Bayesian Topic Modeling of News Articles

New Ideas for Bayesian Sampling/Estimation

JHU ACM Probability and Stochastic Processes EN.625.721

Christopher Puglisi

August 22, 2023

Contents

1 Introduction **3**

2 Theoretical Analysis **3**
2.1 Bayesian Theory and Estimation 3
2.2 Probabilistic Topic Modeling 4
2.2.1 Latent Dirichlet Allocation and Posterior Computation 4
2.3 Approximating Evidence Problem 4
2.3.1 Sampling Based Algorithms 4
2.3.2 Variational Algorithms 5
2.4 Natural Extensions to Latent Dirichlet Allocation 5
2.4.1 Topic Models and Meta-Data 5
2.4.2 Relaxing de Finetti's Assumption: Bigram Topic Models 5
2.4.3 Unknown Number of Mixture Components: Bayesian Non-Parametric Models 6
2.4.4 Introducing Sequential Corpora: Dynamic Topic Modeling 6

3 Mathematical Framework and Pseudocode **7**
3.1 Framework and Pseudo-proof for Latent Dirichlet Allocation 7
3.2 Framework and Pseudocode for Dynamic Topic Models 8

4 Methods **11**
4.1 Data Source 11
4.2 Data Preprocessing 11
4.3 Model Implementation 11
4.4 Code Repository 11

5 Results **12**
5.1 Quantitative Metrics 12
5.2 Qualitative Analysis 12
5.3 Topic Over Time 13
5.4 Deeper Dive: LDA 13
5.5 Comparative Analysis 13

6 Discussion **14**

7 Appendix **15**

1 Introduction

News articles are written for mass consumption at an unprecedented rate - approximately 2-3 million articles are published online and in print each day [7]. Considering the sheer amount of textual information, there arises a need to parse, summarize, relay, and recommend articles in an organized manner. Probabilistic topic modeling is a powerful tool that has become vital to overcome the many challenges that come with this information overload. Topic modeling is a statistical approach to uncover abstract topics within a collection of documents. Considering news articles, one can leverage topic modeling to present organized information in a manner conducive for reader consumption, mine and structure content for publishers, and identify inconsistent narratives outside of the overarching latent topics modeled.

In this project, I research the application of probabilistic topic modeling. I establish a brief background and theoretical framework for model application and apply both a Latent Dirichlet Analysis (LDA) Topic Model and a Dynamic Topic Model (DTM). Differences in approaches and results are presented and discussed.

2 Theoretical Analysis

2.1 Bayesian Theory and Estimation

The foundations of Bayesian Estimation are built on the stance that a given data, parameters, and distributions are best modeled as a probabilities. While the frequentist approach considers parameters to be fixed but unknown quantities, the Bayesian approach infers new understandings of parameters of interest by updating prior beliefs based on new observations. Specifically, the primary mechanism for performing said inference is Bayes Rule.

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta) \times P(\theta)}{P(\text{data})}$$

Here, the probability of observing a parameter given our observations is the posterior, $P(\theta|\text{data})$. The probability of our parameter, $P(\theta)$, is the prior and is defined before employing our inference mechanism by leveraging the current state of knowledge on the parameter to yield a best estimate. The probability of our data given the parameter of interest, $P(\text{data}|\theta)$, is the likelihood, which is the critical component for incorporating new evidence into our analysis. The probability of the data, $P(\text{data})$ is known as the evidence and captures the marginal likelihood.

In many cases, the posterior is intractable thus necessitating sampling and estimation techniques that iteratively refine the probabilistic beliefs with new data. Popular approaches include Markov Chain Monte Carlo (MCMC) methods, like Gibbs Sampling [10]. A more deterministic approach is Variational Inference (VI) [2]. Both are discussed in more detail in later sections.

2.2 Probabilistic Topic Modeling

2.2.1 Latent Dirichlet Allocation and Posterior Computation

LDA is a generative probabilistic model for collections of discrete data. The probabilistic model assumes that documents contain latent, or hidden, topics. As such, a generative process was developed to model how documents are constructed. The topics are assigned as distributions over words in a document, and exist over the entire set of documents, or the corpus. The specific words are thought to be generated in the following process [1].

1. Randomly choose a distribution over topics.
2. For each word in the document
   (a) Randomly choose a topic from the distribution over topics in step #1.
   (b) Randomly choose a word from the corresponding distribution over the vocabulary.

In summary, every document contains a unique distribution of topics, and every word in the document is drawn from one of the distributions of topics. Then the topic is selected from the distribution over topics. Assuming this generative process allows one to infer the hidden topic structure within a corpus. The mathematical framework for LDA is described in Section 3.

2.3 Approximating Evidence Problem

In the original LDA publication, the intractability of the posterior is emphasized as a key issue. To overcome said issue, sampling and estimation algorithms such as Laplace Approximating, Variational Approximation and MCMC are recommended. Below, two key approaches are outlined for approximating inference for LDA.

2.3.1 Sampling Based Algorithms

Common for LDA, sampling based approaches are used to estimate an intractable posterior. The primary approach for sampling leverages MCMC methods, which are a class of algorithms that generate samples from a desired distribution. Samples can be generated by constructing a Markov chain which starts with an initial point and iteratively proposes new points based on the mechanism of the specific type of algorithm leveraged. This approach is statistically sound as the Markov Chain's equilibrium distribution is the desired distribution.

Most commonly used MCMC method is Gibbs Sampling for LDA. For each word in each document given the current assignments of all other words, topic assignments are sampled iteratively to create a Markov Chain that approximates the posterior of topic assignments [10]. Generally, Gibbs sampling follows the structure in Algorithm 2 outlined in detail in the mathematical framework section.

2.3.2 Variational Algorithms

Another common technique used for posterior estimation in LDA, variational algorithms are a suitable alternative to the popular sampling approaches. VI is recommended by the original authors of the LDA paper. Iterative fixed point methods are employed to minimize divergence between the variational distribution and the true distribution. At a high level, VI introduces a family of approximate distributions $q(z|\lambda)$ over the latent variables, parameterized by λ. The goal is to find the best λ such that $q(z|\lambda)$ is close to the true posterior $p(z|x)$ [6]. The essence of the technique is captured in Algorithm 1 [14].

Algorithm 1 Variational Inference for LDA Overview

Initialize variational parameters ϕ (word-topic assignments) and γ (document-topic distributions).
repeat
    for each document **do**
        for each word in document **do**
            Update word's topic assignment ϕ based on current topics and word distributions.
        end for
        Update document's topic distribution γ based on word assignments.
    end for
until change in ϕ and γ is below a threshold or max iterations reached
Return optimized topic assignments ϕ and topic distributions γ.

2.4 Natural Extensions to Latent Dirichlet Allocation

Although the base LDA model remains one of the most popular approaches in probabilistic topic modeling, the many assumptions of LDA implore authors to extend the primary algorithm to more sophisticated approaches.

2.4.1 Topic Models and Meta-Data

One modification of the initial model introduces meta-data into the the topic models. Meta-data includes any relevant information that could provide additional context to the discovery of latent structures beyond the core text data. This includes author, comments on articles, titles, and even hyperlinks referenced [12, 9].

For example, the relational topic model assumes each document is written through the base generative model assumed in LDA, but extends the model to account for the hyperlinks between documents can alter the distance between topic proportions [4]. The idea of enhancing the structure of the topics in a network is a pertinent theme in LDA extensions, which we plan to revisit in our application of new Bayesian estimations.

2.4.2 Relaxing de Finetti's Assumption: Bigram Topic Models

Another natural extension of LDA relaxes de Finetti's Assumption, that is, relaxing the assumption of exchangeablility or that word order does not matter. Within the field

of natural language processing, a variety of computationally intensive and sophisticated methods have been introduced to handle the semantic relationships of words and between word [5, 8], but clever probabilistic models have been introduced to handle word order.

For example, the first extension of LDA to relax de Finetti's Assumption was introduced in the publication "Topic Modeling: Beyond Bag of Words." Note, Bag of Words is the assumption that a document is considered a jumble of unordered and exchangeable words. That said, the author developed a bigram topic model that relaxes de Finetti's Assumption by creating model that generates topics for a word conditional on the prior word. This model and similar models expand the parameter space significantly, but enhanced model performance in a variety of applications [13].

2.4.3 Unknown Number of Mixture Components: Bayesian Non-Parametric Models

Another assumption in LDA is that the number of topics is fixed and should be prespecified prior to model training. However, this may not always be the case as topics can fluctuate under certain circumstances, especially over time. A prime example could be for news articles, in which topics ebb and flow with the news cycle. The assumption is combated by introducing Bayesian nonparametric topic modeling which assumes the documents arise form a mixture model. Teh et al. introduced Hierarchical Dirichlet processes in which the number of mixture components is determined by posterior inference. Instead of introducing a separate Dirichlet process for each group, HDP uses a hierarchical model where the measure for child Dirichlet processes is determined by another Dirichlet process [12].

2.4.4 Introducing Sequential Corpora: Dynamic Topic Modeling

The last notable extension of probabilistic topic modeling considers the significance of document order, relaxing the assumption of document exchangeablity, in order to create an enhanced posterior topical structure. One can naturally see how that for news articles, dynamic modeling could be important to account for the evolution of news stories over time, and how some events influence future events. Introduced by Blei et al., Dynamic Topic Modeling (DTM) drops the assumption that document order does not matter. Then, rather than creating a single distribution over words, a topic is now a sequence of distributions over words [3].

3 Mathematical Framework and Pseudocode

The two models chosen for application on the news data set are LDA for a baseline topic model and DTM for a comparison. As such, the focus on the next section will concern these models. The mathematical frameworks and pseudocode for both LDA and DTM are outlined below.

3.1 Framework and Pseudo-proof for Latent Dirichlet Allocation

Based on our described problem of news article topic modeling, a suitable baseline for topic modeling of news article descriptions would be an implementation of LDA. Let us return to the generative process assumed for documents. As described by Blei et al., LDA assumes for each document is a set of words $\mathbf{w}$ in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$

Each word is the unit of discrete data from a vocabulary $\{1, \ldots, V\}$. Each document is a sequence of N words: $\mathbf{w} = (w_1, w_2, \ldots, w_N)$. The corpus is a collection of M documents: $D = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_M}\}$ [1]. In step one, the Poisson assumption is considered somewhat arbitrary since any distribution that follows a reasonable document length distribution is viable. In step two, the Dirichlet assumption is vital. The distribution is exponential in nature, has finite sufficient statistics, and is a conjugate to the multinomial distribution. Therefore, it facilitates tractable computations and captures the underlying document structure. Step three captures the thematic structure of the document by associating each word with a specific topic. This process helps the model identify the underlying topics within the corpus.

With this fundamental understanding of the assumed generation, we can begin to understand the approach of LDA in more detail. The actual algorithm works *backwards* through the generative process by first initializing each word in every document to one of the of Z topics to yield both document-topic and topic-word counts.

One can see this come together as the joint distribution of topics, words, and topic mixture.

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

This joint distribution yields the probability of a specific combination of topics and words for a given document given the hyper-parameters α and β. Understanding the joint distribution is critical to unravel the generative process for topics and words. Next, by integrating over θ and summing over z, we can find the marginal distribution of a document.

$$p(w|\alpha,\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n,\beta) \right) d\theta$$

This can be thought of as the likelihood of observing a given set of words w in a document. With this marginal distribution, we are able to infer the probability of a given document's words without the topic distribution. Lastly, by taking the product of the marginal probabilities of single documents, we can find the probability of a corpus.

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_d n} p(z_d n|\theta_d) p(w_d n|z_d n,\beta) \right) d\theta_d$$

This equation can be thought of as the likelihood of the entire corpus, given the model's hyper-parameters. This equation is most important for optimization and hyper-parameter tuning. Another critical component of the theoretical underpinnings is tied to De Finetti's representation theorem. As stated before, in LDA it is assumed that topics are infinitely exchangeable within a document. By de Finetti's theorem, the probability of a sequence of words and topics have the following form.

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

We are able to leverage our understanding of the probabilities to illustrate LDA in a graphical format as well (see Figure 1 in appendix). This figure illustrates exactly how parameters influence each other and on the levels that the do. The Dirichlet parameters, α and β influence the document-topic distribution, θ and the observed words w, respectively. Document-topic distribution influences the word-topic assignment, z. These parameters work in concurrence to create topics and establish topic-document frequencies.

Then by leveraging tools discussed previously, like Variational Inference or Gibbs Sampling, samples are gathered using an iterative process. This iterative process helps us understand the tiered structure of LDA by underscoring the fact the parameters α and β are corpus-level, θ is document-level, and z and w are word-level. When the topic assignments become stable, the model is considered to have converged. By examining the model, we can glean and estimate useful information such as the distribution over topics within a document, or words within a topic. Note, Gibbs Sampling is used within this model and the pseudocode is outlined below in Algorithm 2 [11].

To conclude, LDA follows the iterative sampling and inference approach outlined above. LDA tries to reverse-engineer the generative process by adjusting and improving estimates for the topic-word and document-topic distributions iteratively.

3.2 Framework and Pseudocode for Dynamic Topic Models

A natural extension of the LDA baseline analysis introduces sequential corpora. As such, the assumption of document exchangeability is dropped. We can then introduce the data grouped into arbitrarily decided segments of time, which we will refer to as time slices. The documents within each slice are modeled with a K-component topic model. For a K-component model with V terms, let $\beta_{t,k}$ denote the V-vector for topic k in slice t.

Algorithm 2 LDA Gibbs Sampling

Require: words w in documents $\mathbf{w}$
Ensure: topic assignments z and counts $n_{\mathbf{w},k}$, $n_{k,w}$, and n_k

randomly initialize z and increment counters
for each iteration **do**
  for $i = 0$ to $N - 1$ **do**
    word $\leftarrow w[i]$
    topic $\leftarrow z[i]$
    n_d, topic$-= 1$; $n_{\text{word,topic}}-= 1$; $n_{\text{topic}}-= 1$
    for $k = 0$ to $K - 1$ **do**
      compute $p(z = k|\cdot) = \frac{(n_{d,k}+\alpha_k)(n_{k,w}+\beta_w)}{n_{k,w}+}$
    end for
    topic $\leftarrow$ sample from $p(z|\cdot)$
    $z[i] \leftarrow$ topic
    n_d, topic$+= 1$; $n_{\text{word,topic}}+= 1$;$n_{\text{topic}}+= 1$
  end for
end for
return z, $n_{d,k}$, $n_{k,w}$, n_k

With this notation, it is critical to see that the Dirichlet should not be applied to sequential modeling as the distribution is static and does not account for temporal dynamics of data [DTM]. However, we can overcome said challenge by introducing a chain of parameters for each topic $\beta_{t,k}$ following the general form:

$$\beta_t|\beta_{t-1} \sim (\beta_{t-1,k}, \sigma^2 I)$$

Similarly, we are to sample topic proportions using the Dirichlet. This is done by introducing a logistic normal with mean α to express uncertainty over proportions, following the general form:

$$\alpha_t|\alpha_{t-1} \sim (\alpha_{t-1}, \delta^2 I)$$

With this foundation established, we are able to extending upon the original LDA generative process. Note, t denotes a slice of a sequential corpus.

1. Draw topics $\beta_t|\beta_{t-1} \sim (\beta_{t-1}, \sigma^2 I)$
2. Draw $\alpha_t|\alpha_{t-1} \sim (\alpha_{t-1}, \delta^2 I)$
3. For each document:
   (a) Draw $\eta \sim (\alpha_t, \sigma^2 I)$
   (b) For each word:
      (i) Draw $Z \sim \text{Mult}(\pi(\eta))$
      (ii) Draw $W_{t,d,n} \sim \text{Mult}(\pi(\beta_{t,z}))$

As the posterior remains intractable due the nature of the Gaussian and multinomial models, several approximation approaches are recommended including Mean-Field Variational Inference, Structured Variational Inference, Online Variational Inference, and

Variational Kalman Filtering. In the original paper, the latter is chosen and is characterized by the approaches ability to describe how the state of the system evolves over time and relates that state to the observation. As such we can describe the pseudocode as follows:

Algorithm 3 Dynamic Topic Modeling with Variational Inference

Require: Corpus split into time slices: $\text{corpus}_{1:T}$
Require: Number of topics: K
Require: Maximum number of iterations: MaxIter
Ensure: Dynamic topics over time slices: $\theta_{1:T}$

```
Initialize topics θ_1 for the first time slice using LDA.
for t = 2 to T do
    Use θ_{t−1} to initialize the topics θ_t for time slice t.
    for i = 1 to MaxIter do                                   ▷ E-step
        for all documents d in corpus_t do
            Compute the variational distribution q(z_d) using the current estimate of θ_t.
        end for                                               ▷ M-step
        Update θ_t based on the aggregated information from all q(z_d).
        if convergence criteria met then
            Break
        end if
    end for
end for return θ_{1:T}
```

Given a corpus split into the time slices and a fixed a number of topics K, DTM is creating a dynamic set of topics $\theta_{1:T}$ that evolve over time slices. The algorithm starts by initializing the topics θ_1 for the first time slice using Latent Dirichlet Allocation (LDA). Then, for each subsequent time slice t, it initializes the topics θ_t using the topics θ_{t-1} from the previous slice. Within each time slice, topics are updated for each document based on the distributions calculated.

This portion of the code includes the two fundamental parts of the Expectation-Maximization (EM) algorithm, used to find the maximum likelihood estimates of model. The E-step (Expectation step) computes the variational distribution $q(\mathbf{z}_d)$ for each document where $\mathbf{z}_d$ represents the topic assignments for the words in document d. The M-step (Maximization step) takes the variational distributions and uses them to update the topic distribution for that time slice by maximizing some objective function. This process is continued until the dynamic topics converge and all time slices are returned. Further, this process can be generally visualized in Figure 2 in the appendix illustrating chained topic models.

4 Methods

4.1 Data Source

The dataset used for this project was sourced from Kaggle's News Category Dataset, which is available here: News Category Dataset. This dataset contains over 200,000 news articles with 42 unique categories of article types, proving to be ideal for topic modeling experiments.

4.2 Data Preprocessing

Before applying LDA or DTM, the dataset underwent the standard natural language preprocessing steps. The preprocessing pipeline is as follows:

1. **Basic Cleaning:** Removal of special characters, normalization to lower case, sentence splitting.
2. **Contraction Removal:** Replacing contracted forms like "it's" with "it is."
3. **Tokenization:** Conversion of sentences into individual words.
4. **Stop Word Removal:** Elimination of common and non-informative words like "and," "the," "is."
5. **Lemmatization:** Transforming words to root form (e.g., "coding" to "code").

4.3 Model Implementation

The models for LDA and DTM were implemented using the Gensim package. The LDA model was run 7 times testing fixed topic ranging from 3 to 10. This data ranged from 2012-01 through 2022-09. The DTM time slices were monthly. The range of data for the DTM model was truncated from 2021-01 through 2022-09.

4.4 Code Repository

The Jupyter Notebook for the project implementation is available on GitHub at cpuglis1/ProbStoch-I-Project.

5 Results

5.1 Quantitative Metrics

After running the LDA model, it was found that the ideal number of topics is 5. This was done by maximizing coherence, which is a common method for quantitatively describing model fitness. Generally, coherence scores measure the degree to which words in a single topic co-occur across a corpus and is calculated as

$$C_V(t) = \frac{1}{C(N,2)} \sum_{i \neq j} \text{similarity}(w_i, w_j)$$

where $C(N,2)$ is the number of unique pairs of words that can be made from N words, which is $\frac{N(N-1)}{2}$. Among the top N words from a topic, the similarity is calculated for each pair of words (w_i, w_j). The scores are usually averaged together and lie within the range [0,1] and any score above 0.5 is considered strong. LDA models were run with 3-10 fixed topics. In this case, 0.38 is the maximum score. See Figure 3 in the appendix. This will serve as the baseline coherence score.

Given the computational complexity of the DTM model, it was only feasible to run the model with 5 topics. This number of topics yielded an average coherence score of 0.25 for all time slices, which is much lower than the baseline score of 0.38 suggesting that sequential modeling of news data in our case is not an improvement upon LDA topic modeling. That said, given the computational complexity limits and truncated data set chosen for DTM, I believe the results are not conclusive. Let us continue by visualizing topic clusters.

5.2 Qualitative Analysis

As visualized in Figures 4 and 5 in the appendix, we can see how the topics were split in two dimensional space. The LDA clusters 3, 4, and 5 are likely split in additional components and only show overlap in the two dimensions. Further, the top 30 most-salient terms are listed on the right. We can see that terms related to trump, covid-19, and the police dominate the news distributions, which makes intuitive sense. When analyzing the 5 topics further, we can see trends in clustering based on the top 5 most salient terms in each category. This data is visualized in Table 1 in the appendix. We can see that the general overlap of salient terms is high, with political terms dominating nearly every topic. That said, we can begin to see the emergence of some hidden topics. For example, topic 4 seems to have grouped articles discussing the police and civil rights, while topic 3 discusses the election and the close race in Georgia. This suggests the topic model works, but that the data is likely imbalanced with political articles.

Further, we can take a look under-the-hood to see an example of the results. One article was represented in lemmatized form as the following list:

```
['campaign', 'house', 'member', 'trail', 'use', 'representative', 'covid-19',
'due', 'present', 'proxy', 'remotely', 'stopping', 'supposed', 'voting']
```

The probability of belonging to each topic is given as:

```
[6.64451827e-04 1.35225270e-01 6.64451827e-04 6.64451827e-04 8.62781374e-01]
```

We can see that the article was assigned to topic 5, but has some overlap with topic 2. Examining our salient terms, this categorization makes sense as those are the two Political and Covid like clusters.

5.3 Topic Over Time

As DTM produces results visible over each time slice, in this section, the top words within Topic 1 for 5 time-slices were examined for trends. Table 3 in the appendix underscores a changing landscape of political dialogue over time. Beginning with general political discourse regarding republicans and democrats in time slices 1 and 5, we can see an eventual shift to more polarized and personality-centric debate. These topics seem to eventually leaning back towards more general coverage, but this dynamic change could be reflective of significant national events and shifts in media coverage. When compared to LDA in this case, DTM is less coherent overall. However, DTM in some cases offers a valuable temporal perspective for some news articles, which could provide useful for other researchers and analysts.

5.4 Deeper Dive: LDA

As the clusters seem to be dominated by political articles, I extended my work to include a case where articles belong to a single category, 'SCIENCE'. This categories were predetermined by the dataset. For a 5-topic model was created for the $\sim$ 2200 SCIENCE articles, The coherence score jumped to 0.46 indicating that model performance can improve when the models are run on articles that are known to be related.

Table 2 in the appendix relays the most salient terms in each topic for SCIENCE articles. Upon examination, we can see some narratives begin to appear. For example, although many articles are related to space and related studies, we see that unique topics within that category are revealed in Topics 1 and 4. Topic 1 focuses on space exploration and NASA while Topic 4 focuses primarily on articles specifically mentioning celestial bodies. A general research and medical research topic is emerging within Topics 2 and 3, respectively. These narratives lead us to understand the potential impact of topic modeling when applied in the correct context.

5.5 Comparative Analysis

When comparing LDA and DTM on the same corpus, LDA yields higher coherence scores indicating that the topics generated are semantically meaningful and tightly clustered. Further, the significantly higher computational cost of DTM is concerning, as it restricts its feasibility of applications to larger datasets. As such, we can see that a more fitting application for DTM would involve a particular trend, like monitoring articles over an election cycle. This point was partly illustrated through the topic over time analysis where DTM captured political discourse article evolution over time.

Further, the high coherence score for the application of LDA to the SCIENCE category suggests the model could be most effective in certain niches where the semantic landscape is less cluttered by diverging topics. However, DTM's capacity to handle a broader range of topics may make it more suitable for a diverse dataset.

6 Discussion

LDA and DTM offer a unique set of advantages and disadvantages in topic modeling. LDA shows higher coherence overall, which suggests that LDA can be leveraged for more lucid results when analyzing news articles. That said, I am unable to make any definitive conclusions regarding the utility of LDA to this corpus given that the coherence is considered low at 0.5.

While computationally costly and less coherent, DTM is useful for modeling topic evolution in specific cases. This difference in coherence between LDA and DTM models leaves room for further applications of the models, but based on the analysis, a natural next step for this corpus should involve specific narratives. Filtering articles relating to an ongoing news story, one could apply DTM or LDA on those articles to understand the latent ideas hidden within the articles.

In conclusion, the choice between LDA and DTM depends on the specific requirements of the research question. Other probabilistic topic models may be better suited for the problem faced.

7 Appendix

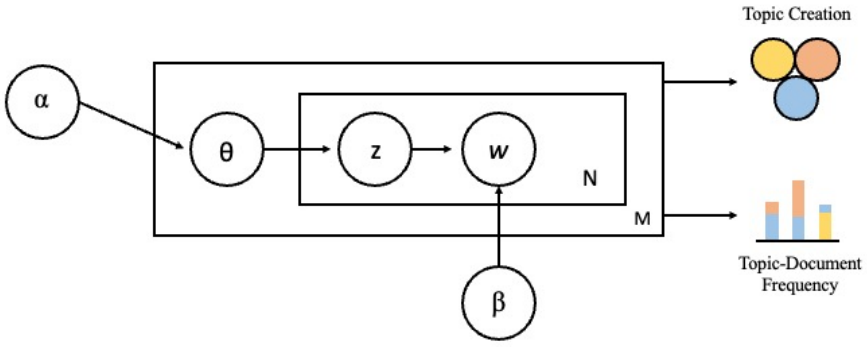

Figure 1: LDA Graphical Model. The Dirichlet parameters, α and β influence the document-topic distribution, θ and the observed words w, respectively. Document-topic distribution influences the word-topic assignment, z. Process leads to topic creation and topic-document frequency distributions.

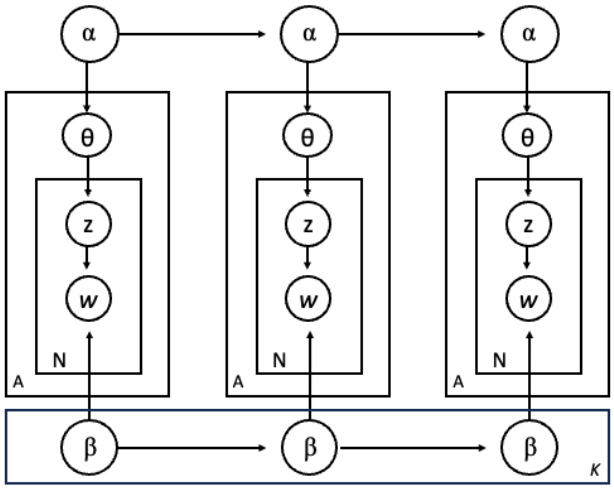

Figure 2: Dynamic Topic Model Graph. Graph representation of a dynamic topic model for three time slices. Parameters $\beta_{t,k}$ evolve over time with the mean parameters α_t.

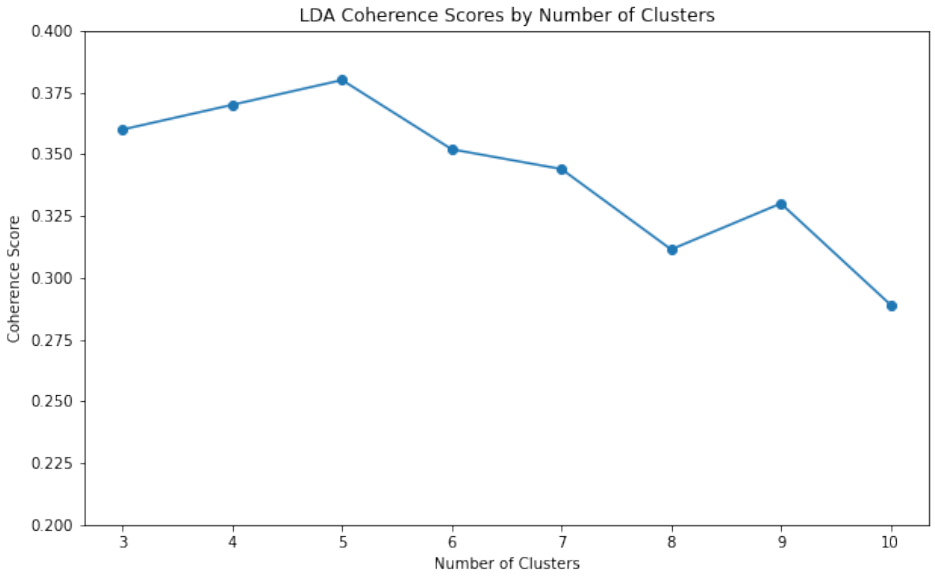

Figure 3: LDA Coherence Score Graph. The Coherence scores for each number of topics are plotted.

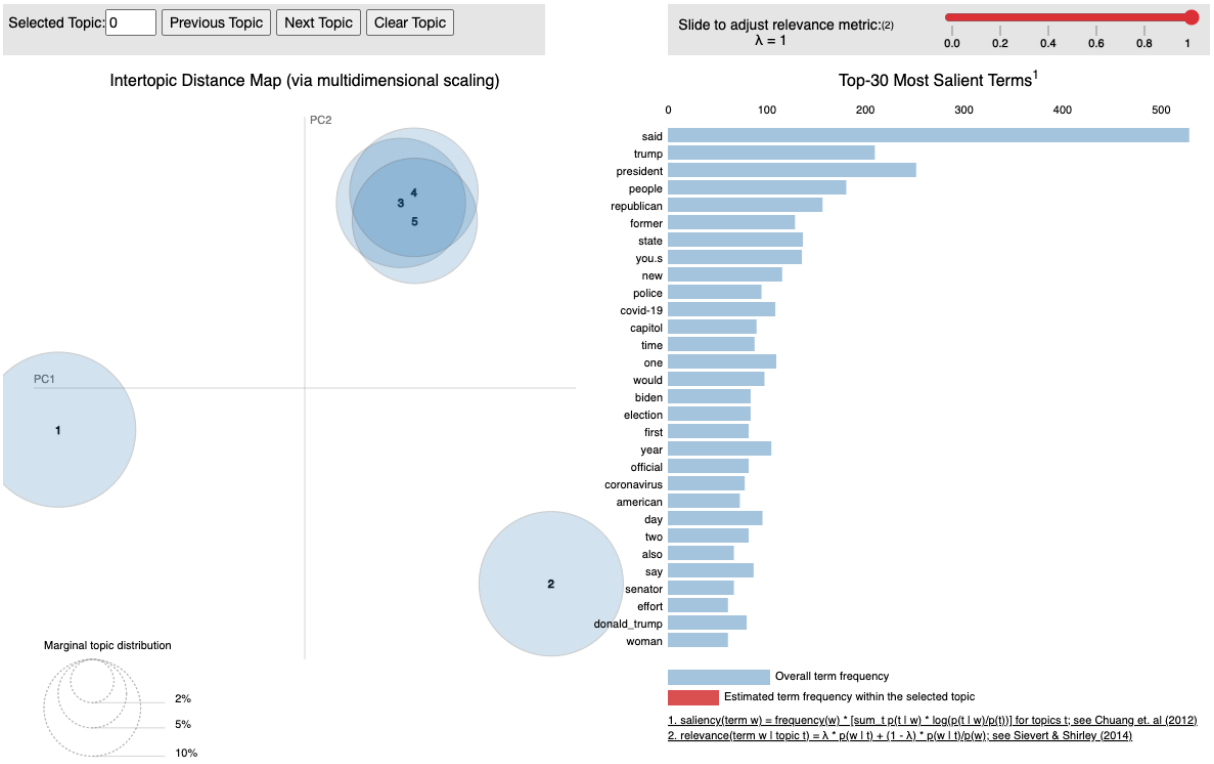

Figure 4: LDA Clusters Overview: On the left, we see the intertopic distance map plotted over principle components one and two. On the right, we see the most salient terms for all topics ranked by overall term frequency.

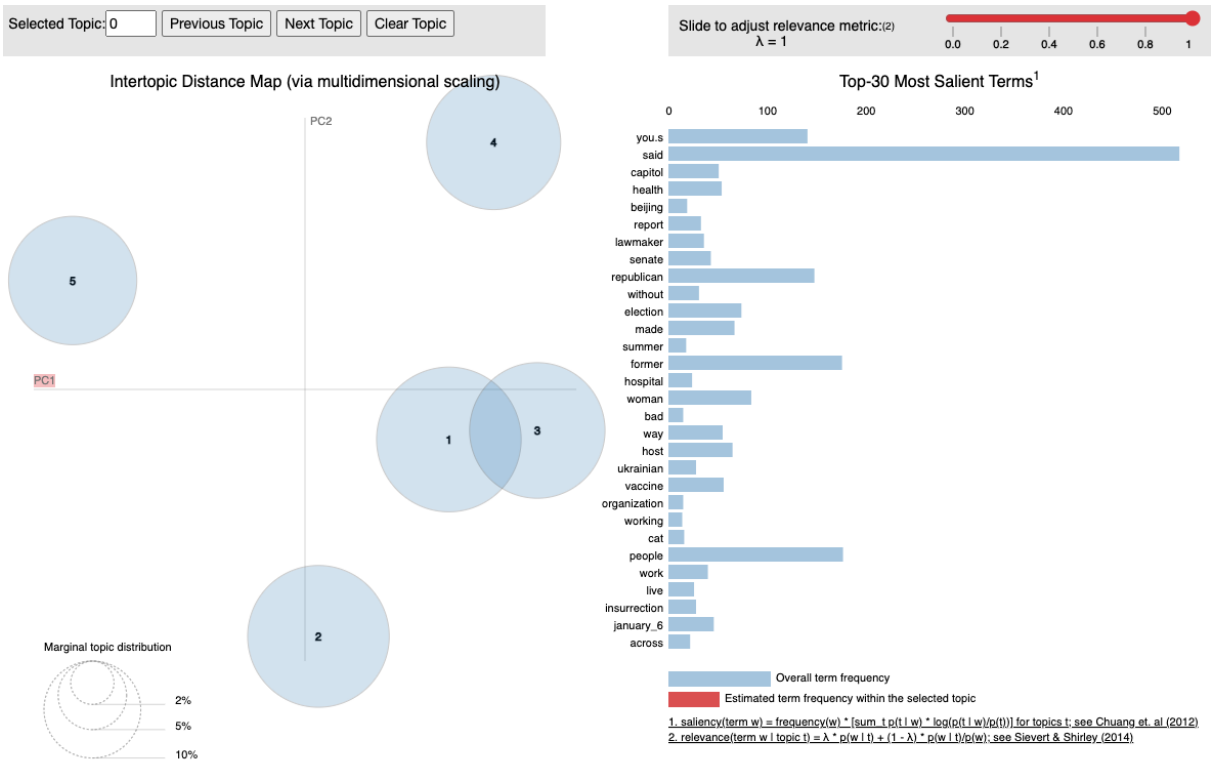

Figure 5: DTM Clusters Overview: On the left, we see the intertopic distance map plotted over principle components one and two for time slice 1. On the right, we see the most salient terms for all topics ranked by overall term frequency for time slice 1.

Topic Number	Top Terms	Notes
1	president, donaldtrump, state, pandemic, attack, congress	General Republican Topics
2	republican, gop, covid-19, vaccine, expert	COVID and Republican Topics
3	biden, election, georgia, policy, democrat	Election Coverage
4	trump, police, foxnews, black, right	Social Rights
5	capitol, cnn, mask, coronavirus, effort	COVID

Table 1: LDA: Notable Salient Terms: For the 5 topics, notable salient terms are listed and arbitrary notes are given discussing potential hidden topics within the data.

Time Slice	Top Terms	Notes
1	said, people, republican, new, day	General Political Discourse
5	republican, democrat, new, police, president	Political Discourse and Controversy
10	democrat, republican, trump, president, former	Higher Trump Coverage
15	trump, president, law, gov, republican	Higher Trump Coverage
20	people, republican, trump, democrat, day	Balanced Political Coverage

Table 2: DTM: Topic 1 Over Time. The top terms in Topic 1 for 5 different time slices are displayed. Arbitrary notes are given discussing potential hidden topics within the data.

Topic Number	Top Terms	Notes
1	nasa, solar, space, planet, spaceship, expedition	Outerspace
2	mission, gallery, period, published, live-science	Publications and Expositions
3	Antibiotic, study, research, report, human	Medical Research
4	moon, sun, star, earth, orbit	Celestial Objects
5	space-station, aurora, evidence, ancient, discovered	Nature or Discoveries

Table 3: LDA: SCIENCE Notable Salient Terms. For the 5 topics, notable salient terms are listed and arbitrary notes are given discussing potential hidden topics within the data.

References

[1] David M Blei. "Latent Dirichlet Allocation". en. In: ().

[2] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". en. In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2017.1285773. URL: https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1285773 (visited on 08/20/2023).

[3] David M. Blei and John D. Lafferty. "Dynamic topic models". en. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 113–120. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143859. URL: http://portal.acm.org/citation.cfm?doid=1143844.1143859 (visited on 08/19/2023).

[4] Jonathan Chang and David M. Blei. "Hierarchical relational models for document networks". In: *The Annals of Applied Statistics* 4.1 (Mar. 2010). arXiv:0909.4331 [stat]. ISSN: 1932-6157. DOI: 10.1214/09-AOAS309. URL: http://arxiv.org/abs/0909.4331 (visited on 08/20/2023).

[5] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. May 2019. DOI: 10.48550/arXiv.1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 08/21/2023).

[6] Matthew D Hoffman. "Stochastic Variational Inference". en. In: ().

[7] *How Many News Articles Are Published Every Day? (2023 Statistics) - The Small Business Blog*. en-US. Section: Data & Statistics. URL: https://thesmallbusinessblog.net/news-articles-published-every-day/ (visited on 08/19/2023).

[8] Alec Radford et al. "Improving Language Understanding by Generative Pre-Training". en. In: ().

[9] Michal Rosen-Zvi et al. *The Author-Topic Model for Authors and Documents*. arXiv:1207.4169 [cs, stat]. July 2012. DOI: 10.48550/arXiv.1207.4169. URL: http://arxiv.org/abs/1207.4169 (visited on 08/20/2023).

[10] A. F. M. Smith and G. O. Roberts. "Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods". en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 55.1 (1993). _eprint: https://onlinelibrary.wiley.com/doi/6161.1993.tb01466.x, pp. 3–23. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1993.tb01466.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1993.tb01466.x (visited on 08/20/2023).

[11] Bambang Subeno, Retno Kusumaningrum, and Farikhin Farikhin. "Optimisation towards Latent Dirichlet Allocation: Its Topic Number and Collapsed Gibbs Sampling Inference Process". en. In: *International Journal of Electrical and Computer Engineering (IJECE)* 8.5 (Oct. 2018), p. 3204. ISSN: 2088-8708, 2088-8708. DOI: 10.11591/ijece.v8i5.pp3204-3213. URL: http://ijece.iaescore.com/index.php/IJECE/article/view/10087 (visited on 08/20/2023).

[12] Yee Whye Teh et al. "Hierarchical Dirichlet Processes". en. In: *Journal of the American Statistical Association* 101.476 (Dec. 2006), pp. 1566–1581. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214506000000302. URL: https://www.tandfonline.com/doi/full/10.1198/016214506000000302 (visited on 08/21/2023).

[13] Hanna M. Wallach. "Topic modeling: beyond bag-of-words". en. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 977–984. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143967. URL: http://portal.acm.org/citation.cfm?doid=1143844.1143967 (visited on 08/21/2023).

[14] David Wingate and Theophane Weber. *Automated Variational Inference in Probabilistic Programming*. arXiv:1301.1299 [cs, stat]. Jan. 2013. DOI: 10.48550/arXiv.1301.1299. URL: http://arxiv.org/abs/1301.1299 (visited on 08/20/2023).