ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1946

Semantic Document Clustering Using NLP

T. Madhu[1]; M. Mallikarjun[2]; P. Charan Teja[3]; K. Rahitya[4]

[1]Assistant Professor of Department CSE, Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, India.

[2,3,4]Students of Department CSE of Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad India.

Publication Date: 2025/06/05

Abstract: This project explores a semantic-based document clustering system designed to group documents based on the similarity of their content. Unlike traditional keyword-based methods, which rely solely on word frequency, this system leverages Natural Language Processing (NLP) to understand and compare the semantic meaning within documents. Using pre-trained language models such as BERT and Sentence-BERT, each document is converted into a dense vector representation that captures its underlying meaning. These vectors enable precise comparison of documents' semantic content, allowing for more accurate clustering. The project employs clustering algorithms such as K-Means and DBSCAN, which group documents into clusters based on similarity. Cosine similarity further ensures that related documents are accurately clustered together. Experimental results demonstrate that this approach produces more coherent and contextually relevant clusters compared to traditional techniques, making it an effective solution for applications in content organization, topic analysis, and information retrieval.

Keywords: *Semantic Document Clustering, NLP, BERT Embeddings, Sentence-BERT, Document Similarity, Content-Based Clustering, Cosine Similarity, K-Means, DBSCAN, Vector Representation, Topic Analysis, Information Retrieval, Dense Vector Embeddings, Pre-trained Language Models, Contextual Clustering.*

How to Site: T. Madhu; M. Mallikarjun; P. Charan Teja; K. Rahitya; (2525) Semantic Document Clustering Using NLP. *International Journal of Innovative Science and Research Technology*, 10(5), 3415-3420. https://doi.org/10.38124/ijisrt/25may1946

I. INTRODUCTION

In the era of rapidly expanding digital content, organizing large volumes of unstructured text data has become a significant challenge. Traditional document clustering techniques, which primarily rely on keyword frequency and syntactic patterns, often fall short in capturing the true semantic relationships between documents. This project introduces a semantic-based document clustering system that addresses these limitations by leveraging advanced Natural Language Processing (NLP) techniques. By utilizing pre-trained language models such as BERT and Sentence-BERT, the system transforms each document into a dense vector representation that encapsulates its contextual and semantic meaning. These embeddings serve as the foundation for accurate comparison and grouping of documents based on their intrinsic content, rather than mere word occurrence. To effectively cluster these semantic vectors, the system employs robust unsupervised learning algorithms like K-Means and DBSCAN, along with cosine similarity to measure the closeness between document vectors. The result is a set of coherent and contextually meaningful clusters that outperform traditional keyword-based approaches in terms of relevance and interpretability. This semantic clustering framework has broad applications in content organization, topic discovery, and intelligent information retrieval, making it a valuable tool for navigating and understandinglarge-scaletextualdatasets.

II. LITERATURE REVIEW

Conventional clustering approaches have largely depended on mechanisms akin to Term frequence- Inverse Document frequence(TF- IDF) and the Bag- of- Words(arc) model to model textbook. Effective in some environments, these processes often fall short of capturing the semantic connotation of words and phrases, functioning in syntactically similar but semantically unrelated clusters. To overcome these limitations, researchers have investigated semantic- apprehensive models for textbook modeling. idle Semantic Analysis(LSA) and idle Dirichlet Allocation(LDA) were the first to go beyond raw word frequencies by adding content modeling and latent point birth. still, these models still parade limitations in model- ing environment, polysemy, and word order. The advent of deep literacy and pre-trained language models has ushered in a remarkable leap in semantic textbook representation. Bidirectional Encoder Representations from Mills(BERT), proposed by Devlin et al.(2018), changed the face of NLP with contextual embeddings that account for both left and right word contexts. judgment- BERT(Reimers & Gurevych, 2019), being a variant of BERT, was actually tailored to induce semantically

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1946

rich judgment and document-position embeddings and hence is primarily applicable to clustering tasks. Recent works have shown that incorporating grounded approaches, together with unsupervised clustering methods like K- Means and DBSCAN, achieve much improved results in achieving the semantic similarity of documents. Cosine similarity, widely advocated within these works, has been successful in capturing the angular distance among high-dimensional embedding vectors. Compared to standard methods, these sophisticated semantic clustering methods have demonstrated improved performance in processes such as content segmentation, news composition grouping, and exploration paper bracket.

III. METHODOLOGY

The methodology of the semantic-based document clustering system involves a series of well-defined steps to ensure accurate and meaningful grouping of documents based on their content. It begins with the collection of raw documents from users, which are stored in a centralized document repository. These documents then undergo text preprocessing, a crucial step that includes tokenization (splitting text into individual words or terms), stopword removal (eliminating common, non-informative words such as "the" and "is"), and normalization (standardizing text by converting it to lowercase, removing punctuation, etc.).

Following preprocessing, each cleaned document is passed through a pre-trained language model such as BERT or Sentence-BERT. These models are capable of capturing the semantic meaning of text and generate dense vector embeddings that represent the context and underlying meaning of each document. These embeddings are then stored as vector representations and used to assess similarity between documents.

To determine how similar documents are to one another, the system applies cosine similarity, which measures the angular distance between vectors. This step is essential in ensuring that documents with similar meanings—regardless of the specific words used—are recognized as related. The similarity scores produced are then used by clustering algorithms such as K-Means or DBSCAN. K-Means organizes documents into a predefined number of clusters based on vector proximity, while DBSCAN groups documents based on density, allowing for dynamic cluster discovery without needing to specify the number of clusters in advance.

Once clustering is complete, the grouped documents are presented as the output. These clusters can then be used in various applications such as organizing large volumes of content, performing topic analysis, or enhancing information retrieval systems. This semantic approach provides more meaningful and context-aware clustering compared to traditional keyword-based methods.

IV. EXISTING SYSTEM

- *Hugging Face Transformers + Clustering Libraries*

The Hugging Face Transformers library offers pre-trained transformer models (e.g., BERT, Robert a, T5) that can generate high-quality embeddings for text. These embeddings can then be used with clustering algorithms (like K-Means, DBSCAN, or Agglomerative Clustering) for semantic document clustering.

- *Google Cloud Natural Language API*

Google Cloud offers a Natural Language API that provides text analysis capabilities, including document classification, entity recognition, and sentiment analysis. Though it doesn't directly provide clustering capabilities, you can use its text embedding models (based on BERT-like architectures) to generate semantic embeddings and then apply clustering.

- *Gen sim (LDA for Topic Modelling)*

It uses Latent Dirichlet Allocation (LDA) to discover topics in a corpus of text, which can be used for clustering. While it is not semantic clustering in the deep learning sense, LDA can group documents based on shared themes or topics, providing a form of semantic clustering.

- *Azure Text Analytics API*

Azure's Text Analytics API provides multiple NLP services such as key phrase extraction, sentiment analysis, and language detection. While it doesn't provide direct clustering functionality, it can be used to extract semantic features from text, which can then be used for clustering.

- *Clustering using Tensor Flow and Ker as*

You can create a custom semantic document clustering system using TensorFlow or Keras. This involves training a model to generate document embeddings and using clustering algorithms on those embeddings. TensorFlow has several pre-trained models (such as BERT or Distil BERT) that can be fine-tuned for document clustering tasks.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1946

V. ARCHITECTURE

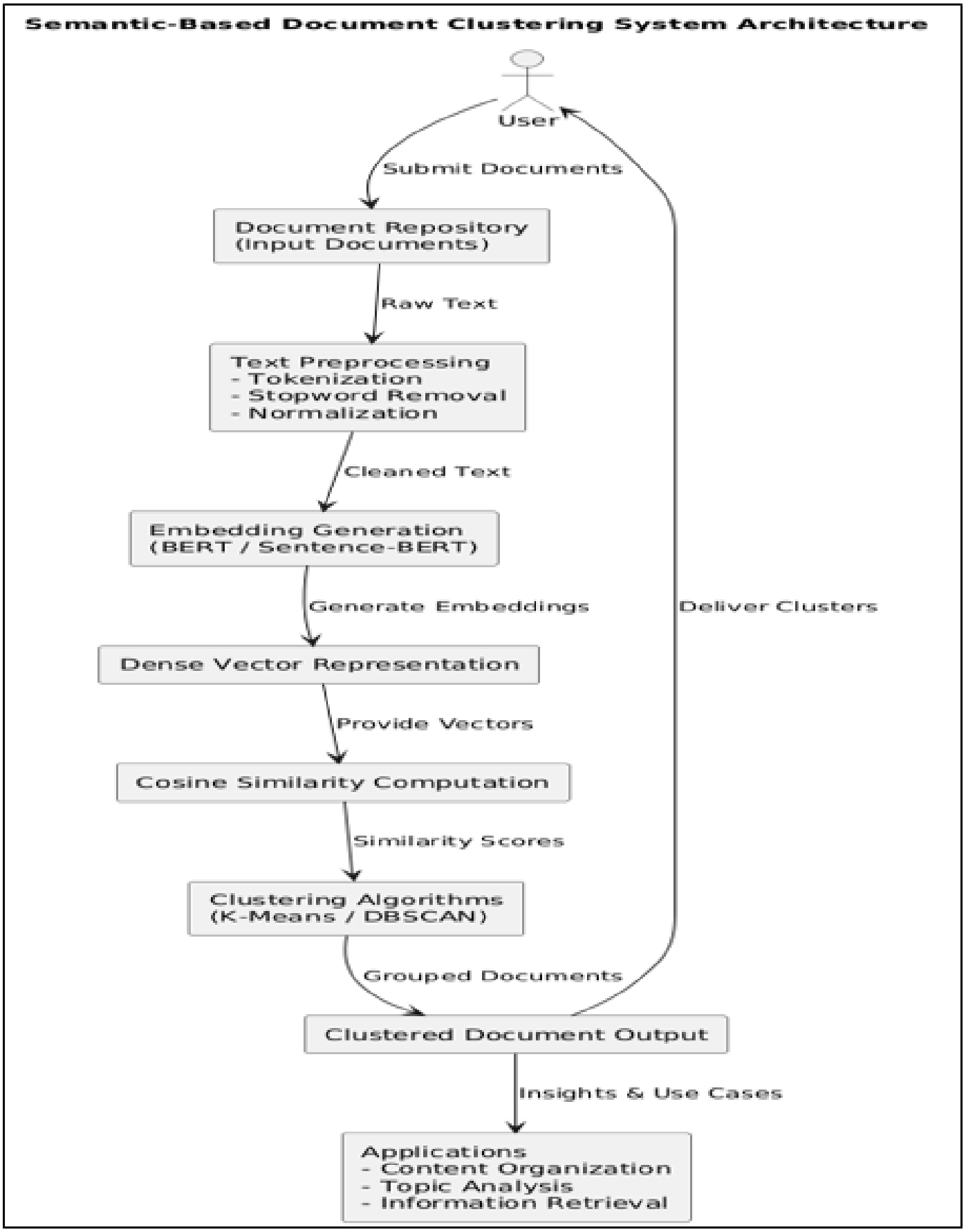

Fig 1 Architecture

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1946

VI. DATA FLOW

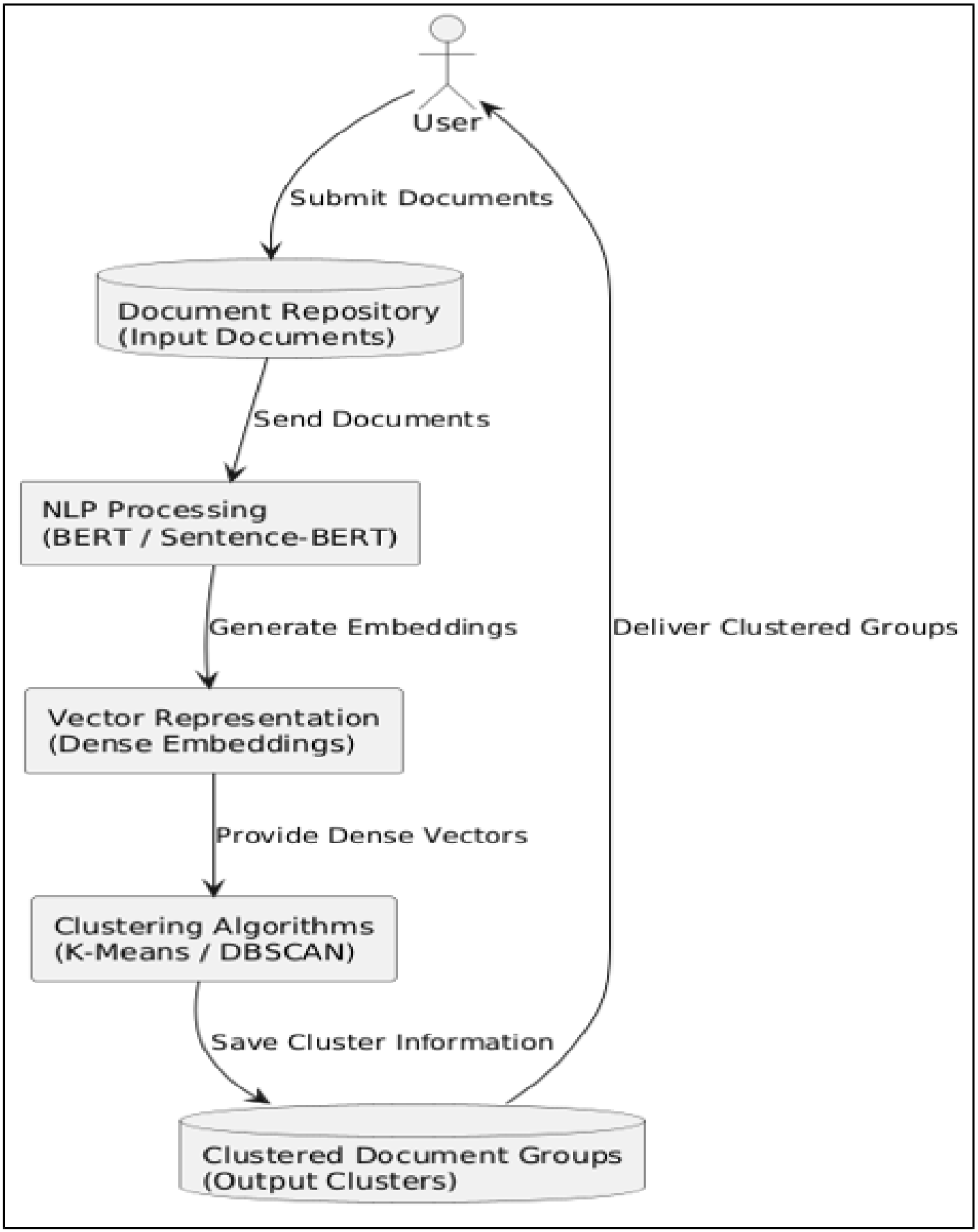

Fig 2 Data Flow

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1946

A user begins by submitting documents into the system. These documents are stored temporarily in a document repository, holding the raw, unprocessed text. The text then goes through a cleaning process where it's broken into words, common words are removed, and everything is standardized to ensure uniformity.

After cleaning, the text is sent to a language model such as BERT or Sentence-BERT. These models understand the meaning of the text and convert it into numerical data called embeddings. These embeddings are stored as dense vectors, which capture the core meaning of each document.

The vectors are compared using cosine similarity, which helps determine how alike the documents are in meaning. This similarity information is passed to a clustering algorithm, such as K-Means or DBSCAN, which groups the documents based on their semantic closeness.

Once grouped, the clustered documents are produced as an output. These clusters are useful for organizing content, identifying topics, or retrieving information more efficiently. Finally, the system provides these organized groups back to the user.

The process begins with a user who initiates the system by submitting documents. These documents are stored in a Document Repository, which acts as the source of input data.

The documents from this repository are then sent to an NLP processing module, where pre-trained language models like BERT or Sentence-BERT are used to extract semantic meaning from the text. This module generates embeddings, which are vector representations of the documents' content.

These dense embeddings are passed to a Vector Representation unit, which serves as an intermediary to organize and prepare the embeddings for clustering.

Next, the vector data is fed into the Clustering Algorithms block, which includes methods like K-Means or DBSCAN. These algorithms analyze the semantic similarities among document vectors and group them into clusters.

The results of the clustering process—i.e., the grouped documents—are stored in the Clustered Document Groups section, which holds the output clusters.

Finally, the clustered information is delivered back to the user, completing the cycle from input documents to organized, semantically meaningful document groups.

VII. ALGORITHAMS BERT

BERT (Bidirectional Encoder Representations from Transformers) is a powerful language model developed by Google that has transformed how computers understand human language. Unlike traditional models that read text in one direction, BERT reads text both ways at the same time, allowing it to grasp the full context of words in a sentence. It is first trained on a large amount of text to learn language patterns and then fine-tuned for specific tasks like sentiment analysis or question answering. BERT's ability to generate rich, contextual word representations makes it highly effective for various natural language processing tasks. However, it requires significant computational resources and can struggle with very long texts. Overall, BERT has set new standards in the field of natural language understanding.

- *Dbscan*

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a powerful clustering algorithm that identifies clusters based on the density of data points in a given space. It operates by defining two main parameters: epsilon (ε), which specifies the radius of the neighborhood around a point, and minPts, the minimum number of points required to form a dense region. The algorithm categorizes points into three types: core points, which have enough neighbors within ε; border points, which are near core points but do not meet the minPts requirement; and noise points, which do not belong to any cluster. This approach allows DBSCAN to effectively discover clusters of arbitrary shapes and sizes, making it particularly useful for datasets with irregular structures. Additionally, it excels in identifying outliers, as it marks low-density regions as noise rather than forcing them into clusters. DBSCAN is widely applied in various fields, including spatial data analysis, anomaly detection, and customer segmentation, due to its robustness and flexibility in handling complex data distributions.

- *T-Sne or Umap*

t-SNE (t-distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection) are both popular techniques for dimensionality reduction, particularly in visualizing high-dimensional data. t-SNE focuses on preserving local structures by converting similarities between data points into probabilities and minimizing the Kullback-Leibler divergence between these distributions in high and low dimensions, which can lead to the formation of distinct clusters but may distort global relationships. In contrast, UMAP is based on manifold learning and aims to preserve both local and global structures by modeling the data as a topological space, using cross-entropy as a loss function and stochastic gradient descent for optimization, making it generally faster and more scalable than t-SNE, especially for larger datasets. Both methods are widely used in fields like machine learning and bioinformatics for exploratory data analysis and visualization. t-SNE (t-distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection) are both widely utilized techniques for dimensionality reduction, particularly effective in visualizing high-dimensional datasets. t-SNE excels at preserving local structures by transforming similarities between data points into probabilities, subsequently minimizing the Kullback-Leibler divergence between these distributions in high and low dimensions. This often results in visually distinct clusters, although it can distort global relationships and is sensitive to hyperparameters like perplexity. On the other hand, UMAP is grounded in manifold learning principles, aiming to maintain both local and global structures by representing the data as a topological space. It employs cross-entropy as a loss function and utilizes stochastic

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1946

gradient descent for optimization, which generally makes UMAP faster and more scalable than t-SNE, particularly for larger datasets. Both methods are extensively applied in various domains, including machine learning, bioinformatics, and exploratory data analysis, providing valuable insights through effective visualization of complex data.

VIII. CONCLUSION

The primary objective of this project, Semantic Document Clustering, was to build an intelligent system capable of automatically grouping similar documents based on their semantic meaning rather than just keyword matching. This objective has been successfully achieved.

Through the integration of deep learning embeddings (Sentence Transformers) and unsupervised clustering techniques (DBSCAN), the system efficiently clusters PDF, DOCX, and TXT files without needing prior labels or training.

The Node.js server provided a simple and user-friendly interface, allowing users to easily upload multiple files and view the clustered results dynamically.

REFERENCES

[1]. M. REIMER M., DODGE J., GILMER J., HOFFMAN M.D., DREDZE M. Sentence-level representations for document classification. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, USA, 2019, pp. 700–707, doi: 10.18653/v1/N19-1070.

[2]. DEVLIN J., CHANG M.W., LEE K., TOUTANOVA K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, 2019, pp. 4171–4186, doi: 10.48550/arXiv.1810.04805.

[3]. REIMERS N., GUREVYCH I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, 2019, pp. 3982–3992, doi: 10.48550/arXiv.1908.10084.

[4]. AGGARWAL C.C., ZHAI C.X. A Survey of Text Clustering Algorithms. *Mining Text Data*, Springer, Boston, MA, 2012, pp. 77–128, doi: 10.1007/978-1-4614-3223-4_4.

[5]. ESTER M., KRIEGEL H.P., SANDER J., XU X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland,1996, pp. 226–231.

[6]. HAN J., PEI J., KAMDAR M. Data Mining: Concepts and Techniques. *Elsevier*, 4th Edition, 2022, pp. 493–508, ISBN: 978-0-12-818148-7.

[7]. LIU J., SHEN X., PAN W., LIU B. Document clustering via topic modeling using BERT embeddings. *Proceedings of the 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, Beijing, 2020, pp. 188–192, doi: 10.1109/ICAICE51518.2020.00047.

[8]. ZHAO W.X., GUO Y., HE Y. A Comparative Study of Deep Learning Models for Semantic Document Clustering. *Information Sciences*, 2021, 576, pp. 55–72, doi: 10.1016/j.ins.2021.07.055.