As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology
Volume 9, Issue 10, pages 28-40.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol9issue10.html
DOI: https://doi.org/10.46501/IJMTST0910004

Unlocking Insights: A Comprehensive Literature Review on Topic Modeling and Text Analysis

C.B.Pavithra[1] | Dr.J.Savitha[2]

[1]Research Scholar, Department of Information Technology, Dr.N.G.P. Arts & Science College, Coimbatore, Tamilnadu, India.
[2]Professor, Department of Information Technology, Dr.N.G.P. Arts & Science College, Coimbatore, Tamilnadu, India.

To Cite this Article
C.B.Pavithra and Dr.J.Savitha. Unlocking Insights: A Comprehensive Literature Review on Topic Modeling and Text Analysis, International Journal for Modern Trends in Science and Technology, 2023, 9(10), pages. 28-40.https://doi.org/10.46501/IJMTST0910004

Article Info
Received: 12 October 2023; Accepted: 23 October 2023; Published: 24 October 2023.

ABSTRACT

Topic modeling is a prominent technique in the field of natural language processing and text analysis that has gained significant attention in recent years. This literature review provides an in-depth exploration of topic modeling, its historical evolution, basic concepts and applications across various domains, challenges, recent advances and future directions. The review begins with an introduction that outlines the importance of topic modeling and sets the scope for the discussion. It then delves into a historical overview of the field, tracing its origins, evolution and the contributions of influential researchers. Following this, the review explains the fundamental concepts of topic modeling, including the core components such as corpora, document-term matrices and topics as well as describes commonly used algorithms like Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA). The exploration of recent advances and trends in topic modeling highlights emerging techniques, the integration of deep learning, cross-domain and multilingual applications and real-world industry use cases. In conclusion, this literature review offers a comprehensive understanding of topic modeling, emphasizing its significance in text analysis and its evolving landscape in research and application domains.

Keywords: Topic Modeling, Natural Language Processing, Text Analysis, Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA) , Text Classification, Clustering and Sentiment Analysis.

1. INTRODUCTION

In today's information-rich world, the ability to efficiently process and extract valuable insights from vast volumes of textual data is of paramount importance. Natural Language Processing **(NLP)** techniques have emerged as indispensable tools in this endeavor, with topic modeling standing out as a pivotal approach. This literature review embarks on a comprehensive exploration of the field of topic modeling, shedding light on its historical evolution, core concepts, multifaceted applications, challenges, recent advancements and promising future directions. As the volume of digital text data continues to expand exponentially, the need for automated methods to organize, categorize and extract meaningful information from text becomes increasingly crucial. Topic modeling, a subfield of NLP, addresses this need by providing a framework for uncovering hidden thematic structures within large collections of text. Understanding the intricacies (details) of topic modeling is not only essential for researchers and practitioners in the field of NLP but also relevant to a wide range of disciplines, including information retrieval, content recommendation and social sciences [1].

This review begins with an exploration of the historical roots of topic modeling, tracing its origins and examining its evolution over time. By highlighting pivotal milestones and acknowledging the contributions of influential researchers, we aim to provide a contextual understanding of how this field has matured. Subsequently, the review delves into the fundamental concepts underpinning topic modeling. We unravel the core components of topic modeling, including corpora, document-term matrices and topics, while elucidating common algorithms such as Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA). These foundational elements serve as the building blocks upon which the topic modeling edifice is constructed.

Topic modeling, as a field of study within natural language processing and machine learning, has witnessed significant evolution and development over the years. This historical overview provides insight into the origins, key milestones and influential figures that have shaped the trajectory of topic modeling research.

- *Early Foundations (Late 20th Century):* The roots of topic modeling can be traced back to the late 20th century, where researchers explored statistical and mathematical methods to analyze text data. One of the earliest approaches, Latent Semantic Analysis (LSA), emerged in the 1990s, aiming to uncover the latent semantic structure in large text corpora.
- *Latent Dirichlet Allocation (LDA) (2003):* A major milestone in the field was the introduction of Latent Dirichlet Allocation by David Blei, Andrew Ng, and Michael Jordan in 2003. LDA revolutionized topic modeling by providing a probabilistic framework for discovering latent topics within a collection of documents. **LDA has since become one of the most widely used topic modeling techniques.**
- *Probabilistic Topic Models (Mid-2000s):* In the mid-2000s, the field of topic modeling saw an influx of probabilistic models beyond LDA. Researchers began developing variations like Dynamic Topic Models (DTM) and Author-Topic Models to account for dynamic topic evolution and authorship attribution, respectively.
- *Applications in Information Retrieval and Recommendation Systems:* As the 2000s progressed, topic modeling started finding practical applications in information retrieval and recommendation systems. Search engines and content recommendation algorithms incorporated topic modeling to improve the relevance of search results and recommendations to users.
- *Emergence of Non-Negative Matrix Factorization (NMF):* In parallel with LDA, Non-Negative Matrix Factorization (NMF) gained popularity as an alternative technique for uncovering latent topics in text data. NMF's ability to provide interpretable results contributed to its adoption in various applications.
- *Deep Learning Integration (Late 2010s):* In recent years, the integration of deep learning techniques, such as neural topic modeling and document embeddings, has reshaped the landscape of topic modeling. These approaches leverage neural networks to capture complex semantic relationships within text data.
- *Cross-Domain and Multilingual Topic Modeling (2010s):* Researchers began exploring topic modeling's applicability across different languages and domains. **Multilingual and cross-domain topic modeling techniques emerged to address the challenges** of working with diverse text data.
- *Ethical Considerations and Bias (2020s):* As topic modeling techniques became more prevalent in real-world applications, ethical concerns about bias and fairness in automated text analysis gained prominence. Researchers and practitioners started addressing these issues in their work.

As we move forward, **topic modeling continues to play a pivotal role in unlocking insights from textual data across various domains**. Despite its much strength, topic modeling is not without its challenges and limitations. This review candidly addresses these issues, including algorithmic complexities, scalability concerns, the subjectivity of evaluation metrics and ethical considerations regarding bias and fairness. Moving forward, the review spotlights recent advances and trends in topic modeling, revealing cutting-edge techniques, the integration of deep learning, cross-domain and multilingual applications and real-world success stories from industry and academia. Finally, we gaze into the future of topic modeling, offering predictions for upcoming research directions, highlighting the potential for interdisciplinary collaborations and addressing the ethical and societal implications of this evolving field. In summary, this literature review embarks on a journey through the realm of topic modeling, with the aim of providing a holistic understanding of its significance in the world of natural language processing and text analysis. By navigating the historical context, core concepts, applications, challenges, advancements and future possibilities, we invite readers to embark on a comprehensive exploration of this dynamic and transformative field.

2. LITERATURE REVIEW

Anton Eklund et al., [2] assessed the performance of **static NTM-CE models, which stands for Neural Topic Modeling by Clustering document Embeddings.** Their findings indicated that these static models performed at a comparable and in some cases, superior level when compared to other topic modeling approaches. However, an essential extension of static topic modeling is the transition to dynamic models, enabling the examination of how topics evolve over time and the identification of emerging and disappearing topics. This research proposal aims to address the complexities of dynamic topic modeling using NTM-CE, both in theory and practice. To achieve this, they present two primary research questions that will guide our exploration of this dynamic topic modeling approach. The research plan outlines four distinct phases, each contributing significantly to our understanding and practical implementation of dynamic topic modeling:

- *Development of a Codebase:* The first phase involves creating a comprehensive codebase specifically designed for evaluating dynamic topic models. This codebase will serve as a valuable resource for assessing the effectiveness and performance of various dynamic models, including NTM-CE.
- *Establishing a General Framework:* The second phase centers on the development of a versatile framework for efficiently constructing dynamic topic models using NTM-CE. This framework will provide researchers and practitioners with a structured methodology for implementing dynamic models effectively.
- *Practical Application and Dataset Insights:* In the third phase, we will apply the framework to a range of diverse datasets, gaining practical insights into its performance and adaptability. This step will provide valuable empirical evidence regarding the real-world application of dynamic topic modeling with NTM-CE.
- *Comprehensive Evaluation:* Evaluation is a critical aspect of this research. We propose to employ a dual evaluation approach, combining quantitative measurements of coherence with human evaluation. Additionally, we will leverage a recently developed evaluation tool to ensure robust and comprehensive assessments.

It is essential to note that the research questions and phases presented in this proposal represent our best efforts based on our current understanding. **Our aim is to thoroughly explore dynamic topic modeling using NTM-CE, encompassing both the theoretical underpinnings and practical** implementation aspects. In summary, this research proposal outlines our intention to delve deeply into the realm of dynamic topic modeling with NTM-CE. By posing two fundamental research questions and delineating the key phases of our investigation, we seek to advance our understanding of this evolving field and make substantial contributions to its theoretical foundations and practical utility.

Manuel V. Loureiro et al., [3] introduced an innovative approach to cluster-based topic modeling that leverages conceptual entities. Traditional topic models aim to uncover the latent structure within a corpus, primarily working with bag-of-words representations of documents. However, within the realm of topic modeling, much of the vocabulary can be either irrelevant for revealing underlying topics or strongly associated with relevant concepts, which can impact the interpretability of the resulting topics. Additionally, these models often exhibit limited expressiveness and demand substantial computational resources due to their reliance on language-specific data. In contrast, entities represent language-agnostic, real-world concepts enriched with rich relational information. To address these limitations, the researchers extracted vector representations of entities from two distinct sources: (i) an encyclopedic corpus utilizing a language model, and (ii) a knowledge base employing a graph neural network. Their research findings consistently demonstrated the superiority of their approach over other state-of-the-art topic models, as evidenced by coherence metrics. Notably, they observed that the explicit knowledge encoded within the graph-based embeddings yielded more coherent topics compared to the implicit knowledge encoded by contextualized embeddings from language models. This investigation delves into the realm of entity-based topic models, grounded in the clustering of vector representations of these entities. **The resulting model, referred to as TEC (Topics as Entity Clusters), internally represents documents using language-agnostic entity identifiers**. This unique approach ensures a single set of topics that can be shared across languages and extended to new languages without compromising the model's performance in existing languages.

Bernadeta Griciute et al., [4] conducted an analysis of Swedish newspaper articles regarding **COVID-19 using two prominent topic modeling techniques: Latent Dirichlet Allocation (LDA) and BERTopic.** The study encompassed a comprehensive corpus consisting of 6,515 articles, spanning a period of approximately one year and two months, from January 17, 2020, to March 13, 2021. The primary objective of this research was to gain insights into the societal impact of the Swedish government's policies concerning the COVID-19 pandemic. To achieve this, the researchers employed topic modeling methodology on a collection of Swedish newspaper articles that covered a year-long period. The research began by providing essential background information on the pandemic, contextualizing the study within the broader scientific research landscape on the topic. It subsequently elucidated the details of the corpus compiled for the analysis and outlined the mathematical models employed, particularly emphasizing the use of Latent Dirichlet Allocation (LDA) and the associated toolkits. **The core of the research lay in the presentation of the topic modeling outputs generated through the application of LDA and dynamic topic modeling (DTM).** These outputs were categorized into several themes, spanning topics related to the virus's origin, strategies employed by the World Health Organization **(WHO),** public health recommendations, scientific biomedical research and economic discussions. The successful application of the LDA method in this research underscored its effectiveness in uncovering and categorizing key themes within the corpus.

Kostadin Cvejoski et al., *[5]* they harnessed recent advancements in neural variation inference to introduce an innovative neural approach to the dynamic Focused Topic Model. This research represents a significant development in the field, as it presents a neural model for tracking topic evolution. This model leverages sequences of Bernoulli random variables to monitor the appearance of topics, effectively dissociating topic activities from their proportions. The **researchers evaluated their model using three diverse datasets, namely the UN general debates, the collection of NEURIPS papers, and the ACL Anthology dataset.** Their findings demonstrated the model's superiority in generalization tasks when compared to state-of-the-art topic models. Moreover, it performed on par with these models in prediction tasks while maintaining a similar parameter count and converging approximately twice as quickly. They have made their source code available online to facilitate the replication of their experiments. This research marks the introduction of the Neural Dynamic Focused Topic Model (NDF-TM), specifically designed for sequentially collected data. Notably, NDF-TM explicitly disentangles dynamic topic proportions from topic activities by incorporating sequences of Bernoulli variables. **The study results consistently revealed coherent and diverse topics, effectively capturing historical events.** Looking ahead, the researchers envision future work involving the integration of NDF-TM with Variation Autoencoders (VAEs) for topic-guided text generation. This promising avenue holds the potential to further advance the capabilities of topic modeling in text analysis and generation tasks.

Hamed Rahimi et al., *[6]* introduced a family of dynamic topic models known as Aligned Neural Topic Models (ANTM). **ANTM represents a novel approach that incorporates advanced data mining algorithms to create a modular framework for the discovery of evolving topics**. This innovative framework maintains the temporal continuity of evolving topics by harnessing time-aware features extracted from documents using advanced pre-trained Large Language Models **(LLMs)**. It also employs an overlapping sliding window algorithm for sequential document clustering. The unique feature of the overlapping sliding window algorithm is its ability to identify varying numbers of topics within each time frame. Moreover, it aligns semantically similar document clusters across different time periods. This alignment process effectively captures emerging trends and fading topics across distinct time spans, resulting in a more interpretable representation of evolving topics. Experiments conducted on four diverse datasets demonstrate that ANTM surpasses probabilistic dynamic topic models in terms of topic coherence and diversity metrics. **Additionally, it enhances the scalability and adaptability of dynamic topic models, making it accessible and versatile for various algorithmic approaches.** Existing dynamic topic models often overlook specific temporal variations within evolving topics, typically configuring a global structure for dynamic topics, such as maintaining the same number of document clusters in each period. Moreover, some of these models face challenges when applied to large archives with extensive vocabularies and struggle to handle short texts, as is often the case in social network analysis applications.

Meisam Dastani et al., *[7]* conducted an extensive analysis of the **evolving topics within scientific publications related to tuberculosis using text mining techniques and co-word analysis**. Their approach was analytically driven, aiming to uncover trends and patterns within this body of literature. The study's statistical population encompassed all global publications pertaining to tuberculosis. To extract the necessary data, the researchers utilized the Scopus citation database, spanning the period from 1900 to 2022. The selection of primary keywords for the search strategy was a result of consultations with subject matter experts and the use of MESH terminology. Data analysis was carried out using the Python programming language and VOSviewer software. The results of their analysis revealed four primary topics within the corpus of tuberculosis-related scientific publications:

- Clinical Symptoms (41.8%)
- Diagnosis and Treatment (28.1%)
- Bacterial Structure, Pathogenicity and Genetics (22.3%)
- Prevention (7.84%)

These findings hold significant value for organizations involved in tuberculosis research and healthcare. They provide insights that can inform decision-making processes, guide the allocation of resources and facilitate the development of programs and guidelines aimed at addressing and combating this disease.

The outcomes of this research not only benefit experts, physicians and decision-makers in their efforts to control tuberculosis but also assist healthcare organizations in planning and executing targeted interventions within regional health centers. Additionally, the World Health Organization (WHO), responsible for monitoring and managing various diseases, including tuberculosis, can utilize the results of this investigation to make informed decisions, track the progress of tuberculosis studies and allocate resources effectively to combat this disease.

Ashwini Pachore et al., *[8]* developed a system designed to collect and process public opinions regarding products as expressed in various forms such as **blog posts, comments, reviews and tweets, with the aim of assessing public sentiment.** Their study incorporates a data preprocessing phase where tweets are cleaned by removing punctuation, special symbols, hashtags and URLs. To extract themes from the collected data, the researchers employ topic modeling techniques, specifically Latent Dirichlet Allocation (LDA). Furthermore, they introduce Principal Components Analysis **(PCA)** as a dimensional reduction strategy, striving to identify the minimal number of Principal Components (PCs) necessary to achieve optimal classification performance. **The study also makes use of K-means clustering to group similar words in tweets, complemented by cluster analysis.** This approach involves representing words as vectors using the GloVe model, enabling the grouping of tweets through K-means clustering. For sequential text analysis, the study employs the BERT paradigm, which involves sequentially reading text inputs. Long Short-Term Memory **(LSTM)** models are applied to anticipate sequences effectively. To maintain semantic associations between words in a low-dimensional embedding space, the researchers utilize Word2Vec, a powerful word embedding technique capable of handling small text corpora with extensive vocabularies. The study further introduces the use of "T-SNE," a dimensionality reduction technique that maps each data point to a two- or three-dimensional space for a more intuitive visualization of high-dimensional data. In addition to these techniques, the researchers evaluate their models and outcomes using various performance metrics, including accuracy, F1-score, and a confusion matrix. These metrics contribute to a more comprehensive validation of the models and their results.

Jing Li et al., *[9]* introduced the Dynamic Graph Convolutional Network (DynGCN), a novel approach that combines spatial and temporal convolutions in an interleaved manner. This unique network architecture incorporates a model adaptation mechanism, enabling the dynamic adjustment of model parameters to align with evolving graph snapshots. Recent advancements in representation learning on graphs have generated significant interest, with Graph Convolutional Networks (GCN) achieving remarkable performance in various graph-related tasks. However, **the majority of existing methods primarily focus on static graphs, overlooking the dynamic nature of real-world graph structures**. While some recent studies have attempted to integrate sequence modeling, such as Recurrent Neural Networks **(RNN),** into the GCN framework, they often struggle to capture the dynamic changes in graph structural information over time. **DynGCN addresses this limitation by effectively extracting both structural dynamism and temporal dynamism in dynamic graphs**. The model is designed to accommodate the evolving nature of real-world graphs, ensuring that it captures both spatial and temporal aspects. The researchers conducted a series of extensive experiments using real-world datasets, focusing on tasks such as link prediction and edge classification. The results of these experiments clearly demonstrate that DynGCN outperforms existing state-of-the-art methods in these tasks, highlighting its effectiveness in capturing and leveraging the dynamic nature of graph data.

Huaqing Cheng et al., *[10]* introduced a novel neural topic model that integrates SBERT and data augmentation techniques. Topic models serve the purpose of extracting consistent themes from extensive corpora for research applications. Recent developments have seen an increasing interest among scholars in combining pre-trained language models with neural topic models. However, **this approach has its limitations, particularly when dealing with short texts, as it often results in low-quality and incoherent topics due to reduced word frequency** and insufficient word co-occurrence in short texts compared to longer ones.
To address these challenges, the researchers proposed a multi-step approach:

- Easy Data Augmentation (EDA): A data augmentation method, EDA, is introduced, which involves keyword combination to mitigate the sparsity issue in short texts.
- Attention Mechanism: An attention mechanism is utilized to prioritize keywords relevant to the topic, reducing the influence of noise words.
- SBERT Model: The SBERT model is trained on a comprehensive and diverse dataset, allowing it to generate high-quality semantic information vectors specifically tailored for short texts.
- Feature Fusion: Augmented data, weighted using the attention mechanism and enriched with high-quality semantic information, undergo feature fusion. These fused features are subsequently input into a neural topic model.

The experimental results, based on an English public dataset, demonstrate the effectiveness of this approach. The model consistently generates high-quality topics, with average scores exhibiting a notable improvement of 2.5% for topic coherence and 1.2% for topic diversity when compared to the baseline model. In summary, Huaqing Cheng and their team have developed an innovative neural topic model that overcomes the challenges posed by short texts. By combining data augmentation, attention mechanisms, SBERT and feature fusion, their model yields higher-quality topics, offering a significant improvement over existing methods in terms of coherence and diversity.

Haichao Sun et al., *[11]* introduce a versatile recommendation algorithm based on Non-negative Matrix Factorization (NMF) with the aim of predicting and suggesting labels for new samples. NMF is a well-established technique in intelligent systems that are commonly used to break down a nonnegative matrix into two distinct factor matrices: a basis matrix and a coefficient matrix. The primary objective of NMF is to ensure that the mathematical operations involving these two matrices closely approximate the original matrix while maintaining algorithm stability and generalization capability. **The authors delve into the generalization performance of NMF algorithms, emphasizing aspects of algorithm stability and providing bounds for generalization errors.** This analysis, referred to as AS-NMF, seeks to enhance the understanding of NMF's behavior in terms of stability and its ability to generalize to new data.

Their approach includes several key steps:

- General NMF Prediction Algorithm: They propose a comprehensive NMF prediction algorithm capable of predicting labels for new samples, accompanied by the definition of a corresponding loss function.
- Stability Assessment: The authors define the stability of the NMF algorithm based on the loss function. **They establish two generalization error bounds by employing uniform stability,** considering scenarios where the parameters are either fixed or not fixed under the multiplicative update rule.
- Framework Establishment: A robust and inclusive framework is established for the analysis and measurement of generalization error bounds for the NMF algorithm.

The experimental results validate the effectiveness of their approach on three widely recognized benchmark datasets. These results demonstrate that AS-NMF not only achieves efficient performance but also surpasses state-of-the-art recommendation models in terms of model stability, showcasing its advantages in recommending tasks. In summary, Haichao Sun and their team have introduced a versatile recommendation algorithm rooted in Non-negative Matrix Factorization (NMF). Their analysis of algorithm stability and generalization error bounds, termed AS-NMF, sheds light on the behavior of NMF in recommendation tasks. Their experimental results further emphasize the efficiency and superior stability of AS-NMF, marking it as a noteworthy advancement in recommendation systems.

Weisi Chen et al., *[12]* undertook a comparison of three cutting-edge topic modeling techniques: Latent Dirichlet Allocation (LDA), Top2Vec, and BERTopic, in a specific context of analyzing news impact on financial markets. They conducted this analysis on a dataset comprising 38,240 news articles, each with an average length of 590 words. News impact analysis has become a prevalent task for finance researchers, involving the categorization of news articles based on themes and sentiments, linking news events to relevant stocks, and gauging the influence of selected news on stock prices. **To enhance the efficiency of news selection and analysis, topic modeling techniques are applied to distill meaningful topics from a large volume of news documents.** In response to this, the researchers propose a service-oriented framework for news impact analysis, known as "News Impact Analysis" (NIA). This framework leverages multiple topic models to create an automated and streamlined news impact analysis process, catering to the needs of finance researchers. The experimental outcomes reveal that BERTopic exhibited superior performance in this specific scenario. It achieved these results with minimal data preprocessing, producing the highest coherence score, offering the best interpretability, and maintaining reasonable computational efficiency. Moreover, the framework was validated through the successful execution of the entire news impact analysis process by a finance researcher, affirming the feasibility and usability of the NIA framework. In summary, Weisi Chen and their team conducted a comprehensive evaluation of topic modeling methods in the context of news impact analysis on financial markets. Their findings indicated that BERTopic was the most effective choice for this task, showcasing its advantages in terms of performance, interpretability and usability within the NIA framework.

Yixin Zou et al., [13] conducted a comprehensive analysis and discussion of 52,891 posts related to digital fashion and virtual fashion, which were published on social networking sites. Their investigation employed various analytical techniques, including k-means clustering analysis, Latent Dirichlet Allocation (LDA) topic modeling and sentiment analysis. **The primary focus of this study was to explore public perceptions, prevailing topics and the development trends** within the digital fashion industry. The study unveiled intriguing insights:

- Public Sentiment: The analysis revealed that both positive and neutral emotions are prevalent in the public's attitude toward digital fashion.

- Diverse Discussion Topics: A wide spectrum of topics emerged from the discussions, highlighting the multifaceted nature of digital fashion.
- Impact of Digital Technology: Innovations in digital technology were identified as pivotal influencers in various aspects of the fashion industry, including job creation, talent demand, marketing strategies, profit models and overall innovation within the fashion-related business ecosystem.

This research not only serves as a valuable reference for scholars and researchers in related fields, providing insights into research methods and potential research directions but also offers a rich dataset and case study for industry practitioners. It stands as a comprehensive resource for understanding the evolving landscape of digital fashion, including public perceptions, key topics, and industry dynamics.

Lu Cheng et al., *[14]* introduce a method that places two distinct sets of topics into a shared, interpretable vector space, utilizing an entropy-based measure to quantify shifts in these topics. This research employs topic modeling through non-negative matrix factorization on cognitive science publications from both before and after 2012. This enables a comprehensive exploration of how the field has evolved since the resurgence of neural networks in the adjacent field of AI/ML. **The study presents case studies on topics that have either faded (dull) away** (e.g., the connectionist/symbolic AI debate) or emerged anew (e.g., the intersection of art and technology). To identify newly emerging and vanishing topics, the research employs a similarity matrix (S) based on cosine similarity, which normalizes vectors to produce values between 0 and 1. **This metric aids an identifying highly similar topic, although it may occasionally lead to the false perception of two disparate topics as nearly identical.** Moreover, **it does not consider the actual number of papers within each topic,** which could influence the extent of one topic's impact on others. The framework developed in this study can be adapted to investigate various fields or historical events marked by significant shifts in thought. Such insights have the potential to drive more efficient and impactful scientific discoveries, benefitting research in diverse domains. In summary, Lu Cheng and their team have introduced an innovative method for understanding the evolution of topics in cognitive science publications. By placing topics into a shared vector space and quantifying shifts, they provide a valuable tool for exploring changes in thought over time, applicable to diverse fields and historical events.

Nikolai Gerasimenko et al., *[15]* delve into recent advancements in topic modeling with the aim of accurately identifying emerging trends at an early stage. The ever-increasing volume of scientific publications and the rapid emergence of new research directions pose a challenge for the scientific community in promptly and automatically identifying trends. In this research, trends are defined as semantically coherent themes characterized by a lexical core that evolves steadily over time and is accompanied by a sharp, often exponential, increase in the number of publications.

Specifically, **the research customizes the conventional ARTM-based approach and introduces an innovative incremental training technique, enabling the model to operate in real-time with data**. Additionally, the researchers curate the Artificial Intelligence Trends Dataset **(AITD),** which includes a collection of early-stage articles and a set of key collocations associated with each trend. The experimental results showcase the superiority of the proposed ARTM-based approach over traditional models such as PLSA, LDA and a neural approach based on BERT representations. It is worth noting that both the models and the dataset are made available for research purposes. The research also outlines the validation process and introduces a method for aligning labeled trends with extracted topics. To validate the efficacy of their approaches in early trend topic detection, the researchers compiled a meticulously (carefully) labeled specialized dataset, AITD. This dataset comprises 91 groups of machine learning and AI articles, each corresponding to a distinct trend topic. **These trends are identified using keywords extracted from publications in top conferences, along with alternative trend names**. In conclusion, Nikolai Gerasimenko and their team have contributed to the field of early trend detection by proposing an ARTM-based approach that excels in extracting emerging trends at their initial stages. This approach, characterized by its real-time capabilities and efficiency, is complemented by the creation of the AITD dataset, further supporting research in this domain.

Federico Ravenda et al., *[16]* introduced a novel model based on Mixtures of Normalizing Flows, designed to dynamically extract topics from a collection of time-varying documents. This innovative model operates by taking embeddings generated by a pre-trained transformer-based language model for each timestamp as input and subsequently learning the parameters of a sophisticated high-dimensional mixture density distribution. **These parameters are then employed to cluster documents into groups, each of which represents a distinct topic**. Importantly, this model continually updates the mixture distribution's parameters at each timestamp and generates topic representations using the TF-IDF procedure. Key features of this proposed model include:

- *Temporal Clustering:* It employs a temporal clustering approach based on mixtures of normalizing flows, allowing the parameters to evolve dynamically over time.
- *Probabilistic Nature:* The model is inherently probabilistic, enabling the inference of topic characteristics not only from qualitative representations but also from the evolving distribution parameters over time. For instance, changes in estimated means can signify topic evolution, while small variances may indicate cohesive clusters representing coherent topics.
- *Suitability for High-Dimensional Data:* **Normalizing flows are particularly well-suited for high-dimensional data,** eliminating the need for intermediate dimensionality reduction steps.

In summary, Federico Ravenda and their team have introduced an advanced model that excels at dynamically extracting topics from evolving documents. This probabilistic approach leverages the power of normalizing flows, providing valuable insights into topic evolution and coherence without the need for dimensionality reduction steps.

Belal Abdullah Hezam Murshed et al., *[17]* conducted extensive research in the domain of Short Text Topic Modeling **(STTM),** a field that has recently garnered significant interest for uncovering coherent latent topics within brief textual content. Their article provides a comprehensive exploration of the current state of STTM algorithms, offering a detailed survey and taxonomy of these algorithms. **The research also comprises a qualitative and quantitative examination of STTM techniques, highlighting their respective strengths and weaknesses**. Furthermore, it includes a comparative analysis of representative STTM models, assessing the quality of topics and their performance. **The study is particularly relevant in the context of social media platforms like Twitter, Facebook, and Weibo, where individuals, groups,** and organizations increasingly rely on these platforms as rich sources of information. Such social media-generated data often takes the form of short, voluminous, sparse, and low-density text. The performance evaluation conducted in this research utilizes real-world Twitter datasets, specifically the Real-World Pandemic Twitter (RW-Pand-Twitter) dataset and the Real-world Cyberbullying Twitter (RW-CB-Twitter) dataset. Evaluation metrics include topic coherence, purity, normalized mutual information **(NMI)** and accuracy. In closing, this research not only provides valuable insights into the current landscape of short text topic modeling but also identifies open challenges and suggests future research directions in this promising field. It serves as a valuable resource for researchers seeking to understand the state-of-the-art in short text topic modeling and for those actively involved in developing new algorithms for this purpose.

Supriya Kinariwala et al., *[18]* introduced a novel model called G_SeaNMF (Gensim_SeaNMF) aimed at enhancing the semantic relationships between words by employing both local and global word embedding techniques. Word embeddings derived from extensive corpora offer valuable general semantic and syntactic information about words, serving as guidance for topic modeling, especially in the context of collections with short text where sparse co-occurrence patterns are common. The proposed model combines SeaNMF (Semantics-assisted Non-negative Matrix Factorization) with the word2vec model from the Gensim library to reinforce the semantic relationships between words. This article explores short text topic modeling techniques based on **DMM** (Dirichlet Multinomial Mixture), self-aggregation and global word co-occurrence. These techniques are evaluated using various measures to assess cluster coherence on real-world datasets, including Search Snippet, Biomedicine, Pascal Flickr Tweet, and TagMyNews. The empirical evaluation demonstrates that the fusion of local and global word embeddings yields more relevant words for each topic, resulting in improved outcomes. **Notably, G_SeaNMF exhibits the highest purity, indicating that topics are consistent and most words are drawn from a single category.** This enhances the accurate assignment of labels to each topic. However, the primary limitation of the proposed G_SeaNMF model is its reliance on a single larger external corpus dataset for obtaining pre-trained word representations. **Additionally, the model's processing speed is relatively slow, balancing the trade-off between implicit and explicit short-text relationships**. Future research directions may focus on further improving the topic model's consistency and exploring automated label generation for each topic.

Marie Uncovska et al ., *[19]* conducted a comparative analysis of user experiences between DiGAs (Digital Health Applications) and non-prescription mHealth apps in Germany. Their analysis encompassed both average app store ratings and written reviews, introducing a pioneering approach that utilizes BERTopic for sentiment analysis and topic modeling within the mHealth research domain. The dataset employed for this study consisted of 15 DiGAs and 50 comparable apps, compiling a total of 17,588 reviews written in the German language. The findings indicate that DiGAs tend to receive higher contemporary ratings in comparison to non-regulated apps (Android: 3.82 vs. 3.77; iOS: 3.78 vs. 3.53; p < 0.01; assessed via non-parametric Mann–Whitney–Wilcoxon test). **Noteworthy factors contributing to a positive user experience with DiGAs include exceptional customer service and personalization (15%) and user-friendly interfaces (13%).** However, DiGAs face their own set of challenges, notably issues related to software bugs (24%) and a somewhat cumbersome registration process (20%). Negative user reviews often express concerns about the effectiveness of therapy (11%). Conversely, non-regulated apps face a primary concern related to pricing, with excessive pricing mentioned by 27% of users. **Interestingly, user attention to data privacy and security is relatively limited in both DiGAs (0.5%) and non-regulated app reviews (2%)**. In conclusion, the study indicates that DiGAs are generally well-received based on ratings and sentiment analysis of user reviews. However, addressing pricing concerns within the non-regulated mHealth sector is of paramount importance. The incorporation of user experience evaluation into the review process could potentially enhance adherence and health outcomes in this context.

Lijimol George et al., *[20]* introduce a hybrid model that combines Bidirectional Encoder Representations from Transformers (BERT) and Latent Dirichlet Allocation (LDA) for in-depth topic modeling with clustering based on dimensionality reduction. Given the computational complexity associated with clustering algorithms, which increases with a higher number of features, this study also explores dimensionality reduction techniques such as PCA, t-SNE and UMAP. **As part of their research, a unified clustering-based framework is proposed, leveraging both BERT and LDA, to extract meaningful topics from extensive text corpora**. To assess the efficacy of this cluster-informed topic modeling framework using BERT and LDA, experiments are conducted simulating user input on benchmark datasets. The results of these experiments demonstrate that clustering with dimensionality reduction enhances the inference of more coherent topics. Therefore, this unified approach, combining clustering with BERT-LDA, proves effective for the development of topic modeling applications. The researchers utilize the CORD-19 dataset after preprocessing and the hybrid model with **UMAP** (Uniform Manifold Approximation and Projection) dimensionality reduction yields comparatively superior results for this input data. **The reduced dimensionality results are then subjected to the k-means clustering algorithm, with the precise number of clusters determined using the Elbow Method**. The model is developed by amalgamating the probabilistic subject assignment vector from the LDA model and sentence vectors extracted from the BERT model. This hybrid approach preserves semantic information and creates contextual topic information, enhancing the effectiveness of the topic modeling process.

Bruno Spilak et al., *[21]* introduce a novel portfolio allocation methodology centered on risk factor budgeting, leveraging convex Nonnegative Matrix Factorization **(NMF)**. Distinguishing itself from classical factor analysis, PCA or ICA, NMF ensures the presence of positive factor loadings, thereby enabling the creation of easily interpretable long-only portfolios. The NMF factors represent distinct sources of risk, resulting in a quasi-diagonal correlation matrix that fosters diversified portfolio allocations. This study assesses the performance of their approach within the context of volatility targeting, applying it to two long-only global portfolios encompassing crypto currencies and traditional assets. **The findings reveal that their method surpasses classical portfolio allocations in terms of diversification and exhibits a superior risk profile compared to hierarchical risk parity (HRP).** To ensure the robustness of their conclusions, Monte Carlo simulations are conducted. Their novel portfolio allocation strategy, named NMF Risk Budget (NMFRB), distributes risk equally among synthetic asset classes and assigns greater risk to assets that exhibit stronger representation. This allocation strategy fosters diversification at both the latent factor and original asset levels. The stability of the factors in their interpretation is demonstrated, highlighting persistent correlations between assets. **Furthermore, these factors consistently deliver effective diversification across various market regimes, even if they are not perfectly uncorrelated**. As a result, this method outperforms HRP in terms of risk-adjusted returns over an extended historical period.

Haein Lee et al., [22] presented his findings from an unsupervised learning algorithm (BERTopic) applied to both LexisNexis and Web of Science datasets. The ESG-related topics identified in the LexisNexis data spanned from industry changes over time to topics related to asset indicators and Asian companies. In contrast, the topics extracted from the Web of Science data primarily revolved around the performance of the energy industry and companies, starting with subjects related to their impact on the bond market. By analyzing the evolution of these topics over time, the study revealed that the concept of efficiency was initially addressed in academia in response to discussions initiated by international organizations like the United Nations. Subsequently, the media began to focus on this content from various angles. The results indicate that, in the trend observed in international news reports, ESG topics and their connection to the asset market were strongly emphasized in the hierarchical composition of topics. This emphasis was particularly pronounced in the time series analysis. Furthermore, it was evident that **topics concerning ESG and the bond market received hierarchical emphasis in international academia trends as well**. This suggests that there is a shared recognition of the economic value of ESG factors from both academic and international media perspectives. This perspective breaks away from the limitations of previous studies that often emphasized the significance of ESG solely at the national level. In essence, these findings underscore the importance of global attention and active social discourse surrounding ESG matters.

Sara J. Weston et al., [23] provide an overview of tools available to researchers for identifying the number and labels of topics in topic modeling. The process involves several steps:

- Narrowing Down Candidate Models: Initially, ***the researchers outline a procedure to reduce a large set of models to a select number of candidate models***. This entails comparing various models using fit metrics such as exclusivity, residuals, variational lower bound, and semantic coherence.
- Comparing a Small Number of Models: From the narrowed-down set of candidate models, a smaller subset is selected based on the research project's goals. Factors such as topic representativeness and solution congruence guide this selection process.
- Labeling Topics: **The authors also discuss tools for labeling topics**. These include identifying frequent and exclusive words associated with each topic, highlighting key examples, and exploring correlations among topics.

In essence, ***the paper offers a comprehensive guide for researchers to effectively identify, evaluate, and label topics in their topic modeling projects.***

3. CONCLUSION

Topic modeling's historical roots trace back to the late 20th century, with early foundations laid by researchers exploring mathematical and statistical methods to extract latent structures from text data. However, it was the introduction of Latent Dirichlet Allocation (LDA) in 2003 that marked a turning point, revolutionizing the way we uncover hidden topics within large text corpora. LDA provided a probabilistic framework that has since become a cornerstone of topic modeling. As we ventured further, we uncovered the rich tapestry of applications where topic modeling has left an indelible mark. ***From text classification and clustering to sentiment analysis, information retrieval and content summarization, topic modeling has consistently proven its versatility and utility.*** Its integration into search engines, recommendation systems, and content analysis tools has enhanced the efficiency and effectiveness of these systems, benefiting users across the digital landscape. **Yet, our exploration also unveiled the challenges and limitations that accompany topic modeling.** Algorithmic complexities, scalability concerns, subjective evaluation metrics, and ethical considerations surrounding bias and fairness have highlighted the need for ongoing research and refinement in this field. In recent years, the integration of deep learning techniques, cross-domain and multilingual applications and the exploration of industry use cases have reshaped the landscape of topic modeling. These developments have positioned topic modeling as a dynamic and transformative field with vast potential. As we peer into the future, we see a promising horizon for topic modeling. **Anticipated research directions include refining existing techniques, exploring interdisciplinary collaborations and addressing the ethical and societal implications of automated text analysis**. Topic modeling is poised to continue its journey as an essential tool for extracting valuable insights from textual data in an ever-evolving digital landscape.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1]. Fu, Q., Zhuang, Y., Gu, J., Zhu, Y., & Guo, X. (2021). Agreeing to disagree: Choosing among eight topic-modeling methods. Big Data Research, 23, Article 100173. https://doi .org/10.1016/j.bdr.2020.100173

[2]. Anton Eklund, Mona Forsman,Frank Drewes, "Dynamic Topic Modeling by Clustering Embeddings from Pretrained Language Models: A Research Proposal", Proceedings of the AACL-IJCNLP 2022 Student Research Workshop, pages 84–91 November 20, 2022. Association for Computational Linguistics

[3]. Manuel V. Loureiro , Steven Derby and Tri Kurniawan Wijaya, "Topics as Entity Clusters: Entity-based Topics from Language Models and Graph Neural Networks", International Joint Conferences on Artificial Intelligence Organization. Survey Track.represenarXiv: 2301.02458v1 , 6 Jan 2023

[4]. Bernadeta Griciute,Lifeng Han,Goran Nenadic,"Topic Modelling of Swedish Newspaper Articles about Coronavirus: a Case Study using Latent Dirichlet Allocation Method",The 11th IEEE International Conference on Healthcare Informatics, Houston, Texas, USA,June , 2023

[5]. Kostadin Cvejoski, Ramses J. Sanchez, Cesar Ojeda, "Neural Dynamic Focused Topic Model", 2023, Association for the Advancement of Artificial Intelligence.

[6]. Hamed Rahimi,Hubert, Naacke,Camelia , Constantin,Bernd Amann,"ANTM: An Aligned Neural Topic Model For Exploring Evolving Topics", In Proceedings of 2023 Association for Computing Machinery. ACM, New York, NY, USA, 11 pages.

[7]. Meisam Dastani , Alireza Mohammadzadeh , Jalal Mardaneh , and Reza Ahmadi,"Topic Analysis and Mapping of Tuberculosis Research Using Text Mining and Co-Word Analysis",Hindawi Tuberculosis Research and Treatment Volume 2022, Article ID 8039046, 7 pages

[8]. Ashwini Pachore and Vrishali Chakkarwar,"Opinion Mining and Tweet Analysis Using Topic Modeling by LDA with BERT and GLOVE Embedding",ACVAIT 2022, AISR 176, pp. 660–673, 2023

[9]. Jing Li, Yu Liu, and Lei Zou,"DynGCN: A Dynamic Graph Convolutional Network Based on Spatial-Temporal Modeling", WISE (2020).

[10]. Huaqing Cheng , Shengquan Liu , Weiwei Sun and Qi Sun, "A Neural Topic Modeling Study Integrating SBERT and Data Augmentation", Appl. Sci. 2023, 13, 4595. https://doi.org/10.3390/app13074595 and https://www.mdpi.com/journal/applsci

[11]. Haichao Sun and Jie Yang,"The Generalization of Non-Negative Matrix Factorization Based on Algorithmic Stability", Electronics 2022, 12, 1147. https://doi.org/10.3390/electronics12051147 and https://www.mdpi.com/journal/electronics

[12]. Weisi Chen, Fethi Rabhi, Wenqi Liao and Islam Al-Qudah,"Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study",Electronics 2023, 12, 2605. Electronics 2023, 12, 2605. https://doi.org/10.3390/electronics12122605 and https://www.mdpi.com/journal/electronics

[13]. Yixin Zou, Ding-Bang Luh and Shizhu Lu,"Public perceptions of digital fashion: An analysis of sentiment and Latent Dirichlet Allocation topic modeling", Frontiers in Psychology , Psychol. 13:986838. doi: 10.3389/fpsyg. 986838, (2022).

[14]. Lu Cheng, Jacob G. Foster and Harlin Lee,"A Simple, interpretable method to identify surprising topic shifts in scientific fields",Frontiers in Research Metrics and Analytics, Front. Res. Metr. Anal. 7:1001754. doi: 10.3389/frma.1001754, (2022).

[15]. Nikolai Gerasimenko,Alexander Chernyavskiy,Maria Nikiforova,Anastasia Ianina,Konstantin Vorontsov,"Incremental Topic Modeling for Scientific Trend Topics Extraction",Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2023”, June, 2023

[16]. Federico Ravenda, Andrea Raballo, Antonietta Mira and Fabio Crestani,"Incremental Mixture of Normalizing Flows for Dynamic Topic Modelling",IIR2023: 13th Italian Information Retrieval Workshop, June 2023, Pisa, Italy

[17]. Belal Abdullah Hezam Murshed, Suresha Mallappa, Jemal Abawajy, Mufeed Ahmed Naji Saif, Hasib Daowd Esmail Al-ariki, Hudhaifa Mohammed Abdulwahab,"Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis", Artificial Intelligence Review (2023) 56:5133–5260

[18]. Supriya Kinariwala , Sachin Deshmukh,"Short text topic modelling using local and global word-context semantic

correlation",Multimedia Tools and Applications (2023) 82:26411–26433, https://doi.org/10.1007/s11042-023-14352-x

[19]. Marie Uncovska , Bettina Freitag, Sven Meister and Leonard Fehring,"Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany",npj Digital Medicine (2023) 6:115 ; https://doi.org/10.1038/s41746-023-00862-3

[20]. Lijimol George, P. Sumathy,"An integrated clustering and BERT framework for improved topic modeling",Int. j. inf. tecnol. (April 2023) 15(4):2187–2195 https://doi.org/10.1007/s41870-023-01268-w

[21]. Bruno Spilak, Wolfgang Karl Hardley, "Risk Budget Portfolios With Convex Non-negative Matrix Factorization", June 2023. Available at SSRN: https://ssrn.com/abstract=4474100 or http://dx.doi.org/10.2139/ssrn.4474100

[22]. Haein Lee, Seon Hong Lee, Kyeo Re Lee and Jang Hyun Kim,"ESG Discourse Analysis Through BERTopic: Comparing News Articles and Academic Papers",Computers, Materials & Continua, 2023, 75(3), 6023-6037. https://doi.org/10.32604/cmc.2023.039104

[23]. Sara J. Weston , Ian Shryock, Ryan Light, and Phillip A. Fisher,"Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial",Advances in Methods and Practices in Psychological Science April-June 2023, Vol. 6, No. 2, pp. 1– 13, sagepub.com/journals-permissions, DOI: 10.1177/25152459231160105