Sorbonne Université

Ecole Doctorale L'EDITE de Paris
Laboratoire d'Informatique de Paris 6 (LIP6)

Thèse de doctorat

Discipline Informatique

Soutenue le 7 November 2024

Par

Hamed Rahimi

Modeling Topic Evolution in Scientific Archives with Deep Learning

Zoltan Miklos Professeur, Université de Rennes	(Rapporteur)
Lynda Tamine-Lechani Professeure, Université Toulouse III	(Rapportrice)
Sylvain Lamprier Professeur, Université d'Angers	(Examinateur)
Benjamin Piwowarski Directeur de Recherche CNRS, Sorbonne Université	(Examinateur)
Bernd Amann Professeur, Sorbonne Université	(Directeur de thèse)
Hubert Naacke Professeur, Sorbonne Université	(Encadrant)

“dédié à mon père.”

DECLARATION

I hereby declare that the work in this dissertation entitled *"Modeling Topic Evolution in Scientific Archives with Deep Learning"* is my own original work and has not been submitted for another examination or assignment, either wholly or excerpts thereof. Furthermore, I confirm that I have acknowledged the work of others by providing detailed references of said work.

Hamed Rahimi
August 22nd, 2024
Sorbonne University, Paris

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my family. To my father, who encouraged me the most to pursue a PhD, and my mother, whose prayers were a constant source of comfort, and whose voice was always reassuring. I also want to thank my grandmother and grandfather, who were always rooting for me, and my uncles, who supported me throughout this journey. A special thanks goes to my brother, who was always next to me, offering unwavering support and companionship.

I am profoundly grateful to my advisors, Professor Bernd Amann and Professor Hubert Naacke, for their invaluable guidance, advice, and patience. They were always there to help me with my problems and guide me through the complexities of my research.

I would like to extend my appreciation to my colleagues, Dr. Camelia Constantin and Dr. David Mimno, for their collaboration and insightful discussions throughout this work.

I am deeply thankful to the Sorbonne Center for Artificial Intelligence (SCAI) for funding this project. I extend my sincere appreciation to Dr. Xavier Fresquet, Dr. Gérard Biau, Nora Roger, and Rakhee Patel for their invaluable support.

Finally, I would like to acknowledge the support and camaraderie of my team members: Sara, Yuhe, Stephan, Amin, Rafael, Anis, Lyes, and Marie-Veronique. Your dedication and teamwork have been a significant part of this accomplishment.

Thank you all for your unwavering support and encouragement.

Abstract

Topic evolution analysis in scientific archives is a critical domain of research that focuses on understanding how scientific topics evolve, emerge, or change over time. This thesis aims to take advantage of advanced deep learning techniques to analyze large collections of time-stamped scientific documents. By tracking the temporal progression of semantically similar documents and their citation relationship, we contribute to this line of research by identifying evolving topics, detecting emerging areas of interest, and observing shifts in scientific paradigms. Our objectives include establishing new baselines for fundamental concepts like "topic" and "evolving topics," developing novel methods for detecting emerging topics and paradigm shifts, and creating improved evaluation metrics for topic models. By addressing challenges such as the difficulty in defining complex notions, lack of standard categorization, and absence of reliable evaluation metrics, this research seeks to push the boundaries of scientific discovery and innovation, providing more nuanced and accurate insights into the evolution of scientific knowledge. This analysis provides valuable insights into the dynamic nature of scientific knowledge, revealing patterns such as emerging topics, integration of insights from different fields, and the rise or decline of specific research areas. The experiments conducted throughout this research have led to the development of several software tools, each designed to address specific challenges in topic modeling and topic evolution analysis. The Contextualized Topic Coherence (CTC) metrics, implemented as a Python package, provide advanced methods for evaluating the interpretability of topic models by leveraging contextual embeddings. The Aligned Neural Topic Models (ANTM) software facilitates the exploration of evolving topics in large-scale text archives, integrating temporal information for dynamic analysis. Additionally, the Automated Topic Emergence Monitoring (ATEM) framework and the QuTE model, both implemented as Python-based tools, support the early detection of emerging topics and the modeling of paradigm shifts, respectively.

Résumé

L'analyse de l'évolution des sujets dans les archives scientifiques est un domaine de recherche crucial qui se concentre sur la compréhension de l'évolution, de l'émergence ou du changement des sujets scientifiques au fil du temps. Cette thèse vise à tirer parti des techniques avancées d'apprentissage profond pour analyser de grandes collections de documents scientifiques horodatés. En suivant la progression temporelle des documents sémantiquement similaires et de leurs relations de citation, nous contribuons à cette ligne de recherche pour identifier les sujets en évolution, détecter les nouvelles zones d'intérêt et observer les changements de paradigmes scientifiques. Nos objectifs incluent l'établissement de nouvelles références pour des concepts fondamentaux comme "sujet" et "sujets en évolution", le développement de nouvelles méthodes pour détecter les sujets émergents et les changements de paradigmes, et la création de métriques d'évaluation améliorées pour les modèles de sujets. En relevant des défis tels que la difficulté de définir des notions complexes, le manque de catégorisation standard et l'absence de métriques d'évaluation fiables, cette recherche cherche à repousser les limites de la découverte et de l'innovation scientifiques, en fournissant des insights plus nuancés et précis sur l'évolution des connaissances scientifiques. Cette analyse offre des insights précieux sur la nature dynamique des connaissances scientifiques, révélant des schémas tels que les sujets émergents, l'intégration des insights provenant de différents domaines et l'ascension ou le déclin de zones de recherche spécifiques. Les expériences menées au cours de cette recherche ont conduit au développement de plusieurs outils logiciels, chacun conçu pour relever des défis spécifiques en matière de modélisation des sujets et d'analyse de l'évolution des sujets. Les métriques de Contextualized Topic Coherence (CTC), implémentées sous forme de package Python, offrent des méthodes avancées pour évaluer l'interprétabilité des modèles de sujets en utilisant des embeddings contextuels. Le logiciel Aligned Neural Topic Models (ANTM) facilite l'exploration des sujets en évolution dans les archives textuelles à grande échelle, en intégrant des informations temporelles pour une analyse dynamique. De plus, le cadre Automated Topic Emergence Monitoring (ATEM) et le modèle QuTE, tous deux implémentés sous forme d'outils basés sur Python, soutiennent respectivement la détection précoce des sujets émergents et la modélisation des changements de paradigmes.

Contents

List of Figures

List of Tables

Chapter

1 Introduction

The evolution of science is a dynamic process marked by the continuous emergence and transformation of ideas and discoveries [1]. Studying science evolution provides valuable insights for researchers to understand how scientific domains evolve. These insights have the potential to transform the research landscape by promoting up-to-date knowledge and innovative research. Additionally, understanding science evolution holds significant implications for research funding and public policy decisions in both academic and industrial contextual settings [2].

Within this broader context, topic modeling plays a crucial role [3, 4], as highlighted in numerous surveys [5]. Topic modeling refers to a statistical technique used to uncover abstract themes from a corpus of text documents, such as scientific archives. Specifically, dynamic topic models [6] enable the study of the evolution of scientific research topics over time.

The primary motivation behind this thesis is to leverage advanced topic modeling and large language models to uncover complex notions of topic evolution and deepen our understanding of the dynamics that shape scientific knowledge. These powerful methods offer unprecedented capabilities for analyzing vast amounts of data, allowing us to gain insights that push the boundaries of scientific discovery and innovation.

Chapter content

1.1 Motivation

Our understanding of the world and scientific knowledge is constantly evolving over time, and the evolution of science has been studied by philosophers for centuries. For instance, Popper [7] considers science evolution as a Bayesian inference process that updates the logical possibility of falsification, whereas Kuhn [1] introduces the notion of paradigm shift, where science evolution is a Darwinian selection process of theories. Paradigm shifts represent fundamental changes in the underlying assumptions and methodologies of a particular field, driving the evolution of topics and fostering the emergence of new research areas within scientific archives. Similarly, the concept of emergence dates back at least to the time of Aristotle (Metaphysics) [8]: *"... the totality is*

not, as it were, a mere heap, but the whole is something besides the parts ...", i.e., the whole is other than the sum of the parts. In science, the emergence of new research topics frequently occurs through the subtle convergence of distinct research areas, resulting in the formation of interdisciplinary fields with blurred boundaries. These phenomena are inherently part of social science, as they reflect the collective efforts and interactions of researchers.

These philosophical theories attempt to explain the evolution of scientific domains and describe their relations and interactions with different semantics. The ability to accurately categorize and quantify these complex notions is crucial for understanding the trajectory of scientific advancement, predicting future research trends, and informing science policy decisions. However, these theories are limited in certain respects, highlighting the need for more comprehensive analytical approaches.

The abstract nature of these philosophical models can make them difficult to empirically validate with analytical methods on concrete data from scientific research and publications. Capturing these notions within scientific archives involves leveraging advanced data mining techniques, network analysis, and machine learning algorithms to detect evolution patterns within the contents and connections. Additionally, incorporating expert knowledge and employing multi-scale temporal analysis can help in distinguishing significant shifts from normal scientific progress.

More recently, to capture and analyze these changes, researchers have turned to the analysis of large scientific archives with advanced computational techniques. One widely used method for analyzing how scientific knowledge evolves over time is Dynamic Topic Modeling [9]. This approach, an extension of traditional topic models, identifies and tracks the progression of topics over time by examining semantically similar documents within temporal archives. By capturing these temporal dynamics, science evolution is mainly studied by topic evolution analysis [10] which involves modeling and discovering how topics change, emerge, or fade over time from time-stamped document collections [11].

1.2 Main Challenges

The dynamic nature of scientific knowledge, coupled with the vast and ever-growing corpus of research, creates significant obstacles in accurately mapping and understanding the evolution of science. This thesis addresses two main challenges in topic-based science evolution analysis as follows.

Modeling Science Evolution A key challenge in analyzing the evolution of science within scientific archives is developing effective models that can connect the information contained in these archives to complex concepts such as research trend analysis, scientific emergence, and paradigm shifts. These models must be capable of accurately representing the research activities within scientific communities and domains, capturing the various research topics across different time periods, and tracking their evolution, including semantic relationships like shared concepts and citations. A crucial aspect of this challenge involves ensuring the quality of this representation, including its precision, diversity, coherence, and completeness. Advanced techniques in text mining and semantic analysis, such as dynamic topic models and citation network analysis, are essential for overcoming this challenge.

Analysing Science Evolution The second challenge involves developing and implementing new methods to analyze and empirically validate philosophical concepts related to science evolution, such as topic emergence [12] and paradigm shifts [13]. Recent advancements in deep

learning, particularly in graph neural networks [14] and Large Language Models (LLM) [15], offer unprecedented opportunities to model or redefine these complex concepts in the context of science evolution analysis. These advanced techniques can capture intricate patterns and relationships within large datasets, providing a more nuanced understanding of how scientific knowledge evolves and how new ideas emerge. The goal of this thesis is to leverage these advancements and establish a baseline for these complex notions in the analysis of science evolution.

1.3 Objectives and Contributions

1.3.1 Formal Framework For Science Evolution Analysis.

The first objective of this thesis is to establish a formal framework for science evolution analysis. This framework involves updating and refining concepts within topic modeling by introducing new notions such as "evolving topic" and "contextualized topic coherence" for analyzing science evolution. By incorporating recent advancements in deep learning, we aim to provide more accurate and nuanced definitions that better capture the complex dynamics of scientific knowledge evolution. This framework will serve as a robust foundation for further analysis and the development of innovative methodologies in the field.

Contributions To achieve this, we propose ANTM, a dynamic topic model that leverages Large Language Models (LLMs) and advanced clustering techniques to capture the temporal dynamics of topics, providing insights into how they evolve and develop over time.

As a second contribution, we introduce CTC, a novel family of topic coherence metrics that leverage LLMs for the more accurate evaluation of topic models. This approach improves our ability to identify the best-performing models for specific scientific corpora and guides fine-tuning for better results, potentially transforming topic modeling in scientific literature analysis.

1.3.2 Science Evolution Analysis.

The second objective of this thesis is to identify and analyze complex concepts in science evolution analysis, such as topic emergence and paradigm shifts. This analysis consists of extracting evolution patterns from complex temporal and semantic representations of topics and their evolution, including the contents and the citation relationships.

Contributions To discover and analyze emerging topics, this thesis proposes ATEM, a novel framework based on analyzing citation activity across evolving topics using dynamic graph embedding techniques to represent dynamic citation contexts. By examining how documents from different topics cite each other over time, ATEM can identify emerging topics that represent new and increasingly prominent ideas or issues within specific fields or broader areas of interest.

Furthermore, we introduce QuTE, an innovative framework for modeling paradigm shifts in scientific evolution. QuTE leverages Reinforcement Learning (RL) techniques, combining ATEM with Q-Learning, to uncover the intricate dynamics of scientific evolution. This approach aims to provide a deeper understanding of how paradigm shifts occur and evolve over time in scientific research.

1.4 Main contributions

CTC In Chapter 4, we introduce the Contextualized Topic Coherence (CTC) metrics, a set of innovative evaluation metrics designed to assess the interpretability of topic models. These metrics leverage contextual embeddings from state-of-the-art language models like BERT to assess the semantic relatedness of words within a topic, providing a more nuanced and accurate measure than traditional topic coherence metrics. We also benchmark various topic models, including Latent Dirichlet Allocation (LDA) and Neural Topic Models, against standard datasets like the 20 Newsgroups and Wikipedia, demonstrating the superior performance of our CTC metrics in capturing the semantic coherence of topics across different archives and models. As noted in Section 1.7, these metrics are implemented as a Python Package and are available to the research community.

ANTM In Chapter 6, we address the challenge of identifying evolving topics in large-scale text archives by proposing a clustering-based dynamic topic model called Aligned Neural Topic Models (ANTM). This model integrates temporal information into the topic modeling process, allowing for the detection and tracking of topic evolution over time. By utilizing advanced clustering algorithms and LLMs, ANTM can dynamically adapt to the emergence of new topics and the decline of old ones, offering a more realistic and accurate depiction of topic dynamics. As mentioned in Section 1.7, this model is implemented as a Python Package and is available to the research community.

ATEM Chapter 9 presents the Automated Topic Emergence Monitoring (ATEM) framework, which integrates dynamic graph embedding methods with LLMs to enable the early identification of emerging topics. This model leverages the structural information of citation networks and the semantic depth of language models to detect nascent research trends at their inception. As mentioned in Section 1.7, this model is implemented as a Python-based framework and is open-sourced for the research community.

QuTE Chapter 10 defines the concept of a paradigm shift in scientific domains and introduces QuTE, a reinforcement learning-based framework that models topic evolution using Q-Learning. QuTE quantifies and predicts the transition probabilities between evolving topics during paradigm shift cycles, capturing the interplay between established paradigms and emerging ideas. By integrating ATEM, QuTE efficiently explores the vast state space of scientific topics, with Q-Learning enabling adaptive learning from observed patterns in the literature.

1.5 Thesis Structure

This thesis is structured into three parts, each focusing on different contributions to the field of topic modeling and science evolution analysis.

Part I establishes a foundational understanding of baseline and state-of-the-art topic models and topic modeling concepts, as discussed in Chapter 2. It further delves into topic evaluation metrics in Chapter 3. Our contribution to this part is detailed in Chapter 4, where we introduce the Contextualized Topic Coherence (CTC) metric. Part II focuses on dynamic topic models, which are crucial for studying topic evolution. Chapter 5 reviews baseline models for dynamic topic modeling, setting the stage for our contribution in Chapter 6, where we introduce the Aligned Neural Topic Model (ANTM) that optimizes the discovery and analysis of evolving topics over time. Part III explores topic evolution analysis in scientific archives, covered in Chapters 7 and 8. Our contributions in this part include the introduction of ATEM in Chapter 9 and

the QuTE model in Chapter 10. And finally, we conclude the thesis in Chapter 11 with some perspectives for future work.

1.6 List of Publications

- Hamed Rahimi, David Mimno, Jacob Hoover, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Contextualized Topic Coherence Metrics. In Findings of the Association for Computational Linguistics: EACL 2024, pages 1760–1773, St. Julian's, Malta. Association for Computational Linguistics.
- Rahimi, H., Naacke, H., Constantin, C., Amann, B. (2024). ANTM: Aligned Neural Topic Models for Exploring Evolving Topics. In: Hameurlain, A., Tjoa, A.M., Akbarinia, R., Bonifati, A. (eds) Transactions on Large-Scale Data- and Knowledge-Centered Systems LVI. Lecture Notes in Computer Science, vol 14790. Springer, Berlin, Heidelberg.
- Rahimi, H., Naacke, H., Constantin, C., Amann, B. (2024). ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives. In: Cherifi, H., Rocha, L.M., Cherifi, C., Donduran, M. (eds) Complex Networks & Their Applications XII. COMPLEX NETWORKS 2023. Studies in Computational Intelligence, vol 1143. Springer, Cham.

ANTM and ATEM were presented at BDA 2023, and the demo software of ATEM has been accepted for presentation at BDA 2024. Additionally, the work presented in Chapter 10 is being prepared for submission.

1.7 List of Software

- Rahimi, H., Naacke, H., Constantin, C., Amann, B. (2023). ANTM: An aligned neural topic model for exploring evolving topics [Computer software]. GitHub.
- Rahimi, H., Hoover, J. L., Mimno, D., Naacke, H., Constantin, C., Amann, B. (2023). Contextualized topic coherence metrics [Computer software]. GitHub.
- Rahimi, H., Amann, B., Naacke, H., & Constantin, C. (2022). ATEM: A topic evolution model for the detection of emerging topics in scientific archives [Computer software]. GitLab.

Part I

Topic Modeling

> *Individuals who break through by inventing a new paradigm are almost always either very young men or very new to the field whose paradigm they change. These are the men who, being little committed by prior practice to the traditional rules of normal science, are particularly likely to see that those rules no longer define a playable game and conceive another set that can replace them.*
>
> Thomas S. Kuhn

Motivation

Understanding topic models is crucial for analyzing the evolution of science, as they offer powerful tools for processing and interpreting large volumes of scientific literature. These models reveal underlying themes, trends, and shifts in research priorities over time, helping to map the structure and interconnectedness of scientific fields. By quantifying the prominence and progression of various topics, topic models provide a data-driven perspective on scientific advancements and the diffusion of ideas.

Organization

In this part of the manuscript, we begin by defining the concepts of topics and topic modeling to establish a clear foundational understanding of these concepts in Chapter 2. In this chapter, we categorize the various types of topic models and provide a structured and detailed overview of the field. This categorization encompasses a range of models, from classical approaches to the latest advancements, highlighting the diversity and evolution of topic modeling techniques. We then delve into the methods used for evaluating topic models and critically assess existing evaluation metrics in Chapter 7. This analysis is crucial as it sheds light on the strengths and limitations of current evaluation methods, offering insights into their efficacy and areas where they may fall short.

Contributions

We contribute to this part of the manuscript by introducing a novel family of topic coherence metrics in Chapter 4, which utilize pre-trained Large Language Models (LLMs) to provide a more nuanced understanding of language and context for assessing topic coherence.

This contribution is the result of a collaboration with David Mimno (Cornell University) and Jacob Hoover (McGill University) and has been published in *Hamed Rahimi, David Mimno, Jacob Hoover, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Contextualized Topic Coherence Metrics. In Findings of the Association for Computational Linguistics: EACL 2024, pages 1760–1773, St. Julian's, Malta.*

Chapter

2 Topic Models

In today's era of big data, the ability to comprehend and interpret vast amounts of information is essential for understanding the evolution of science. This capability enables researchers to uncover hidden patterns, identify emerging trends, and gain insights that drive scientific discovery and innovation. By analyzing large datasets effectively, the scientific community can map the progression of knowledge, track the diffusion of ideas, and make informed decisions that shape the future of research and technological advancement. Topic modeling provides a powerful method for extracting critical information from large volumes of unstructured textual data, such as scientific archives and social media posts.

Chapter content

2.1 Introduction

Topic modeling is a statistical technique that discovers abstract themes from a corpus of text documents [4, 5]. Each topic represents a meaningful concept that can be understood through a group of related words. For instance, a topic focused on *transportation* might be characterized by words like *car*, *bicycle*, and *airplane*. We can define the topics of a corpus in a nutshell as follows.

Definition 2.1.1 (Topic)**.** A *topic* $t \in T$ is a pair (D_t, W_t) where $D_t \subseteq D$ represents a subset of *semantically similar* documents from the corpus D, and $W_t \subseteq V$ is a weighted list of terms from the vocabulary V of the corpus, which serves as the representation of topic t. D_t and W_t are called, respectively, the *topic cluster* and the *topic representation* of topic t.

Figure 2.1 represents topics extracted from a corpus of documents using BERTopic [16] that will be explained in Section 2.3.

This definition is applicable to the topics generated by any topic model extracted from a document archive, regardless of the specific algorithm or approach used. It serves as a universal framework for understanding how topic models function and what outputs they produce.

Definition 2.1.2 (Topic Model)**.** A *topic model* is a computational method that identifies a set of topics T within a corpus D. Each topic $t \in T$ is characterized by two key components:

- Topic-Document Distribution: A probability distribution $P(d|t)$ over documents $d \in D$, given a topic t. This distribution assigns probabilities to each document in the corpus D, indicating its relevance or membership probability in the *cluster* of topic t.

Figure 2.1: *Topics and Topic Representations [17]*

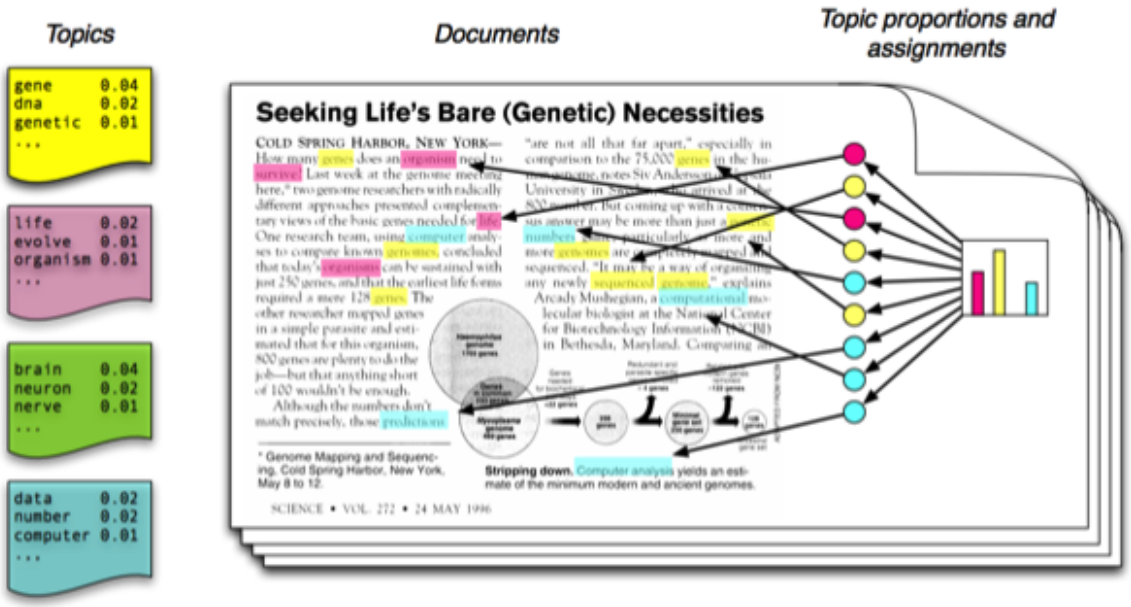


Figure 2.2: *Components of Topic Models [4]*

- Topic-Word Distribution: A probability distribution $P(w|t)$ over words $w \in$ V from the vocabulary V of the corpus, given a topic t. This distribution assigns probabilities to each word in the vocabulary, reflecting its importance or contribution to the *representation* of topic t.

As illustrated in Figure 2.2, topic modeling provides a structured way to discover and describe the underlying thematic structure of a document corpus through these two components. Based on the characteristics of the methodologies used to find these components [18], topic models are widely categorized into Probabilistic Topic Models (PTM) [19, 20] and Neural Topic Models (NTM) [16, 21–24]. PTMs are generative methods that use statistical inference to uncover latent topic structures in document collections, simultaneously deriving topic-document distributions $P(d|t)$ and topic-word distributions $P(w|t)$ from observed word frequencies. On the other hand, NTMs take advantage of numerical optimization techniques and exploit recent neural network approaches to represent topics as weighted word-vectors extracted from a set of semantically similar documents. In the following sections, we will provide a detailed explanation of these methods and introduce the most significant ones. We will also discuss their applications and implications, highlighting their relevance and importance in the context of scientific research.

2.2 Probabilistic Topic Models

Conventional approaches to topic modeling, which embrace probabilistic graphical models such as Latent Dirichlet Allocation (LDA) [19], have been extensively explored over the past two decades. LDA is a popular generative probabilistic model for collections of discrete data such as text corpora. It has found widespread use in natural language processing and information retrieval.

As shown in Figure 2.3, LDA identifies word-topic distribution and document-topic distribution through a Bayesian inference process. Consider a corpus of D of M documents over a fixed vocabulary containing V of N distinct terms. Let $w_{dn} \in \{1, \ldots, V\}$ denote the n^{th} word in the d^{th} document. LDA models D through K topics, denoted as $\beta_{1:K}$, where each topic β_k is a probability distribution over the vocabulary. For each document d, LDA defines a vector of topic proportions, θ_d, where each component θ_{dk} indicates how prevalent the k^{th} topic is in that

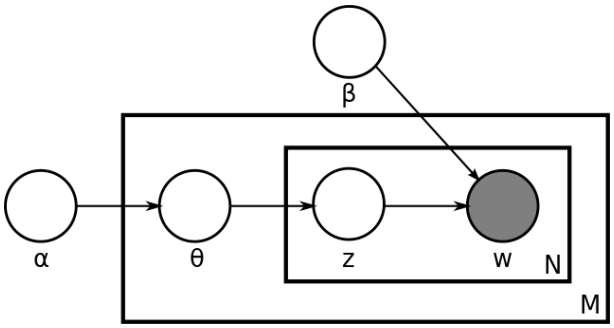


Figure 2.3: *LDA Graphical Model [3]*

document. In the generative process of LDA, each word in a document is assigned to a topic k with probability θ_{dk}, and the word itself is then drawn from the distribution β_k associated with that topic. The generative process for each document is as follows:

- 1. Draw topic proportions θ_d from a Dirichlet distribution.

$$\theta_d \sim \text{Dirichlet}(\alpha_\theta) \tag{2.1}$$

- 2. For each word n in the document:
  - (a) Draw topic assignment z_{dn} from the categorical distribution over topics.

$$z_{dn} \sim \text{Cat}(\theta_d) \tag{2.2}$$

  - (b) Draw the word w_{dn} from the categorical distribution over the vocabulary.

$$w_{dn} \sim \text{Cat}(\beta_{z_{dn}}) \tag{2.3}$$

Here, $\text{Cat}(\cdot)$ denotes the categorical distribution. LDA also imposes a Dirichlet prior on the topics, $\beta_k \sim \text{Dirichlet}(\alpha_\beta)$. The concentration parameters α_β and α_θ of the Dirichlet distributions are model hyperparameters.

LDA has established itself as a leading method in topic modeling, and it has served as the foundation for several advanced techniques that build upon its principles. For instance, [25] propose a generative model based on LDA that mines distinct topics in document collections by integrating the temporal ordering of documents into the generative process. This integration improves identifying the best terms for representing each topic. However, LDA also faces fundamental limitations. First, its static nature does not account for temporal information, such as publication dates, reducing its ability to analyze topic evolution. Second, its bag-of-words representation ignores the contextual semantics of words, which diminishes the precision and interpretability of the generated topic representations. Finally, its implementation requires tuning the optimal number of topics and involves iterative computation, which limits its applicability to large document archives. These limitations hinder LDA's capacity to accurately track and understand how scientific fields develop and change over time. Researchers have proposed various model structures derived from LDA, such as supervised LDA [26] or correlated LDA [27] to address some of these issues such as scalability in handling large datasets. These models fundamentally infer model parameters through Variational Inference [28] or Monte Carlo Markov Chain (MCMC) methods such as Gibbs LDA [22].

Despite the achievements of these probabilistic methods, two notable limitations reduce the scalability of probabilistic methods:

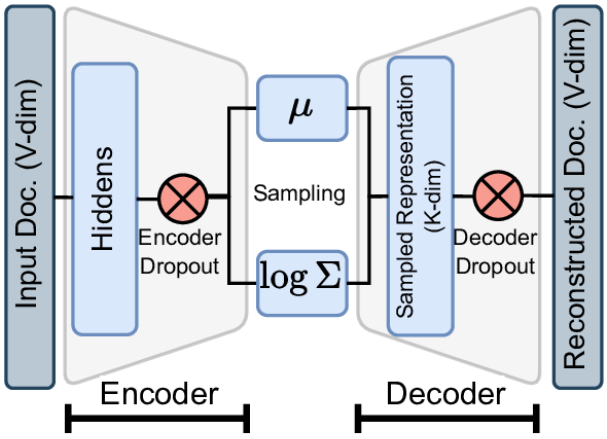


Figure 2.4: *The architecture of a VAE-based NTM [37]*

1. Computation-intensive parameter inference: Probabilistic methods require intricate, model-specific derivations for parameter inference, and as the model's complexity grows, so does the complexity of the inference process. This escalating complexity hinders their capacity to adapt to a wide range of model structures and application scenarios [29].

2. Limited data-parallelism: Probabilistic inference algorithms generally do not naturally support data parallelism for handling large document archives in a scalable way. Although some parallel inference methods have been developed [30–32], these approaches are often tailored to specific models and cannot be easily applied to different model structures or various application contexts.

2.3 Neural Topic Models

Neural Topic Models (NTMs) employ neural networks to discover and represent topics in a collection of documents. NTMs infer model parameters through automatic gradient back-propagation by using deep neural networks to model latent topics [33]. Based on the types of neural networks utilized in the structure of topic models, NTMs can be categorized into the following classes [33].

VAE-based NTMs These models leverage the Variational Autoencoder (VAE) framework [34, 35], which consists of an encoder inference network and a decoder generation network to derive topic-word and topic-document distributions, thereby exposing the underlying thematic structure within a collection of documents. shown in Figure 2.4. ProdLDA [29] is the most popular model of this category [36], in which the encoder infers document topic distributions from the documents, while the decoder generates documents from these inferred topic distributions.

LDA-based NTMs These models combine the generative process of LDA with vector representations of words [38, 39]. LDA-based NTMs define a topic using the embedding component $\beta = W^\top \mathbf{T}$, where $W \in \mathbb{R}^{D \times V}$ denotes the word embeddings, and $T \in \mathbb{R}^{D \times K}$ represents the topic embeddings, with D being the dimension of the embedding space [40].

The Embedded Topic Model (ETM) [41] is the most important model of this category that improves the performance of topic models in terms of quality and predictive accuracy, especially in the presence of large vocabularies. ETM leverages pre-trained word embeddings, such as

Word2Vec [42] or GloVe [43], to initialize W. The notations and mathematical representation of the generative process in ETM can be outlined as follows. Let ρ be an $L \times V$ matrix containing L-dimensional embeddings of the words in the vocabulary, with each column $\rho_v \in \mathbb{R}^L$ representing the embedding of the v^{th} term. ETM uses the embedding matrix ρ to define each topic β_k; specifically, it sets $\beta_k = \text{softmax}(\rho^\top \alpha_k)$, where $\alpha_k \in \mathbb{R}^L$ is the embedding representation of the k^{th} topic, known as the topic embedding. By this, topic embeddings provide a distributed representation of the topics in the semantic space of word embeddings. Finally, ETM incorporates the obtained topic embeddings in its generative process, which is analogous to that of LDA:

- 1. Draw topic proportions $\theta_d \sim \mathcal{LN}(O, \text{I})$.
- 2. For each word n in the document:
  - (a). Draw topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$.
  - (b). Draw word $w_{dn} \sim \text{Cat}(\text{softmax}(\rho^\top \alpha_{z_{dn}}))$

The notation $\mathcal{LN}$ in Step 1 refers to the logistic-normal distribution [44], which transforms Gaussian random variables to the simplex (the space where the components are non-negative and sum to one). In using the word representations $\rho_{1:V}$ in the definition of $\beta_{1:K}$, ETM learns the topics of a corpus in a particular embedding space. The intuition behind ETM is that semantically related words will be assigned to similar topics (since their embedding representations are close, they will interact similarly with the topic embeddings $\alpha_{1:K}$).

Clustering-based NTMs Clustering-based topic models leverage word and document embedding techniques and combine them with clustering algorithms to derive topic-document and topic-word distributions based on document clusters [45].

Top2Vec [24] is the most important model of this category. A key advantage of Top2Vec is that it does not require preprocessing steps such as stopword removal, stemming, or lemmatization. Secondly, it autonomously determines the number of topics, their size, and the words representing each topic. As shown in Figure 2.5, Doc2Vec [46] is utilized to generate word and document embeddings. Given the sparsity of the vector space, dimension reduction is performed using uniform manifold approximation and projection (UMAP) [47] before applying hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [48] to identify dense clusters of documents. The centroid of document vectors in their original dimensions is then computed for each dense region, defining the topic vector. The proximity of word vectors to document vectors indicates the best descriptive words for a document's topic, and the clustering of documents signifies the number of topics.

LLM-based NTMs LLM-based topic models leverage transformer-based Large Language Models (LLMs) [50], which are pre-trained on large-scale archives to capture contextual linguistic features and to generate semantically rich and coherent topic models.

Several approaches combine Variational Autoencoders (VAEs), Large Language Models (LLMs), and the Bag-of-Words (BoW) representation where a document is represented by a vector of word frequencies, disregarding grammar and word order. For instance, the Contextualized Topic Model (CTM) [51] integrates contextual document embeddings generated by Sentence-BERT [52] with BoW features through concatenation to reconstruct the original BoW. Similarly, [53] generates pseudo BoW representations using predictive word probabilities from BERT [15]. These pseudo BoW representations are then exploited alongside the actual BoW to reconstruct both the real and pseudo BoW, facilitating knowledge transfer from BERT to Neural Topic Models (NTMs).

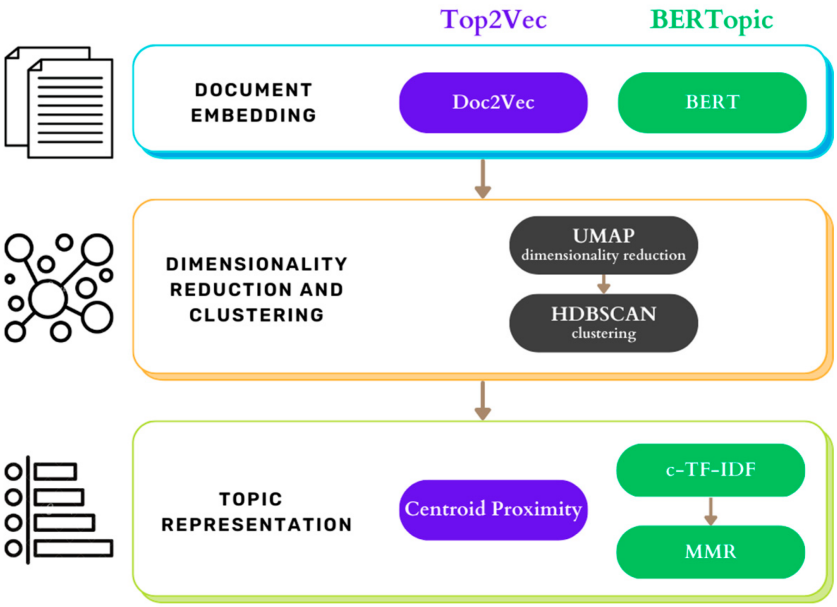


Figure 2.5: *Architecture of Top2Vec and BERTopic [49]*

Another important LLM-based topic modeling method is BERTopic [54], which leverages encoder-only transformers to encode documents into vector representations. BERTopic follows the clustering-based architecture of Top2Vec [24] and employs advanced techniques to derive topic-word distributions and topic-document distributions. As depicted on Figure 2.5, BERTopic first encodes documents into high dimensional vector representations (encoding) using BERT [15] and then reduces the dimension size by UMAP [47]. It then applies HDBSCAN [48] to extract topic-document clusters and calculates the term vector representation (topic-word distribution) via c-TF-IDF [55].

There are also LLM-based topic models that use decoder-only transformers, such as ChatGPT [56], employing a prompt-based framework to uncover latent topics in text collections. One notable example is TopicGPT [57], which provides natural language labels and free-form descriptions for topics, enhancing interpretability and aligning more closely with human categorizations.

RL-based NTMs Reinforcement Learning (RL) is a type of machine learning where an agent learns to make decisions by performing actions in an environment to maximize cumulative rewards through trial and error. Various works take advantage of RL to facilitate the learning process of NTMs. Most of these models use topic evaluation metrics to optimize reward functions that help them to guide the learning of topic modeling [58]. For instance, [59] use sentence embeddings and add a weighting term to the ELBO to track topic diversity and coherence during training.

GNN-based NTMs Graph Neural Networks (GNNs) are neural networks designed to process and analyze data structured as graphs, leveraging the relationships between nodes and edges to learn representations and perform tasks like node classification, link prediction, and graph classification. To construct a document model, several NTMs follow the VAE framework and use GNN in addition to the traditional BoW. For instance, [60–62] transforms documents into biterm graphs connected by word occurrences or TF-IDF values and [63, 64] applies graph topic modeling based on word co-occurrence and semantic word correlation in social media applications.

GAN-based NTMs Generative Adversarial Networks (GANs) are a class of machine learning models where two neural networks, a generator, and a discriminator, are pitted against each other

to generate new, synthetic data that mimics a given dataset. Several studies have explored the use of GANs for topic modeling. [65] proposes the Adversarial-neural Topic Model (ATM) where a generator creates "fake" documents from a random Dirichlet sample, while a discriminator distinguishes these generated documents from real ones. However, ATM cannot infer document-topic distributions, as it directly maps documents to TF-IDF-based word vector representations. To overcome this limitation, [66] introduces bidirectional adversarial training, which allows for the inference of document-topic distributions. Building on this, [67] presents an extension that employs two cycle-consistency constraints to produce more informative representations.

2.4 Applications

Topic models are widely used in exploratory data analysis for organizing, understanding, and summarizing large amounts of text data [68]. In this section, we describe some examples of the usage of topic models for content analysis, content generation, and content recommendation [69].

Content Analysis Content analysis is the most popular application for topic modeling and covers different subtopics such as content classification and sentiment analysis in social media. The content classification TASK includes the automatic categorization of text documents [22], images, or video segments [5] into predefined classes. For example, [70] integrates an NTM with a memory network for enabling simultaneous document classification and topic discovery in short texts through joint training. In a related advancement, [71] introduces a method combining BERT with NTMs to optimize the classification process by minimizing the use of self-attention mechanisms. In sentiment analysis tasks, topic modeling can be utilized to determine the sentiment expressed by user messages within social media applications [72]. [73] propose a multitask mutual learning framework that integrates sentiment analysis with topic detection by aligning topic-word distributions with word-level attention vectors through a process of mutual learning. [74] applies NTMs to analyze discourse in micro-blog conversations, providing insights into public discussions on health-related topics. [75] leverages a dynamic Neural Topic Model to capture how non-pharmacological interventions have been perceived and discussed across various countries and media outlets.

Content Generation Numerous studies have applied topic modeling to generate content that adheres to specific styles or themes [76]. For instance, [77] proposes a text generation model that captures semantic and structural features using a VAE-based NTM. Similarly, [78] utilizes NTMs to address information sparsity in long story generation. They map a short text to a low-dimensional document-topic distribution, from which they sample interrelated words to create a skeleton. This skeleton, along with the short text, is then used as input to a Transformer model to generate longer stories. [79] employs document-topic distributions from NTMs to enhance and control global semantics in text summarization. In the context of language translation, topic modeling is used to improve the accuracy of machine translation systems [80]. Furthermore, [81] introduces a neural hierarchical topic model to discover hierarchical topics within documents, subsequently generating keyphrases guided by these hierarchical topics.

Content Recommendation Similar to early works such as [82], topic models can be integrated with recommendation systems to suggest items to users based on their preferences [83–85]. For instance, [86] combines an NTM with a recommendation system for reviews using a structured auto-encoder. [87] employs a graph-based NTM for citation recommendation and Valero et al [69] use a short text NTM for podcast short-text metadata. In a medical application, [88] proposes a classification-aware NTM that integrates a topic model with a classifier, focusing on

classifying disinformation about COVID-19 to enhance the delivery of effective public health messages.

2.5 Conclusion

In this chapter, we delved into the diverse and evolving landscape of topic modeling approaches, highlighting their various applications across different domains. Probabilistic topic models, especially those based on Latent Dirichlet Allocation (LDA), have long been the cornerstone of topic modeling due to their ability to uncover latent thematic structures in text corpora. These models have been widely adopted and have formed the basis for numerous advancements in the field. However, they are not without their limitations, particularly in handling large-scale datasets and capturing the nuanced, context-dependent meanings of words.

With the recent surge in neural network research and the growing capabilities of deep learning, the field of topic modeling has seen significant innovation. Neural topic models have emerged, leveraging the power of neural networks to go beyond the probabilistic frameworks traditionally used. By exploiting the semantic depth offered by advanced language models, such as those based on transformers, neural topic models are capable of generating richer and more coherent topics, even from vast and complex document archives.

Chapter

3 Topic Model Evaluation

The analysis of the evolution of science is a complex and nuanced endeavor, requiring careful consideration due to the vast collections of documents that vary in size, publication date, and domain. Selecting an optimal topic model for this purpose necessitates a deep understanding of the metrics that can help determine the most suitable model for our specific needs. In this chapter, we will explore topic evaluation metrics for identifying the most effective models for analyzing the dynamic progression of scientific knowledge across different fields and time periods.

Chapter content

3.1 Introduction

Since ground-truth labels are absent in topic modeling tasks, determining reliable and comprehensive evaluation methods for topic models is inherently complex. This complexity arises from the rich semantics of the problem, including the latent structure and high dimensionality of textual data, context dependency, polysemy, homonymy, synonymy, and other linguistic phenomena.

Topic evaluation metrics provide valuable insights into the performance of topic models, each measuring different aspects depending on the application context. Initially, topic models were evaluated using *perplexity* [19], an automated metric that quantifies how well a statistical model predicts a sample of unseen data. Perplexity is computed by taking the inverse probability of the test set, normalized by the number of words in the dataset. A lower perplexity score indicates better predictive performance and more coherent document clustering. However, perplexity has been found to be inconsistent with human interpretability [89], leading to the adoption of *automated topic coherence metrics*. These metrics, such as those based on Point-wise Mutual Information (PMI) [90], assess the semantic relatedness of words within a topic and are believed to align more closely with human judgment. For instance, a topic with words like "dog," "cat," and "pet" would have high coherence. Additionally, *topic diversity metrics* [41] evaluate the variety and distinctiveness of topics generated by topic models. They quantify how different the identified topics are from each other, promoting richer and more comprehensive topic representations. A high diversity score indicates that the model captures a wide range of themes, ensuring that the topics are not overly similar. Each of these metrics – perplexity, coherence, and diversity – offers unique insights, making them useful in different contexts and for different objectives in topic modeling.

3.2 Perplexity

Borrowed from machine learning and language models, where perplexity measures the performance of a statistical model when it faces a new set of data [91], in topic modeling, perplexity evaluates a topic model's performance on a new set of test document [19]. More specifically, it quantifies how well the topic model predicts new documents based on the normalized log-likelihood of held-out test documents.

This is usually done by splitting the dataset into two parts: one for training, and the other for testing. For LDA, a test set is a collection of unseen documents $\mathbf{w}_d$, and the model is described by the topic matrix $\mathbf{\Phi}$ and the hyperparameter α for topic-distribution of documents. For instance, the LDA parameters $\mathbf{\Theta}$ are not taken into consideration as they represent the topic distributions for the documents of the training set, and can therefore be ignored to compute the likelihood of unseen documents. Therefore, we need to evaluate the following log-likelihood (Equation (3.1)) of a set of unseen documents $\mathbf{w}_d$ given the topics $\mathbf{\Phi}$ and the hyperparameter α for topic-distribution θ_d of documents.

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\mathbf{\Phi}, \alpha) = \sum_d \log p(\mathbf{w}_d|\mathbf{\Phi}, \alpha). \tag{3.1}$$

The likelihood of unseen documents can be used to compare models; a higher likelihood implies a better model. The measure traditionally used for topic models is the *perplexity* of held-out documents $\mathbf{w}_d$ defined as:

$$\text{perplexity}(\text{test set } \mathbf{w}) = \exp\left\{-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right\} \tag{3.2}$$

which is a decreasing function of the log-likelihood $\mathcal{L}(\mathbf{w})$ of the unseen documents $\mathbf{w}_d$; the lower the perplexity, the better the model.

However, the likelihood $p(\mathbf{w}_d|\mathbf{\Phi}, \alpha)$ of one document is intractable, which makes the evaluation of $\mathcal{L}(\mathbf{w})$, and therefore the perplexity, intractable as well. [92] derive various sampling methods to approximate this probability.

Discussion: Perplexity is a long-standing metric for evaluating topic models but also has faced criticism due to its poor alignment with human judgment [89, 93, 94]. Log-likelihood computation varies inconsistently across different topic models, leading to challenges in equitable comparisons [95, 96]. Additionally, perplexity may not adequately assess the practical utility of topic models, as users primarily employ them for content analysis rather than document generation.

3.3 Topic Diversity

Topic diversity estimates the diversity of the topic representations within a given set of topics. These metrics are driven by the expectation that the generated topic representations should be diverse and avoid topics with similar representations to efficiently represent the contents of some archive.

The authors of [97] propose *Topic Uniqueness* (TU) which computes the average reciprocal of top word occurrences in topics. In detail given K topics and the top T words of each topic, TU is computed as follows.

$$TU = \frac{1}{K}\sum_{K=1}^{K}\frac{1}{T}\sum_{x_i \in t(k)}\frac{1}{\#(x_i)} \tag{3.3}$$

where $t(k)$ means the top word set of the k-th topic, and $\#(x_i)$ denotes the occurrence of word x_i in the top T words of all topics. TU ranges from $1/K$ to 1.0, and a higher TU score indicates more diverse topics.

The authors of [98] propose *Topic Redundancy* (TR) that calculates the average occurrences of a top word in other topics. Note that the higher TR score means less diverse topics. TR is computed as follows.

$$TR = \frac{1}{K}\sum_{K=1}^{K}\frac{1}{T}\sum_{x_i \in t(k)}\frac{\#(x_i)-1}{K-1} \tag{3.4}$$

The authors of [41] propose the *Proportion of Unique Words* (PUW) method defined as the vocabulary size divided by the total number of words within a set of topics. PUW (Equation (3.5)) measures the semantic relatedness by the proportion of unique words in a list of K topics, where each topic is represented as a list of m words. The proportion is computed by taking the number of unique words divided by the total number of words for each topic. Topics within a low-diversity topic set share many words, whereas a high-diversity topic set contains topics that have few words in common.

$$\text{PUW} = \frac{|\bigcup_{i=1}^{K}\{t_i^r\}_{r=1}^{m}|}{m \cdot K} \tag{3.5}$$

Discussion: These metrics all assess topic diversity based on the uniqueness of individual words, assuming that diversity is optimal when each topic has distinct top words. However, we can challenge diversity metrics because some topics naturally share common words, which complicates reliable diversity evaluation. For instance, the word "chip" could belong to both the "potato chip" and "electronic chip" topics, highlighting that a single word can have multiple, context-dependent meanings. Similarly, "apple" could refer to both the "fruit" and the "company," illustrating how common words might bridge entirely different subject areas. This overlap indicates that the mere presence of unique words within topics is not always a sufficient indicator of true diversity. As a result, relying solely on traditional diversity metrics might lead to an oversimplified understanding of topic separation and cohesion, failing to account for the nuanced ways in which language is used across different domains and contexts. A possible approach to address this issue could involve incorporating more semantic analysis through the use of language models, such as comparing word embeddings instead of merely terms, to better evaluate topic diversity. This method would account for the shared vocabulary that naturally occurs between overlapping or related topics, offering a more nuanced evaluation.

3.4 Topic Coherence

Topic Coherence (TC) metrics assess the interpretability of topics produced by topic models [99]. In essence, they gauge the consistency of words within a given topic, evaluating their interpretive quality and meaningfulness by quantifying the semantic similarity among the words included in that topic. A high TC value indicates that the words in the topic are semantically similar and are likely to co-occur in the same circumstances.

TC metrics are categorized into two classes: automated TC metrics and human-annotated TC metrics [94]. Automated TC metrics estimate the interpretability of topic models concerning various factors such as co-occurrence or semantic similarity of topic words. On the other hand, human-annotated TC metrics are protocols for designing surveys that rate or score the interpretability and semantic coherence of topic models and are often used to validate automated topic coherence metrics [30, 95, 100]. While human-annotated TC metrics incorporate subjective

human judgments and provide a more accurate and nuanced understanding of how well topic models are performing (e.g. in terms of their ability to capture the underlying themes in a text corpus), they are expensive, time-consuming, and require multiple human-subjects to avoid personal biases. On the other hand, automated TC metrics are more cost-effective than human-annotated methods, as they do not require the hiring and training of human annotators, which results in their ability to evaluate large amounts of data and iterate through many model comparisons.

The authors of [30, 101] claim that a method based on the Point-wise Mutual Information (PMI) gives the largest correlations with human ratings. They define UCI, which measures the strength of the association between pairs of words based on their co-occurrence in a sliding window of length-l words. Topic coherence over PMI (TC_{UCI}) is defined as the average of the $\log_2$ ratio of co-occurrence frequency of word w_i^r and w_i^s within a given topic i (Equations (3.6) and (3.7)).

$$\text{TC}_{\text{UCI}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\binom{m}{2}}\sum_{r=2}^{m}\sum_{s=1}^{r-1}\text{PMI}(w_i^r, w_i^s) \tag{3.6}$$

with

$$\text{PMI}(w^i, w^j) = \log_2\frac{P(w^i, w^j)+\epsilon}{P(w^i)P(w^j)} \tag{3.7}$$

where n is the number of topics with m topic words and PMI represents the pointwise mutual information between each pair of words (w_i^r and w_i^s) in the topic i. PMI is computed by taking the logarithm of the ratio of the joint probability of two words $P(w_i^r, w_i^s)$ appearing together to the individual probabilities of the words $P(w_i^r)$, $P(w_i^s)$ occurring separately. Note that $\epsilon = 1$ is added to avoid the logarithm of zero.

The authors of [95] propose UMass, an asymmetric confirmation measure that estimates the degree of coherence between words within a given topic by calculating the log ratio frequency of their co-occurrences in the corpus of documents. UMass counts the number of times a pair of words co-occur in a given corpus and compares this number to the expected number of co-occurrences where words are randomly distributed across the whole corpus. More formally, UMass computes the co-document frequency of word w_i^r and w_i^s divided by the document frequency of word w_i^s (Equation (3.8)).

$$\text{UMass}(w_i^r, w_i^s) = \log\frac{D(w_i^r, w_i^s)+\epsilon}{D(w_i^s)} \tag{3.8}$$

The authors of [100] propose context vectors for each topic word w to generate the frequency of word co-occurrences within windows of ± 1 words surrounding all instances of w (Equation (3.9)). They showed that NPMI [102] has a larger correlation with human topic ratings compared to UCI and UMass. Additionally, NPMI takes into account the fact that some words are more common than others and adjusts the frequency of individual words accordingly [103].

$$\text{NPMI}(w_i^r, w_i^s) = \frac{\log_2\frac{P(w_i^r, w_i^s)+\epsilon}{P(w_i^r)P(w_i^s)}}{-\log_2(P(w_i^r, w_i^s)+\epsilon)} \tag{3.9}$$

While NPMI is generally more sensitive to rare words and can handle small datasets, UMass focuses on fast computation of coherence scores over large corpora. [104] showed that a smaller value of ϵ tends to yield better results than the default value of $\epsilon = 1$ used in the original paper since it emphasizes more the word combinations that are completely unattested.

The authors of [99] propose a unifying framework of coherence measures that can be freely combined to form a configuration space of coherence definitions, allowing their main elementary

components to be combined in the context of coherence quantification. They propose the C_V metric, which uses a variation of NPMI to compute topic coherence over a sliding window of size N and adds a weight γ to assign more strength to more related words (Equation (3.10)). According to [105], the C_V metric is more sensitive to noisy information and dirty data than C_{UMass} and C_{UCI}.

$$C_V(w_i^r, w_i^s) = \text{NPMI}^{\gamma}(w_i^r, w_i^s) \tag{3.10}$$

The authors of [106] and [107] propose the metric TC_{DWR} based on the Distributed Word Representations (DWR) [42, 108] which are better correlated to human judgment. One way to estimate TC_{DWR} is to compute the average pairwise cosine similarity between word vectors in a topic as in Equation (3.11).

$$\text{DWR}(w_i^r, w_i^s) = \frac{w_i^r \cdot w_i^s}{\|w_i^r\| \cdot \|w_i^s\|} \tag{3.11}$$

Another way is to define a topic vector t_i (e.g. by average summation of word vectors in the topic i) and then to calculate the average cosine similarity between topic vectors and topic word vectors in each topic (Equation (3.12)).

$$TC_{DWR}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\cos(w_i^j, t_i) \tag{3.12}$$

where

$$t_i = \frac{1}{m}\sum_{j=1}^{m} w_t^j \tag{3.13}$$

Similarly, the authors of [109] present a coherence measure based on grouping topic words into buckets and using Singular Value Decomposition (SVD) and integer linear programming-based optimization to create coherent word buckets from the generated embedding vectors.

The authors of [110] propose several topic coherence metrics based on topic documents rather than topic words. The approach vectorizes topic documents using several methods such as word embedding aggregation, and computes a coherence score based on the document vectors. [111] proposes an automated evaluation metric for local-level topic models by introducing a task designed to elicit human judgment and reflect token-level topic quality.

Discussion: Automated metrics are intended to align more closely with human judgment, providing a better measure of the interpretability of topic words. The risk of such approximations, however, is that they become the target of optimization rather than the underlying property they were intended to measure. Several recent works suggest that this has occurred especially in the context of neural topic models.

The authors of [112] argue that interpretability is ambiguous and conclude that current automated topic coherence metrics are unreliable for evaluating topic models in short-text data collections and may be incompatible with newer neural topic models. In a similar study, [94] show that topics generated by neural models are often qualitatively distinct from traditional topic models while they receive higher scores from current automated topic coherence metrics. [94] conclude that the validity of the results produced by fully automated evaluations, as currently practiced, is questionable, and they only help when human evaluations cannot be performed. [113] in another recent work shows that neural topic models fail to improve on the traditional topic models such as Gibbs LDA [114, 115] and consider neural topic broken as they do not function well for their intended use.

3.5 Conclusion

This chapter explored various topic evaluation metrics for selecting optimal topic models in scientific document analysis, highlighting the challenges posed by the absence of ground-truth labels and complex linguistic phenomena. We discussed three main metrics: perplexity, which measures predictive performance but lacks alignment with human interpretability; coherence, which assesses semantic relatedness and better aligns with human judgment; and diversity, which evaluates topic distinctiveness. The shift from perplexity to coherence reflects attempts to better capture human-interpretable results. However, recent studies have raised concerns about the reliability of automated metrics, especially for neural topic models, as they may become optimization targets rather than true quality measures [94]. The chapter concludes that topic model evaluation remains a nuanced field, calling for future research to develop more sophisticated approaches that account for language complexities and diverse applications of topic modeling.

Chapter

4 Contextualized Topic Coherence Metrics

In this chapter, we focus on topic coherence metrics, which have become the most commonly used evaluation tools in recent years. We critically examine their limitations and the challenges they present in accurately measuring the semantic consistency of topics. To address these shortcomings, we introduce a new metric called Contextualized Topic Coherence (CTC). This innovative metric enhances the evaluation process by incorporating the context in which words appear, allowing for a more nuanced assessment of topic quality. With this new approach, we re-evaluate the performance of the significant topic models discussed in Chapter 2, providing deeper insights into their effectiveness and reliability.

This work, entitled "Contextualized Topic Coherence Metrics," has been presented at the conference *Findings of the Association for Computational Linguistics: EACL 2024*, held in St. Julian's, Malta [116].

Chapter content

4.1 Introduction

The evaluation of topic models is a crucial aspect of understanding their performance, particularly in capturing the underlying themes of a text corpus. Human-annotated topic coherence (TC) metrics, which incorporate subjective human judgments, provide a more accurate and nuanced assessment of topic models. These metrics are valuable because they reflect human interpretability and understanding. However, they come with significant drawbacks: they are expensive, time-consuming, and require multiple human annotators to mitigate personal biases. In contrast, automated metrics offer a more cost-effective solution. They do not necessitate hiring and training human annotators, allowing for the evaluation of large datasets and facilitating multiple model comparisons efficiently. Automated metrics aim to align closely with human judgment, enhancing the interpretability of topic words. However, a major risk with automated metrics is that they

can become the target of optimization, rather than genuinely measuring the intended properties of topic models.

Recent studies highlight issues with current automated metrics, discussed in Chapter 7, particularly in the context of neural topic models. [112] argue that the interpretability of topic models by itself is inherently ambiguous and conclude that existing automated Topic Coherence (TC) metrics are *unreliable* for evaluating topic models on short-text data collections and may not be compatible with newer neural topic models. Similarly, the author of [94] demonstrates that while neural models often receive higher scores from current automated TC metrics, they tend to produce qualitatively weaker topics from traditional models. In another study, the author of [113] show that neural topic models fail to outperform traditional models like Gibbs LDA [114, 115], discussed in Chapter 2, suggesting that neural topic models do not function well for their intended use. Indeed, this shows that the existing automated topic coherence metrics may not align well with human-based coherence evaluation results, as they struggle to account for the linguistic context awareness introduced by neural topic models.

These findings indicate a demand for new automated coherence measures that are *context-aware* and can *effectively handle neural topic models and short-text datasets*. To address these challenges, we introduce the Contextualized Topic Coherence (CTC) metrics, a family of context-aware TC metrics based on pre-trained Large Language Models (LLMs). Leveraging LLMs enhances the understanding of language at a sophisticated level, incorporating linguistic nuances, contexts, and relationships. CTC metrics are designed to be less susceptible to being misled by meaningless topics that often receive high scores with traditional TC metrics. The contributions are as follows.

- Introduction of CTC Metrics: We introduce Contextualized Topic Coherence (CTC) metrics, a context-aware family of topic coherence metrics based on pre-trained Large Language Models (LLMs). These metrics offer a sophisticated understanding of language, incorporating its nuances and contexts.

- Comprehensive Analysis: We present a thorough analysis demonstrating the validity of CTC metrics compared to traditional TC metrics. This analysis is based on experiments with six topic models, including recent advanced neural topic models (ETM [41], ATM [65], CTM [51], BERTopic [16], Top2Vec [24]) and a dominant traditional topic model (Gibbs LDA [114]), across two different datasets.

- Re-evaluation of Topic Models: We re-evaluate these models using the proposed CTC metrics and demonstrate that CTC metrics perform well on short documents, avoiding the pitfall of high-scoring but meaningless topics. This re-evaluation underscores the effectiveness of CTC metrics in providing a more reliable and context-aware measure of topic coherence.

4.2 Contextualised Topic Coherence

In this section, we introduce Contextualized Topic Coherence (CTC), a new family of topic coherence metrics that benefit from the recent development of Large Language Models (LLM). We present two approaches. The first approach uses encoder-only LLMs to compute contextualized estimates of the Pointwise Mutual Information (CPMI) between topic words. In the second approach, we use decoder-only LLMs such as ChatGPT [56] to evaluate topic coherence by simulating to human-annotated evaluation methods.

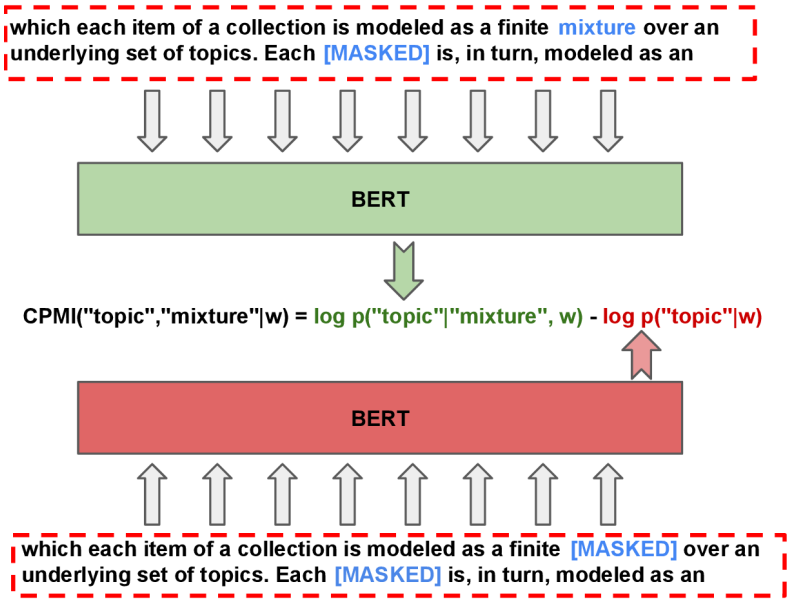


Figure 4.1: *Calculating CPMI for two topic words in a segment of a document.*

4.2.1 Automated CTC

Automated CTC leverages encoder-only Transformers[50] such as BERT[15] and the Contextualized Pointwise Mutual Information (CPMI)[117] metric to calculate topic coherence. By utilizing encoder-only LLMs, CTC captures rich contextual information, improving the assessment of how semantically related the words within a topic are. The incorporation of CPMI further refines this evaluation by measuring the contextualized dependencies between word pairs in a context-sensitive manner. We will show with our experiments in Section 4.3 that this advanced method provides a more accurate and automated means of determining topic coherence, enhancing the reliability of topic model evaluations.

CPMI Recent work by [117] uses conditional PMI estimates to analyze the relationship between linguistic and statistical word dependencies. For example, in a sentence about coffee, a high CPMI between "coffee" and "beans" would suggest that these words are closely related in this context, and knowing one word significantly helps in predicting the other. They propose Contextualized PMI (CPMI) as a new method for estimating the conditional PMI between words *in context* using a pre-trained language model. The CPMI between two words w_i and w_j in a sentence s is defined by the following equation:

$$\text{CPMI}(w_i, w_j \mid s) = \log \frac{p(w_i \mid s_{-w_i})}{p(w_i \mid s_{-w_{ij}})} \tag{4.1}$$

where s is a sentence, s_{-w_i} represents s with one masked word w_i (top in Figure 4.1) and $s_{-w_{ij}}$ is s with two masked words w_i and w_j (bottom in Figure 4.1). The conditional probability $p(w_i \mid s_{-w_{ij}})$ estimates the occurrence probability of w_i in $s_{-w_{ij}}$ based on a pre-trained masked language model (MLM) such as BERT.

We adopt CPMI to introduce a new automated metric called Contextualized Topic Coherence (CTC). CTC computes the CPMI value for each pair of topic words along a sliding window applied to a dataset. This method captures the statistical dependence between these words, reflecting their semantic relatedness and thus providing a measure of topic coherence. For this, the corpus is divided into a set of sliding window segments of length w and overlap k with previous and following segments to compute the average CPMI over all topic word pairs in all window

segments:

$$\mathrm{CTC}_{\mathrm{CPMI}}(T, D, w, k) = \frac{1}{n * \binom{m}{2}} \sum_{i=1}^{n} \sum_{r=2}^{m} \sum_{s=1}^{r-1} \mathrm{CPMI}(w_i^r, w_i^s \mid c^u) \tag{4.2}$$

where $c^u \subset$ corpus D is a window segment with length of w that has k words overlapping with its adjacent window segments, n is the number of topics and m is the number of topic words.

4.2.2 Semi-automated CTC

In this section, we employ instruction-based LLMs[118] such as ChatGPT[56] to replicate the most important human-annotated evaluations for topic models: Word Intrusion Task and Rating Task. By leveraging these advanced LLMs, we aim to automate and enhance the evaluation process, which traditionally relies on labor-intensive human annotations. These models can generate high-quality, context-aware assessments that mirror human judgment. We will show in Section 4.3 that this approach not only demonstrates the efficacy of the topic models in terms of interpretability but also validates the potential of instruction-based LLMs in replicating and "possibly" surpassing human annotation accuracy in the field of topic modeling.

Word Intrusion Task [89] proposed the *topic words intrusion task* to assess topic coherence by identifying a coherent latent category for each topic and discovering the words that do not belong to that category. In this task, human subjects detect *topic intruder words* to assess the quality of topic models and to measure a coherence score that assigns a low probability for intruder words to belong to a topic. We apply this idea by replacing humans with ChatGPT [56] answering to prompts that provide the topic words and ask for a category and intruder words. The prompt is as follows.

> **System prompt:** *I have a topic that is described by the following keywords: [topic-words]. Provide a one-word topic based on this list of words and identify all intruder words in the list concerning the topic you provided. Results be in the following format: topic: <one-word>, intruders: <words in a list>*

Rating Task The *topic rating task* consists of rating topics by their usefulness for a given task (for example, document search). While human topic ratings are expensive to produce, they serve as the gold standard for coherence evaluation [99]. For example, [119] uses human ratings to explore the coherence of topics generated by LDA topics across full texts and abstracts. [120] provides human annotators with a rubric and guidelines for judging whether a topic is useful or useless. The annotators evaluate a randomly selected subset of topics for their usefulness in retrieving documents on a given topic and score each topic on a 3-point scale, where 3=highly coherent and 1=useless (less coherent). Following [120], [100] presented topics without intruder words to Amazon Mechanical Turk to score them on a 3-point ordinal scale. Similar to the intrusion task, we adapt this method to ChatGPT by defining prompts that provide ChatGPT with the topic words and ask it to rate the usefulness of the various topic words for retrieving documents on a given topic. The $\mathrm{CTC}_{\mathrm{Rating}}$ for a topic model is obtained by the average sum of all ratings over all topics. The prompt is as follows.

> **System prompt:** *I have a topic that is described by the following keywords: [topic-words]. Evaluate the interpretability of the topic words on a 3-point scale where 3 = "meaningful and highly coherent" and 0 = "useless" as topic words are usable to search and retrieve documents about a single particular subject. Results be in the following format: score: <score>*

4.3 Experiments

In this section, we validate the effectiveness of CTC and compare the most prominent topic models across different categories. By validating CTC, we aim to assess its reliability and compare it with traditional coherence metrics in capturing the semantic consistency of topics. As a result, we can also conduct a comprehensive comparison between prominent topic models, evaluating their performance. This dual approach not only establishes CTC as a robust evaluation tool but also provides valuable insights into the comparative efficacy of different topic modeling techniques for scientific archives. Moreover, we expect to observe that the baseline metrics (UCI, UMass, NPMI, C_V, DWR) rank topic models differently from CTC. We also expect CTC rankings to favor interpretable topics and handle short text datasets more effectively than the baseline metrics [94, 112]. In contrast, CTC uses a much richer contextualized language model than the statistical models used by the other methods. Therefore, we expect to see that baseline metrics and CTC would differ at extremes of highest or lowest coherency.

4.3.1 Experimental Setup

The experimental setup encompasses datasets from diverse domains, baseline topic models, baseline metrics for evaluation, and specific configurations for CTC to ensure comprehensive and rigorous comparisons across all evaluated models.

Datasets The experiments incorporate two datasets including the 20Newsgroups dataset[121] which is a collection of approximately 20K newsgroup documents, partitioned evenly across 20 different newsgroups describing various topics. This dataset has been extensively studied with many papers published on topic modeling since it provides a diverse set of documents across multiple topics, making it an ideal test bed for evaluating topic modeling algorithms. The second dataset is a collection of 17K tweets by Elon Musk published between 2017 and 2022 by [122]. Each tweet includes a text message, the date and time it was posted, and other metadata such as the number of likes and retweets. The purpose of selecting the second dataset is to evaluate the performance of topic models on short-text data, specifically in terms of their ability to produce coherent topics as measured by baseline and CTC metrics.

Topic Models The experiments involve six different topic models including traditional topic models that are widely used and the new neural topic models that have been recently proposed. These topic models include Gibbs LDA [114], Embedded Topic Model (ETM) [41], Adversarial-neural Topic Models (ATM) [65], Top2Vec [24], and Contextualized Topic Model (CTM) [51], and BERTopic [16], which are discussed with details in Chapter 2. In the case of parametric models, we selected the number of topics based on the values reported by their original papers. Specifically, for the 20Newsgroup dataset, we chose 20, 50, and 100 topics, while for Elon Musk's Tweets, we selected 10, 20, and 30 topics. We maintained the remaining settings at default values as suggested by the main papers. For the reproduction of code for each topic model, we have a GitHub repository with this link that you can use: GitHub Repository.

Topic Coherence Metrics The topics generated by the topic models are evaluated using the proposed Contextualized Topic Coherence (CTC) metrics, which are then compared to the well-established automated topic coherence metrics C_V, UCI, UMass, NPMI, and DWR, as are explained in Chapter 7.

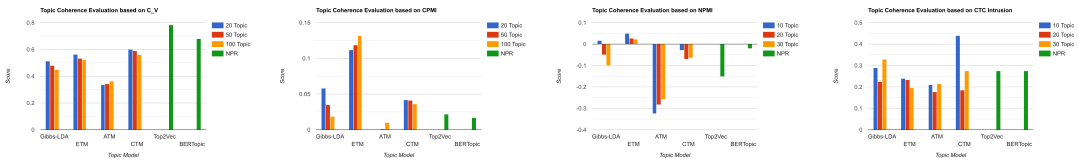
CTC Configuration For CTC_{CPMI}, we segmented the 20Newsgroup and Elon Musk's Tweets datasets into chunks of 15 and 20 words, respectively, without intersections. We then extracted the CPMI for all word pairs in each segment using the pre-trained language models *bert-base-uncased* and *Tesla K80 15 GB GPU* from Google Colab [123]. This pre-computing step took about 7 hours but allowed us to compute CTC_{CPMI} for any topic model in the order of a few seconds. The advantage of pre-computing CPMI between all word pairs is that one can run it on very large datasets and open-source it for CTC_{CPMI} calculation. Note that calculating CPMI only for the word pairs of all topics takes on the order of a few minutes. For evaluating $CTC_{Intrusion}$ and CTC_{Rating}, we made a request for each topic to *ChatGPT* with *GPT 3.5 Turbo*, which cost less than a dollar for all the experiments.

4.3.2 Results

Tables 4.1 and 4.2 represent the results of the evaluation of the topic models obtained from the 20Newsgroup and Elon Musk's Tweets datasets, respectively, using CTC and the baseline metrics. The highest value for each metric is shown in bold to compare the models in terms of concerning metrics. The highest values for each metric within each topic model are noted in *italic* font. This helps us determine the optimal number of topics for all models except Top2Vec and BERTopic, which don't require this input parameter. For example, as illustrated in Table 4.1, the most coherent number of topics for Gibbs LDA is 20 across all coherence metrics, while for ETM, the optimal number is 20 for all metrics except UMass. ATM achieves the best coherence with 100 topics, except with respect to DWR and Rating. Similarly, as illustrated in Table 4.2, for the short-text Elon Musk Tweets dataset, the optimal number of topics follows a comparable pattern. Specifically, for CTC_{CPMI}, the most coherent parameters for Gibbs LDA, ETM, ATM, and CTM are 20, 100, 100, and 20, respectively.

Table 4.1: *Scores of Topic Coherence Metrics on 20Newsgroup dataset.*

Topic Models		Baseline Metrics					CTC Metrics		
	#T	UCI	UMass	NPMI	C_V	DWR	Rating	Intrusion	CPMI
Gibbs LDA	20	*0.260*	*-2.338*	*0.043*	*0.512*	*0.211*	*1.3*	0.225	*9.92*
	50	-0.121	-2.771	0.023	0.479	0.191	1.16	0.220	5.99
	100	-0.690	-3.030	0.002	0.450	0.149	1.14	*0.267*	3.25
ETM	20	*0.478*	-2.08	*0.067*	*0.563*	0.292	0.7	***0.452***	19.16
	50	0.380	***-1.903***	0.054	0.532	*0.330*	1.22	0.348	20.35
	100	0.351	-1.962	0.049	0.522	0.312	*1.23*	0.41	***22.58***
ATM	20	-1.431	-3.014	-0.059	0.338	*0.151*	0.92	0.305	0.03
	50	-0.940	-2.902	-0.046	0.342	0.077	*1.15*	0.275	0.18
	100	*-0.735*	*-2.741*	*-0.032*	*0.362*	0.053	1.12	*0.340*	*1.72*
CTM	20	-1.707	-4.082	0.005	*0.601*	*0.268*	1.25	0.385	*5.93*
	50	*-0.724*	*-3.008*	*0.046*	0.590	0.236	*1.56*	0.380	7.02
	100	-0.926	-3.118	0.027	0.561	0.210	1.31	*0.392*	6.16
Top2Vec	85	***0.910***	*-2.449*	***0.192***	***0.785***	***0.473***	***1.670***	*0.399*	*3.77*
BERTopic	145	*-1.023*	*-5.033*	*0.098*	*0.681*	*0.309*	*1.517*	*0.359*	*2.91*



(a) *20Newsgroup / C_V* **(b)** *20Newsgroup / CPMI* **(c)** *Twitter / NPMI* **(d)** *Twitter / Intrusion*

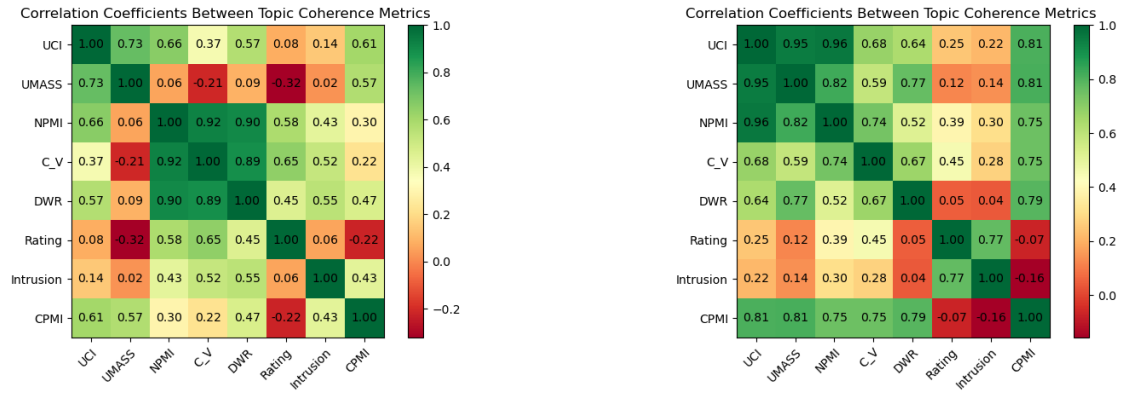
Figure 4.2: *Comparison Between Topic Models based on Topic Coherence Evaluation*

General Observations Before analyzing the results in detail, we examine the relationship between the CTC metrics and the baseline metrics by performing Pearson's correlation coefficient

Table 4.2: *Scores of Topic Coherence Metrics on Elon Musk's Tweets dataset*

Topic Models		Baseline Metrics					CTC Metrics		
	#T	UCI	UMass	NPMI	C_V	DWR	Rating	Intrusion	CPMI
Gibbs LDA	10	*-0.441*	*-3.790*	*0.016*	*0.498*	*0.838*	*1.6*	0.29	*2.19*
	20	-1.834	-5.415	-0.049	0.395	0.798	1.5	0.225	1.04
	30	-3.068	-6.390	-0.099	0.336	0.783	1.466	*0.33*	0.86
ETM	10	***0.205***	-3.209	***0.051***	*0.560*	0.952	1.1	*0.24*	***5.41***
	20	0.155	***-3.079***	0.028	0.538	0.974	*1.433*	0.233	4.48
	30	0.025	-3.215	0.022	0.515	***0.978***	1.05	0.195	4.30
ATM	10	-9.021	-12.859	-0.324	*0.364*	0.730	*1.2*	0.211	-0.004
	20	-7.967	-11.770	-0.283	0.343	0.694	1.1	0.177	*0*
	30	*-7.278*	*-11.301*	*-0.258*	0.350	*0.753*	0.933	*0.214*	-0.03
CTM	10	*-2.614*	*-7.049*	*-0.030*	***0.580***	*0.888*	***2.0***	***0.439***	1
	20	-3.720	-8.336	-0.070	0.534	0.880	1.45	0.185	*3.04*
	30	-3.589	-8.063	-0.064	0.573	0.873	1.766	0.276	2.56
Top2Vec	164	*-6.272*	*-10.536*	*-0.152*	*0.401*	*0.847*	*1.481*	*0.274*	*2.08*
BERTopic	217	*-4.131*	*-11.883*	*-0.020*	*0.432*	*0.541*	*1.539*	*0.276*	*1.52*

analysis [124] on the results from Tables 4.1 and 4.2 similar to [112]. Performing Pearson's correlation coefficient analysis for topic coherence evaluations involves calculating the degree of linear relationship between pairs of coherence scores obtained from different evaluation metrics or methods. This analysis quantifies how closely the coherence assessments derived from various approaches align, providing insights into their consistency and reliability across different topic models. As shown in Figure 4.3(a), for 20Newsgroup, the baseline metrics UCI and UMass are highly correlated with CPMI but not with CTC_{Rating} and $CTC_{Intrusion}$, which are more correlated with the baseline measures NPMI and C_V and DWR (which are also highly correlated). On the other hand, for the short text EM Tweets dataset, Figure 4.3(b) shows that CPMI has a high correlation with all baseline methods, while $CTC_{Intrusion}$ and CTC_{Rating} are completely independent of CPMI and the baseline measures.



(a) *20Newsgroup* **(b)** *Elon Musk Tweets*

Figure 4.3: *Pearson's correlation coefficient on CTC and baseline*

Concerning our expectation that baseline metrics rank topic models differently from CTC metrics, Table 4.1 reports that the baseline metrics (except for UMass) point to Top2Vec while CTC metrics (except for CTC_{Rating}) point to ETM for achieving the highest scores. Similarly, Table 4.2 reports that the baseline metrics (except for C_V) point to ETM while CTC metrics (except for CTC_{CPMI}) point to CTM for achieving the highest scores. These contradictions between CTC and baseline metrics are aligned with our expectations and we will explore them with a meta-analysis of topics generated by these topic models and the scores they have received from CTC and baseline metrics.

Meta-analysis To check the performance of different coherence metrics, we will compare the interpretability of their high and low-scoring topics. Note that CTC metrics observe contextual patterns between topic words, and therefore, we expect them to provide more consistent coherence scores according to the interpretability of the generated topics for all topic models.

20Newsgroup Dataset To verify the consistency of some representative scores in Table 4.1, we examine the topics for 20Newsgroup generated by Top2Vec, which have high and low baseline metrics scores, and ETM, which have high and low CTC metrics scores. Table 4.3 compares the top-2 and bottom-2 topics ranked by C_V and CTC_{CPMI}. The choice of these metrics is motivated by our correlation analysis in Figure 4.3(a), which has the least correlation among CTC and baseline metrics in CTC_{CPMI} and C_V.

First, we notice that the top-2 topics returned by C_V for Top2Vec are not readily interpretable but are statistically meaningful: *dsl, geb, cadre, shameful, jxp* are fragments of an email signature that occurs 82 times, while *tor, nyi, det, chi, bos* are abbreviations for hockey teams. This is not surprising, since Top2Vec produces what we call "trash topics", which is a common problem for clustering-based topic models that cannot handle so-called "trash clusters" [125]. CTC_{CPMI} returns a more coherent ranking for Top2Vec (the top 2 topics appear coherent, while the bottom topics are incoherent for human evaluation). This supports our assumption that traditional topic coherence metrics such as C_V fail to evaluate neural topic models and, in this case, even give the highest scores to trash topics. This happens because they only consider the syntactic co-occurrence of words in a window of text and cannot observe the underlying relationship between topic words. CTC_{CPMI}, on the other hand, can detect these trash topics and score them more accurately because it is supported by LLMs that have rich information about linguistic dependencies between topic words. Therefore, CTC_{CPMI} also might be a good measure to filter "trash topics" obtained by some cluster-based topic model.

The second observation in Table 4.3 is that all eight topics returned for ETM are coherent. This is because ETM, which is a semantically-enabled probabilistic topic model, produces decent topics that are overall highly ranked by CTC_{CPMI} (Table 4.1 and fig. 4.2(b)).

Table 4.3: *Top-2 and bottom-2 topics of* $ETM^{(100)}$ *and Top2Vec on 20Newsgroup*

Topic Model	Ranked By	Topics	C_V	CPMI
$ETM^{(100)}$	Highest C_V	god, christian, people, believe, jesus	0.740	0.017
		drive, card, scsi, disk, mb,	0.739	0.037
	Lowest C_V	book, number, problem, read, call	0.369	0.018
		line, use, power, bit, high	0.458	0.018
	Highest CPMI	year, time, day, one, ago, week	0.559	0.709
		game, year, team, player, play	0.706	0.242
	Lowest CPMI	new, number, also, well, call, order, used	0.340	-0.007
		people, right, drug, state, world, country	0.529	-0.002
Top2Vec	Highest C_V	dsl, geb, cadre, shameful, jxp	0.995	0.009
		tor, nyi, det, chi, bos	0.989	0.012
	Lowest C_V	hacker, computer, privacy, uci, ethic	0.255	-0.0001
		battery, acid, charged, storage, floor	0.344	0.006
	Highest CPMI	mailing, list, mail, address, send	0.792	0.154
		icon, window, manager, file, application	0.770	0.076
	Lowest CPMI	lc, lciii, fpu, slot, nubus, iisi	0.853	-0.004
		ci, ic, incoming, gif, edu	0.644	-0.002

Twitter Dataset In the same way, we verify the consistency of some representative scores in Table 4.2 by checking the interpretability of topics for Elon Musk's tweets generated by ETM, which has high baseline scores, and by CTM, which has high CTC scores. These metrics are among those with the lowest correlation between CTC and baseline metrics in Figure 4.3(b). We compare the top 2 and bottom 2 topics ranked by NPMI and CTC_{Rating} shown in Table 4.4.

A notable finding for CTM topics is that topics ranked highest by the CTC_{Rating} metric tend to be more interpretable compared to those ranked highest by NPMI. Similarly, topics ranked

Table 4.4: *Top-2 and bottom-2 topics of ETM$^{(30)}$ and CTM$^{(30)}$ on Elon Musk's Tweets*

Topic Model	Ranked By	Topics	NPMI	Rating	Intrusion
CTM$^{(30)}$	Highest NPMI	erdayastronaut, engine, booster, starship, amp	0.122	3	0.1
		year, week, next, month, wholemarsblog	0.057	2	0.1
	Lowest NPMI	transport, backup, ensure, installed, transaction	-0.480	2	0.1
		achieving, transition, late, transport, precision	-0.459	1	0.1
	Highest Rating	tesla, rt, model, car, supercharger	-0.152	3	0.5
		spacex, dragon, launch, falcon, nasa	-0.283	3	0.4
	Lowest Rating	ppathole, soon, justpaulinelol, yes, sure	-0.330	1	0.5
		achieving, transition, late, transport, precision	-0.459	1	0.1
ETM$^{(30)}$	Highest NPMI	amp, time, people, like, would, many	0.001	2	0.7
		engine, booster, starship, heavy, raptor	-0.023	2	0.1
	Lowest NPMI	amp, rt, tesla, im, yes	-0.283	1	0.1
		amp, tesla, year, twitter, work	-0.228	1	0.1
	Highest Rating	amp, twitter, like, tesla, dont	-0.186	2	0.8
		amp, time, people, like, would	0.001	2	0.7
	Lowest Rating	amp, tesla, year, twitter, work	-0.228	1	0.1
		amp, tesla, one, like, time	-0.204	1	0.1

lowest by the $\text{CTC}_{\text{Rating}}$ metric tend to be less interpretable compared to those ranked lowest by NPMI. These observations also apply to ETM topics, as the $\text{CTC}_{\text{Rating}}$ metric is not affected by the scarcity of short text records. This is because $\text{CTC}_{\text{Rating}}$ is complemented by a chatbot that mitigates the impact of limited data availability.

It is also interesting to observe that, based on the four topics seen, the topics generated by CTM appear to be more interpretable and coherent than those generated by ETM. This demonstrates the superiority of $\text{CTC}_{\text{Rating}}$ and $\text{CTC}_{\text{Intrusion}}$ over baseline metrics, as we observed in Table 4.2. It also suggests that, for the short text datasets examined inFigure 4.2(d), CTM is ranked higher than ETM by the $\text{CTC}_{\text{Intrusion}}$ metric, potentially due to the contextualized element in its architecture.

4.4 Human Evaluation

The goal of automated topic coherence metrics is to accurately approximate human judgment on topics without the need for expensive, time-consuming studies that require multiple annotators to avoid bias. In this section we compare the proposed metric with human evaluation data provided by [94]. The validity of automated topic coherence metrics depends on their correlation with human scores. However, these evaluations are expensive, time-consuming, and require multiple human subjects to avoid personal bias. In this section, we use the evaluation data provided in [94] and investigate the proposed metric with human scores. This data includes human evaluation scores (intrusion and ranking) for 50 topics generated by three topic models (Gibbs LDA [115], DVAE [29], and ETM [41]) applied on the (New York Times) dataset. We evaluate the generated topics with CTC_{CPMI}, $\text{CTC}_{\text{intrusion}}$ and $\text{CTC}_{\text{ranking}}$, which are comparable to human intrusion and human ranking.

As shown in Table 4.5, human evaluators tend to see little quantifiable difference between Gibbs LDA and DVAE, while traditional metrics show pronounced differences. In contrast, we find that CTC metrics more closely match human preferences (or lack thereof). This result may be simply due to a miscalibration of relative scores. We also report Spearman's Rank Correlation [126] results to assess the strength and direction of the monotonic relationship between the ranking of topics in each metric. The CTC metrics have an overall higher correlation with human ratings than the baseline metrics. We also can examine and compare different coherence metrics by analyzing the topic words of high and low-scoring topics. As shown in Tables 4.6 and 4.7, C_V generates top topics which probably would not be chosen by a human. For example, the topic *inc, 9mo, earns, otc, qtr, rev* gets the highest score, even though it has little clear interpretability. On the other hand, CTC metrics score topics relative to their contextual relationship and are very close to human scores. For example, the topic *film, theater, movie, play, director, movies*

Table 4.5: *Topic Coherence Scores of Gibbs LDA, DVAE, ETM on NYT News*

Metrics		Topic Models (T = 50)		
		Gibbs LDA	DVAE	ETM
Baseline	UCI	1.42	**2.43**	1.01
	UMass	-7.6	-15	**-7.4**
	C_V	0.69	**0.84**	0.60
	NPMI	0.15	**0.25**	0.11
Human	Intrusion	0.71	**0.74**	0.64
	Rating	**2.66**	2.48	2.38
CTC	Intrusion	**2.12**	2.05	2.06
	Rating	0.62	**0.67**	0.64
	CPMI	**4.18**	0.61	3.72

receives the highest score by both CTC and human scoring.

Table 4.6: *Bottom-5 topics among the topics generated by Gibbs LDA, DVAE, and ETM on NYT News*

Botton-5 Sorted by	Model	Topic	Scores		
			C_V	Human	CTC
C_V	DVAE	spade, derby, belmont, colt, spades, dummy, preakness	0.23	1.5	0.4
	ETM	like, making, important, based, strong, including, recent	0.35	2	0.3
	ETM	time, half, center, open, away, place, high	0.37	1.6	0.2
	ETM	today, group, including, called, led, known, began, built, early,	0.37	2	0.3
	Gibbs LDA	people, editor, time, world, good, years, public, long,	0.37	0.1	1.1
Human Score	Gibbs LDA	people, editor, time, world, good, years, public,	0.37	0.1	1.1
	ETM	week, article, page, march, tuesday, june, july	0.57	0.4	1.3
	Gibbs LDA	street, tickets, sunday, avenue, information, free	0.75	0.4	0.3
	ETM	new_york, yesterday, director, manhattan, brooklyn, received	0.49	0.4	1
	Gibbs LDA	bedroom, room, bath, taxes, year, market, listed, kitchen, broker	0.72	0.4	1.3
CTC	Gibbs LDA	city, mayor, state, new_york, new_york_city, officials	0.61	2.5	0.1
	ETM	power, number, control, according, increase, large	0.44	0.9	0.2
	Gibbs LDA	officials, board, report, union, members, agency, yesterday	0.51	0.8	0.3
	ETM	time, half, center, open, away, place, high, day, run	0.37	1.2	0.3
	ETM	net, share, inc, earns, company, reports, loss, lead	0.73	1.8	0.3

4.5 Conclusion

In this chapter, we have introduced Contextualized Topic Coherence (CTC) metrics which leverage pre-trained Large Language Models (LLMs) to offer a nuanced understanding of language and context for estimating topic coherence. Through a comprehensive analysis, we have demonstrated the superiority of CTC metrics over traditional Topic Coherence (TC) metrics, showcasing their effectiveness across a range of advanced neural and traditional topic models. Our experiments with six models, including ETM, ATM, CTM, BERTopic, Top2Vec, and Gibbs LDA, on diverse datasets have validated that CTC metrics excel in capturing meaningful semantic relationships, particularly beneficial for short documents where traditional metrics may falter. This implies that baseline metrics often yield high scores for incoherent topics, while conversely assigning low scores to well-interpretable topics. This re-evaluation underscores CTC metrics as a robust tool for evaluating and improving the quality of topic models in diverse applications such as topic evolution.

Table 4.7: *Top-5 topics among the topics generated by Gibbs LDA, DVAE, and ETM on NYT News*

Top-5 Sorted by	Model	Topic	Scores		
			C_V	Human	CTC
C_V	DVAE	inc, 9mo, earns, otc, qtr, rev	0.98	1.2	0.9
	DVAE	inc, 6mo, earns, otc, rev, qtr	0.98	1.2	1.3
	DVAE	inc, otc, qtr, earns, rev, 6mo	0.97	1.3	0.8
	DVAE	arafat, hamas, gaza, palestinians, west_bank	0.97	2.1	1.5
	DVAE	condolences, mourns, mourn, board_of_directors, heartfelt, deepest	0.97	0.6	1.3
Human Score	Gibbs LDA	film, theater, movie, play, director, films	0.73	3	2.7
	DVAE	skirts, dresses, chanel, couture, fashion	0.91	3	1.3
	DVAE	tenants, tenant, zoning, rents, landlords, developers	0.86	3	1.2
	DVAE	paintings, sculptures, galleries, picasso, sculpture, drawings,	0.91	2.9	2.1
	DVAE	television, network, news, cable, nbc, year, cbs	0.68	2.8	1.9
CTC	Gibbs LDA	film, theater, movie, play, director, films	0.73	3	2.7
	ETM	court, judge, law, case, federal, lawyer, trial	0.80	2.8	2.6
	Gibbs LDA	court, law, judge, case, state, federal, legal,	0.72	2.6	2.2
	Gibbs LDA	music, dance, opera, program, work, orchestra, performance	0.73	1.1	2.1
	ETM	film, movie, story, films, directed, movies, star, character	0.79	2.7	2.1

Part II

Dynamic Topic Modeling

> *Unknown in Paris, I was lost in the great city, but the feeling of living there alone, taking care of myself without any aid, did not at all depress me. If sometimes I felt lonesome, my usual state of mind was one of calm and great moral satisfaction.*
>
> Marie Curie

Motivation

Studying dynamic topic models is crucial for understanding the evolution of science as they allow researchers to analyze how scientific themes and interests shift over time. By capturing the temporal dynamics of topic prevalence, dynamic topic models help identify influential works, pivotal discoveries, and the diffusion of ideas across disciplines. Moreover, they play a significant role in examining the evolution of content within specific domains, enabling a detailed exploration of how the focus and discourse within a field change.

Organization

In this part of the thesis, we explore what dynamic topic models are, providing an overview of their mechanisms and applications in Chapter 5. We delve into the significant models within this domain, emphasizing their advantages, such as the ability to capture temporal shifts in topic prevalence and the evolution of scientific discourse. We also discuss the challenges associated with dynamic topic models, including computational complexity and the difficulty of accurately modeling rapidly changing topics. In Chapter 6, we introduce ANTM, a novel dynamic topic model that incorporates neural topic modeling with advanced dynamic clustering techniques to more precisely capture the temporal dynamics of topics and their interrelationships.

Contributions

Our contributions in this part are twofold. First, in Chapter 6, we introduce ANTM [127, 128], a novel dynamic topic model that incorporates neural topic modeling with advanced dynamic clustering techniques to more precisely capture the temporal dynamics of topics and their interrelationships. We then conduct a comparative analysis of ANTM against other state-of-the-art models, evaluating their performance based on various criteria such as diversity and interpretability. This comparative study aims to highlight the improvements and potential applications of our new model in understanding the evolution of scientific content and domain evolution.

This contribution has been presented at "BDA 2023: 39ème Conférence sur la Gestion de Données – Principes, Technologies et Applications" published in Rahimi, H., Naacke, H., Constantin, C., Amann, B. (2024). *ANTM: Aligned Neural Topic Models for Exploring Evolving Topics.* In: Hameurlain, A., Tjoa, A.M., Akbarinia, R., Bonifati, A. (eds) Transactions on Large-Scale Data- and Knowledge-Centered Systems LVI. Lecture Notes in Computer Science, vol 14790. Springer, Berlin, Heidelberg.

Chapter

5 Dynamic Topic Models

In this chapter, we focus on the concept of dynamic topic model, categorizing and examining their various types. We study their advantages, such as their capacity to track the evolution of topics over time and their utility in uncovering hidden patterns in temporal data. Additionally, we address the challenges associated with these models, including the computational demands and the complexity of accurately modeling dynamic data. By categorizing and scrutinizing these models, we aim to provide a comprehensive understanding of their strengths and limitations, laying the groundwork for further innovations and applications in the study of evolving scientific archives.

Chapter content

5.1 Introduction

Dynamic topic models are the temporal variants of topic models that update their estimates of the underlying topics as new documents are added to the corpus [6, 129]. They can be used to analyze topic evolution and to identify *evolving topics* and temporal patterns in the contents of document archives [130, 131]. More specifically, they can be applied, for example, to discover the evolution of research topics and innovations in scientific archives [132] and to understand trends in public opinion on particular issues [133].

Before introducing dynamic topic models, it is essential to understand the notion of evolving topics that capture the temporal progression of semantically similar documents within an ordered sequence of time periods in an input archive. This concept is formally defined as follows.

Definition 5.1.1 (Evolving Topics)**.** Given a corpus D of *time-stamped documents* and an ordered sequence of possibly overlapping time periods $P = [p^0, ..., p^n]$, an evolving topic is a sequence of *sub-topics* $t = [t^0, ..., t^n]$ such that all documents in $D(t^i)$ have been published during the time window p^i and $D(t) = \cup_{t^i \in t} D(t^i)$, is a cluster of *semantically similar* documents. Figure 5.1 represents two evolving topics from a corpus of time-stamped documents.

Dynamic topic models are a computational approach that captures the temporal changes in topic distributions over a sequence of time periods by producing sets of evolving topics. Formally, dynamic topic models can be defined as follows:

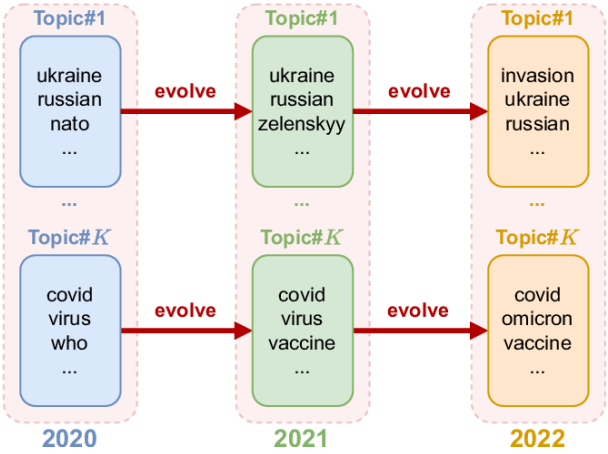


Figure 5.1: *Topic Representation of Dynamic Topic Models [69]*

Definition 5.1.2 (Dynamic Topic Models)**.** Given a corpus D over a sequence of time periods $P = [p^0, ..., p^n]$, a dynamic topic model generates a set of evolving topics $\mathrm{T}_{\mathrm{dynamic}}$, where each evolving topic $t = [t^0, ..., t^n] \in \mathrm{T}_{\mathrm{dynamic}}$ is a sequence of topics as defined in Definition 5.1.1.

Dynamic topics models identify how the weighted list of terms $\mathrm{W_t}$ and the subset of documents $\mathrm{D_t}$ associated with each topic t^i in t change over time, thereby providing insights into the temporal dynamics of the corpus.

Dynamic topic models use different assumptions to statistically characterize the evolution of a document corpus in the form of evolving topics [11]. Like their static counterparts, dynamic topic models can be divided into Probabilistic Dynamic Topic Models (PDTM) and Neural Dynamic Topic Models (NDTM).

5.2 Probabilistic Dynamic Topic Models

Different assumptions regarding the definition of topic evolution are employed to statistically characterize dynamic topic models (DTMs) [11].

Dynamic Latent Dirichlet Allocation (D-LDA) [6] represents one of the earliest PDTMs that incorporates temporal components into topic models for studying topic evolution. D-LDA [6] is a variant of Latent Dirichlet Allocation (LDA) [19] which employs a Bayesian approach to characterize the evolving proportions of topics within documents as time progresses. The generative model of D-LDA differs from LDA in that the topics are time-specific, denoted as $\beta_{1:K}^{(t)}$, where $t \in \{1, \ldots, T\}$ indexes time steps. Additionally, the prior over the topic proportions θ_d depends on the time stamp of document d, denoted as $t_d \in \{1, \ldots, T\}$.

- 1. Draw topic proportions $\theta_d \sim LN(\delta_{t_d}, a^2 I)$.
- 2. For each word n in the document:
  - (a) Draw topic assignment $z_{dn} \sim \mathrm{Cat}(\theta_d)$.
  - (b) Draw word $w_{dn} \sim \mathrm{Cat}(\beta_{z_{dn}}^{(t_d)})$.

where a represents a model hyperparameter and δ_t is a latent variable that governs the prior mean over the topic proportions at time t. To promote smoothness across topics and topic proportions, D-LDA introduces random walk priors for $\beta_{1:K}^{(t)}$ and δ_t.

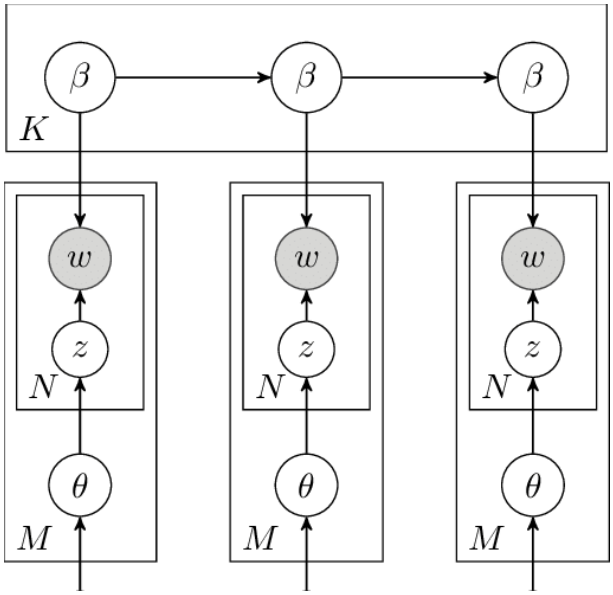


Figure 5.2: *D-LDA Graphical Model*

$$\bar{\beta}_k^{(t)} | \bar{\beta}_k^{(t-1)} \sim N(\bar{\beta}_k^{(t-1)}, \varphi^2 I) \tag{5.1}$$

$$\beta_k^{(t)} = \text{softmax}(\bar{\beta}_k^{(t)}) \tag{5.2}$$

$$\delta_t | \delta_{t-1} \sim N(\delta_{t-1}, \theta^2 I) \tag{5.3}$$

$\bar{\beta}_k^{(t)} \in \mathbb{R}^V$ are transformed topics mapped to the simplex to obtain $\beta_k^{(t)}$. Hyperparameters θ and φ control Markov chain smoothness.

D-LDA has been extended by other approaches to handle the temporal continuity of evolving topics in different ways [134, 135]. For instance, the discrete-time Dynamic Topic Model (dDTM) [136] discretisizes the data into time intervals to make evolving topics more representative of the contents of documents than D-LDA. On the other hand, Continuous-Time Dynamic Topic Model (cDTM) [137] handles any data point in time, regardless of the time resolution to decrease the complexity of variational inference for the dDTM grows quickly as time granularity increases.

Another important PDTM is Topics Over Time (ToT) [130], which relies on the discretization of time. In this model, each topic is associated with a continuous distribution over timestamps. For each generated document, the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp. Thus, the meaning of a particular topic can be relied upon as constant, but the topics' occurrence and correlations change significantly over time. ToT assumes time is inherently continuous and, therefore is characterized by a continuous distribution over timestamps [138].

Although PDTMs have been innovative tools for studying topic evolution [139, 140], they have several limitations. The first of these is scalability, as PDTMs become computationally expensive and time-consuming when dealing with very large datasets [141]. This complexity is influenced by several important factors, including the size of the vocabulary, the number of documents, the number of time periods (time granularity), and the number of topics. Notably, the size of the vocabulary plays a very important role in this regard [142]. There are efforts in the development of scalable PDTMs, such as [143], which extends the class of tractable priors from Wiener processes to the more general class of Gaussian processes. This approach allows the

model to be applied to large collections of text and to explore topics that evolve smoothly over a long period.

In addition, PDTMs cannot fully capture the diversity of topics over time because they assume that topics remain relatively unchanged over time [144]. Therefore, topics are consistent across the dataset and one may not be able to detect fine-grained changes in topics or subtle variations in topic representation over different time periods [145]. Several generative probabilistic topic models attempt to overcome these problems [146–149].

5.3 Neural Dynamic Topic Models

Neural Dynamic Topic Models (NDTMs) use neural networks to create a fixed-length vector representation by formalizing the temporal probability distributions of words and documents [150]. While preserving significant words in the topic descriptions, these methods generate dense clusters of interpretable documents[151]. Despite attempts to solve the problems with PDTMs (scalability, low evolution diversity), PDTMs lack coherence and diversity in representing topics when compared to the methods that take advantage of neural networks [16, 152, 153].

DTM is combined with word embeddings in Dynamic Embedded Topic Models (D-ETM) [41, 154] to improve the performance of topic models by providing a more informative representation of words. D-ETM analyzes time-series documents by changing the topics over time. D-ETM runs ETM for each time period in the data set, passing parameters into the next time period like in D-LDA. The generative process of documents is described as follows:

- 1. For each time step $t \in \{1, \ldots, T\}$:
  - Draw topic distribution for each topic k:

$$\alpha_k^{(t)} \sim \mathcal{LN}(\eta_{t_d}, \gamma^2 \mathrm{I}) \text{ and } \beta_k^{(t)} = \mathrm{softmax}(\rho^T \alpha_k^{(t)}) \tag{5.4}$$

  - Draw topic proportion mean:

$$\eta_t \sim \mathcal{N}(\eta_{t-1}, \delta^2 \mathrm{I}) \tag{5.5}$$

- 2.For each document index $d \in \{1, \ldots, D\}$:
  - Draw topic proportion: $\Theta_d \sim \mathcal{LN}(\eta_{t_d}, \gamma^2 \mathrm{I})$
- 3. For each word index $n \in \{1, \ldots, N_d\} \in d$:
  - Draw topic assignment: $z_{d,n} \sim \mathrm{Cat}(\Theta_d)$
  - Draw word: $w_{d,n} \sim \mathrm{Cat}(\beta_{z_{d,n}}^{(t_d)})$

where $\alpha_k^{(t)}$ and $\beta_k^{(t)}$ represent the topic embedding and word distribution for the k^{th} topic at the t^{th} time step, respectively. The hyperparameters γ, η, and ρ govern the variance of the normal distributions in the model.

D-ETM has been criticized for its time-consuming nature and for producing topic representations with lower diversity and coherence compared to recent models [127]. To address similar concerns, [155] propose an amortized variational inference with self-normalized importance sampling approximation to the word distribution that dramatically reduces the computational complexity and the number of variational parameters in order to handle large vocabularies.

Another NDTM model is the Graph-based Dynamic Topic Model (GDTM) [156] that offers a scalable solution for dynamic topic modeling, particularly in social media contexts. This model

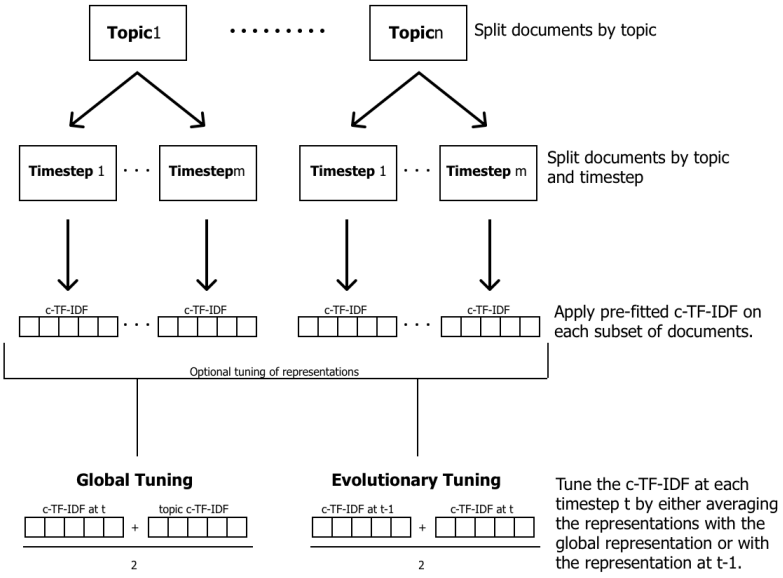


Figure 5.3: *BERTopic process for Dynamic Topic Modeling [16]*

assigns documents to topics by analyzing the overlap in their graph representations and partitions them based on graph density. A limitation of GDTM is that it focuses on outputting document partitions and does not provide the most probable words for each topic. Observe that topic representations can also be computed a posteriori as in Top2Vec and BERTopic (Chapter 2).

BERTopic [16] is a prominent clustering-based model that introduces dynamic topic modeling through a static process, as explained in Chapter 5, while representing topics dynamically over time. BERTopic first computes low-dimensional semantic document embeddings with the Bert language representation model [15] (Embedding) and the UMAP dimension reduction algorithm [47]. As shown in Figure 5.3, it then applies a static clustering technique to extract topic document clusters (Clustering) and segments each document cluster along an overlapping temporal time window (Segmentation) before computing the term vector representation (Representation) with c-TF-IDF [55]. The transformation of static document clusters into dynamic topics through segmentation has an important drawback. While documents may be categorized together under the same global topic, a more fine-grained analysis reveals that they often belong to distinct sub-topics within specific temporal frames. This global clustering followed by segmentation typically generates only one sub-topic for each period, which can be less precise than the multiple sub-topics that would be obtained through a period-wise topic generation process [157]. This problem hampers the ability to capture the temporal variations of dynamic topics, including changes in the number of document clusters or the size of each cluster, as topics emerge and fade over time. Moreover, this limitation diminishes the explainability and the interpretability of evolving topics within specific time periods [158]. This is mainly due to the fact that the representation is derived only from the content of documents assigned to a set within a given time period [159]. Our dynamic topic model presented in Chapter 6, aims to better capture the temporal variations in document clusters, to enhance the explainability and interpretability of evolving topics, and to resolve the issues related to documents being incorrectly categorized under a single global topic.

5.4 Conclusion

In this chapter, we have explored the concept of dynamic topic models, detailing their various types and examining their strengths and challenges. We highlighted their ability to track the evolution of topics over time and to reveal hidden patterns in temporal data. Additionally, we discussed the computational demands and complexities associated with modeling dynamic and multifaceted data. By categorizing and scrutinizing these models, we aimed to provide a comprehensive understanding of their advantages and limitations, setting the stage for further advancements in the analysis of evolving scientific content. The discussion also revealed specific limitations in current dynamic topic modeling approaches, such as difficulties in accurately capturing temporal variations and managing large vocabularies. The next chapter will address these issues by proposing a novel dynamic topic model. This new model aims to enhance the accuracy of topic segmentation over time, improve the explainability and interpretability of evolving topics, and resolve problems related to the misclassification of documents across different temporal frames.

Chapter

6 Aligned Neural Topic Models

Dynamic Topic Models (DTMs) have become crucial tools for understanding the evolution of topics over time in large text corpora. As explained in Chapter 5, DTMs can be broadly categorized into Probabilistic Dynamic Topic Models (PDTMs) and Neural Dynamic Topic Models (NDTMs). Each category has its unique strengths and limitations, particularly in terms of scalability and computational efficiency. Probabilistic Dynamic Topic Models (PDTMs) have been widely adopted for analyzing topic evolution due to their robust statistical foundations. However, when applied to extensive archives with vast vocabularies, PDTMs face significant computational challenges. These challenges primarily stem from the necessity to sample from complex posterior distributions, making PDTMs less scalable and more computationally expensive as the dataset size increases. In contrast, Neural Dynamic Topic Models (NDTMs) have emerged as a promising alternative, leveraging neural network architectures to enhance scalability and efficiency.

In this chapter, we introduce ANTM, a novel family of dynamic topic models designed to capture topic evolution with advanced algorithms. We conduct a comprehensive evaluation of ANTM, assessing the diversity and interpretability of the evolving topics it generates, and compare these results with state-of-the-art dynamic topic models, including recent methods such as D-ETM [154] and BERTopic [16], as well as traditional probabilistic dynamic topic models like TOT [130] and D-LDA [6]. Additionally, we perform a configuration analysis of ANTM to compare runtime and quality scores across different settings, providing valuable insights into the performance of various configurations. Finally, we offer a qualitative analysis of ANTM, discussing its limitations and proposing future research directions to enhance its capabilities.

This work has been presented at BDA'2023 and published in Rahimi, H., Naacke, H., Constantin, C., Amann, B. (2024). *ANTM: Aligned Neural Topic Models for Exploring Evolving Topics.* In: Hameurlain, A., Tjoa, A.M., Akbarinia, R., Bonifati, A. (eds) Transactions on Large-Scale Data- and Knowledge-Centered Systems LVI. Lecture Notes in Computer Science, vol 14790. Springer, Berlin, Heidelberg.

Chapter content

6.1 Introduction

Among the state-of-the-art NDTMs, BERTopic is noteworthy for its innovative approach to topic extraction. Despite their advantages, NDTMs, including BERTopic, are not without limitations [157]. One critical issue arises during the transformation of static document clusters into dynamic topics through segmentation. This process can inadvertently group documents under the same global topic, even when they belong to distinct topics within specific temporal frames [159]. This misclassification poses a significant drawback, especially in scenarios where capturing the temporal nuances of topic evolution is crucial. The resultant clusters may fail to accurately reflect the dynamic nature of topics, thereby impeding the model's ability to monitor changes in the number of document clusters or the size of each cluster as topics emerge and dissipate over time. These limitations highlight the ongoing need for advancements in dynamic topic modeling techniques to better accommodate the complexities of temporal topic variation. To address these challenges, we introduce a new clustering-based dynamic topic model, called the Aligned Neural Topic Models (ANTM).

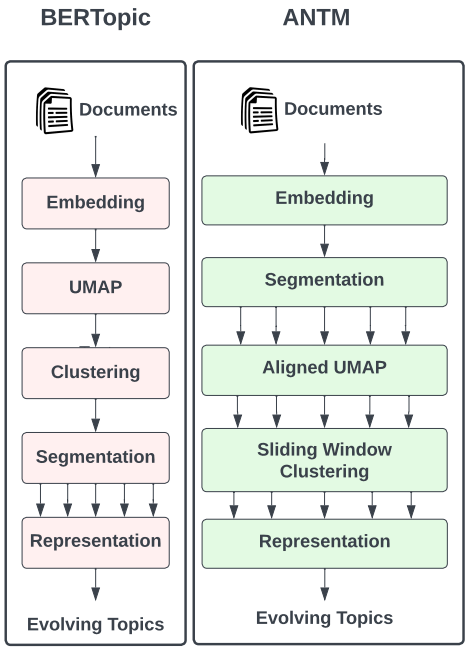


Figure 6.1: *BERTopic and ANTM*

6.2 ANTM

As depicted on the right of Figure 6.1, ANTM employs a process similar to BERTopic but in a different sequence. Specifically, it performs the temporal segmentation step on the document embeddings before the dimensionality reduction step, utilizing Aligned UMAP. The document cluster segments are then reassembled along the time dimension through Sliding Window Clustering to generate dynamic topic document clusters. This approach allows ANTM to consider temporal changes in document content, producing a range of high-quality topics for each period. By implementing an overlapping sliding window algorithm, ANTM can identify collections of related

topics that span multiple periods, yet remain distinct enough to demonstrate evolution within a single domain. Our experiments conducted on three datasets and four dynamic topic models show a considerable improvement in coherence and diversity scores compared to state-of-the-art PDTMs and ADTMs. ANTM offers a significant advancement in the field of topic modeling. This work not only enhances our ability to track and understand topic evolution over time but also sets the stage for future innovations in this domain. As illustrated in Figure 6.2, the ANTM architecture consists of three layers. The first layer leverages advanced pre-trained language models (LLMs) to generate vector representations for each document, capturing its content (Document Vectors). The second layer, referred to as the SWS Splitter, subdivides the document vectors into temporally overlapping subsets (Temporal Document Vectors) and applies the AlignedUMAP algorithm [160] to produce low-dimensional document vector embeddings with global coherence for each subset. Subsequently, the hierarchical density-based clustering algorithm HDBSCAN [161] is applied to each overlapping subsets to generate topic clusters for each time period, which are then aligned to produce dynamic topic clusters (aligned subsets of clusters over all time periods). In the third layer, the Topic Representation Layer, we introduce a contextualized LLM-based Nearest Words approach to generate word representations for the aligned topic clusters. This layer can also incorporate class-based TF-IDF [55]. The following subsections provide a detailed definition and implementation of each layer.

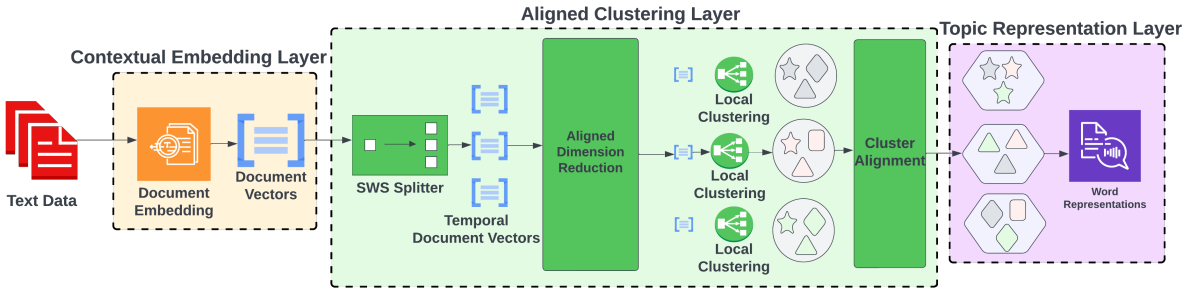


Figure 6.2: *Architecture of ANTM*

6.2.1 Contextual Embedding Layer (CEL)

The Contextual Embedding Layer (CEL) is responsible for providing a vector representation for the document d in the corpus D. More formally, the embedding vector y for a document d is a mapping $CEL : d \mapsto y \in \Re^z$, capturing contextual and semantic information from the corpus D. The document embedding represents words and documents in a low-dimensional feature vector space, where the embedding dimension z is expected to be much smaller than the size of the vocabulary (i.e., the number of unique words in D).

Implementation

CEL takes advantage of pre-trained LLMs (e.g., Data2Vec, GPT-4) to compute a contextualized time-aware vector representation for each document. Pre-trained LLMs capture the meaning and context of a document by using an attention mechanism to consider the context of words in a document. As shown in Equation (6.1), given an input document d, the number u of tokens in d, and h_j the hidden state vector of the j-th token in d, the embedding $y = \text{CEL}(d)$ of the whole document d is then obtained by taking the mean of all the hidden state vectors of the tokens in d.

$$CEL : D \mapsto Y = [\mathbf{h}_z]_{z=1}^{u} = \frac{1}{u} \sum_{z=1}^{u} \mathbf{h}_z \tag{6.1}$$

We argue that LLM-based embeddings implicitly capture the temporal context of the document in addition to capturing its meaning. Indeed, the attention mechanism weights the importance of each word in the context of a document. This context implicitly depends on the publication period of the document since word contexts happen to vary over time.

6.2.2 Aligned Clustering Layer (ACL)

This layer is motivated by the belief that topic evolution occurs in cycles of steps as explained by [162]. Therefore, we discretize these steps by time frames with intersections similar to [6, 131], but we use time-aware algorithms to account for the continuity of time and to model evolving topics that emerge and fade over time. After performing transformer-based document embedding on the corpus D, we can obtain a dynamic vector representation of D by applying a Sliding Window Segmentation (SWS) process that divides D into a series of n overlapping time frames $\{W^1, \ldots, W^n\}$. We then denote the set of documents published in the time frame W^t by $D^t \subseteq D$ and the set of embeddings of D^t by $Y^t = \{CEL(d) | d \in D^t\}$. This layer aims to discover evolving document clusters by sequentially grouping similar documents. This procedure is called Partitioned Clustering (PC).

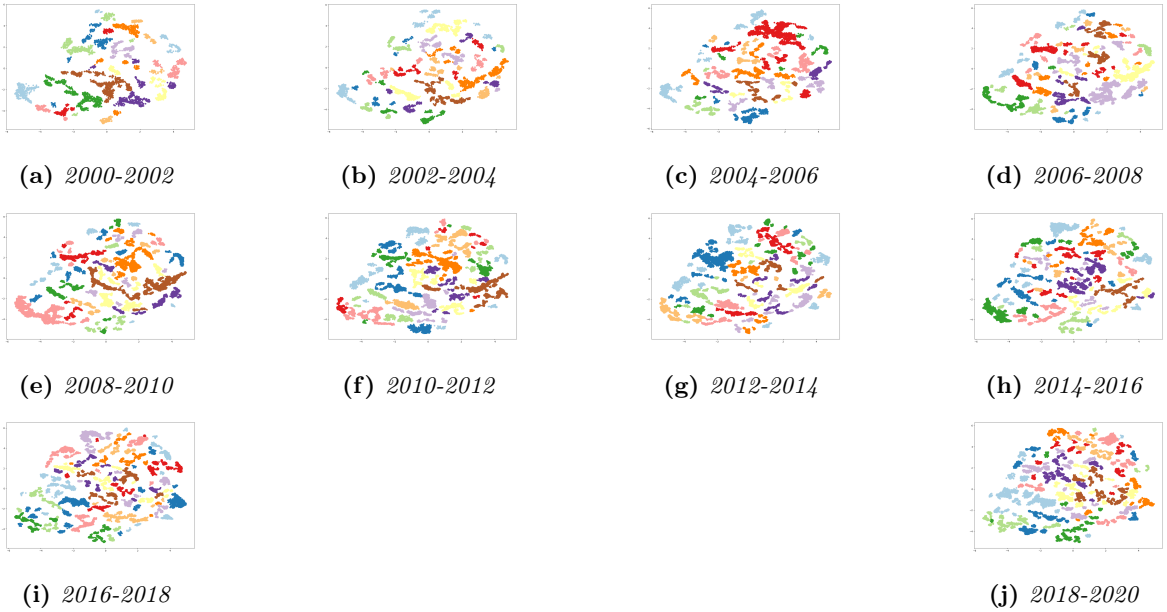
Definition 6.2.1. Let D^t be a set of documents in the t-th time frame of the document corpus D and let Y^t denote the set of embeddings of D^t. Partitioned Clustering $PC : D^t \mapsto \{D_i^t\}$ clusters the documents D^t by their embeddings Y_t and returns a set of *local* document clusters $D_i^t \subseteq D$ for $i = 1$ to k_t where *all document embeddings in* Y_i^t *are similar.* Y_i^t *is the set of embeddings of the documents contained in cluster* D_i^t.

Figure 6.3 shows two-dimensional document embedding vectors clustered by their cosine distance for different time periods. The sequence of figures reveals evolution patterns and trends of document clusters (topics) that may not be apparent when documents are grouped based on their content. The documents from the DBLP dataset are embedded using BERT and sequentially clustered in each time frame. The clusters of all periods are aligned to create evolving topics as described in Figure 6.4.

To obtain a representation of the temporal evolution of these clusters, we apply a final step that aligns the document clusters along all periods using a cluster linkage measure. We can use different cluster linkage measures (single, average, centroid, complete) to estimate the similarity of the vector embedding clusters Y_i^t for local document clusters D_i^t. The result of each alignment is a set of *evolving document clusters* D_k (EC) that contain semantically similar documents from different time periods :

Definition 6.2.2. Let $\{D_i^t\}$ be the set of local document clusters obtained by PC. Cluster Alignment $CA : D \mapsto \{D_k\}$ generates m subsets of documents $D_k \subseteq D$ for $k = 1$ to m and where D_k is the *union* of a set of local document clusters (topics) D_i^t with *similar embedding clusters* Y_i^t.

Definition 6.2.3. Let $\{D_i^t\}$ be the set of local document clusters obtained by PC. Cluster Alignment $CA : D \mapsto \{D_k\}$ generates m subsets of documents $D_k \subseteq D$ for $k = 1$ to m and where D_k is the *union* of a set of local document clusters (topics) D_i^t with *similar embedding clusters* Y_i^t. Figure 6.4 shows the evolving clusters after the alignment.



(a) *2000-2002* **(b)** *2002-2004* **(c)** *2004-2006* **(d)** *2006-2008*

(e) *2008-2010* **(f)** *2010-2012* **(g)** *2012-2014* **(h)** *2014-2016*

(i) *2016-2018* **(j)** *2018-2020*

Figure 6.3: *Partitioned Clusters*

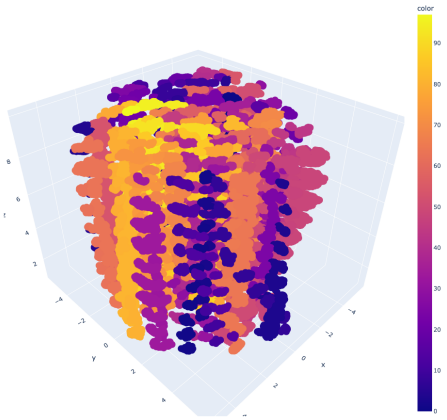


Figure 6.4: *Evolving Clusters*

Implementation

The Aligned Clustering Layer (ACL), similar to the other two layers of ANTM, can be implemented using different clustering and alignment methods combined with scalable algorithms. As shown in Figure 6.2, we implemented the ACL layer with density-based clustering techniques [163] to cluster data points (embedding vectors) based on their distribution in the feature space. These algorithms are particularly useful for identifying clusters of varying densities and arbitrary shapes. The number of obtained clusters is defined by the parameters used to estimate the density. This feature provides an advantage over non-parametric PDTMs that require a fixed number of topics as input. The Aligned Clustering Layer is developed in the following three steps.

Aligned Dimension Reduction To overcome the curse-of-dimensionality problem of similarity-based clustering methods on high-dimensional data, we first reduce the size of document embedding vectors. The proposed solution uses AlignedUMAP [164] to align sequences of UMAP [47] embeddings based on the documents shared between two consecutive time periods. These

overlapping documents act as landmarks for aligning the different vector spaces defined for each period and allow us to obtain globally coherent reduced-dimension vector representations.

Sequential Local Clustering After the aligned dimension reduction, a density-based clustering algorithm is performed on each period data frame to separate all documents of a given period into a set of local document clusters. We use HBDSCAN [161], a hierarchical version of DBSCAN [165] which does not require a predefined number of clusters but generates and selects clusters based on a notion of stability (the hierarchical clustering process slices the cluster hierarchy so that the number of clusters is as close as possible to that of the next level in the hierarchy, while constraining the cluster size to discard clusterings with too large and too few clusters). As shown in Figure 6.3, each such local cluster describes a concept within a time frame.

Cluster Alignment The following step involves aligning consecutive local clusters that were achieved in the previous step. As the document embedding vectors of all time frames are in the same space, we can aggregate each cluster by considering their centroids. We use the centroid linkage method [166] to align document clusters from different periods by clustering the cluster centroids using HDBSCAN. An example of the output for this step is shown in Figure 6.4. All documents within an evolving cluster share the same color and are the results of a local clustering step for each time period (x-y plane) followed by a cluster alignment step along time (z-axis). We can see that aligned clusters that have the same color are rather vertically aligned through consecutive time frames ("slices"). The representation of these Evolving Clusters offers valuable insights into the progression of topics over time, as depicted in Figure 6.9 and Figure 6.8. Figure 6.8 highlights a noteworthy shift in news coverage, from focusing on Boko Haram in Nigeria in 2015, to ISIS in Iraq and Baghdad in 2016, and later to events in Kabul, Afghanistan, involving the Taliban between 2017 and 2019. Meanwhile, Figure 6.9 demonstrates the emergence of terms like CNN, Lung, and Diagnosis between 2018 and 2020, aligning with the COVID-19 pandemic and advancements in computer vision research. Another alternative for HDBSCAN is to use the K-Nearest Neighbors (KNN) clustering method based on cosine similarity.

KNN clustering allows us to introduce a similarity threshold parameter to determine the strength of the alignment links. A high threshold results in fewer alignments, highlighting topics that barely change over time. On the other hand, a low threshold produces more alignments, including topics with significant changes over time and lower similarity values. This enhances the versatility of the topic alignment step and helps to improve the representation quality of evolving topics. It also helps to avoid aligning topics of the same time periods and to focus only on aligning the topics of each time period with their adjacent time periods. Figure 6.5 illustrates an evolving topic about the Ebola outbreak using HDBSCAN. We can see that the topic representation (defined in Section 6.2.3) changes slightly over time, and most of the topic words are stable and repeated in the next time period. At time period 1, every word is considered as new (green color). For the following periods, a word contained in the previous period and the next one is considered as stable (grey color), and a word not in the next period is considered as disappearing (red color). However, if we align topics using KNN which is more flexible than HDBSCAN, we can obtain a richer evolving topic containing a larger set of words about various diseases, see Figure 6.6 where the similarity threshold has been set to 0.6 which is rather low. If we increase the similarity threshold from 0.6 to 0.8, the alignment in Figure 6.6 splits into multiple smaller alignments that showcase the evolution of specific diseases, and the Ebola outbreak is one of them (Figure 6.7). This increases the versatility of the topic alignment process, which can be configured based on application and user preferences.

Figure 6.5: *Evolving Topic regarding Ebola Outbreak based on HDBSCAN topic alignment.*

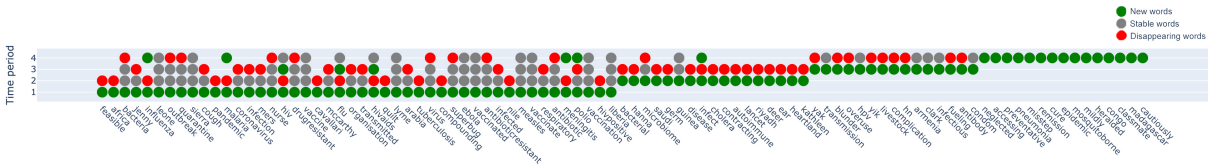


Figure 6.6: *Evolving topic about diseases and outbreaks- Alignment using KNN (0.6 unit threshold)*

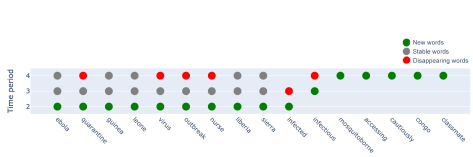
6.2.3 Representation Layer

The Representation Layer (RL) is responsible for generating word representations for each local topic document cluster in each time frame. The goal of this layer is to identify the most relevant terms or term phrases to summarize the main ideas or themes of each document cluster.

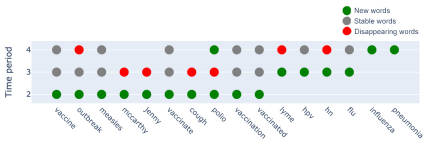
Definition 6.2.4. The topic Representation Layer (RL) computes a list of m terms $\{t_{ij}^r\}_{r=1}^m$ that describe the semantic contents of each document cluster D_{ij}.

Implementation

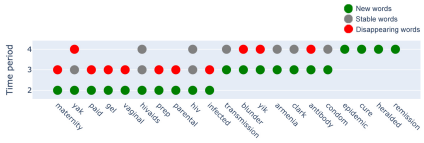
There exist various ways to represent a set of documents by a set of terms [167]. One way is to use a method called class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) [16, 55],



(a) *Evolving topic 1 - Ebola Outbreak*



(b) *Evolving topic 2 - Measles Outbreak*



(c) *Evolving topic 3 - HIV*

Figure 6.7: *Topic alignment using KNN (0.8 unit threshold)*

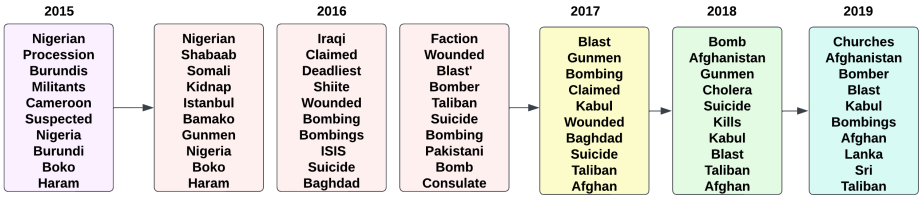


Figure 6.8: *Evolution of New York Times News on Foreign Terrorist Activities.*

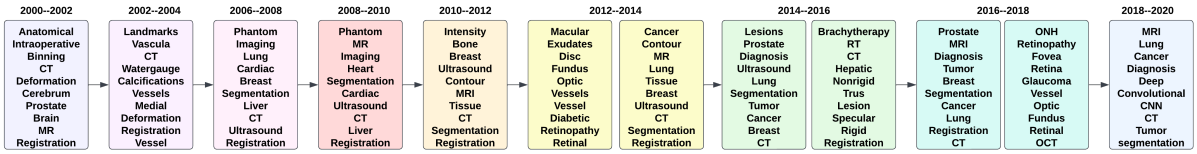


Figure 6.9: *Evolution of Computer Science Research on Medical Science based on DBLP documents.*

which is a variation of TF-IDF that takes into account the class labels of documents. c-TF-IDF weights the terms in a cluster not only by their frequency but also by their relevance to a particular cluster. c-TF-IDF ensures that each word of a topic representation occurs in at least one document of that topic. Examples of outputs for this step are shown in Figure 6.9 and Figure 6.8. We propose to use an LLM-based Nearest Words approach similar to [24] that computes joint word and document embeddings over a document cluster and finds the m closest words from the centroid of the document embeddings. LLM-based Nearest Words generates a vector representation for each document by averaging its word vectors. Subsequently, it derives a contextualized vector representation for each word through max-pooling across all available vectors for that word across all documents. Max pooling captures the diversity of various contexts of a given word across different dimensions. As shown in Figure 6.10, the proposed topic representation method achieves higher median coherence compared to c-TF-IDF and the non-contextualized Nearest Words representations, which use generic word embeddings without considering the context of the words in each document, as proposed in [24].

6.3 Experiments

We compare ANTM with four other dynamic topic models on three datasets. Our experiments aim to demonstrate ANTM's performance in terms of topic coherence and diversity relative to state-of-the-art competitors and to illustrate its capability for exploratory topic evolution analysis.

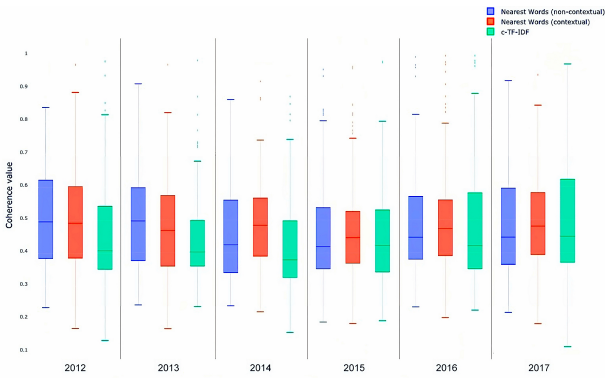


Figure 6.10: *Coherence value distributions of non-contextual Nearest Words, contextual Nearest Words, and c-TF-IDF*

Table 6.1: *Datasets*

Dataset	Documents	Tokens	Vocabulary	Date Range
DBLP	200K	18.5M	144K	2000-2020
arXiv	70K	1.17M	12.6K	2000-2022
NYT News	210K	5M	61K	2012-2022

Besides, we provide a comparative study that examines ANTM under different settings for each layer in different scenarios.

6.3.1 Datasets

We use three datasets in the experiments. The first dataset is the DBLP [168] archive of 200K scientific articles (title and abstract) published between 2000 and 2020. DBLP is an open bibliographic database, search engine, and knowledge graph that archives computer science publications. The second dataset is a 70K sample of documents (title and abstract) extracted from arXiv [169] and published between 2000 and 2022. The last dataset is a collection of 210K articles [170] from the New York Times (NYT) published between 2012 and 2022. The NYT dataset includes historical and current articles on a variety of topics and includes text, images, and metadata such as the article's headline, author, and publication date. The statistics of the datasets are summarized in Table 6.1.

6.3.2 Baseline models

The performance of ANTM is compared to four other dynamic topic models. The first model is DTM [6], which has been widely used for various applications such as public opinion tracking [171]. The second model is D-ETM [154], which extends DTM with word embeddings to improve topic quality and performance. The third topic model is TOT [130], which is a variation of LDA and captures how the co-occurrence patterns of words change over time. The last model is BERTopic [16], a clustering-based topic model that uses a static process for clustering document embeddings while dynamically representing topics.

6.3.3 Evaluation Metrics

We evaluate the performance of each topic model in terms of human interpretability using Topic Coherence (TC) [99, 120, 172] and its diversity by Topic Diversity (TD) [154, 173].

Topic Coherence (TC) TC, as defined in [99] with co-occurrence value C_V, is a variation of Normalized Point-wise Mutual Information (NPMI)[102] that averages the co-occurrence values of word pairs (t_i^r, t_i^s) in all topics i over N topics.

Topic Diversity (TD) For this metric, we use the Proportion of Unique Words (PUW) method [41] defined as the vocabulary size divided by the total number of words within a set of topics. Topics within a low-diversity topic set share many words, whereas a highly-diverse topic set contains topics that have few words in common.

Table 6.2: *Segmentation setting*

Dataset & Time Frame Sizes	Length	Overlap	#Frames
DBLP Documents	3 Years	1 Year	10
arXiv Documents	4 Years	1 Year	10
NYT News	4 Years	1 Year	10

6.3.4 Experimental Setup

ANTM We used two embedding models from SBERT [52] in the contextual embedding layer (CEL) of ANTM. The two configurations with these models are called ANTM-large (using *all-mpnet-base-v2*) and ANTM-mini (using *all-MiniLM-L6-v2*). We also used the Data2Vec pre-trained model (*facebook/data2vec-text-base*) [174] in our study. We then split the document embedding vectors into a series of time frames to explore the change in document contents. The setting of the segmentation step is summarized in Table 6.2. These segmentation values, as suggested in [175], provide a comprehensive view of the data and ensure that changes over time are captured in the analysis.

Afterward, we chose the hyperparameters for dimension reduction (using AlignedUMAP) and sequential document clustering (using PC). As suggested in [16, 24, 160], the cosine similarity metric, with 5-dimensional output, was chosen for the dimensionality reduction setting of AlignedUMAP. We then performed HDBSCAN on each time frame with Euclidean distance and a minimum size of 10 documents per cluster to create a set of semantically similar document clusters for each period (Figure 6.3). Since all document embeddings are in the same vector space, we could then align the generated clusters by again using HDBSCAN on the centroid of all document clusters with Euclidean distance and a minimum number of 2 clusters to obtain evolving clusters for each dataset (Figure 6.4). Finally, we represented the documents of each cluster with a set of $m = 10$ words using the c-TF-IDF (Figures 6.8 and 6.9). The common parameters and methods are chosen similarly to BERTopic for fair comparison.

Baselines DTM, D-ETM, and TOT were configured with 20, 50, and 100 topics respectively, and applied to the titles and abstracts of the DBLP, arXiv, and NYT News datasets. The topic numbers were determined based on the number of topics generated by ANTM in an unsupervised manner. Additionally, we ran BERTopic using the same sentence transformer models as the proposed model for a fair comparison. The default configurations were used for the rest of the hyperparameters of BERTopic.

6.3.5 Results

The performance of dynamic topic models can be analyzed from two perspectives. First, we observe the quality of topics in terms of coherence and diversity within each time frame (period-wise analysis). This aims to assess the ability of dynamic topic models to describe temporal topics in a diverse and coherent manner. The second perspective analyzes the quality of word representations within each evolving topic (topic-wise analysis) and compares the ability of dynamic topic models to represent the evolution of each dynamic topic over time.

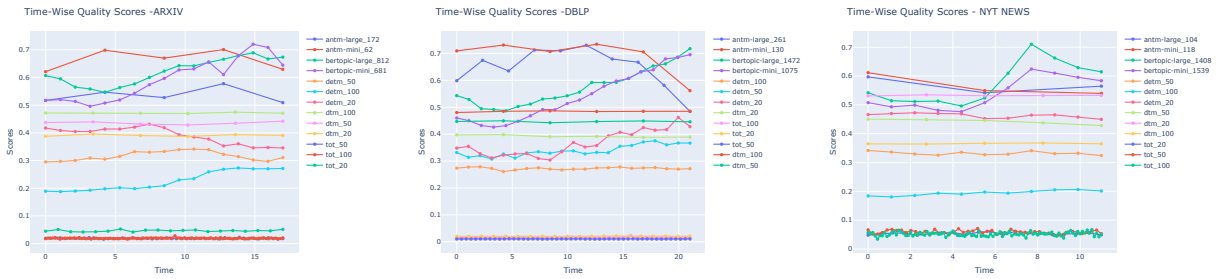
As shown in Table 6.3, ANTM achieves the highest Topic Quality (TC×TD) scores among the different baseline variants in all three datasets. However, ANTM is 2.8 to 4.5 slower than BERTopic depending on the used embedding model and due to the increased number of topic clustering steps. Yet, ANTM's runtime for producing 62 topics is respectively 8, 30, and 43 times

faster than its competitors when generating 50 topics.

Table 6.3: *Performance comparison of ANTM and baselines*

		ArXiv					DBLP					NY Times				
Model	Embedding	#T	TC	TD	TQ	t(s)	#T	TC	TD	TQ	t(s)	#T	TC	TD	TQ	t(s)
DTM	-	20	0.56	0.71	0.40	16194	20	0.58	0.68	0.40	15485	20	0.57	0.95	0.54	3071
	-	50	0.61	0.72	0.44	23897	50	0.61	0.74	0.45	33541	50	0.46	0.98	0.45	6535
	-	100	0.61	0.79	0.48	46195	100	0.64	0.77	0.49	61140	100	0.38	0.99	0.37	13255
TOT	-	20	0.33	0.15	0.05	14071	20	0.33	0.07	0.03	20385	20	0.42	0.13	0.06	2680
	-	50	0.32	0.07	0.02	15833	50	0.33	0.04	0.02	45295	50	0.41	0.15	0.06	6490
	-	100	0.32	0.07	0.02	29205	100	0.33	0.05	0.02	91725	100	0.38	0.14	0.06	12146
D-ETM	W2V	20	0.45	0.88	0.40	2925	20	0.44	0.85	0.37	12352	20	0.48	0.98	0.47	8257
	W2V	50	0.47	0.69	0.32	4625	50	0.49	0.71	0.34	19332	50	0.43	0.79	0.34	12907
	W2V	100	0.49	0.48	0.23	6840	100	0.45	0.61	0.28	36471	100	0.38	0.52	0.20	20133
BERTopic	Mini	681	0.68	0.88	0.60	**192**	1075	0.70	0.78	0.54	**342**	1539	0.60	0.91	0.54	**340**
	Large	812	0.70	0.89	0.63	**90**	1472	0.71	0.81	0.58	**311**	1408	0.63	0.92	0.57	**143**
ANTM	Mini	62	0.72	0.94	0.67	**544**	130	0.77	0.91	0.70	**1398**	118	0.62	0.93	0.57	**1549**
	Large	172	0.59	0.92	0.54	**405**	261	0.70	0.93	0.65	**1869**	104	0.63	0.92	0.57	**1660**

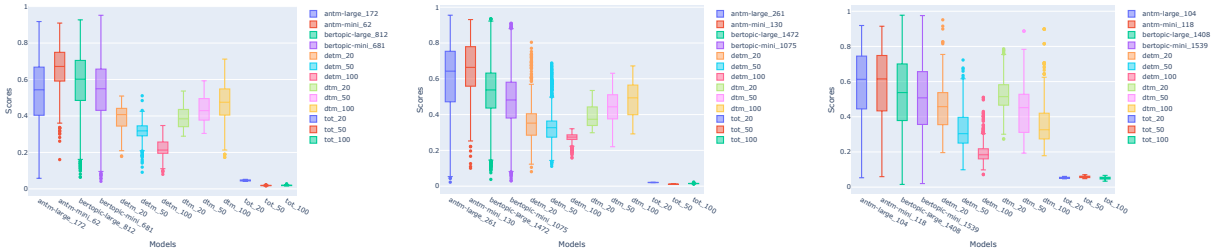
Period-Wise Quality Comparison The goal of the period-wise analysis is to determine whether aligned document clustering, applied separately for each time frame, results in better topic quality per period. As shown in Figure 6.11, ANTM overall has higher quality scores compared to D-ETM, DTM, and TOT in all three datasets. In contrast to BERTopic, which exhibits ascending values across consecutive time frames, our proposed method maintains a stable consistency. The variability observed in BERTopic's performance stems from its global clustering approach with uniform parameters, resulting in uneven document distributions within each cluster. It is noteworthy, though, that the quality of ANTM experienced a decrease in the last time frame. We argue that this decrease can be attributed to the data scarcity for the last time frame within the datasets. Moreover, we observe that DTM generates topics with higher consistency compared to D-ETM, even without the use of embedding models. Lastly, it is worth noting that TOT produces the lowest-quality evolving topics across all the datasets.



(a) *ArXiv* **(b)** *DBLP* **(c)** *NYT News*

Figure 6.11: *Period-wise Quality Comparison*

Topic-Wise Quality Comparison The objective of topic-wise analysis is to assess whether an evolving topic, which contains topics that are semantically close to each other, can nevertheless effectively represent the evolution of the represented document clusters. As shown in Figure 6.12, topic quality (TQ) is calculated for each evolving topic, and their distributions are plotted across each model on all three datasets. These figures lead to various observations. First, as depicted in Figure Figure 6.12 (b) and (c), ANTM consistently generates the highest quality evolving topics in both the DBLP and NYT News datasets. However, in the case of ArXiv, BERTopic outperforms the ANTM-large model in terms of topic-wise quality, although it still falls short of the quality achieved by the ANTM-mini model. The second observation highlights that DTM consistently generates higher-quality topics compared to D-ETM, even without the use of embedding models.



(a) *ArXiv* **(b)** *DBLP* **(c)** *NYT News*

Figure 6.12: *Topic-wise Quality Comparison*

Table 6.4: *Variation Comparison of ANTM*

CEL	ACL				RL					
	SWS		Sliding Window Clustering		c-TF-IDF					
Embedding Model	Window length	Overlap length	AlignedUMAP #Neighbor	HDBSCAN Min Cluster	#Words	#T	TC	TD	TQ	t(s)
all-MiniLM-L6-v2	5	2	25	25	5	60	0.74	0.93	0.70	747
					10	60	0.61	0.89	0.55	752
	4	1	20	20	5	66	0.71	0.93	0.66	**474**
					10	66	0.56	0.90	0.51	**478**
all-mpnet-base-v2	5	2	25	25	5	75	0.74	0.92	0.68	740
					10	75	0.59	0.88	0.52	806
	4	1	20	20	5	80	0.68	0.92	0.63	505
					10	80	0.54	0.89	0.48	547
data2vec	5	2	25	25	5	10	0.40	0.65	0.26	801
					10	10	0.30	0.60	0.18	823
	4	1	20	20	5	8	0.39	0.77	0.30	503
					10	8	0.30	0.73	0.22	531

Lastly, it is worth noting that TOT consistently produces the lowest quality evolving topics across all the datasets.

Variation Analysis As shown in Table 6.4, we compare different configurations of ANTM, similar to an ablation study, by systematically varying the values of one or more algorithm parameters while keeping other aspects constant. By observing how changes in parameter values affect the performance of the model, we can identify the most influential parameters and their optimal settings. In this comparative analysis, we have selected three distinct language models for the contextualized embedding layer, allowing us to investigate the impact of pre-trained models. Additionally, we have explored two distinct configurations for the aligned clustering layer, facilitating a comparison between scenarios involving long time frames with substantial overlap and those with shorter time frames and reduced overlap. Furthermore, we have examined two distinct settings for the representation layer to assess the model's representational capacity concerning the number of words. The details of this comparative analysis are provided in Table 6.4. In our findings, it becomes evident that ANTM's performance significantly improves when utilizing BERT as the embedding layer. Moreover, employing larger time frames with higher overlap yields superior results, and intriguingly, a representation layer comprising five words outperforms one with ten words.

Qualitative Comparison The qualitative comparison in topic modeling is crucial for making sense of the results, refining models, validating them against domain knowledge, and ultimately extracting meaningful insights from textual data. While quantitative metrics provide valuable guidance, qualitative assessment is essential for ensuring the relevance and human interpretability of the topics generated by these models. As shown in Table 6.5, a qualitative comparison between ANTM and the baselines is provided. This comparison includes *Temporal Labeling* (the ability

Table 6.5: *Qualitative Comparison*

	DTM	TOT	D-ETM	BERTopic	ANTM
Temporal Labeling	✔	✗	✔	✔	✔
Non-parametric	✗	✗	✗	✔	✔
Semantization	✗	✗	Word2Vec	LLMs	LLMs
Coherence	Normal	Very Low	Low	High	Very High
Diversity	Low	Low	Normal	High	Very High
Runtime	Very High	High	Normal	Very Low	Low
Complexity	Low	Low	Normal	High	High
Scalability	Very Low	Very Low	Low	Normal	High
Adaptability	Very Low	Very Low	Low	High	High

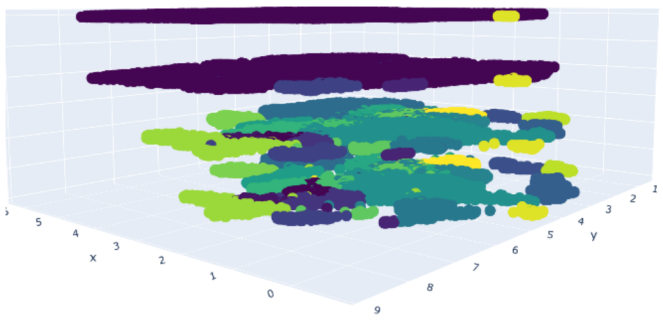


Figure 6.13: *Formation of Non-informative Clusters*

to represent each evolving topic corresponding to a time period), *Semantization* (the ability to understand the semantics behind the text representation), *Topic Coherence* and *Diversity* of labels as discussed in Table 6.3, *Runtime* (the computational efficiency of each model, considering the time required to process large datasets and generate topic models), algorithmic *Complexity* (including the number of hyperparameters and algorithmic complexity), *Scalability*, which assesses the ability to handle large corpora, and *Adaptability* with different algorithms concerning the data and the domain. These aspects are informed by our methodological understanding of different topic modeling approaches, which has allowed us to identify the key dimensions along which these models can be compared and contrasted. By considering these qualitative aspects, researchers and practitioners can gain a deeper understanding of the strengths and limitations of different topic modeling approaches, ultimately selecting the most suitable model for their specific use case.it

Limitation Initial experiments have shown that the hyperparameters need to be chosen carefully to avoid the formation of large topics that group together unrelated topics during certain time periods, as shown in Figure 6.13.

6.4 Conclusion

Existing dynamic topic models ignore certain temporal variations of evolving topics by configuring a global structure for dynamic topics, such as the same number of document clusters in each period. Furthermore, many of these models necessitate computationally expensive operations when dealing with large-scale corpora, which can be a significant bottleneck. These limitations directly affect the observation of topic evolution and reduce the coherence and diversity of evolving topics, which are sequentially represented. In this chapter, we proposed a family of dynamic neural topic models called Aligned Neural Topic Models (ANTM), which combines novel data

mining algorithms to provide a modular framework for discovering evolving topics. Based on a series of experiments on three distinct datasets we can conclude that ANTM outperforms the state-of-the-art dynamic topic models in terms of topic coherence and diversity. In Part III, we will leverage the insights gained from topic models and dynamic topic models to explore the evolution of topics and their impact on the progression of scientific knowledge. By analyzing how topics transform over time, we aim to uncover patterns in scientific advancement and discourse.

Part III

Science Evolution Modeling

> *An intellectual? Yes. And never deny it. An intellectual is someone whose mind watches itself. I like this, because I am happy to be both halves, the watcher and the watched. "Can they be brought together?" This is a practical question. We must get down to it. "I despise intelligence" really means: "I cannot bear my doubts.*
>
> Albert Camus

Motivation

In Parts I and II, we focused on the complexity of topic models, explored various evaluation metrics, and examined dynamic topic models. These foundational concepts have equipped us with a robust framework to comprehend how topics evolve over time. Building on this foundation, we now turn our attention to the analysis of scientific evolution, focusing on the different types and methodologies used to study how scientific ideas and disciplines develop and transform. This progression from understanding basic topic modeling to exploring the dynamic nature of topic evolution sets the stage for a deeper investigation into the patterns and trends that characterize the evolution of science.

Organization

In Chapter 7, we first explore topic evolution through various analysis tasks proposed in the literature. Following this chapter, we delve into the evolution of science in Chapter 8 by using a novel categorization approach for a more nuanced comprehension of scientific progression. Building on these studies, we contribute to the study of science evolution by proposing innovative frameworks in Chapters 9 and 10 that integrate dynamic topic models with citation networks and advanced deep learning techniques such as dynamic graph neural networks and reinforcement learning. These novel frameworks aim to offer more sophisticated and interconnected perspectives on the mechanisms driving scientific evolution.

Contributions

Our work presents innovative deep-learning methodologies for analyzing the evolution of topics within scientific archives. In Chapter 9, we propose a new definition of emerging topics and introduce a unique concept of citation context, which facilitates the early identification of these topics. In Chapter 10, we offer a novel definition of paradigm shift and use Q-Learning to model the transitions in the evolution of topics.

The first contribution has been presented at "BDA 2023: 39ème Conférence sur la Gestion de Données – Principes, Technologies et Applications" and published/presented in *Rahimi, H., Naacke, H., Constantin, C., Amann, B. ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives. In: Cherifi, H., Rocha, L.M., Cherifi, C., Donduran, M. (eds) Complex Networks & Their Applications XII. COMPLEX NETWORKS 2023, Studies in Computational Intelligence, Springer.* Additionally, a demonstration paper on ATEM will be presented at BDA 2024. The second contribution is in preparation for submission.

Chapter

7 Topic Evolution Models and Analysis

Topic evolution analysis has many applications and, before focusing on the evolution of research topics within scientific archives, it is crucial to study the related work in other areas like trend analysis in social media or event tracking in news media. This comprehensive understanding serves as a guide for discovering and investigating the evolution of science, enabling us to draw on methodologies and findings from diverse areas to inform our analysis of scientific progress.

Chapter content

7.1 Introduction

Topic evolution refers to the modeling and discovering how topics change, emerge, or fade over time from time-stamped document collections [11] including academic publications [176], news articles, social media posts [177, 178].

The underlying topic models and analysis tasks vary depending on the underlying application and data. In social networks, topics often manifest as simple terms such as hashtags, which capture the essence of trending discussions. In news media, topics are typically represented by events or entities, leading to the development of event-based topic models that can track news stories and key figures over time. Conversely, in scientific articles, topics are more complex, represented by weighted term vectors that encapsulate specific research topics. Topic analysis tasks include topic trend detection, topic emergence analysis, sentiment analysis, and semantic shift detection. Each domain necessitates a tailored approach to topic modeling, reflecting the unique nature and requirements of the information being analyzed.

It is important to distinguish between dynamic topic models (DTMs, Chapter 5) and topic evolution models. Dynamic topic models identify and track topics over time without being tied to a specific analysis task, providing a broad understanding of how topics evolve in a corpus. In contrast, topic evolution models are designed for targeted analysis, focusing on specific questions or tasks related to topic changes. These models can leverage the evolving topics generated by DTMs to inform their analyses, allowing for a more directed exploration of how topics develop in relation to particular objectives.

In the state of the art, we have identified four primary tasks that are commonly investigated in relation to topic evolution. These tasks include tracking the emergence and development of topics over time, understanding the shifts in topic prevalence, detecting the fusion or splitting of topics, and identifying the underlying factors driving topic changes. Each task addresses a distinct aspect of how topics evolve, highlighting the multifaceted nature of topic evolution modeling and its application-specific requirements. The most important tasks are explained as follows.

7.2 Topic Trend Analysis

Topic trend analysis has been one of the first kinds of analysis conducted to understand topic evolution. Topic trend analysis examines the temporal evolution of various topic properties, such as popularity or interest (measured by "Likes" in social media) and importance or utility (indicated by the number of citations) within a document stream. Topic trend analysis has been used in various scenarios and applied to a variety of document streams, such as newspaper [179], social media [178] and US Patent database [180] to detect emerging trends in scientific literature [181].

Manual trend analysis In the 1970s, identifying trends in topics involved meticulously analyzing the content of published articles [182]. Researchers manually defined topic areas by closely examining article contents, and summarizing trends by tallying the number of articles on a particular topic over time. Although expert-driven content analysis generally produces high-quality results, the overwhelming volume of today's publications makes it impractical to rely solely on human annotation for summarizing or discovering new topics. Topic trends can be analyzed and measured in several ways and we will discuss some of the most important methods below.

Counting the number of documents Most research on tracking topic trends over time has followed a straightforward approach of counting the number of documents associated with each topic on an annual basis [183–185] and analyze how the topic-document distribution evolves over time. Although this method offers a broad perspective, it is crucial to acknowledge its limitations; the approach may oversimplify complex phenomena, potentially obscuring deeper insights into the underlying factors driving topic trends, including the nuanced aspects of topic relevance.

Topic over Time (ToT) The ToT model [130] (Chapter 5) not only functions as a dynamic topic model but also enhances the evolution analysis by incorporating a beta distribution to represent the temporal distribution of each topic. This allows TOT to reveal complex trends, such as specific activity peaks or gradual changes in relevance over time, providing deeper insights into the dynamics of topic relevance. More specifically, each topic is represented by two distributions: a multinomial distribution over words, capturing the vocabulary associated with each topic (similar to LDA), and a beta distribution over timestamps, capturing the temporal distribution of documents associated with each topic. This enables a more nuanced understanding of when topics are most active or relevant.

Document Cluster Tracking In the past decade, clustering-based topic models have had a significant impact on the field of topic modeling and evolution analysis as explained in Chapter 2. Document Cluster Tracking refers to monitoring and the continuous flow of documents to find and follow additional content related to the predefined topics based on a set of initial seed documents and their associated topics.

For instance, [179] explores a supervised Topic Tracking that incorporates an unsupervised cluster-based algorithm. This algorithm assigns each document to a cluster based on document-cluster similarity scores. Initially, if a document is labeled as a seed document for a specific topic, any new document that subsequently joins the same cluster as the seed document is considered a potential candidate for being labeled as on-topic for that particular topic.

[186] investigates how temporal metadata can enhance Topic Detection and Tracking (TDT) systems by clustering news articles based on event timing. The authors develop a neural method that combines temporal and textual information to improve event detection, demonstrating significant improvements in clustering performance through the use of time-aware document embeddings.

7.3 Topic Shift Analysis

Topics are fundamentally clusters of words that collectively convey specific semantic meanings [187]. Topic Shift is a phenomenon described by [188] that refers to the changes in the semantics meanings of topic words over time.

There has been a growing focus in recent research on comprehending and identifying these shifts in word semantics. For instance, the word "cloud" has evolved from primarily referring to a weather phenomenon to commonly describing internet-based storage and services. This shift in word semantics has significant implications for understanding changes in language and communication in the digital age.

Words can undergo semantic changes in two primary ways: they may adopt entirely new meanings that supplant their old ones, or they may gain additional meanings while retaining their original sense [189]. [190] categories studies of these shifts within into two classes: *synonymy detection* and *polysemy detection*. Synonymy detection monitors the use of different words with the same meaning over time while polysemy detection monitors different meanings expressed by the same word over time.

To quantitatively measure these semantic shifts, researchers commonly use distributional semantic models [191–193]. These models assess word similarity through vector space representations, where each word is positioned according to its contextual usage. Previous studies have focused on semantic shifts to enhance applications in natural language processing, such as improving document retrieval through time-aware query expansion in historical datasets. These analyses typically consider words in isolation, without linking them to broader topics. In topic evolution analysis, words that migrate between topics do more than just transfer existing lexical information; they foster the creation of new ideas within their new contexts [194]. A word that moves into a different topic is likely to convey fresh concepts, shaped by the unique context of its new environment. This process of semantic transformation underscores the dynamic nature of language and its capacity for generating novel meanings as words traverse different topical landscapes.

7.4 Topic Evolution Network Analysis

Constructing graphs or networks is one of the widely used methods to study topic evolution. Topic evolution networks illustrate a dynamic visualization of how topics develop, transform, and interrelate over time. These networks not only track the progression and change of individual topics but also capture the intricate relationships and structural shifts (split, merge) that occur among topics as they evolve.

Typically, these transformations are modeled using directed acyclic graphs that link and

align topics identified in documents from various time periods. By examining these networks, researchers can gain insights into the temporal dynamics of topics, understanding both their internal changes and how they influence and are influenced by other topics in the broader context.

Topic evolution networks can represent evolution in different ways. [195] defines a topic as a unit of evolutionary change in content. Topic nodes are discovered with the time of their appearance in the corpus and connected to form a topic evolution graph using a measure derived from the underlying document network. In "phylomemetic" topic networks [196] each topic is represented by a cluster of terms that co-occur in a set of documents in a given time period. Topics from different time periods are connected by a matching operator to build a temporal network for identifying emerging, branching, merging, and declining fields. A similar approach was used in [197] to propose an extended graph model for querying evolutionary patterns. The authors of [198, 199] use a Hierarchical Dirichlet Process-based (HDP) model and a temporal similarity graph(using Bhattacharyya distance on the topic vector representation) to model complex topic changes. [200, 201] also applies a network-based topic evolution approach to represent topics by their past neighbors or ancestors and their structural properties in a temporal sequence of *topic co-occurrence networks*. This representation simultaneously captures content transition and topic correlation and detects complex topic evolution events such as merging, splitting, and emergence. In particular, emerging topics are represented as new nodes, where the context of new topics can be inferred from the contexts and past behavior of their neighboring nodes. TermBall [202] builds a weighted dynamic network of co-occurring keywords and extracts topic subgraphs by performing community detection methods. These subgraphs can identify past topic evolution and predict future topic evolution based on structural and temporal features of topic structures. The dynamic Bibliographic Knowledge Graph (BKG) [203]is the result of the merging of bibliographic entities (PMC) with a biomedical and life sciences thesaurus (MeSH). Using different types of topic, paper, author, and venue rankings, together with pre-trained node embedding and various pooling techniques, it can be applied to predict topic hotness.

The predictive power of evolutionary topic networks is also validated through the use of community detection algorithms, where the Advanced Clique Percolation Method (ACPM) classification algorithm [204] has been proposed to identify emerging topic correlations. Topic clusters with notable recent collaborations, found in the semantically enriched topic evolutionary networks and represented by the core publications and author information, are considered as ancestors of a novel topic in its embryonic stage [205].

7.5 Topic Emergence Detection

One of the most useful analyses of the evolution of science is the detection of topic emergence, which involves the identification of new ideas [206]. Emerging topics are ideas or issues that gain attention or become more prominent in a particular field or area of interest and detecting emerging topics in science has far-reaching implications for society, as it provides a way to track the progression of scientific fields and shape future research and technological development.

Topic emergence detection typically employs unsupervised clustering methods to group similar documents and reveal novel topics as they arise. This task has been described by various communities using different terminologies [207–209] and has been applied in various fields such as business [210, 211], information science [12] or public relations intelligence [212].

7.6 Conclusion

The integration of dynamic topic models and advanced algorithms has significantly enhanced the analysis of topic evolution, providing deeper insights into how subjects change and develop over time. These sophisticated tools allow researchers to track shifts in discourse, identify emerging trends, and understand the underlying factors driving these changes. In this chapter, we reviewed a wide range of studies that explore the evolution of topics on a broad scale, highlighting different methodologies and findings. This comprehensive overview sets the stage for the next chapter, where we will narrow our focus to examine the evolution of science specifically. By delving into this specialized area, we aim to uncover patterns and dynamics unique to scientific progress and innovation.

Chapter

8 Science Evolution Analysis

In this chapter, we focus on the evolution of science, categorizing and investigating it thoroughly. We will focus on the patterns of scientific progress, examining how various fields of study have developed and intersected over time. Moreover, we study various topic evolution models that aim to uncover the underlying trends and shifts within the scientific literature, providing a detailed analysis of how knowledge and innovation have evolved. This focused investigation will offer a deeper understanding of the mechanisms driving scientific advancement and the interconnected nature of different scientific domains.

Chapter content

8.1 Introduction

The evolution of topics within scientific domains has been thoroughly examined in the fields of scientometrics and data mining. There are various philosophical theories and definitions for science evolution. For example, Popper considers science evolution as a Bayesian inference process that updates the logical possibility of falsification [7], or Kuhn introduces the notion of Darwinian epistemology where science evolution is a Darwinian selection process of theories [1] characterized into phases of normal, crisis, revolution, and the new normal phase. These theories try to explain the evolution of scientific domains and describe their relations and interactions with different semantics.

Several approaches to analyzing science evolution in scientific archives have been proposed in the literature. We define a *scientific domain* as a scientific evolving topic that changes over time and we categorize these approaches into two classes based on different factors that contribute to the evolution of scientific domains: Single-Domain (SD) Evolution Analysis and Cross-Domain (CD) Evolution Analysis. These changes include influencing other topics, being influenced by them,

fading away, or becoming prominent. Single-Domain (SD) evolution refers to the development of scientific knowledge and methods within a particular domain independent of external factors, while Cross-Domain (CD) evolution refers to the interaction and cross-fertilization of different scientific domains leading to new insights and discoveries. Single-domain-based approaches are generally able to identify trends in the use of specific terms or phrases [127], but may not capture the broader context and relationships between scientific concepts. On the other hand, cross-domain-based approaches can capture the relationships between scientific articles [213] but are less effective at identifying trends in the usage of specific terms or phrases. In the following section, we thoroughly investigate these categorizations.

8.2 Single-Domain Evolution

Single-domain evolution analysis (SD) aims to describe the progression of published content and citation activity within a specific scientific field. This analysis is frequently conducted through the temporal analysis of topics. The temporal dimension can be inherently integrated into the characteristics of topic models, such as dynamic topic models, or implemented as an additional algorithmic layer on top of a static topic model like LDA. Additionally, some methodologies focus on temporal graph analysis across different domains. In this section, we introduce these methods and explore their state-of-the-art developments.

8.2.1 LDA-based Analysis

The idea of using LDA for science analysis is particularly intriguing because LDA is inherently static and does not account for temporal dynamics in its latent discovery process. This makes the integration of temporal elements an innovative approach to uncovering how topics evolve over time. Indeed, when applying LDA-based topic models to a complete corpus, the outcome is a document-topic matrix. This matrix represents the distribution of topics across each document, effectively indicating how much each topic contributes to the corpus. Given that each document is associated with a timestamp, these contributions can be aggregated over specific time periods to reveal trends in topic popularity. By analyzing these aggregated contributions over time, researchers can track how the prominence of different topics evolves throughout the dataset.

For instance, the authors of [214] apply LDA on the ACL Anthology (1978-2006) to study trends in Computational Linguistics. They found that topics at three major conferences have become increasingly similar over time, supported by analyzing the Jensen-Shannon divergence [215] of their topic distributions.

[216] uses LDA and regression analysis to explore topic evolution in astrophysics from 1992 to 2011. They compare topic popularity and duration between arXiv preprints and Web of Science (WoS) papers. The analysis showed that topics in WoS lose popularity sooner than those in arXiv, and open-access preprints exhibit stronger growth compared to traditional publications.

Some works extend LDA and propose novel architectures to analyze scientific evolution. For instance, [217] investigates the development of Cognitive Science over 30 years using an LDA-based topic model with a topical weighted-contribution method. The authors analyze the weighted contributions of topics year by year among 3,104 articles published from 1980 to 2014. This approach allows them to map the evolving landscape of the field through detailed topic trend analysis.

[114] implements a Gibbs sampling-based version of LDA to analyze a dataset of 28,154 abstracts from the Proceedings of the National Academy of Sciences (PNAS) spanning the years 1991 to 2001. The authors define a technique to determine the optimal number of topics by employing Bayesian model selection focused on log-likelihood [218]. To track how these topics

evolve over time, the authors conduct a trend analysis on the distribution of topics across the documents each year.

8.2.2 DTM-based Analysis

This approach employs dynamic topic models (DTM) that reflect the evolution by discovering latent semantic structures of the documents published in different time periods. For instance, Leap2Trend [219] relies on temporal word embeddings to track the dynamics of similarities between pairs of keywords, their rankings, and the respective uprankings (ascents) over time. Another example is [220] which proposes to use of a time-stamp-based discounting learning algorithm to track online topics by forgetting out-of-date publications.

8.2.3 Graph-based Analysis

Graph-based topic models take advantage of topic networks to analyze the evolution of a single domain [221, 222]. These models are mainly used for the design of digital libraries and social media platforms, as well as the evaluation of scientific contributions and policies [209]. For instance, CiteSpace II [223] is one of the first frameworks for detecting and visualizing evolution patterns in scientific archives. It applies various graph analysis methods such as burst detection and betweenness centrality to analyze highly cited publications within a scientific domain. Another example is the Inheritance Topic Model (ITM) [222], which is an iterative topic evolution learning framework that extends LDA with citation networks. [224] proposes an alternative topic modeling method conscious of the topical correlation in the academic domain by introducing the notion of the common interest authors (CIA1), defining a topic as a set of shared common research interests of a researcher group by citation. Similarly, [225] introduces an LDA-based topic model that utilizes citation counts to measure the impact of topics. The authors define a graphical model called Topical Impact over Time (TIoT), which identifies trending topics and highlights significant papers within a bibliographical database. By this, TIoT captures the changing influence of latent topics over time by analyzing the temporal dynamics present in a corpus of documents.

8.3 Cross-Domain Evolution

Cross-domain evolution is studied by analyzing semantic relationships between scientific evolving topics (domains) and observing their evolution over time. This analysis primarily focuses on the analysis of topic representations (e.g., semantic similarity) and the topic citations (e.g. citation context) of evolving topics. Topic evolution networks (Chapter 7) are a natural tool for this kind of analysis that includes identifying evolution patterns such as topic merging, topic splitting, etc. [226–228].

Research on cross-domain topic evolution is less prevalent compared to studies focused on the evolution within a single domain, such as topic trend analysis. Most existing studies in this area examine the development of scientific fields by analyzing scholarly literature, and tracing how topics evolve within these domains over time. We categorize these studies based on the type of information used in their methodology into the following classes.

8.3.1 Citation-based Analysis

Citation-based analysis leverages document archives that contain additional information, such as citation links or retweets/likes, to construct *document citation networks*. These networks can be used to investigate semantic relationships between topics, such as topic influence [10] and

information flow [209, 229]. By analyzing these networks, it is possible to study more complex trends in the interaction between different topics like novelty [230, 231] and paradigm shift [232, 233].

For instance, [234] seeks to discover the pairwise probabilistic dependency in topics of documents that associate authors from a latent social network. In [235], authors apply node-embedding algorithms to citation networks for community detection and citation prediction. These low-dimensional vectors are mainly used for community detection within citation networks. These approaches lack content and can only describe information in the scale of authors or documents.

[13, 236] proposes the Flow Vergence (FV) gradient to detect paradigm shifts in the scientific literature. Similarly in [209], authors proposed a method to identify dynamic knowledge flow patterns of business method patents with a hidden Markov model. Moreover, the authors of [210] combine citation and co-citation analysis, and social network analysis altogether in order to review trends in the field of business research.

8.3.2 Semantic-based Analysis

These methods are based on the temporal analysis of topic representations (Part I) changing across different time periods.

For instance, [196] introduces a method for reconstructing the dynamics of scientific fields using phylomemetic networks. Topics are defined as directed cliques of co-occurring terms and aligned across different time periods using Jaccard similarity [237]. These networks can then be used to visually trace the evolution of topics and the structural transformations within different scientific domains over time and the approach is illustrated by two extensive case studies on "embryology research" and "networks in biology" corpora.

[238] proposes a framework for identifying emerging topics based on dynamic co-word network analysis. Time-sliced co-word networks are weighted according to co-occurrence term frequency, and a back-propagation neural network is used to predict a future co-occurrence term network. Moreover, DAC [201] is a descendant-aware topic clustering algorithm that generates overlapping clusters as candidates for future emerging topics. In [200], the authors train binary classification models to capture the materialization of emerging topics.

The authors of [190] utilize LDA to extract local topics from each time span within a dataset of information retrieval publications. They investigate the evolution of research topics, analyze the trends of topics across documents, and explore the dynamic processes involving the splitting and merging of topics, highlighting the transfer of knowledge between these evolving topics. Technically, They then aggregate the per-document topic distributions annually to assess the popularity and detect trends in these topics over time. The correlation between local topics and between a local topic and a global topic is measured by cosine similarity [239]. With this methodology, the splitting and merging of local topics indicates the existence of knowledge transfer within a global topic or between global topics.

Similarly, the authors of [199] present a framework based on HDP for uncovering the thematic substance within a corpus of data and tracing its intricate structural transformations over time. This approach leverages metrics such as Hellinger distance [240], BHD [241], and Jaccard similarity [237] and involves the construction of a similarity graph. The authors employ an automated edge elimination mechanism based on a predefined threshold. This process ensures that only connections between sufficiently akin topics across consecutive time periods are preserved. Subsequently, they prune the graph by selecting the appropriate operating point on the cumulative distribution function (CDF) [242], rather than relying on a fixed similarity threshold. Their findings indicated that incorporating a moderate overlap (25-50% of the epoch length) between

successive epochs could substantially improve the interconnectedness of topic nodes within an evolutionary graph.

8.3.3 Hybrid Analysis

These models combine citation links, semantic links, and other network analyses to study topic evolution and their dynamics [190]. For instance, the Emerging Clusters Model [243] uses citation analysis to assess the impact of patents on subsequent technological developments. The basic idea is to accelerate the detection of emerging topics by analyzing the evolution of citing and cited patents to identify clusters around hot topics.

In [244], the authors aim to detect the emergence of new research topics using a hybrid approach based on citation analysis and keywords. [212] explores a research area, Public Relations Intelligence, using a hybrid approach based on the most frequently cited articles, keyword occurrence frequency, and co-occurrence network. Moreover, the authors of [211] define innovation and entrepreneurial ecosystems, and operate a co-citation analysis and network meta-analysis for explaining the trends and features of multiple meta-knowledge. In [12], the authors aim to study the timeline knowledge map through author co-citation analysis and direct clustering algorithm. Furthermore, the authors of [245] integrate citation graphs into the NTMs to forecast the links between scientific articles. Consequently, their model not only analyzes the content but also suggests related articles to users.

8.4 Paradigm Shift Analysis

Beyond single-domain and cross-domain analyses, another important line of research examines paradigm shifts in scientific communities. A paradigm shift is a transformative process where prevailing theories and methodologies face challenges from anomalies and new discoveries. This occurs through the collective efforts of scientists engaged in research, experimentation, and analysis, ultimately overwhelming protective hypotheses and leading to a radical departure from established norms, thus paving the way for new perspectives. This concept can be approached from multiple angles. Cross-domain evolution analysis explores how a topic from one community contributes to another scientific field, potentially catalyzing significant changes and fostering new topics within a hybrid community composed of the original fields. Conversely, single-domain analysis can track paradigm shifts by examining how a community evolves over time, generating new concepts within its boundaries. In the following section, we will provide a comprehensive definition of paradigm shifts and delve deeper into their implications for scientific progress.

A paradigm shift represents a fundamental change in the underlying assumptions and methodologies of a particular scientific field, driving the evolution of topics and fostering the emergence of new research areas within scientific archives [1]. As shown in Figure 8.1, the general pattern behind paradigm shift is described by the Kuhn Cycle [246] which represents the cyclical nature of scientific progress. According to this model, a paradigm shift occurs within the following phases:

- **Pre-Science:** This phase consists of several incompatible and incomplete theories without any consensus on specific research goals shared by a community.
- **Normal Science:** This phase is characterized by a consensus within a community and a dominant paradigm.
- **Model Drift:** During the normal science phase, several anomalies might appear that challenge the existing paradigm within facts that are difficult to explain and initiate a model drift phase.

- **Model Crisis:** In this phase, the community starts debates and enters a crisis mode as the consequence of the paradigm's inability to account for the anomalies.
- **Model Revolution:** In this phase, the scientific revolution takes place, underlying assumptions of the field are re-examined and a new paradigm gets established.
- **Paradigm change (Post-revolution):** This phase establishes a new paradigm's dominance and delivers a normal science to scientists for solving puzzles within the new paradigm.

Recent studies in fields such as innovation policy [248] and revenue management in hospitality and tourism [249] have employed this framework to assess the state of their respective disciplines. While some domains, like innovation policy, appear to be approaching a "crisis stage" that may precede a paradigm shift due to ongoing societal and technological transformations, others, such as revenue management in hospitality and tourism, have not yet progressed beyond the "normal science" phase. These diverse examples illustrate that the applicability and manifestation of paradigm shifts vary across disciplines, reflecting the unique challenges and developments within each field. Researchers are increasingly using scientometric analyses, evolutionary frameworks, and co-citation studies to track these potential shifts and understand the changing landscape of knowledge in their respective domains.

8.4.1 Characteristic

To gain a clearer understanding of paradigm shifts, it is necessary to identify and analyze their key characteristics. As described in [247], the main characteristics of paradigm shift include 1) Collectiveness, 2) Sequentially Cumulative, 3) Causal Interdependence, and 4) Contextual Metamorphosis. These characteristics are explained in detail as follows.

- **Collectiveness:** This characteristic emphasizes the collaborative and cumulative nature of the shift within a topic or group of research publications. In other words, a paradigm shift is driven by a collective group of documents within specific problems, rather than being limited to individual research or isolated instances.
- **Sequentially Cumulative:** Scientific knowledge is cumulative, meaning that discoveries and insights build on previous ones. This cumulative process of knowledge generation is imposed by time in a chronological order that allows for the continuous expansion and refinement of scientific understanding.
- **Causal Interdependence:** The progression of scientific knowledge generation is not only sequential but also interconnected in such a way that each phase influences or causes the next.

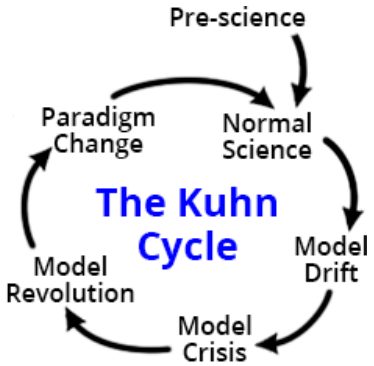


Figure 8.1: *Kuhn Cycle [247]*

- **Contextual Metamorphosis:** As scientific exploration deepens and anomalies emerge, a dominant paradigm undergoes a complete and fundamental transformation, eventually producing a new paradigm that is radically different from the initial paradigm.

By examining these characteristics in detail, we can develop a more comprehensive framework for recognizing and interpreting paradigm shifts across different fields of study. In Chapter 10, we suggest a new definition for paradigm shift and propose a framework that extracts topic evolution from the perspective of paradigm shifts.

8.5 Conclusion

In this chapter, we have focused deeply on the complex landscape of scientific evolution, categorizing and investigating the multifaceted processes that drive the advancement of knowledge. By examining patterns of scientific progress and the interplay between different fields of study, we have highlighted the dynamic nature of science, driven by innovations and cross-disciplinary fertilization. Our investigation was structured around two primary approaches to analyzing science evolution: Single-Domain Evolution Analysis and Cross-Domain Evolution Analysis. Through the lens of Single-Domain Evolution, we examined the progression within specific scientific fields using dynamic topic models, LDA-based methods, and graph-based techniques. These approaches provided insights into the temporal dynamics of topic trends and the underlying structures of scientific knowledge within individual domains. On the other hand, Cross-Domain Evolution Analysis focuses on the interactions between different scientific fields, utilizing citation networks, semantic links, and hybrid approaches to uncover the relationships and influence between topics. This cross-disciplinary perspective revealed the complex web of connections that drive scientific innovation and highlighted the importance of interdisciplinary research in fostering new insights and discoveries. Throughout this chapter, we have seen how advanced algorithms and dynamic models enable us to track the evolution of scientific topics, identify emerging trends, and understand the mechanisms underlying the development and dissemination of knowledge. The integration of temporal elements into these models has proven crucial in capturing the evolving nature of science, providing a nuanced view of how topics emerge, converge, and diverge over time. In the next two chapters, we introduce models, developed from our current findings, that facilitate the discovery of emerging topics and the investigation of paradigm shifts.

Chapter

9 Automatic Topic Emergence Monitoring

The evolution of science is an ever-unfolding narrative, characterized by the relentless pursuit of new knowledge and the refinement of existing theories. This dynamic process is driven by the collaborative efforts of scientists who engage in rigorous research, experimentation, and analysis. Their endeavors are influenced by a myriad of factors, including groundbreaking discoveries, technological innovations, and the steady accumulation of empirical evidence. Gaining a deep understanding of the evolution of science not only enriches the current research landscape but also informs strategic resource allocation. This understanding has profound implications for research funding and public policy decisions, impacting both academic and industrial spheres.

This chapter presents ATEM [250, 251], which aims to advance our understanding of scientific evolution and to provide valuable insights into the identification of emerging topics. The chapter includes the introduction of the ATEM science evolution model, a discussion on the use of ATEM for various science evolution analysis tasks, an exploration of the application of ATEM to emergent topic detection, and the presentation of the implementation and proof of concept.

This work has been presented at BDA'2023 and published in *Rahimi, H., Naacke, H., Constantin, C., Amann, B. (2024). ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives. In: Cherifi, H., Rocha, L.M., Cherifi, C., Donduran, M. (eds) Complex Networks & Their Applications XII. COMPLEX NETWORKS 2023. Studies in Computational Intelligence, vol 1143. Springer.*

Chapter content

9.1 Introduction

One of the most insightful ways to analyze the evolution of science is through the detection of emerging topics. This process involves identifying nascent areas of research and study within scientific disciplines. Emerging topics represent ideas or issues that are gaining traction or becoming increasingly prominent within a specific field or broader area of interest. The ability to detect these emerging topics has significant societal implications. It enables the tracking of scientific progress and helps shape future research agendas and technological advancements.

Various methodologies have been developed to detect emerging scientific themes, each with its own strengths and limitations. Single-domain approaches are adept at identifying trends based on the usage of specific terms or phrases. However, they often fall short of capturing the broader context and the intricate relationships between scientific concepts. In contrast, cross-domain approaches excel at uncovering the relationships between different scientific articles. Yet, these methods may struggle to pinpoint trends in the usage of specific terms or phrases. In this chapter, we introduce a novel framework called ATEM (Automatic Topic Emergence Monitoring) designed to uncover emerging topics through a multifaceted analysis of the evolution of science. This framework extracts evolving topics from a corpus of documents using a dynamic topic model. It then creates a dynamic topic-citation graph by projecting document citation links into the topic space. Finally, it applies a dynamic graph embedding method on the topic-citation graph for representing the dynamic citation context of topics. Based on this context, emerging topics can be defined as couples or sets of evolving topics with similar citation contexts. The motivation behind ATEM lies in the recognition that cross-domain citation links not only reveal semantic relationships between disparate topics but also signal the potential emergence of new interdisciplinary areas. By employing dynamic graph embedding, ATEM is capable of detecting emerging topics and predicting future interdisciplinary trends. The overall architecture of ATEM is depicted in Figure 9.1, and a detailed explanation of its implementation will be provided in Section 9.5.

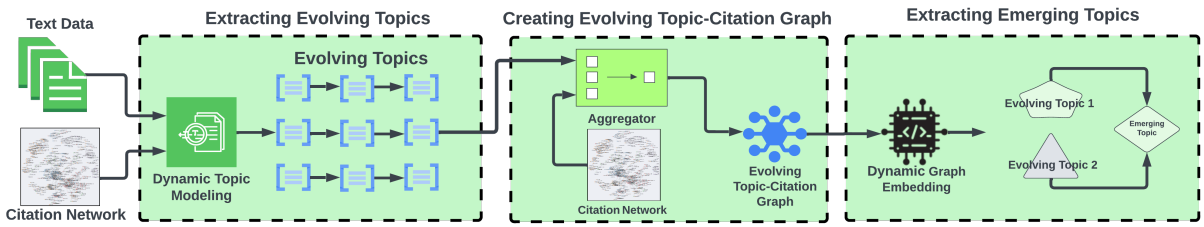


Figure 9.1: *Architecture of ATEM*

9.2 ATEM Evolution Model

ATEM is a general-purpose framework for modeling and analyzing the evolution of topics extracted from scientific archives. ATEM extracts *evolving topics* using dynamic topic models and builds *evolving topic-citation graphs* that connect topics through temporal citation links.

9.2.1 Evolving Topics

Evolving topics (Definition 5.1.1, page page 43) are dynamic subjects that trace the development of a single domain over time, characterized by a temporal structure[127]. This definition is applicable to any topic derived from a document archive using a topic model. The representation of a topic can be precomputed by the topic model or determined post hoc through a labeling

function, such as cTF-IDF[54], applied to the relevant documents. It's important to note that there may be periods without any sub-topics, in which case the document set is considered empty and the topic representation is undefined. For instance, Table 9.1 illustrates an evolving topic with the ID T680C6 (as depicted in Figure 9.2), showing the temporal progression of its sub-topics annually from 2004 to 2018.

Table 9.1: *Distributed word representation for evolving topic T680C6.*

Year	Label
2004	['knn', 'linear classifier', 'nearest neighbors', 'nearest neighbor', 'distributional', 'neighbor classifier']
2005	['knn', 'nearest neighbors', 'euclidian', 'pearson', 'instance based', 'neighbor nn']
2006	['knn', 'instance based', 'neighbor classifier', 'nearest neighbors', 'neighbor nn', 'neighbor knn']
2007	['knn', 'relief', 'nn classifier', 'neighbor nn', 'membership values', 'nearest neighbors']
2008	['knn', 'neighbor nn', 'nearest neighbour', 'nn algorithm', 'nearest neighbors', 'instance based']
2009	['knn', 'neighbor knn', 'nn classifier', 'nearest neighbors', 'neighbor nn', 'text classification']
2010	['nearest neighbors', 'neighbor classification', 'knn', 'metric learning', 'knn classifier', 'neighbor classifier']
2011	['instance selection', 'knn', 'neighbor classifier', 'neighbor classification', 'nearest neighbors', 'instance based']
2012	['knn', 'nearest neighbors', 'neighbor knn', 'test sample', 'instance based', 'nn classifier']
2013	['knn', 'nearest neighbors', 'based nearest', 'knn classifier', 'decision boundary', 'neighbor classifier']
2014	['knn', 'metric learning', 'nearest neighbors', 'nn classifier', 'knn classifier', 'neighbor nn']
2015	['nn classifier', 'knn classifier', 'knn', 'instance based', 'pmc', 'class label']
2016	['knn', 'instance selection', 'knn algorithm', 'knn classification', 'dpc', 'nn classification']
2017	['knn classifier', 'local mean', 'harmonic mean', 'nearest neighbors', 'knn', 'based nearest']
2018	['cent', 'knn classifier', 'knn', 'neighbor method', 'instance selection', 'nearest neighbors']

9.2.2 Evolving Topic-Citation Graph

ATEM aims to discover citation relationships among the evolving topics as an indicator of cross-domain evolution by projecting the structure of citation networks into evolving topics.

Definition 9.2.1 (Evolving Topic Citations)**.** Let $\mathcal{T}$ be a set of evolving topics defined over a document set $\mathcal{A}$ and $E_D(\mathcal{A}) \subseteq \mathcal{A} \times \mathcal{A}$ by a set of citation links defined on $\mathcal{A}$. Then, the topic clusters $D(t_i)$ all topics $t_i \in \mathcal{T}$ and the document citation edges E_D define a set of edges $(t_x, t_y, j) \in E_{\mathcal{T}}$ from an evolving topic t_x to evolving topic t_y if there exists at least one citation from some document in $D(t_x^j)$ to a document in $D(t_y^k)$ where $0 \leqslant k \leqslant j$:

$$E_{\mathcal{T}} = \{(t_x, t_y, j) \mid d \in D(t_x^j), d' \in D(t_y^k), 0 \leqslant k \leqslant j : E_D(d, d')\} \tag{9.1}$$

We can add to each edge (t_x, t_y, i) in $E_{\mathcal{T}}$ a weight w which corresponds, for example, to the number of citations that exist between the documents in $D(t_x^j)$ and $D(t_y^k)$, $0 \leqslant k \leqslant i$.

9.3 ATEM Evolution Analysis

The proposed evolution model can be used to analyze the evolution of topics within the two classes defined in the previous chapter. The first class is single-domain evolution analysis which observes the change within the representation of words and contents, the size of topic clusters, and the number of incoming and outgoing citation links. The second class is called cross-domain evolution analysis and explores the evolution of topic relationships defined by the topic citation network.

Single-Domain Analysis Using ATEM, one can analyze the change within the word representation of a single topic. As shown in Table 9.1, observing this change allows us to explore the semantic transformation in our understanding of a single topic by examining the words and

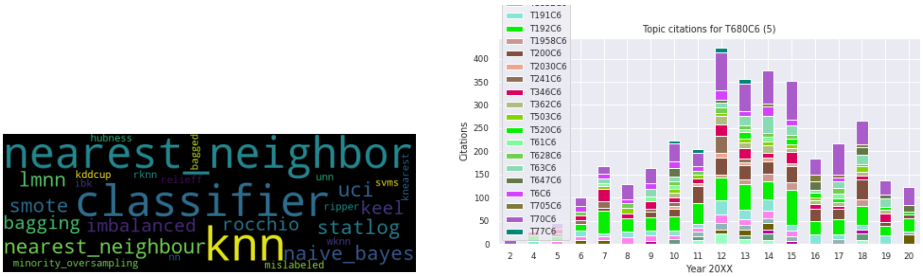


Figure 9.2: *Evolution of citations of evolving topic T680C6.*

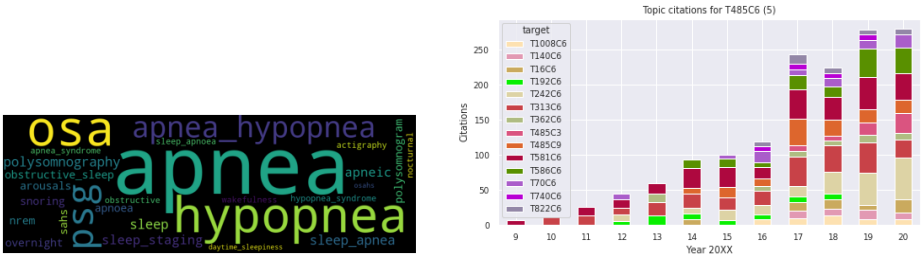


Figure 9.3: *Evolution of citations of evolving topic T485C6*

phrases that are commonly associated with that topic. Besides, one can identify changes in the way people conceptualize a single topic for discussion and research. This kind of information is useful for researchers and companies seeking to stay up-to-date on the way people think about and discuss a particular topic. In addition, ATEM allows us to explore the publication and citation trends of individual topics by observing the document cluster size and the incoming and outgoing citations over time.

Example 1. Figures 9.2 and 9.3 show the evolution of citations to topics T680C6 and T485C6 (with at least 5 citations).

Cross-Domain Analysis Using ATEM, one can observe the evolution of cross-domain relationships between topics. ATEM investigates factors that contribute to the growth or decline of topics by observing the evolution of citation links between topics. Any variation in the existence and number of citation links between topics is a signal of change for the discovery of new knowledge and in particular the emergence of new research topics (see Section 9.4).

Citation links between documents can also be used as a way to trace the development and evolution of ideas over time. By following the citation links between documents, it is possible to see how different ideas or research findings have been built upon or challenged by subsequent work. This can provide insight into the relationships between different topics and how they have evolved over time. Therefore, we define a concept called Citation Context for topics. ATEM identifies emerging disciplines by investigating and comparing the Citation Context of different topics. More exactly, we consider that if two different topics are cited by the same other topics at a given time period, this may indicate the emergence of a new topic defined by the document contents of the cited topics.

Example 2. Figure 9.4 shows the evolution of co-citations to evolving topic T680C6 and T485C6. We can see that both topics have been cited by topic T70C6 (Figure 9.5) in 2009 and from 2012 to 2020. The number of co-citing topics increases in 2012 which might be considered as a first indication that both concepts are the origin of a new emerging research topic about nearest neighbor classifiers and apnea analysis.

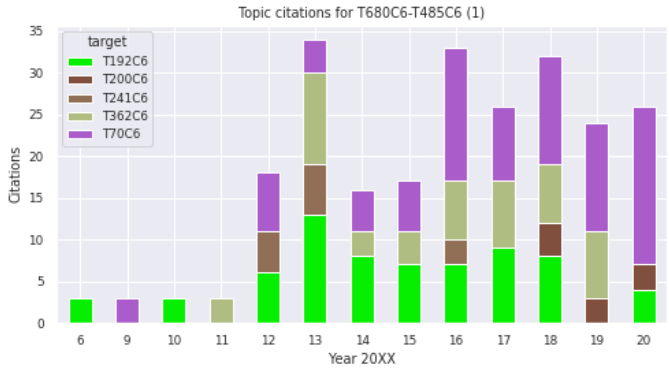


Figure 9.4: *Co-citations of T680C6 and T485C6*

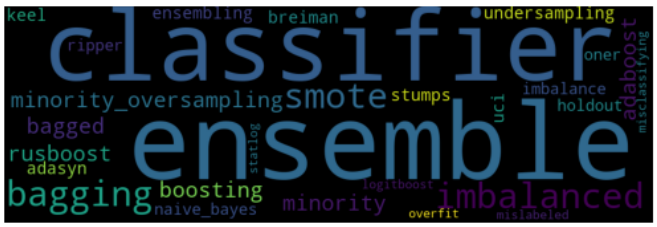


Figure 9.5: *Citing evolving topic T70C6*

9.4 Emerging Topic Detection

One of the main goals of ATEM is to identify emerging research topics based on the topic citation graph. We hypothesize that citations in documents indicate a relationship between the topics discussed in those documents. When a document cites another document, it typically means that the authors of the first document refer to information, ideas, or research from the second document as support or evidence for their work and that the topics discussed in the citing document are influenced or informed by the topics discussed in the second document.

Citation context refers to the ways in which documents in a given topic cite and are cited by documents in other topics. By analyzing and comparing the citation context of multiple topics, it may then be possible to gain a better understanding of the relationships between different ideas and their evolution over time. In other words, *if two or more topics share a high citation context similarity at a given time period, it is reasonable to expect that they might merge into a new topic in the future, providing a basis for the development of new and interdisciplinary ideas.* Thus, the citation context similarity between two topics can be seen as a measure of the likelihood of discovering new interdisciplinary topics in the future by merging these topics.

We assume that two evolving topics t_i and t_j with a *highly similar* citation context at a given time period produce an emerging evolving topic $t_{i,j}$ Based on this assumption, we need to define a similarity measure that allows us to compare the context of two nodes defined by the topic citation graph $E_{\mathcal{T}}$. A graph embedding [252, 253] of a graph G is a mapping function $emb : G \mapsto 2^{\Re^d}$, which aims to represent nodes, edges, subgraphs, or even the entire graph by low-dimensional feature vectors $v \in \Re^d$ that preserve the topological and other contextual information about the encoded entity. The embedding dimension d is expected to be much smaller than the size of the graph $d \ll n$, where n is the number of nodes in G, which allows nodes to be efficiently compared by the encoded properties. Using this technique, our idea is then to represent the context of a topic by its embedding in the citation graph and to generate for each evolving topic a *sequence of graph embeddings* which reflects the evolution of its context in the topic citation graph.

There are several dynamic representation learning methods capable of embedding nodes in a low-dimensional vector space which captures the evolution of the network structure. In our implementation, we use dynamic node embeddings [254], which project each node v in a sequence of graphs into a *sequence* $emb(v)$ *of low-dimensional vectors*. By projecting the topic citation links $E_{\mathcal{T}}$ defined in Section 9.2 on each time period $p^i \in P$ a set of topic edges $E^i_{\mathcal{T}} = \{(t_x, t_y) \mid (t_x, t_y, i) \in E_{\mathcal{T}}\}$, we can produce a *sequence of graphs* $\mathcal{G}(P) = [(\mathcal{T}^i, E^i_{\mathcal{T}}) \mid p^i \in P]$ ordered by periods p^i that reflects the distribution of citations between documents of all topics for all time periods. We hypothesize that the dynamic topic embedding vector $emb(t)$ of a topic t in a dynamic topic citation graph represents the evolution of the citation context of t, and that two topics with similar embedding (citation context) at period p^i are likely to generate new emerging topics. More formally, we can now provide a more precise definition of emerging topics:

Definition 9.4.1 (Emerging Topics)**.** Two evolving topics t_i and t_j define an evolving topic $t_{i,j}$ emerging at time period p^k, if the context distance $dist(t^k_i, t^k_j)$ at period p^k is above a given threshold ϕ and below this threshold before p^k.

In our implementation, we use *cosine*-similarity on the topic embeddings to compute the distance between two topic contexts. Using this definition, we can now detect emerging topics in two ways:

1. K-nearest neighbors of a given topic t: we generate for each evolving topic t and period p^i a set of nearest neighbors with minimal embedding distance higher than a given threshold.
2. Cluster the embeddings of each period: we apply a clustering algorithm on the topic embeddings of each period. Each cluster represents an emerging topic defined by a set of similar topics.

Note that the definition of an emerging topic does not include the definition of its topic cluster and representation. This is consistent with the observation that the "future" documents of a topic emerging at a given time are not defined. However, as we show in our experiments, we can define functions to estimate the past and future documents using the topic model or a search engine like Google Scholar.

Example 3. Figure 9.6 shows the evolution of topics emerging for topic T680C6 in 2013. We can see, for example, topic T661C6 appears in 2013 as a near embedding neighbor of evolving topic T680C6 (with a maximal distance of 0.2).

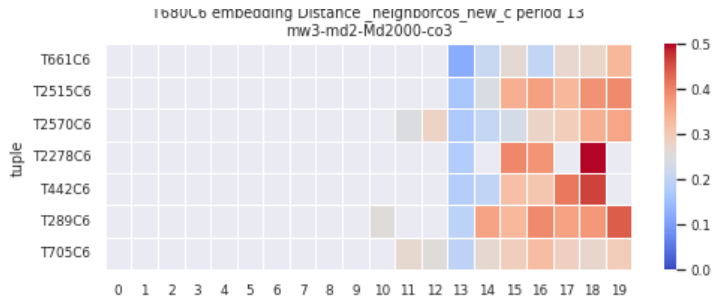


Figure 9.6: *Embedding distance evolution for topic T680C6 at period 2013*

Table 9.2 shows the documents that are common to T680C6 and T661C6. These documents are obtained by taking the intersection of the results of two queries $R(T680C6)$ =['nearest neighbors', 'knn', 'nearest neighbor'] and $R(T661C6)$ =['glycemic', 'hypoglycemia', 'hyperglycemia'] ranked by the average search score. The result shows that most of the top relevant documents for emerging topic (T680C6, T661C6) have been published after its emergence period 2013.

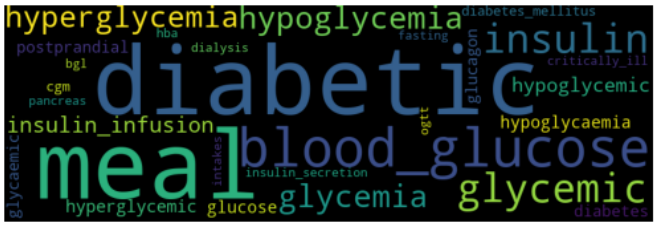


Figure 9.7: *Evolving topic T661C6*

Table 9.2: *Common documents for topic (T680C6,T661C6) emerging in 2013*

Year	Title
2020	Performance evaluation of classification methods with PCA and PSO for diabetes.
2020	An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy
2020	Using Machine Learning to Predict the Future Development of Disease
2019	Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus.
2018	Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers.
2017	Automatic Diagnosis Metabolic Syndrome via a k- Nearest Neighbour Classifier.
2016	Predicting risk of suicide using resting state heart rate.
2015	Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method
2013	Automated detection of diabetes using higher order spectral features extracted from heart rate signals

9.5 Implementation

ATEM is implemented in Python and has been applied to the DBLP[255] dataset of 5M scientific articles published between 2000 and 2020. DBLP is a large open bibliographic database, search engine, and knowledge graph that archives computer science publications. The architecture of ATEM is shown in Figure 9.1. We can distinguish three main steps that will be explained in the following subsections.

9.5.1 Extracting Evolving Topics

The evolving topic extraction workflow consists of four steps, as shown in Figure 9.8. The first two steps generate two types of document clusters (content-based and citation-based). The first type of cluster corresponds to the traditional notion of topic clusters and regroups semantically similar documents. The second type of cluster applies a community detection algorithm to the document citation graph. We assume that the citation graph reflects different research communities, and by using this process, each generated cluster separates documents published and referenced by

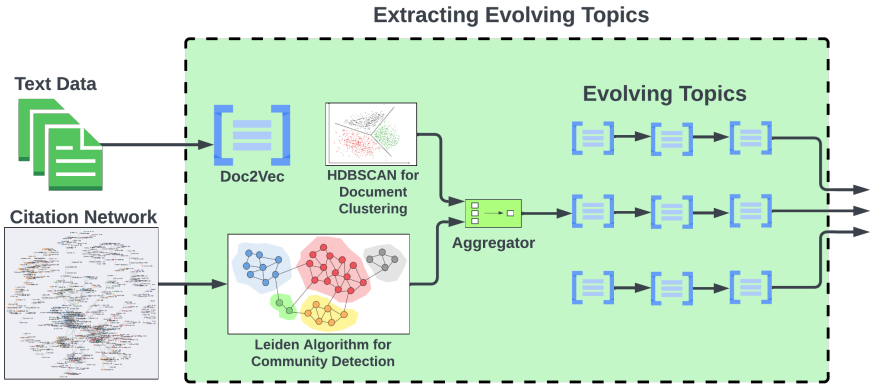


Figure 9.8: *The Implementation of Extracting Evolving Topics.*

different research communities. The third step (cluster aggregation) consists of combining both types of clusters into new clusters corresponding to topic clusters that regroup semantically similar documents published and cited within a scientific community. The last step (evolving topic representation) consists of decomposing each cluster into a temporal sequence of sub-clusters (evolving topics) and computing a representation for each sub-cluster.

Content-based Document Clustering For clustering documents based on their contents, we computed jointly word and document embeddings on the document content with Doc2Vec [46], then reduced the document dimensions with UMAP [256], and finally clustered them with HBDSCAN [48]. This process is similar to existing cluster-based topic modeling techniques such as Top2Vec [24] and BERTopic [16].

Citation-based Document Clustering For clustering documents based on the citation network, we perform the Leiden community detection method [257] on the citation graph. Each document community then reflects the publication activity within a community of authors that are distinguishable from other author communities in the network.

Cluster Aggregation After the two clustering steps, we have two types of document clusters T and C, which are independently informative and valuable. The cluster aggregation step simply consists of taking the intersection of all clusters in T with all clusters in C to obtain a set of non-empty clusters $\mathcal{T}$ where for each cluster $tc_{i,j} \in \mathcal{T}$ is the intersection of some cluster $t_i \in T$ and some cluster $c_j \in C$.

Dynamic Topic Representation In the next step, the topic document clusters $D \in \mathcal{T}$ are divided into $n = |P|$ time frames denoted by $D = (D^1, \ldots, D^n)$ where each D^i is a cluster of documents in period p^i. In this regard, we adopt the dynamic document integration of clusters upon using static time windows. We only keep clusters with a minimal number of 3 documents. Each of these topic clusters is represented in two manners:

- Nearest Words: we compute for each document cluster a centroid vector by averaging over the embeddings of its vectors. The cluster representation is defined by the top-n words corresponding to the n nearest embedding neighbors of the centroid vector.
- Class-based TF-IDF: similar to [54], we regroup the documents of each topic cluster $D(t^i)$ in all time periods p^i and apply TF-IDF to each group to find the top-n word representation for each group.

9.5.2 Creating Topic-Citation Graphs

To generate the evolving topic-citation graph as defined in Section 9.2, we create a node for each evolving topic and a weighted directed edge according to Equation (9.1). Each citation edge in a given time period is weighted by the number of outgoing citations. *This dynamic representation guarantees that the embedding of a topic in a given period depends only on the citations of the current and previous periods.*

9.5.3 Extracting Emerging Topics

To compute the temporal node embeddings on the topic citation graph, we used OnlineNode2Vec[258], which is based on StreamWalk and online second-order similarity. The result is

a temporal embedding for each topic, which can be used to compare evolving topics by their citation context. In particular, it allows us to identify evolving topics (t_x, t_y) when the distance between two evolving topics t_x and t_y is less than a given threshold. In our implementation, we use an approximate nearest neighbor algorithm to generate and rank, for each topic t_x and each period, the set of all nearest neighbors t_y that were not nearest neighbors in a previous period. A collection of generated emerging topics is available at here.

9.6 Proof of concept

The objective of this section is to demonstrate the effectiveness of the proposed framework in identifying emerging topics as compared to co-citation analysis. To achieve this, we categorize the emerging topics into two groups: one based on the embedding representations of the topic-citation graph (referred to as *EmbeddingContext*), and the other based on co-citation analysis of topics connected through citations (referred to as *CitationContext*). These two groups are then compared to each other. To facilitate this comparison, we generate a set of emerging topics from both *EmbeddingContext* and *CitationContext*. We assess the validity of these emerging topics by examining the presence of related documents in the past and future of their discovery. To quantify this, we employ a predictability metric that evaluates the distribution of related documents over time. By scoring the emerging topics based on this metric, we can effectively evaluate their predictive power and performance. Therefore, we first generate a random sample of 200 evolving topics T. For each evolving topic $t \in T$, we generate two sets of $n = 10$ topics in each time period p^i:

1. $EmbeddingContext(t^i)$ contains n topics t_x that are *new* nearest embedding neighbors of t^i at period p^i with a given maximum distance threshold of 0.2 and minimum embedding norm equal to 0.22 to remove noisy embedding vectors (t_x was not a neighbor before period p^i).
2. $CitationContext(t^i)$ contains a random set of n topics t_x connected to the evolving topic t at period p^i by a citation path of maximal length equal to 3.

In each of these sets, each pair $t_e = (t, t_x)$ generated by $t_x \in CitationContext(t^i)$ and $t_x \in EmbeddingContext(t^i)$ is expected to form an emerging topic at period p^i. To explore this expectation, we consider all pairs t_e that were appeared for the first time as emerging in time period p^i with representation $R(t_e) = [R(t^j) \cup R(t_x^j) \mid p^j \in P]$ and a document cluster $D(t_e) = [D(t^j) \cap D(t_x^j)) \mid p^j \in P]$ over all periods p^j in P.

We then look at each of these new topics and investigate their emergence predictability based on the year their papers get published. Therefore, we partition $D(t_e)$ into two subsets: $D_{past}(t_e)$ of documents published before the emergence period of t_e, and $D_{future}(t_e)$ of documents in $D(D)$ published after the emergence period of t_e.

Finally, we quantify the *emergence predictability* $\mathcal{E}$ of each topic pair t_e by defining the following function that measures the distribution of its documents before and after its emergence period:

$$\mathcal{E}(t_e) : \frac{|D_{future}(t_e)| - |D_{past}(t_e)|}{|D(t_e)|} \tag{9.2}$$

meaning (i) when $\mathcal{E}(t_e) = 1$, all documents are published at emergence period of (t, t_e) or afterwards, (ii) when $\mathcal{E}(t_e) = 0$, the same number of documents are published before and after the emergence period and (iii) when $\mathcal{E}(t_e) = -1$, all documents are published before period p. While the number of emerging topics (topic pairs) generated by *CitationContext* increases with

time, we observed that the average number of topics generated by *EmbeddingContext* decreases and the average embedding distance increases.

Figure 9.9 compares the predictability values for emerging topics of *EmbeddingContext* and *CitationContext*. We find that random pairs from *EmbeddingContext* have higher predictability compared to *CitationContext*. Figures 9.10 and 9.11 show the box-plot and violin distribution of predictability values.

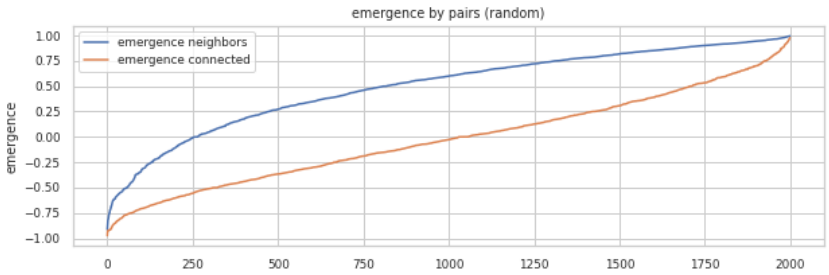


Figure 9.9: *The Average Predictability Values*

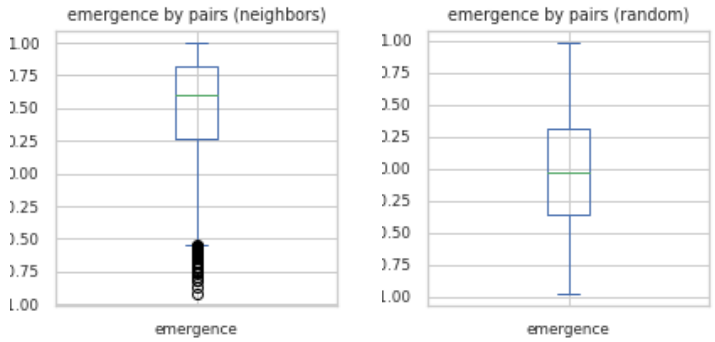


Figure 9.10: *Box-plot distribution of predictability values.*

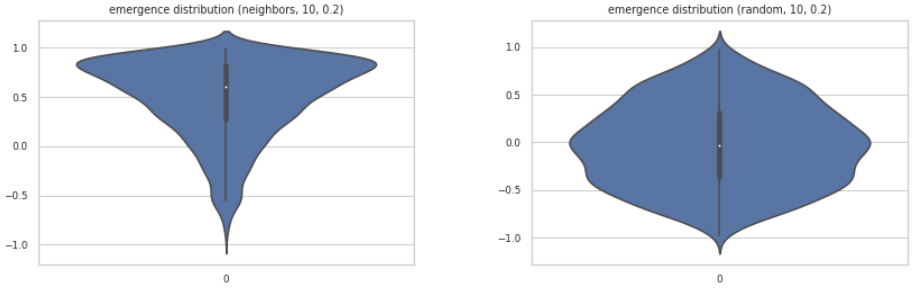


Figure 9.11: *Violin distribution of predictability values.*

By Equation (9.3), we can observe that in average (i) 75% of emerging topics have $1.25/0.75 = 1.66$ times more publications after emergence than before and (ii) 50% of *EmbeddingContext*, have $2.6/0.4 = 6.2$ more publications after emergence than before, whereas the ratio is 1 for *CitationContext*.

$$\frac{|D_{future}(t_e)|}{|D_{past}(t_e)|} = \frac{\mathcal{E}(t_e)+1}{1-\mathcal{E}(t_e)} \tag{9.3}$$

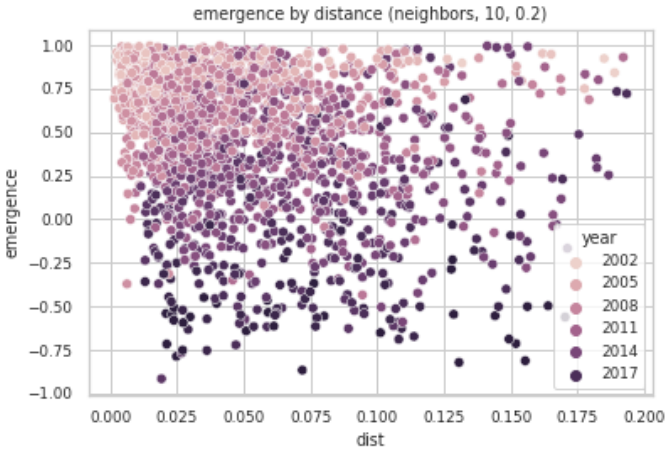


Figure 9.12: *The distance distribution of emerging topics.*

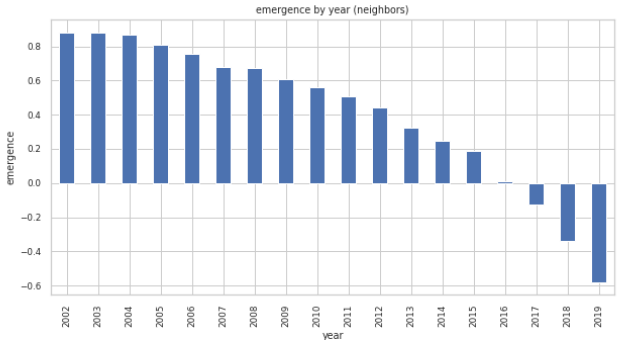


Figure 9.13: *The average of predictability values by year*

9.6.1 Emerging Topic Properties

Figure 9.14 shows the correlation between various parameters that shape the dynamics of emerging topics. We can see that the average embedding distance (*dist*) per period increases in time (*year*). This can be explained by the fact that as we reach the end of the dataset, we have less contextual information about the future, and correspondingly the average embedding distance increases. This signifies that the applied dynamic embedding method estimates that the embedding distance of emerging topic pairs increases after their emergence. However, this conclusion has to be confirmed by a deeper analysis of the bias introduced by the dynamic computation algorithm. Second, the predictability (*emergence*) decreases with increasing distance and *strongly decreases* in time (see also Figure 9.12). This is a natural consequence of the definition of emergence which compares the number of relevant documents before and after the emergence period. This number is also influenced by the "relative length" of the past and the future covered by the archive (as shown in Figure 9.13, the average emergence of topic pairs is positive before 2016 and becomes negative afterward). Finally, we can see that the average cluster size (*all*) of emerging topics is independent of the period, the average predictability, and the average embedding distance.

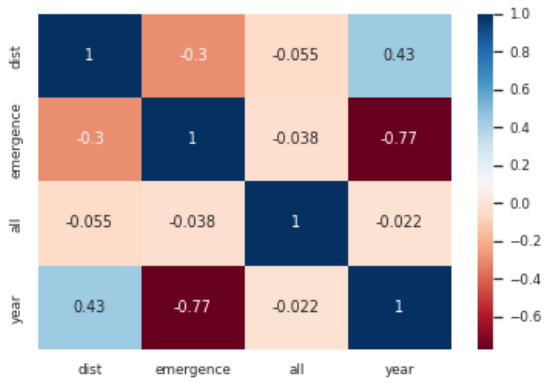


Figure 9.14: *Correlation between Emerging Topic Properties*

Figure 9.15 shows the correlations between the average number of new emerging topics (*EmbeddingContext*) by period (n), the average number of randomly connected pairs (*CitationContext*) by period (c), and the average number of connected emerging topics (intersection of *EmbeddingContext* and *CitationContext*) by period (cn). We can see that the average number of embedding neighbors decreases with time, which is consistent with the observation that the embedding distance increases. The fraction of connected neighbors is independent of the number of neighbors but increases with the number of connected topics.

9.7 Conclusion

In this chapter, we have introduced the ATEM framework, a novel approach to Automatic Topic Emergence Monitoring. ATEM leverages dynamic topic models and graph embedding techniques to uncover emerging topics within scientific archives. By projecting document citation links into the topic space and creating dynamic topic-citation graphs, ATEM provides a comprehensive view of the evolution of scientific topics over time. The ATEM framework offers several key contributions. First, it introduces a robust model for understanding the evolution of science, which is crucial for strategic resource allocation and informed decision-making in research funding and public policy. Second, ATEM enables the detection of emerging topics through a multifaceted analysis of the evolution of science, addressing the limitations of single-domain and cross-domain

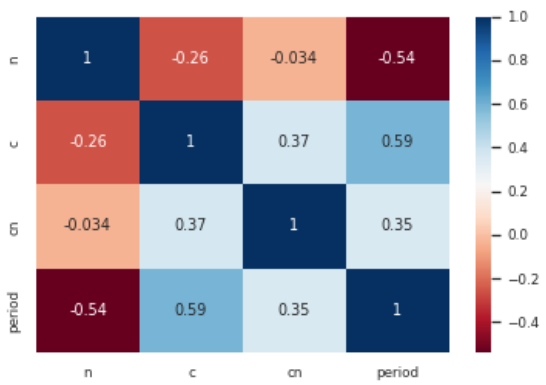


Figure 9.15: *Correlation between Embedding and Citation Context*

approaches. Third, the framework provides a detailed exploration of the application of ATEM to emergent topic detection and presents the implementation and proof of concept.

Through the use of dynamic graph embedding, ATEM successfully identifies emerging topics and predicts future interdisciplinary trends. The experiments conducted demonstrate the effectiveness of ATEM in identifying emerging topics compared to traditional co-citation analysis. The predictability metric used in the experiments highlights the potential of ATEM to anticipate new areas of study and potential breakthroughs. In conclusion, ATEM represents a significant advancement in the field of topic evolution analysis. Its ability to uncover emerging topics and provide valuable insights into the identification of new research directions makes it a powerful tool for researchers, policymakers, and stakeholders in the scientific community. The future development and application of ATEM hold promise for further enhancing our understanding of the dynamic nature of scientific evolution. In the next chapter, we will expand our horizon by investigating paradigm shift.

Chapter

10 Topic Evolution Analysis with Paradigm Shift

Understanding paradigm shifts is crucial for comprehending the dynamic nature of scientific progress and innovation. These shifts often result from discoveries or perspectives that cannot be explained by the existing paradigm. As scientists grapple with these anomalies, they are forced to rethink established theories and methods, leading to the development of new frameworks that better account for the observed phenomena. This rethinking process not only clarifies previously misunderstood or overlooked aspects of a field but also introduces entirely new questions and areas of research. For example, when the heliocentric model of the solar system replaced the geocentric model, it didn't just change how people viewed the universe; it also paved the way for new branches of science, such as astrophysics and celestial mechanics. This chapter presents preliminary findings from a practical experiment aimed at extracting topic evolution patterns based on paradigm shifts. It is important to note that the contributions outlined here are still in progress and have not yet undergone peer review or official validation.

Chapter content

10.1 Introduction

As discussed in Section 8.4, a paradigm shift is a dynamic process wherein established theories and methodologies face challenges from anomalies and breakthrough discoveries. This transformation occurs through the collective efforts of scientists engaged in both theoretical and experimental

research, ultimately leading to the collapse of protective hypotheses. The result is a radical deviation from established norms, paving the way for new perspectives in the field.

Inspired by characteristics discussed in Section 8.4.1, we propose a new definition for paradigm shift and present a novel framework for paradigm shift-based topic evolution discovery. This framework builds on the insights and methodologies developed in ATEM and ANTM discussed in Chapters 6 and 9. The proposed model aims to identify and analyze paradigm shifts, offering a systematic approach to understanding how topics evolve and new scientific domains emerge. The ultimate objective is to identify the transition path between evolving topics by extracting paradigm shifts, as we hypothesize that this approach offers a more accurate depiction of topic evolution compared to traditional citation-based analyses.

10.2 QuTE Framework

10.2.1 Paradigm shift patterns and scores

To establish a framework for extracting paradigm shift patterns, we begin by mathematically defining a paradigm shift pattern.

Definition 10.2.1 (paradigm shift pattern)**.** A *paradigm Shift pattern* ps is a chronologically ordered sequence of topics $ps = [t^0, \ldots, t^n]$ of length n. In this pattern, each topic t^j within the sequence is strongly influenced by the topic that follows it, differs chronologically from those that came before it, and gradually transforms over time into a radically different, metamorphic topic compared to the original one.

As an example, Figure 10.1 illustrates a paradigm shift pattern involving four topics ($n = 4$), transitioning from a topic within the domain of "game theory" to one within "traffic simulation and analysis." Each topic is influenced by, yet distinct from, its predecessor, with the final topic being significantly different from the initial one. This transition highlights the evolving focus from strategic decision-making models to more complex, real-world applications where individual and collective behaviors are simulated to optimize traffic flow and predict congestion patterns.

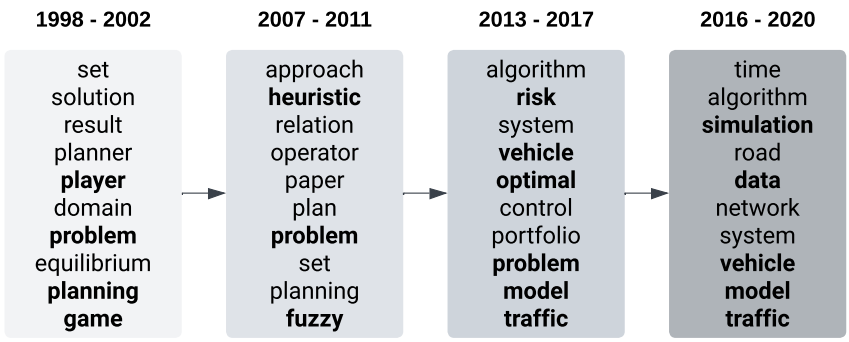


Figure 10.1: *A Paradigm Shift Pattern*

The mathematical formulation of Definition 10.2.1 is a function that assigns a paradigm-shift score to a given sequence of topics ps of length n. This function can rank and retrieve paradigm shift patterns in scientific archives.

Definition 10.2.2 (paradigm shift score)**.** Let $ps = [t^0, \ldots, t^n]$ be a sequence of evolving topic nodes in some topic citation graph $E_{\mathcal{T}}$ (see Equation (9.1) in Definition 9.2.1). Then, the paradigm

shift score $Score(ps)$ of ps is defined as follows:

$$Score(ps) = \sum_{j=2}^{n} \frac{log_2(\mathrm{I}(t^j, t^{j-1}))}{S(t^j, H(t^j)) \times |K^j - K^{(j-1)}|} \tag{10.1}$$

where:

- K^j represents the timeframe represented by topic t^j;
- $H(t^j)$ is the history of the topic t^j in ps, which is equal to the average vector representation of the topics before t^j:
$$H(t^j) = \mathrm{Mean}([V_{t^0}, \dots V_{t^{(j-1)}}]) \tag{10.2}$$
- $I(t^j, t^{j-1})$ estimates the *influence* of topic t^j to previous topic t^{j-1} by the number of citations from topic t^j to tpîc t^{j-1}:
$$\mathrm{I}(t^j, t^{j-1}) = w \text{ where } w \text{ is the number of citations from } t^j \text{ to } t^{j-1} \text{ in } E_{\mathcal{T}} \tag{10.3}$$
- $S(t^i, t^j)$ is the *similarity* between topics t^i and t^j defined by the cosine similarity between the vector representation V_{t^i} and V_{t^j}.
$$\mathrm{S}(t^i, t^j) = \mathrm{cosine}(V_{t^i}, V_{t^j}) \tag{10.4}$$

The main hypothesis behind the paradigm shift score in Definition 10.2.2 is that the more citations exist from recent (high w), but dissimilar (low S), topics t^i to previous topics t^j $(j < i)$, the higher is the likelihood that a substantial wave of anomalies or challenges has been recognized within the outdated scientific framework. We estimate the paradigm change – or the dissimilarity concerning its past – of a topic t^j by the inverse of the similarity between t^j and the average vectors of all previous topics. A lower semantic similarity suggests a higher likelihood of a potential paradigm shift. Moreover, we aim to identify scenarios where the semantic disparity is significant within relatively small time gaps, without ignoring semantic changes within scientific frameworks separated within longer time periods. This is represented by the expression $|K^j - K^{j-1}|$ which represents the time gap between two subsequent topics.

10.2.2 Extracting paradigm shift patterns with reinforcement learning

There are several ways to extract paradigm shift patterns. One approach is to use a greedy algorithm to select a random evolving topic node in the topic-citation graph, and then select other (future) nodes that maximize the defined scoring function. Automation of this approach can be achieved by framing it as a Reinforcement Learning (RL) [259] problem. RL enables the efficient extraction of hidden patterns while training an agent to recognize and identify paradigm shifts.

RL guides the agent in navigating the topic-citation graph, identifying significant shifts in scientific knowledge, and uncovering hidden patterns indicative of paradigm shifts. To implement this RL framework, we define the environment, the agent, the state space, the action space, the reward function, and the transition dynamics as follows.

10.2.3 Environment

The RL environment simulates the evolving topic-citation graph G, which represents the evolution of the scientific literature over time. It is defined by a state space S and an action space A.

State Space

The state space S consists of all possible states (t, k) where t is an evolving topic and k is a time-frame. The current state is represented by (t_c, k_c) which is set to initial state (t_i, k_i) at the start. An episode is denoted as σ which is a list of all previously explored states, initially containing $E(t_i, k_i)$. Each state has 3 attributes that characterize the episodes. These attributes are as follows

- **Similarity** s which is the cosine similarity between current state $E(t_c, k_c)$ and the next state $E(t_n, k_n)$, initially set to 1,
- **Revolution** ρ which represents contextual metamorphosis and is the cosine similarity between initial state $E(t_i, k_i)$ and the next state $E(t_n, k_n)$, initially set to 1,
- **Termination flag** d indicating the end of an episode. initially set to False.

Action Space

The action space is characterized by a *Step Function* that allows the environment to change its current state (t_c, k_c) to the next state (t_n, k_n) by executing the action of moving into the next state (t_n, k_n). Executing the step function adds the embedding of (t_n, k_n) to σ, as well as update s, ρ and d. Setting the termination flag d at *True* prohibits invoking the *Step Function* without *resetting the environment.* However, the Termination flag d is set to *True* in the following prohibited scenarios p:

- **Non-permissible actions:**These actions where $k_n \geqslant k_c$ are not allowed because our model's approach for identifying paradigm shift patterns involves moving from the present to the past without revisiting the same time-frame or jumping forward to the future again. These actions are prohibited to maintain the integrity of the model's chronological analysis, ensuring a unidirectional exploration of time.
- **Non-evident actions**: Since the action space A contains all possible combinations of evolving topics and time-frames, sometimes, the action picked by an agent has no corresponding edge (t_c, t_n, k_n) in the evolving topics-citation graph G, and therefore, are prohibited.
- **Terminal actions** k_n: The episode terminates because the next state is situated in the oldest time-frame, and there are no older states to consider further in the analysis.

Notation	Description
S	Observation space (t, k) of evolving topic t at time-frame k
A	Action space (t, k) to go to evolving topic t at time-frame k
G	Evolving Topics Citation Graph
E	Set of Evolving Topics embeddings
L	Number of time-frames
σ	List of embeddings of previously visited states
r	Score value of going from current to next state
d	Episode termination flag
s	Cosine similarity between (t_c, k_c) and (t_n, k_n)
ρ	Cosine similarity between (t_i, k_i) and (t_n, k_n)

Table 10.1: *Notation used for modeling the environment of the RL Problem*

- **Revolution actions** ρ: When the cosine similarity between the next state and the initial state is below a threshold ρ_h, it signifies the emergence of a normal science that is significantly different from the initial state. In this scenario, there's no need to proceed further, as a substantial shift in the scientific paradigm has already been identified.

Reward Function

The *step function* returns the new current state, a score r calculated via a *reward function*, and the new value of the termination flag d. Resetting the environment consists of picking a new random initial state, emptying σ, and resetting the values of s, ρ, and d. The reward function, which associates a score for each step taken in the environment by an agent, is defined as follows :

$$R(t_c, k_c, t_n, k_n) = \begin{cases} -\infty & \text{if action is prohibited} \\ \frac{log_2(\mathrm{I}(t^n,\ t^c))}{S(t^n,\ H(t^n)) \times |k^n - k^c|} & \text{else} \end{cases}$$

This reward signal tells us how good is the action taken by an agent in terms of finding a potential paradigm shift pattern.

Transition Dynamics

The transition dynamics describes how the environment changes states in response to agent actions. This is governed by the step function, which updates the current state, the episode history, the similarity, revolution, and termination flag based on the chosen action and the environment's constraints. The process starts by initializing the dynamic graph G, set E, and integer L to set up variables and select an initial state (t_i, k_i). It then sets the current state to (t_i, k_i), initializes σ with $E(t_i, k_i)$, and sets ρ, s, and d to their initial values. During each step, the algorithm processes the action (t_n, k_n); if $k_n = 1$, it sets the termination flag d to True. Otherwise, it calculates a reward r, updates ρ, and adjusts σ. If ρ drops below a threshold, d is set to True. The function returns the updated state, reward, and termination flag d. The 'reset' function clears σ and restores ρ, s, and d to their initial states.

To describe the Transition Dynamics more concisely, we have provided the following pseudo-code describing its general functioning :

10.2.4 Agent

In reinforcement learning, an agent interacts with an environment to learn a strategy that maximizes a reward. The agent makes decisions based on observed states, receives feedback in the form of rewards, and updates its strategy over time to optimize its actions. This iterative process involves exploring different actions and gradually improving the agent's decision-making capabilities by learning from the consequences of its actions. In the context of PS pattern extraction, the agent navigates the citation graph and makes decisions about which nodes to explore next in order to uncover patterns indicative of paradigm shifts. To achieve this goal, we employ Q-learning, allowing the agent to store expected future rewards for state-action pairs in a table. This table is updated iteratively based on the rewards obtained from the environment. To optimize results, it balances exploration and exploitation through a decreasing exploration rate ϵ. We use the following notation to describe the agent's behavior.

The outcome of the learning process is captured in a Q-Table, where the states and actions correspond to evolving topics across different time frames. The value of a given state-action pair in this table, once normalized, represents the probability that a paradigm shift path exists, connecting these two evolving topics within their respective time frames. Before engaging in any

Algorithm 1: Transition Dynamics

```
INPUT(Dynamic Graph : G, Set : E, Integer : L)
INITIALIZE A and S using G and L
Randomly choose initial state (t_i, k_i) from S
(t_c, k_c) ← (t_i, k_i)
σ ← [E(t_i, k_i)]
ρ ← 1, s ← 1, d ← False
Function step (action : (t_n, k_n)) :
    if k_n = 1 then
        d ← True
    else
        r ← f(t_c, k_c, t_n, k_n)
        s ← S_C(E(t_c, k_c), E(t_n, k_n))
        ρ ← S_C(E(t_i, k_i), E(t_n, k_n))
        (t_c, k_c) ← (t_n, k_n)
        σ ← σ ∪ [E(t_c, k_c)]
        if ρ ⩽ ρ_h then
            d ← True
    return (t_c, k_c), r, d
Function reset() :
    σ ← empty, ρ ← 1, s ← 1, d ← False
```

training, the agent will carry out a process that will allow it to penalize non-permissible and non-existent state transitions by updating the Q-table. This process is called *Whispering*, in which the agent sets Q-Values to $-\infty$ for any state-action pairs where the action is non-permissible given or there is no existing edge in the graph between the representation of current and next states.

Following that, the agent is capable of selecting an action from the action space A employing a Dynamic-ϵ-Greedy policy and referencing the Q-Table. Subsequently, it executes the selected action within the environment, acquiring the current environment state, the next state (chosen action), and the reward, a tuple referred to as an *experience*. The agent subsequently proceeds to update its Q-Table by leveraging its *learn function* based on this obtained experience. To sum up the structure of the agent, we present the following pseudo-code :

After setting up the environment and the Q-learning agent, the agent enters a training loop consisting of a predetermined number of episodes. During each episode, the agent engages with the environment until the episode concludes via its termination flag d. In each time step of the episode, the agent chooses an action utilizing its Q-Table and an $\epsilon - greedy$ policy. The action is then executed in the environment, leading to the subsequent state and a reward, both of which are employed to update the agent's Q-Table via the *learn* function. After finishing the training loop, we obtain Q^*, the Q-Table we would use to extract paradigm shift patterns. Here's the pseudo-code explaining the interaction between the agent and its environment :

Table 10.2: *Notation used for training the agent*

Notation	Description
Q	Q-Table of dimensions $card(S) \times card(A)$
γ	The discount factor
ϵ	The exploration rate
ϵ_{decay}	Exploration rate decay factor
ϵ_{min}	Minimum exploration rate value
Δ	Dictionary assigning to each state (t, n) its index in the Q-Table

Algorithm 2: Q-Learning Agent

Function *INPUT(Graph : G, Set : S, Set : A)* **:**
    Initialize Q with zeroes
    Execute *whispering* process

Function *select_action(State :* (t_c, k_c)*)* **:**
    $x \leftarrow$ random number between 0 and 1
    if $x \leqslant \epsilon$ **then**
        Randomly Choose action leading to next state (t_n, k_n)
    else
        Select action (t_n, k_n) that gives highest value in current state (t_c, k_c)
    return (t_n, k_n)

Function *learn(experience)* **:**
    $s_c, s_n, r \leftarrow experience$
    $Q[\Delta(s_c), \Delta(s_n)] \leftarrow r + \gamma \cdot \max(Q[\Delta(s_c), :])$
    if $\epsilon > \epsilon_{min}$ **then**
        $\epsilon \leftarrow \epsilon \cdot \epsilon_{decay}$

Algorithm 3: Training the agent on the environment

Initialize environment *env*
Initialize Q-Learning agent *agt*
for _ *in range* $nb_{episodes}$ **do**
    finished = $False$
    $env.reset()$
    while *not finished* **do**
        $c_s \leftarrow env.current\_state$
        $action \leftarrow agt.choose\_action(c_s)$
        $r, d \leftarrow env.step(action)$
        $experience \leftarrow \Delta(c_s), \Delta(action), r$
        $agent.learn(experience)$
        $finished \leftarrow d$

10.3 Experimental Setup

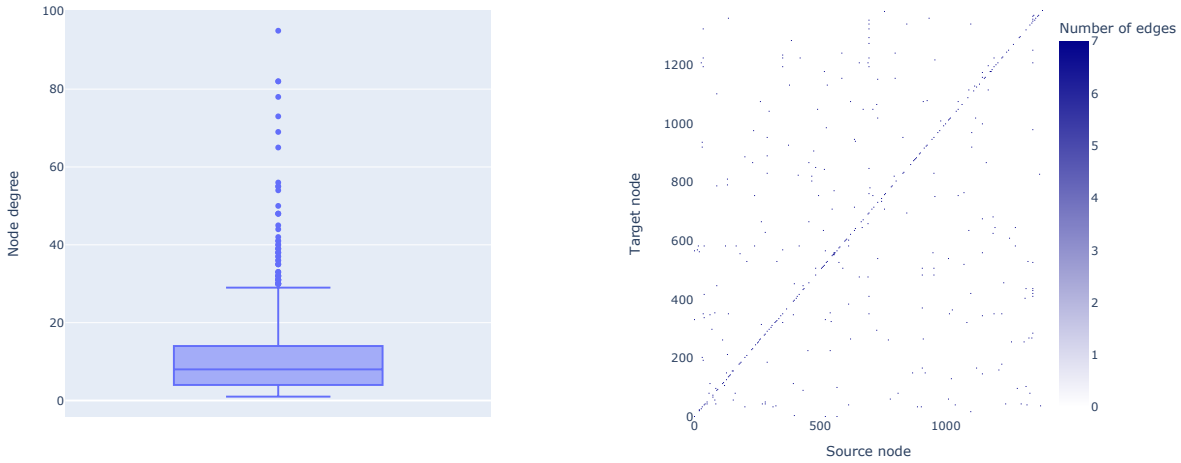
10.3.1 Dataset

We use a DBLP dataset of 500,000 documents dating from 1998 to 2020 for the experiment. Each document is identified by a unique ID and has its list of references, which is also a list of the IDs of other documents it cites. This dataset serves as a basis for analyzing paradigm shifts within the scientific literature, providing a diverse range of documents to explore the evolution of scientific ideas over time.

10.3.2 Creating Evolving Topic-Citation Graph for RL environment

The DBLP dataset undergoes processing via ATEM [250] to construct the evolving topic-citation graph relevant to the environment of the RL problem. ATEM necessitates dynamic topic modeling, for which we've employed ANTM [127]. The ANTM configuration entails embedding through Data2Vec and partitioning documents into seven time-frames of five years each, with a two-year overlap between consecutive frames. The experiment utilizes specific settings aimed at achieving superior dimensionality reduction and clustering quality, outlined as follows:

The evolving topics produced by ANTM are then used to create a citation graph of 1185



(a) *Distribution of node degree* **(b)** *Adjacency matrix of edges*

Figure 10.2: *Evolving Topic-Citation graph statistics*

nodes (each node representing an evolving topic throughout the years), and 6322 citation edges (t_x, t_y, j) where there exists at least one citation from a document within $D(t_x^j)$ to a document withing $D(t_y^k)$ and $0 < k < j < 7$.

To gain insights into the evolving topic-citation graph's structure, we perform statistical analysis to assess its topology. While a majority of nodes self-reference, which is typical within a semantically similar evolving topic, nodes also cite other evolving topics. On average, each node references around 5 to 15 other nodes.

While nodes tend to have the highest number of citations when they reference themselves primarily, it might raise concerns about bias in the RL agent's score function, given that citations serve as the numerator. However, this potential bias is offset by the denominator because there is generally greater similarity within a single evolving topic over time, in contrast to two entirely distinct evolving topics.

10.3.3 Training phase

In this crucial phase, we lay the foundation for our RL environment, leveraging the evolving topic-citation graph achieved in the preceding stage. This graph serves as the dynamic backdrop against which the QuTE agent will learn and adapt. Central to this setup is the injection of Revolution actions, denoted as $\rho_h \leqslant 0.2$. These actions symbolize the emergence of a new scientific paradigm,

Table 10.3: *The configuration of ANTM for Dynamic Topic Modeling*

ANTM config	Parameter	Value
Splitting	Window size	5
	Overlap	2
UMAP dimension reduction	Dimension	5
	Number of neighbours	15
	Minimum distance	0.3
HDBSCAN clustering	Minimum cluster size	10

significantly diverging from the initial state. By incorporating such disruptive events, we aim to equip our agent with the capacity to navigate and thrive in evolving knowledge landscapes. Our training regimen spans an extensive 10 million episodes, denoted as $nb_{episodes} = 10,000,000$. Throughout this training process, we employ an epsilon-greedy exploration strategy, a cornerstone technique in RL, to balance the agent's exploration of new strategies with its exploitation of existing knowledge. Initially, we set epsilon (ϵ) to 1, reflecting maximum exploration. Over time, we gradually decay epsilon using a decay factor ($\epsilon_{decay} = 0.99$) until it reaches a minimum value ($\epsilon_{min} = 0.01$), thus ensuring a gradual shift towards exploitation while retaining a degree of exploration to prevent premature convergence to sub-optimal solutions. Furthermore, we define the discount factor, γ, which plays a crucial role in determining the importance of future rewards. A value of $\gamma = 0.95$ is chosen to balance immediate rewards with future potential gains. This choice helps guide the agent towards making decisions that aim to optimize long-term cumulative rewards by giving some weight to future outcomes without overly discounting them.

10.3.4 Data preparation after training

The Q-table generated after training the QuTE agent has a substantial size of 8295 by 8295. This large dimension presents challenges for result evaluation using the metrics described in Section 10.3.5, as it would require prompting a large language model (LLM) 60 million times to fully test the table. Randomly selecting a subset of states is also not ideal, as it may fail to represent the overall outcome accurately. To simplify the process, we aggregate states over time by applying a summation kernel to the Q-table, summing the values from each evolving topic to others across all time steps. We then normalize these values relative to all topics, reducing the table size to 1185 by 1185.

Summation Kernel: Let $Q \in \mathbb{R}^{8295\times8295}$ be the Q-table and $K \in \mathbb{R}^{7\times7}$ be a kernel where $K_{ij} = 1$ for all i, j. We define $R \in \mathbb{R}^{1185\times1185}$ such that for $i, j \in 0, 1, \ldots, 1184$, $R_{ij} = \sum_{m=0}^{6}\sum_{n=0}^{6} Q_{7i+m,7j+n}$.

This kernel operation represents the convolution of the Q-table with a 7x7 kernel of ones, resulting in an 1185x1185 matrix R. This approximation facilitates the comparison of results obtained from the Q-table to transitions via citation between documents of topics, which is the prevalent method for modeling evolution between two topics. Subsequently, we eliminate self-evolution for each topic, signifying the removal of diagonal values, as the focus lies on deriving evolution from one topic to another distinct topic. Finally, we normalize the values by summation for each state and divide the values by the average of each row of the simplified Q-table.

Normalization Method: Let $R \in \mathbb{R}^{n\times n}$ be the simplified Q-table, where n is the number of rows and columns in R. We initially define a normalized matrix N as $N_{ij} = \frac{R_{ij}-\min(R)}{\max(R)-\min(R)}$ for all $i, j \in 0, 1, \ldots, n-1$. However, we then redefine N as follows: For each row i, we compute $S_i = \sum_{j=0}^{n-1} R_{ij}$. If $S_i \neq 0$, we set $N_{ij} = \frac{R_{ij}}{S_i}$ for all $j \in 0, 1, \ldots, n-1$. If $S_i = 0$, we leave N_{ij} unchanged.

This operation performs row-wise normalization on the matrix R, where each element in a row is divided by the sum of that row, except for rows where the sum is zero. Afterward, we select the subsequent topic as the argmax for each observed evolving topic, facilitating the exploration of evolution patterns between evolving topics across multiple periods with a certain probability distribution. Finally, we employ the TF-IDF method to extract a set of 10 representative words for each identified pattern, aiding in the characterization of these transitions.

10.3.5 Validation metrics

The output of training the Q-learning agent can be represented as a matrix of rewards, denoting the value associated with transitioning between states. To ascertain its veracity, two validation metrics are used:

Contextualized Topic Rating Metrics

$\text{CTC}_{\text{Rating}}$[116] evaluates an evolution path from an state s_i to the next state $s_n = argmax(s_i)$ where s_0 represents the initial state, argmax(s_0) denotes the subsequent state, R represents the word representation, and $CTC(R(s_i, s_n))$ signifies the rating of evolution from state s_i to the next state s_n using Instruction-based LLMs such as ChatGPT and request it to provide evaluations by assigning scores ranging from 0 to 10. For this metric, we utilized "GPT-3.5-turbo," and the specific prompt used in this evaluation method is as follows:

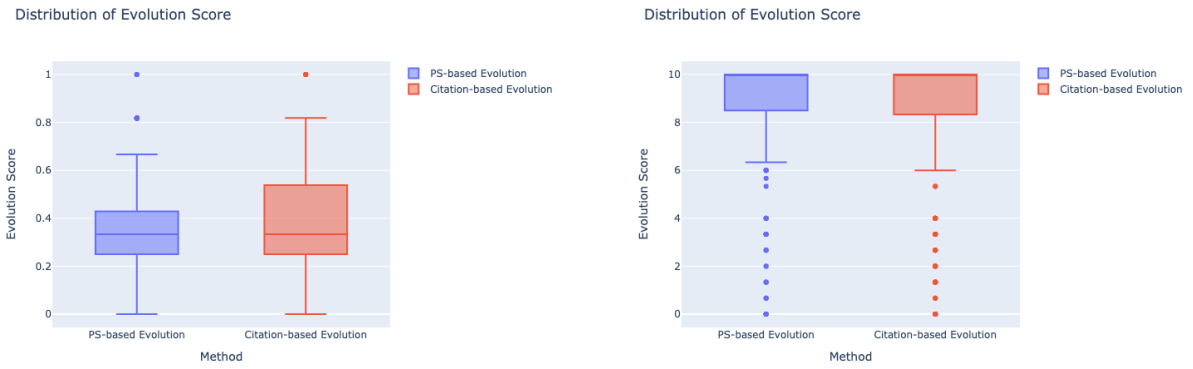
> **System Prompt**: Evaluate the evolution score of the following sequence of two topics, which represent a paradigm shift and scientific transformation from one scientific field to another, and give a score based on a 10-point scale, where 10= " big fundamental change with some similarity and relation" and 0= " no change at all or topics are completely distinguished", how well they match each other and how the pattern can be used to retrieve a change from one topic to another. The two topics are represented with the following keywords in chronological order: [topic-words (1) -> topic-words (2)]. Consider the clarity, coherence, and relevance of each topic to the others and the broader context, and respond in the following format: Score: <score> without explanation.

Jaccard Similarity

The use of Jaccard Similarity ensures that the evaluation considers the alignment between expected transitions and contextualized assessments (the higher the better). This approach offers scalability and flexibility across different tasks and environments, enabling a comprehensive understanding of the agent's performance and facilitating informed decision-making in evaluating its efficacy.

10.3.6 Baseline

To establish a baseline for comparison with our proposed approach, we employ a method that leverages an adjacent topic-citation graph. This method assesses the likelihood of one topic evolving into another, serving as a benchmark against which we can evaluate the effectiveness of our RL-based approach. The process involves computing probability weights derived from the graph's adjacency matrix. Specifically, we aggregate the number of citations between all evolving topics across all time frames, resulting in a new table where each row and column corresponds to an evolving topic. This table, like the simplified Q-table, has dimensions of 1185 by 1185. Since we aggregated data over time, this process might inadvertently amplify the effect of self-evolution within evolving topics, potentially skewing the evaluation results. To mitigate this issue, we first eliminate instances of self-evolution for each topic. Afterward, to normalize our table and accurately calculate the probability of evolution between any two topics, we divide each element in a row by the argmax value of that row. This approach ensures that the influence of self-evolution is removed and the probability distribution across topics is properly scaled. From this matrix, we discern the most probable evolution pattern through time by selecting the argmax value. This methodology not only provides insights into the dynamics of topic evolution but also furnishes a systematic approach to understanding the inter-connectedness and progression of diverse subject



(a) *Jacquard Similarity* **(b)** *CTC*

Figure 10.3: *Distribution of Evolution Scores Achieved by proposed method vs citation-based approaches*

matters within the given domain. Finally, we employ the TF-IDF method to extract a set of 10 representative words for each identified pattern.

10.4 Results

Our proposed method outperforms the traditional citation-based approach in identifying scientific evolution patterns, as evidenced by the comparative data presented in Table 10.4. While the citation-based approach relies solely on citation links, our method employs a more comprehensive analysis, resulting in superior performance. For a more nuanced understanding of this performance difference, we direct readers to Figures 10.3(a) and 10.3(b), which illustrate the distribution of evolution scores achieved by both methods. These figures provide a visual representation of the enhanced capability of our proposed approach in capturing complex relationships and evolutionary trajectories.

Table 10.4: *Comparison of PS-based and Citation-based Methods*

Method	Jacquard Similarity	CTC	Probability	Average
PS-based	0.363	8.559	0.522	0.197
Citation-based	0.388	8.533	0.425	0.172

To illustrate some of these evolutionary transitions, we present the top evolution path that achieved the highest score according to Definition 10.2.2. As shown in Table 10.5, the pattern with the highest score is characterized by a significant similarity among the topic words within the tuple. This table includes the initial and subsequent topic numbers, time-frames, keywords associated with each topic, and transition scores. The topics are ranked by their transition scores, with higher scores indicating a stronger predicted evolution between topics. Despite the substantial overlap among these words, examining the timeline of publications and their corresponding citations reveals a clear evolution over time. This evolution delineates a paradigm shift within the field, driving the community to work on graphs toward new topics and marking a notable transformation in the subject matter over time.

Table 10.5: *Top Identified Topic Transitions*

#	Topic	Time	Topics words	Score
1	344	6	network, disaster, node, communication, approach, routing, wireless, system, protocol, infrastructure	93
	1145	5	network, node, sensor, energy, data, algorithm, routing, protocol, cluster, sink	
2	793	7	packet, classification, algorithm, network, rule, filter, traffic, memory, search, performance	90
	558	6	node, network, key, attack, scheme, sensor, security, protocol, routing, wireless	
3	69	6	access, policy, control, model, system, role, security, rbac, user, constraint	83
	69	5	access, policy, control, security, system, data, information, model, user, privacy	
4	260	4	image, quality, fusion, video, metric, method, visual, proposed, algorithm, measure	82
	260	3	image, quality, fusion, metric, video, method, visual, proposed, model, measure	
5	148	7	graph, processing, algorithm, model, system, partitioning, string, result, distributed, number	82
	745	6	data, system, iot, application, processing, query, stream, performance, distributed, model	
6	319	4	logic, proof, system, program, calculus, rule, set, semantics, sequent, answer	81
	308	3	logic, set, program, simulation, answer, system, problem, reasoning, model, programming	
7	587	6	spectrum, sensing, user, pu, power, secondary, interference, channel, network, primary	81
	608	5	user, channel, proposed, allocation, algorithm, resource, solution, cell, scheme, interference	
8	443	7	writing, learning, citation, research, knowledge, resource, patent, technology, information, data	81
	443	5	patent, paper, research, innovation, field, technology, economy, data, number, network	
9	914	7	array, estimation, doa, algorithm, method, signal, proposed, source, matrix, sparse	80
	20	6	signal, algorithm, matrix, sparse, measurement, sensing, method, recovery, problem, reconstruction	

10.5 Conclusion

In conclusion, this chapter has delved into the critical role of paradigm shifts in the evolution of scientific topics, offering a comprehensive model for systematically identifying and analyzing these shifts. By leveraging the Kuhn Cycle, we have outlined a structured framework for understanding the phases of scientific progress, from pre-science to paradigm change. Our model emphasizes the characteristics of collectiveness, sequential cumulativeness, causal interdependence, and contextual metamorphosis, ensuring a holistic approach to the study of scientific evolution. The application of reinforcement learning (RL) techniques allows for the efficient extraction of hidden patterns within scientific literature, enabling the identification of significant paradigm shifts. Experimental results using the DBLP dataset and techniques like ATEM and ANTM demonstrate the model's effectiveness in capturing the nuanced dynamics of topic evolution, outperforming traditional citation-based methods. The validation metrics used confirm the accuracy and reliability of our findings. Ultimately, the methods and models presented provide valuable insights into the emergence of new research areas and the continuous evolution of scientific knowledge, underscoring the importance of understanding paradigm shifts to foster innovation and drive scientific progress forward.

Chapter

11 Outlooks and Conclusion

The main objective of this thesis is to develop advanced methodologies for analyzing the evolution of scientific knowledge. It aims to establish a formal framework for science evolution analysis by introducing new concepts and incorporating deep learning techniques. The thesis proposes novel approaches such as ANTM for discovering evolving topics, CTC for evaluating topic models, ATEM for identifying emerging topics, and QuTE for modeling paradigm shifts. These contributions seek to provide more accurate and nuanced insights into the complex dynamics of scientific knowledge evolution, ultimately advancing the field of science evolution analysis.

In this chapter, we summarize the main contributions of the thesis and highlight the key advancements of our research, and provide a roadmap for the future of scientific discovery and knowledge organization. This roadmap focuses on identifying emerging topics within scientific archives, exploring their potential trajectories, and demonstrating how advanced models and analyses are poised to drive transformative trends, such as Large Language Models (LLMs). We will discuss promising avenues and anticipate the impacts these innovations may have on the scientific landscape.

Chapter content

11.1 Main Contribution

The contributions of this thesis are as follows.

CTC In Chapter 4, we introduced the Contextualized Topic Coherence (CTC) metrics, a set of innovative metrics designed to evaluate the interpretability of topic models. These metrics leverage LLMs to assess the interpretability of words within a topic, providing a more nuanced and accurate measure compared to traditional coherence metrics. We categorize these metrics into two classes. The first class, Automated CTC, takes advantage of encoder-only transformers such as BERT to predict the probability that topic words would be seen contextually next to each other. The second class, called Semi-automated CTC, utilizes decoder-only transformers and LLMs to simulate important human-based evaluations for scoring and rating topic interpretability. We also benchmarked various recent topic models, including Latent Dirichlet Allocation (LDA) and Neural Topic Models, against standard datasets like the 20 Newsgroups and Wikipedia, demonstrating the superior performance of our CTC metric in capturing the true semantic coherence of topics.

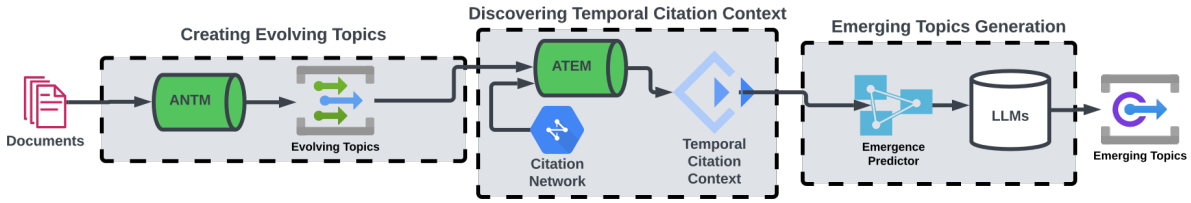


Figure 11.1: *The suggested architecture of generating emerging topics*

ANTM In Chapter 6, we tackled the challenge of identifying evolving topics in large-scale datasets by proposing a clustering-based dynamic topic model called Aligned Neural Topic Models (ANTM). This model integrates temporal information into the topic modeling process, allowing for the detection and tracking of topic evolution over time. By employing advanced clustering algorithms, our approach can dynamically adjust to the emergence of new topics and the decline of old ones, providing a more realistic and accurate depiction of topic dynamics.

ATEM In Chapter 9, we proposed a novel model called Automated Topic Emergence Monitoring (ATEM). This model detects early emerging topics by combining dynamic graph embedding techniques with large language models. This model leverages the structural information of citation networks and the semantic richness of language models to identify nascent research trends at their inception.

QuTE Lastly, in Chapter 10, we defined the concept of a paradigm shift in scientific domains and developed a reinforcement learning-based framework, called QuTE that represents topic evolution with Q-Learning. This framework uses reward mechanisms to identify and promote shifts in research paradigms, facilitating the recognition of groundbreaking changes in scientific thought and practice.

11.2 Future Works

For future work, as shown in Figure 11.1, we suggest building upon our models to address two main challenges in our future work. The first challenge is to extend ATEM by predicting the evolution of topic citations and improving the detection of emerging topics in the future. The second challenge is to improve topic representations of existing and emerging research topics, leveraging advanced large language models for scientific text generation.

11.2.1 Topic Prediction

Predicting the trajectory of scientific progress remains a complex challenge that continues to engage researchers in computer science and related fields, prompting ongoing investigations into novel computational approaches and methodologies. This process involves predicting upcoming scientific topics, assessing the current state of the art, and envisioning potential solutions. We contend that scientific topics often emerge through the collaborative synergy of knowledgeable individuals engaging in dynamic brainstorming sessions where innovative ideas are conceived.

In our thesis, we proposed ANTM to extract evolving topics to represent the dynamics of topics over time [127] and ATEM [250] to detect topic emergence by comparing the evolution of topics citation contexts within citation graphs. The current approach is efficient for detecting topics that emerged in the past. Our first direction for future work is to apply advanced deep learning methods that learn the past and current embedding states in the citation context space

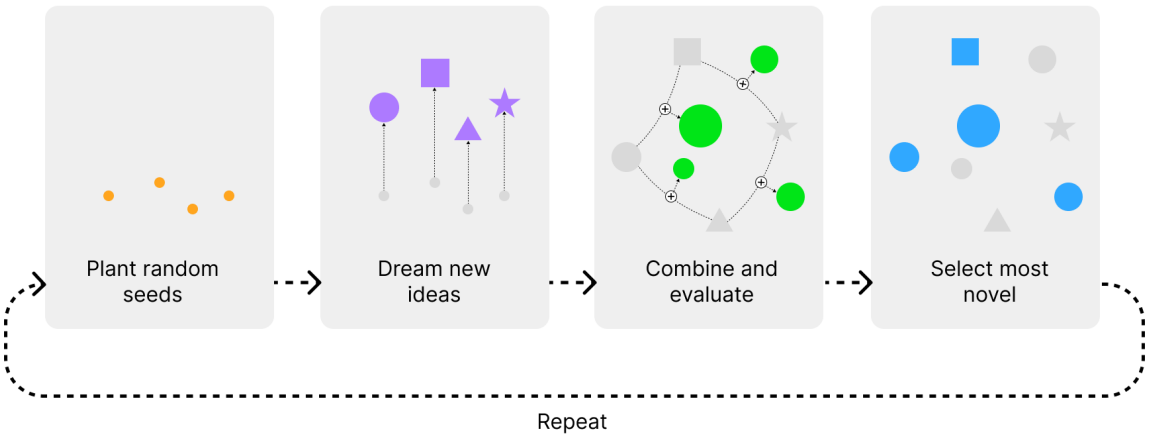


Figure 11.2: *DreamGPT Process [262]*

and produce models to predict the next embedding states in the future. One approach is to use multivariate Kalman filters [260, 261] to estimate the next embedding state and discover estimated topics that will emerge to create new topics. This methodology can create a comprehensive framework that not only identifies current scientific trends but also anticipates future directions, potentially revolutionizing our ability to forecast and navigate the complex landscape of scientific advancement.

11.2.2 Topic Representation Generation

Topic representation is fundamental for the interpretation of scientific research topics. Current methods for computing topic representations are mainly weighted term vectors (word clouds) representing various term-document distributions. This is also true for the representation of emerging topics (Chapter 9) which are obtained by interlacing the topic representations of topics with similar citation contexts. We suggest generating topic representation using LLM for emerging topics that may or may not be identified as such in the scientific literature. For emerging topics that are not yet identified as such by the research communities – meaning there are not yet enough publications to produce significant emergence predictability values (discussed in Chapter 9) – we suggest hallucinating LLMs to generate possible topic representation. LLM hallucination refers to the tendency of LLM models to produce text that appears correct but is actually incorrect or not based on the input given. Hallucination is often seen as a negative aspect of LLMs, but we aim to use it as an advantage for divergent thinking to generate new innovative topics. For example, DreamGPT [262] explores every combination of words to solve specific problems. In this study, we will propose a topic-driven hallucination of LLMs that is goal-directed over DreamGPT, which tries all possibilities.

Bibliography

References for Chapter 1: Introduction

[1] Thomas S Kuhn. *The Structure of Scientific Revolutions*. University of Chicago press, 1962 *Cited on pages 1, 73, 77.*

[2] Ying Ding and Kyle Stirling. "Data-driven discovery: A new era of exploiting the literature and data". In: *Journal of Data and Information Science* 1.4 (2017), pp. 1–9 *Cited on page 1.*

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 *Cited on pages 1, 14.*

[4] David M Blei. "Probabilistic topic models". In: *Communications of the ACM* 55.4 (2012), pp. 77–84 *Cited on pages 1, 11, 13.*

[5] Rubayyi Alghamdi and Khalid Alfalqi. "A survey of topic modeling in text mining". In: *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6.1 (2015) *Cited on pages 1, 11, 18.*

[6] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120 *Cited on pages 1, 43, 44, 49, 52, 57.*

[7] Karl Raimund Popper et al. *Objective Knowledge: An Evolutionary Approach*. Vol. 49. Clarendon press Oxford, 1979 *Cited on pages 1, 73.*

[8] Aristotle. *Metaphysics*. Ed. by W. D. Ross. The specific concept is found in Book VII (Zeta), Part 17. Oxford: Oxford University Press, 1924 *Cited on page 1.*

[9] David M. Blei and John D. Lafferty. "Dynamic Topic Models". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. New York, NY, USA: ACM, 2006, pp. 113–120. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143859 *Cited on page 2.*

[10] Huailan Liu et al. "Mapping the Technology Evolution Path: A Novel Model for Dynamic Topic Detection and Tracking". In: *Scientometrics* 125.3 (2020), pp. 2043–2090 *Cited on pages 2, 75.*

[11] Houkui Zhou, Huimin Yu, and Roland Hu. "Topic evolution based on the probabilistic topic model: a review". In: *Frontiers of Computer Science* 11 (2017), pp. 786–802 *Cited on pages 2, 44, 67.*

[12] Jianhua Hou, Xiucai Yang, and Chaomei Chen. "Emerging Trends and New Developments in Information Science: A Document Co-Citation Analysis (2009–2016)". In: *Scientometrics* 115.2 (2018), pp. 869–892 *Cited on pages 2, 70, 77.*

[13] Thara Prabhakaran et al. "Towards Prediction of Paradigm Shifts from Scientific Literature". In: *Scientometrics* 117.3 (2018), pp. 1611–1644 *Cited on pages 2, 76.*

[14] Bingshan Zhu, Yi Cai, and Haopeng Ren. "Graph neural topic model with commonsense knowledge". In: *Information Processing & Management* 60.2 (2023), p. 103215 *Cited on page 3.*

[15] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018) *Cited on pages 3, 16, 17, 29, 47.*

References for Chapter 2: Topic Models

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 *Cited on pages 1, 14.*

[4] David M Blei. "Probabilistic topic models". In: *Communications of the ACM* 55.4 (2012), pp. 77–84 *Cited on pages 1, 11, 13.*

[5] Rubayyi Alghamdi and Khalid Alfalqi. "A survey of topic modeling in text mining". In: *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6.1 (2015) *Cited on pages 1, 11, 18.*

[15] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018) *Cited on pages 3, 16, 17, 29, 47.*

[16] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022) *Cited on pages 11, 13, 28, 31, 46, 47, 49, 55, 57, 58, 88.*

[17] Maarten Grootendorst. *Topic Modeling with Llama 2*. Accessed: 2023-10-10. 2023 *Cited on page 12.*

[18] Leo Breiman. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical science* 16.3 (2001), pp. 199–231 *Cited on page 13.*

[19] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 *Cited on pages 13, 21, 22, 44.*

[20] Yee Teh et al. "Sharing clusters among related groups: Hierarchical Dirichlet processes". In: *Advances in neural information processing systems* 17 (2004) *Cited on page 13.*

[21] Laure Thompson and David Mimno. *Topic Modeling with Contextualized Word Representation Clusters*. 2020. arXiv: 2010.12626 [cs.CL] *Cited on page 13.*

[22] Mark Steyvers and Tom Griffiths. "Probabilistic topic models". In: *Handbook of latent semantic analysis*. Psychology Press, 2007, pp. 439–460 *Cited on pages 13, 14, 18.*

[23] Seyed Ali Bahrainian, Martin Jaggi, and Carsten Eickhoff. "Self-Supervised Neural Topic Modeling". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3341–3350. DOI: 10.18653/v1/2021.findings-emnlp.284 *Cited on page 13.*

[24] Dimo Angelov. "Top2vec: Distributed representations of topics". In: *arXiv preprint arXiv:2008.09470* (2020) *Cited on pages 13, 16, 17, 28, 31, 56, 58, 88.*

[25] Levent Bolelli et al. "Finding Topic Trends in Digital Libraries". In: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*. 2009, pp. 69–72 *Cited on page 14.*

[26] Jon Mcauliffe and David Blei. "Supervised topic models". In: *Advances in neural information processing systems* 20 (2007) *Cited on page 14.*

[27] David Blei and John Lafferty. "Correlated topic models". In: *Advances in neural information processing systems* 18 (2006), p. 147 *Cited on page 14.*

[28] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877 *Cited on page 14.*

[29] Akash Srivastava and Charles Sutton. "Autoencoding variational inference for topic models". In: *arXiv preprint arXiv:1703.01488* (2017) *Cited on pages 15, 35.*

[30] David Newman, Sarvnaz Karimi, and Lawrence Cavedon. "External evaluation of topic models". In: *Proceedings of the 14th Australasian Document Computing Symposium*. University of Sydney. 2009, pp. 1–8 *Cited on pages 15, 23, 24.*

[31] Yi Wang et al. "Plda: Parallel latent dirichlet allocation for large-scale applications". In: *International Conference on Algorithmic Applications in Management*. Springer. 2009, pp. 301–314 *Cited on page 15.*

[32] Zhiyuan Liu et al. "Plda+ parallel latent dirichlet allocation with data placement and pipeline processing". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011), pp. 1–18 *Cited on page 15.*

[33] Deepak Sharma, Bijendra Kumar, and Satish Chand. "A survey on journey of topic modeling techniques from SVD to deep learning". In: *International Journal of Modern Education and Computer Science* 9.7 (2017), p. 50 *Cited on page 15.*

[34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286 *Cited on page 15.*

[35] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013) *Cited on page 15.*

[36] Yishu Miao, Lei Yu, and Phil Blunsom. "Neural variational inference for text processing". In: *International conference on machine learning*. PMLR. 2016, pp. 1727–1736 *Cited on page 15.*

[37] Suman Adhya, Avishek Lahiri, and Debarshi Kumar Sanyal. "Do neural topic models really need dropout? analysis of the effect of dropout in topic modeling". In: *arXiv preprint arXiv:2303.15973* (2023) *Cited on page 15.*

[38] David E Rumelhart and Adele A Abrahamson. "A model for analogical reasoning". In: *Cognitive Psychology* 5.1 (1973), pp. 1–28 *Cited on page 15.*

[39] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828 *Cited on page 15.*

[40] Yishu Miao, Edward Grefenstette, and Phil Blunsom. "Discovering discrete latent topics with neural variational inference". In: *International conference on machine learning*. PMLR. 2017, pp. 2410–2419 *Cited on page 15.*

[41] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "Topic modeling in embedding spaces". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453 *Cited on pages 15, 21, 23, 28, 31, 35, 46, 57.*

[42] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013) *Cited on pages 16, 25.*

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543 *Cited on page 16.*

[44] Jhon Atchison and Sheng M Shen. "Logistic-normal distributions: Some properties and uses". In: *Biometrika* 67.2 (1980), pp. 261–272 *Cited on page 16.*

[45] Wenchuan Mu et al. "A clustering-based topic model using word networks and word embeddings". In: *Journal of big data* 9.1 (2022), p. 38 *Cited on page 16.*

[46] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in neural information processing systems* 26 (2013) *Cited on pages 16, 88.*

[47] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018) *Cited on pages 16, 17, 47, 53.*

[48] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17*. Springer. 2013, pp. 160–172 *Cited on pages 16, 17, 88.*

[49] Weisi Chen et al. "Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: a comparative study". In: *Electronics* 12.12 (2023), p. 2605 *Cited on page 17.*

[50] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017) *Cited on pages 16, 29.*

[51] Federico Bianchi, Silvia Terragni, and Dirk Hovy. "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 759–766. DOI: 10.18653/v1/2021.acl-short.96 *Cited on pages 16, 28, 31.*

[52] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019) *Cited on pages 16, 58.*

[53] Alexander Hoyle, Pranav Goel, and Philip Resnik. "Improving neural topic models using knowledge distillation". In: *arXiv preprint arXiv:2010.02377* (2020) *Cited on page 16.*

[54] Maarten Grootendorst. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure". In: *arXiv preprint arXiv:2203.05794* (2022). arXiv: 2203.05794 *Cited on pages 17, 83, 88.*

[55] MaartenGr. *cTFIDF - Class-based TF-IDF Implementation in Python*. https://github.com/MaartenGr/cTFIDF. Accessed on: 08 28, 2023. 2022 *Cited on pages 17, 47, 51, 55.*

[56] OpenAI. *ChatGPT: Engaging and Dynamic Conversations*. https://openai.com/blog/chatgpt. 2022 *Cited on pages 17, 28, 30.*

[57] Chau Minh Pham et al. "TopicGPT: A prompt-based topic modeling framework". In: *arXiv preprint arXiv:2311.01449* (2023) *Cited on page 17.*

[58] Lin Gui et al. "Neural topic model with reinforcement learning". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 3478–3483 *Cited on page 17.*

[59] Jeremy Costello and Marek Z Reformat. "Reinforcement learning for topic models". In: *arXiv preprint arXiv:2305.04843* (2023) *Cited on page 17.*

[60] Liang Yang et al. "Graph attention topic modeling network". In: *Proceedings of the web conference 2020*. 2020, pp. 144–154 *Cited on page 17.*

[61] Qile Zhu, Zheng Feng, and Xiaolin Li. "GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model". In: *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2018, pp. 4663–4672 *Cited on page 17.*

[62] Deyu Zhou, Xuemeng Hu, and Rui Wang. "Neural topic modeling by incorporating document relationship graph". In: *arXiv preprint arXiv:2009.13972* (2020) *Cited on page 17.*

[63] Yiming Wang et al. "Extracting topics with simultaneous word co-occurrence and semantic correlation graphs: neural topic modeling for short texts". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021, pp. 18–27 *Cited on page 17.*

[64] Haocheng Wang et al. "Topic model on microblog with dual-streams graph convolution networks". In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2022, pp. 1–8 *Cited on page 17.*

[65] Rui Wang, Deyu Zhou, and Yulan He. "Atm: Adversarial-neural topic model". In: *Information Processing & Management* 56.6 (2019), p. 102098 *Cited on pages 18, 28, 31.*

[66] Rui Wang et al. "Neural topic modeling with bidirectional adversarial training". In: *arXiv preprint arXiv:2004.12331* (2020) *Cited on page 18.*

[67] Xuemeng Hu et al. "Neural topic modeling with cycle-consistent adversarial training". In: *arXiv preprint arXiv:2009.13971* (2020) *Cited on page 18.*

[68] Rob Churchill and Lisa Singh. "The evolution of topic modeling". In: *ACM Computing Surveys* 54.10s (2022), pp. 1–35 *Cited on page 18.*

[69] Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. "A survey on neural topic models: Methods, applications, and challenges". In: *Artificial Intelligence Review* 57.2 (2024), pp. 1–30 *Cited on pages 18, 44.*

[70] Jichuan Zeng et al. "Topic memory networks for short text classification". In: *arXiv preprint arXiv:1809.03664* (2018) *Cited on page 18.*

[71] Yatin Chaudhary et al. "TopicBERT for energy efficient document classification". In: *arXiv preprint arXiv:2010.16407* (2020) *Cited on page 18.*

[72] Jonathan O Cain. "Using topic modeling to enhance access to library digital collections". In: *Journal of Web Librarianship* 10.3 (2016), pp. 210–225 *Cited on page 18.*

[73] Lin Gui et al. "Multi task mutual learning for joint sentiment classification and topic detection". In: *IEEE Transactions on Knowledge and Data Engineering* 34.4 (2020), pp. 1915–1927 *Cited on page 18.*

[74] Jichuan Zeng et al. "What you say and how you say it: Joint modeling of topics and discourse in microblog conversations". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 267–281 *Cited on page 18.*

[75] Yue Li et al. "Global surveillance of covid-19 by mining news media using a multi-source dynamic embedded topic model". In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2020, pp. 1–14 *Cited on page 18.*

[76] Baoxi Liu et al. "A reliable cross-site user generated content modeling method based on topic model". In: *Knowledge-Based Systems* 209 (2020), p. 106435 *Cited on page 18.*

[77] Hongyin Tang, Miao Li, and Beihong Jin. "A topic augmented text generation model: Joint learning of semantics and structural features". In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 2019, pp. 5090–5099 *Cited on page 18.*

[78] Yazheng Yang et al. "Topnet: Learning from neural topic model to generate long stories". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 1997–2005 *Cited on page 18.*

[79] Thong Nguyen et al. “Enriching and controlling global semantics for text summarization”. In: *arXiv preprint arXiv:2109.10616* (2021) *Cited on page 18.*

[80] Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. “Topic models for dynamic translation model adaptation”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2012, pp. 115–119 *Cited on page 18.*

[81] Yuxiang Zhang et al. “Htkg: Deep keyphrase generation with neural hierarchical topic guidance”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 1044–1054 *Cited on page 18.*

[82] Chong Wang and David M Blei. “Collaborative topic modeling for recommending scientific articles”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 448–456 *Cited on page 18.*

[83] Fréderic Godin et al. “Using topic models for Twitter hashtag recommendation”. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW ’13 Companion. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, pp. 593–596. ISBN: 9781450320382. DOI: 10.1145/2487788.2488002 *Cited on page 18.*

[84] Jieying She and Lei Chen. “TOMOHA: TOpic model-based HAshtag recommendation on twitter”. In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW ’14 Companion. Seoul, Korea: Association for Computing Machinery, 2014, pp. 371–372. ISBN: 9781450327459. DOI: 10.1145/2567948.2577292 *Cited on page 18.*

[85] Takeshi Kurashima et al. “Geo topic model: joint modeling of user’s activity area and interests for location recommendation”. In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. WSDM ’13. Rome, Italy: Association for Computing Machinery, 2013, pp. 375–384. ISBN: 9781450318693. DOI: 10.1145/2433396.2433444 *Cited on page 18.*

[86] Babak Esmaeili et al. “Structured neural topic models for reviews”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 3429–3439 *Cited on page 18.*

[87] Qianqian Xie et al. “Graph neural collaborative topic model for citation recommendation”. In: *ACM Transactions on Information Systems (TOIS)* 40.3 (2021), pp. 1–30 *Cited on page 18.*

[88] Xingyi Song et al. “Classification aware neural topic model for COVID-19 disinformation categorisation”. In: *PloS one* 16.2 (2021), e0247086 *Cited on page 18.*

References for Chapter 3: Topic Model Evaluation

[19] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 *Cited on pages 13, 21, 22, 44.*

[30] David Newman, Sarvnaz Karimi, and Lawrence Cavedon. “External evaluation of topic models”. In: *Proceedings of the 14th Australasian Document Computing Symposium*. University of Sydney. 2009, pp. 1–8 *Cited on pages 15, 23, 24.*

[41] Adji B Dieng, Francisco JR Ruiz, and David M Blei. “Topic modeling in embedding spaces”. In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453 *Cited on pages 15, 21, 23, 28, 31, 35, 46, 57.*

[42] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013) *Cited on pages 16, 25.*

[89] Jonathan Chang et al. "Reading tea leaves: How humans interpret topic models". In: *Advances in neural information processing systems* 22 (2009) *Cited on pages 21, 22, 30.*

[90] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999 *Cited on page 21.*

[91] Fred Jelinek et al. "Perplexity—a measure of the difficulty of speech recognition tasks". In: *The Journal of the Acoustical Society of America* 62.S1 (1977), S63–S63 *Cited on page 22.*

[92] Hanna M Wallach et al. "Evaluation methods for topic models". In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 1105–1112 *Cited on page 22.*

[93] He Zhao et al. "Neural topic model via optimal transport". In: *arXiv preprint arXiv:2008.13537* (2020) *Cited on page 22.*

[94] Alexander Hoyle et al. "Is automated topic model evaluation broken? the incoherence of coherence". In: *Advances in neural information processing systems* 34 (2021), pp. 2018–2033 *Cited on pages 22, 23, 25, 26, 28, 31, 35.*

[95] David Mimno et al. "Optimizing semantic coherence in topic models". In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, pp. 262–272 *Cited on pages 22–24.*

[96] Wray Buntine. "Estimating likelihoods for topic models". In: *Asian conference on machine learning*. Springer. 2009, pp. 51–64 *Cited on page 22.*

[97] Feng Nan et al. "Topic modeling with wasserstein autoencoders". In: *arXiv preprint arXiv:1907.12374* (2019) *Cited on page 22.*

[98] Sophie Burkhardt and Stefan Kramer. "Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model". In: *Journal of Machine Learning Research* 20.131 (2019), pp. 1–27 *Cited on page 23.*

[99] Michael Röder, Andreas Both, and Alexander Hinneburg. "Exploring the space of topic coherence measures". In: *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015, pp. 399–408 *Cited on pages 23, 24, 30, 57.*

[100] Nikolaos Aletras and Mark Stevenson. "Evaluating topic coherence using distributional semantics". In: *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*. 2013, pp. 13–22 *Cited on pages 23, 24, 30.*

[101] David Newman et al. "Evaluating Topic Models for Digital Libraries". In: *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. JCDL '10. Gold Coast, Queensland, Australia: Association for Computing Machinery, 2010, pp. 215–224. ISBN: 9781450300858. DOI: 10.1145/1816123.1816156 *Cited on page 24.*

[102] Gerlof Bouma. "Normalized (pointwise) mutual information in collocation extraction". In: *Proceedings of GSCL* 30 (2009), pp. 31–40 *Cited on pages 24, 57.*

[103] Jey Han Lau, David Newman, and Timothy Baldwin. "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 2014, pp. 530–539 *Cited on page 24.*

[104] Keith Stevens et al. "Exploring topic coherence over many models and many topics". In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. 2012, pp. 952–961 *Cited on page 24.*

[105] João Marcos Campagnolo, Denio Duarte, and Guillherme Dal Bianco. "Topic Coherence Metrics: How Sensitive Are They?" In: *Journal of Information and Data Management* 13.4 (2022) *Cited on page 25.*

[106] Sergey I. Nikolenko. "Topic Quality Metrics Based on Distributed Word Representations". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy: Association for Computing Machinery, 2016, pp. 1029–1032. ISBN: 9781450340694. DOI: 10.1145/2911451.2914720 *Cited on page 25.*

[107] Tobias Schnabel et al. "Evaluation methods for unsupervised word embeddings". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 298–307 *Cited on page 25.*

[108] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26 (2013) *Cited on page 25.*

[109] Nitin Ramrakhiyani et al. "Measuring topic coherence through optimal word buckets". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 2017, pp. 437–442 *Cited on page 25.*

[110] Damir Korenčić, Strahil Ristov, and Jan Šnajder. "Document-based topic coherence measures for news media text". In: *Expert systems with Applications* 114 (2018), pp. 357–373 *Cited on page 25.*

[111] Jeffrey Lund et al. "Automatic evaluation of local topic quality". In: *arXiv preprint arXiv:1905.13126* (2019) *Cited on page 25.*

[112] Caitlin Doogan and Wray Buntine. "Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 3824–3848. DOI: 10.18653/v1/2021.naacl-main.300 *Cited on pages 25, 28, 31, 33.*

[113] Alexander Hoyle et al. "Are Neural Topic Models Broken?" In: *arXiv preprint arXiv:2210.16162* (2022) *Cited on pages 25, 28.*

[114] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl_1 (2004), pp. 5228–5235 *Cited on pages 25, 28, 31, 74.*

[115] Andrew Kachites McCallum. "Mallet: A machine learning for languagetoolkit". In: *http://mallet.cs. umass. edu* (2002) *Cited on pages 25, 28, 35.*

References for Chapter 4: Contextualized Topic Coherence Metrics

[15] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018) *Cited on pages 3, 16, 17, 29, 47.*

[16] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022) *Cited on pages 11, 13, 28, 31, 46, 47, 49, 55, 57, 58, 88.*

[24] Dimo Angelov. "Top2vec: Distributed representations of topics". In: *arXiv preprint arXiv:2008.09470* (2020) *Cited on pages 13, 16, 17, 28, 31, 56, 58, 88.*

[29] Akash Srivastava and Charles Sutton. "Autoencoding variational inference for topic models". In: *arXiv preprint arXiv:1703.01488* (2017) *Cited on pages 15, 35.*

[41] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "Topic modeling in embedding spaces". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453 *Cited on pages 15, 21, 23, 28, 31, 35, 46, 57.*

[50] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017) *Cited on pages 16, 29.*

[51] Federico Bianchi, Silvia Terragni, and Dirk Hovy. "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 759–766. DOI: 10.18653/v1/2021.acl-short.96 *Cited on pages 16, 28, 31.*

[56] OpenAI. *ChatGPT: Engaging and Dynamic Conversations*. https://openai.com/blog/chatgpt. 2022 *Cited on pages 17, 28, 30.*

[65] Rui Wang, Deyu Zhou, and Yulan He. "Atm: Adversarial-neural topic model". In: *Information Processing & Management* 56.6 (2019), p. 102098 *Cited on pages 18, 28, 31.*

[89] Jonathan Chang et al. "Reading tea leaves: How humans interpret topic models". In: *Advances in neural information processing systems* 22 (2009) *Cited on pages 21, 22, 30.*

[94] Alexander Hoyle et al. "Is automated topic model evaluation broken? the incoherence of coherence". In: *Advances in neural information processing systems* 34 (2021), pp. 2018–2033 *Cited on pages 22, 23, 25, 26, 28, 31, 35.*

[99] Michael Röder, Andreas Both, and Alexander Hinneburg. "Exploring the space of topic coherence measures". In: *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015, pp. 399–408 *Cited on pages 23, 24, 30, 57.*

[100] Nikolaos Aletras and Mark Stevenson. "Evaluating topic coherence using distributional semantics". In: *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*. 2013, pp. 13–22 *Cited on pages 23, 24, 30.*

[112] Caitlin Doogan and Wray Buntine. "Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 3824–3848. DOI: 10.18653/v1/2021.naacl-main.300 *Cited on pages 25, 28, 31, 33.*

[113] Alexander Hoyle et al. "Are Neural Topic Models Broken?" In: *arXiv preprint arXiv:2210.16162* (2022) *Cited on pages 25, 28.*

[114] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl_1 (2004), pp. 5228–5235 *Cited on pages 25, 28, 31, 74.*

[115] Andrew Kachites McCallum. "Mallet: A machine learning for languagetoolkit". In: *http://mallet.cs. umass. edu* (2002) *Cited on pages 25, 28, 35.*

[116] Hamed Rahimi et al. "Contextualized Topic Coherence Metrics". In: *Findings of the Association for Computational Linguistics: EACL 2024*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024 *Cited on pages 27, 104.*

[117] Jacob Louis Hoover et al. "Linguistic dependencies and statistical dependence". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2941–2963. DOI: 10.18653/v1/2021.emnlp-main.234 *Cited on page 29.*

[118] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744 *Cited on page 30.*

[119] Shaheen Syed and Marco Spruit. "Full-text or abstract? examining topic coherence scores using latent dirichlet allocation". In: *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE. 2017, pp. 165–174 *Cited on page 30.*

[120] David Newman et al. "Automatic evaluation of topic coherence". In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. 2010, pp. 100–108 *Cited on pages 30, 57.*

[121] Ken Lang. "Newsweeder: Learning to filter netnews". In: *Proceedings of the Twelfth International Conference on Machine Learning*. 1995, pp. 331–339 *Cited on page 31.*

[122] Yasir Raza. *Elon Musk Tweets Dataset (17K): Dataset of Elon Musk tweets till now (17K)*. https://www.kaggle.com/datasets/yasirabdaali/elon-musk-tweets-dataset-17k. Version 1. 2023 *Cited on page 31.*

[123] Ekaba Bisong and Ekaba Bisong. "Google colaboratory". In: *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners* (2019), pp. 59–64 *Cited on page 32.*

[124] Philip Sedgwick. "Pearson's correlation coefficient". In: *Bmj* 345 (2012) *Cited on page 33.*

[125] Fosca Giannotti, Cristian Gozzi, and Giuseppe Manco. "Clustering transactional data". In: *Principles of Data Mining and Knowledge Discovery: 6th European Conference, PKDD 2002 Helsinki, Finland, August 19–23, 2002 Proceedings 6*. Springer. 2002, pp. 175–187 *Cited on page 34.*

[126] Leann Myers and Maria J Sirois. "Spearman correlation coefficients, differences between". In: *Encyclopedia of statistical sciences* 12 (2004) *Cited on page 35.*

[127] Hamed Rahimi et al. *ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics*. 2023. arXiv: 2302.01501 [cs.IR] *Cited on pages 41, 46, 74, 82, 101, 108.*

[128] Hamed Rahimi et al. "ANTM: Aligned Neural Topic Models for Exploring Evolving Topics". In: *Transactions on Large-Scale Data-and Knowledge-Centered Systems LVI: Special Issue on Data Management-Principles, Technologies, and Applications*. Springer, 2024, pp. 76–97 *Cited on page 41.*

References for Chapter 5: Dynamic Topic Models

[6] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120 *Cited on pages 1, 43, 44, 49, 52, 57.*

[11] Houkui Zhou, Huimin Yu, and Roland Hu. "Topic evolution based on the probabilistic topic model: a review". In: *Frontiers of Computer Science* 11 (2017), pp. 786–802 *Cited on pages 2, 44, 67.*

[15] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018) *Cited on pages 3, 16, 17, 29, 47.*

[16] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022) *Cited on pages 11, 13, 28, 31, 46, 47, 49, 55, 57, 58, 88.*

[19] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 *Cited on pages 13, 21, 22, 44.*

[41] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "Topic modeling in embedding spaces". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453 *Cited on pages 15, 21, 23, 28, 31, 35, 46, 57.*

[47] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018) *Cited on pages 16, 17, 47, 53.*

[55] MaartenGr. *cTFIDF - Class-based TF-IDF Implementation in Python*. https://github.com/MaartenGr/cTFIDF. Accessed on: 08 28, 2023. 2022 *Cited on pages 17, 47, 51, 55.*

[69] Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. "A survey on neural topic models: Methods, applications, and challenges". In: *Artificial Intelligence Review* 57.2 (2024), pp. 1–30 *Cited on pages 18, 44.*

[127] Hamed Rahimi et al. *ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics*. 2023. arXiv: 2302.01501 [cs.IR] *Cited on pages 41, 46, 74, 82, 101, 108.*

[129] Aly Abdelrazek et al. "Topic modeling algorithms and applications: A survey". In: *Information Systems* 112 (2023), p. 102131. ISSN: 0306-4379. DOI: https://doi.org/10.1016/j.is.2022.102131 *Cited on page 43.*

[130] Xuerui Wang and Andrew McCallum. "Topics over time: a non-markov continuous-time model of topical trends". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 424–433 *Cited on pages 43, 45, 49, 57, 68.*

[131] Jiajun Hu et al. "Modeling the evolution of development topics using dynamic topic models". In: *2015 IEEE 22nd international conference on software analysis, evolution, and reengineering (SANER)*. IEEE. 2015, pp. 3–12 *Cited on pages 43, 52.*

[132] Ke Li, Hubert Naacke, and Bernd Amann. "An Analytic Graph Data Model and Query Language for Exploring the Evolution of Science". In: *Big Data Research* 26 (2021), p. 100247 *Cited on page 43.*

[133] Hao Sha et al. "Dynamic topic modeling of the COVID-19 Twitter narrative among US governors and cabinet executives". In: *arXiv preprint arXiv:2004.11692* (2020) *Cited on page 43.*

[134] Tomoharu Iwata et al. "Online multiscale dynamic topic models". In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, pp. 663–672 *Cited on page 45.*

[135] Lu Ren, David B Dunson, and Lawrence Carin. "The dynamic hierarchical Dirichlet process". In: *Proceedings of the 25th international conference on machine learning*. 2008, pp. 824–831 *Cited on page 45.*

[136] Seyed Ali Bahrainian, Ida Mele, and Fabio Crestani. "Modeling discrete dynamic topics". In: *Proceedings of the Symposium on Applied Computing*. 2017, pp. 858–865 *Cited on page 45.*

[137] Chong Wang, David Blei, and David Heckerman. "Continuous time dynamic topic models". In: *arXiv preprint arXiv:1206.3298* (2012) *Cited on page 45.*

[138] Ike Vayansky and Sathish AP Kumar. "A review of topic modeling methods". In: *Information Systems* 94 (2020), p. 101582 *Cited on page 45.*

[139] Derek Greene and James P Cross. "Exploring the political agenda of the european parliament using a dynamic topic modeling approach". In: *Political Analysis* 25.1 (2017), pp. 77–94 *Cited on page 45.*

[140] Fang Yao and Yan Wang. "Tracking urban geo-topics based on dynamic topic model". In: *Computers, Environment and Urban Systems* 79 (2020), p. 101419 *Cited on page 45.*

[141] Arnab Bhadury et al. "Scaling up dynamic topic models". In: *Proceedings of the 25th International Conference on World Wide Web*. 2016, pp. 381–390 *Cited on page 45.*

[142] Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. "Discovering diverse and salient threads in document collections". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2012, pp. 710–720 *Cited on page 45.*

[143] Patrick Jähnichen et al. "Scalable Generalized Dynamic Topic Models". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by Amos Storkey and Fernando Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, Apr. 2018, pp. 1427–1435 *Cited on page 45.*

[144] Yezheng Liu et al. "Dynamic Topic Model for Tracking Topic Evolution and Measuring Popularity of Scientific Literature". In: *2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC)*. IEEE. 2021, pp. 315–320 *Cited on page 46.*

[145] Rob Churchill. "Percolation-based topic modeling for tweets". In: *KDD Conference (WISDOM'20). San Diego, CA, USA*. 2020 *Cited on page 46.*

[146] Xing Wei, Jimeng Sun, and Xuerui Wang. "Dynamic Mixture Models for Multiple Time-Series." In: *Ijcai*. Vol. 7. 2007, pp. 2909–2914 *Cited on page 46.*

[147] Tomoharu Iwata et al. "Online Multiscale Dynamic Topic Models". In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10. Washington, DC, USA: Association for Computing Machinery, 2010, pp. 663–672. ISBN: 9781450300551. DOI: 10.1145/1835804.1835889 *Cited on page 46.*

[148] Arnab Bhadury et al. "Scaling up Dynamic Topic Models". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 381–390. ISBN: 9781450341431. DOI: 10.1145/2872427.2883046 *Cited on page 46.*

[149] Elaine Zosa and Mark Granroth-Wilding. "Multilingual dynamic topic model". In: *RANLP 2019-Natural Language Processing a Deep Learning World* (2019) *Cited on page 46.*

[150] Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents". In: *International conference on machine learning*. PMLR. 2014, pp. 1188–1196 *Cited on page 46.*

[151] Venkatesan T Chakaravarthy et al. "Efficient scaling of dynamic graph neural networks". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021, pp. 1–15 *Cited on page 46.*

[152] Qiang Gao et al. "Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec". In: *Scientometrics* 127.3 (2022), pp. 1543–1563 *Cited on page 46.*

[153] Anton Eklund, Mona Forsman, and Frank Drewes. "Dynamic Topic Modeling by Clustering Embeddings from Pretrained Language Models: A Research Proposal". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*. 2022, pp. 84–91 *Cited on page 46.*

[154] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "The dynamic embedded topic model". In: *arXiv preprint arXiv:1907.05545* (2019) *Cited on pages 46, 49, 57.*

[155] Federico Tomasi, Mounia Lalmas, and Zhenwen Dai. "Efficient inference for dynamic topic modeling with large vocabularies". In: *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*. Ed. by James Cussens and Kun Zhang. Vol. 180. Proceedings of Machine Learning Research. PMLR, Aug. 2022, pp. 1950–1959 *Cited on page 46.*

[156] Kambiz Ghoorchian and Magnus Sahlgren. "GDTM: graph-based dynamic topic models". In: *Progress in Artificial Intelligence* 9.3 (2020), pp. 195–207 *Cited on page 46.*

[157] Jaakko Eskonen. "Dynamic Topic Modeling and Clustering: Dynamic Topic Modeling and Clustering of Occupational Health and Safety Publications". MA thesis. Tampere University, 2022 *Cited on pages 47, 50.*

[158] Ibai Guillén-Pacho, Carlos Badenes-Olmedo, and Oscar Corcho. "Dynamic Topic Modelling for Exploring the Scientific Literature on Coronavirus: An Unsupervised Labelling Technique". In: *ResearchSquare (preprint)* (2023) *Cited on page 47.*

[159] Roman Egger and Joanne Yu. "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts". In: *Frontiers in sociology* 7 (2022), p. 886498 *Cited on pages 47, 50.*

References for Chapter 6: Aligned Neural Topic Models

[6] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120 *Cited on pages 1, 43, 44, 49, 52, 57.*

[16] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022) *Cited on pages 11, 13, 28, 31, 46, 47, 49, 55, 57, 58, 88.*

[24] Dimo Angelov. "Top2vec: Distributed representations of topics". In: *arXiv preprint arXiv:2008.09470* (2020) *Cited on pages 13, 16, 17, 28, 31, 56, 58, 88.*

[41] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "Topic modeling in embedding spaces". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453 *Cited on pages 15, 21, 23, 28, 31, 35, 46, 57.*

[47] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018) *Cited on pages 16, 17, 47, 53.*

[52] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019) *Cited on pages 16, 58.*

[55] MaartenGr. *cTFIDF - Class-based TF-IDF Implementation in Python*. `https://github.com/MaartenGr/cTFIDF`. Accessed on: 08 28, 2023. 2022 *Cited on pages 17, 47, 51, 55.*

[99] Michael Röder, Andreas Both, and Alexander Hinneburg. "Exploring the space of topic coherence measures". In: *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015, pp. 399–408 *Cited on pages 23, 24, 30, 57.*

[102] Gerlof Bouma. "Normalized (pointwise) mutual information in collocation extraction". In: *Proceedings of GSCL* 30 (2009), pp. 31–40 *Cited on pages 24, 57.*

[120] David Newman et al. "Automatic evaluation of topic coherence". In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. 2010, pp. 100–108 *Cited on pages 30, 57.*

[130] Xuerui Wang and Andrew McCallum. "Topics over time: a non-markov continuous-time model of topical trends". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 424–433 *Cited on pages 43, 45, 49, 57, 68.*

[131] Jiajun Hu et al. "Modeling the evolution of development topics using dynamic topic models". In: *2015 IEEE 22nd international conference on software analysis, evolution, and reengineering (SANER)*. IEEE. 2015, pp. 3–12 *Cited on pages 43, 52.*

[154] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "The dynamic embedded topic model". In: *arXiv preprint arXiv:1907.05545* (2019) *Cited on pages 46, 49, 57.*

[157] Jaakko Eskonen. "Dynamic Topic Modeling and Clustering: Dynamic Topic Modeling and Clustering of Occupational Health and Safety Publications". MA thesis. Tampere University, 2022 *Cited on pages 47, 50.*

[159] Roman Egger and Joanne Yu. "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts". In: *Frontiers in sociology* 7 (2022), p. 886498 *Cited on pages 47, 50.*

[160] *How to use AlignedUMAP* *Cited on pages 51, 58.*

[161] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. "Density-based clustering based on hierarchical density estimates". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2013, pp. 160–172 *Cited on pages 51, 54.*

[162] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012 *Cited on page 52.*

[163] Attri Ghosal et al. "A short review on different clustering techniques and their applications". In: *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* (2020), pp. 69–83 *Cited on page 53.*

[164] Mohammad Tariqul Islam and Jason W Fleischer. "Manifold-aligned Neighbor Embedding". In: *arXiv preprint arXiv:2205.11257* (2022) *Cited on page 53.*

[165] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *KDD*. Vol. 96. 1996, pp. 226–231 *Cited on page 54.*

[166] Angur Mahmud Jarman. "Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method". In: *Georgia Southern University* (2020) *Cited on page 54.*

[167] Olivier Gracianne, Anaïs Halftermeyer, and Thi-Bich-Hanh Dao. "Presenting an event through the description of related tweets clusters". In: *34th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2022, Macao, China, October 31 - November 2, 2022*. Ed. by Marek Z. Reformat, Du Zhang, and Nikolaos G. Bourbakis. IEEE, 2022, pp. 1283–1290. DOI: 10.1109/ICTAI56018.2022.00194 *Cited on page 55.*

[168] Michael Ley. "The DBLP computer science bibliography: Evolution, research issues, perspectives". In: *String Processing and Information Retrieval: 9th International Symposium, SPIRE 2002 Lisbon, Portugal, September 11–13, 2002 Proceedings 9*. Springer. 2002, pp. 1–10 *Cited on page 57.*

[169] Colin B Clement et al. "On the Use of ArXiv as a Dataset". In: *arXiv preprint arXiv:1905.00075* (2019) *Cited on page 57.*

[170] Yuval Pinter, Cassandra L Jacobs, and Max Bittker. "NYTWIT: A dataset of novel words in the New York Times". In: *arXiv preprint arXiv:2003.03444* (2020) *Cited on page 57.*

[171] Zhihua Yan and Xijin Tang. "Exploring evolution of public opinions on Tianya club using dynamic topic models". In: *Journal of Systems Science and Information* 8.4 (2020), pp. 309–324 *Cited on page 57.*

[172] Hamed Rahimi et al. *Contextualized Topic Coherence Metrics*. 2023. arXiv: 2305.14587 [cs.CL] *Cited on page 57.*

[173] Takako Hashimoto et al. "Analyzing temporal patterns of topic diversity using graph clustering". In: *The Journal of Supercomputing* 77 (2021), pp. 4375–4388 *Cited on page 57.*

[174] Hugging Face. *Hugging Face*. https://huggingface.co. Accessed: 2023-02-01. 2021 *Cited on page 58.*

[175] Ashton Anderson, Dan Jurafsky, and Dan McFarland. "Towards a computational history of the acl: 1980-2008". In: *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. 2012, pp. 13–21 *Cited on page 58.*

References for Chapter 7: Topic Evolution Models and Analysis

[11] Houkui Zhou, Huimin Yu, and Roland Hu. "Topic evolution based on the probabilistic topic model: a review". In: *Frontiers of Computer Science* 11 (2017), pp. 786–802 *Cited on pages 2, 44, 67.*

[12] Jianhua Hou, Xiucai Yang, and Chaomei Chen. "Emerging Trends and New Developments in Information Science: A Document Co-Citation Analysis (2009–2016)". In: *Scientometrics* 115.2 (2018), pp. 869–892 *Cited on pages 2, 70, 77.*

[130] Xuerui Wang and Andrew McCallum. "Topics over time: a non-markov continuous-time model of topical trends". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 424–433 *Cited on pages 43, 45, 49, 57, 68.*

[176] Scott Jensen et al. "Generation of topic evolution trees from heterogeneous bibliographic networks". In: *Journal of informetrics* 10.2 (2016), pp. 606–621 *Cited on page 67.*

[177] Muhammad Abulaish and Mohd Fazil. "Modeling topic evolution in twitter: An embedding-based approach". In: *IEEE Access* 6 (2018), pp. 64847–64857 *Cited on page 67.*

[178] Michael Mathioudakis and Nick Koudas. "Twittermonitor: trend detection over the twitter stream". In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 2010, pp. 1155–1158 *Cited on pages 67, 68.*

[179] Martin Franz et al. "Unsupervised and supervised clustering for topic tracking". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 2001, pp. 310–317 *Cited on pages 68, 69.*

[180] April Kontostathis et al. "A survey of emerging trend detection in textual data mining". In: *Survey of text mining: Clustering, classification, and retrieval*. Springer, 2004, pp. 185–224 *Cited on page 68.*

[181] Yi-Ning Tu and Jia-Lang Seng. "Indices of novelty for emerging topic detection". In: *Information processing & management* 48.2 (2012), pp. 303–325 *Cited on page 68.*

[182] John W Lounsbury et al. "An analysis of topic areas and topic trends in the Community Mental Health Journal from 1965 through 1977". In: *Community mental health journal* 15 (1979), pp. 267–276 *Cited on page 68.*

[183] Katy Börner, Chaomei Chen, and Kevin W Boyack. "Visualizing knowledge domains". In: *Annual review of information science and technology* 37.1 (2003), pp. 179–255 *Cited on page 68.*

[184] Qi He et al. "Detecting topic evolution in scientific literature: how can citations help?" In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, pp. 957–966 *Cited on page 68.*

[185] Ding Zhou et al. "Topic evolution and social interactions: how authors effect research". In: *Proceedings of the 15th ACM international conference on Information and knowledge management*. 2006, pp. 248–257 *Cited on page 68.*

[186] Hang Jiang et al. "Topic Detection and Tracking with Time-Aware Document Embeddings". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 16293–16303 *Cited on page 69.*

[187] Thomas Griffiths et al. "Integrating topics and syntax". In: *Advances in neural information processing systems* 17 (2004) *Cited on page 69.*

[188] Winfred P Lehmann. *Historical linguistics: An introduction*. Routledge, 2013 *Cited on page 69.*

[189] Derry Tanti Wijaya and Reyyan Yeniterzi. "Understanding semantic change of words over centuries". In: *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*. 2011, pp. 35–40 *Cited on page 69.*

[190] Baitong Chen et al. "Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval". In: *Journal of Informetrics* 11.4 (2017), pp. 1175–1189 *Cited on pages 69, 76, 77.*

[191] Kristina Gulordava and Marco Baroni. "A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus." In: *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*. 2011, pp. 67–71 *Cited on page 69.*

[192] William L Hamilton, Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change". In: *arXiv preprint arXiv:1605.09096* (2016) *Cited on page 69.*

[193] Yoon Kim et al. "Temporal analysis of language through neural language models". In: *arXiv preprint arXiv:1405.3515* (2014) *Cited on page 69.*

[194] Paul Iles, Anita Ramgutty-Wong, and Maurice Yolles. "HRM and knowledge migration across cultures: Issues, limitations, and Mauritian specificities". In: *Employee Relations* 26.6 (2004), pp. 643–662 *Cited on page 69.*

[195] Yookyung Jo, John E. Hopcroft, and Carl Lagoze. "The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus". In: *Proceedings of the 20th International Conference on World Wide Web*. 2011, pp. 257–266 *Cited on page 70.*

[196] David Chavalarias and Jean-Philippe Philippe Cointet. "Phylomemetic Patterns in Science Evolution—the Rise and Fall of Scientific Fields". In: *PloS one* 8.2 (2013), e54847. ISSN: 19326203. DOI: 10.1371/journal.pone.0054847 *Cited on pages 70, 76.*

[197] Ke Li, Hubert Naacke, and Bernd Amann. "An Analytic Graph Data Model and Query Language for Exploring the Evolution of Science". In: *Big Data Research* 26 (2021), p. 100247 *Cited on page 70.*

[198] Victor Andrei and Ognjen Arandjelović. "Complex Temporal Topic Evolution Modelling Using the Kullback-Leibler Divergence and the Bhattacharyya Distance". In: *EURASIP Journal on Bioinformatics and Systems Biology* 2016.1 (Dec. 2016). ISSN: 1687-4153. DOI: 10.1186/s13637-016-0050-0 *Cited on page 70.*

[199] Adham Beykikhoshk et al. "Discovering Topic Structures of a Temporally Evolving Document Corpus". In: *Knowl. Inf. Syst.* 55.3 (June 2018), pp. 599–632. ISSN: 0219-1377. DOI: 10.1007/s10115-017-1095-4 *Cited on pages 70, 76.*

[200] Sukhwan Jung and Aviv Segev. "Identifying a Common Pattern within Ancestors of Emerging Topics for Pan-Domain Topic Emergence Prediction". In: *Knowledge-Based Systems* 258 (Dec. 2022), p. 110020. ISSN: 09507051. DOI: 10.1016/j.knosys.2022.110020 *Cited on pages 70, 76.*

[201] Sukhwan Jung and Aviv Segev. "DAC: Descendant-aware Clustering Algorithm for Network-Based Topic Emergence Prediction". In: *Journal of Informetrics* 16.3 (Aug. 2022), p. 101320. ISSN: 17511577. DOI: 10.1016/j.joi.2022.101320 *Cited on pages 70, 76.*

[202] Christine Balili et al. "TermBall: Tracking and Predicting Evolution Types of Research Topics by Using Knowledge Structures in Scholarly Big Data". In: *IEEE Access* 8 (2020), pp. 108514–108529 *Cited on page 70.*

[203] Chaoguang Huo, Shutian Ma, and Xiaozhong Liu. "Hotness Prediction of Scientific Topics Based on a Bibliographic Knowledge Graph". In: *Information Processing & Management* 59.4 (2022), p. 102980 *Cited on page 70.*

[204] Angelo A. Salatino, Francesco Osborne, and Enrico Motta. "AUGUR: Forecasting the Emergence of New Research Topics". In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL '18. New York, NY, USA: ACM, 2018, pp. 303–312. ISBN: 978-1-4503-5178-2. DOI: 10.1145/3197026.3197052 *Cited on page 70.*

[205] Angelo A Salatino, Francesco Osborne, and Enrico Motta. "How Are Topics Born? Understanding the Research Dynamics Preceding the Emergence of New Areas". In: *PeerJ Computer Science* 3 (2017), e119 *Cited on page 70.*

[206] Kevin W. Boyack et al. "Characterizing the Emergence of Two Nanotechnology Topics Using a Contemporaneous Global Micro-Model of Science". In: *Journal of Engineering and Technology Management*. Special Issue on Emergence of Technologies: Methods and Tools for Management 32 (Apr. 2014), pp. 147–159. ISSN: 0923-4748. DOI: 10.1016/j.jengtecman.2013.07.001 *Cited on page 70.*

[207] Shilpy Sharma, David A Swayne, and Charlie Obimbo. "Trend Analysis and Change Point Techniques: A Survey". In: *Energy, Ecology and Environment* 1 (2016), pp. 123–130 *Cited on page 70.*

[208] Henry Small, Kevin W Boyack, and Richard Klavans. "Identifying Emerging Topics in Science and Technology". In: *Research policy* 43.8 (2014), pp. 1450–1467 *Cited on page 70.*

[209] Yoonjung An, Mintak Han, and Yongtae Park. "Identifying Dynamic Knowledge Flow Patterns of Business Method Patents with a Hidden Markov Model". In: *Scientometrics* 113.2 (2017), pp. 783–802 *Cited on pages 70, 75, 76.*

[210] Dennys Eduardo Rossetto et al. "Structure and Evolution of Innovation Research in the Last 60 Years: Review and Future Trends in the Field of Business through the Citations and Co-Citations Analysis". In: *Scientometrics* 115.3 (2018), pp. 1329–1363 *Cited on pages 70, 76.*

[211] Chao Zhang and Jiancheng Guan. "How to Identify Metaknowledge Trends and Features in a Certain Research Field? Evidences from Innovation and Entrepreneurial Ecosystem". In: *Scientometrics* 113.2 (2017), pp. 1177–1197 *Cited on pages 70, 77.*

[212] Alba Santa Soriano, Carolina Lorenzo Álvarez, and Rosa María Torres Valdés. "Bibliometric Analysis to Identify an Emerging Research Area: Public Relations Intelligence—a Challenge to Strengthen Technological Observatories in the Network Society". In: *Scientometrics* 115.3 (2018), pp. 1591–1614 *Cited on pages 70, 77.*

References for Chapter 8: Science Evolution Analysis

[1] Thomas S Kuhn. *The Structure of Scientific Revolutions*. University of Chicago press, 1962 *Cited on pages 1, 73, 77.*

[7] Karl Raimund Popper et al. *Objective Knowledge: An Evolutionary Approach*. Vol. 49. Clarendon press Oxford, 1979 *Cited on pages 1, 73.*

[10] Huailan Liu et al. "Mapping the Technology Evolution Path: A Novel Model for Dynamic Topic Detection and Tracking". In: *Scientometrics* 125.3 (2020), pp. 2043–2090 *Cited on pages 2, 75.*

[12] Jianhua Hou, Xiucai Yang, and Chaomei Chen. "Emerging Trends and New Developments in Information Science: A Document Co-Citation Analysis (2009–2016)". In: *Scientometrics* 115.2 (2018), pp. 869–892 *Cited on pages 2, 70, 77.*

[13] Thara Prabhakaran et al. "Towards Prediction of Paradigm Shifts from Scientific Literature". In: *Scientometrics* 117.3 (2018), pp. 1611–1644 *Cited on pages 2, 76.*

[114] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl_1 (2004), pp. 5228–5235 *Cited on pages 25, 28, 31, 74.*

[127] Hamed Rahimi et al. *ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics*. 2023. arXiv: 2302.01501 [cs.IR] *Cited on pages 41, 46, 74, 82, 101, 108.*

[190] Baitong Chen et al. "Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval". In: *Journal of Informetrics* 11.4 (2017), pp. 1175–1189 *Cited on pages 69, 76, 77.*

[196] David Chavalarias and Jean-Philippe Philippe Cointet. "Phylomemetic Patterns in Science Evolution—the Rise and Fall of Scientific Fields". In: *PloS one* 8.2 (2013), e54847. ISSN: 19326203. DOI: 10.1371/journal.pone.0054847 *Cited on pages 70, 76.*

[199] Adham Beykikhoshk et al. "Discovering Topic Structures of a Temporally Evolving Document Corpus". In: *Knowl. Inf. Syst.* 55.3 (June 2018), pp. 599–632. ISSN: 0219-1377. DOI: 10.1007/s10115-017-1095-4 *Cited on pages 70, 76.*

[200] Sukhwan Jung and Aviv Segev. "Identifying a Common Pattern within Ancestors of Emerging Topics for Pan-Domain Topic Emergence Prediction". In: *Knowledge-Based Systems* 258 (Dec. 2022), p. 110020. ISSN: 09507051. DOI: 10.1016/j.knosys.2022.110020 *Cited on pages 70, 76.*

[201] Sukhwan Jung and Aviv Segev. "DAC: Descendant-aware Clustering Algorithm for Network-Based Topic Emergence Prediction". In: *Journal of Informetrics* 16.3 (Aug. 2022), p. 101320. ISSN: 17511577. DOI: 10.1016/j.joi.2022.101320 *Cited on pages 70, 76.*

[209] Yoonjung An, Mintak Han, and Yongtae Park. "Identifying Dynamic Knowledge Flow Patterns of Business Method Patents with a Hidden Markov Model". In: *Scientometrics* 113.2 (2017), pp. 783–802 *Cited on pages 70, 75, 76.*

[210] Dennys Eduardo Rossetto et al. "Structure and Evolution of Innovation Research in the Last 60 Years: Review and Future Trends in the Field of Business through the Citations and Co-Citations Analysis". In: *Scientometrics* 115.3 (2018), pp. 1329–1363 *Cited on pages 70, 76.*

[211] Chao Zhang and Jiancheng Guan. "How to Identify Metaknowledge Trends and Features in a Certain Research Field? Evidences from Innovation and Entrepreneurial Ecosystem". In: *Scientometrics* 113.2 (2017), pp. 1177–1197 *Cited on pages 70, 77.*

[212] Alba Santa Soriano, Carolina Lorenzo Álvarez, and Rosa María Torres Valdés. "Bibliometric Analysis to Identify an Emerging Research Area: Public Relations Intelligence—a Challenge to Strengthen Technological Observatories in the Network Society". In: *Scientometrics* 115.3 (2018), pp. 1591–1614 *Cited on pages 70, 77.*

[213] Mário Cordeiro et al. "Evolving Networks and Social Network Analysis Methods and Techniques". In: *Social media and journalism-trends, connections, implications* (2018), pp. 101–134 *Cited on page 74.*

[214] David Hall, Dan Jurafsky, and Christopher D Manning. "Studying the history of ideas using topic models". In: *Proceedings of the 2008 conference on empirical methods in natural language processing*. 2008, pp. 363–371 *Cited on page 74.*

[215] Bent Fuglede and Flemming Topsoe. "Jensen-Shannon divergence and Hilbert space embedding". In: *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.* IEEE. 2004, p. 31 *Cited on page 74.*

[216] Beibei Hu et al. "A lead-lag analysis of the topic evolution patterns for preprints and publications". In: *Journal of the Association for Information Science and Technology* 66.12 (2015), pp. 2643–2656 *Cited on page 74.*

[217] Uriel Cohen Priva and Joseph L Austerweil. "Analyzing the history of cognition using topic models". In: *Cognition* 135 (2015), pp. 4–9 *Cited on page 74.*

[218] Hugh Chipman et al. "The practical implementation of Bayesian model selection". In: *Lecture Notes-Monograph Series* (2001), pp. 65–134 *Cited on page 74.*

[219] Amna Dridi et al. "Leap2trend: A Temporal Word Embedding Approach for Instant Detection of Emerging Scientific Trends". In: *IEEE Access* 7 (2019), pp. 176414–176428 *Cited on page 75.*

[220] Satoshi Morinaga and Kenji Yamanishi. "Tracking Dynamics of Topic Trends Using a Finite Mixture Model". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2004, pp. 811–816 *Cited on page 75.*

[221] Chunlei Ye et al. "Mapping the topic evolution using citation-topic model and social network analysis". In: *2015 12th international conference on fuzzy systems and knowledge discovery (FSKD)*. IEEE. 2015, pp. 2648–2653 *Cited on page 75.*

[222] Qi He et al. "Detecting topic evolution in scientific literature: how can citations help?" In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, pp. 957–966 *Cited on page 75.*

[223] Chaomei Chen. "CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature". In: *Journal of the American Society for Information Science and Technology* 57.3 (2006), pp. 359–377. ISSN: 1532-2890. DOI: 10.1002/asi.20317 *Cited on page 75.*

[224] Sukhwan Jung and Wan Chul Yoon. "An Alternative Topic Model Based on Common Interest Authors for Topic Evolution Analysis". In: *Journal of Informetrics* 14.3 (Aug. 2020), p. 101040. ISSN: 17511577. DOI: 10.1016/j.joi.2020.101040 *Cited on page 75.*

[225] Zhiya Zuo and Kang Zhao. "A graphical model for topical impact over time". In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. 2018, pp. 405–406 *Cited on page 75.*

[226] Mario Coccia. "General properties of the evolution of research fields: a scientometric study of human microbiome, evolutionary robotics and astrobiology". In: *Scientometrics* 117.2 (2018), pp. 1265–1283 *Cited on page 75.*

[227] Mario Coccia. "The evolution of scientific disciplines in applied sciences: dynamics and empirical properties of experimental physics". In: *Scientometrics* 124.1 (2020), pp. 451–487 *Cited on page 75.*

[228] Mario Coccia, Saeed Roshani, and Melika Mosleh. "Scientific developments and new technological trajectories in sensor research". In: *Sensors* 21.23 (2021), p. 7803 *Cited on page 75.*

[229] Meng Zhao et al. "Lsif: A system for large-scale information flow detection based on topic-related semantic similarity measurement". In: *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. Vol. 1. IEEE. 2015, pp. 417–424 *Cited on page 76.*

[230] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. "Newsjunkie: providing personalized newsfeeds via analysis of information novelty". In: *Proceedings of the 13th international conference on World Wide Web*. 2004, pp. 482–490 *Cited on page 76.*

[231] Jiajia Huang et al. "A probabilistic method for emerging topic tracking in microblog stream". In: *World Wide Web* 20 (2017), pp. 325–350 *Cited on page 76.*

[232] Dejian Yu and Bo Xiang. "Discovering knowledge map and evolutionary path of HRM and ER: using the STM combined with Word2vec". In: *International Journal of Manpower* (2023) *Cited on page 76.*

[233] Angelo Salatino. *Early detection of research trends*. Open University (United Kingdom), 2019 *Cited on page 76.*

[234] Ding Zhou et al. "Topic Evolution and Social Interactions: How Authors Effect Research". In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. CIKM '06. New York, NY, USA: Association for Computing Machinery, Nov. 2006, pp. 248–257. ISBN: 978-1-59593-433-8. DOI: 10.1145/1183614.1183653 *Cited on page 76.*

[235] P Chen and Sidney Redner. "Community Structure of the Physical Review Citation Network". In: *Journal of Informetrics* 4.3 (2010), pp. 278–290 *Cited on page 76.*

[236] Hiran H Lathabai, Thara Prabhakaran, and Manoj Changat. "Centrality and flow vergence gradient based path analysis of scientific literature: A case study of biotechnology for engineering". In: *Physica A: Statistical Mechanics and its Applications* 429 (2015), pp. 157–168 *Cited on page 76.*

[237] Paul Jaccard. "The distribution of the flora in the alpine zone. 1". In: *New phytologist* 11.2 (1912), pp. 37–50 *Cited on page 76.*

[238] Lu Huang et al. "Tracking the Dynamics of Co-Word Networks for Emerging Topic Identification". In: *Technological Forecasting and Social Change* 170 (Sept. 2021), p. 120944. ISSN: 00401625. DOI: 10.1016/j.techfore.2021.120944 *Cited on page 76.*

[239] Amit Singhal et al. "Modern information retrieval: A brief overview". In: *IEEE Data Eng. Bull.* 24.4 (2001), pp. 35–43 *Cited on page 76.*

[240] Ernst Hellinger. "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen." In: *Journal für die reine und angewandte Mathematik* 1909.136 (1909), pp. 210–271 *Cited on page 76.*

[241] Anil Bhattacharyya. "On a measure of divergence between two statistical populations defined by their probability distribution". In: *Bulletin of the Calcutta Mathematical Society* 35 (1943), pp. 99–110 *Cited on page 76.*

[242] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020 *Cited on page 76.*

[243] Anthony Breitzman and Patrick Thomas. "The Emerging Clusters Model: A Tool for Identifying Emerging Technologies across Multiple Patent Systems". In: *Research Policy* 44.1 (Feb. 2015), pp. 195–205. ISSN: 00487333. DOI: 10.1016/j.respol.2014.06.006 *Cited on page 77.*

[244] Angelo A Salatino, Francesco Osborne, and Enrico Motta. "How Are Topics Born? Understanding the Research Dynamics Preceding the Emergence of New Areas". In: *PeerJ Computer Science* 3 (2017), e119 *Cited on page 77.*

[245] Haoli Bai et al. "Neural relational topic models for scientific article analysis". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018, pp. 27–36 *Cited on page 77.*

[246] Dudley Shapere. "The structure of scientific revolutions". In: *The Philosophical Review* 73.3 (1964), pp. 383–394 *Cited on page 77.*

[247] Thwink.org. "Kuhn Cycle". In: (2023) *Cited on page 78.*

[248] Serhat Burmaoglu and Ozcan Saritas. "An evolutionary analysis of the innovation policy domain: is there a paradigm shift?" In: *Scientometrics* 118.3 (2019), pp. 823–847 *Cited on page 78.*

[249] Basak Denizci Guillet. "An evolutionary analysis of revenue management research in hospitality and tourism: is there a paradigm shift?" In: *International Journal of Contemporary Hospitality Management* 32.2 (2020), pp. 560–587 *Cited on page 78.*

References for Chapter 9: Automatic Topic Emergence Monitoring

[16] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022) *Cited on pages 11, 13, 28, 31, 46, 47, 49, 55, 57, 58, 88.*

[24] Dimo Angelov. "Top2vec: Distributed representations of topics". In: *arXiv preprint arXiv:2008.09470* (2020) *Cited on pages 13, 16, 17, 28, 31, 56, 58, 88.*

[46] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in neural information processing systems* 26 (2013) *Cited on pages 16, 88.*

[48] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17*. Springer. 2013, pp. 160–172 *Cited on pages 16, 17, 88.*

[54] Maarten Grootendorst. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure". In: *arXiv preprint arXiv:2203.05794* (2022). arXiv: 2203.05794 *Cited on pages 17, 83, 88.*

[127] Hamed Rahimi et al. *ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics*. 2023. arXiv: 2302.01501 [cs.IR] *Cited on pages 41, 46, 74, 82, 101, 108.*

[250] Hamed Rahimi et al. "ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives". In: *International Conference on Complex Networks and Their Applications*. Springer. 2023, pp. 332–343 *Cited on pages 81, 101, 108.*

[251] Hamed Rahimi et al. *ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives*. 2023. arXiv: 2306.02221 [cs.IR] *Cited on page 81.*

[252] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. "A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications". In: *IEEE Transactions on Knowledge and Data Engineering* 30.9 (2018), pp. 1616–1637 *Cited on page 85.*

[253] Palash Goyal and Emilio Ferrara. "Graph Embedding Techniques, Applications, and Performance: A Survey". In: *Knowledge-Based Systems* 151 (2018), pp. 78–94 *Cited on page 85.*

[254] Sedigheh Mahdavi, Shima Khoshraftar, and Aijun An. "Dynnode2vec: Scalable Dynamic Network Embedding". In: arXiv:1812.02356. arXiv, Feb. 2019. DOI: 10.48550/arXiv.1812.02356. arXiv: 1812.02356 [cs, stat] *Cited on page 86.*

[255] Michael Ley. "The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives". In: *String Processing and Information Retrieval: 9th International Symposium, SPIRE 2002 Lisbon, Portugal, September 11–13, 2002 Proceedings 9*. Springer. 2002, pp. 1–10 *Cited on page 87.*

[256] Leland McInnes, John Healy, and James Melville. "Umap: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426* (2018). arXiv: 1802.03426 *Cited on page 88.*

[257] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. "From Louvain to Leiden: Guaranteeing Well-Connected Communities". In: *Scientific reports* 9.1 (2019), pp. 1–12 *Cited on page 88.*

[258] Ferenc Béres et al. "Node Embeddings in Dynamic Graphs". In: *Applied Network Science* 4.1 (Dec. 2019), p. 64. ISSN: 2364-8228. DOI: 10.1007/s41109-019-0169-5 *Cited on page 88.*

References for Chapter 10: Topic Evolution Analysis with Paradigm Shift

[116] Hamed Rahimi et al. "Contextualized Topic Coherence Metrics". In: *Findings of the Association for Computational Linguistics: EACL 2024*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024 *Cited on pages 27, 104.*

[127] Hamed Rahimi et al. *ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics*. 2023. arXiv: 2302.01501 [cs.IR] *Cited on pages 41, 46, 74, 82, 101, 108.*

[250] Hamed Rahimi et al. "ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives". In: *International Conference on Complex Networks and Their Applications*. Springer. 2023, pp. 332–343 *Cited on pages 81, 101, 108.*

[259] Richard S Sutton and Andrew G Barto. "Reinforcement learning: An introduction". In: *Robotica* 17.2 (1999), pp. 229–235 *Cited on page 97.*

References for Chapter 11: Outlooks and Conclusion

[127] Hamed Rahimi et al. *ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics*. 2023. arXiv: 2302.01501 [cs.IR] *Cited on pages 41, 46, 74, 82, 101, 108.*

[250] Hamed Rahimi et al. "ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives". In: *International Conference on Complex Networks and Their Applications*. Springer. 2023, pp. 332–343 *Cited on pages 81, 101, 108.*

[260] Emmanuel de Bézenac et al. "Normalizing kalman filters for multivariate time series analysis". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2995–3007 *Cited on page 109.*

[261] Rahul G. Krishnan, Uri Shalit, and David Sontag. *Deep Kalman Filters*. 2015. arXiv: 1511.05121 [stat.ML] *Cited on page 109.*

[262] DivergentAI. *DreamGPT*. https://github.com/DivergentAI/dreamGPT. Accessed on June 07, 2023. 2023 *Cited on page 109.*