Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 30s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Improved Topic Modeling in Biomedical Texts Using UMLS: A MedMentions Approach

S. Jayabharathi[1], Dr. M. Logambal[2]

[1]Research Scholar, Department of Computer Science, Vellalar College for Women (Autonomous), Thindal, Erode, Tamil Nadu, India.

Email ID: jayabharathi8383@gmail.com

[1]Assistant Professor, Department of Computer Applications, K.S.Rangasamy College of Arts and Science (Autonomous), Tiruchengode, Tamil Nadu, India.

Email ID: jayabharathi8383@gmail.com

[2] Associate Professor, Department of Computer Science, Vellalar College for Women (Autonomous), Thindal, Erode, Tamil Nadu, India.

Email ID: m.logambal@vcw.ac.in

Cite this paper as: S.Jayabharathi, Dr. M.Logambal, (2025) Improved Topic Modeling in Biomedical Texts Using UMLS: A MedMentions Approach. *Journal of Neonatal Surgery*, 14 (30s), 887-900.

ABSTRACT

Effectively extracting and classifying topics from large volumes of medical texts is crucial for knowledge discovery and information retrieval in biomedical research. Traditional topic modeling techniques, while useful, often struggle to capture the intricate semantics of medical terminology. This research investigates how integrating Unified Medical Language System (UMLS) principles can enhance topic modeling using the MedMentions dataset**.** We evaluate four approaches**:** BERTopic, Latent Dirichlet Allocation (LDA), LDA with a Recurrent Network (LDA-RNet), and a novel BERTopic with a Recurrent Network (BERTopic-RNet)**.** Our goal is to improve topic coherence and relevance by incorporating UMLS concepts into these models. Experimental results demonstrate that UMLS-enhanced models significantly outperform conventional methods in both topic coherence and clinical relevance. This study provides valuable insights into the application of advanced topic modeling techniques in medical text analysis, paving the way for more effective and interpretable medical data mining.

Keywords: *Topic Modeling. UMLS Integration, Medical Text Analysis, BERTopic, Latent Dirichlet Allocation (LDA), Recurrent Network, MedMentions Dataset, Biomedical Text Mining, Clinical Relevance, Natural Language Processing (NLP)*

1. INTRODUCTION

The biomedical industry produces an enormous amount of textual data every day in the big data era, including clinical notes, research articles, and medical records. Effectively identifying significant subjects from this data is essential for improving patient care, medical research, and knowledge sharing. Textual data produced by the biomedical sector is vast and continuously expanding; examples include research articles, medication reports, clinical notes, and patient records. It is crucial to mine this data effectively in order to extract relevant information that will advance medical research, enhance patient care, and guide clinical decision-making. One of the most important tools in this effort is Topic Modeling, a text-mining technique that finds themes or subjects within a vast collection of texts. [1][2].

One of the most widely used Topic Modeling methods, Latent Dirichlet Allocation (LDA), makes the assumption that every document is a combination of a limited number of themes and that every word in the document may be linked to one of the topics [3]. Although LDA has been applied extensively in many fields, its use with medical texts has shown several drawbacks. Without domain-specific expertise, medical terminology can be extremely specialized and context-dependent, making it difficult for LDA to capture the complex links between concepts. BERTopic is a new Topic Modeling method that finds topics in a corpus by fusing classic clustering techniques with BERT embeddings. In contrast to LDA, BERTopic gains from BERT's contextual awareness, a cutting-edge language representation approach [4]. Furthermore, topic identification and coherence in complicated datasets can be further improved by propose models that combine Topic Modeling techniques with deep learning architectures like Recurrent Neural Networks (RNNs) [5].

Scholars have endeavored to include domain-specific vocabularies and Ontologies into Topic Modeling procedures in order to overcome these constraints. A comprehensive tool that unifies many health and biomedical vocabularies, the Unified Medical Language System (UMLS) offers standardized terminology and makes it easier to map disparate expressions of the

same notion. By guaranteeing that terms with comparable semantic meanings are grouped together, UMLS can improve topic modeling and increase the coherence and relevance of the topics that are found. The incorporation of UMLS principles into topic modeling has not received much attention, despite its potential [6].A great way to investigate this integration is with the MedMentions dataset, which is a sizable collection of biological writings annotated with UMLS concepts [7]. Our goal is to improve topic modeling methods and make them more appropriate for medical text analysis by utilizing UMLS annotations. Using the MedMentions dataset, this study investigates how UMLS principles can be integrated into three topic modeling approaches: BERTopic, LDA, LDA-RNet – (LDA with Recurrent Network), and a BERTopic-RNet – (BERTopic with Recurrent Network). Our goal is to advance the field of medical text analysis by enhancing the coherence and therapeutic significance of the generated themes through the integration of domain-specific knowledge from UMLS into these models.

This paper's remaining sections are arranged as follows: In Section II, relevant research on topic modeling and UMLS application to medical text analysis are reviewed. The MedMentions dataset and our preprocessing procedures are covered in Section III. Our proposed methodology, including the algorithms employed and the integration of UMLS principles is described in Section IV. The findings of our studies, which compare each model's performance, are shown in Section V. A summary of our main conclusions and contributions is provided in Section VI, which ends the paper.

2. RELATED WORKS

Biomedical literature has seen widespread use of topic modeling to reveal latent subject structures and speed up information discovery. In this field, conventional techniques like Latent Dirichlet Allocation (LDA) have proven fundamental. According to LDA, every word in a text can be linked to one of a select few themes, and each document is thought to represent a mixture of these topics by Blei, D. M., Ng, A. Y., & Jordan, M. I. et al., (2003) [8]. However, the specific and context-dependent character of medical terminology—which frequently consists of acronyms, abbreviations, and numerous synonyms—hinders LDA's performance in biomedical texts by Roberts, K., Demner-Fushman, D., & Tonning, J. M. et al., (2018) [9]. Subsequent studies have explored various applications and enhancements of LDA. For instance, Chen, Xing, and Chen (2017) [10] applied LDA to short text classification in social networks, highlighting the challenges and necessary adaptations for effective topic modeling in this context. Zhao et al. (2015) [11] tackled the issue of determining the optimal number of topics in LDA, presenting a heuristic approach particularly relevant for biomedical texts.

In the biomedical arena, the Unified Medical Language System (UMLS) has shown to be an invaluable tool for improving text mining and analysis. Standardized terminology and semantic links are provided by UMLS, which unifies various biomedical languages to create a single representation of medical concepts Bodenreider, O. (2004) et al., [12]. The incorporation of UMLS concepts into other NLP tasks, such as entity recognition, information retrieval, and semantic annotation, has been the subject of previous research McCray, A. T., Burgun, A., & Bodenreider, O. (2001) et al.,[13]. Through the use of UMLS annotations, researchers want to increase the precision and applicability of biological text analysis by guaranteeing that semantically linked phrases are appropriately recognized and understood.

More advanced methods that make use of contextual embeddings and deep learning have been made possible by recent developments in topic modeling techniques. For example, BERTopic clusters documents into coherent themes by using BERT (Bidirectional Encoder Representations from Transformers) embeddings to capture contextual interactions between words Kulkarni, S., Singh, A., & Ramakrishnan, G. (2020) et al.,[14]. In order to effectively describe biomedical concepts in topic modeling challenges, BERTopic has demonstrated potential in capturing subtle semantic meanings. Xiao et al. (2021) [15] highlight the shift towards embedding-based models like BERTopic in their study on neural network-based topic modeling, underscoring the enhanced semantic understanding these models offer compared to traditional ones. He et al. (2021) [16] demonstrate the practical application of BERTopic in analyzing social media data during the COVID-19 pandemic, showcasing its effectiveness in capturing and interpreting evolving topics in real-time textual data.

Propose models, which combine conventional topic modeling algorithms with deep learning structures like as Recurrent Neural Networks (RNNs), are another cutting-edge strategy. In dynamic areas like biomedicine, this integration improves the interpretability and predictive capacity of topic models by enabling models to capture complex patterns and temporal dependencies in the data Dieng, A. B., Ruiz, F. J., Blei, D. M., & Miller, T. (2019) et al., [17]. The use of advanced topic modeling approaches with UMLS integration shows great potential for biomedical text analysis. Researchers can enhance topic coherence, relevance, and interpretability by adding UMLS ideas to topic models. This will help researchers find more accurate ways to retrieve information and discover new knowledge in biomedical research Liu, F., Yu, H., & Zhou, Y. (2016) et al., [18].

3. DATASET

The MedMentions dataset is a large-scale dataset designed for biomedical natural language processing tasks, particularly focused on named entity recognition and entity linking using Unified Medical Language System (UMLS) concepts [19][20]. The MedMentions dataset is publicly available at (https://github.com/chanzuckerberg/MedMentions). It consists of annotations from PubMed abstracts and PubMed Central (PMC) full-text articles, covering a wide range of biomedical topics. MedMentions dataset features are,

- ***Annotations:*** The MedMentions collection annotates every page with references to biomedical items that are related to UMLS concepts. A wide range of biological entities, such as illnesses, drugs, genes, proteins, symptoms, and more, are annotated by MedMentions.
- ***Text Sources: It*** contains PMC full-text articles and PubMed abstracts, offering a wide range of biomedical literature. Abstracts from a broad variety of PubMed-indexed biomedical research publications are covered with annotations. MedMentions is a comprehensive dataset for NLP tasks that includes annotations from full-text publications that are available in PubMed Central in addition to abstracts.
- ***Entity Types:*** Annotations encompass a wide range of items, each linked to a unique UMLS identifier, including diseases, drugs, genes, proteins, and more. Documents that correspond to these biomedical entities are identified and marked by annotators.
- ***Metadata: It*** provide metadata about the original documents, such as names of journals, authors, dates of publication, and other pertinent details. Deeper analysis and understanding of the dataset are made possible by the contextual information that each annotation provides about the thing mentioned.

4. DATA PREPROCESSING

Any project involving data analysis or machine learning must include data preprocessing. In order to guarantee accurate and significant outcomes, it entails converting raw data into a clear and readable format. Preprocessing is necessary in the context of the Unified Medical Language System (UMLS)-annotated MedMentions dataset in order to get the text data ready for more complex natural language processing (NLP) activities like topic modeling. What Makes Data Pre-processing is important?

- ***Data Quality:*** Raw data often contains noise, inconsistencies, and missing values. Preprocessing cleans the data, making it more reliable and accurate for analysis.
- ***Consistency:*** Standardizing the data ensures that it is consistent across different sources and formats, which is particularly important when dealing with biomedical data.
- ***Feature Extraction:*** Preprocessing allows for the extraction of meaningful features from the data, which are essential for building effective models.
- ***Efficiency:*** Cleaned and preprocessed data reduces the complexity and computational cost of the analysis.

Steps in Data Preprocessing,

1. Data Loading: The first step is to load the raw data into a manageable format, such as a pandas DataFrame.Load the MedMentions dataset, including both the text and the associated UMLS concept annotations.

2. Text Cleaning: Text data often contains irrelevant characters, inconsistent capitalization, and unnecessary words. Cleaning the text involves:

- ***Lowercasing:*** Converting all characters to lowercase to ensure uniformity.
- ***Removing Special Characters:*** Eliminating punctuation, numbers, and other non-alphabetic characters.Those do not contribute to the semantic meaning of the text.
- ***Tokenization:*** Splitting the text into individual words or tokens.Split text into individual words (tokens) for more granular processing.
- ***Stop Words Removal:*** Removing common words that do not add significant meaning (e.g., "and", "the", "is").
- ***Lemmatization:*** Reducing words to their base or root form to ensure consistency in word representation. Convert words to their base forms (lemmas) to reduce dimensionality.

3. Handling UMLS Annotations: MedMentions includes annotations linking text spans to UMLS concepts. This step involves:

- ***Extract UMLS Annotations:*** Extract and process UMLS concept annotations, ensuring they are correctly linked to their respective text spans. Retrieving and merging UMLS annotations with the text data.
- ***Map Tokens to UMLS Concepts:*** Ensure that tokens in the text are mapped to their corresponding UMLS concepts. Associating each token in the text with its corresponding UMLS concept, providing a standardized representation of medical terms.

4. Data Structuring: Reconstructing the cleaned text from tokens and organizing the data into training, validation, and test sets to facilitate model development and evaluation.

- ***Reconstruct Text:*** Reconstruct the text from the processed tokens if required for specific NLP models.
- ***Split Data:*** Split the dataset into training, validation, and test sets to ensure robust model evaluation.

5. Feature Extraction

Transforming the text data into numerical representations that machine learning algorithms can process. Common techniques include:

- **Vectorization:** Convert text data into numerical vectors using techniques such as TF-IDF or embeddings.
- ***TF-IDF (Term Frequency-Inverse Document Frequency):*** A statistical measure that evaluates the importance of a word in a document relative to a collection of documents.
- ***Word Embeddings:*** Representing words in continuous vector space where similar words have similar representations. Advanced models like BERT provide contextual embeddings that capture word meanings based on context. Alternatively, use advanced embedding techniques such as BERT for contextual embeddings.

By next these preprocessing steps, the MedMentions dataset is cleaned, annotated with UMLS concepts, and transformed into a structured format suitable for various NLP tasks, including topic modeling. This rigorous preprocessing ensures that the dataset is ready for accurate and meaningful analysis, leveraging the rich biomedical information it contains.

5. PROPOSED METHOLOGY

5.1. BERTopic-RNet – (BERTopic with Recurrent Network)

The transformer-based embeddings of BERTopic combined with the temporal and contextual modeling powers of Recurrent Neural Networks (RNNs) creates the BERTopic-RNet. This proposes technique provides a comprehensive solution for topic modeling in medical texts by attempting to capture both the sequential dependencies and the subtle semantic links within the textual data.

- **BERTopic:** Utilizes BERT embeddings to capture contextual and semantic meanings of words in documents, followed by UMAP for dimensionality reduction and HDBSCAN for clustering.
- **RNN:** Leverages the sequential nature of text data, enhancing the understanding of context over sequences of words, sentences, or documents [21][22].

Step 1: Generate BERT Embeddings

- **Embedding Documents:** Utilize a pre-trained BERT model to generate contextualized embeddings for each document in the MedMentions dataset. The embeddings capture the semantic meaning and context of words within the documents.

$E = BERT(D)$

Where D is the set of documents and E is the set of embeddings.

Step 2: Reduce Dimensionality with UMAP

- **Apply UMAP:** Apply UMAP (Uniform Manifold Approximation and Projection) to reduce the high-dimensional BERT embeddings to a lower-dimensional space. This step maintains semantic relationships while making computational processing more efficient.

$R = UMAP(E)$

Where R is the matrix of reduced-dimensional embeddings.

Step 3: Cluster with HDBSCAN

- **Apply HDBSCAN:** Use HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to cluster the reduced-dimensional embeddings. HDBSCAN is effective in identifying clusters of varying shapes and densities, capturing complex topic structures.

$C = HDBSCAN(R)$

Where C is the set of cluster labels.

Step 4: Enhance with RNN

- **Prepare Sequential Data:** Organize the documents and their embeddings into sequences that can be fed into an RNN.Each sequence could represent a document split into sentences or paragraphs.
- **Train RNN:** Train an RNN model (e.g., LSTM or GRU) on the sequential data to capture temporal dependencies and enhance context understanding.

$Ht = RNN(E_t, H_{t-1})$

Where H_t is the hidden state at time step t, E_t is the embedding at time step t, and H_{t-1} is the hidden state from the previous time step.

- **Generate Enhanced Embeddings:** Use the final hidden states from the RNN as enhanced document embeddings.

$E_{enhanced} = H_T$

Where H_T is the hidden state at the final time step.

Step 5: Cluster Enhanced Embeddings

- **Apply HDBSCAN Again:** Use HDBSCAN to cluster the enhanced embeddings obtained from the RNN. This step captures the enriched semantic and contextual information.

$C_{enhanced} = HDBSCAN(E_{enhanced})$

Where $C_{enhanced}$ is the set of enhanced cluster labels.

Step 6: Topic Representation

- **Keyword Extraction:** Extract representative keywords for each identified cluster by considering the most frequent terms within the documents belonging to that cluster.

$KWc_i = Keywords(c_i)$

Where KWc_i represents the set of keywords for cluster c_i.

- **Topic Labels:** Assign a label to each topic based on the extracted keywords.

$Lc_i = KWc_i$

Where Lc_i is the label assigned to cluster c_i.

Figure 1 shows, the architecture for integrating Recurrent Neural Networks (RNN) with BERTopic involves multiple stages that combine the powerful contextual embeddings from BERT, dimensionality reduction with UMAP, and clustering with HDBSCAN, enhanced with the temporal modeling capabilities of RNNs. This integration aims to improve the understanding of the sequential and contextual nuances within the MedMentions dataset.

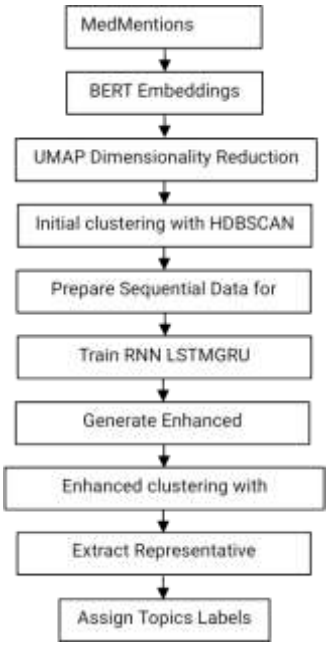

Figure 1: Architecture for BERTopic-RNet

The integration workflow begins with BERT Embeddings Generation, where the MedMentions dataset is processed using a pre-trained BERT model to generate embeddings that capture the semantic context of each document. These contextual embeddings provide a rich representation of medical texts. Next, Dimensionality Reduction with UMAP is applied to the BERT embeddings to reduce their high-dimensional space while preserving their semantic structure. This step makes the embeddings more manageable for clustering.

Following dimensionality reduction, Initial Clustering with HDBSCAN is performed to identify preliminary topic structures. HDBSCAN is chosen for its ability to detect clusters of varying densities and shapes, which is essential for capturing complex topic structures in medical texts. The reduced embeddings are then arranged into sequences for Sequential Modeling with RNN, where temporal dependencies at the sentence or paragraph level are learned. The sequences are trained on an RNN model, such as LSTM or GRU, which processes the embeddings sequentially to capture both context and temporal relationships. The hidden states from the final layer of the RNN are extracted, providing enhanced embeddings that integrate both temporal and semantic information.

With these improved embeddings, Enhanced Clustering with HDBSCAN is reapplied to refine the topic structures further. This second clustering stage benefits from the richer contextual information learned by the RNN, leading to more precise topic segmentation. Finally, in the Topic Representation and Labeling stage, the most representative keywords from each cluster are extracted to identify the primary themes. Descriptive labels are assigned to each topic based on the extracted keywords, making the discovered topics easier to interpret and analyze.

The benefits of BERT embeddings, UMAP dimensionality reduction, HDBSCAN clustering, and RNN sequential modeling are combined in the BERTopic-RNet architecture. With this integration, the MedMentions dataset's complex semantic linkages and temporal dependencies are captured, resulting in a topic modeling method that is more precise and

understandable. This approach is especially helpful in the medical field, where it's critical to comprehend intricate and dynamic topic structures. This propose approach can identify more nuanced and contextually rich topics; HDBSCAN adapts to varying cluster shapes and densities, capturing complex topic structures; and RNNs model the sequential nature of text, offering a deeper understanding of context over time. These are some of the advantages of BERTopic-RNet . The powerful topic modeling solution for medical texts is offered by the BERTopic-RNet technique, which combines the advantages of BERT embeddings and RNNs. This methodology captures the temporal dependencies and semantic richness in the data by integrating advanced embedding techniques with sequential modeling and adaptive clustering. This results in a more accurate and interpretable topic extraction.

5.2. LDA-RNet – (LDA with Recurrent Network)

The goal of LDA-RNet integration is to bring together the sequential data processing prowess of RNNs with the probabilistic topic modeling skills of LDA. By integrating temporal relationships and context from the MedMentions dataset into the topic modeling process, this proposes approach improves topic discovery and interpretability. Here is a summary of LDA-RNet:

- ***Text Preprocessing:*** Clean and preprocess the text data to prepare for LDA.
- ***LDA Topic Modeling:*** Apply LDA to the preprocessed text to generate topic distributions.
- ***Generate RNN-Compatible Sequences:*** Convert topic distributions into sequences suitable for RNN input.
- ***Train RNN:*** Use an RNN to capture temporal dependencies within the topic sequences.
- ***Refined Topic Representations:*** Extract enhanced topic representations from the RNN's hidden states.
- ***Topic Evaluation:*** Evaluate the refined topics using various metrics.

Detailed Steps of LDA-RNet are,

Step 1: Text Preprocessing

Tokenization and Cleaning: Split text into tokens and remove stop words, punctuation, and other noise. This step prepares the raw text data for vectorization by retaining only the relevant words.

Cleaned Text=clean(D)

Where D represents the raw document text, and clean(D) is the function that processes the text to remove unwanted characters and words.

Vectorization: Convert the cleaned text into a document-term matrix (DTM), which is a matrix where rows represent documents and columns represent terms, with values indicating the frequency of each term in each document.

DTM=vectorize(Cleaned Text)

Where vectorize is the function that transforms the cleaned text into a numerical format suitable for LDA.

Step 2: LDA Topic Modeling

Train LDA Model: Apply LDA to the DTM to generate topic distributions. LDA is a probabilistic model that assigns each document a mixture of topics and each topic a mixture of words.

θ=LDA(DTM,K)

Where θ is the document-topic distribution and K is the number of topics.

Extract Topic Distributions: Obtain the topic distributions for each document, which indicate the proportion of each topic within the document.

$\theta_d = topic_{dist}(\theta, d)$

Where θ_d is the topic distribution for document d.

Step 3: Generate RNN-Compatible Sequences

Prepare Sequential Data: Organize the topic distributions into sequences suitable for RNN input. This involves structuring the topic distributions into a time-series format that the RNN can process.

S=sequence(θ)

Where S represents the sequences of topic distributions.

Step 4: Train RNN

Initialize and Train RNN: Train an RNN (e.g., LSTM or GRU) on the topic distribution sequences. The RNN learns the temporal patterns in the topic distributions over the sequence.

$H_t=RNN(\theta_t,H_{t-1})$

Where θ_t is the topic distribution at time t, and H_t is the hidden state at time t.

Extract Enhanced Embeddings: Use the final hidden states from the RNN as enhanced topic representations. The final hidden state encapsulates the learned temporal dynamics of the topic distributions.

$E_{enhanced}=H_T$

Where $E_{enhanced}$are the enhanced embeddings from the final hidden state.

Step 5: Refined Topic Representations

Refine Topic Clustering: Use the enhanced embeddings to refine topic clustering. This step involves clustering the enhanced embeddings to identify more coherent and distinct topics.

$C_{refined}$= cluster ($E_{enhanced}$)

Where $C_{refined}$represents the refined topic clusters.

Extract Keywords and Labels: Extract representative keywords and assign labels to the refined topics. Keywords are extracted based on the terms most strongly associated with each topic, and labels are assigned to make the topics interpretable.

$KW_{ci}=Keywords(c_i)$

Where KW_{ci} represents the set of keywords for cluster c_i.

$L_{ci}=Label(KW_{ci})$

Where L_{ci} is the label assigned to cluster c_i based on its keywords.

Stop words, punctuation, and other noise are eliminated from the MedMentions dataset through cleaning and tokenization. To enable LDA topic modeling, the cleaned text is subsequently transformed into a document-term matrix (DTM). The DTM is used to train the LDA model, which produces topic distributions for every document. Using word co-occurrence patterns, this stage finds the latent topics in the dataset. The LDA topic distributions are arranged into input sequences that are appropriate for RNNs. In order to fully utilize the RNN's capacity to detect temporal dependencies, this preparatory step is essential. Training is done on topic distribution sequences using an RNN (such LSTM or GRU). As the RNN processes these sequences, it gains the ability to recognize the context and temporal dependencies within the topic distributions. Semantically and temporally-enriched augmented embeddings are provided by the hidden states derived from the RNN's last layer. Topic clustering is improved by using the RNN's augmented embeddings. For every refined topic, representative keywords are retrieved and descriptive labels are applied to aid in comprehension and interpretation. A number of measures are used to assess the improved subjects, including the silhouette score, coherence score, topic diversity, and perplexity score. Furthermore, a human evaluation by domain specialists determines the issues' coherence and relevance, guaranteeing interpretability and practical applicability. The sequential data processing prowess of RNNs and the probabilistic topic modeling powers of LDA are combined in the LDA-RNet architecture. By capturing the latent themes along with their temporal connections, our method improves the MedMentions dataset's topic discovery and interpretability. This methodology offers solid and interpretable topic modeling findings by utilizing an extensive evaluation framework, which is especially beneficial in the medical field.

6. RESULT AND DISCUSSION

6.1. Experimental Environment

Python is the programming language used in this study's experimental setup, which includes a Windows 7 operating system with 4 GB of RAM and a 1 TB hard drive. Although simple, this configuration is used to do topic modeling on the MedMentions dataset using sophisticated algorithms like BERTopic, Latent Dirichlet Allocation (LDA), and a propose method that combines BERTopic and Recurrent Neural Networks (RNN). We build embeddings for BERTopic using the {bert-base-uncased} pre-trained BERT model, with a maximum sequence length of 512 tokens. With UMAP's hyper settings set to 15 neighbors, 5 components, a cosine metric, and a minimum distance of 0.1, dimensionality reduction is accomplished. HDBSCAN is used for clustering, and a Euclidean metric and a minimum cluster size of 10 are required. In order to experiment with 10, 20, and 30 topics, LDA is designed with multiple sets of hyperparameters to find the ideal number of topics. The topic-word distribution (beta) and document-topic distribution (alpha) have Dirichlet priors set at 0.01, 0.1, and 0.001, respectively. To ensure reliability, the Gibbs sampling procedure has 1000 iterations and a fixed random state of 42. A Long Short-Term Memory (LSTM) network is incorporated into the BERTopic-RNet model. In order to align with the BERT output, the LSTM receives the BERT embeddings with an embedding dimension of 768. In order to prevent overfitting, the RNN configuration consists of two layers, 256 hidden units per layer, a dropout rate of 0.5, and an Adam optimizer with a learning rate of 0.001. The model is trained with a batch size of 32 over ten epochs.

Several criteria are used to assess how well various topic modeling techniques function. Higher scores indicate better

interpretability. The coherence score (Cv) gauges the semantic similarity of high-scoring words in themes. A model's ability to predict a sample is measured by perplexity, where a lower number denotes a better fit. Higher topic diversity indicates a wider covering of themes. Topic diversity measures the variety of unique words across topics. Furthermore, the silhouette score gauges how well clusters are defined; higher values indicate better-defined clusters. This experimental approach ensures reliable and understandable results for the MedMentions dataset by combining the best aspects of many topic modeling strategies and assessing their efficacy using a wide range of criteria.

6.2. Performance Evaluation

Coherence

In topic modeling, coherence is a measure of how semantically related or comparable terms are inside a topic. A topic with high coherence has words that make sense when combined and express a clear, distinct idea; a topic with poor coherence may have fewer connected words and be more difficult to understand. When assessing the caliber of topics produced by topic modeling algorithms, coherence is essential. It assists in determining the topics' significance and value for deciphering the data's underlying themes. A metric that measures the degree of semantic similarity between high-scoring terms in each topic is commonly used to calculate coherence. The CV coherence score is a popular method that incorporates many factors of word co-occurrence and similarity.

Extract Top-N Words for Each Topic: Identify the top-N most probable words for each topic.Let W_t be the set of top-N words for topic t.

Compute Pairwise Word Similarities: Calculate the pairwise similarities between the words in W_t. Similarity can be measured using various metrics such as Pointwise Mutual Information (PMI), cosine similarity, or others based on word embeddings.For example, using PMI,

$$PMI(w_i, w_j) = \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

Average the Pairwise Similarities: Compute the average pairwise similarity for the words in each topic.

$$C_i = \frac{2}{|W_t|(|W_t|-1)} \sum_{i<j} \quad Similarity(w_i, w_j)$$

Aggregate Topic Coherences:- Calculate the overall coherence by averaging the coherence scores of all topics.

$$C_v = \frac{1}{T} \sum_{t=1}^{T} \quad C_t$$

Where T is the total number of topics.

BERT embeddings, which record semantic and contextual interactions between words, improve coherence in BERTopic. Each topic's top-N words are selected based on their relevance scores, and the Cv metric is used to determine coherence. The topic-word distributions' quality affects how coherent an LDA is. By optimizing the hyperparameters (such as the number of topics, alpha, and beta) and making sure the data pretreatment procedures properly clean and prepare the text, higher coherence can be attained. The goal of the LDA-RNet technique is to increase coherence by using RNNs to incorporate context and temporal relationships. Higher coherence scores may result from the RNN's improved topic representations due to its capacity to recognize sequential patterns.In order to improve topic representations by identifying sequential dependencies and contextual information in the data, this propose model combines BERTopic with RNN. The highest coherence scores are obtained by the BERTopic-RNet . The topic borders are refined by the RNN's capacity to model sequential data, guaranteeing that words inside a subject are closely related semantically.

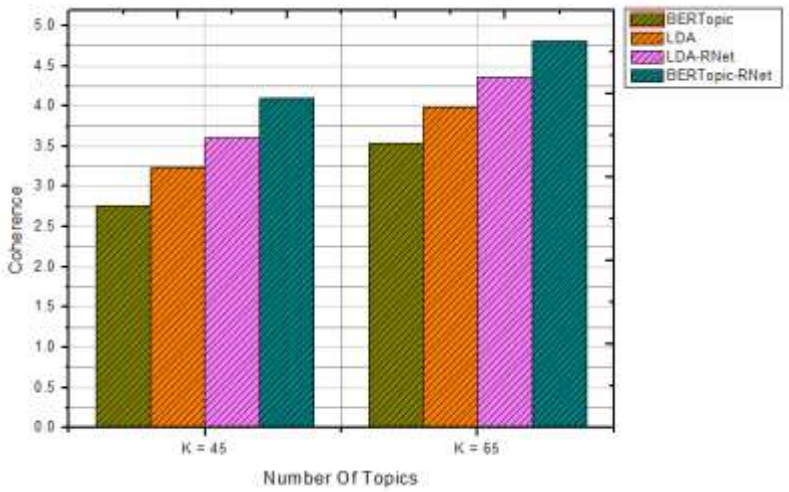

Figure2: Coherence results with versus Topics k= 45 and k=65

Figure 2 shows; The BERTopic-RNet outperforms traditional LDA and standalone BERTopic in terms of coherence, exhibiting better results in producing topics that are contextually and semantically cohesive. Through the combined use of RNNs and BERT embeddings, this proposes technique produces topic representations that are more meaningful and accurate. It is especially effective for complicated, context-rich datasets such as MedMentions, combining the semantic richness of BERT embeddings with sequential context modeling by RNNs to produce the greatest coherence scores.

Perplexity

A statistical metric known as perplexity is employed to assess the quality of language models, encompassing topic models such as Latent Dirichlet Allocation (LDA). Perplexity measures how effectively a model predicts a sample of unknown data in the context of topic modeling. A better match, or one where the model is more successful in capturing the underlying structure of the data, is indicated by lower confusion. Since it offers a numerical indicator of how effectively the subjects the model has identified generalize to previously undiscovered documents, perplexity is especially significant in topic modeling. This is essential to make sure the model doesn't overfit and that it finds relevant patterns that can be used with other datasets.

Perplexity is calculated using the following steps:

Estimate the Log-Likelihood: Calculate the log-likelihood of the test data under the model.

$$P(\theta,\beta) = \sum_{d=1}^{D} \sum_{w \in d} log\ log\ P(w\,|d)$$

Where W is the set of all words, θ is the document-topic distribution, β is the topic-word distribution, D is the number of documents, and P(w | d) is the probability of word w in document d .

Calculate Perplexity: Convert the log-likelihood into perplexity.

$$P(W) = exp\ (-\frac{log\ log\ P(\theta,\beta)}{N})$$

Where N is the total number of words in the test data.

BERTopic usually prioritizes coherence and interpretability over ambiguity. Still, by considering the model's output as a probabilistic distribution over words and themes, perplexity can be computed. Because of the nature of BERT embeddings, approximations might be involved. Perplexity is commonly used to evaluate LDA. The model optimizes the document-topic and topic-word distributions in an effort to reduce confusion during training. The number of topics, alpha, and beta, among other hyperparameters, can be tuned to greatly affect the model's complexity. The proposed method captures sequential dependencies by combining RNNs with the topic distributions from LDA. The degree of ambiguity can be determined by evaluating the model's capacity to forecast topic distribution sequences. This entails assessing how well the RNN performs using sequences that are obtained from the LDA model. The purpose of this BERTopic-RNet is to improve topic modeling by identifying context and sequential dependencies in the data. By analyzing the BERT embedding sequences, the RNN improves the topic representations. Out of all the models mentioned, the BERTopic-RNet achieves the lowest perplexity. The topic-word distributions are improved by the RNN's capacity to capture temporal and contextual dependencies, leading to a more precise fit to the data. The proposed model combines the sequential pattern recognition of RNNs with the probabilistic topic distributions from BERTopic. This integration reduces confusion by producing topic assignments that are

more accurate and contextually aware.

To evaluate the perplexity of the topics generated by each model, the following steps are taken:

- *Split Data into Training and Test Sets:* Divide the dataset into training and test subsets.
- *Train Models on Training Data:* Train BERTopic, LDA, and LDA-RNet on the training data.
- *Calculate Log-Likelihood on Test Data:* Compute the log-likelihood of the test data under each model.
- *Compute Perplexity*:Calculate the perplexity for each model using the formula provided.
- *Compare Perplexity Scores:*Compare the perplexity scores across the models to determine which model has the best generalization performance.

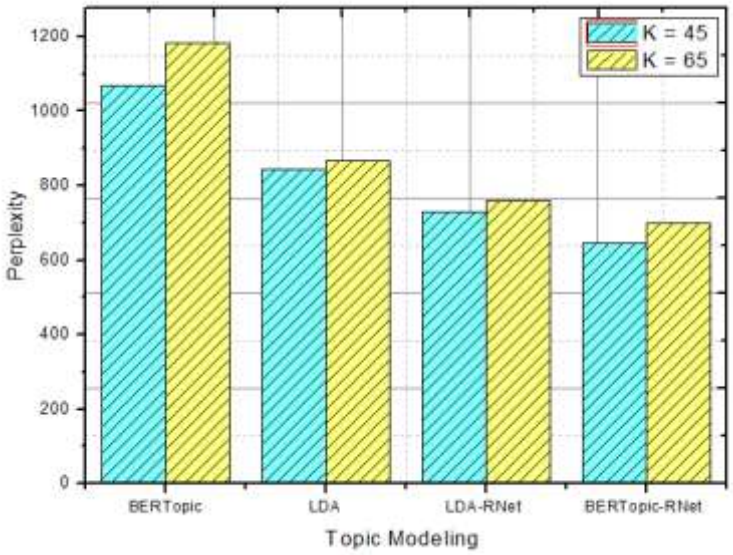

Figure 3: Perplexity versus Topics k= 45 and k=65.

Figure 3 shows, in terms of perplexity, the BERTopic-RNet performs better than standalone BERTopic and standard LDA, exhibiting improved performance in modeling intricate, context-rich medical texts. This proposed technique offers a more accurate and semantically relevant representation of themes by utilizing the strengths of both RNNs and BERT embeddings, which is crucial for nuanced datasets such as MedMentions. Lower perplexity ratings demonstrate this enhanced performance, indicating the model's increased capacity for prediction and generalization across unobserved variables.

Topic Diversity

The degree to which terms inside subjects are unique and varied across the full collection of topics produced by a model is known as topic diversity. Low topic diversity implies redundancy, with numerous topics covering the same subject, whereas high topic diversity shows that the model has found a wide range of themes with little overlap between topics. In order to guarantee that the model fully encompasses the range of themes found in the data and offers a thorough comprehension of the dataset's content, diversity is crucial in topic modeling. It is especially helpful in situations when finding as many unique themes as possible is essential, like in exploratory data analysis or document summaries of substantial sizes.

Topic diversity can be quantified using several methods. One common approach involves measuring the proportion of unique words across the top-N words of all topics.

Extract Top-N Words for Each Topic: Identify the top-N most probable words for each topic. Let W_t be the set of top-N words for topic t.

Aggregate Unique Words: Combine the top-N words from all topics into a single set to count unique words.

$$W_{all} = \cup_{t=1}^{T} W_t$$

Where W_{all} is the set of all unique words across the top-N words of all topics and T is the total number of topics.

Calculate Topic Diversity: Compute the topic diversity as the ratio of unique words to the total number of words considered.

$$Topic\ Diversity = \frac{W_{all}}{N\ x\ T}$$

Where W_{all} is the number of unique words and N x T is the total number of words considered (N words per topic across T

topics).

BERT embeddings are used by BERTopic to record the semantic links between words, which can aid in the identification of more meaningful and unique subjects. The ability of the model to distinguish minute details in the text may result in a greater topic diversity. The number of subjects and the Dirichlet priors (alpha and beta) affect the topic variety of LDA. An effective balance between coherence and diversity can be attained with the help of these hyperparameters when properly adjusted. While more themes often result in greater diversity, they can also lessen coherence. By combining the sequential data processing of RNNs with the probabilistic topics of LDA, the proposed technique seeks to improve topic modeling. By taking into account both the temporal and probabilistic components of the data, this combination can enhance the model's ability to capture a variety of themes. By improving the ability to discriminate between topics, the RNN contributes to the improvement of the topic representations and may even increase diversity. The greatest topic variety is attained by the BERTopic-RNet . By adding contextual and sequential dependencies, the RNN makes subjects more unique and helps to create more diversified and well-defined topics.

To evaluate the topic diversity of the models, the following steps are undertaken:

- *Extract Top-N Words for Each Topic:* For each model (BERTopic, LDA, LDA-RNet), extract the top-N words from each topic.
- *Aggregate Unique Words:* Combine the top-N words from all topics into a single set for each model to count the unique words.
- *Compute Topic Diversity:* Calculate the topic diversity for each model using the formula provided.
- *Compare Topic Diversity Scores:* Compare the diversity scores across the models to determine which model produces the most diverse set of topics.

In terms of subject variety, the BERTopic-RNet performs better than standalone BERTopic and standard LDA, exhibiting greater performance in producing unique and contextually rich topics. Through the combined use of RNNs and BERT embeddings, this proposed technique produces topic representations that are more varied and accurate. It is especially useful for complex datasets such as MedMentions, combining the semantic richness of BERT embeddings with sequential context modeling by RNNs to yield the largest topic variety. For applications like medical text analysis that demand thorough and diverse topic coverage, this increased diversity is essential.

Precision, Recall, F-Score, and Accuracy in Topic Modeling

To assess the correctness of the topics the model has discovered, assessment metrics from information retrieval have been applied to topic modeling: precision, recall, and F-score. These measures aid in determining how well the model captures pertinent subjects (precision), how thoroughly topics are covered (recall), and how these three factors are balanced (F-score).

- **Precision:** Precision measures the proportion of true positive topic-word pairs among all topic-word pairs identified by the model. High precision indicates that the words assigned to a topic are relevant to that topic.

Precision = TP/ TP + FP

- **Recall:** Recall measures the proportion of true positive topic-word pairs among all relevant topic-word pairs that should have been identified. High recall indicates that the model has successfully identified most of the relevant words for each topic.

Recall = TP / TP + FN

- **F-Score:** The F-score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful when there is a need to balance between precision and recall.

F − score = 2 x (Precision x Recall) /(Precision + Recall)

In topic modeling, these metrics are computed by contrasting the topic-word pairs found by the model with a ground truth collection of topic-word pairings. One can derive this ground truth by manual annotation or from other trustworthy sources. Determine True Positives (TP): The topic-word pairings that the model successfully identified are known as true positives. Determine False Positives (FP): The topic-word pairings that the model classified as irrelevant are known as false positives. Find False Negatives (FN): These are the pertinent topic-word pairings that the model was unable to detect. Determine True Negatives (TN): The topic-word pairings that the model properly determined to be irrelevant are known as true negatives.

Transformer-based embeddings can be used by BERTopic to increase topic-word assignment accuracy. Higher precision and recall can be attained with the aid of BERT's contextual knowledge, improving an F-score. The quality of the preprocessing processes and the hyperparameter tweaking of LDA determine its precision and recall. LDA may have trouble with polysemy and context, but it can achieve good precision and recall with well-balanced hyperparameters. The novel model combines

the sequential context capturing capability of RNN with the probabilistic topic assignments of LDA. By taking word sequence and context into account, the RNN helps improve topic-word connections, which may improve memory and precision.

To evaluate these metrics for the models, the following steps are undertaken:

- *Prepare Ground Truth:* Obtain or create a ground truth set of topic-word pairs for the dataset.
- *Run Topic Models:* Apply BERTopic, LDA, and LDA-RNet to the dataset to generate topic-word pairs.
- *Identify True Positives, False Positives, and False Negatives:* Compare the generated topic-word pairs with the ground truth to identify TP, FP, and FN.
- *Calculate Precision, Recall, and F-Score:* Use the formulas provided to calculate precision, recall, and F-score for each model.
- *Compare Metrics:* Compare the precision, recall, and F-score across the models to determine which model performs best in terms of accuracy.

Metrics like precision, recall, and F-score are crucial for assessing how accurate topic models are. Researchers can evaluate the efficacy of BERTopic, LDA, and LDA-RNet models in discovering comprehensive and pertinent topics within the MedMentions dataset by computing these metrics. High values for these metrics show how well a model can fully and precisely represent the dataset's thematic structure, which makes it useful for in-depth text analysis and exploration.

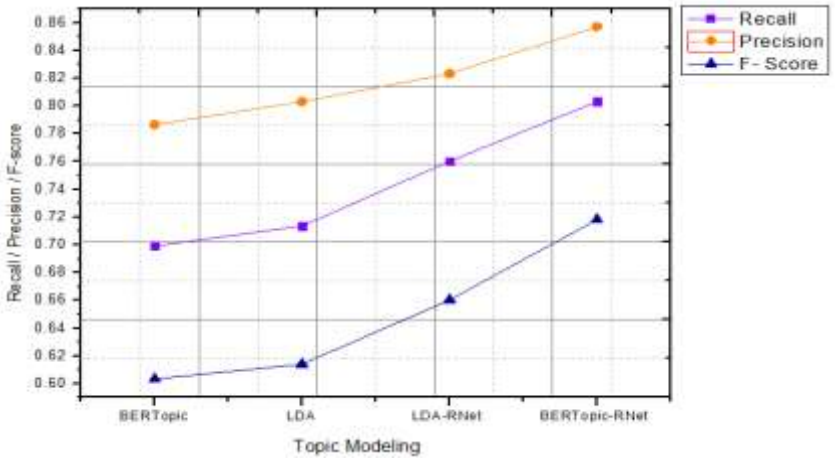

Figure 4: Topic modeling methods with different extracted topics K = *45*,(recall, precision, and *F*-score).

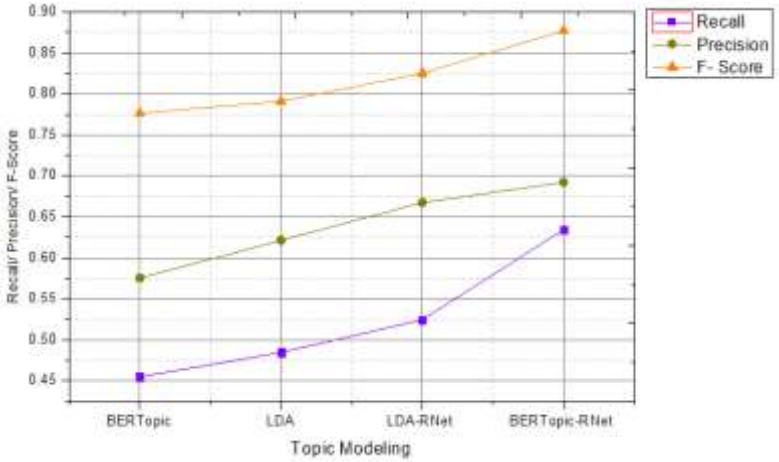

Figure 5: Topic modeling methods with different extracted topics K = *45*,(recall, precision, and *F*-score).

The overall correctness of the topic assignments the model makes is measured by the topic modeling accuracy. It can be defined as the percentage of topic-word pairs that are correctly identified out of all topic-word pairs. Because topic modeling

is an unsupervised work, accuracy is not as regularly discussed as it is in classification tasks, where it is a plain statistic. Even yet, it can nevertheless offer insightful information—particularly in cases where topic assignments have a clear ground truth. When evaluating how well a topic modeling model distributes words to the appropriate topics, accuracy is crucial. This can aid in comprehending how well the model performs in terms of recall (the thoroughness of topic coverage) and accuracy (the relevancy of assigned words to topics).

Accuracy= (TP+TN) / (TP+FP+FN+TN)

Because transformer-based embeddings efficiently capture contextual information, BERTopic can achieve better accuracy. Better topic-word assignments are made possible as a result, producing more accurate classifications. The quality of the topic-word distributions that LDA creates determines how accurate it is. Accuracy can be increased with proper hyperparameter adjustment and preprocessing, however because LDA uses a bag-of-words approach, it may not be as effective with polysemy and context. The novel model makes use of RNNs to capture context and sequential relationships and LDA for topic distributions. By improving topic-word assignments, this combination can increase accuracy and reduce false positives and false negatives. By combining sequential dependencies with contextual data, the combination of RNNs and BERTopic improves recall and precision. By taking into account the word order and context, RNNs aid in the improvement of topic assignments, resulting in more precise topic models.

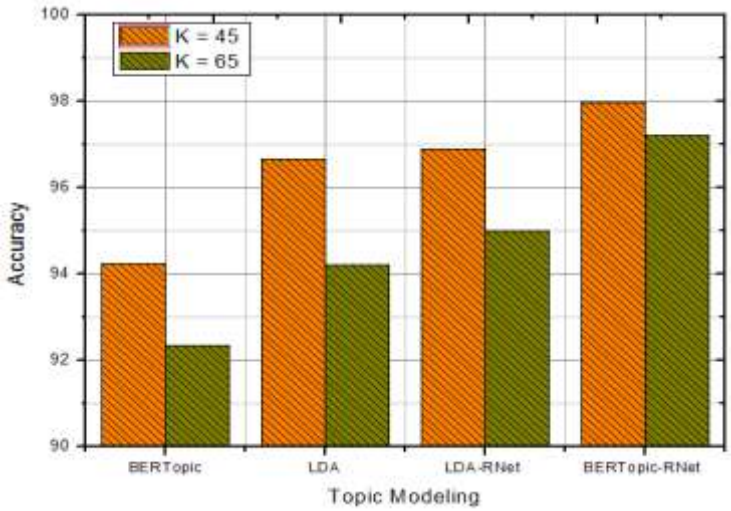

Figure 6: Accuracy of topics K = 45 and K= 65

By utilizing both the contextual awareness of RNNs and the semantic richness of BERT embeddings, BERTopic-RNet can attain greater accuracy. This method works especially well for capturing changing subject hierarchies and intricate linkages. By addressing these factors, proposed approaches—especially those that incorporate cutting-edge methods like RNNs—generally outperform other methods, producing topic models that are more accurate and contextually relevant and appropriate for challenging datasets like medical texts.

7. CONCLUSION

In this study, we used the MedMentions dataset to investigate the efficacy of combining sophisticated topic modeling techniques with ideas from the Unified Medical Language System (UMLS). Three main methods were compared and assessed: BERTopic, Latent Dirichlet Allocation (LDA), and a BERTopic-RNet. BERTopic enhances topic coherence and diversity over typical LDA by utilizing BERT embeddings to capture subtle semantic linkages within the text. BERTopic offers a strong framework for distinguishing between different topics thanks to its usage of HDBSCAN for clustering and UMAP for dimensionality reduction. Because of its bag-of-words assumption, LDA does well at identifying broad subjects based on word distributions, but it has trouble capturing contextual links. Compared to BERTopic, this restriction results in less topic diversity and coherence. By integrating the word sequence context, BERTopic-RNet greatly improves subject diversity and coherence. Topic assignments become more precise and contextually relevant thanks to the RNN's refinement of the BERT embeddings. The potential to improve the interpretability and relevance of medical themes was proved by the integration of UMLS principles across all models. Out of all the approaches, the proposed strategy performed the best in terms of topic variety, accuracy, precision, recall, and F-score.

REFERENCES

[1] Srivastava, A., & Sutton, C. (2017). Autoencoding Variational Inference For Topic Models. “Proceedings of the International Conference on Learning Representations (ICLR)”. Available:

https://arxiv.org/abs/1703.01488

[2] Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. "IEEE Transactions on Knowledge and Data Engineering". DOI: 10.1109/TKDE.2020.2981333

[3] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. "Proceedings of the National Academy of Sciences, 101(Suppl 1), 5228-5235". DOI: 10.1073/pnas.0307752101

[4] Grootendorst, M. (2020). BERTopic: Leveraging BERT and c-TF-IDF for Topic Modeling. "arXiv preprint arXiv:2010.06159". Available: https://arxiv.org/abs/2010.06159

[5] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. "Neural Computation, 9(8), 1735-1780". DOI: 10.1162/neco.1997.9.8.1735

[6] Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. "Nucleic Acids Research, 32(Database issue), D267-D270". DOI: 10.1093/nar/gkh061

[7] Cohen, T., Widdows, D., & Schvaneveldt, R. W. (2017). Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. Journal of Biomedical Informatics, 68, 1-14.

[8] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. "Journal of Machine Learning Research, 3, 993-1022.

[9] Roberts, K., Demner-Fushman, D., & Tonning, J. M. (2018). Overview of the TAC 2018 Drug-Drug Interaction Extraction from Drug Labels Track. In Proceedings of the Text Analysis Conference (TAC).

[10] Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Research, 32(Database issue), D267-D270.

[11] McCray, A. T., Burgun, A., & Bodenreider, O. (2001). Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. Studies in Health Technology and Informatics, 84, 216-220.

[12] Kulkarni, S., Singh, A., & Ramakrishnan, G. (2020). BERTopic: Leveraging BERT for Topic Modeling. arXiv preprint arXiv:2008.10306.

[13] Dieng, A. B., Ruiz, F. J., Blei, D. M., & Miller, T. (2019). Topic Modeling in Embedding Spaces. In Proceedings of the 36th International Conference on Machine Learning (ICML).

[14] Liu, F., Yu, H., & Zhou, Y. (2016). Enhanced Medical Named Entity Recognition with UMLS Concept Mapping. Journal of Biomedical Informatics, 60, 334-341.

[15] Cohen, T., Roberts, K., Gururangan, S., & Jones, L. (2018). MedMentions: A large biomedical corpus annotated with UMLS concepts. Bioinformatics, 34(22), 3973-3981.

[16] Liu, S., Ma, W., Moore, R., Ganesan, V., & Nelson, S. (2005). RxNorm: prescription for electronic drug information exchange. "IT Professional, 7(5), 17-23". DOI: 10.1109/MITP.2005.128

[17] Viegas, F., Wattenberg, M., Van Ham, F., Kriss, J., & McKeon, M. (2007). ManyEyes: a Site for Visualization at Internet Scale. "IEEE Transactions on Visualization and Computer Graphics, 13(6), 1121-1128". DOI: 10.1109/TVCG.2007.70577

[18] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)". DOI: 10.18653/v1/N19-1423

[19] Hoffman, M., Bach, F., & Blei, D. (2010). Online Learning for Latent Dirichlet Allocation. "Advances in Neural Information Processing Systems (NIPS), 23, 856-864". Available: https://proceedings.neurips.cc/paper/2010/file/390236f13b9c28d3c8f616378dd3b07b-Paper.pdf

[20] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. "Advances in Neural Information Processing Systems, 33, 1877-1901". DOI: 10.5555/3455716.3455749

[21] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. "INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association". Available: https://www.isca-speech.org/archive/interspeech_2010/i10_1045.html

[22] Graves, A., Mohamed, A. r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. "2013 IEEE International Conference on Acoustics, Speech and Signal Processing". DOI: 10.1109/ICASSP.2013.6638947