Topic Modeling in Cardiovascular Research Publications

Edgar Navarro
3dgarnavarro@gmail.com
San Diego State University
San Diego, California, USA

Hajar Homayouni
hhomayouni@sdsu.edu
San Diego State University
San Diego, California, USA

ABSTRACT

Topic modeling plays a crucial role in cardiovascular health research as it enables the identification of key themes and subtopics within the domain, facilitating knowledge discovery and focused research. In this study, we developed a topic modeling pipeline for cardiovascular health research publications using BERT, UMAP, and clustering algorithms. We first used BERT to generate embeddings for each document, which captured the semantic meaning of the words used in the publications. We then used UMAP to reduce the dimensionality of the embeddings and create a low-dimensional representation of the data. Finally, we applied multiple clustering algorithms (HBDSCAN, K-means) to cluster the documents into topics based on their low-dimensional representations. We also evaluated these results with traditional modeling methods such as LDA. Our evaluation showed that the pipeline with HBDSCAN as the clustering algorithm outperformed the other pipeline configuration with k-means while also outperforming the LDA model. Our findings suggest that this pipeline can be a valuable tool for researchers and practitioners in the field of cardiovascular health to identify key topics and trends in medical research publications.

CCS CONCEPTS

• **Computing methodologies** → *Information extraction.*

KEYWORDS

Topic Modeling, Transformers, Clustering

ACM Reference Format:
Edgar Navarro and Hajar Homayouni. 2023. Topic Modeling in Cardiovascular Research Publications. . ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Cardiovascular disease is a significant global health concern, affecting millions of individuals worldwide and contributing to a substantial burden on healthcare systems. As medical research in this field continues to generate vast amounts of data, it becomes increasingly challenging for researchers to extract meaningful insights and identify key topics from the growing body of literature.

Traditional methods of topic modeling, such as Latent Dirichlet Allocation (LDA) [2] and Non-Negative Matrix Factorization (NMF) [4], may not be optimal for capturing the complexity and nuances present in cardiovascular health research publications. The need for more advanced and specialized topic modeling techniques in the medical domain has led to the exploration of novel approaches that leverage the power of pre-trained language models. One such model is Bidirectional Encoder Representations from Transformers (BERT) [3], which has demonstrated exceptional performance in a wide range of natural language processing tasks. By leveraging its contextual understanding and representation capabilities, we aim to develop a pipeline specifically tailored for cardiovascular health research publications.

The objective of this research is to propose a pipeline combining BERT, Uniform Manifold Approximation and Projection (UMAP) [12], and a clustering algorithm to identify key topics within cardiovascular health research literature. By harnessing the strengths of BERT and incorporating advanced clustering techniques, we aim to improve the accuracy, interpretability, and scalability of topic modeling in this domain. Through our pipeline, we seek to enable researchers to gain valuable insights, uncover hidden relationships, and track emerging trends within the vast landscape of cardiovascular health literature.

We will present the methodology used in our pipeline, discuss the results of our experiments, and evaluate the performance of our approach against other commonly used topic modeling algorithms. Furthermore, we will analyze the interpretability of the identified topics and discuss the implications of our findings. By addressing the limitations of traditional topic modeling methods and leveraging the advancements in pre-trained language models, we hope to contribute to the field of cardiovascular health research and facilitate knowledge discovery in this critical domain.

2 METHODOLOGY

2.1 Preprocessing

The dataset used in this study was obtained from the PubMed library, maintained by the National Library of Medicine (NLM). It consists of publications specifically related to cardiovascular diseases. The collection includes a total of 86,904 documents, presenting a vast spectrum of topics in cardiovascular health. Topics span from molecular aspects such as endothelial cells, to anatomical and clinical studies involving terms like coronary, myocardial, and ventricular. The diversity of this dataset is further reflected in the work of over 227,172 unique authors, demonstrating the wide range of researchers in the cardiovascular health research field. For our analysis, we focused solely on the titles of these publications. To ensure efficient and meaningful topic extraction, the text data underwent preprocessing using Spacy and the NLTK library. This preprocessing phase involved three main steps: tokenization, stopword removal, and lemmatization.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym KDD, August 06–10, 2023, Long Beach, CA
© 2023 Association for Computing Machinery.

Tokenization[19], a crucial step in text preprocessing, refers to the process of breaking down the text into individual words called tokens. By converting raw text into a structured format, tokenization simplifies various tasks such as word frequency counting, n-gram extraction, and part-of-speech tagging, all of which contribute to a more comprehensive understanding of the text. Consequently, tokenization plays an essential role in enabling downstream natural language processing tasks.

Stopword removal[16] entails the elimination of common words such as "the," "and," and "is," which do not contribute significant meaning to the context and may introduce noise during topic modeling. Additionally, domain-specific words such as "cardiac," "cardiovascular," "heart," "disease," "patient," "study," and "risk" were removed. This was due to their frequent occurrence and strong association with the cardiovascular domain, which could potentially dominate the resulting topics and hinder the identification of more specific or nuanced topics.

Lemmatization[9] is a process that involves converting words into their base or dictionary form, known as the lemma. For instance, the words "running," "ran," and "runner" would be lemmatized to the word "run." This step standardizes the text and helps to consolidate similar terms, ensuring a more accurate representation of the underlying topics.

By applying tokenization, stopword removal, and lemmatization, the preprocessing phase effectively reduces noise in the data and sets the stage for efficient topic modeling. These preprocessing steps are essential for reducing noise in the data and allowing the pipeline to focus on extracting relevant and coherent topics.

2.2 BERT for Semantic Understanding and Representation Learning

BERT [3, 10] is a pre-trained language model that captures the semantic meaning of words and phrases within a context. In the context of cardiovascular health research, this model can understand complex medical terminologies and relationships within the text. BERT is designed to capture the semantic meaning of words within their context by considering both the preceding and following words in a sentence.

Traditional language models, such as Word2Vec [13] and GloVe [14], learn representations in a unidirectional manner, either from left to right or right to left. In contrast, BERT learns context-aware representations by simultaneously considering the words that come before and after a given word in a sentence. This bidirectional approach allows BERT to capture a deeper understanding of the relationships between words and their context.

BERT is based on the Transformer architecture, which was introduced in the paper titled "Attention is All You Need" by Vaswani et al.[2017] [18]. The Transformer model incorporates self-attention mechanisms that process input sequences in parallel, facilitating the capture of long-range dependencies in text. The architecture is composed of an encoder stack and a decoder stack, both consisting of multiple layers. However, for its bidirectional representation learning, BERT specifically leverages the encoder stack.

We utilized BERT to convert the text data into a high-dimensional vector representation, this high-dimensional vector captures the semantic meaning of the tokens. The specific BERT model used in this study is "*all-mpnet-base-v2*",[1] chosen for its efficiency and proven performance in various NLP tasks.

2.3 UMAP for Dimensionality Reduction

UMAP is a dimensionality reduction technique that preserves the structure of high-dimensional data while reducing it to a lower-dimensional space. The dimensionality reduction step using UMAP is crucial due to the high-dimensional nature of BERT embeddings. BERT generates embeddings of 768 dimensions for each word, making the analysis and visualization of the data complex. By reducing the dimensions to a manageable level (usually 2 or 3), we can better visualize and comprehend the data, uncovering the underlying structure of the data and the relationships between the topics.

It achieves this through two main steps. Firstly, UMAP constructs a fuzzy graph where each data point becomes a node, and the edges represent the similarities between the points. These similarities are computed based on the euclidean distance between the points, with closer points having stronger connections. This fuzzy graph captures the nuanced nature of relationships by assigning weights to the edges, indicating the strength of the connection between points.

Secondly, UMAP optimizes the embedding, which means finding a suitable lower-dimensional representation of the data that preserves the structure encoded in the fuzzy graph. It does this by minimizing a mathematical objective function known as cross-entropy:

$$CE = -\sum_{i=1}^{N}\sum_{j=1}^{N}(w_{ij} \cdot \log\left(\frac{w_{ij}}{q_{ij}}\right))$$

In this equation, N represents the total number of data points in the data set. The indices i and j denote individual data points, ranging from 1 to N. The weight assigned to the edge connecting data points i and j in the fuzzy graph is denoted by w_{ij}, which reflects the strength of the relationship between the two points. On the other hand, q_{ij} represents the expected probability of the edge between data points i and j in the lower-dimensional space. This probability is estimated based on the distribution of points in the high-dimensional space.

The cross-entropy (CE) measures the discrepancy between the pairwise similarities of the points in the high-dimensional space, as defined by the fuzzy graph, and their counterparts in the lower-dimensional space. By iteratively adjusting the positions of the data points in the lower-dimensional space using stochastic gradient descent (SGD) [7], UMAP aims to minimize this cross-entropy and find an embedding that captures both local and global structures in the data.

In the domain of cardiovascular research, the high-dimensional BERT-generated embeddings can be challenging to analyze and visualize. By applying UMAP to the BERT generated embeddings, we facilitate more effective visualization and analysis, enabling the identification of meaningful clusters and patterns within the data

[1]The BERT model "*all-mpnet-base-v2*" is a multilingual model that has been pre-trained on a large corpus of text data from multiple languages. It has shown strong performance in various natural language processing (NLP) tasks, including text classification, named entity recognition, and sentiment analysis.

set. The UMAP model used in this study was configured with the following parameters:

- n_components=5
- n_neighbors=15
- min_dist=0.0

n_components (int): This parameter determines the number of dimensions in which the data will be represented after the dimensionality reduction. In this case, the value is 5, which means that the high-dimensional BERT embeddings will be reduced to a 5-dimensional representation. A lower-dimensional representation aids in visualization and analysis while preserving the essential structure of the data.

n_neighbors (int): This parameter controls the balance between the local and global structure of the data in the reduced representation. A smaller value will prioritize preserving the local structure, while a larger value will focus on maintaining the global structure. In this study, the value is 15, which means that UMAP will consider 15 nearest neighbors when constructing the reduced representation. This value provides a balance between local and global structure, allowing for meaningful patterns to be revealed in the cardiovascular health research data set.

min_dist (float): This parameter determines the minimum distance between points in the reduced representation. A smaller value will result in a denser clustering of points, while a larger value will lead to more uniformly spaced points. In this case, the value is 0.0, which means that UMAP will allow points to be arbitrarily close to each other in the reduced representation. This configuration encourages the formation of tight clusters, which can help to identify meaningful groups of documents within the cardiovascular health research literature.

3 COMPARATIVE ANALYSIS

We aim to compare the applications of different clustering algorithms in our topic modeling pipeline. Additionally, we compare the use of LDA, a popular topic modeling algorithm, with our pipelines utilizing the aforementioned clustering algorithms.

3.1 Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

HDBSCAN [11] is a density-based clustering algorithm. It determines the number of clusters automatically and can handle clusters with varying densities and shapes. HDBSCAN works by estimating the density of data points and constructing a hierarchy of clusters called a condensed tree. It uses density-reachability and mutual reachability to link data points based on their densities. The algorithm condenses the clusters, merges them based on mutual reachability distances, and identifies stable clusters.

In the context of cardiovascular health research, this algorithm allows for the identification of unique patterns and trends within the data, even in the presence of noise. We used HDBSCAN to cluster the lower-dimensional representations produced by UMAP, ensuring high-quality topic extraction. The HDBSCAN model used in this study was configured with the following parameters:

- min_samples=20
- min_cluster_size=50

min_samples: This parameter determines the minimum number of samples required for a region to be considered a core point. A core point is a point that has at least "min_samples" within a specified distance. If a region has fewer samples than the "min_samples" threshold, it is not considered a core point.

min_cluster_size: This parameter sets the minimum number of samples required for a cluster to be considered valid and significant. HDBSCAN eliminates clusters that have fewer samples than the "min_cluster_size" threshold. Clusters with a size smaller than this value are labeled as noise or outliers.

3.2 K-means

K-means [17] clustering is an iterative algorithm used to divide a data set into K distinct clusters, where K is a predefined number. The algorithm works by minimizing the sum of squared distances between data points and the centroids of their assigned clusters. It starts by randomly selecting K initial cluster centroids, which serve as the centers of the clusters. Each data point is then assigned to the nearest centroid based on a distance metric, typically the Euclidean distance. After the initial assignment, the centroids are updated by recalculating the mean position of the data points assigned to each cluster. This process of assignment and centroid update is repeated iteratively until convergence, where the centroids stabilize, or a stopping criterion is met. The final result is a clustering solution where each data point is assigned to one of the K clusters based on its proximity to the corresponding centroid.

Similarly to LDA, K-means clustering requires a pre-determined number of clusters set as hyper-parameter as well. K-fold Cross Validation (KCV) [1] was used to compare different the performance of each model with differing number of clusters with CV coherence [15] being used as the metric. Using the results from K-fold validation, the elbow method was employed to find the optimal number of clusters. The optimal number of clusters found was 25 clusters.

3.3 LDA

LDA is a widely used probabilistic generative model for topic modeling. LDA assumes that each document in a collection is a mixture of various topics, and each topic is characterized by a distribution of words. The goal of LDA is to estimate these topic-word distributions and document-topic proportions based on the observed word frequencies in the documents. This is typically done using variational inference, an iterative algorithm that approximates the posterior distribution of the latent variables. By optimizing these parameters, LDA uncovers the underlying topic structure within a collection of documents, allowing for applications such as topic modeling, document clustering, and information retrieval.

While HBDSCAN does not require the amount of clusters specified, LDA requires a pre-determined number of clusters set as a hyperparamter. The model with the number of clusters with the highest coherence scores was chosen. As shown in Figure 1[2], the optimal number of topics was determined to be 42.

[2]A variation of the elbow method is used to work with LDA. [8]

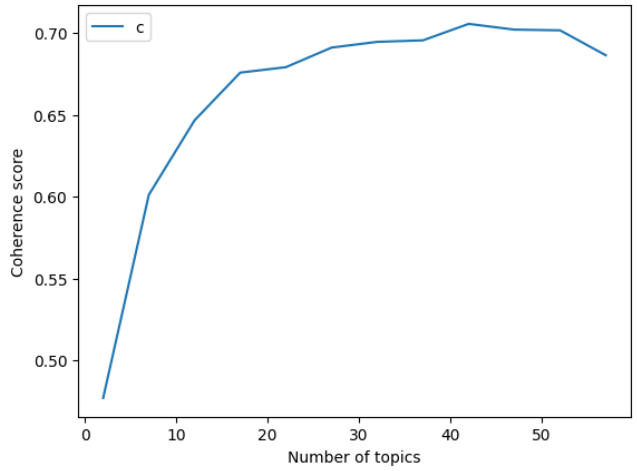

Figure 1: Coherence Scores by Number of Topics: Analyzing the Relationship Between Topic Count and Coherence in LDA

4 EVALUATION METRICS AND RESULTS

4.1 Coherence Score

Our primary evaluation metric was the CV coherence score, a popular and widely used metric for assessing the quality of topics generated by topic modeling algorithms [15]. The CV coherence score gauges the semantic coherence of high scoring words within each topic, with higher coherence scores indicating more coherent and interpretable topics.

Given a topic with N words denoted by $\{w_1, w_2, ..., w_N\}$, the CV coherence is calculated using the following formula:

$$CV_{coherence} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\log\left(\frac{C(w_i,w_j)+1}{D(w_i)}\right)}{-\log\left(\frac{D(w_j)}{D}\right)}$$

Here, $C(w_i, w_j)$ represents the number of documents in which words w_i and w_j both appear, and $D(w_i)$ is the number of documents in which word w_i appears. D represents the total number of documents. The numerator $\log\left(\frac{C(w_i,w_j)+1}{D(w_i)}\right)$ serves as a smoothing of the log of the conditional probability of occurrence of word w_j given word w_i, while the denominator $-\log\left(\frac{D(w_j)}{D}\right)$ is the log of the probability of occurrence of word w_j. The resulting score measures the degree of semantic similarity between the top-scoring words within the topic.

We calculated the CV coherence scores for different pipeline configurations, varying the clustering algorithm used. For performance comparison, the coherence score for the LDA model was also computed.

4.2 Topic Interpretability

In addition to the quantitative evaluation metric, we also assessed the interpretability of the topics generated by each pipeline configuration. We manually inspected the top words in each topic and examined their semantic coherence and relevance to cardiovascular health research. Topics that exhibited clear and meaningful themes were considered more interpretable.

4.3 Results

Table 1: Evaluation Results: CV Coherence Scores for Different Clustering Algorithms and LDA

Algorithm	Number of Clusters	CV Coherence
HDBSCAN	170	0.75
K-means	25	0.61
LDA	42	0.68

The results of our experiments are presented in table 1. It shows the CV Coherence scores obtained for each clustering algorithm, along with the corresponding number of clusters used in each algorithm. The pipeline using HDBSCAN achieved the highest CV Coherence score, indicating the highest degree of topic coherence among the three algorithms.

The higher CV Coherence score obtained by HDBSCAN suggests that it was able to identify more coherent and meaningful topics within the cardiovascular health research publications. The adaptive nature of HDBSCAN, which can handle clusters with varying densities and shapes, might have contributed to its superior performance in capturing the complex patterns and structures present in the data. The K-means-based pipeline had the lowest CV Coherence score, suggesting that it struggled to identify highly coherent topics within the dataset. Topic modeling with LDA didn't perform much better, indicating that it wasn't able to uncover relevant topics.

The following are the prominent topics found using the LDA method: coronary artery, risk factors, smooth muscle. For the HBDSCAN method, the prominent topics identified were heart failure, cardiac function, and risk assessment. Lastly, the key topics identified using the K-means method were blood pressure, nitric oxide, endothelial cell.

5 VISUALIZATION

Visualization plays a crucial role in understanding and interpreting the results of topic modeling. It enables us to gain insights into the relationships between topics, the distribution of terms within topics, and the overall structure of the topic space. In our study, we employed various visualization techniques to facilitate the exploration and interpretation of the generated topics.

5.1 Heat Map

Figure 2 provides a heat map visualization of the similarity scores between topics. This visualization showcases the pairwise similarity between topics using cosine similarity. The cosine similarity function calculates the cosine of the angle between two vectors, which represents their similarity. In our case, the vectors are topic representations, and the similarity is calculated based on the distribution of terms across topics.

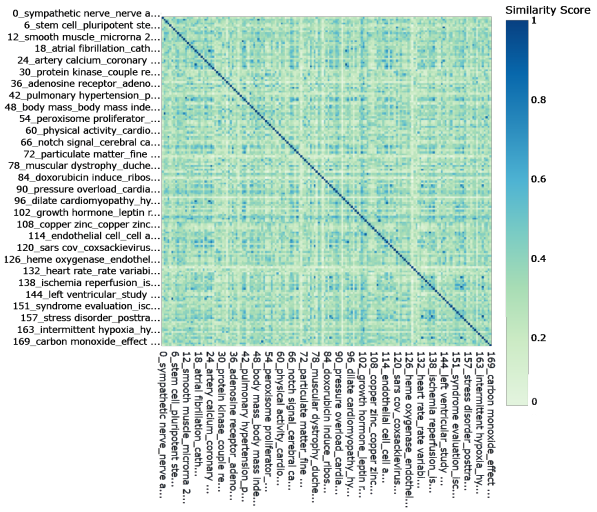

Figure 2: Topic Similarity Heat map: Visualizing the Interrelationships Among Topics

The cosine similarity between two topics, A and B, is computed as follows:

$$\text{similarity}(A, B) = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A_i and B_i represent the weights of the i-th term in topics A and B respectively, and n is the total number of terms.

The heat map employs a color gradient, where dark blue regions represent highly similar topics and lighter green regions indicate more distinct topics. By examining the heat map, we can identify clusters of related topics and observe the overall coherence of the topic space. The heat map provides a global perspective on the relationships between topics, aiding in the identification of dominant themes and topic overlaps.

5.2 Class-Based TF-IDF (c-TF-IDF) Graph

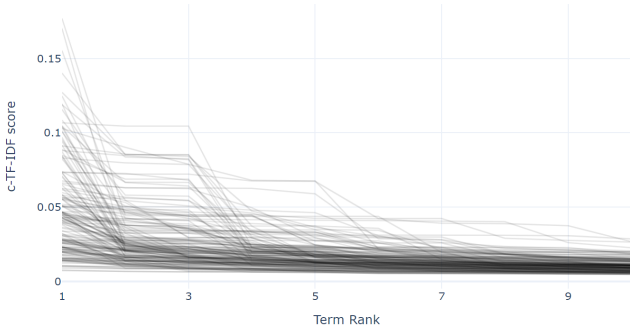

Figure 3: Term Importance within Topics: Examining the Decline of c-TF-IDF Score with Increasing Term Rank

We utilized c-TF-IDF [6] scores to analyze the relevance of terms within each topic. The c-TF-IDF score reflects the importance of terms within topics, with highly relevant terms having higher scores.The c-TF-IDF score ($W_{x,c}$) for a word x in class c is calculated using the equation:

$$W_{x,c} = \|\text{tf}_{x,c}\| \times \log\left(1 + \frac{A}{f_x}\right)$$

Here, $\text{tf}_{x,c}$ represents the frequency of word x in class c, while A corresponds to the average number of words per class. The term f_x denotes the frequency of word x across all classes.

The graph presented in Figure 3 visually represents the trends of c-TF-IDF scores, providing an intuitive understanding of the importance of terms within topics. This analysis enables us to identify key terms that play a significant role in characterizing each topic, offering valuable insights into the domain of cardiovascular health research.

5.3 Cluster Map

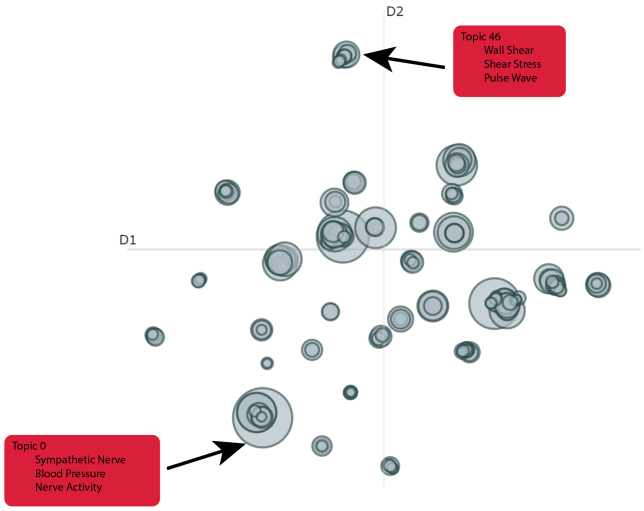

Figure 4: Intertopic Distance Map: Visualizing Clusters and Relationships

Figure 4 represents a cluster map visualization, which provides a two-dimensional representation of the various clusters identified by the pipeline. The cluster map depicts the intertopic distance map, showing the organization and separation of topics. The intertopic distance map provides an understanding of the overall structure of the topics and clusters in the data set. It helped identify closely related clusters, explore topic overlaps and separations, and also to gain a global perspective on the organization of topics within the corpus. This visualization helped with interpreting and analyzing the results of the topic modeling process and provide a basis for further refinement of the topic model.

6 RELATED WORKS

In the field of cardiovascular research, several studies have explored the application of topic modeling techniques to the domain. These

studies have aimed to uncover key topics and identify emerging trends in the field of cardiovascular health.

One notable study by Gal et al. [2019] [5], titled "Hot topics and trends in cardiovascular research", aimed to identify the hot topics and emerging trends in cardiovascular research. The authors conducted a comprehensive analysis of a large data set of cardiovascular research publications to gain insights into the current research landscape in the field.

They applied LDA to a data set of research articles and extracted the most prevalent topics, which were then analyzed and interpreted by domain experts. The findings of the study revealed several prominent topics in cardiovascular research, including cardiac imaging, risk factors, treatment modalities, and preventive strategies. However, while the study provided valuable insights into cardiovascular research topics, it had some limitations that we aim to address in our research.

One limitation is the use of LDA as the sole topic modeling algorithm. Although LDA is widely used and effective in many cases, it may not capture the complex relationships and nuances present in the cardiovascular health literature. By incorporating BERT and UMAP in our pipeline, we can leverage the power of contextual word embeddings and dimensionality reduction to better capture the semantic meaning and structure of the text data.

One additional limitation stems from the inherent delay between the publication of research and ongoing advancements in the field. Consequently, the identified topics may not fully capture the most recent breakthroughs, emerging challenges, or evolving trends within cardiovascular research.

7 CONCLUSION

In this study, we developed a topic modeling pipeline for cardiovascular health research publications using BERT, UMAP, and different clustering algorithms. The pipeline aimed to identify key topics and trends within the literature, providing valuable insights for researchers and practitioners in the field.

Through our evaluation using CV Coherence as the metric, we found that the pipeline with HDBSCAN as the clustering algorithm outperformed the other pipeline configurations with LDA and K-means. HDBSCAN demonstrated superior topic coherence and the ability to capture complex patterns and structures within the dataset. The LDA-based pipeline also performed well, highlighting its effectiveness in identifying relevant topics. However, the K-means-based pipeline had the lowest performance, suggesting limitations in capturing highly coherent topics.

Future work could involve exploring different pre-trained language models and clustering algorithms to further enhance the performance and interpretability of the topic modeling pipeline. Additionally, incorporating additional features such as author information, publication year, and citation network analysis may provide further insights and context for the identified topics. Furthermore, extending the application of our pipeline to other areas of medical research can contribute to knowledge discovery in diverse domains.

In conclusion, our topic modeling pipeline, leveraging BERT, UMAP, and HDBSCAN, demonstrates promising results for analyzing cardiovascular health research publications. By uncovering key topics and trends, the pipeline offers valuable support for researchers, practitioners, and decision-makers in the field, ultimately advancing our understanding of cardiovascular health and improving patient outcomes.

ACKNOWLEDGMENTS

Possible through CAHSI and NSF

REFERENCES

[1] Davide Anguita, Alessandro Ghio, Sandro Ridella, and Dario Sterpi. 2009. K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines. 291–297.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[4] Cédric Févotte and Jérôme Idier. 2010. Algorithms for nonnegative matrix factorization with the beta-divergence. *CoRR* abs/1010.1763 (2010). arXiv:1010.1763 http://arxiv.org/abs/1010.1763

[5] Diane Gal, Bart Thijs, Wolfgang Glänzel, and Karin R Sipido. 2019. Hot topics and trends in cardiovascular research. *European Heart Journal* 40, 28 (06 2019), 2363–2374. https://doi.org/10.1093/eurheartj/ehz282 arXiv:https://academic.oup.com/eurheartj/article-pdf/40/28/2363/46640575/eurheartj_40_28_2363.pdf

[6] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs.CL]

[7] Shun ichi Amari. 1993. Backpropagation and stochastic gradient descent method. *Neurocomputing* 5, 4 (1993), 185–196. https://doi.org/10.1016/0925-2312(93)90006-O

[8] Haider Khalid and Vincent Wade. 2020. Topic Detection from Conversational Dialogue Corpus with Parallel Dirichlet Allocation Model and Elbow Method. arXiv:2006.03353 [cs.IR]

[9] Divya Khyani, BS Siddhartha, NM Niveditha, and BM Divya. 2021. An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology* 22, 10 (2021), 350–357.

[10] M. V. Koroteev. 2021. BERT: A Review of Applications in Natural Language Processing and Understanding. arXiv:2103.11943 [cs.CL]

[11] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2, 11 (2017), 205. https://doi.org/10.21105/joss.00205

[12] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (2018), 861. https://doi.org/10.21105/joss.00861

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL]

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

[15] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.

[16] Serhad Sarica and Jianxi Luo. 2021. Stopwords in technical language processing. *PLOS ONE* 16, 8 (08 2021), 1–13. https://doi.org/10.1371/journal.pone.0254937

[17] Kristina P. Sinaga and Miin-Shen Yang. 2020. Unsupervised K-Means Clustering Algorithm. *IEEE Access* 8 (2020), 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[19] Jonathan J Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *COLING 1992 volume 4: The 14th international conference on computational linguistics*.