Extending LDA functionality using cosine similarity in tracking the COVID-19 Publications

By

Jessica Osekowsky

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Masters of Science in computer science

Middle Tennessee State University

July, 2023

Thesis Committee:

Dr Cen Li, Chair

Dr. Suk Seo

Dr Medha Sarkar

ACKNOWLEDGEMENTS

I would like to thank Dr. Cen Li for helping me every step of the way on this thesis and being patient with me during this process. I would also like to thank Dr. Seo and Dr. Sarkar for being on my thesis advisory committee. I would also like to thank Mallory and Rachel from the MTSU Writing Center. Their insight has been invaluable and they have been an absolute pleasure to work with.

ABSTRACT

Data is being created at an alarming rate, and it is becoming unrealistic to gather important information in a timely fashion without the use of machine learning techniques. The COVID-19 pandemic is one instance where the medical community came together and generated a large amount of data in a short period of time to gain a better understanding of the issues at hand. In this research, a new process called LDASine was developed that extended the Latent Dirichlet Allocation methodology for tracking topic changes over time. Two experiments were conducted to test the viability of LDASine. The first experiment involved determining which number of topics produced the most unique topics for three different time periods. The second experiment involved associating topics from different time periods and analyzing the changes between topics using the LDASine process. The results of the experiments proved the viability of LDASine as a process to analyze how topics change over time and determine which number of topics produced unique topics for a given measure of time.

List of Figures

LIST OF SYMBOLS AND TERMS

Machine Learning (ML) - a type of algorithm that is intended to emulate human intelligence.

Corpus - A collection of documents and is defined as a sequence of documents such that: $C = (D_1, D_2, ..., D_n)$, where D is a document and n is the total number of documents.

Topic Modeling - A machine learning technique used to identify patterns, or topics, within a corpus.

Latent Dirichlet Allocation (LDA) - A generative probabilistic model used for topic modeling technique that focuses on a static collection of documents.

Dynamic Topic Modeling (DTM) - A generative probabilistic model used for topic modeling technique that focuses on collection where time is a consideration to capture how topics evolve over time.

Time Slice - A period of time.

Dirichlet Distribution - A probability distribution used by LDA to determine the probability of a topic at various levels.

Gaussian Distribution - A probability distribution used by DTM to determine the probabilities of topics at various levels. This distribution is also chained to affect the probability distribution of over models in different time slices.

Word - the i-th index of a unit-basis vector used to describe the vocabulary for a corpus described as $V = (w_1, w_2, ..., w_n)$ where n is the total number of words in the corpus.

Document - A collection of words described as $D = (w_1, w_2, ..., w_n)$ where n is the total number of words in the document.

Posterior - The probability of a hypothesis, such as the likelihood of a word belonging to a topic.

Joint Distribution - The probability distribution for two or more random variables.

Topic - A distribution over words.

α - The topic distribution at the corpus level.

β - The topic distribution at the word level.

θ - The topic distribution at the document level.

CHAPTER 1.

INTRODUCTION

Information is being created at an unprecedented rate. At the rate information is being created, it is difficult for a single person to keep up to date without the assistance of machine learning tools. Machine learning is a type of algorithm that is intended to emulate human intelligence[12]. Machine learning has a wide range of applications and when it is used to gain a better understanding and analysis of text, it is often referred to as text mining. There are several different areas of text mining methods: information retrieval, summarization, and topic modeling. For example, Text mining techniques have been used to help track down changes in Twitter topics over time, and gain a deeper understanding of medical, and agriculture documents[1, 11, 20].

Topic modeling is a text mining technique. It involves gathering information from large collections of documents, or a corpus, to identify unique topics within the documents. Topic modeling also has application outside of text data such as finding patterns within DNA sequences[17]. In addition to discovering topics, topic modeling can also be used to retrieve information from documents and classify documents according to topics[2].

Two popular topic modeling techniques are the Latent Dirichlet Allocation (LDA) and Dynamic Topic Modeling (DTM). The LDA is a generative probabilistic algorithm that can find latent topics within a collection of documents. However, the algorithm does not highlight how topics change over time[3]. Whereas, DTM is a generative probabilistic algorithm that does highlight how topics change over time[3]. Both techniques provide valuable insight into large collections of text data, but both have the strict requirement of needing to know the number of topics within a collection beforehand[6]. This makes it difficult to explore all topics within a collection of documents over a period of time and has the potential to leave out topics. There have been attempts at analyzing topics

over time for a collection of documents, though, the existing techniques do not give a full picture of all topics within each time slice analyzed[18, 19]. One paper focused on using LDA and cosine similarity for data visualization to find short-lived topics, but only focused on exploring a fixed number of topics without a framework for finding the appropriate number of topics and has the potential to miss less prominent but important topics[10].

While DTM and other topic modeling techniques using LDA provide valuable insight within the field of topic modeling for documents published over time, both focus on topics at the corpus level rather than the topics relevant per each time slice. This reduces the visibility of all topics and focuses on topics that have a bigger presence within the whole corpus. The focus on topics at the corpus level makes it difficult when exploring a collection of documents that are over a period to identify all topics, especially topics that only appeared in a shorter time period. To solve these issues, we propose a new process called LDASine:

- Explore the use of cosine similarity and a word dictionary to determine the number of topics to produce unique topics
- Explore the use of cosine similarity and a word dictionary when associating topics from different LDA models to track how the topics change over time

To demonstrate the new process, we applied LDASine to track explores a dataset collected by the Allen Institute for AI in response to the COVID-19 pandemic, which includes scientific research on documents about viruses in the coronaviruses family in response to the fast-evolving SARS-CoV-2, or COVID-19, pandemic[23]. The collection includes over 280,000 scholarly documents. The dataset is also referred to as CORD-19. The goal of the dataset was to help reduce information overload by using machine learning techniques to better assist medical professionals[22].

The next chapter discusses LDA, DTM, and cosine similarity to help gain a deeper understanding of the topic modeling algorithm. Chapter Three presents the methodologies used for the experiments conducted in this research. Chapter Four reviews the results of the experiments. Last, Chapter Five summarize the findings of this research.

CHAPTER 2.

BACKGROUND

Topic modeling is a valuable machine learning tool used to gain a better understanding of a large collection of documents. What exactly is topic modeling? Topic modeling is an unsupervised machine learning technique that is used to gain an overall idea of the themes within a collection of documents[15]. Topic modeling is described as a probability distribution over a vocabulary, which is defined to be the vocabulary are all words found within a collection of documents[15]. In the context of topic modeling, topics are a collection of words with probabilities assigned to the words. Since topic modeling is an unsupervised machine learning technique, the topics will not have a natural label applied to them, especially in the case of Latent Dirichlet Allocation(LDA) and Dynamic Topic Modeling(DTM)[9]. Table 1 shows an example of a topic. Someone could look at Table 1 and guess the theme for the topic could be "house pets", but since it was generated via an unsupervised machine learning process, the person viewing the topics will only see the pattern the technique found.

Table 1: Example of a topic model topic

Word	Probability
Cat	0.213
Dog	0.20
Hamster	0.05
Ferret	0.001

There are several topic modeling techniques. Two popular techniques are Latent Dirichlet Allocation and Dynamic Topic Modeling. We will explore both topic modeling techniques in this chapter. Cosine similarity will also be explored in this chapter. Cosine similarity is a means of calculating similarity and has a wide range of applications.

LATENT DIRICHLET ALLOCATION

The Latent Dirichlet Allocation (LDA) is a generative probabilistic model used for topic modeling, document classification, and collaborative filter[6]. The idea behind LDA originated from the shortcomings of other information retrieval methods in classification, detection, summarization, similarity, and relevance judgments[6]. LDA relies on the idea that documents are composed of several topics to varying degrees[4]. Topics are a distribution over a fixed vocabulary. Topics also do not have a natural label given to them because LDA is an unsupervised machine learning technique[3]. When using LDA, the number of topics must be known before any processing can occur[6]. Table 2 shows an example output of a topic using LDA for topic modeling. The topic consists of words and probabilities of the words belonging to a topic[3].

Table 2: Example topic output produced by an LDA model

Word	Probability
protein	0.04
membrane	0.035
expression	0.02
pathway	0.001
compound	0.0001

It is crucial to understand several terms regarding LDA: word, topic, and document. A word is defined as a unit-basis vector where the i-th index has a value of one to denote the word in the vocabulary[6]. Similarly, a document, D, is defined as a sequence of words such that $(w_1, w_2, ..., w_n)$ where n is the total number of words in the document and w_i is the i-th word in the vocabulary of the document[6]. Likewise, a corpus, or a collection of documents,C, is defined as a sequence of documents such that: $(D_1, D_2, ..., D_n)$, where D_i is the i-th document in a collection and n is the total number of documents[6].

LDA relies heavily on probability distributions when determining the distribution of topics of a word or document, more specifically the Dirichlet distribution[6]. The Dirichlet

distribution is a continuous probability distribution over a simplex [7]. The simplex is defined as a set of positive numbers that sum up to one [7]. LDA utilizes the Dirichlet distribution because the k-dimensional simplex can be used to represent the probability distribution of topics[6]. The Dirichlet distribution has a probability distribution that can be described as: f(x) = $\frac{\prod_{k=1}^{K}\gamma(a_k)}{\gamma(\prod_{k=1}^{K}a_k)}\prod_{k=1}^{K}x_k^{a_k-1}$ Where the parameter a is greater than 0 for $a_1,...,a_k$ [7]. Figure 1 shows the graphical representation of the probability distribution for the Dirichlet distribution using a three dimensional simplex. Figure 1 shows a graphical representation of the Dirichlet distribution with a three dimensional simplex.

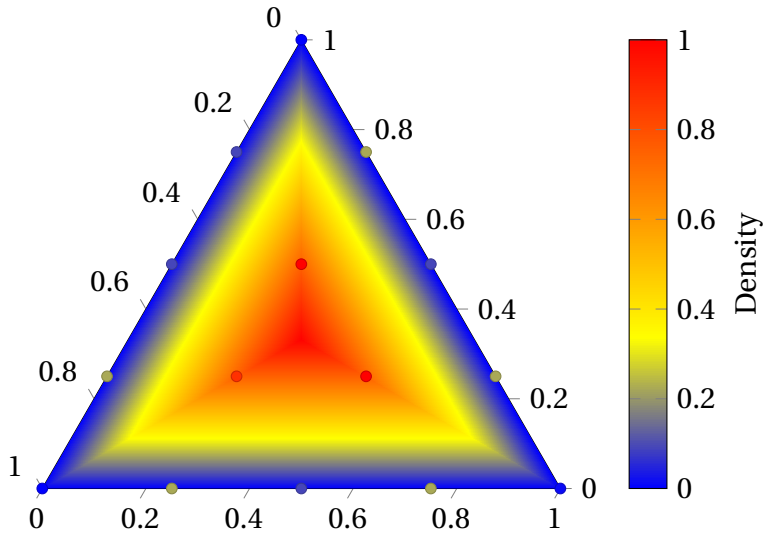

Figure 1: The visual representation of the Dirichlet Distribution using a three dimensional simplex

LDA uses several parameters that use the Dirichlet distribution. Those parameters are α and β. The α parameter is used to determine the topic distribution at the corpus level[6]. The β parameter determines the distribution of topics at the word level[6]. The application of these parameters will be explained further when going over the algorithm LDA uses to derive topics. The variables α, β, and the number of topics selected during model initialization can drastically change the quality of the topics derived from the corpus. If too few or too many topics are selected, the topics may not make sense or

Algorithm 1: High level overview of LDA algorithm

```
High level overview of LDA algorithm (Corpus, α, β);
Input  : a collection of documents Corpus, the document level topic distribution
         used for initialization, k-dimensional Dirichlet distribution for the word
         level topic distribution
for document in Corpus do
    for word in document do
        word.topic = AssignRandomTopic()
while i < maxIterations do
    for document in Corpus do
        for word in document do
            word = ReassignTopic(word,α,β)
```

important topics may be missed.

LDA is a generative probabilistic process that includes a few steps for deriving topics from a collection of documents[3]. The first step in the process is assigning a random topic to every word in every document in the collection[14]. Below is the pseudo-code for LDA:

After initializing each word with a random topic is completed, the frequency of each word that appear in each topic and the frequency of the topics within each document is collected[14]. These values are used during the topic reassignment process. The primary goal of the topic reassignment process is to reassign the topics of every word[14]. At a high level, the topic reassignment process includes calculating the probability of a word belonging to each topic[14]. The topic with the highest conditional probability is the new topic assignment for the given word[14].

The conditional probability of a topic is calculated as the following:

$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$

where β is the topic distribution at the word level, α is the topic distribution at the corpus level K is the number of topics selected, θ represents the topic distribution

found during the initialization process using the frequency of topics assigned to a given document, z represents the topic, w represents the word[6]. The topic reassignment process involves iterating through each word in each document to reassign topics using the conditional probability described above[14]. Topic reassignment is repeated until the number of iterations is met.

LDA can derive topics from a collection of documents, but it cannot show how the topics change over time[5]. DTM is the next step in topic modeling that solves the issue of tracking how topics change over time. The next section will cover DTM.

DYNAMIC TOPIC MODELING

Dynamic topic modeling (DTM) was proposed to extend the functionality of LDA by showing how a fixed number of topics change over time[5]. Like LDA, DTM is a generative process[5]. There are three distinct differences between DTM and LDA. The first is the use of Gaussian distributions instead of Dirichlet distribution for α and β [5]. The second is the assumption that the order in which words appear in document does matter for topic assignment[5]. Last, DTM uses a model per time slice instead of one single model for the whole corpus[5].

DTM makes use of hyper parameters α and β similar to LDA. However, instead of using Dirichlet distributions, the Gaussian distribution is used. The Gaussian distribution is a continuous distribution and is also called the normal distribution. The Gaussian distribution is a symmetrical distribution and is often used as an approximation[8]. The probability distribution for the Gaussian distribution can be described as: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\mu^2})$ where μ is the mean of the distribution and σ is the standard deviation[8]. Figure 2 shows an example of the probability distribution.

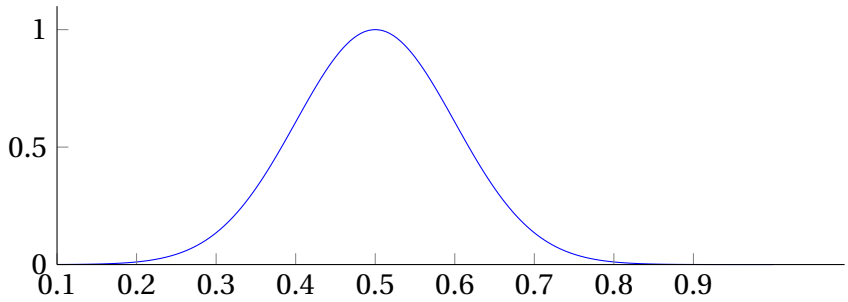

Figure 2: Visual of a Gaussian distribution with a single variable

Figure 2 shows a symmetrical bell curve where the majority of the values are near the center of the curve. The graph in the figure shows a single variable distribution. The Gaussian distribution can be used with multiple variables similar to the Dirichlet distribution. To allow for multiple variables, μ is treated as a k-dimensional vector.

The overall algorithm used for DTM is very similar to the algorithm for LDA. The biggest difference is the use of the Gaussian distribution, the need to order documents by a date, and the Gaussian distribution for α and β needing to be updated after each time slice has been processed[5]. The first step in the DTM algorithm is splitting documents up into time slices by a date[5]. Time slices can be by day, week, or month. Each time slice goes through an initial topic assignment using the α parameter for each word in every document per time slice[5]. Algorithm 2 shows the pseudo-code for DTM.

Next, the second step is the topic reassignment. Each word is reassigned a topic until the number of iterations has been met. Once the first time slice has completed processing, α and β are updated to include the known distribution of topics at the corpus and word level from the recently finished time slice. The α and β are then used in the processing of the subsequent time slice.

One of the drawbacks of DTM is the need to select the number of topics before model initialization. The number selected is used for all topics found within the corpus. Topics that have a small presence in the overall corpus but have a meaningful presence in a

Algorithm 2: Dynamic Topic modeling topic assignment initialization

```
Overview of DTM algorithm (Corpus, α, β);
Input  : a collection of documents Corpus, the corpus level Gaussian Distribution
         variable, the word level Gaussian Distribution variable
for timeSlice in TimeSlices do
    for document in timeSlice of CollectionOfDocuments do
        for word in document do
            word.topic = AssignRandomTopic()
while i < maxIterations do
    for timeSlice in TimeSlices do
        for document in timeSlice in CollectionOfDocuments do
            for word in document do
                word.topic = null word = ReassignTopic(word,αβ)
        UpdateAlphaSimplex()
        UpdateBetaSimplex()
```

particular time slice have the potential to be missed.

COSINE SIMILARITY

Cosine similarity is a technique used to compare similarity between vectors[13]. Cosine similarity has many applications within the machine learning space, including but not limited to text similarity, classification, information retrieval, and sentiment analysis[21]. Cosine similarity works by calculating the angle of two normalized vectors via the dot product between vectors [16]. Figure 3 shows the visualization of the angles between two vectors.

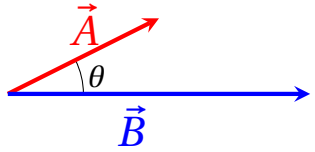

Figure 3: Visualization of the angle between two vectors.

Cosine similarity works by calculating the angle of two normalized vectors via the dot

product between vectors [16]. The formula for cosine similarity is shown below:

$$\cos\theta = \frac{A \cdot B}{||A|| \times ||B||}$$

A and B are two different vectors and do not have to have the same dimension. The values produced by the cosine similarity measure range from -1 to 1, where a value closer to zero means that the two vectors being compared are not similar, but the closer to a value of one, the more similar the two vectors are [21]. Though a value of negative one is possible, typically means that the two vectors are opposites of each other[21]. When using cosine similarity for text comparison, values fall between 0 and 1[21].

CHAPTER 3.

METHODOLOGY

Dynamic Topic Modeling (DTM) gives the user the ability to track the changes in topics over time[5]. However, the topics during model initialization are topics at the corpus level. It has the potential to exclude a topic that had a presence in a particular time slice but not throughout whole corpus. For example, if DTM is initialized with 3 topics and documents are split into five time slices and the second time slice contains a fourth topic, the fourth topic would not be taken into consideration for topic evolution. Figure 4 illustrates the above scenario.

Figure 4: Topics from five neighboring time slices, where the topic in bold is a topic only found in the particular time slice

Similarly, The existing methods using Latent Dirichlet Allocation (LDA) to track how topics change over time are also limited. Those methods focus on a single topic that happens within a short amount of time rather than all topics within a time slice[18, 19]. LDA can be expanded to better capture the relevant events at the time slice level rather than the corpus level.

To address the issues found in DTM and LDA when exploring the evolution of topics over time, we propose a new process. This process involves finding the appropriate number of topics, associating topics from different time slices, and finally analyzing the changes. This process is called LDASine and will allow for easier detection of new topics, determining when topics end, and tracking how topics evolve over time. The process of tracking how topics evolve over time is achieved by expanding upon the existing LDA process by adding two additional steps.

LDASine has five steps. The first step is a standard text pre-process of removing stop

words, non-alphabetical characters, words that appear too frequently, and words that appear too few times. The second step is utilizing a dictionary over the corpus vocabulary. Since LDA and DTM both produce topic vectors that show the most probable words in order, making it difficult to compare topics from different models. The main reason is that all words would not exist between all time slices. The use of this standard data science structure facilitates a fair comparison across models. The third step is determining the number of unique topics within a time slice. The Fourth step is the process of connecting topics from different time slices using cosine similarity to measure the similarity between models. Lastly, the topics between different time slices are analyzed to gain a deeper understanding of how the topics evolve over time. Figure 5 illustrates the steps of LDASine.

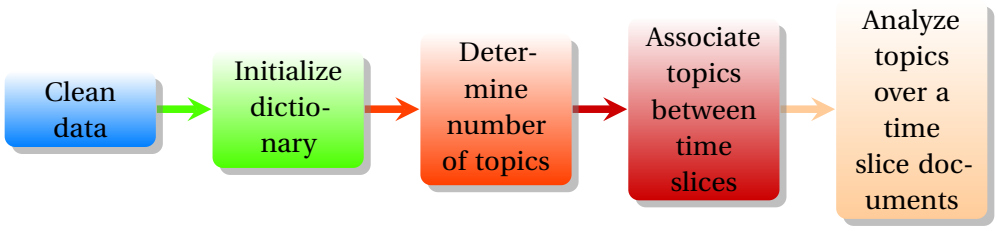

Figure 5: Overview of LDASine for tracking topics in documents over time

DATA CLEANING

The first step in LDASine is data cleaning. Documents are removed if they are not written in English. All non-alphabetical characters, numbers, and punctuation are removed. Words with fewer than three characters and stop words are removed. Next, different inflected forms of words are grouped. This process is called lemmatization. Last, words that appear too often and too few times are removed.

DETERMINE NUMBER OF TOPICS FOR A TIME SLICE

The third step is determining which number of topics produces the most unique topics per time slice. LDASine involves using a dictionary to create a vector populated

with the probabilities each word has of belonging to a given topic in the order the words are found in the dictionary. The cosine similarity value for all topic comparisons per LDA initialized using a specified number of topics are averaged together. The lower cosine similarity value signifies that the topics have little in common. The LDA model with the lowest average of cosine similarity values is the model with the most unique topics. Table 3 shows the average cosine similarity calculation for LDA models. In this example, the LDA model uses three topics, pairwise topics are compared using the cosine similarity to measure how similar they are. The numbers in bold are the values averaged. This process is completed with LDA models with different K values. The model with the lowest cosine similarity average is the model with the most unique topics.

	Topic 1	Topic 2	Topic 3
Topic 1	1.00	0.01	0.02
Topic 2	**0.2**	1.00	0.03
Topic 3	**0.05**	**0.03**	1.00

Table 3: Cosine similarity between topics using an LDA model with three topics. The average of this for this example is 0.09333333333

ASSOCIATING TOPICS BETWEEN TIME SLICES

For this step, we associate topics from neighboring time slices to prepare us for the topic analysis step. Topics from both time slices being compared are initialized into the comparison object. Next, each topic from both time slices are compared using cosine similarity. Table 4 shows an example of associating topics from neighboring time slices. The cosine similarity for all topics from both LDA models is calculated. The topics from the two time slices that have the highest cosine similarity are associated with each other. For example, in Table 4, Topic 1 from Time Slice 1 and Topic 1 from Time Slice 2 share a high cosine similarity value of 0.854273. Topics with the highest cosine similarity value are associated with each other. This would suggest that Topic 1 from time slice 2 is the

continuation of Topic 1 from time slice 1 given the high cosine similarity value.

Table 4: Example Cosine similarity values for topics between two neighboring time slices

	Time Slice 1 Topic 1	Time Slice 1 Topic 2	Time Slice 1 Topic 3
Time Slice 2 Topic 1	**0.854273**	0.023985	0.070024
Time Slice 2 Topic 2	0.181268	**0.938203**	0.01047
Time Slice 2 Topic 3	0.061483	0.01333	**0.77604**

Conversely, if a topic does not have a high cosine similarity value with any topic from a subsequent time slice, the topic could be assumed to have no presence in the subsequent time slice. Table 5 illustrates this example where Topic 2 from Time Slice 1 did not have a high cosine similarity value with any topic from Time Slice 2. This could also be interpreted as a new topic emerging in time slice 2.

Table 5: Example Cosine similarity values for topics between two neighboring time slices where a topic from Time Slice 1 did not have a presence in subsequent time slice

	Time Slice 1 Topic 1	Time Slice 1 Topic 2	Time Slice 1 Topic 3
Time Slice 2 Topic 1	**0.792234**	0.03947	0.01032
Time Slice 2 Topic 2	0.0123	**0.1**	0.0392
Time Slice 2 Topic 3	0.0743	0.0392	**0.8234**

ANALYZE TOPICS OVER A PERIOD OF TIME

For this step LDASine, all documents are separated into time slices based on the publication date. For this research, documents are broken up by months into slices, though smaller time slices would also work. For each time slice, an LDA model is generated with a given number of topics. The number of topics is based on the uniqueness of each topic.

At this point, the analysis of how the topics change between time slices can begin. This process is completed by using the dictionary to create a comparison vector and cosine similarity similar to the previous step except instead of a small cosine similarity, we're looking for a large cosine similarity. Since topics change from time slice to time slice,

it is not expected the topics to be identical but have a higher degree of similarity. Through this analysis, we get an overall picture of how topics evolved over time and phenomena such as short-lived topics that only appear for a brief amount of time.

CHAPTER 4.

EXPERIMENT AND RESULTS ANALYSIS

The experiments for this research are intended to help determine unique topics and the evolution of the topics using the LDASine approach developed. Two experiments have been performed to show a clearer association between topics in different time slices which help determine how topics change across a time slices. Both experiments have the same preprocessing and initialization steps. The first experiment is to determine whether LDASine can be used to determine unique topics within a time slice. The second experiment is to determine how topics evolve across time slices.

THE DATA

The CORD-19 dataset consists of scholarly literature over research on COVID-19, SARS-CoV-2, and other coronaviruses from January 1st, 2020, through January 1st, 2021. The dataset can be found at https://www.semanticscholar.org/cord19 and is a combination between full text and abstracts. For this research, the full text documents were used. Documents are separated by publication date into each month of the year in 2020. For example, the January 2020 time slice only contains documents published within that month. The CORD-19 dataset consisted of metadata file that included the title, abstract, date published, and file path to the full text documents. In total, from January 2020 through December 2020 there were a total of 56,902 documents.

The first step is data cleaning. Documents are removed if they are not written in English. Then, all non-alphabetical, such as punctuation and numbers, characters are removed. Words with fewer than three characters and stop words are also removed, and lemmatization is completed. The gensim library was used for the LDA model, the NLTK library was used for its stopword collection and also the lemmatization process, and the langdetect library was used for the removal of non-English documents.

INITIALIZATION OF THE DICTIONARY

The second step of LDASine is to create a common object for text mining called a dictionary. We use a standard python dictionary object and initialize the dictionary with all remaining words and initialize the value in the dictionary to zero to create a base instance. This dictionary is converted using the word probabilities into a vector format with the use of the Numpy library.

EXPERIMENT ONE - FIND UNIQUE NUMBER OF TOPICS

The primary goal of Experiment One is to determine the number of unique topics. A unique topic is a topic that shares the fewest amount of terms possible with other topics from the same Latent Dirichlet Allocation (LDA) model. The results of this experiment will set the foundation for the second experiment. To determine the number of unique topics, cosine similarity is used to compare topics generated from the same LDA model. Table 6 shows an example output of the topics from the same LDA model being compared against each other.

Table 6: Cosine similarity between topics using an LDA model with three topics. The average of this for this example is 0.21

	Topic 1	Topic 2	Topic 3
Topic 1	1.00	0.01	0.02
Topic 2	**0.6**	1.00	0.03
Topic 3	**0.02**	**0.01**	1.00

We will focus on values found in the lower triangle portion of the matrix in bold since the values repeat across the diagonal. The values in bold are averaged. This averaging process is repeated using LDA models initialized with other numbers of topics. The model with the lowest cosine similarity average is selected. Figure 6 illustrates the process for step one of LDASine.

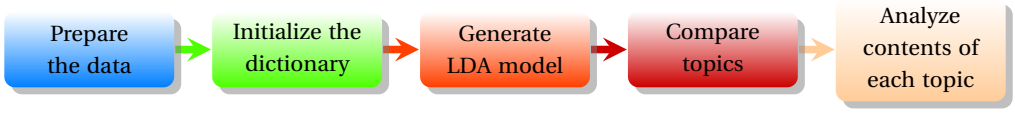

Figure 6: Overview of Experiment One

The first step includes data pre-processing. The second step is the creation of the dictionary that will be used to compare topics. The third step is the creation of the LDA model for a given time slice. The third step also includes initializing the LDA model with different number of topics and then those topics are compared using cosine similarity. Then, the topics in each model will be analyzed for number of words in common between topics and the cosine similarity value.

Three different months will be compared in this experiment: February, March, and December of 2020. Each time slice will be compared using three and five topics. Each topic initialized in the standard topic vector format will be compared using cosine similarity. The topics in the matrix with the lowest and highest value will be compared word by word. This is to determine whether there is an association between the cosine similarity value and the number of common words.

RESULT ANALYSIS OF Experiment One

The goal of Experiment One is to determine whether a word dictionary and cosine similarity can be used to determine topic uniqueness. The word dictionary included 3000 unique words. To achieve this, LDA models were initialized with three or five topics. Three and five topics were used during model initialization for this experiment because on average, three topics had the lowest cosine similarity average between topics. Conversely, five topics generally had a much higher cosine similarity value. Figure 7 illustrates the cosine similarity averages used to determine how many topics are to be used during model initialization.

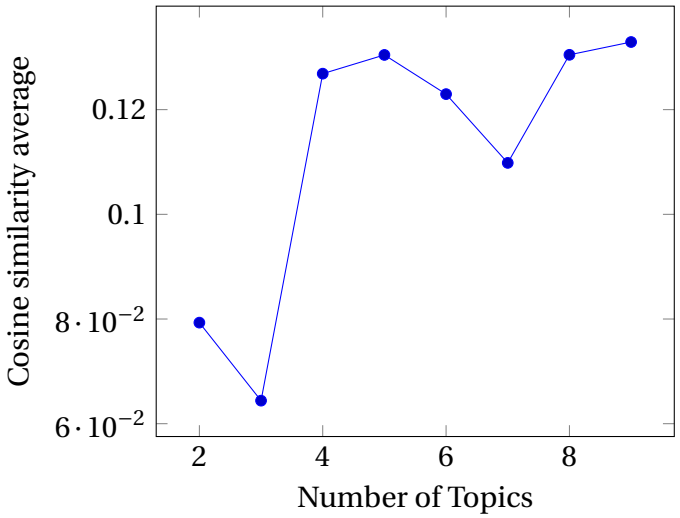

Figure 7: Cosine similarity averages between topics using 2 through 9 topics during LDA model initialization

Scientific publications in the months of February, March, and December 2020 have been used for this experiment. The first time slice analyzed is the month of February. Table 6 shows the cosine similarity values for each topic when initializing an LDA modeling with three topics.

Table 7: Cosine similarity value for topics using three topics for LDA model generation for the month of February 2020

	Topic 1	Topic 2	Topic 3
Topic 1	1.00	0.037129	0.082165
Topic 2	0.037129	1.00	0.062099
Topic 3	0.082165	0.062099	1.00

All cosine similarity values are relatively low which suggests each topic have very little in common. The highest cosine similarity value is found between topic 1 and topic 3 at 0.082165. The values found below the diagonal are used for the average and is 0.06046433333. For each topic, the top 20 words are used to compare how a high cosine similarity value correlates with words in common. The top 20 words are used for analysis because those words are most representative of the topics, though the topic probability

vector has a dimension of 3000. Since the topic probability vector has a larger dimension than the top 20 words, this could result in the cosine similarity value being higher or lower depending on the probabilities of each word. The top 20 words were selected to get an overall idea of the most probable words in each topic.

Table 8 shows the top 20 words for all three topics along with the probabilities of each word belonging to it's respective topic. The words are ordered by the probability in descending order. Topic 1 has words related to public response and safety. Topic 2 has words related to biological research. Topic 3 has words related to medical cases. The three topics displayed do not have any words in common and appear to be over three separate subjects. The cosine similarity values verified that the topics have little in common.

Table 8: Top 20 words from LDA model generated using three topics for February 2020

Topic 1	Topic 2	Topic 3
social - 0.008516902	protein - 0.021653706	child - 0.00853149
public - 0.0068633063	gene - 0.015302758	pneumonia - 0.0060951808
network - 0.005939236	vaccine - 0.014874592	lung - 0.005776479
participant - 0.0057265433	sequence - 0.012623769	mortality - 0.00554679
policy - 0.005421846	strain - 0.0122643	influenza - 0.0052694296
contact - 0.0047689946	antibody - 0.012016157	ncov - 0.0043596975
government - 0.004168912	expression - 0.010578392	blood - 0.004221844
disaster - 0.0037230144	mouse - 0.008469161	therapy - 0.0042156754
student - 0.0036938838	host - 0.00743101	trial - 0.0041353484
infectious - 0.0034308133	acid - 0.0070838756	acute - 0.0035671443
national - 0.0034218093	binding - 0.006640296	january - 0.0034936878
simulation - 0.003386693	immune - 0.0065106694	wuhan - 0.0033921178
training - 0.0031032849	animal - 0.0064637586	ventilation - 0.0033500898
intervention - 0.0030105037	site - 0.005560891	fever - 0.0032572441
medium - 0.0030085838	structure - 0.0049241795	antibiotic - 0.0032455595
pathogen - 0.0030045256	genome - 0.004745385	epidemic - 0.0031241248
professional - 0.0029003667	assay - 0.0047168513	infant - 0.003084938
healthcare - 0.0028921652	receptor - 0.0046790955	incidence - 0.0030460686
nurse - 0.0028695487	membrane - 0.004603735	injury - 0.0030395647
international - 0.0027785038	antiviral - 0.0043917336	adult - 0.0028320649

Table 9 shows the comparison between topic 1 and topic 2 as well as topic 1 and topic 3. When comparing topic 1 and topic 2 together, neither topic has any words in

common. This corresponds with the cosine similarity value comparing topic 1 and topic 2 being 0.037129, which would indicate very little similarity. The reason behind the cosine similarity value having a small value and not zero could be because every word has a probability of belonging to the topic and potentially words with less of a probability of belonging to either topic have a similarity probability. When comparing topic 1 and topic 3 together, neither topic have any words in common. This corresponds with the cosine similarity value comparing topic 1 and topic 3 being 0.082165, which would indicate very little similarity.

Table 9: Top 20 words for topic 1 compared to topic 2 and topic 1 compared to topic 3 for the month of February

Topic 1	Topic 2
social	protein
public	gene
network	vaccine
participant	sequence
policy	strain
contact	antibody
government	expression
disaster	mouse
student	host
infectious	acid
national	binding
simulation	immune
training	animal
intervention	site
medium	structure
pathogen	genome
professional	assay
healthcare	receptor
nurse	membrane
international	antiviral

Topic 1	Topic 3
social	child
public	pneumonia
network	lung
participant	mortality
policy	influenza
contact	ncov
government	blood
disaster	therapy
student	trial
infectious	acute
national	january
simulation	wuhan
training	ventilation
intervention	fever
medium	antibiotic
pathogen	epidemic
professional	infant
healthcare	incidence
nurse	injury
international	adult

Overall, when using three topics to initialize the LDA model, all topics had very low cosine similarity values. The low cosine similarity values between topics corresponds to the lack of common words between topics. If the cosine similarity value were higher, we

could potentially see some words in common.

The next comparison is using an LDA model initialized with 5 topics for the month of February 2020. Table 10 shows the cosine similarity values for topics for the LDA model initialized with five topics. The average values is 0.115216. This value is almost double the cosine similarity value when three topics were used for LDA model generation. This could indicate some words in common in the top 20 words from the topics. Topic 1 and topic 2 had a 0.444677 cosine similarity value. It is expected that those two topics have some commonality. Topic 3 and topic 4 also had a noticeably larger cosine similarity than other values in the same column at 0.243133.

Table 10: Cosine similarity value for topics using five topics for LDA model generation for the month of February 2020

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1.00	0.444677	0.050198	0.039289	0.013572
Topic 2	0.444677	1.00	0.083001	0.043548	0.040607
Topic 3	0.050198	0.083001	1.00	0.243133	0.069719
Topic 4	0.039289	0.043548	0.243133	1.00	0.124416
Topic 5	0.013572	0.040607	0.069719	0.124416	1.00

Table 11 shows the top 20 words for each of the five topics. Topic 1 and topic 2 have words related to biological research. Topic 3 and topic 4 have words related to medical case studies, however, topic 4 has some words related to public safety like topic 5. Topics 1 and 2 also have 6 words in common: "acid", "mouse", "expression", "host", "protein", and "gene". Topic 2 and topic 3 only share one word which was "influenza". Topics 3 and 4 share two words in common: "pneumonia" and "blood". Topics 4 and 5 had three words in common: "Wuhan", "SARS", and "NCoV". Topics 1 and 2 had a higher cosine similarity value at 0.444677 corresponding with having the most words in common. Topics 3 and 4 and topics 4 and 5 also share this pattern. This suggests that a higher cosine similarity value does corresponds to a higher number of words in common across LDA topics and a

lower cosine similarity value corresponds to topics with fewer or no words in common as seen in the LDA model initialized with three topics shown below.

Table 11: Top 20 most probable words for five topics from the February time slice for the month of February 2020

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
protein	protein	child	lung	public
membrane	vaccine	influenza	pneumonia	social
expression	sequence	mortality	injury	contact
pathway	gene	trial	ncov	epidemic
compound	strain	therapy	blood	network
structure	antibody	infant	chest	ncov
inhibitor	genome	antibiotic	ventilation	infectious
acid	animal	adult	acute	january
mouse	mouse	fluid	pulmonary	national
concentration	expression	pathogen	fever	participant
cancer	host	blood	lesion	policy
activation	assay	incidence	muscle	wuhan
antiviral	binding	antimicrobial	syndrome	government
host	immune	cohort	surgery	screening
lipid	site	resistance	sars	surveillance
gene	domain	sepsis	imaging	disaster
receptor	acid	admission	failure	estimate
surface	specie	pneumonia	wuhan	international
signaling	antigen	dose	onset	student
tissue	influenza	feeding	pressure	sars

Topic 1 had the highest cosine similarity value with topic 2. Both topics shared 6 words in common and had the highest cosine similarity value of 0.444677. Table 12 shows the words in common highlighted. Since the cosine similarity value is relatively higher, it is be expected there would be some similarity. The words that each topic has in common have varying positions, though there are a few words that share the same position, such as protein and mouse. Protein was also the most likely word to belong in both topics as well. Because of the words in common and some of the wording, both topics appear to have a similar theme that a person could identify as something related to biological research to some degree. This would not be ideal for comparing how topics evolve since

both topics are so similar, it would be difficult to tell which topic from another time slice was connected to which topic especially if the topic in the next time slice had a drastic change in theme. On the other end of the cosine similarity value comparison, topic 1 and topic 5 had the smallest cosine similarity value for topics compared to topic 1. Neither topic has words in common. Neither topic appears to have a common theme. This would be expected since the cosine similarity value for the comparison of the two topics was a low value at 0.013572. Compared to the topic 1 and topic 2 comparison, the higher the cosine similarity value the more likely a topic will have words in common and share a similar theme.

Table 12: The top 20 words for topic 1 compared to topic 2 and topic 1 compared to topic 5 for the month of February

Topic 1	Topic 2
protein	**protein**
membrane	vaccine
expression	sequence
pathway	**gene**
compound	strain
structure	antibody
inhibitor	genome
acid	animal
mouse	**mouse**
concentration	**expression**
cancer	**host**
activation	assay
antiviral	binding
host	immune
lipid	site
gene	domain
receptor	**acid**
surface	specie
signaling	antigen
tissue	influenza

Topic 1	Topic 5
protein	public
membrane	social
expression	contact
pathway	epidemic
compound	network
structure	ncov
inhibitor	infectious
acid	january
mouse	national
concentration	participant
cancer	policy
activation	wuhan
antiviral	government
host	screening
lipid	surveillance
gene	disaster
receptor	estimate
surface	international
signaling	student
tissue	sars

The second time slice analyzed is from March 1st, 2020 to April 1st, 2020. Table 13 shows the cosine similarity value topic comparing the topics from an LDA model

initialized with three topics.

Table 13: Cosine similarity value for topics using three topics for LDA model generation for the month of March 2020

	Topic 1	Topic 2	Topic 3
Topic 1	1	0.104539	0.045152
Topic 2	0.104539	1	0.009611
Topic 3	0.045152	0.009611	1

The average for the cosine similarity values is 0.0531. The highest cosine similarity value from this time slice is when topic 1 and topic 2 are compared at 0.104539. The lowest cosine similarity value from this time slice is 0.009611 for the topic 2 and topic 3 comparisons. Given the high and low cosine similarity values are small, it would be expected that the topics would have very little similarity to each other. Table 14 shows the top 20 most probable words in each topic.

Table 14: Top 20 words from LDA model generated using three topics for March 2020

Topic 1	Topic 2	Topic 3
child	site	energy
contact	expression	international
fever	acid	participant
therapy	protein	public
acute	host	city
sars	sequence	social
animal	domain	quality
Wuhan	inhibitor	staff
epidemic	gene	healthcare
influenza	strain	training
February	vaccine	digital
January	receptor	user
injury	binding	emergency
chest	animal	epidemic
mortality	antibody	paper
lung	compound	score
pneumonia	mouse	intervention
blood	structure	technology
syndrome	concentration	government
trial	immune	medium

Topic 1 has words that suggest a theme of medical case studies during a time slice. Topic 2 has themes of general biological research terms. Topic 3 has themes of policy. At a glance, all three topics have their unique themes. Topic 1 and topic 3 had the highest cosine similarity value between topics at 0.082165 and only had one word in common which was epidemic. The word "epidemic" had a higher probability of being in topic 1 than in topic 3 which could explain the slightly higher cosine similarity value. The lowest cosine similarity value was between topics 2 and topic 3 and neither topic shared any words in common, which corresponds to the cosine similarity value. Topic 1 and topic 2 had a cosine similarity value of 0.045152 and shared a word in common which was "animal", though the word had a much higher probability of belonging to topic 1 than topic 2. Table 15 shows the words in common between topic 1 and topic 2, as well as between topic 1 and topic 3.

Table 15: Top 20 words for topic 1 compared to topic 3 and topic 1 compared to topic 2 for the month of March

Topic 1	Topic 3
child	energy
contact	international
fever	participant
therapy	public
acute	city
sars	social
animal	quality
Wuhan	staff
epidemic	healthcare
influenza	training
February	digital
January	user
injury	emergency
chest	**epidemic**
mortality	antibody
lung	score
pneumonia	intervention
blood	technology
syndrome	government
trial	medium

Topic 1	Topic 2
child	site
contact	expression
fever	acid
therapy	protein
acute	host
sars	sequence
animal	domain
Wuhan	inhibitor
epidemic	gene
influenza	strain
February	vaccine
January	receptor
injury	binding
chest	**animal**
mortality	antibody
lung	compound
pneumonia	mouse
blood	structure
syndrome	concentration
trial	immune

Similar to the topic comparison for the February 2020 time slice, the cosine similarity values were small, and the topics had very little in common. The topics themselves had distinct themes when analyzing the topics word by word. The March 2020 time slice did have an increased highest cosine similarity value compared to February 2020 highest value, and it was reflected in that the topics compared had a word in common, though each LDA topic had individual themes.

Table 16 shows the cosine similarity values between topics using LDA model with five topics for model initialization. The average cosine similarity value is 0.1050. This average is higher than when three topics were used for model initialization and this is reflected in that there are several topics with a higher cosine similarity value.

Table 16: Cosine similarity value for topics using Five topics for LDA model generation for the month of March 2020

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	0.089573	0	0.148861	0.143149
Topic 2	0.089573	1	0.02564	0.410936	0.018908
Topic 3	0	0.02564	1	0.011463	0.158323
Topic 4	0.148861	0.410936	0.011463	1	0.043227
Topic 5	0.143149	0.018908	0.158323	0.043227	1

The highest cosine similarity value between topics 2 and 4 when five topics were used for model initialization was 0.410936. Topic 1 had high cosine similarity values with topic 4 and topic 5 at 0.148861 and 0.143149 respectively. Topic 1 also had a cosine similarity value of zero with topic 3. Topic 2 had the lowest cosine similarity value with topic 5 at 0.018908. Topic 3 had the highest cosine similarity value at 0.158323 with topic 5 and the lowest value with topic 4 at 0.011463. Topic 4 had the lowest cosine similarity value at 0.011463 with topic 3. Topic 5 had the lowest cosine similarity value with topic 2 at 0.018908. Table 17 shows the top 20 words for each topic.

Table 17: Top 20 words from LDA model generated using five topics for March 2020

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
failure	site	energy	expression	estimated
child	acid	question	acid	public
lesion	protein	participant	protein	prevention
fever	host	public	inhibitor	city
therapy	spike	article	gene	contact
ventilation	sequence	social	food	staff
laboratory	domain	network	vaccine	healthcare
acute	inhibitor	quality	receptor	province
admission	strain	framework	activation	travel
pressure	gene	literature	water	march
pulmonary	vaccine	digital	animal	emergency
ards	receptor	online	production	Wuhan
injury	binding	learning	antibody	epidemic
chest	assay	user	pathway	infectious
mortality	antibody	search	tissue	influenza
lung	compound	student	mouse	estimate
pneumonia	entry	paper	cancer	February
blood	genome	tool	membrane	score
syndrome	residue	technology	concentration	intervention
trial	structure	medium	immune	January

The words in Topic 1 suggest a theme of medical case studies and general symptoms, whereas topic 2 words suggest a theme of biological medical research. Topic 3 has words that would seem to appear when you are reading a research paper, though this topic's theme seems less clear than others evaluated. Topic 4 has words that suggest biological research like topic 2. The theme for topic 5 is not very distinctive as the other topics. Topic 5 makes references to the months of January and February as well as references to places. This could suggest the theme is about the emerging pandemic in January and February. Table 18 shows the comparison between topic 2 and topic 4 and topic 2 and topic 5, the highest and lowest cosine similarity values for topic 1.

In Table 17, the words in bold are the words in common between the topics in each table. Topic 2 and topic 4 shared 8 words in common: "acid", "protein", "inhibitor", "gene", "vaccine", "receptor", and "antibody". The words in common correspond with

Table 18: Top 20 words for topic 2 compared to topic 4 and topic 5 for the month of March

Topic 2	Topic 4
site	expression
acid	**acid**
protein	**protein**
host	**inhibitor**
spike	**gene**
sequence	food
domain	**vaccine**
inhibitor	**receptor**
strain	activation
gene	water
vaccine	animal
receptor	production
binding	**antibody**
assay	pathway
antibody	tissue
compound	mouse
entry	cancer
genome	membrane
residue	concentration
structure	immune

Topic 2	Topic 5
site	estimated
acid	public
protein	prevention
host	city
spike	contact
sequence	staff
domain	healthcare
inhibitor	province
strain	travel
gene	march
vaccine	emergency
receptor	Wuhan
binding	epidemic
assay	infectious
antibody	influenza
compound	estimate
entry	February
genome	score
residue	intervention
structure	January

the cosine similarity value being on the higher end. While both topic 2 and topic 4 share a theme of biological research, both topics share 7 words in common and the words have similar placement for the top 20 words. "Acid" and "protein" both share the same placement at the top two and top three most likely words in both topics. "Antibodies" for topic 4 had a slightly higher placement in topic 2 in the top 20 most likely words. On the other end of the spectrum, topic 2 and topic 5 share no words in common, and this is reflected in the cosine similarity value is 0.018908. Topic 1 had a similar cosine similarity value to topic 4 and topic 5.

Overall, the average cosine similarity value average for five topics is higher than when three topics were used during initialization. This was reflected in that two topics in

the given topic model had a greater degree of similarity in word theme, actual words in common, and cosine similarity. When three topics were used for model initialization, the topics had a greater degree of individual themes and lower individual cosine similarity values as well as a lower average cosine similarity. These findings are like those found when comparing topics for the month of February 2020. This result confirms to the intuition when the same group of documents are divided into more smaller groups, there is a higher chance of topics sharing similar themes.

The third time slice analyzed is between December 1st, 2020 and January 1st, 2021. Like the previous two comparisons, models using three and five topics were used. Table 19 shows the cosine similarity values when comparing topics when three topics were used for model initialization.

Table 19: Cosine similarity value for topics using three topics for LDA model generation for December 2020

	Topic 1	Topic 2	Topic 3
Topic 1	1	0.064964	0.093004
Topic 2	0.064964	1	0.004694
Topic 3	0.093004	0.004694	1

The average cosine similarity value between pairwise topics is 0.05422. This suggests that the topics do not have many similarities. The highest cosine similarity value is 0.093004 between topic 1 and topic 3. The lowest cosine similarity value is 0.004694 between topic 2 and topic 3. Table 20 shows the top 20 words in each topic.

Table 20: Top 20 words from LDA model generated using three topics for December 2020

Topic 1	Topic 2	Topic 3
incidence	expression	child
severity	acid	question
cohort	protein	participant
therapy	host	public
laboratory	sequence	survey
acute	gene	social
admission	surface	lockdown
cancer	vaccine	healthcare
injury	receptor	online
mortality	binding	behavior
lung	assay	student
score	antibody	anxiety
pneumonia	compound	education
median	pathway	family
chronic	tissue	score
blood	mouse	intervention
diabetes	lung	mental
prevalence	structure	policy
syndrome	concentration	life
trial	immune	government

All three topics appear to have very little in common thematically. Topic 1 has words that suggest a theme of medical case studies. Topic 2 has themes of biological research. Topic 3 has themes of behavioral case studies. Table 21 shows the comparison between the topics with the highest and lowest cosine similarity value.

Table 21: Top 20 words for topic 1 compared to topic 3 and topic 2 compared to topic 3 for the month of December

Topic 1	Topic 3
incidence	child
severity	question
cohort	participant
therapy	public
laboratory	survey
acute	social
admission	lockdown
cancer	healthcare
injury	online
mortality	behavior
lung	student
score	anxiety
pneumonia	education
median	family
chronic	**score**
blood	intervention
diabetes	mental
prevalence	policy
syndrome	life
trial	government

Topic 2	Topic 3
expression	child
acid	question
protein	participant
host	public
sequence	survey
gene	social
surface	lockdown
vaccine	healthcare
receptor	online
binding	behavior
assay	student
antibody	anxiety
compound	education
pathway	family
tissue	score
mouse	intervention
lung	mental
structure	policy
concentration	life
immune	government

Topic 1 and topic 3 had the highest cosine similarity value at 0.093004 and shared one word in common which was "score". The word has a higher probability of belonging to topic 1 than topic 3, but all other topics do not share much of a theme. Topic 2 and topic 3 had the lowest cosine similarity value between topics at 0.004694 and the topics do not have words in common. Overall, the cosine similarity value between topics is low and is reflected in the themes between the topics, though topic 1 and topic 2 shared one word in common.

Table 22 shows the cosine similarity values when five topics are used for model generation. The average cosine similarity value is 0.09524. This cosine similarity value is lower than previous comparison using five topics. The highest cosine similarity value between topics is 0.323618 between topic 2 and topic 4. The lowest cosine similarity value

is 0.009775 between topics 2 and topic 3. The second highest cosine similarity value is 0.286553 between topic 3 and topic 5.

Table 22: Cosine similarity value for topics using Five topics for LDA model generation for December 2020

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1	0.106481	0.014616	0.034958	0.068618
Topic 2	0.106481	1	0.009775	0.323618	0.01033
Topic 3	0.014616	0.009775	1	0.067456	0.286553
Topic 4	0.034958	0.323618	0.067456	1	0.030092
Topic 5	0.068618	0.01033	0.286553	0.030092	1

Overall, the cosine similarity values are lower than previous models using five topics for initialization, so it would be expected there are fewer words in common between topics. Table 23 shows the top 20 words for each topic.

Table 23: Top 20 words from LDA model generated using five topics for December 2020

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
severity	expression	environment	site	depression
cohort	acid	crisis	energy	child
therapy	protein	access	image	participant
ventilation	host	public	protein	public
laboratory	gene	cost	sequence	stress
acute	activation	social	network	psychological
admission	receptor	food	vaccine	survey
pulmonary	vaccine	program	algorithm	contact
cancer	binding	online	binding	social
injury	assay	learning	feature	lockdown
mortality	animal	sector	performance	healthcare
lung	antibody	market	compound	physical
score	pathway	student	simulation	anxiety
pneumonia	tissue	education	parameter	item
median	mouse	national	application	score
chronic	lung	policy	residue	family
blood	membrane	technology	size	intervention
diabetes	cytokine	economic	detection	mental
syndrome	concentration	government	structure	respondent
trial	immune	communication	prediction	scale

Topic 1 has themes of medical case studies. Topic 2 has themes of biological research. Topic 3 has themes of policy. The overall theme in this topic is less evident. Topic 4 has slight themes of biological research, though not in the same sense as topic 2. Topic 5 has themes of behavioral case studies. Topic 2 and topic 4 had the highest cosine similarity value and this is reflected in that both topics have some degree of words having a biological research theme to both. Topic 3 and topic 5 have some degree of similarity in that both topics have a theme of some sort of case study. Table 24 shows the top 20 words for both topic 2 and topic 4 as well as for topic 2 and topic 3 to compare the most similar topics to the least similar topics.

Topic 2 and topic 4 share a word theme related to biological research. They shared three words in common: "protein", "vaccine", and "binding". The ordering of the words in common are similar. The word "protein" was one of the top 3 words most likely to

Table 24: Top 20 words for topic 2 compared to topic 4 and topic 2 compared to topic 3 for the month of December

Topic 2	Topic 4
expression	site
acid	energy
protein	image
host	**protein**
gene	sequence
activation	network
receptor	**vaccine**
vaccine	algorithm
binding	**binding**
assay	feature
animal	performance
antibody	compound
pathway	simulation
tissue	parameter
mouse	application
lung	residue
membrane	size
cytokine	detection
concentration	structure
immune	prediction

Topic 2	Topic 3
expression	environment
acid	crisis
protein	access
host	public
gene	cost
activation	social
receptor	food
vaccine	program
binding	online
assay	learning
animal	sector
antibody	market
pathway	student
tissue	education
mouse	national
lung	policy
membrane	technology
cytokine	economic
concentration	government
immune	communication

belong to topic 2 while "protein" was a top 4 word in topic 4. "Vaccine" was a top 7 word most likely to belong to topic 4 and a top 8 word in topic 2. The word "binding" had the same placement for both topics as well. Though the cosine similarity value is lower than previous high cosine similarity value comparisons, there is a commonality between of the both topics; however to a lesser degree and corresponds with the cosine similarity value between both topics. Topic 2 and topic 3 had the lowest cosine similarity value between topics and the topic had no words in common nor did the topics have a similar word theme.

Table 25 shows the top 20 words from the comparison between the next highest cosine similarity value, which is between topics 3 and 5 and the lowest between topic 1 and

Table 25: Top 20 words for topic 3 compared to topic 5 and topic 1 compared to topic 3 for the month of December

Topic 3	Topic 5
environment	depression
crisis	child
access	participant
public	**public**
cost	stress
social	psychological
food	survey
program	contact
online	**social**
learning	lockdown
sector	healthcare
market	physical
student	anxiety
education	item
national	score
policy	family
technology	intervention
economic	mental
government	respondent
communication	scale

Topic 1	Topic 3
severity	environment
cohort	crisis
therapy	access
ventilation	public
laboratory	cost
acute	social
admission	food
pulmonary	program
cancer	online
injury	learning
mortality	sector
lung	market
score	student
pneumonia	education
median	national
chronic	policy
blood	technology
diabetes	economic
syndrome	government
trial	communication

topic 3. Topic 3 and topic 5 had a cosine similarity value of 0.286553. The topics have two words in common: "public" and "social". The word "public" was in the topic 4 most probable words in both topics. The word "social" was more probable in belonging to topic 3 than to topic 4. Regarding theme, both topics seem to have a theme of studying a large group of people. Given the cosine similarity value of 0.286553, the topics only have a small degree of similarity reflected in the top 20 words in the topics. Topic 1 and topic 3, on the other hand, have no words in common and topic 1 has a more defined theme than topic 3 which has words suggesting medical case study.

Unlike previous comparisons with models using five topics for model initialization, there are not two topics with a high degree of similarity, and this is reflected in the

individual cosine similarity between topics. The highest cosine similarity average is attributed to two groups of topics having a higher cosine similarity than other cosine similarity values. This suggests the two groups of topics with an elevated cosine similarity value could be subtopics of a more general topic that could be found using four topics for modal initialization. Overall, the cosine similarity values between topics does correspond with similarity between the top number of words between topics.

EXPERIMENT TWO - TRACKING TOPIC CHANGES OVER TIME

The primary goal of Experiment Two is to determine whether LDASine can be used to determine the unique topics within each time slice and how the topics evolve from time slice to subsequent time slice. The first step in Experiment Two is analyzing the contents of each time slice using different numbers of topics for LDA model generation to determine which number of topics best captures the topic evolution between time slices. In this experiment, three topics were selected for model initialization. This is because of the result found in Experiment One that when three topics were used during model initialization, topics were generally more unique. Figure 8 shows the overview of the steps for Experiment Two.

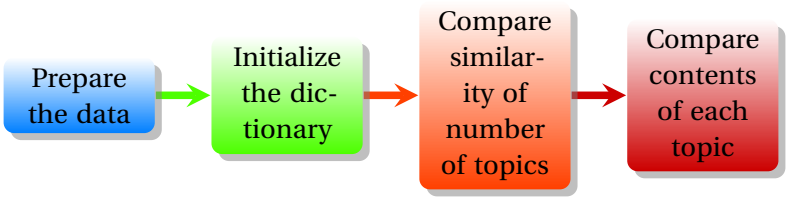

Figure 8: Overview the steps conducted for Experiment Two

The third step in Experiment One consists of gathering the cosine similarity values from the topics from the time slices being analyzed. The topics from both time slices are looped to compare the cosine similarity and the value is saved in the matrix. Once all the cosine similarity values are collected, the cosine similarity matrix is evaluated. For each row, the column with the highest cosine similarity value are the determining factor on

which topic is most similar from the other time slice. The values with the highest and lowest cosine similarity value are evaluated for similarity. Table 26 shows an example output from Experiment Two.

Table 26: Example output from Experiment Two

	Topic 1, time slice 1	Topic 2, time slice 1	Topic 3, time slice 1
Topic 1, time slice 2	**0.9**	0.01	0.02
Topic 2, time slice 2	0.01	**0.8**	0.03
Topic 3, time slice 2	0.02	0.03	**0.7**

Table 27 shows the example output of the topics from a single time slice. When the time slices are analyzed, both time slice topics are compared for the number of common words and if that correlates with the cosine similarity value.

Table 27: Example output of topics

Topic 1	Topic 2	Topic 3
Participant	Mouse	Pneumonia
Technology	Gene	Wuhan
User	Expression	Epidemic
Medium	Sequence	Fever
Energy	Antibody	Influenza
Digital	Acid	February
Government	Binding	Trial
Paper	Receptor	Chest
City	Host	January
Training	Vaccine	Child
Network	Compound	Admission
Student	Strain	Sars-CoV2
Communication	Animal	Pressure
Policy	Immune	Ventilation
Epidemic	Inhibitor	Median
Online	Site	Onset
Learning	Assay	illness
Search	Pathway	province
Framework	Concentration	Mild
Article	Genome	Cough

RESULTS ANALYSIS OF Experiment Two

In Experiment Two, the aim is to determine whether cosine similarity can help identifying similar topics from different time slices. First, the highest cosine similarity value for each topic from two time slices are identified. Then the top 20 words in each topic are analyzed. The first two time slices analyzed are the March and April 2020 time slices. Table 28 shows the cosine similarity values between these two time slices.

Table 28: Cosine similarity value for topics in March compared against topics in April using three topics for both time slices

	March Topic 1	March Topic 2	March Topic 3
April Topic 1	0.854273	0.023985	0.070024
April Topic 2	0.181268	0.938203	0.01047
April Topic 3	0.061483	0.01333	0.77604

March topic 1 and April topic 1 had a high cosine similarity value at 0.889769. March topic 2 had a high cosine similarity together with April topic 2 at 0.938203. March topic 3 had a high cosine similarity with April topic 3. April topic 2 and March topic 2 had the highest cosine similarity value out of all the cosine similarity values. Table 29 shows the comparisons between March topic 1 and April topic 1 sharing the highest cosine similarity value, and April topic 3 with the lowest cosine similarity value.

March and April topic 1 have words that suggest a theme of medical case studies. March topic 1 has more mentions of time slices and April topic 1 has more words related to medical case studies. Overall, the topics are very similar. March topic 1 and April topic 1 had 14 words in common. The ordering of the words is similar as well for a few words, especially for: "chest", "mortality", "lunch", "pneumonia", "blood", "syndrome", and "trial". In both topics those words are lower down on the probability but still in the top 20. When comparing March topic 1 and April topic 3, briefly, both topics shared very little thematically. The topics had two words in common but April topic 3 had more words that

Table 29: Top 20 words for March topic 1 compared to April topic 1 and March topic 1 compared to April topic 3

March Topic 1	April Topic 1
child	swab
contact	**child**
fever	**contact**
therapy	**fever**
acute	**therapy**
sars	ventilation
animal	laboratory
wuhan	**acute**
epidemic	admission
influenza	**wuhan**
february	**influenza**
january	mild
injury	cancer
chest	**chest**
mortality	**mortality**
lung	**lung**
pneumonia	**pneumonia**
blood	**blood**
syndrome	**syndrome**
trial	**trial**

March Topic 1	April Topic 3
child	crisis
contact	participant
fever	public
therapy	team
acute	**contact**
sars	social
animal	school
wuhan	staff
epidemic	healthcare
influenza	mask
february	worker
january	procedure
injury	member
chest	emergency
mortality	**epidemic**
lung	student
pneumonia	family
blood	intervention
syndrome	policy
trial	government

would suggest the topic was related to medical staff and procedures. Table 30 shows the comparisons between the highest and lowest cosine similarity value for March topic 2.

March topic 2 and April topic 2 had the highest cosine similarity value at 0.938203 and shared 17 words in common. The top 6 words of the two topics are in the same order. Other words in both topics have similar ordering as well. Both topics have words that suggest biological research themes. The similarity in both topics corresponds to the cosine similarity value. March topic 2 compared to April topic 3 had the lowest cosine similarity value at 0.01333 and shared no words in common. March topic 2 has strong themes in biological research while April topic 3 has themes of medical staff procedures. The lack of similarity in the words in the topics corresponds with the low cosine similarity

Table 30: Top 20 words for March topic 1 compared to April topic 1 and March topic 1 compared to April topic 3

March Topic 2	April Topic 2
site	**site**
expression	**expression**
acid	**acid**
protein	**protein**
host	**host**
sequence	**sequence**
domain	**gene**
inhibitor	**strain**
gene	**vaccine**
strain	**receptor**
vaccine	**binding**
receptor	**animal**
binding	**antibody**
animal	tissue
antibody	**mouse**
compound	genome
mouse	lung
structure	**structure**
concentration	**concentration**
immune	**immune**

March Topic 2	April Topic 3
site	crisis
expression	participant
acid	public
protein	team
host	contact
sequence	social
domain	school
inhibitor	staff
gene	healthcare
strain	mask
vaccine	worker
receptor	procedure
binding	member
animal	emergency
antibody	epidemic
compound	student
mouse	family
structure	intervention
concentration	policy
immune	government

value.

March topic 3 and April topic 3 had the highest cosine similarity value at 0.77604. Out of all the March and April topics, it was the lowest highest cosine similarity value. The two topics share a common theme of crisis procedure or protocol and share 9 words in common. A few words did have a similarity probability of belonging to both topics such as: staff and healthcare. Other words shared similar, but not the same placement in the top 20 words in both topics. March topic 3 and April topic 3 had the least similarity at 0.01047. The topics had no words in common, April topic 2 has strong biological research themes in words unlike March topic 3. Overall, the cosine similarity value between the topics corresponds with the lack of similarity found in the top 20 words. Overall, when

Table 31: Top 20 words for March topic 1 compared to April topic 1 and March topic 1 compared to April topic 3

March Topic 3	April Topic 3
energy	crisis
international	**participant**
participant	**public**
public	team
city	contact
social	**social**
quality	school
staff	**staff**
healthcare	**healthcare**
training	mask
digital	worker
user	procedure
emergency	member
epidemic	**emergency**
paper	**epidemic**
score	student
intervention	family
technology	**intervention**
government	policy
medium	**government**

March Topic 3	April Topic 2
energy	site
international	expression
participant	acid
public	protein
city	host
social	sequence
quality	gene
staff	strain
healthcare	vaccine
training	receptor
digital	binding
user	animal
emergency	antibody
epidemic	tissue
paper	mouse
score	genome
intervention	lung
technology	structure
government	concentration
medium	immune

comparing the March and April 2020 time slice, cosine similarity value can help identify the topics in common from different time slices and once those topics are identified, the difference can help a user to identify how the topic changed over time.

The next two time slices analyzed are the March and December 2020 time slices. This time slice comparison is to show how drastically topics can change over a wider range of time. Table 32 shows the cosine similarity values between the two time slices.

Table 32: Cosine similarity value for topics in March compared against topics in April using three topics for both time slices

	March Topic 1	March Topic 2	March Topic 3
December Topic 1	0.697161	0.087533	0.00961
December Topic 2	0.061507	0.933066	0.037827
December Topic 3	0.012642	0.09257	0.695277

March topic 1 and December topic 1 had a high cosine similarity value at 0.697161. This value is lower than the cosine similarity value found in the March and April topic 1 comparison. The decrease in cosine similarity can be explained by the larger gap in time between time slices. March topic 2 and December topic 2 had a high cosine similarity at 0.933066. March topic 3 and December topic 3 had a high cosine similarity value at 0.695277. Table 33 shows the top 20 most probable words for March topic 1 between December topic 1, the highest cosine similarity value, and December topic 3, the lowest cosine similarity value.

March and December topic 1 have words in common that have a theme of medical case studies. March topic 1 has more mentions of months, such as January and February, and references to locations. December topic 1 has more words that are related to general medical case studies that have no mention of location or time frame. This could suggest the topic changed from focusing at specific time periods or location to a more general medical case study theme. Both topics shared 9 words in common. Those words are: "therapy", "acute", "injury", "mortality", "lung", "pneumonia", "blood", "syndrome", and "trial". Compared to the March and April topic 1 comparison, there are fewer words in common. Also, the ordering of the words for topic 1 in March and December are different. The majority of the words in common for March are lower on the top 20 most probable words, suggesting those words became more representative of the topic in later months. This corresponds with the high cosine similarity between topics, but lower

Table 33: Top 20 words for March topic 1 compared to December topic 1 and March topic 1 compared to December topic 3

March Topic 1	December Topic 1
child	incidence
contact	severity
fever	cohort
therapy	**therapy**
acute	laboratory
sars	**acute**
animal	admission
Wuhan	cancer
epidemic	**injury**
influenza	**mortality**
February	**lung**
January	score
injury	**pneumonia**
chest	median
mortality	chronic
lung	**blood**
pneumonia	diabetes
blood	prevalence
syndrome	**syndrome**
trial	**trial**

March Topic 1	December Topic 3
child	**child**
contact	question
fever	participant
therapy	public
acute	survey
sars	social
animal	lockdown
Wuhan	healthcare
epidemic	online
influenza	behavior
February	student
January	anxiety
injury	education
chest	family
mortality	score
lung	intervention
pneumonia	mental
blood	policy
syndrome	life
trial	government

cosine similarity when compared to the March and April comparison. March topic 1 and December topic 3 do not have a strong common theme. March topic 1 and December topic 3 had one word in common between both topics: "child". This corresponds with the low cosine similarity value.

Table 34 shows the top 20 words for March topic 2 compared against December topic 2 and December topic 3. March and December topic 2 both have strong themes of biological research. Both topics shared 15 words in common, which corresponds with the high cosine similarity value. Both topics had the following words in common: "expression", "acid", "protein", "host", "sequence", "gene", "vaccine", "receptor", "binding", "antibody", "compound", "mouse", "structure", "concentration", and "immune". The

Table 34: Top 20 words for March topic 2 compared to December topic 2 and March topic 2 compared to December topic 3

March Topic 2	December Topic 2
site	**expression**
expression	**acid**
acid	**protein**
protein	**host**
host	**sequence**
sequence	**gene**
domain	surface
inhibitor	**vaccine**
gene	**receptor**
strain	**binding**
vaccine	assay
receptor	**antibody**
binding	**compound**
animal	pathway
antibody	tissue
compound	**mouse**
mouse	lung
structure	**structure**
concentration	**concentration**
immune	**immune**

March Topic 2	December Topic 1
site	incidence
expression	severity
acid	cohort
protein	therapy
host	laboratory
sequence	acute
domain	admission
inhibitor	cancer
gene	injury
strain	mortality
vaccine	lung
receptor	score
binding	pneumonia
animal	median
antibody	chronic
compound	blood
mouse	diabetes
structure	prevalence
concentration	syndrome
immune	trial

theme of biological research has been a strong and common theme throughout all the time slice comparisons. The biggest difference between both topics is December topic 2 has "vaccine" and "antibody" in a higher placement than in March topic 2, which could suggest the research around this time centered slightly more around vaccine research. March topic 2 and December topic 1 had the lowest cosine similarity value compared and shared no words in common or theme. This corresponds with the cosine similarity value.

Table 35 shows the top 20 words between March topic 3 compared against December topic 3 and December topic 1. March topic 3 and December topic 3 share a common theme of crisis procedure or protocol. Both topics share 9 words in common: "participant", "public", "social", "healthcare", "score", "intervention", and "government". March

Table 35: Top 20 words for March topic 3 compared to December topic 3 and March topic 3 compared to December topic 1

March Topic 3	December Topic 3
energy	child
international	question
participant	**participant**
public	**public**
city	survey
social	**social**
quality	lockdown
staff	**healthcare**
healthcare	online
training	behavior
digital	student
user	anxiety
emergency	education
epidemic	family
paper	**score**
score	**intervention**
intervention	mental
technology	policy
government	life
medium	**government**

March Topic 3	December Topic 1
energy	incidence
international	severity
participant	cohort
public	therapy
city	laboratory
social	acute
quality	admission
staff	cancer
healthcare	injury
training	mortality
digital	lung
user	**score**
emergency	pneumonia
epidemic	median
paper	chronic
score	blood
intervention	diabetes
technology	prevalence
government	syndrome
medium	trial

topic 3 has more of a focus on international response or crisis procedure while December topic 3 has words that suggest a theme of the affects of lockdown on mental health for different areas of life. March topic 3 and December topic 3 had a lower cosine similarity value when compared to March topic 3 and April topic 3 and this is reflected in the fewer words in common and lower cosine similarity value. The biggest difference between March topic 3 and December topic 3 is that December topic 3 has more words in the top 20 related to mental health. This suggests that the topic evolved from a general pandemic respond to the affects the resulting government policies had on individuals. March topic 3 and December topic 1 had the lowest cosine similarity value between both topics. Neither topic shared any word in common which corresponds with the low cosine

similarity value.

In Experiment Two, two time slices were analyzed using the methods in LDASine. The two time slices were the month of March and April in 2020. April topic 1 was the associated subsequent topic to March topic 1. April topic 2 was the associated subsequent topic to March topic 2. April topic 3 was the associated subsequent topic to March topic 3. Each associated topic was analyzed by using the top 20 words to determine if the cosine similarity value corresponded with the words in common in the top 20 words. The topics with high cosine similarity value corresponded with more similar themes. The placement of the words in subsequent topics were also analyzed to determine how the topic changed.

CHAPTER 5.

CONCLUSIONS AND FUTURE WORK

In this research, a new process called LDASine was developed that extended the LDA methodology for tracking topic changes over time. Two experiments were conducted to test the viability of LDASine. The goal of the first experiment was to determine whether the use of a word dictionary and cosine similarity can be used to determine which number of topics to use during LDA model initialization that would produce the unique topics. A unique topic was defined as a topic that had the fewest number of words in common and overall theme between topics derived from the LDA model. The first step in Experiment One included the initialization of the word dictionary. The word dictionary was then used to create a vector with the probabilities of every word found in the corpus. This allowed cosine similarity to be used to compare topics. The cosine similarity value was calculated between topics derived from same LDA model. Topics with high and low cosine similarity values were analyzed. The goal of the analysis was to determine whether high cosine similarity values correlated to having more words in common within the top 20 words or theme in each topic. The result of this experiment confirmed that the lower the cosine similarity value between two topics, the fewer words in common which results in less theme similarity. The results of the first experiment set up the foundation for the second experiment.

There were two goals for Experiment Two. The first goal was to determine whether cosine similarity and the use of the word dictionary could be used to associate topics between neighboring time slices. The second goal of Experiment Two included analyzing the changes in the topics from subsequent time slices. The steps to completing the first goal included the initialization of the word dictionary. The creation of the word dictionary was used as a standard format for comparing topics between time slices. Then, the cosine similarity value between topics across time slices were then calculated. Topics between

time slices that had high and low cosine similarity values were analyzed. The topics between time slices having a higher cosine similarity resulted in having a high degree of common theme and words. Topics with high cosine similarity values were analyzed for changes in the top 20 words.

The main contribution of this thesis is that LDASine is shown to be capable of determining the number of topics used for LDA model initialization and associate topics in neighboring time slices to help facilitate topic tracking. Both steps individually provide value. The first step gives the user the ability to help narrow down the number of topics to initialize LDA models, which can have a drastic affect on the readability of topics. The second step helps the user associate topics from subsequent time slices. It is flexible in selecting the size of LDA models in each time slice. This also helps users analyze topics by giving a measure of how different topics can be theme wise using cosine similarity.

There are a few paths LDASine could be explored in the future. The first potential area of research is using LDASine to explore subtopics by filtering out documents with a high degree of a specific topic and using the steps of LDASine to derive subtopics. This experiment could be completed on a static collection of documents or a collection of documents organized by date that would be split up into time slices. The subtopics could also be analyzed for how subtopics changed over time.

The second area of exploration for LDASine is experimenting with another dataset. The dataset used in this research had a very narrow topic, which was the COVID-19 pandemic. Other dataset that include a wider range of distinct topics would be interesting to explore. An example would be a collection of newspapers over the years. Similarly, LDASine could be explored further using a dataset that has a longer time frame than the CORD-19 dataset, which was a year. LDASine has no limit on how large or small a time slice needs to be, though in this research, time slices were split per month. This research could be expanded by exploring larger time slices, either with a dataset with a narrow

topic or one with more topics. Lastly, LDASine could be expanded upon by creating an objective way of labeling topics without the need of a person interpreting the topics word by word.

BIBLIOGRAPHY

[1] Adeline Abbe, Cyril Grouin, Pierre Zweigenbaum, and Bruno Falissard. Text mining applications in psychiatry: a systematic literature review. *International journal of methods in psychiatric research*, 25(2):86–100, 2016.

[2] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1), 2015.

[3] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[4] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, apr 2012.

[5] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

[7] Jordan Boyd-Graber. Probability distributions: Continuous. `http://users.umiacs.umd.edu/~jbg/teaching/INST_414/05c.pdf`, 2018.

[8] Jordan Boyd-Graber. Probability distributions: Continuous. `http://users.umiacs.umd.edu/~jbg/teaching/INST_414/05b.pdf`, 2018.

[9] Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017.

[10] Francine Chen, Patrick Chiu, and Seongtaek Lim. Topic modeling of document metadata for visualizing collaborations over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 108–117, 2016.

[11] Brett Drury and Mathieu Roche. A survey of the applications of text mining for agriculture. *Computers and electronics in agriculture*, 163:104864, 2019.

[12] Issam El Naqa and Martin J Murphy. *What is machine learning?* Springer, 2015.

[13] Wael H Gomaa, Aly A Fahmy, et al. A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18, 2013.

[14] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23, 2010.

[15] Jian Pei Jiawei Han, Micheline Kamber. *Data Mining*. Elsevier Science Ltd, 2012.

[16] Baoli Li and Liping Han. Distance weighted cosine similarity measure for text classification. In *Intelligent Data Engineering and Automated Learning–IDEAL 2013: 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings 14*, pages 611–618. Springer, 2013.

[17] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22, 2016.

[18] Andreas Niekler and Patrick Jähnichen. Matching results of latent dirichlet allocation for text. In *Proceedings of ICCM*, pages 317–322, 2012.

[19] Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. Rollinglda: An update algorithm of latent dirichlet allocation to construct consistent time series from textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2337–2347, 2021.

[20] Hao Sha, Mohammad Al Hasan, George Mohler, and P Jeffrey Brantingham. Dynamic topic modeling of the covid-19 twitter narrative among us governors and cabinet executives. *arXiv preprint arXiv:2004.11692*, 2020.

[21] P Sunilkumar and Athira P Shaji. A survey on semantic similarity. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–8. IEEE, 2019.

[22] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[23] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K. Funk, Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Brandon Stilson Stilson, Alex D Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.