

Table 5. Proportion of Each Word

Word Proportions in Topic O
33% **zoonotic**
33% **diseases**
33% **coronavirus**

Imagine that we are now discovering topics in documents instead of sentences, and we have two documents with the following words:

Table 6. Simple Topic Modeling Example With Two Documents:

Document X	Document Y
Coronavirus	Coronavirus
Coronavirus	Coronavirus
Zoonotic	Fever
Zoonotic	Patient
Diseases	Patient

Step 1 We tell the algorithm how many topics we think there are. We either use an informed estimate (e.g., results from a previous analysis), or simply trial-and-error. In trying different estimates, we pick the one that generates topics to your desired level of interpretability (i.e., coherence), or the one yielding the highest statistical certainty (i.e., log likelihood). In our Table 6 example, the number of topics might be inferred just by eyeballing the documents.

Step 2 The algorithm will assign every word to a temporary topic. Topic assignments are temporary as they will be updated in Step 3. Temporary topics are assigned to each word in a semi-random manner. This also means that if a word appears twice, each word may be assigned to different topics. Note that in analyzing actual documents, stop words (e.g., “the”, “and”, “my”) are removed and not assigned to any topics.

Step 3 (iterative) The algorithm will check and update topic assignments, looping through each word in every document. For each word, its topic assignment is updated based on two criteria: How prevalent is that word across topics? And how prevalent are topics in the document?

To understand how these two criteria work, imagine that we are now checking the topic assignment for the word “coronavirus” in Document Y. See Table 7.

Table 7. Topic Assignment for the word "coronavirus" in Document Y

Topic	Document X	Topic	Document Y
O	Coronavirus	?	Coronavirus
O	Coronavirus	O	Coronavirus
O	Zoonotic	O	Fever
O	Zoonotic	S	Patient
O	Diseases	S	Patient

Step 3a. How prevalent is the word across topics? Since “coronavirus” words appear in both documents, and comprise nearly half of all remaining Topic O words (but zero percent of Topic S words), a “coronavirus” word picked at random would more likely be about Topic O. See Table 8.

Table 8. Step 3a. Topic Assignment for the word "coronavirus" in Document Y

Topic	Document X	Topic	Document Y
O	**Coronavirus**	?	Coronavirus
O	**Coronavirus**	**O**	**Coronavirus**
O	Zoonotic	O	Fever
O	Zoonotic	S	Patient
O	Diseases	S	Patient

Step 3b. How prevalent are topics in the document? Since the words in Document Y are assigned to Topic O and Topic S in a 50/50 ratio, any

subsequent “coronavirus” word seems equally likely to be about either topic. See Table 9.

Table 9. Step 3b. Topic Assignment for the word "coronavirus" in Document Y

Topic	Document X	Topic	Document Y
O	Coronavirus	?	Coronavirus
O	Coronavirus	**O**	**Coronavirus**
O	Zoonotic	**O**	**Fever**
O	Zoonotic	**S**	**Patient**
O	Diseases	**S**	**Patient**

Weighing conclusions from the two criteria, we would assign the “coronavirus” word of Document Y to Topic O. Document Y might then be a document on patient symptoms.

The process of checking topic assignment is repeated for each word in every document, cycling through the entire collection of documents multiple times. This iterative updating is the key feature of LDA that generates a final solution with coherent topics.[9]

3.2 Non-negative Matrix Factorization (NMF)

The aforementioned example uses LDA to explain how topic modeling works for a non-technical reader. For our analysis, we use another NLP technique that allows for greater resolution of emerging topics, so that we do not miss important themes in our topic discovery.

Non-negative Matrix Factorization (NMF) is a Linear-algebraic model, that factors high-dimensional vectors into a low-dimensionality representation. Similar to Principal Component Analysis (PCA), NMF takes advantage of the fact that the vectors are non-negative. By factoring them into the lower-dimensional form, NMF forces the coefficients to also be non-negative.

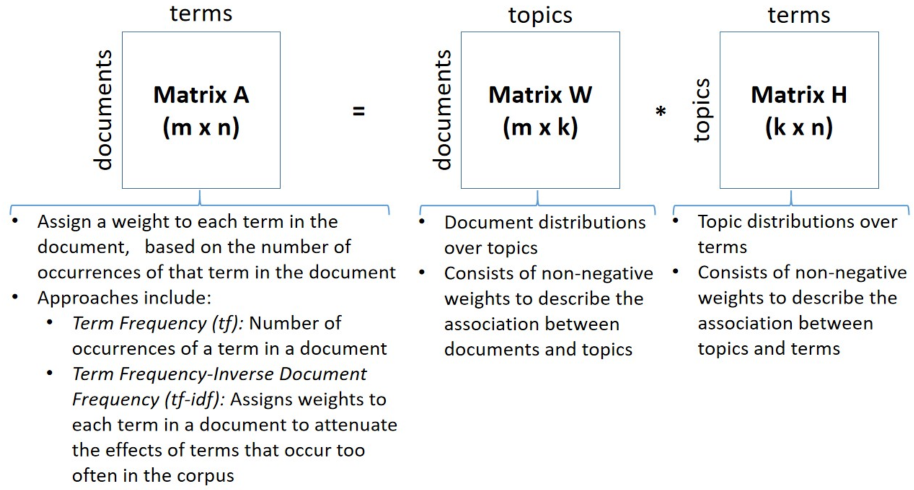
There are pros and cons of both techniques. NMF and LDA are similar in terms of creating topics from a corpus using unsupervised machine learning. However, Non-negative Matrix Factorization (NMF) has become a widely used tool for high dimensional data, like text analysis, as it automatically extracts sparse and meaningful features from a set of non-negative data vectors [22]. It does this using non-negative weights (common approaches include Term Frequency - Inverse Document Frequency [TF-IDF]). Not wanting to miss key (but perhaps less frequent) terms, this is the approach we used in our analysis. Table 10 describes the steps used in NMF.

[9] Additonal information on how topics are assigned can be found at http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

Table 10. Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization Steps [11]
1. Matrix A = Matrix W * Matrix H
2. A = The document-term matrix (dtm) for corpus of m documents and n terms (m X n)
3. NMF takes A as an input and factorizes it into two non-negative matrices W and H having k dimensions (i.e., number of topics)
4. W = The document-topic (m x k) matrix where the rows of W provide weights for the input documents relative to the k topics - these values indicate the strength of association between documents and topics.
5. H = The topic-term (k x n) matrix where the columns of the matrix H provide the weights for the terms (columns) relative to the topics (rows). By ordering the values in a given column and selecting the top-ranked terms, we can produce a description (theme) of the corresponding topic.
6. Proportionally scale the document-topic matrix: To make the analysis and visualization of NMF components similar to that of LDA's topic proportions, we scale the document-topic matrix such that the topic values associated with each document sum to one.
7. Proportionally scale the topic-term matrix: Scale the topic-term matrix such that the term values associated with each topic sum to one.

Consider the bag-of-words matrix representation where each row corresponds to a word, and each column to a document. NMF will produce two matrices W and H. The columns of W can be interpreted as basis documents (bags of words). The interpretation we give to such a basis document is considered a topic, consisting of sets of words found simultaneously in different documents. H tells us how to sum contributions from different topics to reconstruct the word mix of a given original document. Figure 3 illustrates the matrices of NMF and their functions.



- Assign a weight to each term in the document, based on the number of occurrences of that term in the document
- Approaches include:
  - *Term Frequency (tf):* Number of occurrences of a term in a document
  - *Term Frequency-Inverse Document Frequency (tf-idf):* Assigns weights to each term in a document to attenuate the effects of terms that occur too often in the corpus

- Document distributions over topics
- Consists of non-negative weights to describe the association between documents and topics

- Topic distributions over terms
- Consists of non-negative weights to describe the association between topics and terms

Fig. 3. Non-negative Matrix Factorization (NMF).

4 COVID-19 Open Research Dataset (CORD-19)

In this section, we present the data we used in our research. The COVID-19 Open Research Dataset (CORD-19) is a compendium of all available coronavirus research. This dataset was created by the Allen Institute for AI in partnership with the Chan Zuckerberg Initiative, Georgetown University's Center for Security and Emerging Technology, Microsoft Research, IBM, and the National Library of Medicine - National Institutes of Health, in coordination with The White House Office of Science and Technology Policy [21].[10] As of June 9, 2020, the CORD-19 dataset contained 138,794 articles, corresponding to all currently available COVID-19 and coronavirus-related research (e.g., SARS, MERS, etc). Figure 4 illustrates the data sources of CORD-19.

This dataset is intended to mobilize researchers to apply recent advances in natural language processing to generate new insights in support of the fight against this infectious disease. The corpus is updated regularly as new research is published in peer-reviewed publications and archival services like bioRxiv, medRxiv, and others. [11]

CORD-19 sources include: PubMed's open access corpus using COVID-19 and coronavirus research, COVID-19 research articles from a corpus maintained

[10] Information on how the CORD-19 dataset was compiled and how it is curated can be found at https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[11] Additional information on CORD-19 data collection sources can be found at https://www.semanticscholar.org/paper/CORD-19/A-The-Covid-19-Open-Research-Dataset-Wang-Lo/bc411487f305e451d7485e53202ec241fcc97d3b/figure/0

by the WHO bioRxiv and medRxiv pre-prints using the same query as PMC (COVID-19 and coronavirus research), papers and preprints collected by Semantic Scholar, an AI-powered research tool for scientific literature, based at the Allen Institute for AI. Metadata are harmonized and deduplicated, and document files are processed to extract full text. A comprehensive metadata file exists for coronavirus and COVID-19 research articles with links to PubMed, Microsoft Academic and the WHO COVID-19 database of publications. Note: Use metadata from the comprehensive file when available instead of parsed metadata in the dataset. The dataset may contain multiple entries for PMC IDs when supplementary materials are available. The full repository is linked to WHO database of publications on coronavirus disease and other resources.[12]

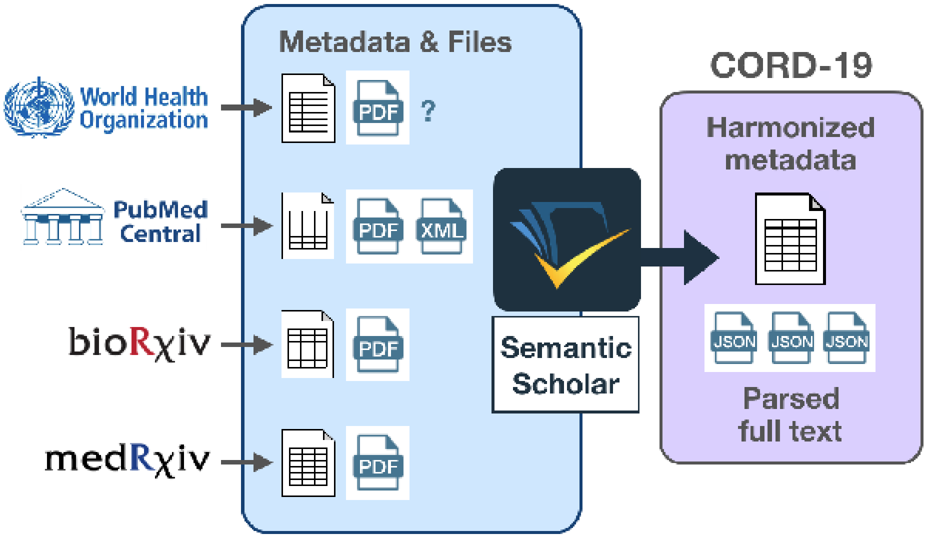


Fig. 4. Sources of papers included in CORD-19. Papers and preprints are collected by Semantic Scholar. Metadata are harmonized and deduplicated, and document files are processed to extract full text.

The CORD-19 research challenge includes key scientific questions of interest, which are drawn from the NASEM's SCEID (National Academies of Sciences, Engineering, and Medicine's Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats) research topics and the World Health Organization's Research and Development Blueprint COVID-19. Many of these questions are suitable for text mining. See Table 11 for the CORD-19 Research Challenge questions of interest.

[12] More information on CORD-19 data sources can be found at https://www.semanticscholar.org/cord19/download

Table 11. CORD-19 Questions of Interest

Questions of Interest From the CORD-19 Research Challenge
What is known about transmission, incubation, and environmental stability?
What do we know about natural history, transmission, and diagnostics for the virus?
What have we learned about infection prevention and control?
What do we know about COVID-19 risk factors?
What do we know about virus genetics, origin, and evolution?
What do we know about vaccines and therapeutics?
What do we know about non-pharmaceutical interventions?
What do we know about diagnostics and surveillance?
What has been published about medical care?
What has been published about ethical and social science considerations?
What has been published about information sharing and inter-sectoral collaboration?

While our research does not answer these questions directly, we created topics that map extremely well to the above questions of interest. In doing so, researchers interested in these topics can use our mapping to quickly obtain the relevant articles to study these questions of interest.

The CORD-19 dataset is organized as a single large file with each row representing a research article. Along with a master identifier for each row, each article contains identifiers from the various sources of research data (e.g., PubMed, ArXiv, WHO, etc) for cross-reference. The abstracts of each article are contained in a single field. These metadata are listed in Table 12.

Table 12. Metadata Descriptions for CORD-19 Dataset.

Field Name	Metadata Description
cord_uid	A string-valued field that assigns a unique identifier to each CORD-19 paper. This is not necessarily unique per row, as in the case of an article sourced from multiple sources.
sha	A List[string]-valued field that is the SHA1 of all PDFs associated with the CORD-19 paper. Most papers will have either zero or one value here, but some papers will have multiple.
source_x	A List[string]-valued field that are the soures for this paper. Semicolon-separated. For example, 'ArXiv; Elsevier; PMC; WHO'. There should always be at least one source listed.
title	A string-valued field for the paper title.
doi	A string-valued field for the paper DOI.
pmcid	A string-valued field for the paper's ID on PubMed Central. Should begin with PMC followed by an integer.
pubmed_id	An integer-valued field for the paper's ID on PubMed.
license	A string-valued field with the most permissive license we've found associated with this paper. Possible values include: 'cc0', 'hybrid-oa', 'els-covid', 'no-cc', 'cc-by-nc-sa'.
abstract	A string-valued field for the paper's abstract.
publish_time	A string-valued field for the published date of the paper. This is in yyyy-mm-dd format. Not always accurate as some publishers will denote unknown dates with future dates like yyyy-12-31.
authors	A List[string]-valued field for the authors of the paper. Each author name is in Last, First Middle format and semicolon-separated.
mag_id	Deprecated, but originally an integer-valued field for the paper as represented in the Microsoft Academic Graph.
who_covidence_id	A string-valued field for the ID assigned by the WHO for this paper. Format looks like #72306.
arxiv_id	A string-valued field for the arXiv ID of this paper.
pdf_json_files	A List[string]-valued field containing paths from the root of the current data dump version to the parses of the paper PDFs in JSON format. Example: 'document_parses/pdf_json/4eb6e165ee705e2ae2a24ed2d4e-67da42831ff4a.json'.
pmc_json_files	A List[string]-valued field. Same as above, but corresponding to the full text XML files downloaded from PMC, parsed integero the same JSON format as above.
url	A List[string]-valued field containing all URLs associated with this paper. Semicolon-separated.
s2_id	A string-valued field containing the Semantic Scholar ID for this paper. (e.g., s2_id=9445722 corresponds to http://api.semanticscholar.org/corpusid:9445722).

5 Model Design

Our novel approach to this research is illustrated in Figure 5. We began with the CORD-19 corpus of 138,794 articles (as of June 9, 2020). In pre-processing, we first eliminated duplicates, Since our text mining was based on the text of the abstracts, we eliminated papers without abstracts. At the time of this paper, 98.89% of the papers were in English, so we elected to remove those papers that were not written in English. After these steps, our examined dataset contained 116,284 articles. We ran topic modeling on these documents, and assigned these documents to primary topics. Our model expands on this idea to include an adaptive topic classifier. Figure 5 illustrates our approach to the research.

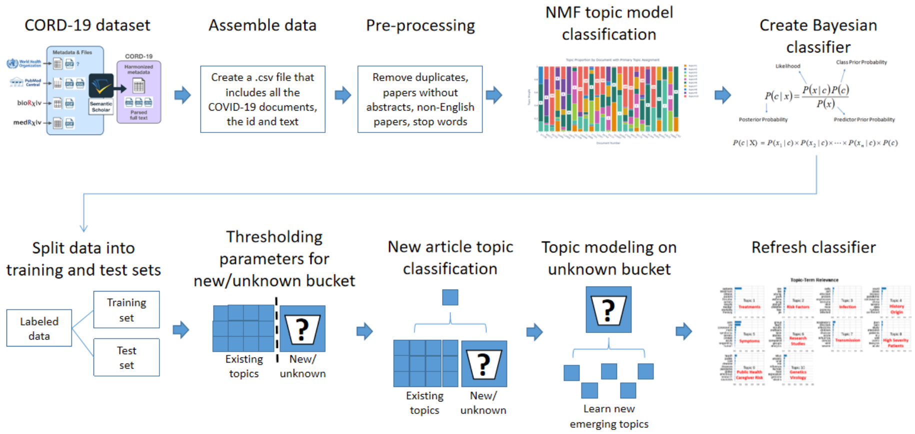

CORD-19 dataset
World Health Organization
PubMed Central
bioRχiv
medRχiv
Metadata & Files
Semantic Scholar
CORD-19
Harmonized metadata
Parsed full text
Assemble data
Create a .csv file that includes all the COVID-19 documents, the id and text
Pre-processing
Remove duplicates, papers without abstracts, non-English papers, stop words
NMF topic model classification
Topic Proportion by Document with Primary Topic Assignment
Document Number
Create Bayesian classifier
Likelihood
Class Prior Probability
$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$
Split data into training and test sets
Labeled data
Training set
Test set
Thresholding parameters for new/unknown bucket
Existing topics
New/ unknown
New article topic classification
Existing topics
New/ unknown
Topic modeling on unknown bucket
Learn new emerging topics
Refresh classifier
Topic-Term Relevance
Topic 1 Treatments
Topic 2 Risk Factors
Topic 3 Infection
Topic 4 History Origin
Topic 5 Symptoms
Topic 6 Research Studies
Topic 7 Transmission
Topic 8 High Severity Patients
Topic 9 Public Health Caregiver Risk
Topic 10 Genetics Virology

Fig. 5. Novel Framework for Topic Modeling, Article Classification, and New Topic Identification (Crane, Friedrich)

We conducted topic modeling as a precursor to building a supervised classification model. We chose Non-negative Matrix Factorization (NMF) to identify categories of primary topics across our corpus of documents, as it automatically extracts sparse and meaningful features from a set of non-negative data vectors.

Our Bayesian classifier incorporates the prior probabilities of each possible class label. We encoded topic categories using embeddings, based on the most relevant terms in each topic category, and we built the new classification model using these embeddings. We then applied Laplace smoothing to give a probability estimate of the term x, given class y. The number of occurrences of term x in the documents from class y, plus 1 for smoothing, divided by the total number of terms in the documents for class y, plus the number of terms in the vocabulary across the entire corpus (for all classes). This returns the posterior probability for each class (or document) given the feature vector x document [16].[13]

Using these topics and labels as our classifier, when a new document comes in, we apply a posterior probability to assign that observation into a topic probability. If the observation exceeds a threshold for our existing topics, we put it into an unknown bucket. As this unknown bucket grows, we run topic modeling on it to learn about new or emerging topics. We then refresh our overall topic model to include these new emerging topics. The primary topic is assigned to the topic with the largest weight in each document.

For our accuracy score, we chose to use Coherence. In linguistics, Coherence is what makes a text semantically meaningful. It is especially dealt with in text linguistics. Coherence is achieved through syntactical features such as the use of deictic, anaphoric and cataphoric elements or a logical tense structure, as well as presuppositions and implications connected to general world knowledge.[14] In other words, coherence helps us to make sense of topics that are easy to interpret and distinct from other topics.

Given the number and level of parameter settings that we need to try, we implemented a grid search technique to compare the model outputs against each other. Table 14 lists the tuning parameters and the settings that were tested to determine the optimal model, based on coherence score. Our three parameters contained five, three, and three levels, respectively, so our 5 x 3 x 3 grid generated 45 different models to determine the best combination. For our success criteria, we chose the model that generated topics yielding the highest interpretability (i.e., coherence score). [15] [16]

[13] More information on Laplace smoothing can be found at https://bookdown.org/rdpeng/advstatcomp/laplace-approximation.html

[14] For more information on Coherence in linguistics, see https://en.wikipedia.org/wiki/Coherence_(linguistics)

[15] More information on the alpha hyperparameter used in our grid search can be found at https://datascience.stackexchange.com/questions/199/what-does-the-alpha-and-beta-hyperparameters-contribute-to-in-latent-dirichlet-a

[16] Additional discussion on regularization ratios used in our grid search can be found at https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

Table 14. Grid Search Parameters and Settings for Optimal Model Output

Grid Search Parameters and Settings
Number of topics: [10, 15, 20, 25, 30] – These are the number of topics that we will iterate through for each combination of parameter values. We chose these because we speculated that 10 topics may make interpretability too general, while 30 topics may introduce too much overlap between topics, and 20 being in the middle of this range, to test if there was a linear relationship with our evaluation criteria (coherence).
Alpha: [0.01, 0.05, 0.1] – This is our initial belief about this distribution, and is the parameter of the prior distribution.
L1 Ratio: [0.1, 0.5, 0.9] - L1 Regularization has the effect of shrinking the less important feature's coefficient to zero, or remove them altogether. This can work well for feature selection in case we have a large number of features.

Figure 6 validates that our highest performing topic model (with a coherence score of 0.4754) contains 10 topics. And, Table 15 is an example deployment of our topic-to-article reference lookup database. The Primary Topic is assigned from the highest proportion of that topic's weight, relative to all other candidate topics for each document. The Primary Topic Weight column shows the proportion of Primary Topic in each document.

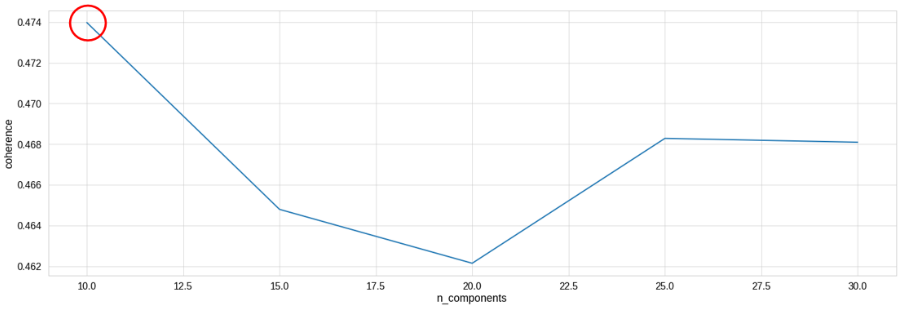


Fig. 6. Best Coherence by Number of Topics: 10 topics, 0.4754 coherence

Table 15. Example Topic-to-Article Reference Lookup Database

Original Document	Primary Topic Terms	Primary Topic	Primary Topic Weight
a summary of informa...	health public care pandemic...	8	0.310342
porcine transmissible...	virus influenza human host...	9	0.629781
successful long-term...	health public care pandemic...	8	0.310342
reductionist approach dissecting...	health public care pandemic...	8	0.310342
response covid-19 pandemic...	health public care pandemic...	8	0.310342
hypercholesterolaemia risk fac...	group study results methods...	5	0.398522
bat-origin coronaviruses...	health public care pandemic...	8	0.310342
clostridium difficile positive...	cell infection protein expression...	2	0.601955
possible association between cov...	covid cases disease pandemic...	3	0.703295
chloroquine and hydroxychloro...	patients treatment clinical hospital...	0	0.421012

Figure 7 illustrates how the primary topic assignment is labeled for each document, based on the topic with the highest proportional weight for each document. For example, the y-axis is the proportion of weight (from 0 to 1) of each topic in a particular document.

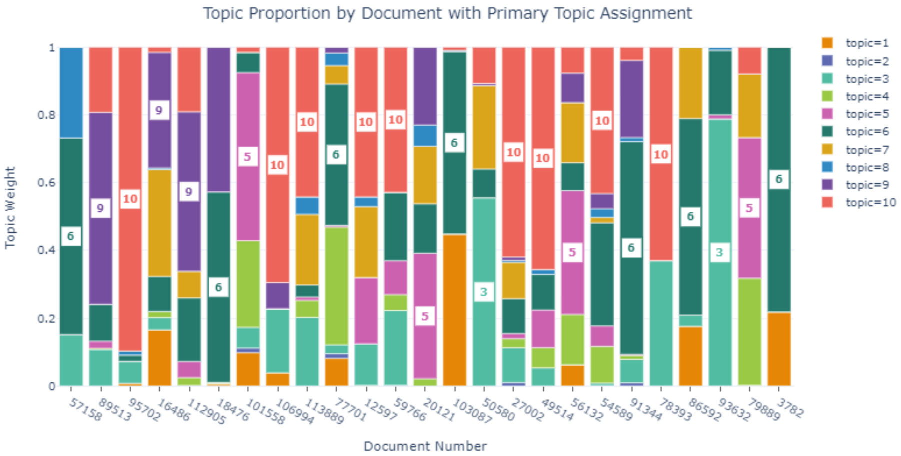


Fig. 7. Primary Topic Assignment by Document. The x-axis represents individual documents and they-axis represents the proportion of weight (from 0 to 1) of each topic in the document.

6 Results and Analysis

From our analysis of the CORD-19 dataset, our best topic model generated ten key topics, representing the model with the highest interpretability (i.e., coherence) score. Figure 8 illustrates the topic-term relevance and top key words associated with each of the ten topics, along with our chosen topic names, and Table 16 provides a short descriptive summary of each of these topics.

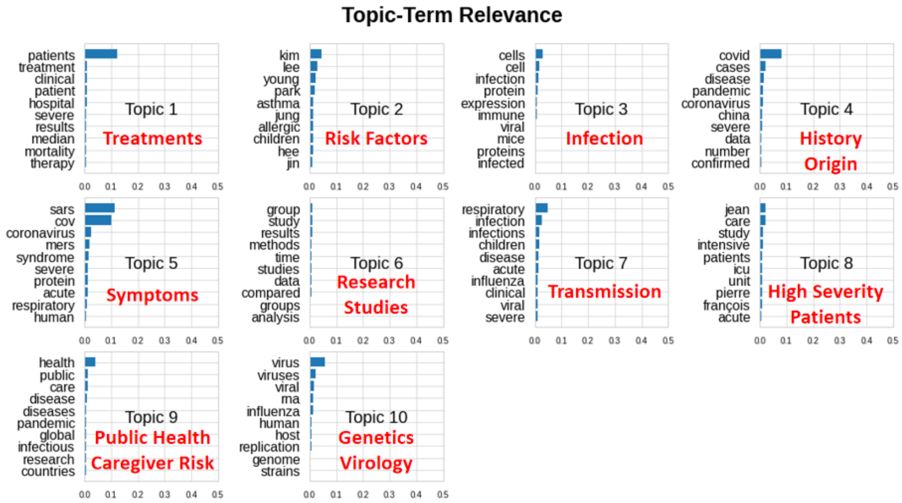


Fig. 8. Topic-Term Relevance, Best Model, n = 10 topics.

Table 16. CORD-19 Primary Topics

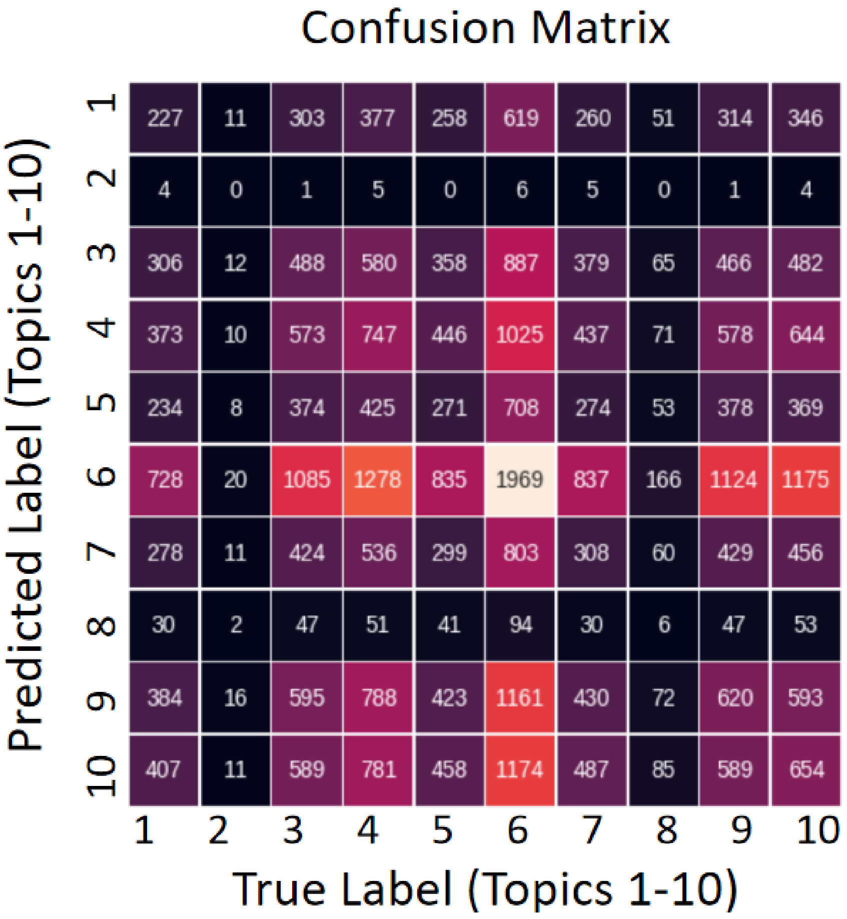
Summary Of Topic Generated By Our Model
Topic 1 refers to articles on patient treatment in hospitals and health/mortality
Topic 2 is a consolidation of risk factors
Topic 3 pertains to infection pathology at the virus and cellular level
Topic 4 relates to articles on the origin and growth rate of the pandemic
Topic 5 groups articles on symptoms of other coronaviruses
Topic 6 contains patient research studies and clinical trials
Topic 7 refers to transmission patterns
Topic 8 pertains to high severity patients with intensive care needs
Topic 9 deals with articles related to public health and caregiver risk
Topic 10 contains articles related to genetics and virology

Bringing it all together, Table 17 connects the CORD-19 Research Challenge questions of interest with our primary topics. Our novel framework produced topics that map well to the CORD-19 questions of interest.

Table 17. Questions of Interest That Map to Topics

Question of Interest	Primary Topic Name	Primary Topic
What is known about transmission, incubation, and environmental stability?	Transmission	7
What do we know about natural history, transmission, and diagnostics for the virus?	History and Origin	4
What have we learned about infection prevention and control?	Infection	3
What do we know about COVID-19 risk factors?	Risk Factors	2
What do we know about virus genetics, origin, and evolution?	Genetics and Virology	10
What do we know about vaccines and therapeutics?	Treatments	1
What do we know about non-pharmaceutical interventions?	Treatments	1
What do we know about diagnostics and surveillance?	Symptoms	5
What has been published about medical care?	Public Health and Caregiver Risk	9
What has been published about ethical and social science considerations?	Public Health and Caregiver Risk	9
What has been published about information sharing and inter-sectoral collaboration?	Research Studies	6

In addition, we created a Bayesian classifier for new or subsequent research documents. We evaluated these results with a confusion matrix (see Figure 9), which compared the true label (the original topic we observed, from prior model) against the predicted label (the topic we think the new article belongs to). Ideally, we would expect to see higher numbers on the diagonal from the top left to the bottom right. As we can see in our confusion matrix heatmap, the prediction model performed reasonably well only on Topic 6 (groups of articles on research studies, methods, and data analysis).



	1	2	3	4	5	6	7	8	9	10
1	227	11	303	377	258	619	260	51	314	346
2	4	0	1	5	0	6	5	0	1	4
3	306	12	488	580	358	887	379	65	466	482
4	373	10	573	747	446	1025	437	71	578	644
5	234	8	374	425	271	708	274	53	378	369
6	728	20	1085	1278	835	1969	837	166	1124	1175
7	278	11	424	536	299	803	308	60	429	456
8	30	2	47	51	41	94	30	6	47	53
9	384	16	595	788	423	1161	430	72	620	593
10	407	11	589	781	458	1174	487	85	589	654

Fig. 9. Confusion Matrix Heatmap. Higher numbers are correlated with lighter colors (better results), while lower numbers are correlated with darker gradient colors (worse results). Ideally, we would expect to see higher numbers on the diagonal from the top left to the bottom right. This would show a high correlation between the predicted label and the true label (note that the 0-9 labels correspond to topics 1-10, respectively). As we can see in our confusion matrix heatmap, the prediction model performed reasonably well only on Topic 6 (groups of articles on research studies, methods, and data analysis).

There are many reasons why this type of model is difficult to use in a prediction setting, namely the overlapping of terms. Recall how in our tutorial topic example when the word "coronavirus" appeared in multiple topics, and we used a probability weighting to place the word into the best topic, based on our knowledge.

Also, we cannot expect the original topics to remain static over time, especially as new information comes in. Continued accuracy and refinement suggests that we rebuild the topic model as each new document arrives. But this creates the challenge of an ever-shifting set of topics-to-articles, requiring our reference mechanism to be constantly updated. While this results in a more accurate database, we must consider the practical usage for a researcher whose topics may be changing over time, and balance our desire for accuracy with the researcher's desire for use and reproducibility. For these reasons, incremental topic classification has room for refinement.

This is illustrated in Figure 10, where each vertical bar on the X axis represents an individual research article. Colored sections of each bar represent the different possible topics for each article. The label on each vertical bar is the primary topic that was assigned, based on the proportional weight of that topic, relative to other topics. In some cases, the weight of the assigned primary topic rivals the secondary topic by only a tiny margin. For example, Document 103087 in the middle of the chart, contains a green proportion of the bar (representing Topic 6, at 53 percent) that is just slightly larger than the orange proportion of the bar (representing Topic 1, at 45 percent). In other cases, documents contain a multitude of topics, and a topic can be assigned with only a small proportional weight, as long as that weight is higher than all the others.

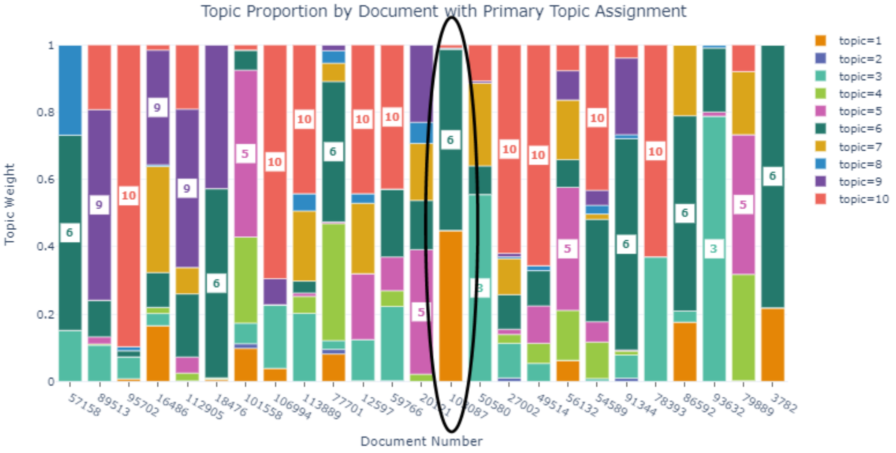


Fig. 10. Potentially ambiguous topic classification due to close proportions and overlap

Another way to visualize topic overlap, particularly when exploring high-dimensional data is called t-distributed Stochastic Neighbor Embedding (t-SNE) (see Figure 11). Introduced by Van Der Maaten and Hinton in 2008, the technique has become widespread in the field of machine learning, since it can create compelling two-dimensional "maps" from data with high dimensionality. The goal of t-SNE is to take a set of points in a high-dimensional space and find a faithful representation of those points in a lower-dimensional space, typically the 2D plane. The algorithm is non-linear and adapts to the underlying data, performing different transformations on different regions. [17]

As we can also see in this view, there is enough overlap between topics that accurately predicting (where the predicted label = the true label) can be very difficult to achieve. This is yet another explanation for why our prediction model has room for improvement.

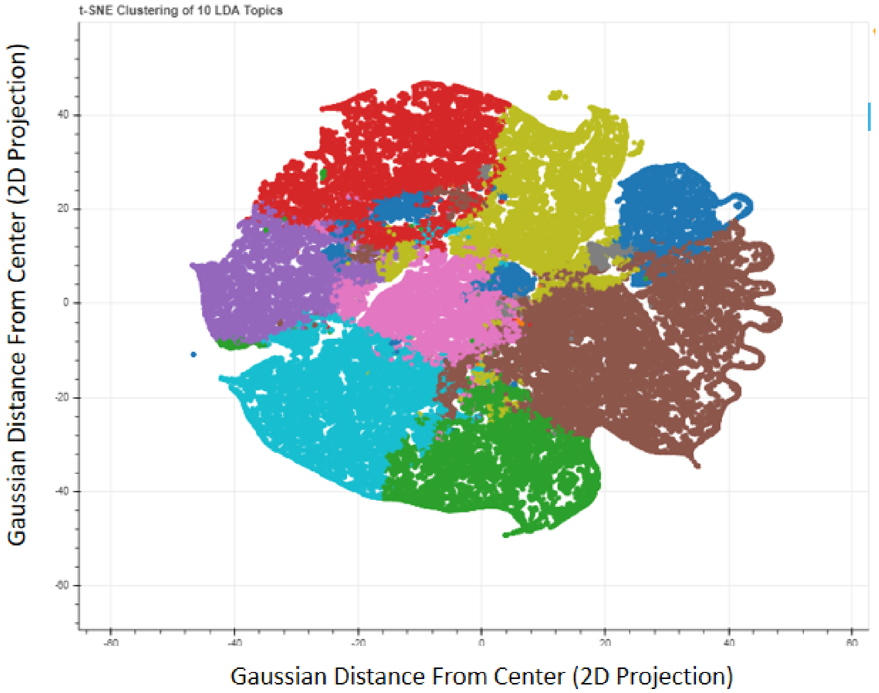


Fig. 11. t-SNE Topic Clusters for our Best Model, n = 10 topics

[17] More on t-SNE visualization and methodology can be found at https://distill.pub/2016/misread-tsne/

7 Ethical Considerations

When building and deploying machine learning systems, it is important to understand the ethical considerations of the work. Ethics are not a separate business objective bolted on after a system has been deployed. They are part of system performance. Only by embedding ethical principles into our applications and processes can we build systems that people can trust.

The IEEE Code of Ethics provides us with ethical standards for AI systems, such as transparency, understanding, and the societal implications of conventional and emerging technologies, including intelligent systems.[18]

Data and insights belong to the people and businesses who created them. Organizations that collect, store, manage, or process data have an obligation to handle it responsibly. Also, knowing how a model arrives at an outcome is key to trust. To improve transparency, developers must define how to build, deploy, and manage these systems through scientific research.

Controlling for bias is essential when dealing with complex models, whose outcomes are stated as fact. Unbiased models and a spirit of diversity and inclusion are necessary to create fair machine learning systems, which can mitigate, rather than accelerate, our existing prejudices. [19]

If more rigor is required in an organization to establish ethical behavior as a default, installing a process likely increases the awareness, as it assigns responsibility and improves consistency of procedure and outcome. An ethics review board for companies and individuals (as already implemented by universities for use in the context of the approval of experiments with human and animal subjects) included in the discussion of new tools, products, services, or planned research experiments should be considered, very much like it already exists in university environments in an experimental natural science context [13].

8 Conclusions

In this paper, we present a novel framework and system for the identification of primary research topics from within a corpus of related publications, the classification of individual publications according to these topics, and the results of the application of our framework and system to the COVID-19 Open Research Dataset (CORD-19). Unlike other indexed or searchable databases, our method does not rely on keywords to be tagged for simple searches, but uses an innovative unsupervised topic modeling approach to classify new and emerging topics from the entire corpus, without the need for tagging or other supervised methods.

Using our system, we identify ten primary topics for the CORD-19 articles existing as of June 2020. These topics accurately map to key questions of interest for the medical community, based on interpretability (i.e., coherence score). The

[18] More discussion on ethics in technology can be found in the IEEE Code of Ethics. https://www.ieee.org/about/corporate/governance/p7-8.html

[19] IBM Watson has developed extensive references, papers, and best practices on ethical AI. https://www.ibm.com/watson/ai-ethics/

ten identified primary topics cover the breadth of the essential research questions that need to be answered in order to understand and find a cure or vaccine for COVID-19.

Our novel framework is applicable to a broad range of research corpora, particularly when the corpus is large, growing over time, and researchers need a means to quickly distill down the corpus into their specific area(s) of interest.

References

1. Alexander, A., Tan, A., Evans, G., Allen, J.: Infection control for the otolaryngologist in the era of severe acute respiratory syndrome. Journal of Otolaryngology pp. 281–287 (2003)
2. Andersen, K., Rambaut, A., Lipkin, W., Holmes, E., Garry, R.: The proximal origin of sars-cov-2. National Medicine pp. 450–452 (2020)
3. Bai, Y., Yao, L., Wei, T.: Presumed asymptomatic carriertransmission of covid-19. JAMA p. 323 (2020)
4. Blei, D.M.: Probabalistic topic models. Communications of the ACM **55(4)**, 77–84 (2012)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
6. Chan, J., Li, K., To, K., Cheng, V., Chen, H., Yuen, K.: Is the novel human betacoronavirus 2c emc/2012(hcov-emc) the beginning of another sars-like pandemic? Infection pp. 477–489 (2012)
7. Doremalen, N.V., Bushmaker, T., Morris, D.: Aerosol andsurface stability of sars-cov-2 as compared with sars-cov-1. New England Journal of Medicine pp. 1564–1567 (2020)
8. Fehlman, W.: Applying topic models to uncover unknown unknowns in unstructured data. In: Applying Topic Models to Uncover Unknown Unknowns in Unstructured Data. The Data Science Conference, University of Chicago (6 2017)
9. Hao, P., Zhong, W., Song, S., Fan, S., Li., X.: Is sars-cov-2 originated from laboratory? a rebuttal to the claim of formation via laboratory recombination. Emerging Microbes Infection pp. 545–547 (2020)
10. Jones, K., Patel, N., Levy, M.: Global trends in emerging infectious diseases. Nature pp. 990–993 (2008)
11. Kuang, D., Choo, J., Park, H.: Non-negative matrix factorization for interactive topic modeling and document clustering. IEEE Computer Society (2013)
12. Leidner, J.L., Plachouras, V.: Ethical by design: Ethics best practices for natural language processing. Proceedings of the First Workshop on Ethics in Natural Language Processing pp. 30–40 (2017)
13. Leidner, J.L., Plachouras, V.: Ethical by design: Ethics best practices for natural language processing. Proceedings of the First Workshop on Ethics in Natural Language Processing pp. 30–40 (2017)
14. Li, J., You, Z., Wang, Q.: The epidemic of 2019-novel-coronavirus (2019-ncov) pneumonia and insights for emerging infectious diseases in the future. Microbes Infection pp. 80–85 (2020)
15. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, 40 W. 20 St. New York, NY United States (7 2008)
16. Miller, J.B., Gelman, A.: Laplace's theories of cognitive illusions, heuristics, and biases. IEEE Computer Society (2018)

17. Miriam M. Lango, M., Gelman, A.: How did we get here? short history of covid-19 and othercoronavirus-related epidemics. Wiley Periodicals pp. 1535–1538 (2020)
18. Parrish, C., Holmes, E., Morens, C.: Cross-species virustransmission and the emergence of new epidemic diseases. Microbiol Molecular Biological Review pp. 457–470 (2008)
19. Peeri, N., Shrestha, N., Rahman, M.: The sars, mersand novel coronavirus (covid-19) epidemics, the newest andbiggest global health threats: what lessons have we learned? International Journal of Epidemiology (2020)
20. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R.M., Liu, Z., Merrill, W., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D.M., Weld, D.S., Etzioni, O., Kohlmeier, S.: Cord-19: The covid-19 open research dataset. ArXiv **abs/2004.10706** (2020)
21. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R.M., Liu, Z., Merrill, W., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D.M., Weld, D.S., Etzioni, O., Kohlmeier, S.: Cord-19: The covid-19 open research dataset. ArXiv (2020)
22. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval pp. 267–273 (2003)
23. Yang, S., Cao, P., Du, P.: Early estimation of the case fatalityrate of covid-19 in mainland china: a data-driven analysis. Annals of Translated Medicine p. 128 (2020)
24. Zhou, P., Yang, X., Wang, X.: A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature pp. 270–273 (2020)