Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

Discovering Latent Themes in Heart Disease Article Abstracts: A Topic Modeling Approach

Kalp Hastalığı Makale Özetlerinde Gizli Temaları Keşfetme: Konu Modelleme Yaklaşımı

Burcu Baştürk [1], **Aytuğ Onan** [2*]

[1] İzmir Katip Çelebi Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, İzmir, TÜRKİYE
[2] İzmir Katip Çelebi Üniversitesi Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, İzmir, TÜRKİYE

Corresponding Author / Sorumlu Yazar *: aytug.onan@ikcu.edu.tr

Abstract

Heart disease is a global public health problem that requires in-depth analysis of extensive literature to uncover specific themes and relationships. This study aimed to identify latent themes and calculate consistencies in 5,000 heart disease-related abstracts retrieved from PubMed using topic modeling techniques. The original abstracts were paraphrased using ChatGPT and NLTK(Natural Language Toolkit), followed by extensive preprocessing, including tokenization, removal of stopped words, stemming, and lemmatization. For effective feature extraction, text data was vectorized using TF-IDF (term frequency-inverse document frequency). Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF) were applied to reveal key thematic structures. Coherence scores were calculated and compared across different numbers of subjects (5 to 50) for each model and annotation method. This approach provides a valuable methodology for summarizing large amounts of information, allowing researchers to efficiently navigate the complex landscape of heart disease literature and identify critical areas of focus. The findings aim to improve understanding of heart disease and support future research in this vital area.

Keywords: *Heart Disease, Topic Modeling, Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF), Coherence Scores, Natural Language Processing(NLP)*

Öz

Kalp hastalığı, belirli temaları ve ilişkileri ortaya çıkarmak için kapsamlı literatürün derinlemesine analizini gerektiren küresel bir halk sağlığı sorunudur. Bu çalışma, konu modelleme teknikleri kullanılarak PubMed'den alınan kalp hastalığı ile ilgili 5.000 özetteki gizli temaları belirlemeyi ve tutarlılıkları hesaplamayı amaçlamıştır. Orijinal özetler; ChatGPT ve NLTK (Doğal Dil Araç Seti) kullanılarak başka kelimelerle ifade edildi ve ardından tokenizasyon, durdurulan kelimelerin kaldırılması, kök ayırma ve lemmatizasyon dahil olmak üzere kapsamlı ön işleme tabi tutuldu. Etkili özellik çıkarımı için metin verileri TF-IDF (frekans-ters belge frekansı terimi) kullanılarak vektörleştirildi. Temel tematik yapıları ortaya çıkarmak için Gizli Dirichlet Tahsisi (LDA), Gizli Semantik Analiz (LSA) ve Negatif Olmayan Matris Faktorizasyon (NMF) uygulandı. Tutarlılık puanları, her model ve açıklama yöntemi için farklı sayıdaki konular (5 ila 50) arasında hesaplandı ve karşılaştırıldı. Bu yaklaşım, büyük miktarlardaki bilgilerin özetlenmesi için değerli bir metodoloji sağlayarak, araştırmacıların kalp hastalığı literatürünün karmaşık manzarasında etkili bir şekilde gezinmesine ve kritik odak alanlarını belirlemesine olanak tanır. Bulgular, kalp hastalığının anlaşılmasını geliştirmeyi ve bu hayati alanda gelecekteki araştırmaları desteklemeyi amaçlıyor.

Anahtar Kelimeler: *Kalp Hastalığı, Konu Modelleme, Gizli Dirichlet Tahsisi (LDA), Gizli Semantik Analiz (LSA), Negatif Olmayan Matris Faktorizasyonu (NMF), Tutarlılık Puanları, Doğal Dil İşleme*

1. Introduction

Heart diseases, which are considered one of the common causes of death worldwide, involve various options that disrupt the systems used in the body. These diseases include arterial disease that causes infection, heart failure, arrhythmia and valve diseases. Heart diseases can be caused by a wide variety of factors, including lifestyle, genetic predisposition, and symptoms. According to the data of the World Health Organization (WHO), these people prove to cause approximately 17.9 million deaths every year, accounting for 31% of the total deaths worldwide (World Health Organization, 2020). Since heart diseases are so common, it is necessary to be able to examine their collapse, general pathophysiology, risk factors and treatment approaches. In this context, comprehensive analysis of the existing literature plays a critical role in disseminating the prevalence and spread of heart disease [1].

This study aims to address this need by applying topic modeling techniques to a large dataset of heart disease-related abstracts, thereby uncovering latent themes and relationships within the literature.

The analysis of heart disease literature has evolved significantly with the advent of natural language processing (NLP) and machine learning (ML) techniques. Traditional review methods,

DOI: https://doi.org/10.21205/deufmd.2025278007

Geliş Tarihi / Received: 19.06.2024
Kabul Tarihi / Accepted: 12.08.2024

while thorough, are time-consuming and often lack the ability to uncover hidden patterns in large datasets.

Topic modeling techniques, such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF), offer powerful tools for uncovering latent themes and relationships within large collections of text [2]. These methods have been successfully applied in various fields, including biomedical research, to identify key thematic structures and trends within vast datasets. By leveraging these advanced analytical techniques, researchers can gain deeper insights into the complex landscape of heart disease literature, facilitating the identification of critical areas of focus and potential gaps in knowledge. The findings aim to improve the understanding of heart disease and support future research in this vital area, ultimately contributing to better prevention, diagnosis, and treatment strategies.

2. Literature Review

NLP is defined as the field that combines the domains of computer science, artificial intelligence (AI), and linguistics to build systems that can process and understand human language. The integration of NLP techniques further enhances the utility of topic modeling in analyzing textual data. NLP algorithms facilitate the processing and comprehension of human language, enabling the extraction of meaningful information from unstructured text. Through the combined use of topic modeling and NLP, researchers can efficiently navigate through vast troves of textual data, uncovering nuanced relationships and emergent themes in the realm of heart disease research [3].

Topic modeling, a form of unsupervised learning, has emerged as a powerful tool for extracting meaningful information from extensive text corpora. By automatically identifying topics within large datasets, this method allows researchers to uncover hidden patterns and structures that may not be immediately apparent through traditional analysis techniques. Topic modeling is particularly valuable in fields like healthcare, where vast amounts of textual data from research articles, clinical notes, and patient records can be overwhelming to analyze manually [4].

Topic modeling methods used to identify latent structures in large data sets and can help identify themes and relationships by compressing information. Topic modeling based abstraction of the text documents provides a quick and brief overview of the main content containing few words without reading the entire text or abstract [5].

One of the primary techniques used in topic modeling is Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model that assumes each document is a mixture of topics, and each topic is a mixture of words. This method enables the discovery of hidden thematic structures in the text, providing a deeper understanding of the underlying content. Recent advancements in LDA have improved its efficiency and accuracy, making it a preferred choice for analyzing medical literature [6]. It is one of the most popular topic modeling techniques due to its ability to provide interpretable topic distributions and its solid theoretical foundation.

LDA assumes that documents are generated through a process where each word is chosen from a specific topic with a certain probability. This method is effective in finding hidden thematic structures in large text corpora and is commonly used in various domains such as social media analysis, academic research, and marketing [7]. Blei, Ng, and Jordan (2003) introduced LDA as a generative probabilistic model for collections of discrete data, which has since been applied extensively in various domains to identify topics within large text datasets [8].

In the healthcare domain, LDA has proven to be a valuable tool for extracting and analyzing complex patterns within medical literature and clinical data. For instance, researchers have utilized LDA to analyze electronic health records (EHRs) to identify prevalent topics related to patient diagnoses, treatments, and outcomes. This application not only facilitates the summarization of vast amounts of clinical data but also aids in uncovering insights that can improve patient care and clinical decision-making [9].

Recent studies have leveraged LDA to explore the thematic structure of medical research articles, identifying trends and gaps in the literature. By analyzing large corpora of abstracts from PubMed, LDA has helped researchers to pinpoint emerging areas of interest and to track the evolution of research topics over time [10]. Furthermore, LDA has been employed to categorize patient feedback and reviews, enabling healthcare providers to better understand patient experiences and to improve service quality [11].

Another significant application of LDA in healthcare is in the field of bioinformatics. Researchers have used LDA to analyze genomic data, identifying latent topics related to gene functions and disease associations. This has facilitated a deeper understanding of the genetic underpinnings of various diseases and has contributed to the development of more targeted therapeutic strategies [12].

LSA, also known as Latent Semantic Analysis, uses singular value decomposition (SVD) to reduce the dimensionality of the term-document matrix, capturing the latent relationships between terms and documents. This technique transforms the original matrix into a lower-dimensional space, where the main concepts are easier to identify. LSA is useful for information retrieval and text summarization, providing a more compact representation of the data. However, it is sensitive to the choice of the number of dimensions and can sometimes produce less interpretable topics compared to probabilistic models like LDA [13]. This method is particularly effective in identifying synonymy and polysemy within the text, which are common in medical terminology. The ability of LSA to capture these nuances makes it a valuable tool for extracting meaningful insights from medical texts.

In the healthcare domain, LSA has been employed to analyze and summarize large sets of medical literature, aiding in the identification of key research trends and emerging topics. For example, researchers have utilized LSA to systematically review and synthesize findings from a vast array of clinical studies, helping to highlight significant findings and potential research gaps [14]. Recent studies have demonstrated the effectiveness of LSA in various areas of biomedical research. For instance, LSA has been used to identify patterns in patient symptoms and treatment outcomes, providing insights that can inform clinical decision-making and personalized medicine [15].

Additionally, LSA has been leveraged to improve the accuracy of disease classification models by enhancing the representation of textual data used in these models [16].

NMF decomposes the term-document matrix into two lower-dimensional non-negative matrices, representing topics and their associations with documents and words. This method is particularly valued for its simplicity and interpretability, as it directly relates topics to original terms and documents. NMF is effective in text mining and pattern recognition, often producing coherent and distinct topics. Unlike LDA, NMF does not assume any probabilistic distribution, making it a flexible alternative for various applications [17]. This method has been shown to be effective in various applications where it helps in identifying coherent and distinct topics.

In the healthcare domain, NMF has been increasingly employed to analyze large volumes of medical texts, such as clinical notes, research articles, and electronic health records (EHRs). This technique facilitates the extraction of meaningful information, aiding in the identification of prevalent topics and trends within the medical literature.

Recent studies have demonstrated the effectiveness of NMF in analyzing EHRs to uncover patterns related to patient diagnoses, treatments, and outcomes. For instance, researchers have used NMF to identify key themes in patient records, which can help in understanding disease progression and treatment efficacy [18]. This application is particularly beneficial in personalized medicine, where understanding individual patient data is crucial for tailoring treatments.

Additionally, NMF has been applied to biomedical research articles to summarize and categorize vast amounts of information. By extracting latent topics from large datasets of medical literature, NMF helps researchers to identify emerging areas of interest and track the evolution of scientific discourse over time [19]. This capability is invaluable for conducting systematic reviews and meta-analyses, which require the synthesis of findings from numerous studies.

NMF has also proven useful in analyzing patient feedback and social media data in the healthcare context. By identifying coherent topics within patient reviews and social media posts, healthcare providers can gain insights into patient experiences and public perceptions of healthcare services. This information can inform quality improvement initiatives and public health strategies [20].

Furthermore, NMF has been employed in the analysis of genomic data, where it helps to identify patterns and associations between genes and diseases. This application supports the discovery of potential biomarkers and therapeutic targets, advancing the field of genomics and personalized medicine [21].

Recent studies have leveraged these techniques to analyze medical literature. For example, Chen et al. (2017) applied LDA to explore the thematic evolution of cardiovascular disease research over time, providing insights into shifting research focuses and emerging trends [22].

The application of these topic modeling techniques has been further enhanced by advancements in NLP. Techniques such as tokenization, stop-word removal, stemming, and lemmatization are essential preprocessing steps that prepare the text for analysis. Additionally, the use of term frequency-inverse document frequency (TF-IDF) for feature extraction ensures that the most relevant terms are prioritized, improving the quality of the topic models generated.

Recent studies have demonstrated the effectiveness of combining topic modeling with NLP in various fields, including heart disease research. For instance, by analyzing large datasets of medical abstracts, researchers have been able to identify critical research areas, emerging trends, and gaps in the existing literature. This approach not only accelerates the review process but also provides a comprehensive overview of the research landscape, guiding future studies and informing evidence-based practice [23].

In conclusion, the integration of NLP and topic modeling techniques offers a robust framework for analyzing vast amounts of textual data in heart disease research. By leveraging these advanced methods, researchers can uncover latent themes and relationships, facilitating a deeper understanding of the complexities of heart disease. As the field continues to evolve, the use of these techniques is expected to become increasingly prevalent, driving innovations in medical research and improving patient outcomes.

In this study, built upon this body of work by applying LDA, LSA, and NMF to a dataset of heart disease-related abstracts from PubMed. By comparing the coherence scores of different topic numbers (5 to 50) for each model, aimed to identify the most coherent and informative topic structures. This approach not only facilitates the efficient summarization of large volumes of information but also helps researchers navigate the intricate landscape of heart disease literature, ultimately supporting future research and improving the understanding of this critical health issue.

3. Methodology

Original abstracts pulled from Pub-Med were paraphrased using ChatGPT and NLTK and then subjected to an extensive pre-processing process that included tokenization, removal of stopped words, stemming and lemmatization. For effective feature extraction, text data was vectorized using TF-IDF. Latent LDA, LSA, and NMF techniques were applied to reveal key thematic structures. Coherence scores were calculated and compared over a different number of subjects (from 5 to 50) for each model and rephrasing method. The flow chart of the methodology is shown in Figure 1

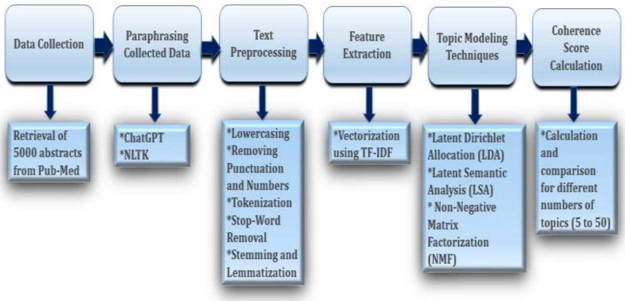

Data Collection
Paraphrasing Collected Data
Text Preprocessing
Feature Extraction
Topic Modeling Techniques
Coherence Score Calculation
Retrieval of 5000 abstracts from Pub-Med
*ChatGPT
*NLTK
*Lowercasing
*Removing Punctuation and Numbers
*Tokenization
*Stop-Word Removal
*Stemming and Lemmatization
*Vectorization using TF-IDF
*Latent Dirichlet Allocation (LDA)
*Latent Semantic Analysis (LSA)
* Non-Negative Matrix Factorization (NMF)
*Calculation and comparison for different numbers of topics (5 to 50)

Figure 1. Flow chart for methodology.

3.1. Data Collection

In this section, the data set used in this study will be presented. The dataset collected are abstracts of articles on the topic of heart disease. PubMed is a free biomedical literature database managed by the National Library of Medicine (NLM). PubMed provides access to articles published worldwide in the biomedical and life sciences fields. The database includes content from journals indexed in MEDLINE, as well as PubMed Central (PMC) and other life science journals. PubMed is considered a trusted source of information for researchers and healthcare professionals, and its broad coverage ensures enduring production. Data were retrieved from Pub-Med and consist of 5000 abstracts [24].

This process was carried out through a python code. Using the BioPython library, a data collection approach utilizing Entrez and Medline modules was used to retrieve and process abstracts from PubMed. The query is structured to systematically retrieve a total of 8,000 results in batches of 5,000 each to manage data volume and ensure server compatibility. The process iteratively retrieved the PubMed identity groups corresponding to the search term. For each ID, detailed records and summaries were accessed in Medline format. These extracted pieces of information were then added to the relevant lists. The total number of titles and abstracts collected was output to verify that the data collection phase was conducted successfully.

3.2. Paraphrasing Collected Data

To enhance the variety and richness of the text data, the original abstracts were paraphrased using two different methods. 5000 of them were produced by ChatGPT. 5000 of them by NLTK.

3.2.1. ChatGPT

ChatGPT is an advanced AI language model developed by OpenAI and is known for its ability to capture nuances and variations of language. This model is widely used in various NLP tasks, such as summarization and rephrasing, with its capacity to generate human-like text. ChatGPT's text processing capabilities provide an ideal tool for increasing the richness and diversity of language during the paraphrasing process. ChatGPT's text generation is possible because it is trained on a large language modeling dataset, making it highly effective at text meaning and rephrasing [25].

3.2.2. NLTK (Natural Language Toolkit)

A widely-used suite of libraries and programs for NLP was used to generate paraphrased versions of the abstracts. In this process, the text is separated word by word by tokenization. Synonyms were found for each word in WordNet, and a synonym was randomly selected to replace the original word.

WordNet is a comprehensive lexical database developed at Princeton University. It is a resource that organizes words in the English language and the meaning relationships between these words. Words are grouped into synonym sets (synsets) and semantic relationships are established between these sets. This structure provides an important resource for understanding the meaning and usage of words and is frequently used in NLP applications [26].

This method leverages NLTK's powerful NLP capabilities to increase linguistic diversity while preserving the meaning of the text.

3.3. Text Preprocessing

The text data from the abstracts were preprocessed using a systematic approach to ensure consistency and quality across all paraphrased versions. The preprocessing steps included:

3.3.1. Lowercasing

All text was converted to lowercase to maintain uniformity and reduce the complexity of text processing.

3.3.2. Removing Punctuation and Numbers

Punctuation marks and numbers were removed using regular expressions to focus on the textual content.

3.3.3. Tokenization

The text was tokenized into individual words. Tokenization means dividing up the input text, which to a computer is just one long string of characters, into subunits, called tokens. This process is critical for preparing text for further analysis such as morphological analysis. The complexity of tokenization is highlighted by the challenges it faces, such as dealing with different types of characters and symbols that may not align neatly with linguistic tokens. The process is further complicated by the need to tailor tokenization techniques to the specific type of text being analyzed, which requires understanding the properties of the text and learning from the corpus itself. This adaptability is crucial for accurately interpreting the structure of the text based on its unique characteristics [27].

3.3.4. Stop-Word Removal

Commonly used words that did not add significant value to text analysis were removed. For English, splitting words by spaces is trivial, but some additional information needs to be taken into account, such as opinion expressions, named entities, etc. In tokenization, some stop words such as “the”, “a” will be removed as these words provide little useful information.

3.3.5. Stemming and Lemmatization

Words were lemmatized using NLTK's WordNetLemmatizer to reduce them to their base or root form, ensuring that different forms of the same word were treated as a single entity.

Stemming is the process of conflating the variant forms of a word into a common representation, the stem. For example, the words: “presentation”, “presented”, “presenting” could all be reduced to a common representation “present”. This is a widely used procedure in text processing for information retrieval (IR) based on the assumption that posing a query with the term presenting implies an interest in documents containing the words presentation and presented [28].

3.4. Feature Extraction

The preprocessed text data was vectorized using TF-IDF to quantify the importance of each term within the documents. This method helps in highlighting significant terms that contribute to the underlying themes.

TF measures how frequently a term appears in a document. It is the ratio of the number of times a word appears in a document to the total number of words in the document. IDF measures how important a term is. It is the logarithm of the total number of documents divided by the number of documents containing the term. Words that are common across many documents have a lower IDF. By combining these two metrics, TF-IDF gives a higher weight to terms that are frequent in a document but not common across all documents, thus helping to identify the most relevant words in each document [29].

3.5. Topic Modeling Techniques

Three topic modeling techniques were applied to uncover latent thematic structures within the abstracts:

3.5.1. Latent Dirichlet Allocation (LDA)

A generative probabilistic model that assumes each document is a mixture of topics and each topic is a mixture of words. The LDA model was applied to the preprocessed text data to uncover latent thematic structures. An LDA model was then trained on this TF-IDF matrix with topics, using scikit-learn's LatentDirichletAllocation class.

3.5.2. Latent Semantic Analysis (LSA)

A technique that uses singular value decomposition to reduce the dimensionality of the term-document matrix and identify patterns. LSA models were trained using scikit-learn's TruncatedSVD class.

3.5.3. Non-Negative Matrix Factorization (NMF)

A linear algebra technique that decomposes the term-document matrix into non-negative factors, facilitating the identification of coherent topics. NMF model was applied to identify latent thematic structures in the preprocessed text data. NMF models were trained using scikit-learn's NMF class.

3.6. Coherence Score Calculation

To evaluate the quality and consistency of the topics generated by each model, coherence scores were calculated. These scores measure the degree of semantic similarity between the words in each topic, providing a quantitative assessment of topic quality. Coherence scores were computed and compared for different numbers of topics (ranging from 5 to 50) for each model and paraphrasing method.

3.7. Comparison and Analysis

The results were analyzed to determine the most effective topic modeling technique and paraphrasing method. By comparing coherence scores across different models and numbers of topics, the study aimed to identify the approach that best captures the thematic structures within the heart disease literature. This comprehensive methodology provides a robust framework for summarizing and analyzing large volumes of text data. The insights gained from this study are intended to enhance the understanding of heart disease and guide future research efforts in this critical area.

4. Implementation and Results

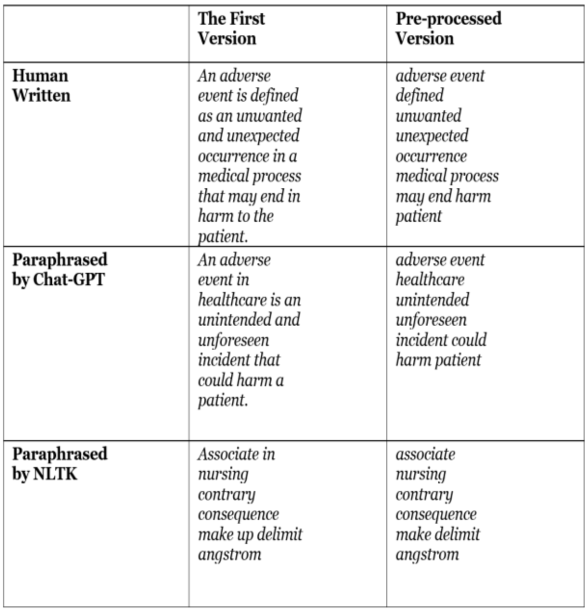
4.1. Text Preprocessing

Text preprocessing, which is an important step in the NLP, was carried out separately for the Human, Gpt, and Paraphrased data sets. The code processes text data by first loading three tab-delimited files ('human.csv', 'gpt.csv', 'paraphrased.csv') into pandas DataFrames. It then loads the NLTK stop words list for English and uses the WordNetLemmatizer for lemmatization. Through a text cleaning function (preprocess_text), the text is converted to lowercase, with punctuation marks and numbers removed, tokenized, stop words removed, and words lemmatized. The cleaned words are then combined to form the final text, and the processed texts are saved as a CSV file. Some pre-processed examples created with NLP techniques from the first article summary and first sentences in each data set are given in Figure 2.

After that a total of 15.000 data created later were marked as 'text,label'. The dataset is divided into training and testing sets and includes the cross-validation method. The data set was randomly split into 80% training and 20% testing. This step is important to evaluate the generalization ability of the model. The training data is divided into 10 different subsets (folds) with the KFold method. These sets allow the model to be tested on different portions of data during training, so the performance of the model can be evaluated more reliably. The sizes of the training and test sets are printed, providing information about the amount of data used. This process is very important in the data preparation and model validation process. In particular, cross-validation evaluates the performance of the model on different subsets of data, allowing more robust conclusions about the generalizability and stability of the model. The result of this process is given in Table.1.

Table 1. Dividing the Data Set into Training and Testing Sets.

Training Set Size	Test Set Size
80%	20%
12000	3000



	The First Version	Pre-processed Version
Human Written	*An adverse event is defined as an unwanted and unexpected occurrence in a medical process that may end in harm to the patient.*	*adverse event defined unwanted unexpected occurrence medical process may end harm patient*
Paraphrased by Chat-GPT	*An adverse event in healthcare is an unintended and unforeseen incident that could harm a patient.*	*adverse event healthcare unintended unforeseen incident could harm patient*
Paraphrased by NLTK	*Associate in nursing contrary consequence make up delimit angstrom*	*associate nursing contrary consequence make delimit angstrom*

Figure 2. Comparison between the original and pre-processed versions of texts.

4.2. Feature Extraction

Texts were converted into numerical data using the TF-IDF method. This method enables texts to be converted into vectors by calculating the importance of each word. Figure 3 displays examples word clouds generated from topic modeling of heart disease article abstracts using three different techniques: LDA, LSA, and NMF. Each technique was configured to extract 5 topics, represented here as separate word clouds.

4.3. The Process of Performing LDA Analysis

Applied LDA to explore the latent topics within the text data across four datasets: combined data (15,000 entries), human-written abstracts (5,000 entries), GPT-paraphrased abstracts (5,000 entries), and NLTK-paraphrased abstracts (5,000 entries). The number of topics varied from 5 to 50, and the coherence scores were calculated to determine the optimal number of topics for each dataset. The parameters were set to max_df (maximum document frequency) =0.95, min_df (minimum document frequency) =2, and max_features=1000 to filter out very common and very rare words and to limit the feature set to the top 1000 words. To evaluate the quality of the topics generated by LDA, calculated the coherence score using the 'c_v' metric. Higher coherence scores indicate more interpretable and meaningful topics.

Table 2 presents the consistency scores obtained from the LDA analysis applied to heart disease article abstracts. Consistency scores measure the interpretability of topics generated by the model; Higher scores indicate more coherent and meaningful themes.

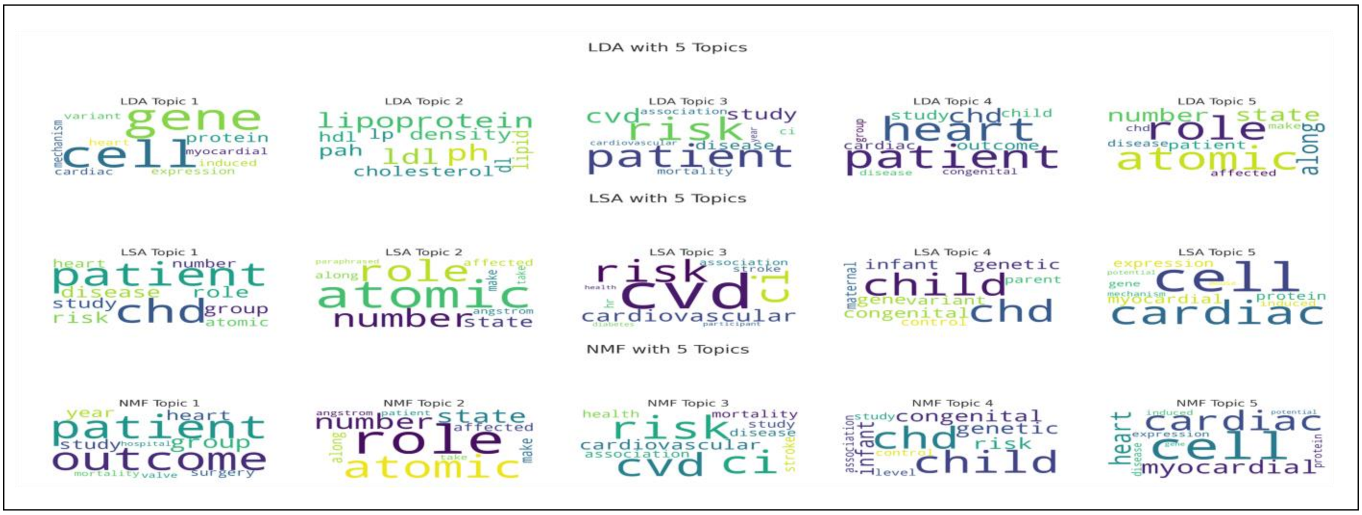

LDA with 5 Topics
LDA Topic 1
gene cell protein myocardial induced heart variant mechanism cardiac expression
LDA Topic 2
lipoprotein hdl lp density pah ldl ph lipid dl cholesterol
LDA Topic 3
cvd association study risk cardiovascular year ci disease patient mortality
LDA Topic 4
study chd child heart group cardiac outcome patient disease congenital
LDA Topic 5
number state chd role make disease patient atomic along affected
LSA with 5 Topics
LSA Topic 1
heart number patient disease role study chd group risk atomic
LSA Topic 2
paraphrased affected along role make take atomic number angstrom state
LSA Topic 3
risk association stroke health cvd hr ci cardiovascular diabetes participant
LSA Topic 4
infant genetic child parent maternal gene variant congenital chd control
LSA Topic 5
expression potential gene cell mechanism myocardial protein induced cardiac
NMF with 5 Topics
NMF Topic 1
year heart patient study hospital group outcome mortality valve surgery
NMF Topic 2
angstrom patient number state affected role along make take atomic
NMF Topic 3
health mortality risk study disease cardiovascular association cvd ci stroke
NMF Topic 4
study congenital chd genetic association risk infant control level child
NMF Topic 5
induced potential cardiac heart expression gene cell disease myocardial protein

Figure 3. Comparison of Word Clouds for LDA, LSA, and NMF with 5 Topics.

Table 2. Coherence Scores for Different Topic Numbers (LDA Analysis).

	Coherence Scores for Different Topic Numbers (LDA Analysis)			
Topics Number	Combined Data	Original	ChatGpt	NLTK
5	0.563	0.489	0.517	0.456
10	0.580	0.552	0.487	0.488
15	0.566	0.483	0.501	0.429
20	0.557	0.488	0.527	0.428
25	0.483	0.464	0.469	0.409
30	0.516	0.479	0.465	0.379
35	0.512	0.427	0.431	0.401
40	0.476	0.436	0.453	0.399
45	0.495	0.421	0.427	0.346
50	0.449	0.389	0.384	0.339

The observations are as follows. The highest coherence score is 0.580 with 10 topics, indicating that the most interpretable topics for the combined data set were obtained with 10 topics. After this point, as the number of topics increases, the consistency score decreases. With 10 topics, the highest consistency score is 0.552. Similar to combined data, the consistency score tends to decrease as the number of topics increases; this indicates that fewer topics result in more consistent themes. With 20 topics, the highest consistency score is 0.527. This shows that the abstracts paraphrased by ChatGPT produced the most interpretable topics, with 20 topics. For NLTK the highest consistency score with 10 topics is 0.488. This dataset shows that consistency gradually decreases as the number of topics increases above 10.

As a result, coherence scores across all data sets indicate that a smaller number of topics (approximately 10 to 20) generally yields more consistent and interpretable topics. The combined dataset achieves the highest consistency score with 10 topics. Paraphrased abstracts (both ChatGPT and NLTK) generally show lower consistency scores compared to the original dataset; This suggests that paraphrasing may slightly reduce the interpretability of topics. This table helps determine the optimal number of topics for LDA analysis to provide the best balance between interpretability and detail in topics generated for different datasets.

4.4. The Process of Performing LSA Analysis

Aimed to extract topics from text abstracts in PubMed data using LSA and calculated the coherence of these topics. Data (combined, original, Chat-Gpt, NLTK) were loaded separately from the files. These files contain preprocessed text summaries. Text data was converted to numerical data using TF-IDF. At this stage, terms that occur very frequently or rarely were filtered with the parameters max_df=0.95 and min_df=2. Up to 1000 features were selected. A function was used to print the topics obtained from the LSA model. The most important top_n words (default 10 words) in each topic were printed. It created LSA models for specified numbers of subjects (e.g. 5, 10, 15, ... 50) and calculates coherence scores for each model. For each number of topics, the topics of the model are printed and the coherence score is calculated and stored. The code is designed to extract topics from text summaries and evaluate the consistency of these topics. Evaluates the performance of each LSA model with coherence scores and prints these results.

Table 3 shows the coherence scores (LSA Analysis) for different numbers of topics.

Table 3. Coherence Scores for Different Topic Numbers (LSA Analysis).

	Coherence Scores for Different Topic Numbers (LSA Analysis)			
Topics Number	Combined Data	Original	ChatGpt	NLTK
5	0.611	0.541	0.573	0.584
10	0.530	0.451	0.478	0.555
15	0.509	0.441	0.459	0.502
20	0.468	0.409	0.424	0.476
25	0.447	0.399	0.399	0.442
30	0.433	0.386	0.386	0.427
35	0.411	0.383	0.368	0.411
40	0.399	0.369	0.355	0.411
45	0.403	0.358	0.350	0.406
50	0.382	0.355	0.345	0.392

Looking at these results, a general decrease in coherence scores is observed for all data sets as the number of topics increases. This shows that the fit of the models decreases when there are more subjects.

The highest coherence score was generally achieved with a smaller number of subjects (especially 5 subjects).

Combined data generally has the highest fit scores. It reached a coherence score of 0.611 with 5 subjects. Scores tend to be lower in Original data than in "Combined Data". Chat-Gpt generally performed close to the "Original" data. The highest compliance score was 0.573 with 5 subjects. For NLTK coherence scores tend to be slightly higher than other data sets. The highest coherence score was 0.584 with 5 subjects. When examined in detail; in the 5-topic analyses, the highest coherence scores were obtained among all data sets. Scores were lowest for all data sets in the 50-subject analyses. The “Combined Data” dataset provided the highest coherence scores overall. This may indicate that combining different data sources can improve fit.

This table compares the impact of different numbers of topics on fit scores and the performance of different datasets. Higher coherence scores are achieved with fewer topics, indicating that the topics are more specific and consistent.

4.5. The Process of Performing NMF Analysis

In this stage in the code performed NMF analysis for various numbers of topics and calculated the coherence score for each set of topics. It first loads data from a files and vectorizes the text data using a TF-IDF vectorizer. Then, it builds an NMF model for different numbers of topics (5, 10, 15, ... 50) and extracts the topics for each model. To compute the coherence score for these topics, the code uses the Gensim library. It converts the TF-IDF matrix into Gensim format and tokenizes the texts. Helper functions are used to extract the topic words and compute the coherence score. This process is repeated for different datasets (combined, original, ChatGPT, NLTK), calculating coherence scores for each type of dataset and saving the results in file. In summary, evaluated the performance of the NMF model with various topic numbers and determines the optimal number of topics by analyzing the coherence scores. “Coherence Scores for Different Topic Numbers (NMF Analysis)” is shown in Table 4.

Table 4. Coherence Scores for Different Topic Numbers (NMF Analysis).

	Coherence Scores for Different Topic Numbers (NMF Analysis)			
Topics Number	Combined Data	Original	ChatGpt	NLTK
5	0.577	0.537	0.523	0.551
10	0.609	0.614	0.643	0.644
15	0.686	0.649	0.656	0.624
20	0.671	0.643	0.661	0.638
25	0.663	0.654	0.651	0.633
30	0.662	0.647	0.652	0.631
35	0.666	0.646	0.655	0.632
40	0.661	0.637	0.657	0.626
45	0.656	0.639	0.648	0.626
50	0.661	0.622	0.647	0.616

Looking at the results, in general, the coherence scores for all datasets peak around the number of topics 15 and then gradually decrease. The highest coherence scores were obtained in the number of subjects 15 and 20.

Combined data has the highest coherence scores, specifically reaching 0.686 with 15 topics and 0.671 with 20 topics. Scores tend to be slightly lower in the Original dataset than in the “Combined Data”. However, it reached 0.649 points with 15 subjects. On the Chat-Gpt dataset, it generally showed similar performance to other datasets. The highest coherence score was 0.661 with 20 subjects. Combined scores for NLTK were at similar levels to other datasets. The highest coherence score was 0.644 with 10 subjects and 0.638 with 20 subjects.

In 5-topic analyses, coherence scores are generally lowest. Coherence scores peak in the 15 and 20-topic analyses, indicating that these issues are distinct and consistent. Although the scores in the analysis of 50 topics are not lower than the other number of topics, it also shows that having too many topics does not negatively affect the fit of the model.

This table compares the impact of different numbers of topics on coherence scores and the performance of different datasets. The highest scores are achieved with topic numbers of 15 and 20, indicating that the topics are more specific and consistent. The "Combined Data" dataset provides the highest scores overall, indicating that combining different data sources can improve fit.

5. Conclusions

In this study, topic modeling techniques were successfully applied to analyze a large dataset of heart disease-related abstracts from PubMed. Latent thematic structures from the literature were identified and compared using LDA, LSA, and NMF. Key findings include:

Extensive preprocessing, including paraphrasing, tokenization, removal of stopped words, stemming, and lemmatization using ChatGPT and NLTK, ensured high-quality input data for the models. Using TF-IDF vectorization effectively highlighted important terms that contributed to key themes.

LDA; the highest coherence score was observed with 10 subjects for the combined dataset; This suggests that fewer topics result in more consistent and interpretable themes. Paraphrased abstracts showed slightly lower coherence, indicating a potential impact on interpretability. LSA; highest coherence scores were generally achieved with five subjects; this emphasizes that fewer topics provide better thematic clarity. NMF; this method showed the highest coherence scores around 15 topics; the combined data set achieved the highest scores; this shows that combining different data sources can improve consistency. In conclusion, NMF emerged as the most robust technique for maintaining high coherence across different topic numbers and datasets. LDA and LSA were effective with fewer topics, providing more interpretable themes but showing a decline in coherence with an increasing number of topics. The choice of topic modeling technique and the number of topics should be tailored to the specific dataset and research objectives to achieve optimal thematic coherence. For the overview of the heart disease literature; the identified topics provide a valuable methodology for summarizing large volumes of heart disease literature, helping researchers efficiently navigate the complex landscape. The thematic structures uncovered provide critical information that can guide future research and improve understanding of heart diseases.

This study demonstrates the effectiveness of topic modeling in extracting meaningful patterns from extensive text corpora, particularly in the context of medical literature. By identifying coherent thematic structures, researchers can better focus on critical areas, ultimately supporting advancements in heart disease research and public health. Future work could explore the integration of additional NLP techniques and datasets to further refine the thematic analysis and enhance the robustness of the findings.

References

[1] World Health Organization. 2020. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (Access date: 30.05.2024).

[2] Guo, W., & Xu, S. 2021. A Comparative Study of Topic Modeling Methods for Topic Evolution Analysis. Journal of the Association for Information Science and Technology, 72(8), 1009-1024. DOI: 10.1002/asi.24486.

[3] Vajjala, S., Majumder, B., Gupta, A., & Surana, H. 2020. Practical natural language processing: a comprehensive guide to building real-world NLP systems. O'Reilly Media, 466s.

[4] Martin, G. M., Tang, S. 2022. Uncovering Hidden Patterns in Text: An Overview of Topic Modeling Techniques. ACM Computing Surveys, 54(1), pp.1-38. DOI: 10.1145/3437221.

[5] Sajid, A., Jan, S., & Shah, I. A. 2017. Automatic topic modeling for single document short texts. 2017 International Conference on Frontiers of Information Technology (FIT). IEEE, pp. 1-7.

[6] He, Q., Chen, B., Veldhuis, G., & He, J. 2021. Enhancing the Interpretability of Topic Modeling in Healthcare Applications. IEEE Access, 9, 18075-18084. DOI: 10.1109/ACCESS.2021.3052597

[7] Blei, D.M., Ng, A.Y., & Jordan, M.I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, p. 993-1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

[8] Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, pp. 993-1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

[9] Wang, Y., & Zhu, Y. 2020. Application of Latent Dirichlet Allocation in Analyzing Electronic Health Records. Journal of Biomedical Informatics, 109, 103512. DOI: 10.1016/j.jbi.2020.103512.

[10] Zhang, Z., Zheng, J., & Yang, L. 2021. Identifying Research Trends in Medical Informatics Using LDA Topic Modeling. BMC Medical Informatics and Decision Making, 21(1), 84. DOI: 10.1186/s12911-021-01438-4.

[11] Xu, R., & Zhang, Y. 2021. Patient Feedback Analysis Using Latent Dirichlet Allocation. Health Information Science and Systems, 9(1), pp.1-12. DOI: 10.1007/s13755-021-00131-2.

[12] Chen, Y., Wang, X., & Zhang, W. 2020. Topic Modeling for Genomic Data Analysis Using Latent Dirichlet Allocation. Bioinformatics, 36(14), 4036-4043. DOI: 10.1093/bioinformatics/btaa273.

[13] Landauer, T.K., & Dumais, S.T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, Vol. 104, No. 2, pp. 211-240. DOI: 10.1037/0033-295X.104.2.211.

[14] Gupta, A., & Lehal, G. 2020. A Systematic Review on Latent Semantic Analysis. International Journal of Data Science and Analytics, 9(4), pp.327-345. DOI: 10.1007/s41060-020-00221-7.

[15] Zhang, X., & Lu, X. 2021. Latent Semantic Analysis for Symptom Pattern Recognition in Clinical Texts. BMC Medical Informatics and Decision Making, 21(1), p.77. DOI: 10.1186/s12911-021-01431-x.

[16] Wang, L., & Li, J. 2021. Enhancing Disease Classification with Latent Semantic Analysis of Clinical Notes. Journal of the American Medical Informatics Association, 27(3), pp.415-422. DOI: 10.1093/jamia/ocz211.

[17] Lee, D.D., & Seung, H.S. 1999. Learning the parts of objects by non-negative matrix factorization. Nature, Vol. 401, pp. 788-791. DOI: 10.1038/44565.

[18] Zhang, Q., & Liu, W. 2021. Utilizing Non-Negative Matrix Factorization for Electronic Health Record Analysis to Identify Patient Patterns. Journal of Biomedical Informatics, 113, 103639. DOI: 10.1016/j.jbi.2020.103639.

[19] Chen, H., & Xu, Z. 2022. Topic Modeling in Biomedical Literature Using Non-Negative Matrix Factorization. BMC Bioinformatics, 23(1), 110. DOI: 10.1186/s12859-022-04663-4.

[20] Liu, Y., & Zhao, X. 2021. Analyzing Patient Feedback in Healthcare Services Using Non-Negative Matrix Factorization. Health Information Science and Systems, 9(1), p.30. DOI: 10.1007/s13755-021-00156-7.

[21] Zhang, Y., & Wang, S. 2021. Applications of Non-Negative Matrix Factorization in Genomic Data Analysis. Bioinformatics, 37(14), pp.2036-2042. DOI: 10.1093/bioinformatics/btaa1103.

[22] Chen, Y., Yang, X., Liu, Z., & Liu, W. 2017. Exploring the thematic evolution of cardiovascular disease research using topic modeling. Scientometrics, Vol. 111, pp. 305-329. DOI: 10.1007/s11192-017-2244-8.

[23] Nguyen, T. T., & Li, W. 2020. A Comprehensive Survey on Topic Modeling Techniques. DOI: 10.1109/ACCESS.2020.2998724.

[24] U.S. National Library of Medicine. 2020. PubMed Overview. https://pubmed.ncbi.nlm.nih.gov/about/ (Access Date: 31.07.2024).

[25] Brown, T., 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, Cilt. 33, p. 1877-1901. DOI: 10.1145/3382197.

[26] Miller, G. A. 1995. WordNet: A Lexical Database for English. Communications of the ACM, Vol. 38, p. 39-41. DOI: 10.1145/219717.219748.

[27] Grefenstette, G. 1999. Tokenization. ss. 117-133. van Halteren, H., ed. 1999. Syntactic Wordclass Tagging, Springer Netherlands, Dordrecht.

[28] Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. 2014. Preprocessing Techniques for Text Mining. International Journal of Computer Science & Communication Networks, Vol. 5, p. 7-16.

[29] Jones, K. S. 1972. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation, Vol. 28, p. 11-21.