International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I6P101

Volume 74 Issue 6, 1-16, June 2026

Original Article

An Enhanced Topic Extraction Model for Medical PubMed Documents using State-of-the-Art Algorithms

K.T. Mathuna[1], I. Elizabeth Shanthi[2]

[1,2]*Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamilnadu, India.*

[1]*Corresponding Author : 19phcsf001@avinuty.ac.in*

Received: 08 June 2024 Revised: 11 April 2025 Accepted: 20 April 2026 Published: 27 June 2026

Abstract *- The rapid growth of medical literature databases represents both a challenge and an opportunity for pharmacovigilance. Medical abstracts are full of specialized terms and complex sentences that make extracting meaningful insights on the adverse effects of drugs very challenging. This paper addresses the critical problem of extracting relevant topics related to drug adverse effects from PubMed medical abstracts using advanced topic modelling methods. It enhances the four-topic modelling with two optimization algorithms to improve topic extraction and assesses their performance, such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN), combined with grid search and Bayesian optimization algorithms. The experimental results show that LDA optimized with Bayesian optimization gives the highest coherence score, 0.605, which is better than other models. Coherent results, as shown in a complex comparison table, reveal the performance of each model and optimization method.*

Keywords *- Topic Modeling, LDA, LSA, LSTM, RNN, and Grid Search.*

1. Introduction

The biomedical research field is evident by the quickly increasing number of publications, which poses a challenge and an opportunity for pharmacovigilance. The challenge is based on the large amount of information, and the opportunity is to uncover essential information about drug safety, which is buried in the complex vocabulary of biomedical research abstracts, an essential part of the biomedical databases like PubMed. These documents are characterized by a large amount of unstructured and sporadic information, displayed in several forms like medical documents, research studies, digital health files, and clinical case studies [1]. Deciphering the complex landscape of biomedical data goes beyond just interpreting the technical jargon, and it requires the utilization of sophisticated computing technologies. Topic modeling is a segment of Natural Language Processing (NLP) to derive insights from an unstructured dataset [2]. It effectively segments massive volumes of text into clear and systematically organized themes. However, the distinct challenges posed by medical texts necessitate more than the standard approaches to topic modeling.

Traditional and advanced methods are being used side by side recently to identify topics as a solution to present challenges. Conventional topic modeling techniques rely on statistical and probabilistic approaches and are based on the assumption that each document in a corpus is a combination of several topics. A topic is characterized by a specific distribution of words. The different models for traditional topic modeling, such as Latent Dirichlet Allocation (LDA) [3], Latent Semantic Analysis (LSA) [4], Probabilistic Latent Semantic Analysis (PLSA) [5], and Non-negative Matrix Factorization (NMF) [6], are known for revealing the latent thematic structures in large texts. Deep learning for topic modeling means the use of advanced neural network architectures to get topics from large text datasets [7]. These methods are far more intricate than traditional topic modeling techniques, and they can handle the subtleties and complexities of language better. Recently, the deep learning and neural network technologies leading to the rise of are Convolutional Neural Networks (CNNs) [8], Recurrent Neural Networks (RNNs) [9] and the derivatives of these such as Long Short-Term Memory (LSTM) [10], Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM) [11, 12] which provide advanced capabilities but generally require further optimization for the efficient handling of the complexities and nuances of the biomedical language. It can reveal very subtle patterns and relationships in the data, thus allowing for more precise and deeper topic extraction.

The research gap for this study is the limited investigation of the optimization strategies that are integrated with both traditional and advanced topic modeling techniques for biomedical abstracts. In general, the existing studies have been

applying these models individually or without strong optimization methods; they have left a lot of space for improvement in terms of coherence and accuracy. This work is a significant step towards filling a crucial gap in research by not only formulating but also investigating optimization strategies that are combined with both conventional and advanced techniques of topic modeling for biomedical abstracts. The main contribution of this work is a new approach to the integration of four topic modeling techniques-LDA, LSA, LSTM, and RNN-with two optimization algorithms, that is, grid search and Bayesian optimization, for better topic extraction and performance evaluation.

The key contribution of the proposed work includes:

- The performance of four topic modeling techniques (LDA, LSA, LSTM, and RNN) on biomedical abstracts is exhaustively analyzed.
- The integration of grid search and Bayesian optimization algorithms for enhancing topic extraction and achieving better coherence scores.
- A comparative analysis of the results will provide insight into the efficacy of each model and optimization method.

These research experiments have demonstrated promising outcomes, particularly emphasizing the effectiveness of the various model categories in this context. The paper structure is outlined below: Section 2 offers an overview of the existing literature. Section 3 outlines the methodology being proposed. Section 4 examines the findings, providing both comparative analysis and detailed discussions. Section 5 provides the paper's conclusion, exploring potential future work.

2. Related Works

The advancement of topic modeling strategies in biomedicine and pharmaceuticals has been a major factor in deepening the extraction of insightful information from large volumes of unstructured text data. In this section, the current research that has been instrumental in the development and application of different topic modeling approaches in the medical field is discussed. Understanding the challenges, techniques, and results of these related works, the objective is to position the proposed study in the context of the analysis of the medical literature and to emphasize the sophisticated facets of the proposed approach. The following points highlight the essential innovations from the research, which serve as a basis for further work and allow us to delve deeper into the potential of advanced topic extraction models in the field of medical PubMed abstracts. Junaid Rashid et al. [13] proposed a new way of topic modeling for biomedical text documents. The authors discussed the difficulty of getting the information from different and sometimes unstructured texts of biomedicine sources, such as electronic health records, scientific papers, and case summaries. The authors devised a new method that combines hybrid inverse document frequency with fuzzy K-means clustering.

Stefano Sbalchiero et al. [14] focused their attention on model fitting as a main theme in topic modeling research, especially the extent to which their methods could be applied to long texts. They explored the relationship between the length of the text and the number of topics, as well as the introduction of a mathematical formula to represent this relationship. Through experiments with Latent Dirichlet Allocation, they analysed the impact of varying topic numbers on the outcomes of corpus analysis and introduced a novel methodology tailored for analysing long texts.

Sandhya Avasthi et al. [15] provided a comprehensive exploration of topic modeling in the context of biomedical literature. It examines cutting-edge methodologies such as Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM), Hierarchical Dirichlet Process (HDP), Dirichlet Multinomial Regression (DMR), and the Hierarchical Pachinko Allocation (HPA) models. The paper focuses on text mining from PubMed articles related to adolescent substance use and depression. Using these models, a comparative analysis is made based on log-likelihood, perplexity, and topic coherence measures.

Tao Chen, Mingfen Wu, and Hexi Li [16] focused on enhancing medical relation extraction using deep learning. It discusses the use of a pre-trained model, specifically BERT (Bidirectional Encoder Representations from Transformers), combined with a 1D Convolutional Neural Network for fine-tuning. This approach is tested on three datasets: BioCreative V CDR task, TCM literature, and 2012 i2b2 task for the extraction of temporal relationships in clinical settings, which shows significant improvements in F1 score. This research contributes to the advancement of NLP techniques in biomedical applications.

Abhyuday N et al. [17] investigated the application of Bidirectional Recurrent Neural Networks (BRNNs) in identifying medical events within Digital Health Records. It compares the performance of RNNs with traditional Conditional Random Fields (CRFs) and highlights the superiority of RNNs in learning long-term dependencies in data. The authors revealed the difficulties of sequence labeling in EHRs that are made up of noisy and unstructured texts and showed how efficiently RNNs can be used to solve these problems. The work highlighted the pivotal role that cutting-edge machine learning techniques play in medical informatics and pharmacovigilance. Nizar A. Ahmed et al. [18] focused on improving multiclass classification of biomedical documents related to cardiovascular disease. For representing the features, they used a mix of Bag of Words (BoW) and Word Embeddings (WE) with several statistical weighting schemes. The study makes use of Bidirectional Long Short-Term Memory (BLSTM) in DNNs and evaluates its

performance on the MIMIC III and PubMed datasets. Karami, A. et al., [19] explored the problem of sifting through textual medical documents and electronic health records to find the most relevant pieces of information. They came up with a brand-new topic modeling technique called Fuzzy Latent Semantic Analysis (FLSA), which is geared towards making the automatic extraction of themes from health and medical corpora more efficient. FLSA confronts the redundancies that have been identified in these corpora and offers a totally new way for figuring out the number of topics.

Jelodar, H. et al., [20] focused their work on topic modeling, a crucial area of text mining that aims to find new data and relationships in the textual documents. The authors emphasized the significance of LDA, which is a generally referred technique in topic modeling, and investigated a variety of models derived from LDA. The study serves as an exhaustive survey of research articles from 2003 to 2016 that concentrate on topic modeling via LDA, analyzing the development of research, the trends, and the conceptual framework. In summary, the current studies highlight major improvements in biomedical text analysis using topic modeling and deep learning. Nevertheless, there is a considerable difference that has been left unacknowledged in the integration of optimization algorithms with both conventional and advanced modeling techniques. This research goes beyond by bridging the gap through a comprehensive assessment of biomedical abstracts-oriented optimized topic modeling methods, therefore, being a source of new knowledge on their usefulness and accuracy in practice.

3. Materials and Methods

The section details the approach that was to be used, including a deep dive into four sophisticated models for topic extraction, and Figure 1 describes the methodology flow. The experimental research proceeded with the acquisition of a set of medical PubMed abstracts that were fetched through a tailored query and recorded in a CSV format. The data set was divided into three different sentiments, namely positive, negative, and neutral. The study has concentrated on the category of negativity, which means those texts that express a negative sentiment or a negative tone and consist of 107 abstracts, as shown in Table 1 below.

A validation split of 20% was utilized, allocating 80% of the dataset for training and 20% for validation purposes. The dataset was split randomly by Keras during the training call to provide unbiased validation metrics for deep learning models. The methodology proposed is divided into four key steps, and each is described in subsequent sections. Four topic modeling techniques-LDA, LSA, LSTM, and RNN-were selected based on their prominence in the literature and their capability in handling complex language structures. Each model was implemented using Python libraries such as Gensim, scikit for LDA and LSA, and Keras for LSTM and RNN. Hyperparameters for each model were rigorously optimized via grid search and Bayesian optimization. The rationale for selecting hyperparameters such as the number of topics is 10, learning rates, batch sizes, epochs, and neuron counts involved maximizing coherence and minimizing perplexity scores on the validation set.

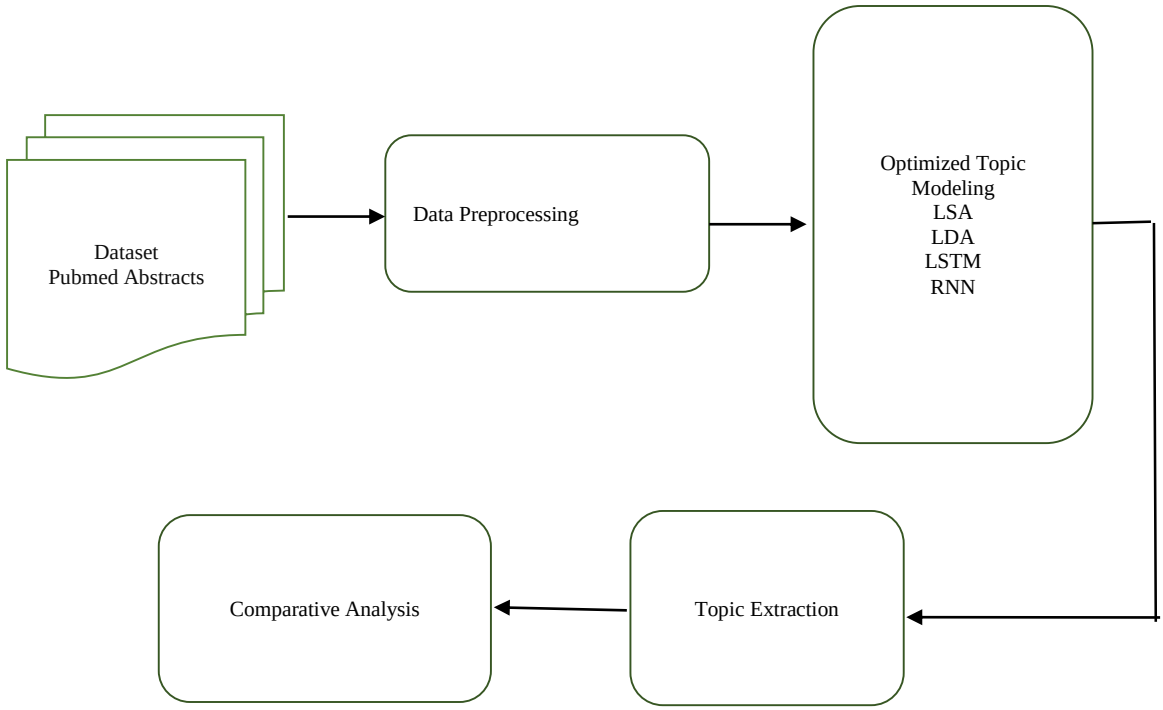


Fig. 1 Proposed methodology flow

Table 1. Data set

S. No	Id	Abstract
0	1	Small interfering RNA (siRNA) for silencing gene...
1	182	There are limited data on cardiovascular efficacy...
2	236	Crushed oral tablets, when injected intravenously...
...	...	...
105	6312	Complex diseases are associated with a wide ra...
106	6399	Cardiac arrest in Wolff-Parkinson-White (WPW) ...

3.1. Data Preprocessing

The significance of data preprocessing is the foundational step, ensuring that topic modeling techniques effectively identify the underlying topics in a dataset. In this research, data preprocessing involves several crucial steps that aim at refining the data, raising its quality, saving time for further steps, and improving the results of the model. For this study, a specific preprocessing method is adopted as depicted in Figure 2.

Tokenization, breaking down articles or sentences into smaller elements like words or tokens. It was followed by the Stop Word and Punctuation Removal process, which aimed to eliminate common words that typically do not contribute significant meaning to the text (e.g., "and,""is,""the"), along with the removal of special characters from the dataset. Lastly, the Data Unicode process is adopted to normalize the text, which ensures that the data is consistent and ready for further analysis.

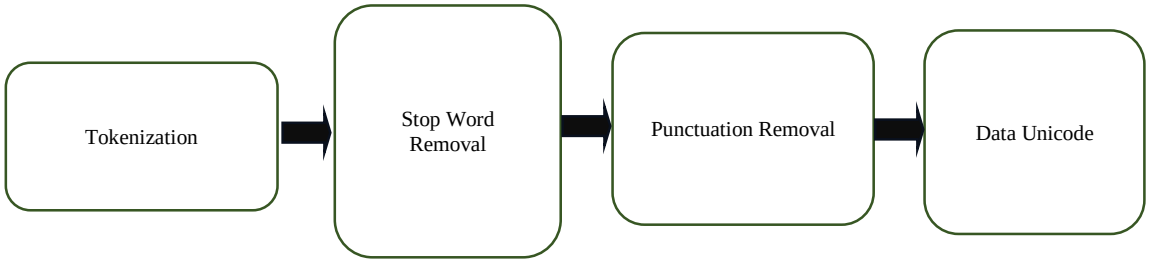


Fig. 2 Data preprocessing

3.2. Topic Modeling

Topic modeling involves identifying topics within a document corpus. Unlike clustering, which assigns documents to topics, topic modeling generates a word composition for each Topic based on the documents [21]. It refines each document's topics using this composition of words. This study delves into two kinds of topic modeling: machine learning and deep learning, as shown in Figure 3.

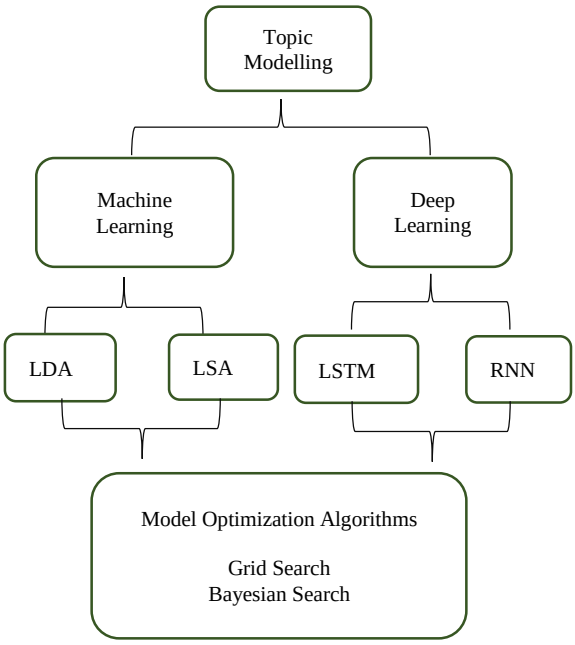


Fig. 3 Different types of proposed topic modeling

A well-known machine learning approach, like LDA and LSA, is utilized, which is enhanced with specialized optimization algorithms. Likewise, for deep learning, LSTM and RNN models are used for topic extraction from pre-processed PubMed abstracts. Respectively, these four models are optimized by Grid search and Bayesian search optimization algorithms to ensure they are fine-tuned, enhancing their topic modeling abilities and aligning them precisely with the specific data and requirements of tasks.

Grid search is a commonly used method that systematically observes all possible combinations of hyperparameters within a predefined grid. This method is often defined as exhaustive or brute-force, and it assesses every potential hyperparameter configuration stated in the grid. Bayesian optimization is a well-organized method for optimizing complex functions that are costly to evaluate and have multiple boundaries.

It operates by creating and constantly updating a probabilistic model of the objective function, which is based on current evaluations [22, 23]. This model is utilized to pick the next most promising points for assessment by optimization of an acquisition function. These optimization methods are at the heart of how complex machine learning and deep learning models can be used to their full capacity.

3.2.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is a popular technique in topic modeling. It functions based on the concept that each document consists of a blend of different topics and every Topic is linked to specific words. The objective is to determine the topics a document covers based on the words in the content. It is predicated on the idea that document sharing in topics will likely use a related set of words.

Therefore, LDA assigns documents to a probability distribution across latent topics, with each Topic being a probability distribution of words [24, 25]. It serves a dual purpose in text analysis, as it identifies distinct topics from a given corpus, and it gives discovered topics to the corpus in the documents simultaneously. This is effectively summarized in the provided schematic diagram in Figure 4, which illustrates the LDA process [26].

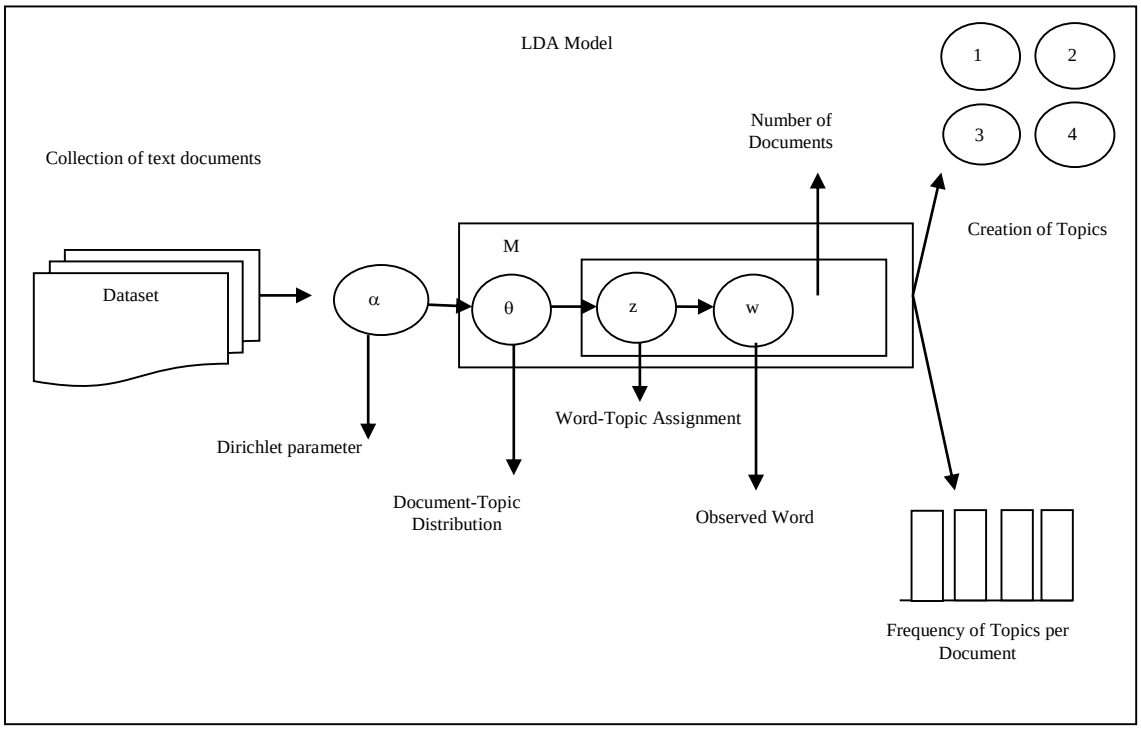


Fig. 4 LDA process

The LDA model generative process for each PubMed abstract ai ∈ A, which can be outlined as follows.

A. Generate parameters for every Topic $\beta_k \sim$ Dirichlet (φ), for k ∈ {1…K}

B. In relation to all abstracts:

1. Choose the distribution of Topic: $\theta_m \sim$ Dirichlet (α)
2. For every one of the N words w_n:
   i. Choose a topic $z_{mn} \sim$ Multinomial (θ_m)
   ii. Choose a word $w_n \sim$ Multinomial (β_k) based on p ($w_n|z_{mn},\beta_k$)

where β is the parameter at the dataset level, and α is sampled once during dataset generation. K represents the

count of topics. For each word in every abstract, word-level variables z_{mn} and w_{mn} are allocated once, whereas the dataset-level variable θm is tested once for all abstracts. φk denotes the word probability distribution for topic k. The LDA algorithm time complexity is $O(N_{itr}KN\bar{I})$, where N_{itr} represents the iteration count, K denotes the topic quantity, N signifies the overall count of abstracts, and $\bar{I}$ indicates the average no of abstracts in the dataset. Algorithm 1 outlines a structured approach to optimizing a topic modeling process using LDA while evaluating the quality of the topics through coherence scores.

Algorithm 1: Optimized Latent Dirichlet Allocation Topic Modeling with Coherence Evaluation

Input: A set of PubMed abstracts A, a range of topics K_range=10, and the number of iterations N_itr
Output: Optimized final model with topic assignments for each abstract and coherence score

1. Begin
2. Initialize grid search with a range of hyperparameters for K within K_range.
3. Initialize Bayesian optimization with a prior distribution over the hyperparameters.
4. For each configuration of K in the hyperparameter space:
   a. Implement Bayesian optimization to adjust hyperparameters based on the earlier iterations' performance.
   b. Assume K topics for the corpus.
   c. For every abstract m in the corpus A:
      i. Assign every word in the abstracts arbitrarily to one of the K topics.
   d. For every abstract m:
      i. For every word w in abstracts m:
         - Compute $pd(w|t_k)$, word probability w conditional on topic t_k.
         - Compute $pd(t_k|a_i)$, topic probability t_k conditional on abstracts a_i.
      ii. Update the probability of Topic for word w in abstracts a_i:
         - $pd(w|t_k, a_i) = pd(w|t_k) * pd(t_k|a_i)$.
   e. Loop over every word in all abstracts:
      i. Reassign the current word topic based on $pd(w|t_k, a_i)$.
   f. Repeat steps a and e until N iterations are complete.
5. After the iterations, calculate the coherence value for the LDA model to evaluate topic quality.
6. Assess the LDA model effectiveness for each K by measuring the coherence score.
7. Choose the model that has the highest coherence score as the ideal model.
8. Consolidate the results to form the final optimized model with topic distributions.
9. End

3.2.2. Latent Semantic Analysis (LSA)

Latent Semantic Analysis is a method aimed at uncovering topics in abstracts through transforming their textual content into matrices of word-topic and document-topic relationships. Fundamentally, it seeks to break down a matrix of document-terms into two separate matrices, such as one connecting document to topics and another linking topics to terms [4]. Pubmed Abstracts consist of a collection of texts that must be converted into a machine-understandable format, such as numerical data. The transformation of documents into numerical vectors is called the document term matrix, which is accomplished using methods like count vectorizer, TF-IDF, Word2Vec, and others. The discovery of latent topics through LSA encompasses applying Singular Value Decomposition (SVD), which is a method for matrix factorization, to the document-term matrix used to decrease its dimensionality based on features of latent topics. [27] The LSA process is illustrated in the diagram below 5.

The initial step involves producing a document-term matrix: For *m* abstracts and *n* words in the vocabulary, an *m*×*n* matrix *A* is constructed, where rows are dedicated to abstracts and columns denote individual words. The basic implementation of LSA is to allow the model to count the raw frequency of word occurrences, where the y^{th} word in the x^{th} document for each matrix entry is calculated. Normally, the raw counts are insufficient because they overlook the contextual importance of each word in a document. Hence, LSA models commonly employ TF-IDF scores instead of raw counts in the Document-Term Matrix (DTM). TF-IDF stands for term frequency-inverse document frequency, which calculates a weight for term *y* within document *x* in the following way:

$$wd_{x,y} = tc_{x,y} X \log\left(\frac{N}{af_x}\right) \quad (1)$$

Where $tc_{x,y}$ signifies the count of times *x* appears in *y*, af_x indicates the count of abstracts containing *x*, and *N* is the total count of abstracts. Utilize truncated SVD to lower the dimensionality of *X*, breaking it down into three matrices [28], namely matrix *U*, *S*, and V^T (the transpose of matrix *V*).

$$A=USV^T \quad (2)$$

$$U \in R^{(i\ x\ k)}, V \in R^{(j\ x\ k)} \quad (3)$$

Where *A* is the primary matrix with dimensions *i*×*j*, with *i* indicating Terms and *j* denoting the abstracts. *k* represents the total number of topics. *U* is a matrix illustrating the document vectors within abstracts, sized *i*×*k*, *V* is a matrix reflecting the topic vectors, with dimensions *j*×*k*, and *S* is a diagonal matrix sized *k*×*k*. Algorithm 2 below outlines an optimized approach to Latent Semantic Analysis (LSA) for topic modeling.

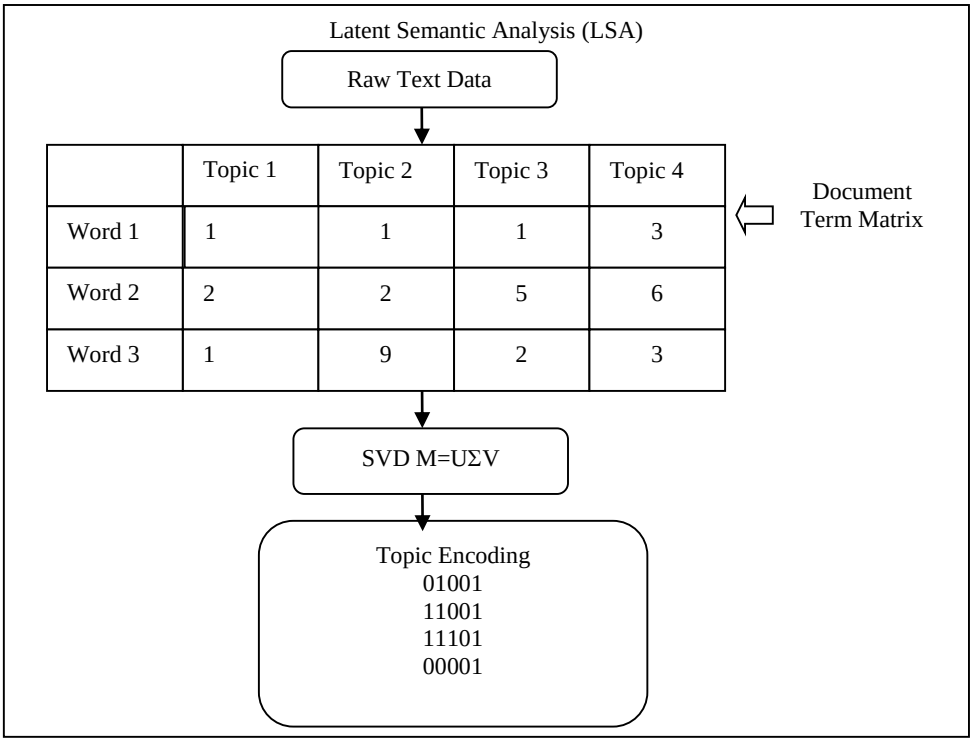


Fig. 5 Process of LSA

Algorithm 2: Optimized Latent Semantic Analysis Topic Modeling with Coherence Evaluation

Input: A set of PubMed abstracts A, a range of potential topic numbers K_range=10, and the number of iterations N itr
Output: Optimized final model with singular value decomposition components and coherence score

1: Begin
2: Initialize grid search with a variety of hyperparameters for the quantity of topics K within K_range.
3: Initialize Bayesian optimization with a prior distribution over the hyperparameters.
4: For each configuration of K in the hyperparameter space:
   a. Perform Bayesian optimization to refine hyperparameters based on previous iterations' performance.
   b. Construct a term-document matrix from the corpus A.
   c. Utilize SVD to split the term-document matrix into K separate topics.
5: Calculate the document-topic and topic-term matrices.
6: For each K, calculate the coherence value to evaluate the semantic similarity of top words in each Topic.
7: Measure the LSA model performance for every K value by the coherence score.
8: Choose the model that achieves the highest coherence score as the best model.
9: Compile the findings into the best optimized model with topic distributions.
10: End

3.2.3. Long short-term memory (LSTM)

Long Short-Term Memory (LSTM) networks are an advanced model of Recurrent Neural Networks, which excel in identifying and retaining long-term associations in a sequence of data. LSTM analyzes and processes textual data to extract and identify the underlying themes or topics within a corpus of documents. LSTMs for topic modeling work by processing text data sequence-by-sequence, preserving information from previous inputs through their memory cells. [29] This allows them to maintain a context and understand the relevance of words and phrases within the document text, which is crucial for accurate topic detection.

Diagram 6 shows the proposed LSTM model architecture to handle text data, mainly for the topic extraction task. At first, the model gets the data at the Input Layer. After that, the data goes to the Embedding Layer, where words are converted into vector representations that are understandable to the machine. The next GRU Layer, being a variant of RNN, further handles the embeddings and also gets the temporal relationships in the data. After that, a Max Pooling layer is there to extract the most important features for each segment of the sequence from the information that it has condensed. Most importantly, the design of the system features two types of optimization techniques: Grid Search and Bayesian Optimization.

Grid Search is a hyperparameter optimization technique that exhaustively tries every combination of the hyperparameters to find the one that leads to the best result of

the model [30]. On the other hand, Bayesian Optimization uses a probabilistic approach, estimating the model performance for different hyperparameters and selecting only those that are most likely to improve the model for an actual experiment. The last layer of this network is Topic Extraction, where the features that are processed and optimized are utilized to recognize and extract the topics from the input text. If the topic extraction optimized model is further enhanced with Adam optimization and K-means clustering, then it would perform better and also have improved capabilities for post-processing. The first paragraphs of this file present an overview of Algorithm 3, which is a detailed description of the proposed LSTM Model usage for topic modeling and the primary aim of deriving topics from a text corpus. It is also about the text preprocessing, the training of the LSTM model with Adam optimization, and the post-processing with K-means clustering to identify and group topics.

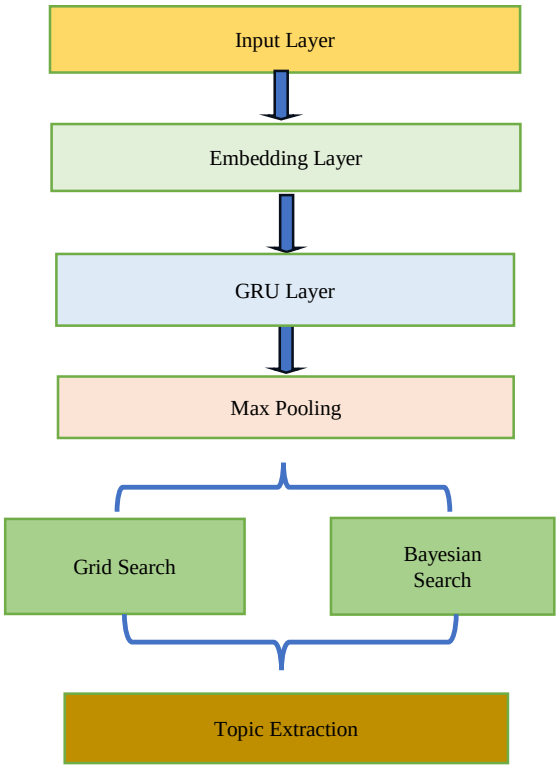


Fig. 6 Proposed LSTM model architecture

Algorithm 3: Proposed LSTM-based Topic Modeling for Topic Extraction

Input: A corpus of text documents D, Number of desired topics T=10, Number of LSTM training epochs E =50, Hyperparameters for the LSTM model H_params
Output: A set of T topics extracted from the corpus

1: Begin
2: Preprocess the text documents in D to convert words to numerical data suitable for the LSTM (e.g., via tokenization and embedding).
3: Initialize the LSTM model with hyperparameters H_params.
4: To train the LSTM model using Adam optimization to regulate weights:
   a. For every epoch ep from 1 to Ep:
      i. For every document dc in DC:
         - Input the preprocessed text data into the LSTM.
         - LSTM's memory cells can capture dependencies and contextual information in the text.
      ii. Revise the LSTM weights using the Adam optimizer.
5: After training, pass the document representations through the final layer to get topic probabilities or embeddings.
6: Apply a topic extraction method, such as clustering with K-means, to group the document embeddings into T topics.
7: For every cluster (Topic), recognize the most representative terms or phrases.
8: Assign topics to every document based on the next cluster in the embedding space.
9: Return the group of T topics, each with its connected group of representative terms and documents.
10: End

3.2.4. Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) offer a flexible way to model topics, giving them the power to handle data sequences. In comparison to the conventional methods, RNNs take into account the word order, which is especially helpful for grasping the context and meaning in language processing tasks [31]. Such sequential processing ability thus allows RNNs to be considered as a tool to get topics from texts, the word relationships in which play a major role in the general Topic. Figure 7 shows the RNN model architecture for topic extraction from text data, along with the processes of Adam optimization and K-means clustering.

The architecture starts with the Input Layer, where the model gets raw text data. The data is subsequently handed over to the Embedding Layer that links the words with the vectors from the high-dimensional space, thus it is able to grasp the semantic meaning of the words. After the embedding, the data goes through a Bi-Directional LSTM layer. [32] The cell is very effective in getting patterns from both the past and the future contexts of the data, which is necessary for understanding the sequence and the structure of the text. Once the LSTM is done, the Dense Layer acts like a fully connected neural network, merging features extracted by the LSTM layer into a new form, which is then usable for classification or regression tasks [33].

Grid Search is used to explore different hyperparameters to find the optimal model parameters systematically. At the same time, Bayesian Optimization offers a probabilistic

method for hyperparameter tuning by using previous evaluation results to forecast and, therefore, select the most suitable parameters for the model. Once the model is trained, Adam Optimization is used during the learning process. Finally, the features learned by the network are clustered using K-means. This unsupervised learning algorithm categorizes data points that are similar, which efficiently organizes comparable topics together, based on the attributes extracted from the text. This step is crucial for organizing the topics into coherent clusters, making the interpretation of the results more intuitive and actionable. Below is a detailed description of Algorithm 4, which details the usage of an Optimized RNN for topic modeling to identify topics within a text corpus. Each document is processed through the RNN to learn its structure and context. After training, the outputs (hidden states) are clustered to form topics, and representative terms are identified for each Topic.

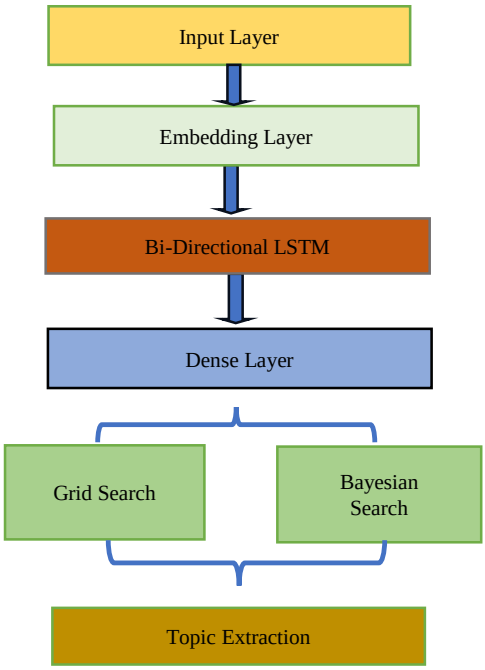


Fig. 7 Proposed RNN model architecture

Algorithm 4: Optimized RNN-based Topic Modeling

Input: A corpus of abstract text A, No of topics T=10, Count of RNN iterations N_iter=50, Hyperparameters to the RNN model H_params

Output: A set of T topics extracted from the corpus

1: Begin
2: Preprocess the text in Abstracts A to convert words to numerical data (e.g., word embeddings).
3: Initialize the RNN model with hyperparameters H_params.
4: For each document d in A:
   a. Sequentially input the preprocessed text data into the RNN.
   b. Use the hidden states of the RNN to learn the sequential patterns in the text.
5: After N_iter iterations, use the final hidden states of the RNN to represent the Abstracts.
6: Implement a clustering algorithm (e.g., k-means) upon the abstract representations to group them into T topics.
7: For each cluster representing a topic:
   a. Identify and record the most representative terms.
8: Return the set of T topics, each with its corresponding set of representative terms.
9: End

3.3. Topic Extraction

Topic extraction identifies and extracts underlying themes or topics from PubMed abstracts. This process involves analyzing the words and phrases within the abstract to uncover patterns of topics that frequently occur together. This study implemented the four topic extraction models, such as LDA, LSA, LSTM, and RNN, and the extracted results are shown in Table 2. Each model reveals different facets of the dataset and focuses on various aspects related to cardiac disease.

4. Results and Discussion

The study conducted experiments to evaluate the four refined models-LDA, LSA, LSTM, and RNN-on the PubMed medical abstract dataset. The dataset was selected because it represents a diverse set of biomedical research articles from different periods, journals, conferences, and various subfields of biomedicine. Essentially, the PubMed dataset is a comprehensive database comprising abstracts of medical journals, conference papers, and other biomedical writings. In general, it covers almost all areas of the biomedical literature, thus providing a vast amount of information on topics such as clinical trials, disease mechanisms, and new drug developments. As such, it is extensively used by pharmacovigilance researchers. The dataset enables them to collect evidence on the efficacy of drugs, their side effects, and safety in general, which, in turn, leads to the advancement of public health. To corroborate the findings, a qualitative study was also done on the selected abstracts, checking the quality and usefulness of the extracted topics. According to the analysis, the topics that LDA with Bayesian optimization extracted were more domain-relevant and coherent, matching accepted medical terminology and concepts.

The models were gauged and compared based on a metric known as topic coherence value, which serves as a yardstick for determining the ideal number of topics within a model [34]. Topic coherence measures how human interpretation aligns with the number of topics generated, providing insights into the optimal topic quantity through coherence scores. Specifically, the CV method is employed, which calculates coherence based on the Normalized Pointwise Mutual Information (NPMI) of the maximum relevant words within a sliding window framework.:

Table 2. Topic extraction results for each model

LDA	LSA
Topic: 0, ("cardiac", "treatment", "clinical", "heart", "ventricular", "drugs", "common", "risk") Topic: 1, ("cardiac", "patients", "study", "drugs", "myocardial", "fibrosis", "risk", "drug", "stroke", "heart")…… Topic: 9, ("cardiac", "heart", "drugs", "cardiovascular", "patients", "clinical", "associated", "effects")	Topic 0, ("cardiac", "drug", "patient", "heart", *"study", "disease", "effect", "fibrosis", "failure", "treatment") Topic 1, ("cardiac", "patient", "fibrosis", "drug", "study", "risk", "ci", "clinical", "stroke", "expression")…… Topic 9, ("lqt", "study", "disease", "heart", "cardiac", "rdn", "treatment", "therapy", "drug", "group")
LSTM	**RNN**
Cluster 1 top words: ['inhibitors', 'ktr', 'heart', 'failure', 'diabetes', 'disease', 'mellitus', 'sglt2', '2', 'ptdm'] Cluster 2 top words: ['cardiovascular', 'drugs', 'copd', 'disease', 'new', 'vascular', 'clinical', 'use', 'molecular', 'studies'] …… Cluster 10 top words: ['patients', '95', 'ci', 'i2', 'ric', 'bmscs', 'lenvatinib', 'studies', 'a1bs', 'development']	Cluster 1 top words: ['cardiac', 'nps', 'effect', 'potential', 'alm', 'heart', 'drug', 'mortality', 'receptor', 'studies'] Cluster 2 top words: ['cardiac', 'drugs', 'heart', 'fibrosis', 'patients', 'cardiovascular', 'clinical', 'ric', 'hfpef', 'mechanisms'] …… Cluster 10 top words: ['cardiac', 'heart', 'fibrosis', 'drugs', 'disease', 'treatment', 'failure', 'diseases', 'models', 'growth']

$$NPMI(wd_i, wd_j) = \sum_{j}^{N-1} \frac{log\frac{PB(wd_i,wd_j)}{P(wd_i),P(wd_j)}}{-\log PB(wd_i,wd_j)} \quad (4)$$

In the given context, $PB\ (wd_i)$ represents the likelihood of the word wd_i randomly occurring in the abstracts. At the same time, $PB\ (wd_i,wd_j)$ signifies the joint probability of each word wd_i and wd_j occurring together in abstracts by chance. The term N signifies the cumulative count of distinct choices of words available within the collection 1,2,..., $w1$,$w2$,...,wn.

Figure 8 shows the distribution of documents in terms of word count for the entire dataset. Figure 9 is a word cloud for "Topic 1," which visually represents the frequency of words within a specific topic. Larger words, such as "cardiac,""expression,"" model," and "heart," suggest these are key terms within the Topic, indicating a likely focus on cardiac expression models of heart conditions. The presence of words like "fibrosis,""cells,""tissue," and "stress" implies a medical or biological context, possibly relating to research in cardiac diseases, their mechanisms, and the effects of stress on the heart.

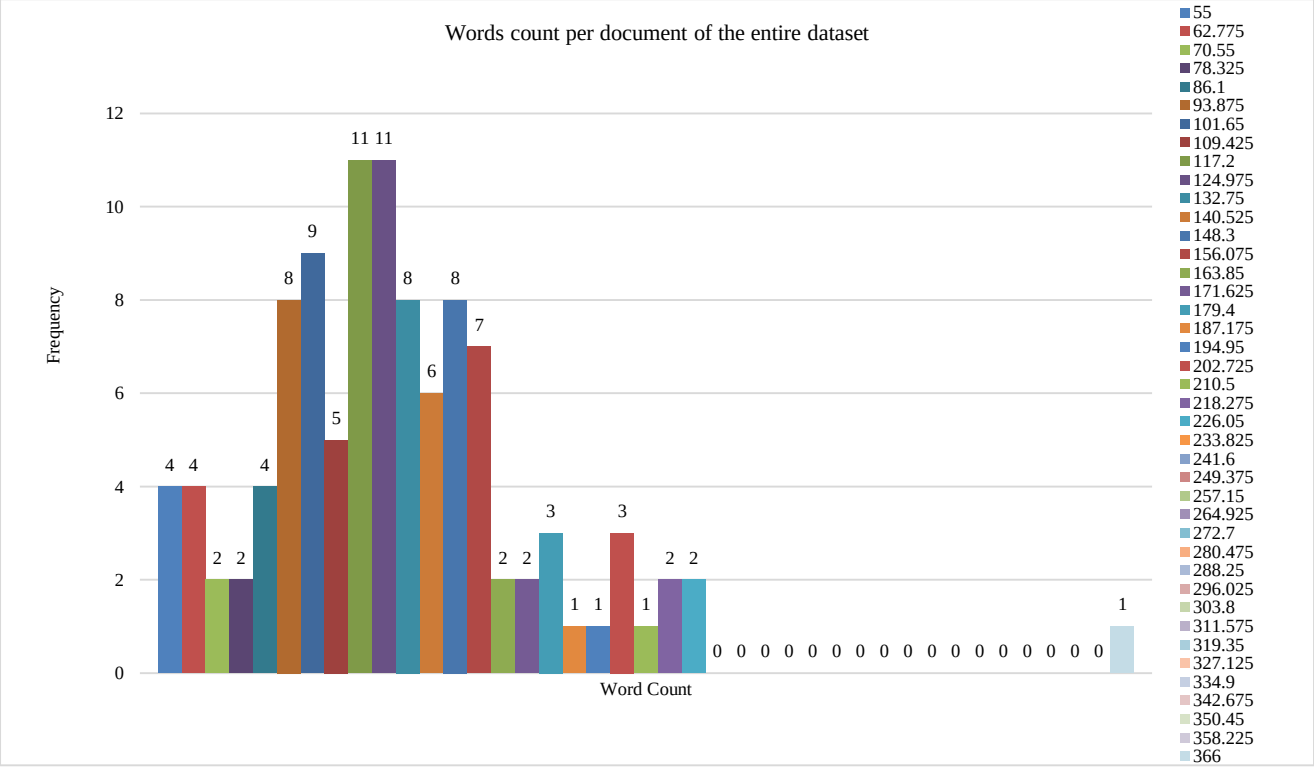
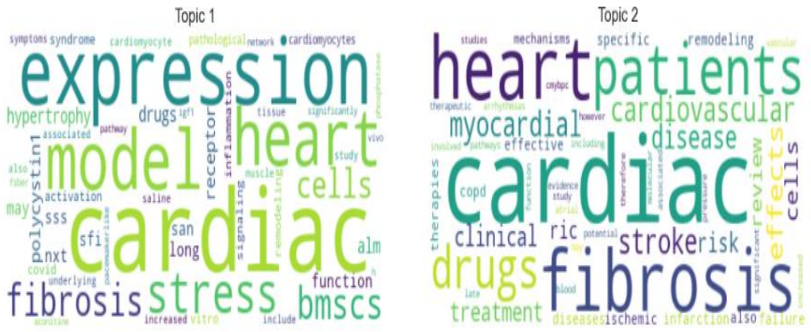


Fig. 8 Histogram of word count per document of the entire dataset



(a) Topic 1 **(b) Topic 2**

Fig. 9 Word cloud representation of LDA for topics 1 and 2

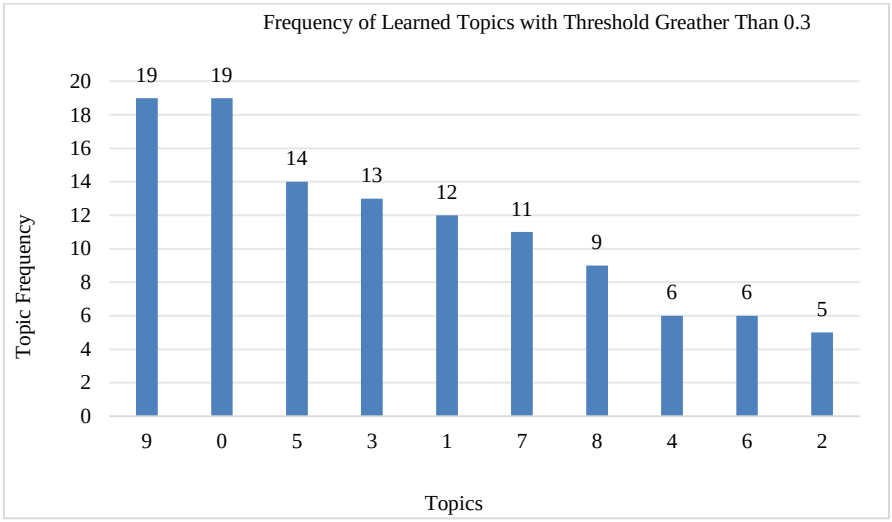


Fig. 10 Topic frequency of LDA

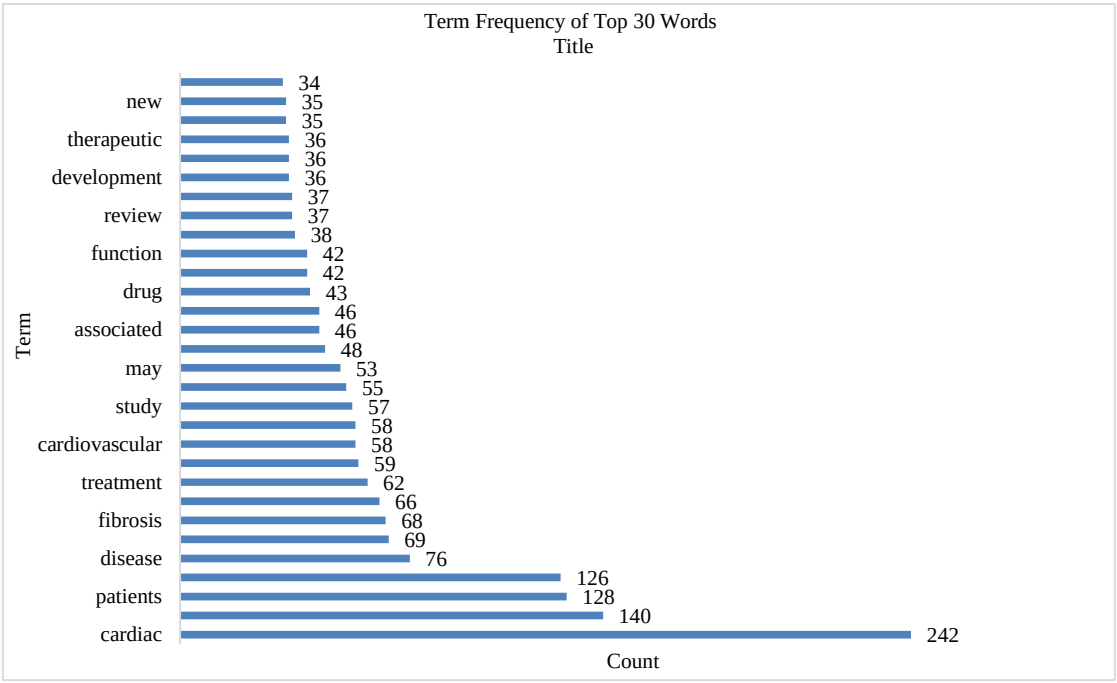
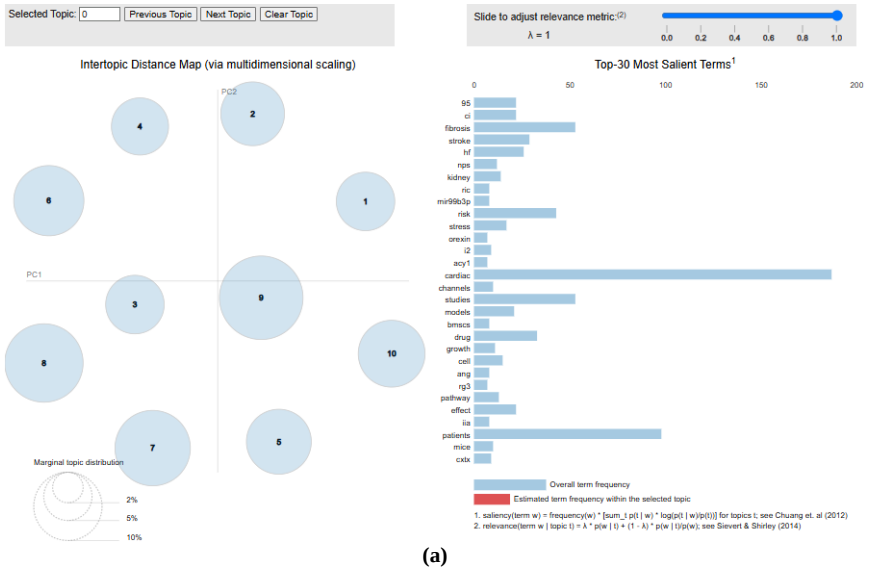


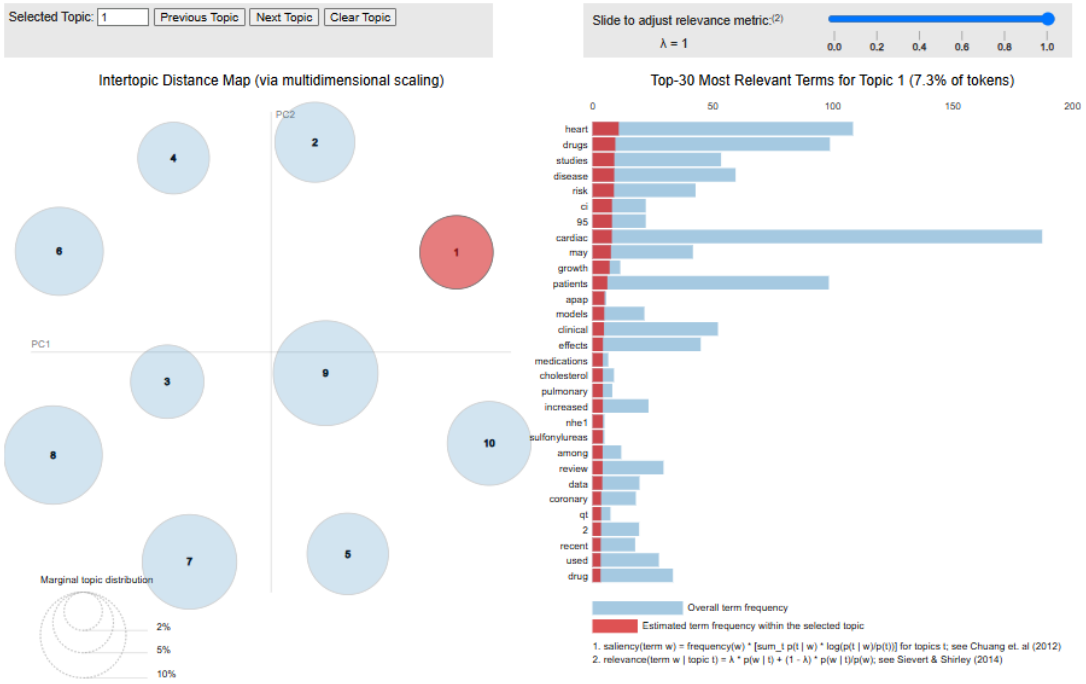
Fig. 11 Topic frequency of LDA

The bar chart in Figure 10 depicts the frequency of learned topics from a dataset where the threshold for topic inclusion is set above 0.3. Each bar represents a different topic, and the height indicates the frequency count of that Topic within the dataset. The chart shows a descending frequency, indicating that some topics appear more frequently than others. Figure 11 presents the top 30 word frequencies, where “cardiac”, “patients”, “heart”, and “disease” were the most frequent words. The word “cardiac” is the one that appears the most, suggesting that the dataset is primarily related to the cardiovascular system.

Figure 12 shows the pyLDAVi visualization of the topic modeling carried out by LDA. (a) The bubbles on the left depict the distribution of topics within the dataset, while the bars of sky-blue color on the right indicate the occurrence rate of terms throughout the corpus. This visualization is tailored for electronic viewing, where a more detailed view can be obtained by zooming in. (b) The left side of the figure has a red bubble indicating Topic 1, which is the main focus. Hence, the red bars on the right show the approximate occurrences of the top 30 most frequent keywords composing the first Topic.



(a)



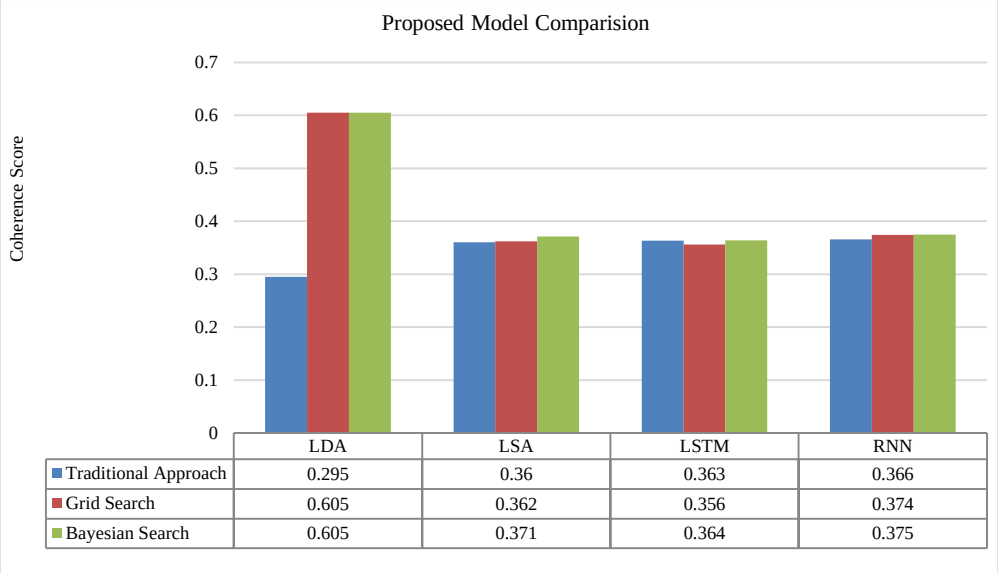
(b)

Fig. 12 Visualization using pyLDAViz visualizing the topic modeling performed through LDA: (a) Overall topic distribution, and (b) Selected topic distribution.

The chart provided in Figure 13 compares the coherence values of four different models-LDA, LSTM, RNN, and LSA-each of which was trained using three different methods. The traditional method, grid search optimization, and Bayesian optimization were evaluated for their effectiveness in improving topic modeling. From the improvements in the LDA model alone, it can be seen that optimization techniques, in particular grid search and Bayesian optimization, have a substantial effect on the performance of the LDA model. The LSTM model only shows a slight improvement with Bayesian optimization, and the RNN and LSA models are also improved by both optimization methods, but to a lesser degree. The comparison here is a perfect example of how machine learning advanced optimization strategies can be used to fine-tune model accuracy and topic coherence. Topic diversity is the basis for ensuring that the themes are distinct and comprehensive within the PubMed abstract dataset. High topic diversity indicates that the topics cover a wide range of themes or subjects without much overlap, while low topic diversity suggests that the topics are more similar and may overlap significantly in terms of the words they contain. It is calculated by the following equation.

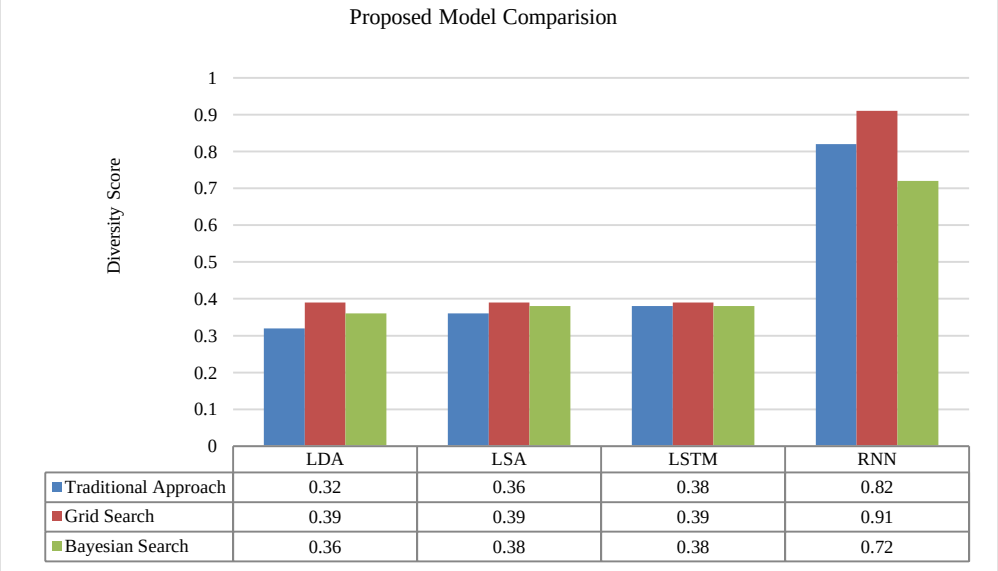
$$TopicDiversity(d) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} Dist(t_i, t_j) \quad (5)$$

Where n is the total number of topics, t_i,t_j represent topics i and j respectively, and Dist(t_i,t_j)) is used to calculate the distance between topics t_i and t_j. Figure 14 shows the topic diversity score for each model.



	LDA	LSA	LSTM	RNN
Traditional Approach	0.295	0.36	0.363	0.366
Grid Search	0.605	0.362	0.356	0.374
Bayesian Search	0.605	0.371	0.364	0.375

Fig.13 Proposed model comparison based on coherence value



	LDA	LSA	LSTM	RNN
Traditional Approach	0.32	0.36	0.38	0.82
Grid Search	0.39	0.39	0.39	0.91
Bayesian Search	0.36	0.38	0.38	0.72

Fig. 14 Proposed model comparison based on diversity value

In Table 3, the comparative analysis of the proposed topic modeling technique with existing research is provided. It highlights different studies of LDA-based topic modeling techniques on various datasets, measuring their effectiveness through coherence value. The proposed Enhanced LDA Model achieved the highest coherence value of 0.605 and demonstrates its superior ability to extract meaningful topics from PubMed abstracts when compared to other models.

Table 3. Comparison of the proposed model with existing research

Paper References	Dataset	Coherence Value
Muhammad Inaam et al [35], 2023	Healthcare article	0.588
Amna Meddeb et al [36], 2022	Twitter dataset	0.396
Rahul Kumar Gupta [37], 2022	Published Articles	0.483
Sivanandham et al. [38], 2021	Published Articles	0.365
Proposed Study - Enhanced LDA Model	Pubmed Articles	0.605

The proposed system findings have a great deal of practical influence, especially in the area of pharmacovigilance, where it is very important to have precise and fast topic extraction from a large volume of biomedical texts. As a result, the improved topic modeling techniques disclosed in this study may be used to a great extent to render drug safety supervision more effective and efficient, thus giving the earliest possible and situationally appropriate information about the occurrence of side effects.

5. Conclusion

This research is a successful example of the utilization of state-of-the-art topic modeling methods on the PubMed medical abstracts dataset. Through experimentation, optimization algorithms such as Grid Search and Bayesian Optimization were used to a great extent to improve the performance of Latent Dirichlet Allocation (LDA) models. The experimental results serve as a testament to the capabilities of machine learning and deep learning methods in parsing vast medical literature to fetch relevant topics for more detailed and contextually rich analysis.

The precision of adverse drug reaction identification, as well as the speed, can be enormously improved by the use of advanced topic extraction techniques, which in turn have a direct impact on clinical decision-making processes and regulatory oversight in pharmacovigilance.

This study was limited by the relatively small dataset size and the complexity of tuning multiple hyperparameters across various models. Future research should consider larger datasets, other optimization algorithms, and extended testing in real-world pharmacovigilance applications.

References

[1] Mengqian Wang et al., "A Systematic Review of Automatic Text Summarization for Biomedical Literature and EHRs," *Journal of the American Medical Informatics Association*, vol. 28, no. 10, pp. 2287-2297, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Usman Naseem et al., "Benchmarking for Biomedical Natural Language Processing Tasks with a Domain Specific ALBERT," *BMC Bioinformatics*, vol. 23, no. 1, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003. [Google Scholar]

[4] Scott Deerwester et al., "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990. [CrossRef] [Google Scholar] [Publisher Link]

[5] Thomas Hofmann, "Probabilistic Latent Semantic Indexing," *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, United States, pp. 50-57, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[6] Daniel D. Lee, and H. Sebastian Seung, "Algorithms for Non-Negative Matrix Factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 1-7, 2000. [Google Scholar] [Publisher Link]

[7] Chun Yen Lee, and Yi-Ping Phoebe Chen, "Prediction of Drug Adverse Events using Deep Learning in Pharmaceutical Discovery," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1884-1901, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yankang Jing et al., "Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era," *The American Association of Pharmaceutical Scientists*, vol. 20, no. 3, pp. 1-22, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yuan Luo, "Recurrent Neural Networks for Classifying Relations in Clinical Notes," *Journal of Biomedical Informatics*, vol. 72, pp. 85-95, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] Jiebin Chu et al., "Using Neural Attention Networks to Detect Adverse Medical Events from Electronic Health Records," *Journal of Biomedical Informatics*, vol. 87, pp. 118-130, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] M. Schuster, and K.K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997. [CrossRef] [Google Scholar] [Publisher Link]

[12] Tongxuan Zhang et al., "Adverse Drug Reaction Detection Via a Multihop Self-Attention Mechanism," *BMC Bioinformatics*, vol. 20, no. 1, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Junaid Rashid et al., "Topic Modeling Technique for Text Mining Over Biomedical Text Corpora through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering," *IEEE Access*, vol. 7, pp. 146070-146080, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Stefano Sbalchiero, and Maciej Eder, "Topic Modeling Long Texts and the Best Number of Topics. Some Problems and Solutions," *Quality and Quantity*, vol. 54, no. 4, pp. 1095-1108, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Sandhya Avasthi, Ritu Chauhan, and Debi Prasanna Acharjya, "Topic Modeling Techniques for Text Mining Over a Large-Scale Scientific and Biomedical Text Corpus," *International Journal of Ambient Computing and Intelligence*, vol. 13, no. 1, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Tao Chen, Mingfen Wu, and Hexi Li, "A General Approach for Improving Deep Learning-based Medical Relation Extraction using a Pre-Trained Model and Fine-Tuning," *Journal of Biological Databases and Curation*, vol. 2019, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Abhyuday N. Jagannatha, and Hong Yu, "Bidirectional RNN for Medical Event Detection in Electronic Health Records," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, pp. 473-482, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[18] Nizar Ahmed, Fatih Dilmaç, and Adil Alpkocak, "Classification of Biomedical Texts for Cardiovascular Diseases with Deep Neural Network using a Weighted Feature Representation Method," *Healthcare*, vol. 8, no. 4, pp. 1-15, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Amir Karami et al., "Fuzzy Approach Topic Discovery in Health and Medical Corpora," *International Journal of Fuzzy Systems*, vol. 20, no. 4, pp. 1-12, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[20] Hamed Jelodar et al., "Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, A Survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 1-40, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[21] Pooja Kherwa, and Poonam Bansal, "Topic Modeling: A Comprehensive Review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, pp. 1-16, 2020. [Google Scholar]

[22] Antonio Candelieri, "A Gentle Introduction to Bayesian Optimization," *2021 Winter Simulation Conference (WSC)*, Phoenix, AZ, USA, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Xilu Wang et al., "Recent Advances in Bayesian Optimization," *ACM Computing Surveys*, vol. 55, no. 13S, pp. 1-36, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Stuart J. Blair, Yaxin Bi, and Maurice D. Mulvenna, "Aggregated Topic Models for Increasing Social Media Topic Coherence," *Applied Intelligence*, vol. 50, no. 1, pp. 138-156, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Daniel Maier et al., "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology," *Communication Methods and Measures*, vol. 12, no. 2-3, pp. 93-118, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Neha, Latent Dirichlet Allocation (LDA) and Topic Modeling using Gensim and Sklearn, 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/

[27] Harshit Agarwal, Topic Modelling Using LDA and LSA with Python Implementation, Enjoyalgorithms.Com, 2022. [Online]. Available: https://www.enjoyalgorithms.com/blog/topic-modelling-using-lda-lsa

[28] Ihsan Ahsanu Amala, Donni Richasdy, and Mahendra Dwifebri Purbolaksono, "Telkom University News Topic Modeling using Latent Semantic Analysis (LSA) Method on Online News Portal," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, pp. 110-115, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[29] Hamed Jelodar et al., "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or Covid-19 Online Discussions: NLP using LSTM Recurrent Neural Network Approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2733-2742, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ishaani Priyadarshini, and Chase Cotton, "A Novel LSTM-CNN Grid Search based Deep Neural Network for Sentiment Analysis," *The Journal of Supercomputing*, vol. 77, no. 12, pp. 13911-13932, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[31] Mohammad Ehsan Basiri et al., "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for Sentiment Analysis," *Future Generation Computer System*, vol. 115, pp. 279-294, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[32] Minyong Shi, Jingyi Huang, and Chunfang Li, "Entity Relationship Extraction based on BLSTM Model," *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, Beijing, China, pp. 266-269, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[33] Beakcheol Jang et al., "Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism," *Applied Sciences*, vol. 10, no. 17, pp. 1-14, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[34] Feng Yi, Bo Jiang, and Jianjun Wu, "Topic Modeling for Short Texts via Word Embedding and Document Correlation," *IEEE Access*, vol. 8, pp. 30692-30705, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[35] Muhammad Inaam ul haq, and Qianmu Li, "Revealing the Trends in the Academic Landscape of the Health Care System using Contextual Topic Modelling," *Data Intelligence*, vol. 5, no. 4, pp. 923-946, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] Amna Meddeba, and Lotfi Ben Romdhane, "Using Topic Modeling and Word Embedding for Topic Extraction in Twitter," *Procedia Computer Science*, vol. 207, pp. 790-799, 2022. [CrossRef] [Google *Scholar*] [Publisher Link]

[37] Rahul Kumar Gupta et al., "Prediction of Research Trends using LDA based Topic Modeling," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 298-304, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[38] S. Sivanandham et al., "Analysing Research Trends using Topic Modelling and Trend Prediction," *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP*, Springer, Singapore, vol 1325, pp. 157-166, 2021. [Google Scholar] [Publisher Link]