iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73119

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Multilingual Text Document Clustering and Classification

Divya Katta[1], Dr. M. Dhanalakshmi[2]
[1]*M. Tech, Data Science, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad, UCESTH, India*
[2]*Professor of IT Dept & Deputy Director of DILT, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad, UCESTH, India*

Abstract: ***The increasing volume of digital content in multiple languages has created a strong need for intelligent systems that can organize and retrieve multilingual documents efficiently. This project introduces a comprehensive pipeline for clustering and semantic search of multilingual text documents, supporting English, Hindi, and Telugu. The system begins by accepting PDF documents and identifying their language using the langdetect library. This is followed by language-specific preprocessing, including Unicode normalization, sentence tokenization, punctuation removal, stopword elimination, and lemmatization (for English). After preprocessing, the cleaned texts are transformed into semantic embeddings using the paraphrase-multilingual-MiniLM-L12-v2 model from Sentence Transformers. These embeddings are then passed through Agglomerative Clustering based on cosine distance to group similar documents. The clustered results are projected onto a two-dimensional space using UMAP for visualization and further analyzed using cosine similarity heatmaps. To enhance clustering, the system incorporates a semantic search module that retrieves top documents across languages using cosine similarity between query and document embeddings. The system's effectiveness is demonstrated through metrics evaluating both language detection accuracy and clustering performance, supported by visualization techniques.***

Keywords: Multilingual, Language Detection, Text Preprocessing, Semantic Embeddings, Transformer Models, Agglomerative Clustering.

I. INTRODUCTION

In today's multilingual digital landscape, managing and analysing text documents across different languages poses a significant challenge. Traditional monolingual systems struggle with scalability when applied to languages such as Hindi, Telugu, and English—commonly used in a linguistically diverse country like India. This project presents an unsupervised system for clustering and semantic search of multilingual documents without relying on labelled data. The system detects the document language, applies language-specific preprocessing, and generates semantic embeddings using a multilingual transformer model. These embeddings are clustered using Agglomerative Clustering based on cosine similarity, with visualization provided through UMAP and heatmaps. The system also supports semantic search and classification of new documents. Performance is evaluated using language detection accuracy and clustering quality via Silhouette Scores. The approach is language-independent, scalable, and adaptable to real-world applications in education, governance, and multilingual content management.

II. RELATED WORK

Traditional document clustering systems are predominantly monolingual and rely on shallow techniques such as Bag-of-Words or TF-IDF, which fail to capture semantic meaning, especially across different languages. These systems often require separate pipelines for each language, increasing complexity and limiting scalability. While some approaches translate documents into a common language before processing, this can introduce errors and loss of meaning. Others depend on supervised models that require large amounts of annotated data, making them impractical for multilingual, real-world applications.

Recent research focuses on semantic representations using transformer-based models. Sentence-BERT by Reimers and Gurevych (2019) introduced sentence-level embeddings for semantic clustering and multilingual similarity. Angelov (2020) proposed Top2Vec, combining embeddings, UMAP, and topic modeling for unsupervised clustering. UMAP, developed by McInnes et al. (2018), is widely used for dimensionality reduction and visualization of high-dimensional semantic spaces.

Litschko et al. (2021) demonstrated the use of agglomerative clustering on multilingual embeddings for cross-lingual tasks. Jauhiainen et al. (2019) reviewed various language identification techniques including LangDetect, which supports over 50 languages. Mersha et al. (2024) highlighted the integration of BERT embeddings, cosine similarity, and UMAP for unsupervised multilingual document grouping.

These studies highlight the shift toward semantic, language-independent techniques—laying the foundation for systems like the one proposed in this paper.

III. METHODOLOGY

The proposed system follows an unsupervised pipeline to cluster and retrieve multilingual text documents in English, Hindi, and Telugu. The pipeline includes six main stages: language detection, language-specific preprocessing, semantic embedding generation, clustering, visualization, and semantic-based retrieval.

A. Language Detection

Each input document is first converted from PDF to raw text. The system then applies an automatic language detection algorithm using the Lang detect Python library. Based on the textual characteristics, each document is classified as English (EN), Hindi (HI), or Telugu (TE). This step is crucial to apply accurate preprocessing routines tailored to each language.

B. Preprocessing

Text preprocessing prepares the raw input for semantic analysis. It includes:

1) Sentence segmentation: Splitting paragraphs into individual sentences.
2) Tokenization: Breaking sentences into words or subwords.
3) Stopword removal: Eliminating frequent but non-informative words.
4) Punctuation and case normalization: Removing special characters and converting to lowercase.
5) Lemmatization (English only): Reducing words to their root forms using WordNetLemmatizer.

Preprocessing uses NLTK for English and Indic NLP Toolkit for Hindi and Telugu.

C. Semantic Embeddings

To capture the meaning of each document, the system uses a multilingual Sentence-BERT model (paraphrase-multilingual-MiniLM-L12-v2). Each document Di is transformed into a fixed-length vector $\vec{E}(Di) \in R^{384}$, where 384 is the embedding dimension:

$$\vec{E}(Di) = f(Di)$$

Here, f(·) represents the Sentence-BERT embedding function. Documents with similar meaning are positioned close to each other in the embedding space—even if they are written in different languages.

D. Document Clustering

To group similar documents, the system uses Agglomerative Hierarchical Clustering based on cosine distance between embedding vectors. Cosine similarity between two documents Di and Dj is computed as:

$$\text{cos\_sim}(Di,Dj) = E(D_i).\ E(D_j) \ / \ \| E(D_i)\| \ \| E(D_j)\|$$

Cosine distance (used for clustering) is: $\text{cos\_dist}(Di,Dj) = 1 - \text{cos\_sim}(Di,Dj)$

The algorithm starts by treating each document as a separate cluster and then merges the most similar pairs until a desired number of clusters is reached.

E. Visualization

To assess clustering effectiveness, the high-dimensional document vectors are projected into 2D space using UMAP (Uniform Manifold Approximation and Projection). Each point in this space represents a document, and color-coded clusters reveal topic groupings. Additionally, cosine similarity matrices are visualized as heatmaps to observe document similarity patterns.

Let $\vec{E}(Di)$ be the original embedding and $U(Di) \in R2$ be its 2D projection:

$$U(Di) = \text{UMAP}(E(Di))$$

F. Semantic Search

The system includes a semantic search feature. A user query Q is embedded as $\vec{E}(Q)$, and cosine similarity is calculated between the query and each document $\vec{E}(Di)$. Top documents with highest similarity scores are retrieved:

$$\text{cos\_sim}(Q,Di) = E(Q).\ E(D_i) \ / \ \| E(Q)\| \ \| E(D_i)\|$$

This allows retrieval of contextually relevant documents even across different languages.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

IV. RESULTS AND EVALUATIONS

This section presents the performance evaluation of the proposed multilingual clustering and semantic search system. Results are reported for language detection, clustering quality, visualization, and semantic search accuracy using standard metrics and graphical outputs.

A. Language Detection Performance

A total of 15 multilingual documents (5 English, 5 Hindi, 5 Telugu) were evaluated. The system accurately identified the language of each document using the Lang detect library.

1) Total Documents: 15
2) Detection Accuracy: 100%
3) Confusion Matrix: No misclassification across EN, HI, and TE

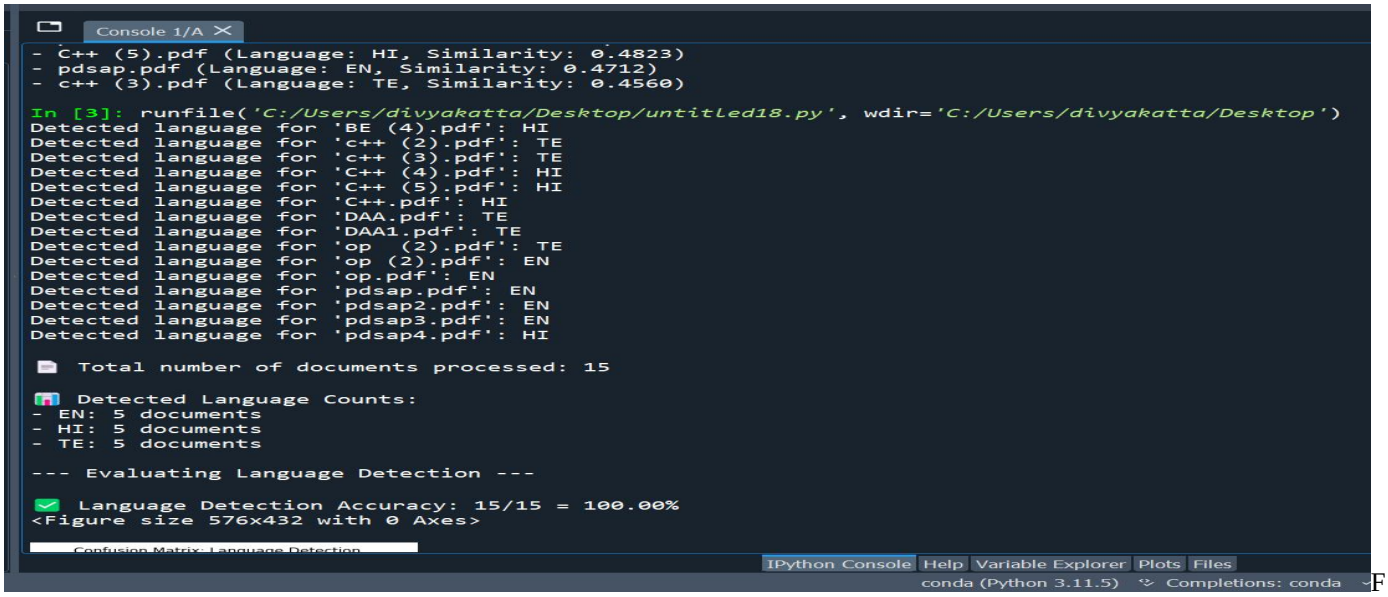


Figure No 1-Language Detection Output and Document Count Summary.

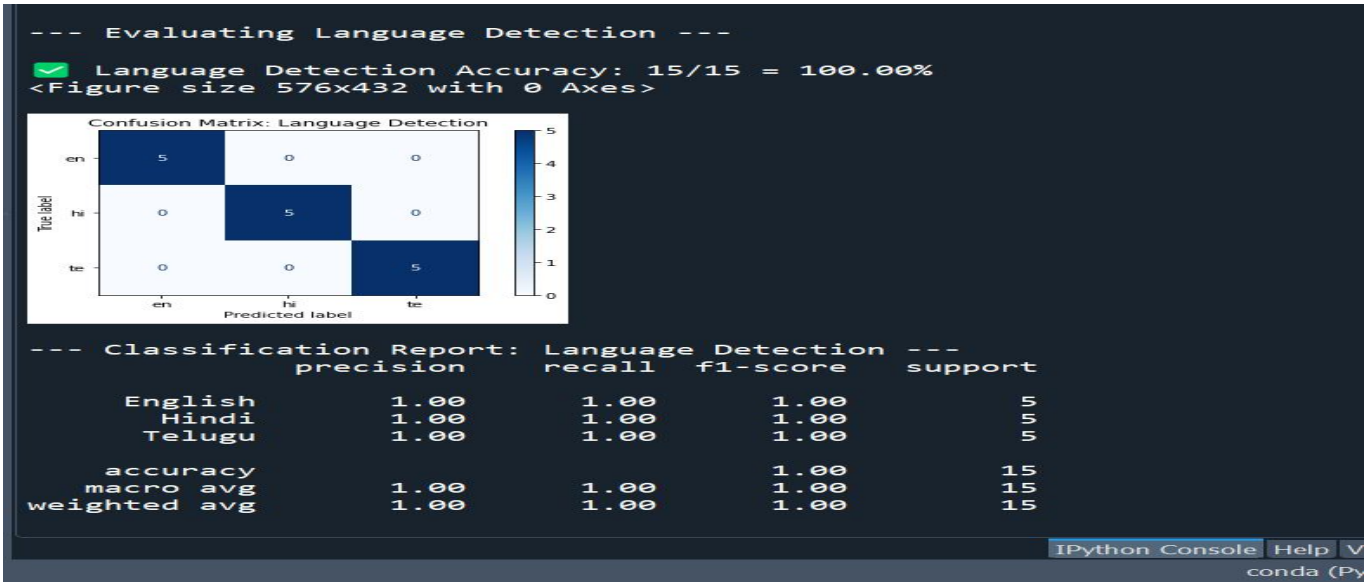


Figure No 2- Confusion Matrix and Classification Report for Language Detection.

B. Clustering Evaluation

The proposed system applies Agglomerative Clustering on semantic embeddings to group multilingual documents based on content similarity. The evaluation includes cluster count, Silhouette Score, UMAP visualization insights, and intra-cluster cosine similarity. For the English documents, the system formed three clusters with a Silhouette Score of 0.3366, indicating reasonably well-separated groups. Cluster 0 contained *pdsap2.pdf* and *pdsap3.pdf*, Cluster 1 included *op.pdf* and *op (2).pdf*, and Cluster 2 had *pdsap.pdf*. The UMAP visualization showed clear separation among clusters, particularly between documents discussing similar technical topics. The highest cosine similarity was observed between *op.pdf* and *op (2).pdf*, scoring 0.8531, confirming strong semantic closeness.

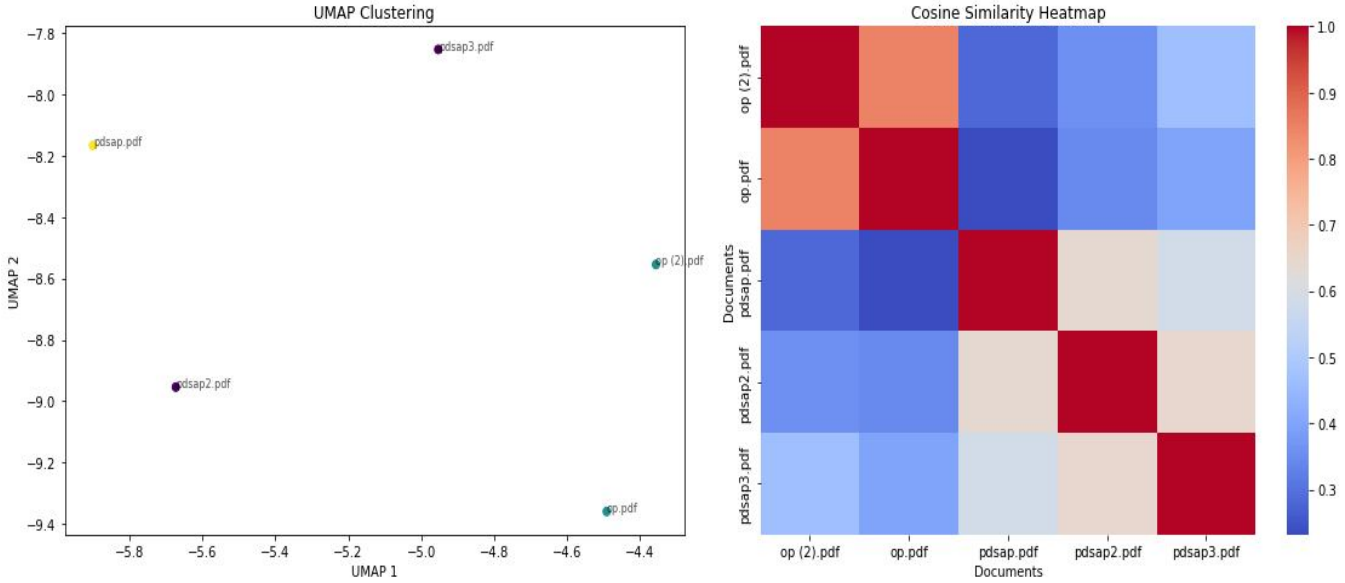


Figure No 3-UMAP-Based Cluster Visualization and Cosine Similarity Heatmap for English Documents

In the case of Hindi documents, the clustering algorithm again formed three clusters with a Silhouette Score of 0.2502. Cluster 0 grouped *C++ (4).pdf*, *C++ (5).pdf*, and *C++.pdf*, indicating a shared programming theme. Cluster 1 included *pdsap4.pdf*, and Cluster 2 contained *BE (4).pdf*. The UMAP output displayed a dominant cluster with two smaller outliers. The cosine similarity between *C++ (5).pdf* and *C++.pdf* was **0.8476**, indicating moderate semantic similarity within the main cluster.

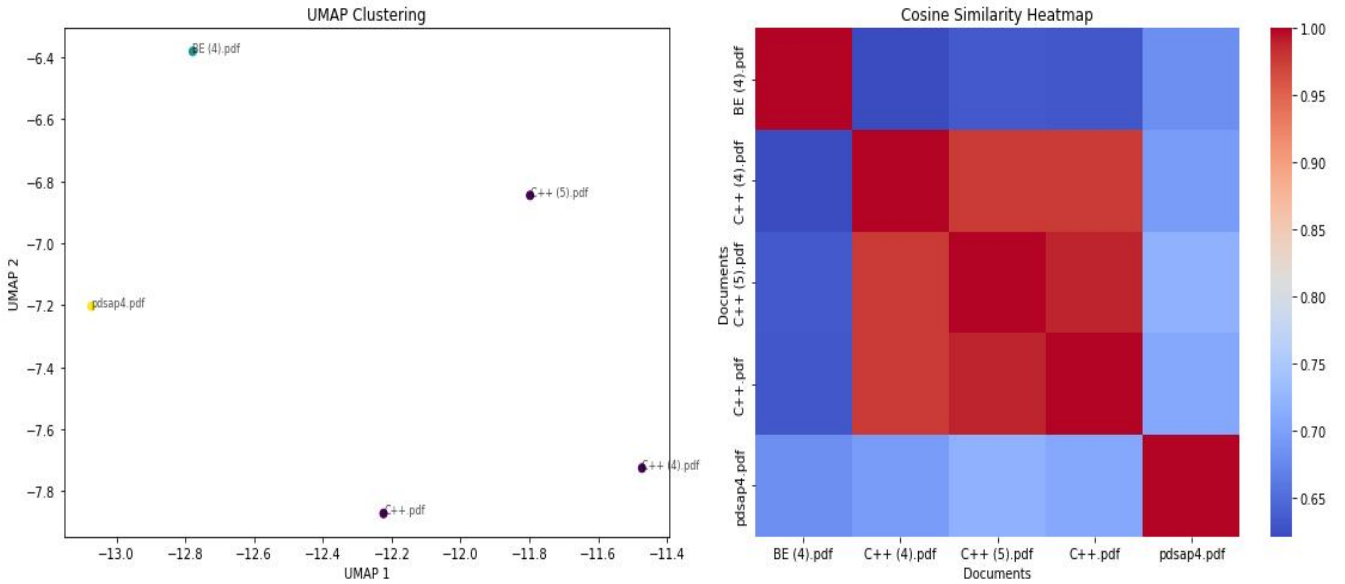


Figure No 4-UMAP-Based Cluster Visualization and Cosine Similarity Heatmap for Hindi Documents

For the Telugu documents, the system achieved the best clustering performance with a Silhouette Score of **0.4107**, showing strong cluster cohesion. Cluster 0 consisted of *c++ (2).pdf* and *c++ (3).pdf*, while Cluster 1 grouped *BE (3).pdf* and *pdsap (3).pdf*. *BE (2).pdf* formed a separate cluster (Cluster 2). UMAP visualizations revealed clear thematic grouping with minimal overlap. The highest cosine similarity was found between *c++ (2).pdf* and *c++ (3).pdf*, scoring 0.9124, indicating strong semantic alignment

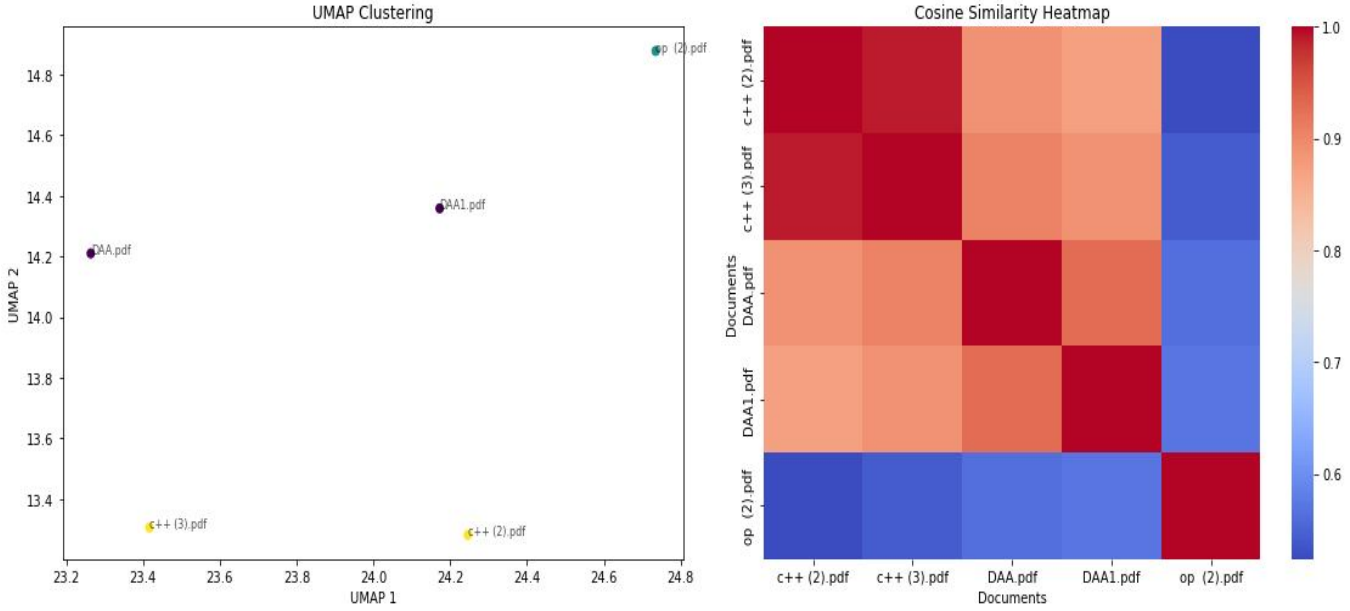


Figure No 5-UMAP-Based Cluster Visualization and Cosine Similarity Heatmap for Telugu Documents

C. Semantic Search Results

Semantic search was tested with the cross-lingual query:

Query: *"Operating System Concepts"*

Top 3 Results

Rank	File Name	Detected Language	Cosine Similarity
1	op.pdf	EN	0.5933
2	op (2).pdf	EN	0.5628
3	op (2).pdf	TE	0.4968

Table No 1-Top 3 retrieved documents for the query "Operating System Concepts" with cosine similarity**.**

D. Metric Summary

Task	Metric Used	Value(s)
Language Detection	Accuracy	100%
Clustering (EN)	Silhouette Score	0.3366
Clustering (HI)	Silhouette Score	0.2502
Clustering (TE)	Silhouette Score	0.4097
Semantic Search	Cosine Similarity	~0.59, ~0.56, ~0.49 (sample search)

Table No 2-Summary of Evaluation Metrics for System Performance

V. CONCLUSIONS

This project successfully developed a multilingual system for clustering and semantic search of text documents in English, Hindi, and Telugu. The primary goal was to enable meaningful grouping and retrieval of documents across languages using a robust, language-aware pipeline. The first phase of the pipeline focused on language detection, which achieved an impressive 100% accuracy, correctly identifying all 15 test documents (5 each in EN, HI, and TE) as per their respective languages. This step ensured

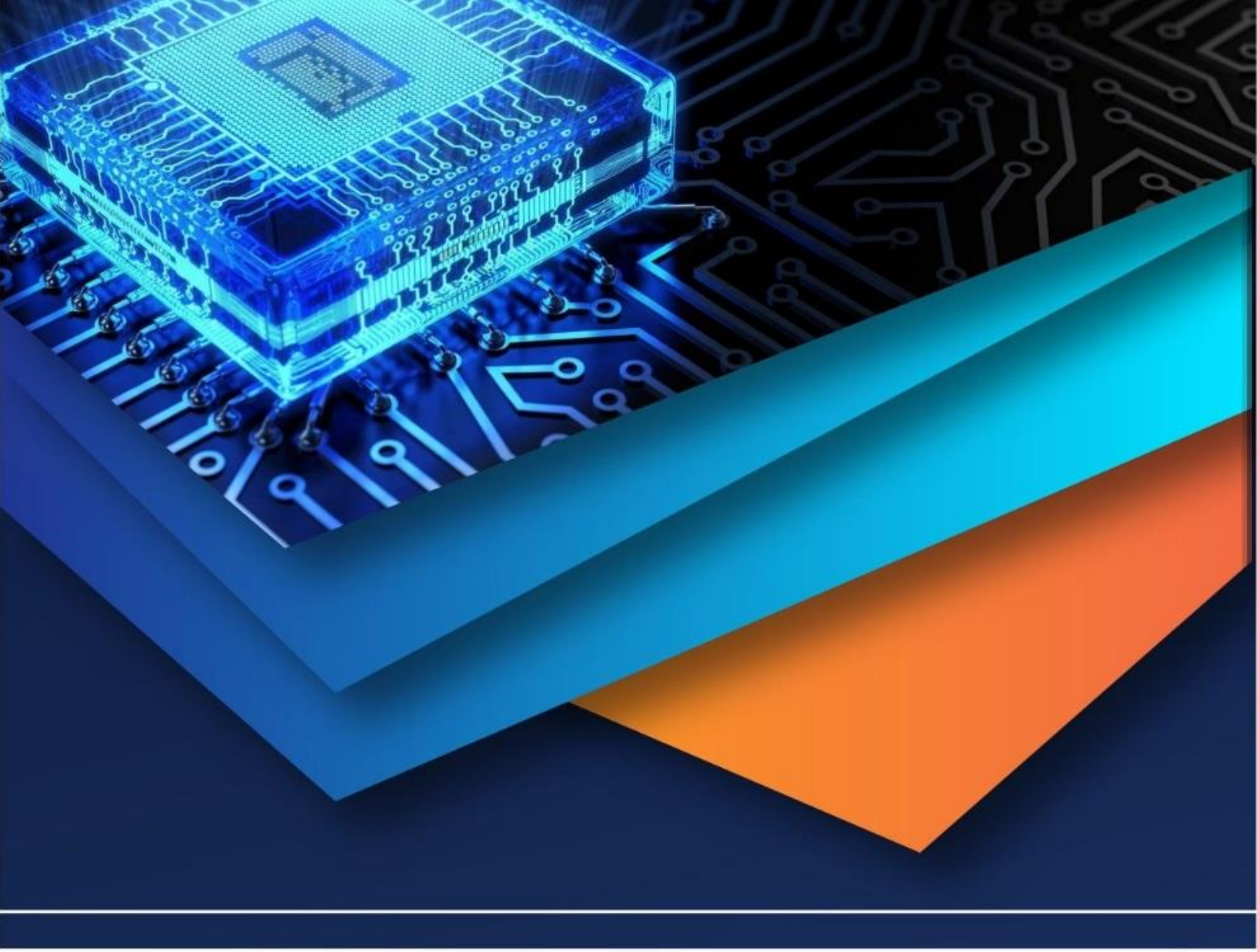
that subsequent processing was language-specific and tailored to the document content. In the second phase, Agglomerative Clustering was applied to semantically embed and group the documents within each language. The clustering quality was evaluated using the Silhouette Score, which ranged from 0.25 to 0.41:

1) English documents: 0.336
2) Hindi documents: 0.250
3) Telugu documents: 0.409

These values indicate moderately meaningful clustering, with the Telugu documents showing the strongest intra-cluster cohesion. Visualizations using UMAP helped confirm the spatial separation of clusters, while cosine similarity heatmaps supported the content similarity between documents within each cluster. Finally, a semantic search engine was integrated into the system. For a sample query like *"operating system concepts"*, the top 3 retrieved documents spanned across English and Telugu languages with cosine similarity scores of 0.5933, 0.5628, and 0.4968, respectively. This demonstrated the model's ability to find semantically relevant documents across different languages, leveraging BERT embeddings for meaningful comparison.

REFERENCES

[1] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084
[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019).BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.
[3] McInnes, L., Healy, J., & Melville, J. (2018).UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426.https://arxiv.org/abs/1802.03426
[4] Scikit-learn: Machine Learning in Python.Pedregosa, F., et al. (2011). Journal of Machine Learning Research, 12, 2825–2830.
[5] Langdetect - Language Detection Library in Python.https://pypi.org/project/langdetect/
[6] Natural Language Toolkit (NLTK). Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc. https://www.nltk.org
[7] Agglomerative Clustering — scikit-learn Documentation.
[8] NPTEL Online Courses – Video Transcript Dataset Source. https://nptel.ac.in
[9] Indic NLP Library.Kunchukuttan, A. https://github.com/anoopkunchukuttan/indic_nlp_library



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY