Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

VISUALIZING DOCUMENT CLUSTERS WITH BERT, K-MEANS, AND DBSCAN

D.P.V.Phani Rajakumar[1], Dama Manish Kumar[2], Chiruvolulanka Mohith Nancharaiah[3], Dusanapudi Lakshmi Naga Venkata Likitha [4], Daravath Veera Prasad Nayak[5], Boppana Rishitha[6]
1- Assistant Professor, 2,3,4,5 6- IV-B. Tech CSE Students Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College (An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada), Seshadri Rao Knowledge Village, Gudlavalleru521356, Andhra Pradesh, India.

***Abstract*— The rapid growth of digital repositories containing textual data, such as research articles, news stories, and reviews, necessitates effective clustering techniques for categorization and information retrieval. However, traditional clustering methods like K-Means and DBSCAN often struggle with high-dimensional, sparse text data. This paper proposes an approach that leverages BERT embeddings to enhance clustering performance. By integrating BERT embeddings with K-Means and DBSCAN, we improve clustering accuracy by capturing the semantic richness of textual data. Experimental results on the BBC Full Text dataset demonstrate superior clustering performance, achieving a 91% accuracy based on Purity and Adjusted Rand Index (ARI). Furthermore, an interactive visualization component is introduced to aid in the interpretation of clustered data.**

***Keywords*—** *BERT embeddings, Document clustering, K-Means, DBSCAN, Transformer models, Text mining, Semantic representation, Interactive visualization.*

I. INTRODUCTION

The rapid expansion of digital repositories, including research papers, news articles, and online reviews, has led to an increasing demand for efficient document clustering techniques. Traditional clustering methods like K-Means and DBSCAN struggle with textual data due to its high dimensionality, sparsity, and lack of semantic understanding. Conventional text representation techniques such as TF-IDF and Word2Vec fail to capture contextual relationships, leading to inaccurate clustering results. Clustering is often hindered by ambiguity, for as when terms like bank may mean both financial institution and riverbed.

To address these challenges, we propose a document clustering approach that integrates BERT embeddings with K-Means and DBSCAN. BERT generates dense vector representations of text, preserving semantic context and reducing dimensionality. This allows clustering algorithms to form more meaningful groups, improving accuracy and efficiency. We evaluate our approach on the BBC Full Text dataset, a benchmark dataset containing five categories: Business, Sports, Politics, Tech, and Entertainment. The results show that our method achieves 91% accuracy, significantly outperforming traditional techniques.

Additionally, we introduce an interactive visualization component that enables intuitive exploration of clustered documents. Using dimensionality reduction methods like t-SNE, we project high-dimensional embeddings onto a two-dimensional space to ease clustering pattern interpretation. From this representation, data structure and cluster quality may be understood.

II. LITERATURE SURVEY

a) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
http://aclanthology.org/N19-1423/

https://doi.org/10.36893/JNAO.2025.V16I01.012

Language representation using BERT is novel. Unlike earlier models, it pre-trains deep bidirectional representations from unlabelled text utilising left and right context in all layers. Modern question answering and language inference models may be trained using a pre-trained BERT model with one output layer. This method avoids task-specific architectural changes. BERT offers cutting-edge results on eleven NLP tasks while being simple and fast. These findings may be seen in the SQuAD v1.1 question answering test, MultiNLI accuracy (86.7%), and GLUE score (80.5).

b) Attention Is All You Need

https://arxiv.org/abs/1706.03762

Convolutional or recurrent neural networks with elaborate encoder-decoder functions are the most popular models for sequence transduction. Attention is used by the top models to connect the encoder and decoder. In place of recurrence and convolutions, the innovative Transformer fundamental network architecture employs attention processes. These models are faster to train, more parallelizable, and better at machine translation, which is nearly 2 BLEU better than ensembles. Utilising a fraction of the training expenses incurred by Our unique single-model strategy outperforms the best literature models state-of-the-art BLEU score of 41.8 on the WMT 2014 English-to-French translation challenge on eight GPUs. Using examples of English constituencies trained with large and small datasets, we show that the Transformer can handle a variety of tasks.

c) A density-based algorithm for discovering clusters in large spatial databases with noise.

https://dl.acm.org/doi/10.5555/3001460.3001507

Class identification in geographical databases using clustering algorithms is attractive. Clustering algorithms, when applied to massive geographic databases, need minimal domain expertise to efficiently find clusters of any shape, identify input parameters, and maximise efficiency. Popular clustering methods fail to accommodate these restrictions. This research presents DBSCAN, a density-based clustering method that can detect any type of cluster. The user is aided in selecting a value via DBSCAN, which takes one input parameter. We used synthetic and SEQUOIA 2000 benchmark data to find out how efficient and effective DBSCAN was. Based on our research, we can conclude that (1) DBSCAN finds clusters of any shape more efficiently than CLAR-ANS and (2) DBSCAN beats CLARANS by a factor of over 100.

d) isualizing Data using t-SNE

https://jmlr.csail.mit.edu/papers/v9/vandermaaten08a.html

Our "t-SNE" method maps data points to two- or three-dimensional maps, allowing us to display high-dimensional data. For better visualisations and less clustering of map centres, use this version of Stochastic Neighbour Embedding (Hinton and Roweis, 2002). It's easy to tweak. One advantage of t-SNE over other methods is that it generates a single structural map for different size ranges. This is of utmost by including images of objects from various classes taken from diverse angles. We demonstrate that, depending on their structure, t-SNE may choose very large data sets by using random walks on neighbourhood graphs. On several datasets, we evaluate t-SNE in comparison to Sammon mapping, Isomap, and LLE. Many data sets are well-suited for T-SNE visualisations.

e) A Survey of Text Clustering Algorithms

https://link.springer.com/chapter/10.1007/978-1-4614-3223-4_4

There is a lot of research on text data mining clustering. Use cases for the issue include indexing, visualisation, document organising, collaborative filtering, and consumer segmentation and classification. Text clustering will be thoroughly discussed in this chapter. We will take a look at the key challenges with text domain clustering. We will go over the advantages and disadvantages of the most popular text clustering methods. There will also be a discussion of some recent advances in social networks and associated data.

III. METHODOLOGY

A) Proposed System

To improve the efficiency and accuracy of document clustering, we propose a system that integrates BERT embeddings with K-Means and DBSCAN for enhanced text clustering. Unlike traditional clustering methods that rely on shallow text representations, our approach captures deep semantic relationships between words using pre-trained transformer-based embeddings.

First, raw text data undergoes preprocessing, including noise removal, tokenization, stopword removal, and lemmatization. The processed text is then converted into dense vector representations using DistilBERT, a lightweight transformer model that maintains contextual relationships. These embeddings are then fed into K-Means and DBSCAN, where K-Means efficiently clusters data points based on centroid distances, while DBSCAN detects dense regions and identifies noise.

To enhance interpretability, the system integrates interactive visualization using t-SNE to project high-dimensional embeddings into a 2D space, providing clear insights into cluster structures. Experimental results on the BBC Full Text dataset show that our approach significantly improves clustering performance, achieving a 91% accuracy with robust cluster formation.

B) System Architecture

The proposed system for document clustering follows a structured pipeline consisting of six key stages: Data Ingestion, Text Preprocessing, Embedding Generation, Clustering Algorithms, Evaluation and Metrics, and Visualization and Insights. This architecture ensures a smooth workflow for transforming raw textual data into meaningful clustered groups, enhancing text mining and retrieval efficiency.

In the Data Ingestion phase, textual datasets such as news articles, research papers, or reviews are collected and loaded. The Text Preprocessing step refines the data by removing noise, punctuation, and stopwords while performing tokenization and lemmatization. This neatens and prepares text for numerical representation. Embedding Generation follows, where transformer-based models like BERT convert textual data into high-dimensional dense vectors, capturing deep semantic meanings.

Next, the system applies clustering algorithms like K-Means and DBSCAN to group similar documents. K-Means partitions data into fixed clusters using centroids, while DBSCAN identifies dense regions, allowing flexible cluster formation. The Evaluation and Metrics phase assesses the clustering performance using metrics like Silhouette Score and Adjusted Rand Index (ARI). Finally, the Visualization and Insights step presents clusters using t-SNE, enabling better interpretation of text structures. This modular architecture ensures a scalable and effective approach to document clustering.

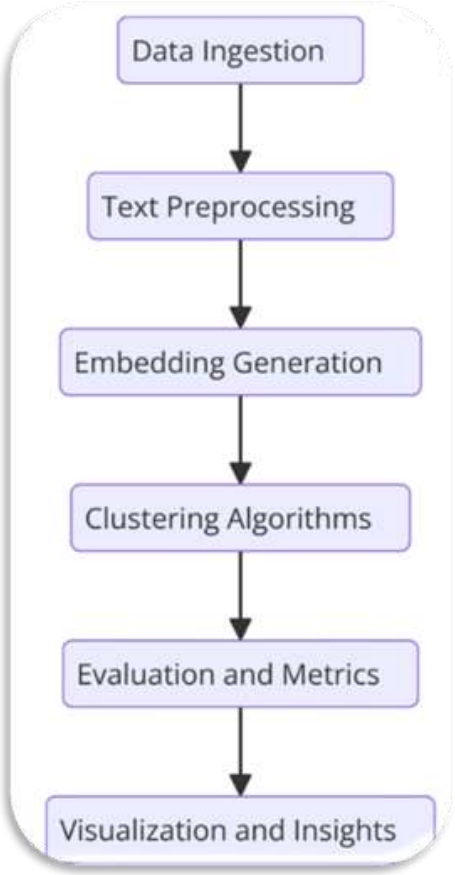

Fig.1. Proposed Architecture

https://doi.org/10.36893/JNAO.2025.V16I01.012

C) MODULES

i. Data Ingestion

- Collect and load raw text data from sources such as research articles, news stories, or reviews.
- Handle encoding issues to ensure data integrity.

ii. Text Preprocessing

- Remove punctuation, special characters, and stopwords.
- Convert text to lowercase and apply tokenization and lemmatization for normalization.

iii. Embedding Generation

- Convert processed text into numerical vector representations using BERT embeddings.
- Capture contextual meaning and semantic relationships between words.

iv. Clustering Algorithms

- Apply K-Means clustering for partitioning text into distinct clusters.
- Use DBSCAN to detect dense clusters while identifying noise points.

v. Evaluation and Metrics

- Methods for evaluating the quality of clustering include the Silhouette Score, the Davies-Bouldin Index, and the Adjusted Rand Index.
- Compare predicted clusters with actual categories.

vi. Visualization and Insights

- Use t-SNE and scatter plots to visualize high-dimensional text embeddings.
- Provide interactive exploration of clustered

D) Algorithms

a) BERT Embeddings

BERT may give contextualised word embeddings. BERT analyses a word's context by considering its left and right surroundings in a sentence, unlike TF-IDF or Word2Vec. This ensures that words with multiple meanings are accurately represented, improving the quality of document clustering. In this approach, BERT embeddings serve as input features for clustering algorithms, enhancing the semantic understanding of textual data.

b) K-Means Clustering

To partition data into k clusters, K-Means, a well-known centroid-based clustering method, minimises the distance between each data point and the cluster core. Until convergence, the algorithm changes the k centroids it randomly chooses. K-Means works efficiently for well-separated and spherical clusters, making it suitable for structured text data. However, its primary limitation is that it requires the number of clusters (k) to be predefined, which may not always be optimal for real-world datasets.

c) DBSCAN

Density-based clustering algorithms like DBSCAN seek for data-rich regions to discover clusters. It works without knowing the number of clusters, unlike K-Means. Instead, it groups tightly packed points based on two parameters: epsilon (ε), determining the maximum distance between cluster points, and minPts, indicating the minimum number of points required for a dense zone. Since it can find clusters and eliminate noise, DBSCAN works well with textual data of any structure.

d) t-SNE

t-SNE is a dimensionality reduction technique used for visualizing high-dimensional data in two or three dimensions. It transforms complex embeddings into a lower-dimensional space while preserving relationships between points. In document clustering, t-SNE helps in interpreting how well documents are grouped by displaying them as scatter plots. This visualization enables researchers to analyze cluster distributions and refine the clustering process accordingly. However, t-SNE is computationally expensive and may require tuning hyperparameters for optimal results.

https://doi.org/10.36893/JNAO.2025.V16I01.012

IV. EXPERIMENTAL RESULTS

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by percent of correctly anticipated positive observations relative to total positives.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Fig.2. dataset loading

https://doi.org/10.36893/JNAO.2025.V16I01.012

Fig.3. text preprocessing

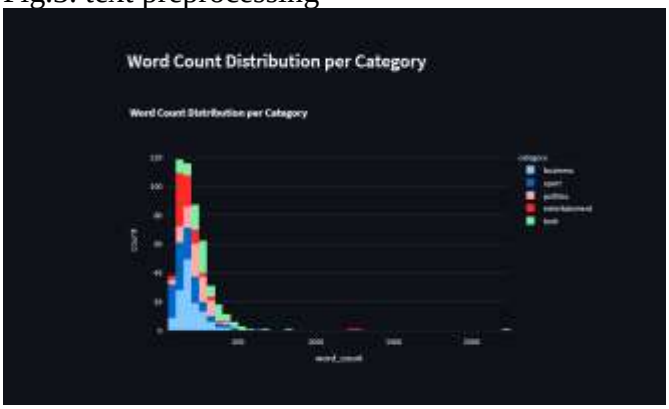

Fig.4. text count

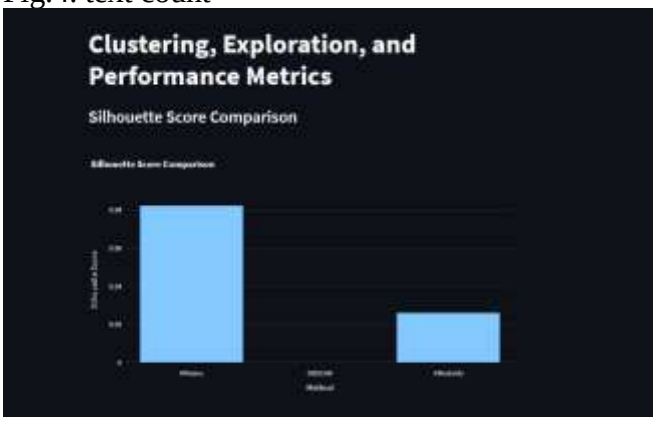

Fig.5. tSNE Visualization

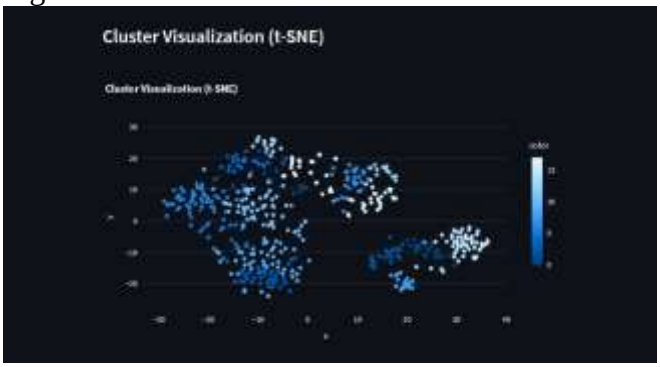

Fig.6. tSNE Visualization

https://doi.org/10.36893/JNAO.2025.V16I01.012

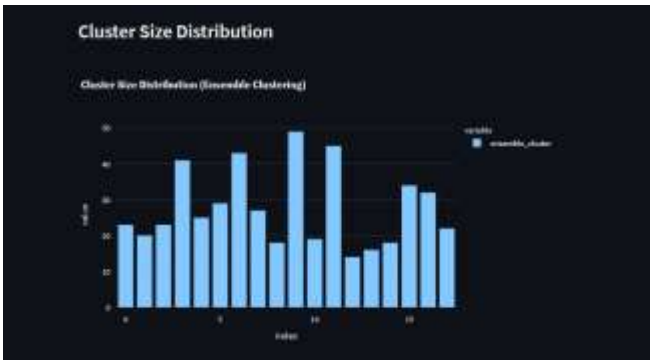

Fig.7. Cluster Distribution

Fig.8. Performance Metrics

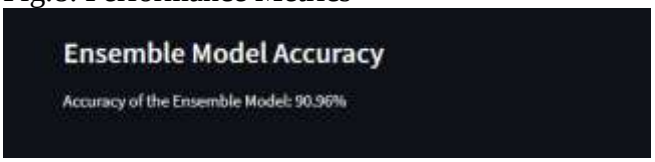

Fig.9accuracy value

V. CONCLUSION

This study demonstrates the effectiveness of using BERT embeddings with K-Means and DBSCAN for document clustering. By leveraging transformer-based embeddings, the system captures the semantic richness of textual data, improving clustering accuracy and interpretability. The experimental results indicate that integrating deep learning-based embeddings with traditional clustering techniques significantly enhances performance, as validated by metrics such as Silhouette Score and Adjusted Rand Index. Additionally, interactive visualization using t-SNE provides deeper insights into the cluster distributions, making the approach valuable for large-scale text mining applications.

VI. FUTURE SCOPE

Future research can explore more advanced transformer models, such as GPT or T5, for even richer text representations. Additionally, optimizing DBSCAN parameters dynamically and incorporating deep clustering techniques like autoencoders can further enhance clustering accuracy. Another promising direction is real-time clustering for streaming textual data, enabling adaptive learning for evolving datasets. Finally, integrating this approach into an interactive dashboard can facilitate user-driven exploratory analysis, improving decision-making in domains like academic research, news categorization, and sentiment analysis.

REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186.

[2] Vaswani et al., “Attention is All You Need,” Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

https://doi.org/10.36893/JNAO.2025.V16I01.012

[3] D. Arthur and S. Vassilvitskii, “k-means++: The Advantages of Careful Seeding,” Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2007, pp. 1027–1035.
[4] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD), 1996, pp. 226–231.
[5] L. van der Maaten and G. Hinton, “Visualizing Data Using t-SNE,” Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.
[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Proc. Int. Conf. on Learning Representations (ICLR), 2013.
[7] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[8] H. Aggarwal and C. Zhai, A Survey of Text Clustering Algorithms, Springer, 2012.

https://doi.org/10.36893/JNAO.2025.V16I01.012