Unsupervised Analysis of Scientific Paper Abstracts: Exploring Clustering Algorithms

ABSTRACT- The number of scientific publications published online has increased significantly in tandem with the rise of the Internet. The quantity of online articles has significantly expanded in recent years. As the quantity of scientific papers continues to grow exponentially, researchers are having challenges organising and generating relevant insights from the vast amount of information accessible. Unsupervised clustering techniques are applied in this research to categorise scientific papers based on the abstract. Centroid Based Clustering, Spectral Clustering, Distribution Based Clustering (DBSCAN), Grid Based Clustering, and hybrid LDA-BERT KMeans Clustering are the clustering algorithms explored in this study. Results showed that Centroid-Based Clustering resulted in 15 clusters with a silhouette score of 0.0038, Spectral Clustering produced 10 clusters with a silhouette score of 0.0046, DBSCAN formed a single cluster only, Grid-Based Clustering achieved 82 clusters with a silhouette score of 0.3061, and Hybrid LDA-BERT KMeans Clustering yielded 6 clusters with a silhouette score of 0.3951. Possible reasons that the algorithms are less effective compared to one another in this specific task are also investigated.

KEYWORDS: Clustering Algorithms, Natural Language Processing, Dimensionality Reduction, Unsupervised Learning, Artificial Intelligence.

I. INTRODUCTION

The number of scientific articles published online has considerably expanded along with the growth of the Internet. Online articles have steadily increased in number during the past few years. It is frequently difficult to correctly match articles during the editorial process—to match the article to the expertise of a particular editor—despite giving keywords and creating abstracts. Additionally, it poses a challenge for scholars who wish to keep updated on the most relevant papers.

Clustering method is applied to put scientific papers with related subjects together. It can discover underlying thematic linkages and spot groups of publications that discuss similar research themes or domains by examining the abstracts. It can also be used to visualise the clustered papers, allowing academics to quickly spot interesting clusters and acquire new insights.

With this initiative, the aim is to advance the study of scientific literature analysis, providing researchers with an invaluable tool for sifting through the vast amount of information available. Researchers can fully comprehend research trends, discover relevant articles, and make decisions regarding study areas and collaborative projects by utilising clustering and visualisation approaches.

II. PROBLEM STATEMENT

Since 1996, it is predicted that at least 64 million academic papers have been published, with the rate of newly published publications increasing over time [1]. Researchers encounter difficulty in organising and deriving useful insights from the large amount of information available as the number of scientific publications continues to expand dramatically. Manually classifying and sorting documents is time-consuming and labour-intensive, necessitating the use of automated and intelligent technologies to streamline the process.

The initiative intends to overcome the constraints of traditional rule-based approaches and human-intensive procedures by utilising unsupervised learning techniques. Instead, it tries to harness the power of algorithms to uncover patterns and groupings within abstracts without relying on preconceived categories or predefined classifications.

This would not only improve document organisation efficiency but would also allow for a more accurate and informative analysis of the scientific scene. Researchers can better navigate a sea of knowledge by enabling efficient and intelligent clustering of papers, resulting in more informed decision-making, faster information retrieval, and higher productivity.

However, when authors publish scientific papers, tags or labels are often added to indicate the field to which the paper belongs. So, one might wonder why clustering scientific papers is still necessary. Besides serving the purpose of aiding in indexing and searching, clustering scientific papers can also:

A. Identify groups of articles that are topically related to each other: Clustering can group similar articles together based on their content.

B. Assessing the quality level and magnitude of existing evidence on a topic: Researchers can get a sense of how much research has already been done on a particular topic and how strong the evidence is. This can help to inform decisions about future research directions and help to identify areas where more research is needed.

For example, if a particular cluster contains a large number of articles, this may indicate that there is a significant body of research on that topic. Conversely, if a cluster contains only a few articles, this may suggest that there is a gap in the literature that needs to be addressed.

C. Filtering large article corpora during systematic literature reviews: Help researchers quickly filter through a large number of articles to find the ones that are relevant to their research question during the literature review stage.

D. Quickly identify scientific communities and provide an iterative perspective for the linear systematic literature review methodology so far: Through the analysis of these clusters, researchers gain valuable insights into distinct subfields within larger areas of study. When a cluster contains numerous articles on a particular subfield within a broader domain, it signals the presence of a dedicated scientific community focusing on that specific subfield.

To illustrate, imagine a researcher interested in exploring the realm of artificial intelligence (AI). By applying clustering techniques to group AI-related articles and examining the resulting clusters, one may discover that one cluster comprises papers centred on natural language processing, while another cluster encompasses papers focused on computer vision.

This understanding of AI's various subfields can steer the researcher's future research endeavours and empower them to make meaningful contributions to the existing knowledge base. For instance, they might opt to concentrate their research efforts on a specific AI subfield or aim to integrate insights from multiple subfields to cultivate a more comprehensive understanding of AI as a whole.

III. OBJECTIVES

This study explored 5 Clustering Algorithms: Spectral Clustering, Centroid Based Clustering, Distribution Based Clustering, Grid Based Clustering and a hybrid LDA-BERT KMeans Clustering. The objectives are:

1. To apply the 5 clustering algorithms to group scientific papers with similar topics together.

2. To develop a clustering algorithm based on the 5 unsupervised learning techniques.

3. To discover insights and keywords in the scientific literature's abstract using the 5 clustering algorithms.

4. To evaluate the performance of the 5 clustering algorithms using appropriate metrics.

IV. LITERATURE REVIEW

A. Centroid Based Clustering

Centroid-based clustering groups the data into non-hierarchical clusters. The most popular centroid-based clustering algorithm is k-means. Although efficient, centroid-based algorithms are sensitive to beginning conditions and outliers. It is best suited for situations in which the clusters are substantially spherical and equally proportioned. It is most effective when the data can be partitioned into clusters with roughly similar variances. [2] conducted a study on various different types of centroid based clustering algorithms such as k-means, k-medoids, CLARA, CLARANS, generalised k-harmonic means and fuzzy c-means. These algorithms are used on par with primitive numerical data to categorical data. Among all the algorithms, k-means is the most sensitive to noise and outliers [2]. [3] reviewed state-of-the-art Deep Learning-based approaches for cluster analysis that are based on representation learning, which they hope to be useful, particularly for bioinformatics research. The evaluation metrics used in their research are Normalised Mutual Information, Rand Index and accuracy [3].

B. Spectral Clustering

The connectedness between the data points is used to create the clustering in spectral clustering, a variation of the clustering technique. It forecasts the data into lesser dimensions of space to cluster the data points using the eigenvalues and eigenvectors of the data matrix. It is built on the concept of a graph representation of data, where the nodes represent the data points, and the edges show how similar the data points are to one another. It is best suited for tasks where the data has a graph-like structure or when clusters are not necessarily spherical and may have complex relationships. [4] introduced a three-way decision-based approach to spectral clustering in order to make it insensitive to noise.

Results demonstrated that this approach outperformed classical spectral clustering by an average of 30% [4]. [5] presented a new method for Recommendation Based on Embedding Spectral Clustering in Heterogeneous Networks (RESCHet), which uses the embedding spectral clustering method, whose similarity matrix is generated by a heterogeneous embedding approach. Experiments carried out on three open benchmark datasets have demonstrated that RESCHet outperforms current leading methods in a significant manner [5].

C. Distribution-based Clustering

The distributed-based clustering technique separates data into Gaussian-like distributions, where the probability of a point belonging to a distribution decreases as it moves away from the distribution's centre. When there are a limited number of distributions, the Gaussian distribution becomes more important as it maximises data distribution. [6] evaluated this clustering technique using both external and internal validation indices. External indices measure the closeness between the clustering algorithm's output and the appropriate dataset partitioning, while internal indices assess clustering quality without external input.

The Jaccard Index, Adjusted Rand Index, Silhouette Index, and Dunn Index were used as validation indices. Distribution-based clustering algorithms are primarily designed for numerical data based on statistical features, as indicated by [7]. This may not be suitable for natural language processing (NLP) tasks like topic modelling due to high-dimensional feature spaces and limitations in capturing linguistic features.

Transforming textual data into numerical representations using methods like TF-IDF or word embeddings may not fully capture semantic meanings [8]. This method assumes specific probability distributions, which may not always apply to NLP data with irregular word or subject distributions [9]. In specific domains like image processing, computer vision, and finance, distribution-based clustering can be effective, as demonstrated by [10] for image segmentation and [11] for analysing time series financial data.

D. Grid Based Clustering

Grid-based clustering algorithms quantize data into cells and operate on a grid structure. Examples include STING, WaveCluster, and CLIQUE [12]. They use statistical parameters to assess cell relevance to a query. Several evaluation methods like Silhouette coefficient, Dunn index, Davies-Bouldin index, and Calinski-Harabasz index are used [13]. Grid-based clustering is valuable for its ability to handle various cluster shapes and robustness to noise. However, it struggles with large datasets due to high temporal complexity [13].

Determining the optimal grid size is also challenging [14]. In contrast, grid-based clustering is less suitable for textual data, as it lacks semantic understanding [15]. Topic modelling requires semantic insight, which grid-based methods don't provide [16]. High-dimensional data poses challenges like feature sparsity and the curse of dimensionality for grid-based clustering [17].

Using grid-based approaches for textual data may lead to subpar results [18]. Additionally, grid-based clustering struggles with non-linear cluster boundaries common in Natural Language Processing tasks [16]. However, grid-based clustering excels in domains where spatial relationships are crucial, such as network traffic segmentation and handling uncertain or time series data [19].

E. K-Means with LDA-Bert hybrid

Latent Dirichlet Allocation (LDA) is one of the most used topic modelling techniques that can automatically detect topics from a vast collection of text documents [20]. However, LDA-based topic models alone do not always yield promising results. Clustering, an effective unsupervised machine learning algorithm, is extensively used in applications like extracting information from unstructured textual data and topic modelling.

A hybrid model combining Bidirectional Encoder Representations from Transformers (BERT) and LDA in topic modelling with clustering based on dimensionality reduction has been studied in detail [20]. The complexity of clustering algorithms increases with the number of features, prompting the use of Principal Component Analysis (PCA), t-SNE, and UMAP-based dimensionality reduction methods.

K-means and LDA has been employed to facilitate the automated annotation and classification of manufacturing webpages, thereby enhancing the intelligence of supplier discovery and knowledge acquisition tools [21]. In the biomedical domain, where text data plays a pivotal role, topic modelling techniques have been developed to address the redundancy in biomedical text documents and to discover precise topics. Techniques combining inverse document frequency and machine learning fuzzy k-means clustering algorithms have been proposed to enhance the quality of text mining in this domain [22].

LDA is a statistical method for topic modelling, while BERT is a deep learning-based model for understanding the context of words in a sentence. When combined, they offer a powerful approach for topic modelling. Integrating this hybrid model with K-Means clustering and dimensionality reduction techniques can further enhance the extraction of meaningful topics from large text corpora.

V. DATA UNDERSTANDING

The dataset for this study was sourced from Kaggle, comprising a compilation of high-quality and informative academic publications. It represents a vast and diverse dataset suitable for various research and analysis projects. The compilation encompasses publications from a range of disciplines such as computer science, mathematics, physics, and medicine.

The dataset has a total of 1 million rows, however, due to large amounts of computational resources required for NLP tasks, 1000 samples are selected using random sampling method. Before selecting the random samples, filters are also applied to select data from the most recent years (2017) and has more than 50 citations.

Table 1: Dataset Fields

id	A unique identifier for each paper
title	The title of the research paper
authors	The list of authors involved in the paper
venue	The journal or venue where the paper was published
year	The year when the paper was published
n_citation	The number of citations received by the paper
references	A list of paper IDs that are cited by the current paper
abstract	The abstract of the paper

Unsupervised learning approaches are used to cluster the scientific papers based on their abstract. Therefore, only the abstract column will be retained whereas the others will be dropped.

VI. METHODOLOGY

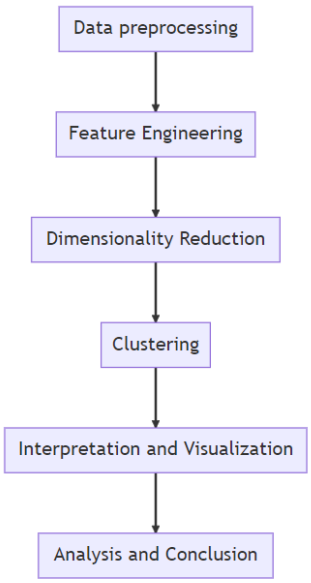

Figure 1: Data Clustering Flowchart

A. Data preprocessing: includes the steps of removing duplicate rows, removing stop words, removing punctuation, and converting to lowercase.

B. Feature engineering: uses TF-IDF and word-embedding techniques.

C. Dimensionality reduction: uses Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP).

D. Clustering: uses Spectral Clustering, Centroid Based Clustering, Distribution Based Clustering, Grid Based Clustering and hybrid LDA-BERT KMeans Clustering.

E. Interpretation and visualisation: interpreting silhouette score and distribution of clusters as well as making appropriate interpretation.

F. Analysis and conclusion: analyse the final outcome and make conclusions.

VII. MODELLING

The workflow of the project depends on the goals and the specific challenges of the dataset at hand. The proposed approach is:

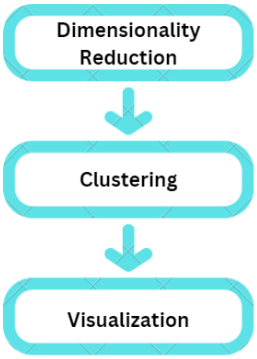

Figure 2: Modelling Flowchart

A. Dimensionality Reduction: Often, datasets have high-dimensional features that can be noisy or correlated. Dimensionality reduction techniques like PCA, t-SNE, or UMAP can help reduce the dimensionality of the data, which can improve the performance of clustering algorithms by focusing on the most important features and reducing the "curse of dimensionality." Additionally, some clustering algorithms, such as KMeans, work better in lower-dimensional spaces due to their reliance on distances.

B. Clustering: Once the dimensionality of the data is reduced, clustering algorithms are applied to the reduced data to identify patterns or groups.

C. Visualisation: The resulting clusters are visualised to represent the data in 2 dimensions. Techniques like t-SNE or UMAP are often used for this purpose because they can capture nonlinear structures and are optimised for visual representation. Other than that, the top words for each cluster are also identified to provide insights.

In other cases, clustering might be performed directly on the original high-dimensional data if it is believed that all features are important and there is a risk of losing critical information during dimensionality reduction.

If the primary goal is to get the best clustering performance, then dimensionality reduction is often done before clustering. If the primary goal is visualisation, then clustering might be done first on the original data, followed by dimensionality reduction to achieve a 2D or 3D representation for visualisation.

VIII. RESULTS AND DISCUSSIONS

The primary metric for evaluating an unsupervised clustering algorithm is the silhouette score. This score quantifies how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a high score indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

The relationship between the silhouette score and the number of clusters is crucial in determining the optimal clustering configuration. Generally, as the number of clusters increases, the silhouette score initially improves, indicating more distinct and well-separated clusters.

However, beyond a certain point, further increasing the number of clusters can lead to a decrease in the silhouette score, suggesting overfitting or the creation of too many, potentially unnecessary clusters. The optimal number of clusters is often chosen by evaluating the silhouette scores across a range of cluster counts and selecting the one that offers the highest score, indicating a balance between cluster cohesion and separation.

Refer to Figure 3, while the number of clusters formed and the visualisation of these clusters also can provide insights into the performance of the algorithm, they are more qualitative measures and cannot be easily quantified as the silhouette score does.

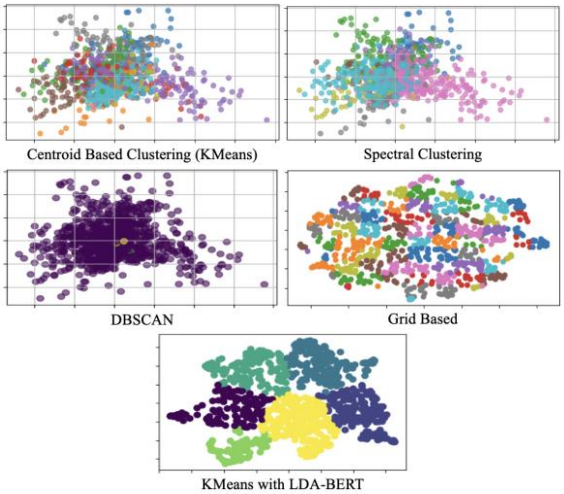

Figure 3: Visualisation of Clusters for All Algorithms

Refer to Table 2 and Table 3, LDA-BERT with KMeans showed the best clustering performance, and the top terms, or keywords in each cluster are identified. The cluster labels are manually annotated based on the theme exhibited by the keywords in each cluster.

Table 2: Top Terms Within Each Cluster Identified by LDA-BERT

Cluster Label	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10
Human -Robot Interaction	social	robot	user	student	learning	different	information	medium	human	design
Mathematical Modeling	graph	time	function	equation	finite	linear	control	delay	numerical	matrix
Energy Systems	power	energy	control	vehicle	channel	time	frequency	signal	scheme	phase
Cloud & Mobile Networking	network	cloud	service	user	scheme	mobile	node	energy	wireless	device
Medical Imaging & Treatment	image	patient	segmentation	information	brain	imaging	clinical	drug	classification	medical
Multimedia Processing	image	video	based	networking	learning	information	level	state	recognition	time

Table 3: Results of All Algorithms with Silhouette Scores And Number of Clusters Formed

Method	Silhouette Score	No. of Clusters	Clusters Keyword	Overall Comment
Centroid Based Clustering (KMeans)	0.0038	15	The terms within each cluster do not seem to have a clear and distinct theme.	With a silhouette score of 0.0038, the clusters are not well-separated and the interpretation of the clusters is not straightforward. Therefore, this method is unsuitable for clustering scientific papers based on abstracts.

Spectral Clustering	0.0046	10	The terms within each cluster do not seem to have clear and distinct themes.	With a silhouette score of 0.0046, although slightly better than KMeans, the clusters are still not well-separated and the interpretation of the clusters is not straightforward. Therefore, this method is unsuitable for clustering scientific papers based on abstracts.
DBSCAN	-	8	-	Almost all scientific papers are grouped into 1 single cluster. Therefore, this method is unsuitable for clustering scientific papers based on abstracts.
Grid Based	0.3061	82	The terms within each cluster do not seem to have a clear and distinct theme.	With a silhouette score of 0.3061, the clusters can be seen to be separated. Upon further inspecting the keywords in each cluster, some are showing a theme. However, there are too many clusters formed, and some keywords are repeated in other clusters. Therefore, this method is unsuitable for clustering scientific papers based on abstracts.
KMeans with LDA-BERT	0.3951	6	The terms within each cluster have clear and distinct themes, the themes can also be identified.	With a silhouette score of 0.3951, the clusters can be seen to be separated. Upon further inspecting the keywords in each cluster, they are distinct and exhibit a theme. Therefore, this method is suitable for clustering scientific papers based on abstracts.

IX. CONCLUSION

K-Means clustering combined with LDA-BERT outperforms other algorithms in clustering scientific papers based on abstracts. The choice of vectorisation, dimensionality reduction, and clustering algorithms significantly influence the results.

While most models in this study used TF-IDF vectorization, the LDA-BERT hybrid stood out. It combines LDA's topic modelling with BERT's feature enrichment, offering both topic discovery and contextual depth. However, it tends to overuse certain domain-specific stopwords, necessitating domain analysis for their removal.

UMAP excels in dimensionality reduction for text data due to its nonlinear nature and ability to preserve both local and global data structures. This makes it adept at capturing the complex relationships in textual data, especially when compared to linear methods like PCA.

In conclusion, K-Means clustering with LDA-BERT and UMAP dimensionality reduction yields the best results for clustering scientific abstracts. While this combination reveals meaningful text patterns, the overuse of domain-specific stopwords in the LDA-BERT method requires attention. This highlights the need for domain-specific considerations in text analysis to ensure optimal clustering and NLP project outcomes.

X. REFERENCES

[1] "Number of Academic Papers Published Per Year – WordsRated," Jun. 01, 2023. https://wordsrated.com/number-of-academic-papers-published-per-year/#:~:text=It%20is%20 estimated%20that%20 at

[2] S. Kumar Uppada, "Centroid Based Clustering Algorithms-A Clarion Study." Available: https://ijcsit.com/docs/Volume%205/vol5issue06/ijcsit2014050688.pdf

[3] M. R. Karim et al., "Deep learning-based clustering approaches for bioinformatics," Briefings in Bioinformatics, vol. 22, no. 1, pp. 393–415, Feb. 2020, doi: https://doi.org/10.1093/bib/bbz170.

[4] S. Khan, O. Khan, N. Azam, and I. Ullah, "Improved spectral clustering using three-way decisions," Information Sciences, vol. 641, pp. 119113–119113, Sep. 2023, doi: https://doi.org/10.1016/j.ins.2023.119113.

[5] Saman Forouzandeh, Kamal Berahmand, Razieh Sheikhpour, and Y. Li, "A new method for recommendation based on embedding spectral clustering in heterogeneous networks (RESCHet)," Expert Systems with Applications, vol. 231, pp. 120699–120699, Nov. 2023, doi: https://doi.org/10.1016/j.eswa.2023.120699.

[6] M. Z. Rodriguez et al., "Clustering algorithms: A comparative approach," PLOS ONE, vol. 14, no. 1, p. e0210236, Jan. 2019, doi: https://doi.org/10.1371/journal.pone.0210236.

[7] Eshref Januzaj, H.-P. Kriegel, and M. Pfeifle, "DBDC: Density Based Distributed Clustering," pp. 88–105, Jan. 2004, doi: https://doi.org/10.1007/978-3-540-24741-8_7.

[8] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," International Journal of Computer Applications, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: https://doi.org/10.5120/ijca2018917395.

[9] M. Indu and K. V. Kavitha, "Review on text summarization evaluation methods," 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS), Bangalore, India, 2016, pp. 1-4, doi: 10.1109/RAINS.2016.7764406.

[10] R. Lu, A. Zlateski, and H. Seung, "Large-scale image segmentation based on distributed clustering algorithms." arXiv preprint arXiv:2106.10795, Jun. 2021, doi: https://doi.org/10.48550/arXiv.2106.10795

[11] H. Kargupta, W. Huang, K. Sivakumar, and E. Johnson, "Distributed Clustering Using Collective Principal Component Analysis," Knowledge and Information Systems, vol. 3, no. 4, pp. 422–448, Nov. 2001, doi: https://doi.org/10.1007/pl00011677.

[12] "What is Grid Based Methods?," www.tutorialspoint.com. https://www.tutorialspoint.com/what-is-grid-based-methods#:~:text=The%20grid (accessed Dec. 28, 2023).

[13] J. Guo et al., "An improved density-based approach to risk assessment on railway investment," Data Technologies and Applications, Nov. 2021, doi: https://doi.org/10.1108/dta-11-2020-0291.

[14] D. R. Edla and P. K. Jana, "A grid clustering algorithm using cluster boundaries," 2012 World Congress on Information and Communication Technologies, Trivandrum, India, 2012, pp. 254-259, doi: 10.1109/WICT.2012.6409084.

[15] W. Wang, J. Yang, and R. R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining," pp. 186–195, Aug. 1997.

[16] Célia Hirèche, Habiba Drias, and Hadjer Moulai, "Grid based clustering for satisfiability solving," Applied Soft Computing, vol. 88, pp. 106069–106069, Mar. 2020, doi: https://doi.org/10.1016/j.asoc.2020.106069.

[17] O. AYTUĞ, K. SERDAR, and B. HASAN, "LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis," International Journal of Computational Linguistics and Applications, vol. vol. 7, no. 1, 2016, no. 0976–0962, pp. 101–119, June. 2016.

[18] S. Sia, A. Dalmia, and S. J. Mielke, "Tired of Topic Models? Clusters of Pre Trained Word Embeddings Make for Fast and Good Topics too!," ACLWeb, Nov. 01, 2020. https://aclanthology.org/2020.emnlp-main.135/ (accessed Aug. 01, 2022).

[19] J. Oyelade et al., "Data Clustering: Algorithms and Its Applications," 2019 19th International Conference on Computational Science and Its Applications (ICCSA), St. Petersburg, Russia, 2019, pp. 71-81, doi: 10.1109/ICCSA.2019.000-1.

[20] George, L., Sumathy, P. An integrated clustering and BERT framework for improved topic modelling. Int. j. inf. tecnol. 15, 2187–2195 (2023). https://doi.org/10.1007/s41870-023-01268-w

[21] Peyman Yazdizadeh Shotorbani, F. Ameri, Boonserm Kulvatunyou, and N. Ivezic, "A Hybrid Method for Manufacturing Text Mining Based on Document Clustering and Topic Modeling Techniques," IFIP advances in information and communication technology, pp. 777–786, Jan. 2016, doi: https://doi.org/10.1007/978-3-319-51133-7_91.

[22] J. Rashid et al., "Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering," in IEEE Access, vol. 7, pp. 146070-146080, 2019, doi: 10.1109/ACCESS.2019.2944973.