Topic Clustering of Journal Articles Using LDA and K-Means

1st Samson Kinyanjui
dept. Computer Science
Dedan Kimathin University of Technology
Nyeri, Kenya
sammainah98@gmail.com

2nd Benson Kituku
dept. Computer Science
Dedan Kimathi University of Technology
Nyeri, Kenya
benson.kituku@dkut.ac.ke

***Abstract*— Topic modeling is Natural Language Processing (NLP) technique, which aims at identifying underlying topics in the collection of documents. One of the most popular method of generating topics is Latent Dirichlet Allocation (LDA), which describes the distribution of topics in a large textual dataset. However, topic models constructed solely based on LDA are not always able to distinguish clear and distinct topics. The aim of this project was to enhance the performance of the topic modelling task by combining LDA and K-means clustering. To prove the effectiveness of the proposed hybrid technique, experiments were conducted on the CORD 19 dataset. In this paper, the topics were first generated using LDA and then K-means was employed to cluster the topics into unique clusters. A Grid search method was used to decide on the best parameters for the LDA model. The hybrid model was performance was analyzed using the Silhouette Score of (0. 7194), Davies-Bouldin Index of (0. 4454) and Calinski-Harabasz Index of (140872. 8767). These metrics show that our method has the potential of generating more coherent topics and more unique topics. Specifically, our finding indicates that the clustering based topic modelling techniques might greatly enhance the topic modelling in many disciplines.**

***Key Terms*—Latent Dirichlet allocation (LDA), topic modelling, k-means clustering, gridsearch,CORD-19 Dataset,Unstructured Text**

I. INTRODUCTION

Topic modeling is a statistical technique that identifies clusters or groupings of related words within a text using unsupervised machine learning. This method leverages the semantic patterns in text to understand unstructured data without training or predefined tags [3]. By detecting topics within a large text corpus, topic modeling can produce concise summaries that highlight the most common topics.

Latent Dirichlet Allocation (LDA) is one of the most widely used techniques for topic modeling. It has been successfully used to classify articles into coherent topics based on word distributions. For instance [4], demonstrated the effectiveness of LDA in categorizing COVID-19-related articles into distinguishable topics. Their study highlighted the potential of combining topic models with other techniques, such as word embeddings and support vector classifiers, to enhance the classification accuracy of scientific articles.

Similarly, [5] introduced a framework for identifying primary research topics within the COVID-19 Open Research Dataset (CORD-19) using Non-negative Matrix Factorization (NMF) and Bayesian classifiers. This approach underscores the power of unsupervised learning methods in extracting meaningful features from large text corpora. However, while these methodologies have proven successful in topic classification, they do not address the potential benefits of combining topic modeling with clustering techniques to enhance the granularity and interpretability of topic groupings.

The huge amount of text material in electronic format and the human inability to read vast amounts of text on the subject have presupposed the increasing use of automatic topic modeling systems. There are several approaches that can be used to select topics from the text, images, and videos. [7]. In NLP ap- plications, these techniques are frequently used to

process large volumes of text and learn the topics that underlie any text in the process. [6] [8] and [9]. However, not all topic modeling techniques are suitable for all types of data. For instance, algorithms used in social media hidden topic retrieval may not be effective for modeling research articles due to different word patterns, document length, and sparsity factors [10]. Therefore, it is crucial to develop methods tailored to specific datasets and applications.

This study aims at integrating LDA with K-means clustering to make topic modeling systems more topic-specific and more easily interpretable at the level of specific journal articles. Specifically, by leveraging the CORD-19 dataset, this work aims to perform unsupervised clustering of articles and extract their thematic similarity to understand the key areas explored during the COVID-19 pandemic.

The rest of the paper is organized as follows: Section 2 describes the literature related to our study Section 3 explains the research method, Section Section Section 4 presents the results of the experiment Section 5 discusses the implication of the study and the future research.

II. Literature Review

Topic Modelling (TM) is an important area of research in Natural Language Processing (NLP), TM aims at identifying salient topics in text so that such text can be better understood without necessarily going through the texts in their entirety. There has been a surge in textual data, which keeps and will continue growing at an exponential rate, it is nearly impossible to read ev- erything to find important topics. Relevant insights may be extracted quickly by employing robust topic modelling techniques. Major NLP difficulties are significantly addressed by such algorithms, which help with organizing,finding, and summarizing large amounts of textual data.

The Vector Space Model (VSM) was initially prevalent in text processing. However, its limitations in capturing statistical structure across documents led to the development of Latent Semantic Analysis (LSA) [11]. LSA creates conceptual sets for documents and words, facilitating the exploration of links between documents and their terms by examining word relationships within a text collection. Using singular value decomposition (SVD), LSA identifies latent relationships between concepts and words. Although LSA effectively reduces data, especially in large corpora, it struggles with nonlinear relationships due to its linear model.

Latent Dirichlet Allocation (LDA) emerged to address LSA's limitations by using a generative probabilistic model. LDA views documents as mixtures of topics and topics as mixtures of words, accommodating complex, nonlinear relationships [12]. Various LDA modifications, including Anchor-free Correlated Topic Modeling, Nested Hierarchical Dirichlet Processes, Supervised Topic Models, and Correlated Topic Models, have been proposed to enhance its functionality [13] [14].

[15] Proposed a sentiment classification model for the tourism domain where two methods, including sentiment lexicon and machine learning, were employed. In their article "Sentiment Classification of Tourism Based on Rules and LDA Topic Model," they gathered texts from Sina microblogs and hotel travel reviews and then built a tourism emotional word list. It was comparing subjects and reference sentiment score of single-sentence texts to a specific threshold which left them in possessing of only those sentences which are unilaterally emotional and they employed the Naive Bayes NB algorithm for preliminary classification.. They then employed LDA to refine the sentiment model, integrating both sets of results. This hybrid method improved accuracy by (0.1385) over the sentiment lexicon method and by (0.0851) over the machine learning approach alone, demonstrating the model's effectiveness for classifying tourism texts.

[16] proposed a Domain-oriented Topic Discovery based on Features Extraction and Topic Clustering (DTD-FETC) to analyze open-source domain-specific web content and identify emerging topics in real-time. They developed three feature extraction methods: ITFIDF-LP for keywords, LDA-SLP for subject words, and a named entity feature extraction method, utilizing the HAC algorithm based on vector product similarity. Their results showed high precision, recall, and F1 scores, demonstrating the DTD-FETC system's ability to filter and aggregate web content for specific security threat topics, aiding rapid response and defense against cyber threats.

[17] introduced CORD19STS, a semantic tex-

tual similarity dataset for COVID-19. Addressing the lack of in-domain knowledge in previous STS datasets, they sampled one million sentence pairs using four strategies and annotated them through Amazon Mechanical Turk (AMT). The Sen-SCI-CORD19-BERT model, a fine-tuned BERT version, was used to compute similarity scores. It achieved notable performance improvements in both unsupervised and supervised settings, highlighting the dataset's potential to enhance NLP applications in the medical domain.

[18] explored the relationship between COVID-19 and intellectual disabilities using the COVID-19 Open Research Dataset (CORD-19). They utilized term frequency-inverse document frequency (TF-IDF) analysis combined with K-means clustering to analyze over 44,000 scholarly articles, including more than 29,000 with full text. From this, they identified 259 articles containing terms related to intellectual disabilities. Their clustering revealed five distinct topics: mental health (51 articles), viral diseases (58 articles), diagnoses and treatments (52 articles), maternal care and pediatrics (51 articles), and genetics (47 articles). This analysis provided insights into the intersection of COVID-19 and intellectual disabilities and highlighted significant research gaps.

[19] Performed a comprehensive scientometric study of its content as well as the relevance of CORD-19 dataset in the light of the current pandemic.. The methodology involved comparing CORD-19 with the Web of Science database, using citation anal- ysis, text analysis, and enrichment with data from Dimensions, Altmetric, Twitter, and other sources. The main objective was to understand the scope and composition of CORD-19. The dataset includes 138,058 publications, with a significant increase in research output during major coronavirus outbreaks, especially in 2020 due to COVID-19. The study identified three main research clusters within the dataset: coronavirus outbreaks and public health, molecular biology and immunology, and respiratory viruses, covering (0.43) of all publications. Altmetrics analysis showed that (0.63) of CORD- 19 publications received social media attention, a higher proportion compared to previous studies. The temporal analysis revealed shifts in research focus from molecular biology and immunology before the 2003 SARS outbreak to broader topics, including epidemics and public health, during and after major outbreaks. This scientometric overview underscores the dataset's extensive coverage and the evolving nature of coronavirus research.

[19] Applied a combination of BERT and LDA that aimed to improve topic modeling via clustering and dimensionality reduction methods. The method- ology involved feature extraction using TF-IDF; topic modeling using BERT for sentence embedding and LDA for topic identification. To enhance the clustering efficiency, various dimensionality reduction methods like PCA, t-SNE, and UMAP were incorporated. K-means algorithm was employed for clustering of the reduced-dimensional data. The data set employed in the paper was CORD-19, which contains more than 900,000 research articles on COVID-19 and coronavirus. The primary goal of this work was to create an integrated approach for enhancing the overall coherence of the topics extracted from the massive textual data. Results demonstrated that the BERT-LDA technique, coupled with dimensionality reduction, yielded more coherent topics, as evi- denced by improved Silhouette Scores (SIL): for LDA with PCA 0. 33239, t-SNE 0. 346303, and UMAP 0. 37624 for BERT with PCA 0. 46211, t-SNE 0. 369380, and UMAP 0. 48761 andThe proposed combined BERT-LDA model with PCA (0. 50843), t-SNE (0. 471261), and UMAP (0. 51998) outperforms the baseline models, emphasizing that the proposed approach is suitable for topic modeling tasks.

III. METHODOLOGY

The proposed approach is shown in Fig 1.

A. CORD-19 Dataset

The dataset used in this study is COVID-19 Open Research Dataset (CORD-19), representing a substantial openly available collection of research papers on COVID-19 [1].CORD-19 dataset was as resulted from collaboration between the National Library of Medicine (NLM) at the National Institutes

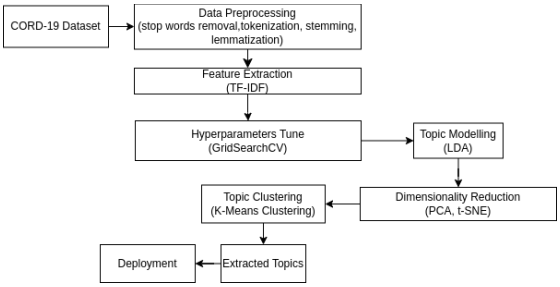

Fig. 1. Proposed integrated Clustering and LDA based topic modelling framework has been described in the following figure 1 which is known as Block Diagram.

of Health (NIH), the Chan-Zuckerberg Initiative, Microsoft Research, IBM Research, Kaggle, and the Centre for Security and Emerging Technology at Georgetown University provide this dataset in metadata and full texts of COVID-19 publications and preprints, which are updated daily.

B. Data Preprocessing

Data preparation processes are used to convert raw data into machine learning modellable format(vectors). Real-world data typically includes defects, inconsistencies, and missing sections in various patterns or trends, as well as a huge number of errors. Preprocessing data is a dependable strategy for addressing these difficulties since it prepares raw data for later processing. Incoming data was first processed through a variety of steps, including Data Cleaning, stopword removal, tokenization, stemming, and lemmatization as shown in Fig 2.

C. Data Preprocessing

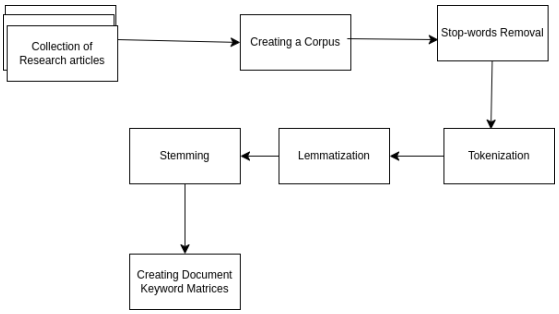

Fig. 2. Block diagram of the proposed Text Processing Methods.

1) Stop-words Removal: Stop words are minor words in a phrase that do not contribute to its meaning. As a result, they may be omitted without changing the sentence's meaning. The NLTK library offers a collection of such stop words that may we used to remove them from our text and create a list of relevant word tokens [21]. This elimination method helped in emphasizing the key terms in the text.

2) Tokenization: This is the process of breaking longer sequence of text or longer sentences into smaller text called tokens [22].

3) Stemming: Stemmers are basic programs that attempt to locate a word's root for uses like clustering. The well-known Porter stemmer approach, for instance, detects that "university" and "universities" have the same stem and combines them together [21] [23].

4) Lemmatization: In NLP, lemmatization is a method of text normalization that converts words into their base or root form. Its principal role is to group different inflected word forms into a single base form with the same semantic meaning. In essence, lemmatization simplifies the work by uniting words with similar roots or lemmas but different inflections or meanings, allowing them to be handled as a single entity. Its purpose is to eliminate inflectional affixes and prefixes, allowing words to be presented in their basic form as found in dictionaries [24].

5) Feature Extraction: Several feature extraction techniques applied on the extracted and processed text data to extract the characteristics of the documents. As we discuss in the context of the NLP domain, there are a number of ways to do this [25]. The features can be extracted from the document and converted into vector representation using one of the ususal methods named as TF-IDF which stands for Term frequency- inverse Document frequency. TF-IDF is a weighted method often used in NLP and Information Retrieval to know the importance of a particular term in a document with respect to a set of documents. Text vectorization allows words in a document to be given numerical values according to their importance. One issue with scoring word frequency is that very common words begin to dominate the document, but they may not contain as

much ”informational content” to the model as rarer but possibly domain-specific words [26]. One technique is to rescale the frequency of terms based on how frequently they appear in all texts, penalizing frequent words that appear in all publications. The TF-IDF Formula is a follows: The Term Frequency-Inverse Document Frequency (TF-IDF) score for a term t in a document d is given by:

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \times \log \frac{N}{\text{df}(t)} \quad (1)$$

where:

- $\text{tf}(t, d)$ is the term frequency of term t in document d,
- N is the total number of documents,
- $\text{df}(t)$ is the document frequency of term t (i.e., the number of documents containing term t).

TF-IDF scores highlight words that are more interesting and distinct in a given document [25].

D. Topic Modelling

1) Optimal number of parameters (Grid-search): By methodically examining a preset range of hyperparameter values, the Grid-Search methodology can be used to determine the ideal parameters for Latent Dirichlet Allocation (LDA) models. Im- portant factors for LDA include the quantity of subjects, The first measure is referred to as the document-topic density (alpha) and the second is the topic-word density (beta). Given- Grid-search involves fitting of LDA model for every possible value of these parameters by cycling through all of them. Evaluation with respect to each model involves the use of a scoring mechanism that could be a coherence score or a perplexity score. The identified parameter set is the one with all parameters that provide the highest score.. This thorough search makes sure that the selected parameters produce the best topic model, improving the subjects that are found and their interpretability and relevance.

2) LDA: The CORD-19 dataset’s vectorized texts are subjected to a topic modelling approach following preprocessing and feature extraction. A standard LDA technique is utilized for topic modelling TM, LDA postulates that each document is the byproduct of few latent topics, and that every individual word has been generated by a distinct topic.. The approach allocates each word in a document to a separate topic, mapping documents to a list of topics based on the words they contain, regardless of the sequence in which the words appear. LDA uses Bayesian inference to identify the underlying topics in a corpus of texts, with the goal of discovering the subjects that best represent the documents’ content [29].

E. Clustering

This strategy involves organizing related data points into groups without supervision. Clustering is used to efficiently categorize subjects within these groups . This clustering approach uses the k-Means algorithm to partition unlabelled datasets into discrete clusters. k-Means, a well-known partitional clustering method, is commonly employed in topic modelling. Silhouette scores are used to measure the performance of clustering, and it is observed that using a mixture of topic modeling by decreasing dimension clusters is effective.

F. Hybrid Algorithm

Algorithm 1 LDA and Clustering Algorithm

0: **Input:** Pandemic Dataset
0: **Output:** Cluster topics
0: **Step 1:** A represents the initial documents derived from CORD-19, where $j = 1, \ldots, m$ and m is the total number of documents.
0: **Step 2:** Create cleaned documents A_c from A by applying stop word removal, tokenization, stemming, and lemmatization.
0: **Step 3:** feature extraction A_f from A_c utilizing term weighting (TF-IDF).
0: **Step 4:** Implement the topic modeling method TM on A_c to obtain Tp_TM .
0: **Step 5:** Apply clustering algorithm to Tp_TM to get Cl.
0: **Step 6:** Perform components reduction on Cl using PCA to get Cp_PCA.
0: **Step 7:** Perform components reduction on Cl using t-SNE to get Cp_tSNE.
0: **Step 8:** Calculate Silhouette Index to assess the model’s effectiveness. =0

G. Terminology

- A - Articles
- A_c - Cleaned documents
- A_f - Feature vectors of documents
- TM - Topic modeling
- Tp_TM - Topics derived from the topic extraction method
- Cl - Clusters obtained from the clustering algorithm
- Cp_PCA - Components after PCA reduction
- Cp_tSNE - Components after t-SNE reduction

IV. RESULTS AND DISCUSSION

A. Experiment Setup

This project was accomplished using Jupyter notebook and was done on core i5 machine with 8gb Ram.

B. CORD-19 dataset

The CORD-19 dataset Initially comprised 51,078 scholarly articles, initially containing 18 columns related to bibliographic information, metadata, and content, including titles and abstracts. After preprocessing, unnecessary columns such as identifiers, licenses, and authors were removed, resulting in a refined dataset focused on the title and abstract of each article. Null values were also dropped, yielding a final dataset of 42,202 entries with two columns: ”title” and ”abstract.” This gave us a curated dataset that is essential for topic modelling and clustering, providing a comprehensive collection of scientific literature abstracts, essential for analysing thematic patterns and trends using LDA and K-means algorithms .

C. Text Preprocessing

The text preprocessing procedure involved several steps to prepare the dataset for topic modelling clustering. Initially, the text was tokenized and converted to lowercase. Stopwords were removed to eliminate common, non-informative words. Next, lemmatization was applied using WordNetLemmatizer to reduce words to their base forms. After this, stemming was done to reduce words further to their base stemmed form using the stemmer. The processed tokens were then joined back into a string format. This preprocessing was applied to a subset of 30,000 entries from the dataset, both for the ”title” and ”abstract” columns. The resulting dataset contains cleaned and standardized text, with titles and abstracts transformed into a more analysable format, exemplified by entries like ”airborn rhinoviru detect effect ultraviolet irradi” and ”background: rhinoviru ,common caus upper respiratori tract infect.” This thorough preprocessing enhances the dataset’s suitability for subsequent topic modelling and clustering tasks.

D. Feature Extraction

TF-IDF vectorization was utilized to modify the pre-processed dataset’s combined corpus of titles and abstracts for the feature extraction process. The text was transformed into a matrix by the TfidfVectorizer, where each row was assigned to a document and each column to a phrase that was weighted according to how frequently that term appeared in the corpus and how important it was in the document. A sparse matrix that represented the importance of every term was the outcome of this.TF-IDF matrix was transformed into a Data Frame for simpler analysis, and feature names were derived from the terms that were taken out of the corpus. Columns were prefixed appropriately to differentiate phrases from abstracts and titles.

E. Determining the optimal number of topics for LDA model

Using GridSearchCV with a parameter grid to explore various values for the number of topics (components), the best number of components for Latent Dirichlet Allocation (LDA) was found. Topics (5, 10, 15, 20) were represented by values in the parameter grid. All CPU cores available for parallel processing were used in the GridSearch initialization, with cross-validation set to three folds (n_jobs=-1). The best parameters result showed that five topics were the optimal number after fitting the grid search to the TF-IDF matrix. With a log-likelihood score, -1007960.2 was the best result from the grid search. Better fitting models are shown by less negative values in the negative score,

which is indicative of the log-likelihood in LDA models. This result suggests that an LDA model with 5 topics is optimal for this dataset, providing a balance between model complexity and performance as evaluated through cross-validation.

F. LDA model results

All things considered; the dataset's unique theme clusters were successfully found by the LDA model. There are distinct themes associated with topics 0 through 3 that are relevant to particular fields including virology, immunology, public health, and medical research.The lack of concentration in topic 4, on the other hand, may point to areas that require additional refinement or data cleaning. In order to facilitate deeper insights and research in pertinent domains, these results offer an organized means of exploring and comprehending the underlying patterns and subjects included in the dataset.

from Fig 3 Each topic represents a distribution of words that are most likely to occur together within documents. Here is a detailed explanation

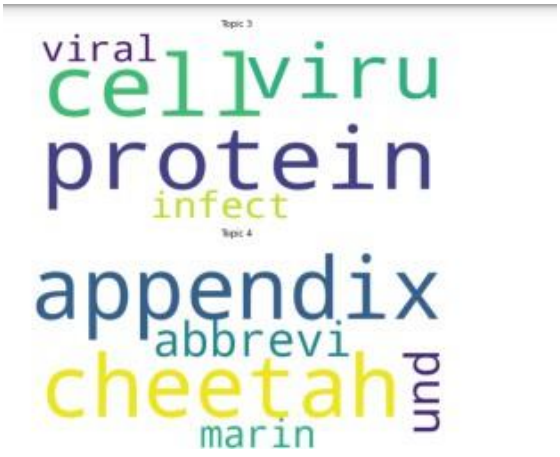

Fig. 3. Topic Wordcloud

below.Topic 0 The Medical Research topic focuses on medical terms related to angiotensin,renin, and other biological processes, possibly indicating research in cardiovascular or biochemical studies. Topic 1 Public Health and Epidemiology topic revolves around terms like disease, health, COVID-19, influenza, and outbreak, suggesting discussions related to public health emergencies, infectious diseases, and epidemiological studies.Topic 2 Respiratory Viruses-This topic includes terms like respiratory, virus, coronavirus, SARS, and acute respiratory syndrome, indicating a focus on studies related to respiratory infections, viruses, and syndromes. Topic 3 Virology and Immunology topic covers terms such as cell, protein, virus, RNA, immunology, and expression, indicating research on viral infections, cellular mechanisms, and immune responses. Topic 4 Miscellaneous Terms-This topic appears more diverse and less coherent, with terms like appendix, cheetah, cocaine, and others. This could indicate documents covering a range of subjects or possibly noise in the dataset.

G. Topic Clustering using k-means

The Elbow Method is utilized to find the ideal k for the clustering intend. It is clear when using the elbow method plot that evaluates the cost function for the different k values.

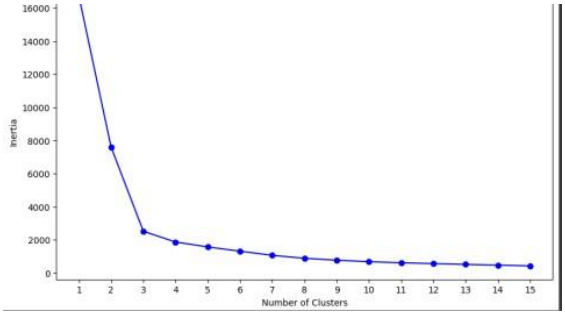

Fig. 4. Elbow plot showing the appropriate number of Clusters

According to the elbow approach, there is a progressive decrease in inertia as the number of clusters grows. Even if there isn't a clear elbow point, the inertia decrease significantly slows down after three clusters, suggesting that three clusters account for a sizable amount of the variance in the data.

H. Visualizing clusters

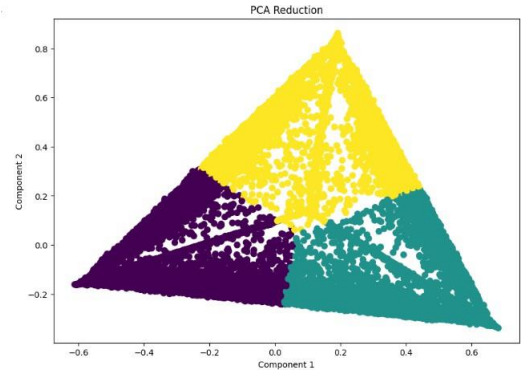

Fig. 5. K-means Clustering with PCA

The provided PCA plot visualizes clusters in a two-dimensional space, showcasing three distinct groups represented by different colors (purple, yellow, and teal). The x-axis and y-axis shows the first and second principal components, respectively, capturing the most significant variance in the data. Each point in the plot corresponds to a data point from the original high-dimensional space, now represented in this reduced 2D PCA space. The clear separation between clusters indicates that the original data has distinct groupings, preserved even after dimensionality reduction. This visualization aids in validating the clustering results, as distinct clusters in the PCA plot suggest effective clustering. Additionally, it provides insight into the data's structure by reducing complexity while retaining essential variance.

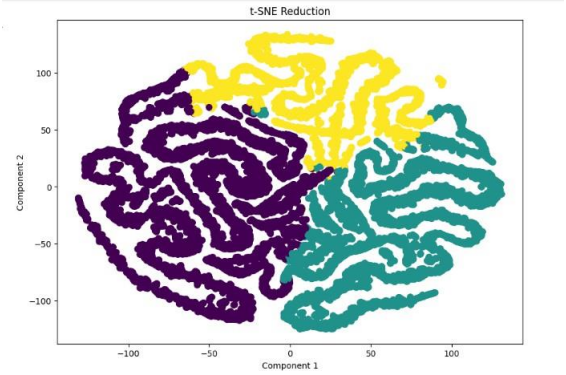

Fig. 6. K-means Clustering with t-SNE

The x-axis and y-axis shows the two components derived from t-SNE, a technique used for dimensionality reduction that emphasizes preserving local structures within the data. Each point corresponds to a data point from the original high-dimensional space. The intricate patterns and shapes formed by the points highlight t-SNE's capability to capture non-linear relationships and complex data structures. The clear separation between clusters suggests effective clustering, validating the results. Unlike PCA, which focuses on capturing variance, t-SNE excels at maintaining the local structure, making it particularly useful for visualizing and interpreting high-dimensional data with complex relationships.

I. Interpreating the clusters

The clusters are shown in a two-dimensional space using PCA and t-SNE techniques. Each cluster is represented by a distinct hue. The two figures shows that the clusters are reasonably separated from one another, while there may be some areas

that overlap. Cluster Features Cluster 0 (Virology and Infectious Diseases)-Terms like influenza, virus, infection, and respiratory system imply that virology, respiratory infections, and associated illnesses are the main topics of discussion. Disease, health, patient, and COVID-19 are terms that signify studies about general medical research, public health, and patient outcomes in Cluster 1 (Medical Research and Global Health). Cluster 2 (Immunology and Immunization)-Terms like immune system, respiratory tract, and immunization imply studies about immunology, specifically about immunological response and early-life susceptibility.

J. Results and Discussion of Topic Clustering

The findings of clustering were evaluated using both the Silhouette Score(SIL) , Davies Bouldin index(DBI), and Calinski-Harabasz Index (CHI). The Silhouette Coefficient, which ranges from -1 to 1, measures the efficiency of a clustering process. Clusters are divided and distinct from one another.

The Silhouette Score $S(i)$ for a single sample i is calculated as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where:

- $a(i)$ is the average distance between i and all other points in the same cluster.
- $b(i)$ is the minimum average distance between i and points in a different cluster, minimized over clusters.

The mean Silhouette Score for all samples is then:

$$\text{Silhouette Score} = \frac{1}{n} \sum_{i=1}^{n} S(i)$$

The Davies-Bouldin Index (DBI) is given by:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)}$$

where:

- k is the number of clusters.
- σ_i is the average distance of all points in cluster i to the centroid of cluster i.
- $d(c_i, c_j)$ is the distance between the centroids of clusters i and j.

The Calinski-Harabasz Index (CHI) is defined as:

$$CHI = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \cdot \frac{n - k}{k - 1}$$

where:

- $\text{Tr}(B_k)$ is the trace of the between-cluster dispersion matrix.
- $\text{Tr}(W_k)$ is the trace of the within-cluster dispersion matrix.
- n is the total number of samples.
- k is the number of clusters.

TABLE I
METRICS FOR LDA+K-MEANS AND NMF+K-MEANS CLUSTERING

Metric	LDA+K-Means	NMF+K-Means
Silhouette Score	0.7194	0.2688
Davies-Bouldin Index	0.4454	1.4027
Calinski-Harabasz Index	140872.8767	14129.4853

Even with three clusters, the silhouette score of 0.7194 shows a significant amount of separation between them, indicating that the clustering quality is still rather strong. The uniqueness and well-separation of the clusters are indicated by the Davies-Bouldin index of 0.4454, which supports the efficacy of the clustering strategy. A high CHI value of 140872.8767 indicates that the clusters are well-separated and compact, implying that the clustering structure was appropriate and the clustering algorithm performed well.

In contrast, the NMF-KMeans approach exhibits lower clustering quality metrics. The silhouette score of 0.2688 indicates less separation between clusters, which suggests that the clusters are not as distinct. The Davies-Bouldin index of 1.4027 indicates a higher degree of cluster overlap, reflecting less effective separation. The CHI value of 14129.4853 for NMF-KMeans is significantly lower compared to LDA-KMeans, further suggesting that the clusters formed are less compact and less well-separated.

Overall, LDA+K-Means demonstrates superior clustering performance relative to NMF+K-Means,

with higher silhouette scores, better Davies-Bouldin index values, and a much higher Calinski-Harabasz Index. This comparison highlights that LDA+K-Means provides more distinct and well-separated clusters, making it a more effective clustering strategy in this context.

V. CONCLUSION AND FUTURE WORKS

Coherent Topic clusters about virology and infectious diseases, medical research, global health, immunology, and immunization were identified using LDA and K-Means to cluster the data into three distinct clusters. These clusters offered significant insights into the dataset, facilitating focused analysis in the Medical field. The quality and isolation of the detected clusters are validated by the high silhouette score, low Davies-Bouldin index score, and high Calinski-Harabasz Index score. This highlights how the clusters can be used in organizing and understanding the underlying structure of a huge text corpus.

Improving the understandability and usage of these clusters in many research and domain-specific situations will be the goal of future improvements or additional analysis In the future, topics can be deduced using the Deep Neural Networks (DNNs) and Transformers-based models. With the use of deep learning-based algorithms and state of the art transformers described in the literature review , the study can be expanded to confirm the effectiveness of the suggested topic modeling framework.

REFERENCES

[1] Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., ... and Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. ArXiv.

[2] Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., ... and Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. ArXiv.

[3] Levity AI, https://levity.ai/blog/what-is-topic-modeling

[4] Bartolome´, L. C., Melchor, J. A. E., and Arenas-Garc´ıa, J. (2023, May). ITMT: Interactive Topic Model Trainer. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 43-49).

[5] Crane, L., Lui, L. M., Davies, J., and Pellicano, E. (2021). Autistic parents' views and experiences of talking about autism with their autistic children. Autism, 25(4), 1161-1167.

[6] Alghamdi, A. G., El-Saeid, M. H., Alzahrani, A. J., and Ibrahim, H. M. (2022). Heavy metal pollution and associated health risk assessment of urban dust in Riyadh, Saudi Arabia. PLoS One, 17(1), e0261957.

[7] Qian, S., Zhang, T., Xu, C., and Shao, J. (2015). Multi-modal event topic model for social event analysis. IEEE transactions on multimedia, 18(2), 233-246.

[8] Alghamdi, R., and Alfalqi, K. (2015). A survey of topic modeling in text mining. Int. J. Adv.Comput. Sci. Appl.(IJACSA),6(1).

[9] Kherwa, P., and Bansal, P. (2019). Topic modeling: a comprehensive review. EAI Endorsedtransactions on scalable information systems, 7(24).

[10] Craig, L., and Churchill, B. (2021). Dual-earner parent couples' work and care during COVID-19. Gender, Work and Organization, 28, 66-79.

[11] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42, 177-196.

[12] Campbell, J. C., Hindle, A., and Stroulia, E. (2015). Latent Dirichlet allocation: extracting topics from software engineering data. In The art and science of analyzing software data (pp. 139-159). Morgan Kaufmann.

[13] Blei, D., and Lafferty, J. (2006). Correlated topic models. Advances in neural information processing systems, 18, 147.

[14] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. Journal of Machine Learning Research.

[15] Chen, B., Fan, L., and Fu, X. (2019, November). Sentiment classification of tourism based on rules and LDA topic model. In 2019 International Conference on Electronic Engineering and Informatics (EEI) (pp. 471-475). IEEE.

[16] Lu, X., Zhou, X., Wang, W., Lio, P., and Hui, P. (2020). Domain-oriented topic discovery based on features extraction and topic clustering. IEEE Access, 8, 93648-93662.

[17] Guo, X., Mirzaalian, H., Sabir, E., Jaiswal, A., and Abd-Almageed, W. (2020). Cord19sts: Covid-19 semantic textual similarity dataset. arXiv preprint arXiv:2007.02461.

[18] Tummers, J., Catal, C., Tobi, H., Tekinerdogan, B., and Leusink, G. (2020). Coronaviruses and people with intellectual disability: an exploratory data analysis. Journal of Intellectual Disability Research, 64(7), 475-481.

[19] Colavizza, G., Costas, R., Traag, V. A., van Eck, N. J., van Leeuwen, T., and Waltman, L. (2021). A scientometric overview of CORD-19. Plos one, 16(1), e0244839.

[20] Lijimol, G., and Sumathy, P. (2023). An integrated clustering and BERT framework for improved topic modeling. International Journal of Information Technology, 1-9.

[21] Sharma, A., Aggarwal, R., and Alawadhi, R. (2023). A Comparative Study of Text Summarization using Gensim, NLTK, Spacy, and Sumy Libraries. Journal of Xi'an Shiyou University, Natural Science Edition, 19.

[22] Zouhar, V., Meister, C., Gastaldi, J. L., Du, L., Sachan, M., and Cotterell, R. (2023). Tokenization and the noiseless channel. arXiv preprint arXiv:2306.16842.

[23] IBM ,2023:https://www.ibm.com/docs/en/watson-explorer/11.0.1?topic=bases-stemming.

[24] Engati (2022) https://www.engati.com/glossary/lemmatization

[25] Remawati, D., Noersasongko, E., and Marjuni, A. (2024, February). Mental Health Detection with TF-IDF Feature Extraction. In 2024 IEEE International Conference on

Artificial Intelligence and Mechatronics Systems (AIMS) (pp. 1-6). IEEE.
[26] Cahyani, S. N., Saraswati, G. W. (2023). IMPLEMENTATION OF SUPPORT VECTOR MACHINE METHOD IN CLASSIFYING SCHOOL LIBRARY BOOKS WITH COMBINATION OF TF-IDF AND WORD2VEC. Jurnal Teknik Informatika (Jutif), 4(6), 1555-1566.
[27] Ali, L., Wajahat, I., Amiri Golilarz, N., Keshtkar, F., and Bukhari, S. A. C. (2021). LDA-GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine. Neural Computing and Applications, 33, 2783-2792.
[28] Sinaga, K. P., and Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.
[29] de Campos, L. M., Ferna´ndez-Luna, J. M., Huete, J. F., Ribadas-Pena, F. J., and Bolan˜os, N. (2024). Information retrieval and machine learning methods for academic expert finding. Algorithms, 17(2), 51.