Journal of Technology and Informatics (JoTI)
Vol. 5, No. 2, April 2024
P-ISSN 2721-4842
E-ISSN 2686-6102 DOI: 10.37802/joti.v5i2.618

Topic Modeling for Evolving Textual Data Using LDA, HDP, NMF, BERTOPIC, and DTM With a Focus on Research Papers

C.B. Pavithra[1], **J. Savitha**[2]
[1,2]Department of Information Technology, Dr.N.G.P. Arts & Science College, Coimbatore, Tamilnadu, India
E-mail: c.b.pavithramsc2004@gmail.com[1], savithaj@drngpasc.ac.in[2]
*Corresponding author: E-mail: c.b.pavithramsc2004@gmail.com

Abstract: *As the volume of academic literature continues to burgeon, the necessity for advanced tools to decipher evolving research trends becomes increasingly apparent. This study delves into the utilization of topic modeling techniques—specifically Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), Non-negative Matrix Factorization (NMF), BERTopic, and Dynamic Topic Modeling (DTM)—applied to a dynamic corpus of research papers. Our research endeavors to confront the challenges posed by capturing temporal dynamics, evolving terminology, and interdisciplinary themes within academic literature. Through a comprehensive comparative investigation of these models, we assess their efficacy in extracting and tracking research topics over time. While DTM exhibited the highest term topic probability, its inclusion of non-meaningful words proved to be a hindrance to its suitability. Conversely, NMF, HDP, LDA, and BERTopic demonstrated comparable performance in topic extraction. Surprisingly, DTM emerged as the most effective model in our research, showcasing its prowess in navigating the intricacies of evolving research trends.*

Keywords: *BERTopic; Dynamic Topic Modeling (DTM); Evolving Textual Data; Hierarchical Dirichlet Process (HDP); Latent Dirichlet Allocation (LDA)*

INTRODUCTION

The rapid expansion of academic literature presents a daunting challenge in understanding the dynamic landscape of research trends. As scholarly output continues to escalate, the demand for advanced analytical tools becomes ever more pressing. Topic modeling, a formidable technique in natural language processing, has proven its effectiveness in revealing hidden thematic structures within textual data [1]. In this context, our research focuses on the application of five distinct topic modeling approaches Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), Non-negative Matrix Factorization (NMF), BERTopic, and Dynamic Topic Modeling (DTM) with a specific emphasis on their performance in capturing the temporal dynamics of research topics within a diverse corpus of academic papers. The exponential growth of scholarly publications in recent years has transformed the academic landscape, presenting both opportunities and challenges [2]. The sheer volume and diversity of research output across various disciplines have made it increasingly difficult for researchers, practitioners, and decision-makers to keep pace with emerging trends and evolving knowledge domains. In this context, the application of advanced computational methods, such as topic modeling, becomes essential to distill meaningful insights from large corpora of textual data.

Globalization, technological progress, and enhanced collaboration among researchers have fueled a surge in academic literature, making knowledge extraction challenging due to the overwhelming volume of publications. Traditional search methods struggle to discern patterns, identify seminal works, and track the evolution of ideas effectively [3]. Research topics evolve with scientific inquiry, accommodating emerging disciplines, interdisciplinary trends, and evolving academic terminology. Traditional static models often fall short in capturing dynamic nature and subtle changes in research themes. Thus, there's a demand for methodologies adept at adaptively modeling and uncovering latent structures within evolving textual data. [4]. Topic modeling, within natural language processing, provides a promising solution for handling vast and dynamic textual datasets. These models extract latent topics, enhancing comprehension of underlying themes and trends. However, their effectiveness in capturing the temporal dynamics of evolving research topics, especially in academic literature, is still being explored. [5].

This research is positioned to make substantial contributions by:

1. Evaluating the efficacy of both traditional and advanced topic modeling techniques within the realm of evolving research papers.
2. Revealing valuable insights into the temporal progression of research topics and the capacity of various models to adapt to shifting scholarly landscapes.
3. Establishing groundwork for improving knowledge discovery processes, thereby assisting researchers, educators, and decision-makers in navigating the ever-changing terrain of academic literature.

Our primary objectives are threefold. Firstly, we aim to investigate the challenges posed by evolving textual data, particularly in the realm of research papers. Secondly, we introduce and implement five distinct topic modeling techniques LDA, HDP, NMF, BERTopic, and DTM to discern their effectiveness in capturing evolving research topics. Thirdly, we seek to provide a comparative evaluation of these models, shedding light on their respective strengths and weaknesses in the context of dynamic academic literature.

This research endeavors (activities) to answer key questions:

1. How do traditional and advanced topic modeling techniques perform in the analysis of a dynamic corpus of research papers?
2. Can these models effectively capture the temporal aspects and evolving trends in academic research?
3. What insights can be gleaned from the comparative analysis of LDA, HDP, NMF, BERTopic and DTM in the context of research paper datasets?

In the subsequent sections, we delve into the existing literature, outline our methodology, present the models employed and discuss the results and implications of our findings. This study contributes to the broader understanding of topic modeling applications in the analysis of evolving textual data, particularly within the intricate domain of academic research literature.

METHOD

As the volume of research papers continues to grow exponentially, the need for effective tools to distill and comprehend the underlying themes becomes paramount. Topic modeling, a branch of natural language processing (NLP) has emerged as a powerful computational technique for uncovering latent structures within large textual corpora. In the context of research papers, topic modeling serves as a valuable method for revealing the inherent thematic structures, trends, and shifts in scholarly discourse. Topic modeling refers to a suite of algorithms designed to identify topics present in a collection of documents without the need for prior annotation or human supervision. The fundamental assumption is that each document is a mixture of topics, and each topic is a mixture of words. The goal is to extract these latent topics and their associated word distributions, providing a succinct representation of the major themes within a corpus [6].

The primary purpose of applying topic modeling to research papers is to facilitate the automatic discovery of prevalent themes, trends, and relationships embedded in the vast and diverse scholarly literature. By discerning topics and their evolution over time, researchers gain a deeper understanding of the prevailing concerns, emerging subfields and interdisciplinary intersections within their domain of study.

Challenges in Modeling Evolving Research Topics

Modeling evolving research topics poses unique challenges that stem from the dynamic nature of scholarly discourse, the emergence of new fields, and the constant evolution of research paradigms. Traditional topic modeling approaches, designed for static corpora, encounter limitations when applied to datasets characterized by temporal shifts and changing trends [7]. This section outlines key challenges in effectively capturing evolving research topics and underscores the need for advanced methodologies.

Table 1: Methodologies and Challenges in Modeling Evolving Research Topics

Advanced methodologies	Challenge	Implication
Rapid Changes in Research Focus	The rapid evolution of research fields and the emergence of new disciplines result in sudden shifts in focus.	Traditional models may struggle to adapt quickly, leading to the risk of overlooking nascent research trends and failing to capture the latest developments.
Emergence of Interdisciplinary Fields	Research is increasingly interdisciplinary, spanning traditional disciplinary boundaries.	Models must be capable of identifying and accommodating interdisciplinary connections, which can be challenging for algorithms designed with a single discipline focus.
Need for Adaptive Models	The static nature of traditional topic models may hinder their ability to adapt to changes over time.	Adaptive models are essential to capture the evolving nature of research topics, ensuring accurate representation and timely identification of emerging themes.
Evolving Terminology and Concepts	The introduction of new terminology and conceptual frameworks requires models to dynamically update their understanding.	Failure to adapt to evolving language may lead to misinterpretations of topics and hinder the accurate representation of emerging research trends.
Temporal Aspects of Topic Evolution	Understanding the temporal dynamics of topic evolution is crucial for tracking the life cycle of research themes.	Traditional models may lack the temporal granularity needed to capture how topics emerge, evolve and decline over time.
Data Sparsity and Noise	Sparse datasets and noisy	Models must be resilient to noise

Advanced methodologies	Challenge	Implication
	information, common in evolving research domains can impact the robustness of models.	and capable of extracting meaningful patterns from datasets with varying levels of information density.
Evaluation Metrics for Temporal Coherence	Traditional evaluation metrics may not adequately capture the temporal coherence of evolving research topics.	Novel metrics and methodologies are needed to assess the performance of models in capturing the dynamic nature of scholarly discourse.

Applications in Various Domains

Topic modeling finds applications across various domains within the realm of research papers:

1. *Literature Review Automation:* Automated topic modeling aids researchers in conducting comprehensive literature reviews by efficiently identifying and summarizing the key themes across a vast body of work.
2. *Trend Analysis:* By analyzing the temporal evolution of topics, researchers can gain insights into the emergence and fading of research trends, facilitating proactive engagement with evolving fields.
3. *Interdisciplinary Exploration:* Topic modeling enables the identification of interdisciplinary connections within research papers, revealing how different domains converge and influence each other.
4. *Recommendation Systems:* In academic databases and repositories, topic modeling can be employed to enhance recommendation systems, suggesting relevant papers based on shared thematic content.

While traditional topic modeling algorithms have proven effective in extracting themes from static datasets, the dynamic nature of research papers necessitates continuous refinement and adaptation of these methods [8]. This analysis builds upon existing research by applying a diverse set of topic modeling techniques likeLDA, HDP, NMF, BERTopic, and DTM to a dynamic corpus of research papers, aiming to advance our understanding of evolving textual data within the scholarly domain. The subsequent sections detail the methodology employed, the selection and implementation of each model, and the comparative analysis of their performance in capturing evolving research topics.

TOPIC MODELING MODELS

In this section, we introduce five distinct topic modeling models such as Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), Non-negative Matrix Factorization (NMF), BERTopic, and Dynamic Topic Modeling (DTM). Each model brings unique characteristics and capabilities to the analysis of evolving textual data, particularly within the context of research papers.

Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model that assumes each document in a corpus is a mix of topics, and each topic is a mix of words. It works by assigning a probability distribution of topics to each document and a probability distribution of words to each topic. LDA has been widely adopted for topic modeling due to its simplicity and interpretability [9]. However, its static nature may limit its effectiveness in capturing the temporal dynamics of evolving research topics.

Latent Dirichlet Allocation (LDA) Algorithm:

Step 1: Initialization:

1. For each document *d* in the corpus:
   Assign a distribution of topics $\theta_d \sim$ Dirichlet(α), where α is the hyper parameter for the Dirichlet prior on document-topic distributions.
2. For each word w in document d:
   1. Assign a topic $z_{d,w} \sim$ Multinomial(θ_d), where $z_{d,w}$ is the topic assignment for word w in document d.
   2. Assign a word w to topic $z_{d,w}$ based on the topic-word distribution $\phi z_{d,w} \sim$ Multinomial(β), where β is the hyper parameter for the Dirichlet prior on topic-word distributions.

Step 2: Iterative Process:

2.1 For each iteration until convergence:

1. For each document d and each word w in document d:
2. Compute $P(z_{d,w} = k \mid \text{all other } z)$, the probability that word w in document d belongs to topic k.
3. $$P(z_{d,w} = k \mid all\ other\ \ z)\ \alpha\ (n_{(t)}^{d,k} + \alpha\ /\ \sum_k n(t)d,k + \alpha\ )\ x\ (n_{(t)}^{w,k} + \beta\ /\ \sum_k n(t)w,k + \beta) \quad (1)$$

   Where:
   1. $n_{(t)}^{d,k}$ is the number of words in document d assigned to topic k up to iteration t.
   2. $n_{(t)}^{w,k}$ is the number of times word w is assigned to topic k up to iteration t.
   3. α is the Dirichlet hyper parameter for document-topic distributions.
   4. β is the Dirichlet hyper parameter for topic-word distributions.
   5. Sample a new topic assignment $z_{d,w}$ based on the computed probabilities.

Step 3: Output

After convergence, output the inferred topic assignments and the learned document-topic and topic-word distributions.

$$P(\theta_d,\phi_k \mid all\ z) \propto (n_{(T)}^{d,k} + \alpha / \sum_k n(T)d,k\ +\ \alpha\ ) \times (n_{(T)}^{w,k} + \beta / \sum_k n(T)w,k\ +\ \beta) \quad (2)$$

Where T is the total number of iterations.

Note:

1. T is the number of iterations.
2. $n_{(T)}^{d,k}$ is the number of words in document d assigned to topic k at the end of iteration T.
3. $n_{(T)}^{w,k}$ is the number of times word w is assigned to topic k at the end of iteration T.

The algorithm aims to discover the latent topics in a collection of documents and the distribution of words associated with each topic.*Initialization:* For each document, assign a mixture of topics. The number of topics is a parameter specified by the user. For each word in the document, assign it to one of the topics. *Iterative Process:* Iterate through each document and each word in the document multiple times. During each iteration, reassign the word to a different topic based on the current distribution of topics in the document and the distribution of words in the topic. Update the topic assignments for all words in all documents iteratively.*Output:* After a sufficient number of iterations, the model converges, and the final assignments represent the discovered topics for each document and the distribution of words for each topic. *Probability Distributions:* The outcome of LDA is two probability distributions are Document-Topic Distribution: The probability of each topic in each document and Topic-Word Distribution: The probability of each word in each topic. *Inference:* Once trained, the model can be used for inference. Given a new document, LDA can infer the distribution of topics in the document and the distribution of words in each topic. *Hyper parameters:* LDA has hyper parameters that need to be set, such as the number of topics, the Dirichlet priors for document-topic and topic-word distributions, and the number of iterations [10][11].

Hierarchical Dirichlet Process (HDP)

The Hierarchical Dirichlet Process (HDP) is an extension of the Latent Dirichlet Allocation (LDA) model, designed to address some of the limitations of LDA, particularly in cases where the number of topics is unknown or may change over time**.** HDP introduces a hierarchical structure that allows for an infinite number of topics, providing a more flexible framework for capturing the complexities of real-world data [12]. Here is an overview of the HDP algorithm.

HDP Algorithm:

Step 1: Initialization:

For each document d in the corpus:

1. Assign a global topic distribution G0 ~ Dirichlet(γ), where γ is a hyper parameter controlling the strength of the global distribution.

For each document d and each word w in document d:

1. Assign a document-specific topic distribution $\theta_d \sim \text{Dirichlet}(G_0)$.
2. Assign a topic $z_{d,w} \sim \text{Multinomial}(\theta_d)$, representing the global topic assignment for word w in document d.
3. Assign a word w to topic $z_{d,w}$ based on the topic-word distribution $\phi z_{d,w} \sim \text{Multinomial}(\beta)$, where β is a hyper parameter for the Dirichlet prior on topic-word distributions.

Step 2: Iterative Process:

2.1 For each iteration until convergence:

For each document d and each word w in document d:

Compute $P(z_{d,w} = k \mid \text{all other } z)$, the probability that word w in document d belongs to topic k.

$$P(z_{d,w} = k \mid all\ other\ z) \propto (n_{(t)}^{d,k} + \alpha / \sum_k n(t)d,k\ +\ \alpha\ ) \times (n_{(t)}^{w,k} + \beta / \sum_k n(t)w,k\ +\ \beta\ ) \quad (3)$$

Where:

1. $n_{(t)}^{d,k}$ is the number of words in document d assigned to topic k up to iteration t.
2. $n_{(t)}^{w,k}$ is the number of times word w is assigned to topic k up to iteration t.
3. α is the Dirichlet hyper parameter for document topic distributions.
4. β is the Dirichlet hyper parameter for topic-word distributions.
5. Sample a new topic assignment $z_{d,w}$ based on the computed probabilities.

For each topic k:

Update the global topic distribution G_0 based on the documents assigned to topic k and the global hyper parameter γ.

Step 3: Output:

After convergence, output the inferred topic assignments, the learned document-specific topic distributions, and the global topic distribution.

$$P(\theta_d,\phi_k \mid all\ z) \propto (n_{(T)}^{d,k} + \alpha / \sum_k n(T)d,k\ +\ \alpha\ ) \times (n_{(T)}^{w,k} + \beta / \sum_k n(T)w,k\ +\ \beta) \quad (4)$$

Where T is the total number of iterations.

The Hierarchical Dirichlet Process (HDP) is an advanced probabilistic model and an extension of the Latent Dirichlet Allocation (LDA) algorithm, designed to address the challenges associated with the dynamic nature of topics in a corpus. Unlike LDA, HDP allows for an infinite number of topics, adapting more naturally to situations where the number of underlying themes is unknown or changes over time. In the initialization step, each document is assigned a global topic distribution sampled from a Dirichlet distribution. For each word in a document, a document-specific topic distribution is sampled from the global distribution. The iterative process involves refining these distributions based on the observed data, allowing topics to emerge and evolve organically. Importantly, HDP introduces a hierarchical structure that facilitates sharing of topics among documents, capturing the complexity of **real-world scenarios** where documents may exhibit diverse themes. This hierarchical approach provides a more flexible and adaptive framework, making HDP well-suited for applications in which the underlying structure of topics is intricate and may vary across different subsets of the

data. The output of the HDP algorithm includes the inferred topic assignments, document-specific topic distributions, and the global distribution, offering a rich representation of the latent thematic structures present in the corpus [13].

Non-negative Matrix Factorization (NMF)

NMF is a matrix factorization technique that factorizes a matrix into two lower-dimensional matrices, each containing only non-negative values. In the context of topic modeling, NMF identifies latent topics and their associated word distributions. NMF is known for its interpretability and ability to capture parts-based representations. While traditionally applied to static datasets, adaptations of NMF for dynamic corpora have been proposed to address evolving research trends [14]. Here's an explanation of the NMF algorithm:

NMF Algorithm:

Step 1: Initialization:

1. For a given document-term matrix V of dimensions m X n, where m is the number of documents and n is the number of terms.
2. Initialize two non-negative matrices W and H with random or predefined non-negative values.
3. Set a target rank k, representing the desired number of topics.

Step 2: Iterative Process:

For each iteration until convergence:

1. Update matrix W by solving the following optimization problem:
   W >= 0, Vapprox WH
2. Update matrix \(H\) by solving the optimization problem:
   H >= 0, V approx WH

   Minimize the difference between the original matrix V and the product WH by adjusting the values in matrices W and H while ensuring non-negativity.

Step 3: Output:

After convergence, the matrices W and H represent the factorization of the original matrix V.

1. Matrix W (of dimensions m X k contains the document-topic distribution, where each column represents the strength of each document in each topic.
2. Matrix H (of dimensions k X n) contains the topic-term distribution, where each row represents the importance of each term in each topic.

NMF is particularly suited for topic modeling due to its ability to generate non-negative, interpretable factorizations. The algorithm iteratively refines the factorization by minimizing the difference between the original data and the product of the factorized matrices. The resulting matrices provide insights into the distribution of topics across documents and the distribution of terms across topics, offering a clear and interpretable representation of the latent thematic structures in the given corpus [15]. NMF is widely used in various applications, including text mining, image processing and bioinformatics, where parts-based representations are valuable for data analysis and interpretation [16].

Bidirectional Encoder Representations from Transformers (BERTopic)

BERTopic is a novel approach to topic modeling that leverages transformer-based embeddings, such as BERT (Bidirectional Encoder Representations from Transformers), to enhance the accuracy and interpretability of topic extraction. This methodology, which builds upon the strengths of BERT embeddings, addresses some of the limitations of traditional topic modeling algorithms, especially in capturing nuanced semantic relationships within textual data [17]. Here's an overview of the BERTopic approach.

BERTopic Algorithm:

Step 1: BERT Embeddings:

1.1. *Embedding Documents:*

1. Utilize a pre-trained BERT model to generate contextualized embeddings for each document in the corpus.
2. The embeddings capture the semantic meaning and context of words within the documents.
3. Let D be the set of documents, and BERT (d) be the BERT embedding for document d.

1.2. *Dimensionality Reduction:*

1. Apply dimensionality reduction techniques, such as UMAP (Uniform Manifold Approximation and Projection) or t-SNE (t-Distributed Stochastic Neighbor Embedding), to reduce the high-dimensional BERT embeddings to a lower-dimensional space.
2. This step helps maintain the semantic relationships while making the computational processing more efficient. Let X is the matrix of reduced-dimensional embeddings.

Step 2: Clustering:

2.1. *Density-Based Clustering:*

1. Apply HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), a density-based clustering algorithm.
2. HDBSCAN is effective in identifying clusters of varying shapes and densities, making it suitable for capturing complex topic structures.

2.2. *Topic Discovery:*

1. Utilize clustering algorithms, such as HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), to group similar documents based on the reduced-dimensional BERT embeddings.
2. HDBSCAN is particularly effective in identifying clusters of varying shapes and densities, making it suitable for diverse and dynamic topics.

Step 3: Topic Representation:

3.1. *Keyword Extraction:*

1. Extract representative keywords for each identified cluster by considering the most frequent terms within the documents belonging to that cluster.
2. The keywords provide a succinct representation of the main themes within each discovered topic.
3. For each cluster c_i in C, extract representative keywords by considering the most frequent terms within the documents in c_i.

4. Let KW_{ci} represent the set of keywords for cluster c_i.

3.2. *Topic Labels:*

1. Assign a label to each topic based on the extracted keywords, making it easier for users to interpret and comprehend the content encapsulated by each cluster.
2. Assign a label to each cluster based on the extracted keywords.
3. Let Lc_i be the label assigned to cluster c_i.
4.

Advantages of BERTopic:

1. Semantic Understanding: BERT embeddings capture the semantic relationships between words, leading to a more nuanced understanding of document content.
2. Context-Aware Embeddings: Contextual embeddings provided by BERT ensure that the representation of each word considers its context within the document.
3. Adaptive Clustering: HDBSCAN, used for clustering is capable of adapting to varying cluster shapes and densities, providing a more flexible approach to topic discovery.
4. Interpretability: By extracting representative keywords and assigning labels to topics, BERTopic enhances the interpretability of the discovered topics, facilitating user understanding.

BERTopic is a cutting-edge approach to topic modeling that harnesses the power of BERT embeddings and advanced clustering techniques. By incorporating semantic understanding and adaptability in clustering, BERTopic contributes to more accurate and interpretable research topic analysis, particularly in domains where capturing nuanced relationships and dynamic topic structures is crucial [18]. This methodology has demonstrated effectiveness in various applications, including research paper analysis, document clustering and thematic content extraction.

Dynamic Topic Modeling (DTM)

DTM is designed explicitly for modeling the evolution of topics over time. Unlike traditional topic models, DTM accounts for temporal dependencies in the data, allowing it to capture how topics change and transition from one state to another**.** DTM is well-suited for tracking the temporal dynamics of evolving research topics, making it a valuable tool for understanding the progression of themes within a corpus of research papers. Dynamic Topic Modeling (DTM) is a method used to model the evolution of topics in a corpus over time. It is an extension of Latent Dirichlet Allocation (LDA) that incorporates time as a parameter, allowing for the analysis of how topics change and transition across different time periods [19]. Below is a simplified algorithm for Dynamic Topic Modeling with each step and relevant formulas:

DTM Algorithm:

Step 1: Initialization:

Initialize Parameters:

1. Set the number of topics K, the number of time periods T, and other hyper parameters.
2. Initialize matrices for document-topic proportions $\theta_{d,t}$ and topic-word probabilities $\phi_{k,t}$.

Step 2: Iterate Over Time Periods:

For each time periodt from 1 to T:

2.1.1 Initialization for Time Period t:

Set initial values for $\theta_{d,t}$ and $\phi_{k,t}$ based on the results from the previous time period.

2.1.2 For each iteration until convergence:

1. Update document-topic proportions $\theta_{d,t}$ based on document content and the current estimate of $\phi_{k,t}$.

$$\theta_{d,t} \propto \exp(\psi(\gamma_d) + \sum_w n_{d,w,t}\, \phi_{w,k,t}) \quad (5)$$

2. Update topic-word probabilities \(\phi_{k,t}\) based on the words in the documents and the current estimate of \(\theta_{d,t}\).

$$\theta_{w,k,t} \propto \exp(\psi(\beta_d) + \sum_d n_{d,w,t}\, \phi_{d,t}) \quad (6)$$

2.1.3 Normalize Parameters:

Normalize $\theta_{d,t}$ and $\phi_{k,t}$ to ensure they sum to 1.

Step 3: Output:

3.1 Topic Evolution:

The result is a set of topic proportions $\theta_{d,t}$ and topic-word probabilities $\phi_{k,t}$ for each time period.

Notes:

1. γd and βw are hyper parameters associated with document d and word w, respectively.
2. $n_{d,w,t}$ represents the count of word w in document d during time period t.
3. $\psi(.)$ denotes the digamma function.
4. The algorithm iterates over time periods, updating topic proportions and word probabilities for each time slice, allowing for the modeling of dynamic topic evolution.

This algorithm provides a high-level overview of the DTM process. The actual implementation may involve additional considerations, such as convergence criteria, handling of hyper parameters, and optimization for efficiency. The goal is to capture how topics change over time in a dynamic corpus. Dynamic Topic Modeling (DTM) is an advanced technique in natural language processing that extends Latent Dirichlet Allocation (LDA) to account for the temporal evolution of topics within a corpus. The algorithm aims to discover how topics change and transition over different time periods. In the initialization step, parameters such as the number of topics K, the number of time periods T, and hyper parameters are set. The iterative process involves updating document-topic proportions $\theta_{d,t}$ and topic-word probabilities $\phi_{k,t}$ for each time period. This update is performed by considering the content of documents, the distribution of words, and the temporal context [20]. The model iterates over time periods, refining the estimates of topic proportions and word probabilities until convergence. The output is a set of evolving topic distributions, revealing how topics shift and emerge over time. Formulas incorporating the digamma function and word-document counts guide the update process. DTM

is particularly valuable in scenarios where topics exhibit temporal dynamics, providing insights into the changing thematic structures within a dynamic corpus. The implementation of DTM involves careful consideration of convergence criteria, hyper parameter tuning, and efficient optimization to effectively capture the nuanced evolution of topics over time [21].

Summary: Each of the introduced models brings unique strengths to the task of topic modeling in the context of evolving research papers. LDA and HDP provide a foundational understanding of topics and their hierarchical relationships. NMF excels in capturing non-negative, interpretable topic representations. BERTopic harnesses transformer-based embeddings for robust and context-aware topic extraction. DTM is tailored explicitly for modeling temporal changes in topics, providing insights into the evolving nature of research areas. In the following sections, these models will be implemented and evaluated to assess their effectiveness in capturing evolving research topics within a dynamic corpus of research papers.

RESULTS AND DISCUSSION

Data preprocessing

Data preprocessing is a critical phase in the data analysis and machine learning workflow, involving several key steps to ensure that raw data is transformed into a suitable format for analysis or model training. Initially, data cleaning addresses some issues such as missing values, duplicate records, and outliers, either through imputation, removal or outlier handling strategies. The transformation phase follows, encompassing tasks such as feature scaling to standardize numerical features, encoding categorical variables and processing text data by tokenization, removing stop words and applying techniques like stemming or lemmatization. Handling imbalanced data is important for classification tasks, involving methods to balance class distributions. Feature engineering may introduce new variables to enhance model performance and dimensionality reduction techniques like Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP) etc., can be applied to reduce high-dimensional data. Additionally, data splitting into training and testing sets is necessary for model evaluation. Other considerations include addressing noise, handling skewed (titled) data through transformations, normalizing numerical data, and appropriately dealing with time series data, such as incorporating lag features. The process is not one-size-fits-all, and the choice of preprocessing steps depends on the nature of the data and the specific goals of the analysis or modeling task. Python libraries like scikit-learn offer tools for various preprocessing tasks, streamlining the overall data preparation process.

Dataset

"Advanced Topic Modeling for Research Articles 2.0": In the vast landscape of scientific articles available online, researchers face challenges in locating pertinent information**.** The abundance of research content makes it increasingly challenging to identify relevant articles. Tagging and topic modeling offer a solution by providing a clear means of identifying research articles, facilitating the recommendation and search processes. Building upon our previous efforts, where we organized a Hackathon on Independence Day to predict topics for articles in the test set, this Live Hackathon takes us a step further. Now, our focus is on predicting the tags associated with each article. The task involves predicting tags for a set of research articles based on their abstracts in the test set. It's important to note that a single research article may have multiple tags. The research article abstracts are derived from four main topics: Computer Science, Mathematics, Physics and Statistics. Data set Download from https://www.kaggle.com/datasets/abisheksudarshan/topic-modeling-for-research-articles/

In this experiment, we conducted experiments using various Topic Modeling (TM) methods on widely employed public text datasets for the 29 research topic tasks and short conversations from the Research Articles 2.0, as outlined in Table 2.

Table 2: Statistics of Our Involved Datasets

Dataset	Description
Advanced Topic Modeling for Research Articles 2.0 29-Research Topics	14,000 documents Average document length: 60 **Topics:** Computer Science, Mathematics, Physics, Statistics, Analysis of PDEs, Applications, Artificial Intelligence, Astrophysics of Galaxies, Computation and Language, Computer Vision and Pattern Recognition, Cosmology and Non galactic Astrophysics, Data Structures and Algorithms, Differential Geometry, Earth and Planetary Astrophysics, Fluid Dynamics, Information Theory, Instrumentation and Methods for Astrophysics, Machine Learning, Materials Science, Methodology Number Theory, Optimization and Control, Representation Theory, Robotics, Social and Information ,Networks, Statistics Theory, Strongly Correlated Electrons, Superconductivity, Systems and Control

Performance Evaluation

Our research work assessed the quality and performance of five commonly used TM techniques, employing statistical measures such as precision, recall, and F-score for accuracy verification across different numbers of features (f = 50 and 500). Additionally, determining the optimal number of topics to extract from the corpus is a critical user-driven decision. In our experiment, we extracted four topics (t = 15 and 25), and the calculations for recall, precision, and F-score.

Recall (R), a standard information retrieval metric, gauges (estimate) the proportion of relevant items among the recommended items.

$$\text{Recall} = tp / tp + fn \quad (7)$$

Precision (P) is a widely used information retrieval metric, quantifying the proportion of retrieved recommended items to the actual relevant items.

$$\text{Precision} = tp/ tp + fp \quad (8)$$

The F-score (F) serves as a comprehensive measure of retrieval effectiveness, calculated by combining two key metrics in text mining: recall and precision.

F-score = Precision x Recall / Precision + Recall

It is essential to note the definitions of true positive (TP), representing the number of keywords correctly identified as a topic; false positive (FP), denoting the number of non-keywords incorrectly identified as a topic; true negative (TN), signifying the number of non-keywords accurately identified as non-topics; and false negative (FN), indicating the number of topics erroneously identified as non-topics.

During our data extraction phase, our objective is to extract topics from clusters of input data. As previously stated, we conducted multiple iterations of our second evaluation, varying the number of features (f) and topics (t). Specifically, we considered f values of 50 and 500 and t values of 15 and 25. Our initial findings on the performance and accuracy of topics are presented in Table 3, showcasing the application of common standard metrics relevant to Topic Modeling (TM) methods in the context of the 29-research topics.

Table 3: Performance of Involved Topic Modeling Methods with Different Extracted Topics *t = 15 and t =25*, (Average Value of Recall, Precision, And *F*-Score)

Topic Modeling Methods	Number of Topics					
	15			25		
	Recall	Precision	F- score	Recall	Precision	F- score
LDA	0.352425	0.462542	0.514256	0.438565	0.652487	0.754852
HDP	0.363524	0.452541	0.529868	0.458967	0.662542	0.762542
NMF	0.355124	0.456515	0.536582	0.462856	0.658957	0.759862
BERTopic	0.342563	0.435242	0.525259	0.459965	0.647213	0.745286
DTM	0.387542	0.542425	0.556856	0.478546	0.695241	0.775869

Our observation reveals that each Topic Modeling (TM) method employed in our investigation possesses distinct strengths and weaknesses. Throughout our comprehensive evaluation, we found that the outcomes of all the methods exhibited a comparable level of performance. DTM stands out by producing the highest term topic probability among all the models. However, it presented challenges as it included non-meaningful words, resembling domain-specific stop words, which hindered its suitability for further processing. Non-negative Matrix Factorization (NMF), Hierarchical Dirichlet Process (HDP), Latent Dirichlet Allocation (LDA) and BERTopic Models, all these models demonstrated similar levels of performance in terms of topic extraction outcomes. Recall/Precision/F-Score (R/P/F) Statistical Scores: Notably, the R/P/F statistical scores were comparatively lower for all the models, indicating areas for improvement in terms of precision, recall, and overall effectiveness in capturing relevant topics. Probabilities Range: Across all evaluated TM methods, probabilities ranged from 0 to 1, reflecting the confidence levels of the models in associating words with topics. LDA methods excelled in generating well-learned descriptive topics, showcasing a strength in capturing the semantic nuances of words in the corpus.

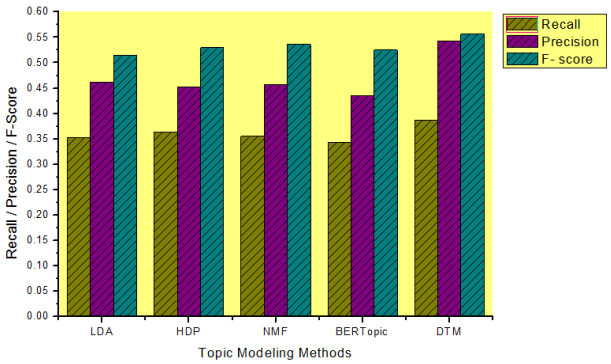

Figure 1: Performance of Involved Topic Modeling Methods with Different Extracted Topics *t = 15*, (Average Value of Recall, Precision And *F*-Score)

This was particularly evident when compared to certain methods, such as Latent Semantic Analysis (LSA), which struggled to create compact semantic representations of words. Additionally, in our detailed statistical measure results presented in Table 4. In this research work, DTM exhibited superior performance in comparison to other TMs with similar outcomes. These findings highlight the nuanced strengths and weaknesses of each TM method and underscore the importance of considering various metrics and evaluation criteria to comprehensively assess their performance in extracting meaningful topics from the dataset.

Recall, an important information retrieval metric, measures the proportion of relevant items (research topics) that were successfully identified by the model. In the case of DTM being the best result, a high recall implies that DTM excelled in capturing a significant portion of the actual research topics present in the dataset. Precision is a metric that gauges the accuracy of the identified topics by measuring the ratio of relevant items to the total items recommended by the model.

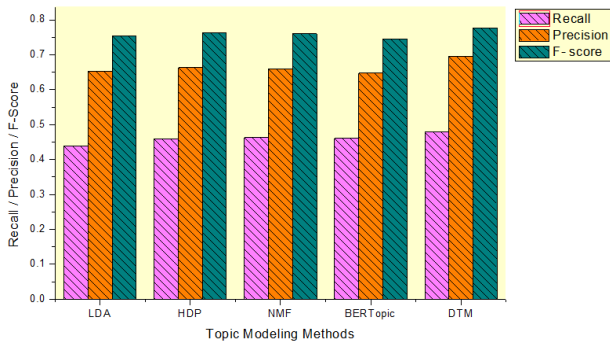

Figure 2: Performance of Involved Topic Modeling Methods with Different Extracted Topics *t = 25*, (Average Value of Recall, Precision, And *F*-Score)

In this experiment, DTM demonstrated high precision, indicating that a substantial portion of the identified topics were indeed relevant to the research context. The F-score, calculated as the harmonic mean of precision and recall, offers a balanced assessment of a model's overall effectiveness in topic identification. Given that DTM yielded the best results, its high F-score underscores the model's ability to achieve a harmonious balance between precision and recall, indicating robust performance in capturing relevant research topics. In summary, the evaluation of recall, precision, and F-score in our topic modeling experiment underscores (highlight) the superior performance of Dynamic Topic Modeling (DTM) in identifying and characterizing research topics within the dataset when compared to other models such as LDA, HDP, NMF, and BERTopic. These metrics collectively highlight the effectiveness and accuracy of DTM in the context of extracting meaningful and relevant research topics.

Coherence on a Research Paper dataset: Coherence is a common metric used to assess the interpretability and quality of topics generated by topic modeling algorithms. It measures the semantic similarity between high-scoring words within a topic, providing an indication of how well the topics capture meaningful associations. Here's a brief overview of how each model Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), Non-negative Matrix Factorization (NMF), BERTopic, and Dynamic Topic Modeling (DTM) might perform based on coherence on a research paper dataset.

1. ***Latent Dirichlet Allocation (LDA):*** LDA typically produces topics with moderate coherence. The choice of the number of topics K can significantly impact coherence. It's common to observe improved coherence with a moderate number of topics, as too few or too many can lead to less interpretable results.
2. ***Hierarchical Dirichlet Process (HDP):*** HDP often performs well in terms of coherence, as it automatically adapts the number of topics. Its hierarchical structure can lead to coherent subtopics. However, the interpretation might be challenging due to the hierarchical nature.
3. ***Non-negative Matrix Factorization (NMF):*** NMF tends to generate topics with high coherence. Its non-negativity constraints often result in more interpretable topics. The choice of the number of topics and other hyper parameters can influence coherence.
4. ***BERTopic:***BERTopic, leveraging BERT embeddings and clustering, tends to produce topics with high coherence. The semantic understanding provided by BERT embeddings contributes to meaningful topic representations.
5. ***Dynamic Topic Modeling (DTM):*** DTM's coherence can vary based on the evolution of topics over time. It may perform well in capturing temporal coherence, but this depends on how well the model adapts to changes in topics across different time periods.

In our experimental setup, we designate the default number of topics as K = 100. Specifically tailored parameter settings are applied to enhance the performance of each model. For Latent Dirichlet Allocation (LDA), we opt for $\alpha = 0.1$ and $\beta = 0.01$, leveraging a weak prior to yield improved results for short texts. HDP and NMF we adhere to default hyper parameter configurations. More explicitly, we define parameters $\alpha = 0.1$, $\lambda = 0.1$, and $\beta = 0.01$ for DTM, while setting $\beta = 0.1$ for BERTopic. In this case of LDA, HDP, NMF, BERTopic and DTM are executed for 1000 iterations. Moreover, we ensure result consistency and independence from random initial states by setting the seed for the random number generator to 0 for NMF.

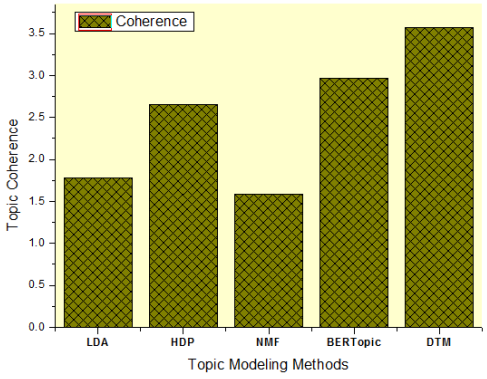

Figure 3: Topic Coherence Results with Research Article

Table 4: Topic Coherence Results with Research Article

Topic Modeling Methods	Coherence
LDA	1.7859
HDP	2.6523
NMF	1.5896
BERTopic	2.9656
DTM	3.5684

Analyzing the results presented in Table 4. and Figure 3, it becomes evident that one model exhibits superior performance provided especially Dynamic Topic Modeling (DTM), highlighting the effectiveness of DTM in extracting topics from short texts compared **to** other models. Notably, when contrasted with traditional methods such as Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), Non-negative Matrix Factorization (NMF), and BERTopic, our models demonstrate significant enhancements, indicating a more robust discovery of coherent topics. To delve deeper into the comparatively lower performance of NMF in all cases, we conduct a visualization of the top keywords within each topic. This examination reveals that several top keywords (e.g., 'computer,' 'algorithm,' and 'theory') exhibit semantic correlation but do not tend to co-occur in the same document. This semantic discordance may contribute to the suboptimal performance of NMF. Additionally, the complexities in capturing word semantic relationships within the context of Research Articles may play a role in the observed disparities.

CONCLUSION AND SUGGESTION

In conclusion, applying a suite of topic modeling algorithms, including Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), Non-negative Matrix Factorization (NMF), BERTopic, and Dynamic Topic Modeling (DTM) to a research paper dataset involves a multi-faceted approach to uncovering latent thematic structures. Each algorithm offers unique strengths and considerations, and their performance can be evaluated through a combination of quantitative metrics and qualitative assessments. Latent Dirichlet Allocation (LDA), a widely used probabilistic model, is effective in discovering topics with moderate coherence, with careful consideration of the number of topics playing a crucial role. Hierarchical Dirichlet Process (HDP) excels in adaptability to the number of topics and often generates coherent subtopics, though its hierarchical nature can pose interpretational challenges. Non-negative Matrix Factorization (NMF) tends to produce highly interpretable topics with high coherence, leveraging non-negativity constraints. BERTopic, incorporating BERT embeddings and clustering, offers high coherence and semantic understanding, enhancing the interpretability of topics. Dynamic Topic Modeling (DTM) designed for temporal analysis, captures the evolution of topics over time, providing insights into the changing thematic structuresin a dynamic research paper corpus. Coherence metrics serve as valuable quantitative indicators of the quality of topics generated by each model, but they should be complemented by domain expert validation and visual exploration of representative terms.

REFERENCES

[1]. Singhal, T., Liu, J., Blessing, L.T.M.,Lim, K.H.,"Analyzing scientific publications using domain-specific word embedding and topic modeling", IEEE International Conference on Big Data (Big Data),2021, pp. 4965–4973.

[2]. Alghamdi, R., and Alfalqi, K," A survey of topic modeling in text mining", Int. J. Adv. Comp. Sci. Appl. 6,2015, pp. 147–153.

[3]. Debortoli, S., Müller, O., Junglas, I., &vom Brocke, J. Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. Communications of the Association for Information Systems (CAIS), Vol. 39, No.7, 2016, pp.110-135.

[4]. Michal Rosen-Zvi, MarkSteyvers,"The Author-Topic Model for Authors and Documents", UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July,2004,

[5]. Anupriya P, Karpagavalli S,"LDA based topic modeling of journal abstracts", International Conference on Advanced Computing and Communication Systems. IEEE; 2015. pp. 1–5.

[6]. Anantharaman, A., Jadiya, A., Siri, C. T. S., Bharath Nvs, A., and Mohan, B. "Performance evaluation of topic modeling algorithms for text classification," in 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (Tirunelveli),2019

[7]. Neogi, P. P. G., Das, A. K., Goswami, S., and Mustafi,"Topic modeling for text classification in Emerging Technology in Modelling and Graphics", Advances in Intelligent Systems and Computing, Vol. 937, Singapore: Springer,2020,pp. 395–407.

[8]. Newman D, Noh Y, Talley E, Karimi S, Baldwin T,"Evaluating topic models for digital libraries", Proceedings of the 10th Annual joint conference on digital libraries, 2010. pp. 215–224.

[9]. David M. Blei, Andrew Y. Ng, and Michael I. Jordan,"Latent Dirichlet allocation". Journal of Machine Learning Research,2003, Vol.3, No.1, pp.993–1022.

[10]. E. Linstead, C. Lopes, P. Baldi," An application of latent Dirichlet allocation to analyzing software evolution ", Proceedings of the 7th International Conference on Machine Learning and Applications, ICMLA '08, 978-0-7695-3495-4, IEEE Computer Society, Washington, DC, USA (2008), pp. 813-818

[11]. Zoghbi, S., I. Vulic, and M.-F. Moens, "Latent Dirichlet allocation for linking user-generated content and e-commerce data". Information Sciences, 2016. 367: pp. 573-599.

[12]. John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan, "Nested hierarchical Dirichlet processes", IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(2):pp.256–270.

[13]. Y. Teh, M. Jordan, M. Beal, and D. Blei,"Hierarchical dirichlet processes. Journal of the American Statistical Association", 2006, 101(576): pp.1566-1581.

[14]. Jun Li, Jos´e M Bioucas-Dias, Antonio Plaza, and Lin Liu, "Robust collaborative nonnegative matrix

factorization for hyperspectral unmixing," IEEE Transactions on Geoscience and Remote Sensing,2016, Vol. 54, No. 10, pp. 6076–6090.

[15]. Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park, "Weakly supervised nonnegative matrix factorization for user-driven clustering. Data Mining and Knowledge Discovery 29, 6 (2015), pp.1598–1621.

[16]. M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, R.J. Plemmons, "Algorithms and Applications for Approximate Non-Negative Matrix Factorization", Comput. Stat. Data Anal. Vol.52. No.1. (2007), pp.155–173.

[17]. Grootendorst, M.R.,"BERTopic: Neural topic modeling with a class-based TF-IDF procedure". ArXiv, abs/2203.05794.2022

[18]. Xiaohui Yan, Jiafeng Guo, Yanyan Lan, XueqiCheng,"Abiterm topic model for short texts", Proceedings of the 22nd international conference on World Wide Web,May 2013,pp. 1445–1456

[19]. D. M. Blei and J. D. Lafferty," Dynamic topic models", Proceedings of the 23rd International Conference on Machine Learning,2006,pp.113–120.

[20]. Ren, L., Dunson, D. B., & Carin, L. ,"The dynamic hierarchical Dirichlet process", Proceedings of the 25th International Conference on Machine Learning, 2008,pp. 824–831.

[21]. J. F. Canny and T. L. Rattenbury, "A Dynamic Topic Model for Document Segmentation," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2006-161, Dec. 2006.