SRM UNIVERSITY
AP, Amaravati

Multimodal Clustering Using NLP + Deep Learning

Submitted By

Vijay Krishna P.J (Roll No. AP22110010756)

Supervisor: Dr. Bala Venkateswarlu

Department of Computer Science and Engineering

SRM University AP

Multimodal Clustering Using NLP + Deep Learning

ABSTRACT

This project presents a Multimodal Clustering Framework that performs unsupervised learning on both text and image data using machine learning and deep learning techniques. For the text modality, the BBC News dataset is processed using Natural Language Processing (NLP) techniques such as tokenization, Part-of-Speech (POS) tagging, lemmatization, stopword removal, and TF-IDF vectorization with n-grams. Multiple clustering algorithms including K-Means, Agglomerative, DBSCAN, Spectral Clustering, and Gaussian Mixture Models (GMM) are applied and evaluated using Silhouette Score to determine the best performing model. For the image modality, the CIFAR-10 dataset is processed using a pre-trained MobileNetV2 deep learning model for feature extraction, followed by K-Means clustering for grouping visually similar images. PCA is used for visualization in both modalities. The results confirm that meaningful topic-based clusters are formed for BBC news articles, and visually coherent clusters are generated for CIFAR-10 images. This project demonstrates an efficient and scalable approach for multimodal data organization using unsupervised learning.

Keywords— Multimodal Clustering, TF-IDF, K-Means, MobileNetV2, Deep Learning, Unsupervised Learning, NLP, Image Clustering

I. INTRODUCTION

With the exponential growth of digital content, data is increasingly available in multiple formats such as text, images, audio, and video. Manual organization of such large-scale data is inefficient and impractical. This has motivated the adoption of unsupervised machine learning techniques, particularly clustering, to automatically group similar data instances.

Traditional clustering systems focus on single-modality data, either text or images. However, real-world applications such as news recommendation systems, media indexing platforms, digital libraries, and intelligent search engines require the ability to process and analyze multimodal data jointly.

The current best-known approaches in text clustering involve TF-IDF vectorization, Word Embeddings, and Transformer-based embeddings followed by clustering algorithms like K-Means, Hierarchical Clustering, and DBSCAN. For image clustering, state-of-the-art methods rely on deep convolutional neural networks (CNNs) such as ResNet, VGG, and MobileNet for automatic feature extraction, followed by traditional clustering algorithms.

This project integrates these modern techniques into a unified multimodal clustering framework that handles both textual and visual data efficiently.

II. PROBLEM STATEMENT

The continuous generation of heterogeneous data in the form of text documents and images makes it difficult to organize, search, and analyze content effectively. Existing systems generally operate on a single data modality and require labeled datasets for supervised learning. There is a need for a scalable unsupervised framework that can automatically discover patterns and group both textual and visual data without relying on labeled information.

Hence, the problem addressed in this work is:

To design and implement a multimodal unsupervised clustering system capable of processing and clustering both text documents and images using machine learning and deep learning techniques.

III. NOVELTY CLAIMS

- The system integrates both NLP-based text clustering and deep learning-based image clustering within a single unified pipeline.
- Unlike traditional text clustering systems that use only unigrams, this project uses TF-IDF with unigrams, bigrams, and trigrams for richer semantic representation.
- Automatic best model selection using Silhouette Score is applied instead of manual model selection.

- The image pipeline adopts a pre-trained MobileNetV2 CNN for lightweight and efficient deep feature extraction.
- The entire pipeline supports low-RAM execution, making it suitable for standard personal computers and laptops.
- Automatic PDF report generation is integrated directly into the project.

Novelty 1: Use of POS-based filtering (Nouns + Verbs) combined with TF-IDF n-grams for enhanced topic modeling.
Novelty 2: MobileNetV2-based lightweight deep feature extraction optimized for CPU-based systems.

IV. PROPOSED METHODOLOGY

A. Text Clustering Pipeline

Step 1: Load BBC News Dataset from compressed archive
Step 2: Text Preprocessing

- Lowercasing
- Special character removal
- Tokenization
- POS filtering (Nouns & Verbs)
- Stopword removal
- Lemmatization

Step 3: Feature Extraction using **TF-IDF (n-grams 1–3)**
Step 4: Clustering using:

- K-Means
- Agglomerative Clustering
- DBSCAN
- Spectral Clustering
- Gaussian Mixture Model (GMM)

Step 5: Model Evaluation using **Silhouette Score**
Step 6: Automatic Cluster Naming using **Top TF-IDF Keywords**
Step 7: Visualization using **PCA**
Step 8: Report Generation using **ReportLab (PDF)**

B. Image Clustering Pipeline

Step 1: Load CIFAR-10 Images
Step 2: Image Preprocessing

- Resize to 224×224
- Normalization
- Tensor conversion

Step 3: Deep Feature Extraction using MobileNetV2 (Pre-trained)
Step 4: Clustering using K-Means
Step 5: Visualization using PCA

Algorithm (Text Clustering - Simplified Pseudocode)

1. Load all text documents
2. Preprocess text
3. Compute TF-IDF features
4. Apply clustering algorithms
5. Compute Silhouette Score
6. Select best clustering model
7. Extract top keywords per cluster
8. Visualize clusters
9. Save results and generate report

V. DATASET DESCRIPTION

1. BBC News Dataset (Text Dataset)

- Source: Kaggle
- Categories: Business, Politics, Sports, Technology, Entertainment
- Format: Text files (.txt)
- Usage: Topic discovery using TF-IDF and clustering algorithms

2. CIFAR-10 Dataset (Image Dataset)

- Source: Torchvision Library
- Classes: Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, Truck
- Number of Images: 50,000 (training)
- Size: 32 × 32 RGB images
- Usage: Deep feature-based unsupervised image clustering

VI. RESULTS & DISCUSSION

1) Text Clustering Results

- Among all clustering algorithms, K-Means achieved the highest Silhouette Score.
- The automatically extracted keywords clearly matched BBC News topics such as:

- Politics
- Sports
- Business
- Technology
- Entertainment

- PCA visualization confirmed **well-separated topic clusters**.

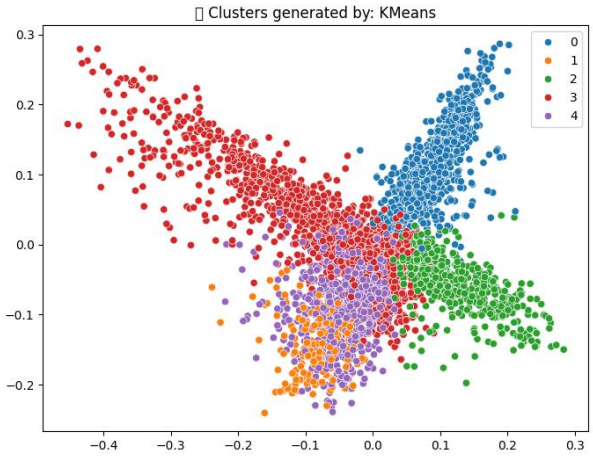

Figure 6.1: K-Means

Figure 6.1 shows the clustering output generated by the **K-Means algorithm**, where documents are grouped into five distinct clusters based on textual similarity.

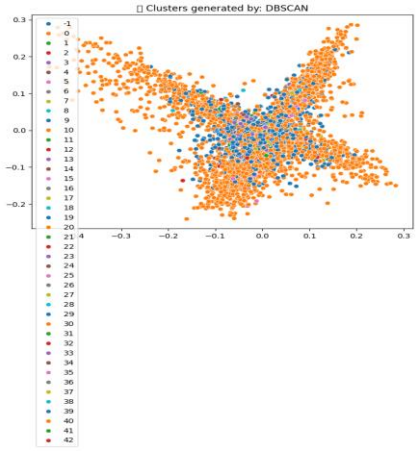

Figure 6.2: DBSCAN

Figure 6.2 represents the clusters formed using **DBSCAN**, which additionally identifies noise and outlier documents.

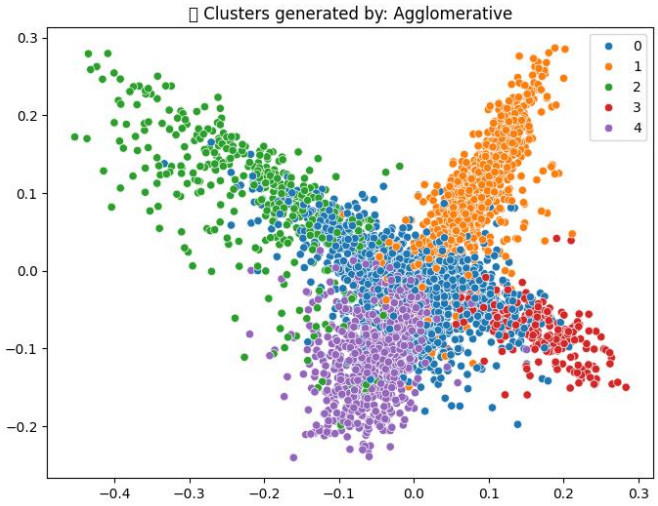

Figure 6.3 Agglomerative

Figure 6.3 illustrates the results of Agglomerative Hierarchical Clustering, where a slightly overlapping yet meaningful cluster structure can be observed.

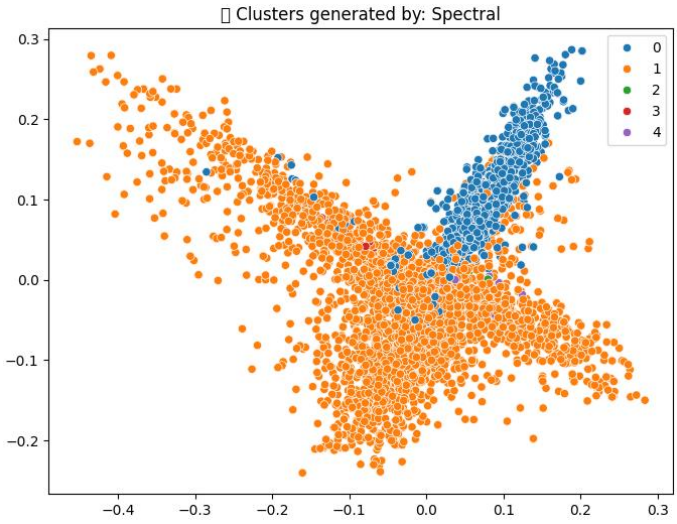

Figure 6.4 Spectral

Figure 6.4 presents the Spectral Clustering results, which show clear nonlinear separations among clusters.

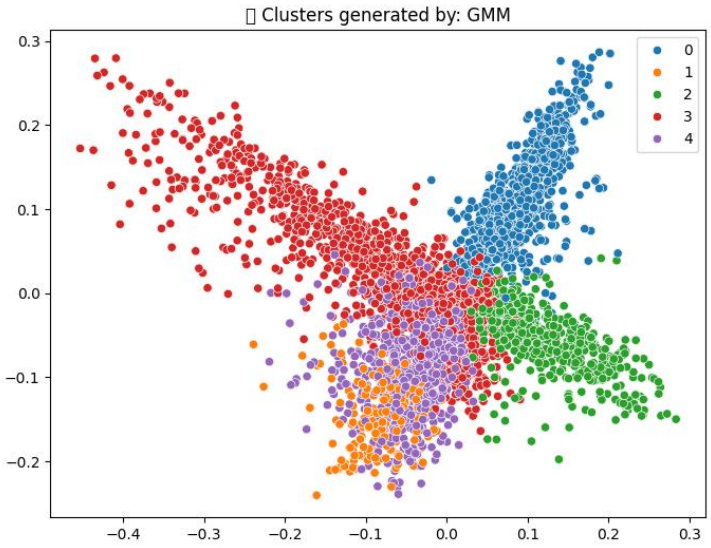

Figure 6.5 GMM

Figure 6.5 displays the output of the **Gaussian Mixture Model (GMM)**, where documents are assigned to clusters probabilistically.

2) Image Clustering Results

- MobileNetV2 successfully extracted deep visual features from CIFAR-10 images.
- K-Means clustering grouped images into visually meaningful clusters such as:
  - Animals
  - Vehicles
  - Ships and Aircraft
- PCA visualization showed clear spatial separation of clusters.

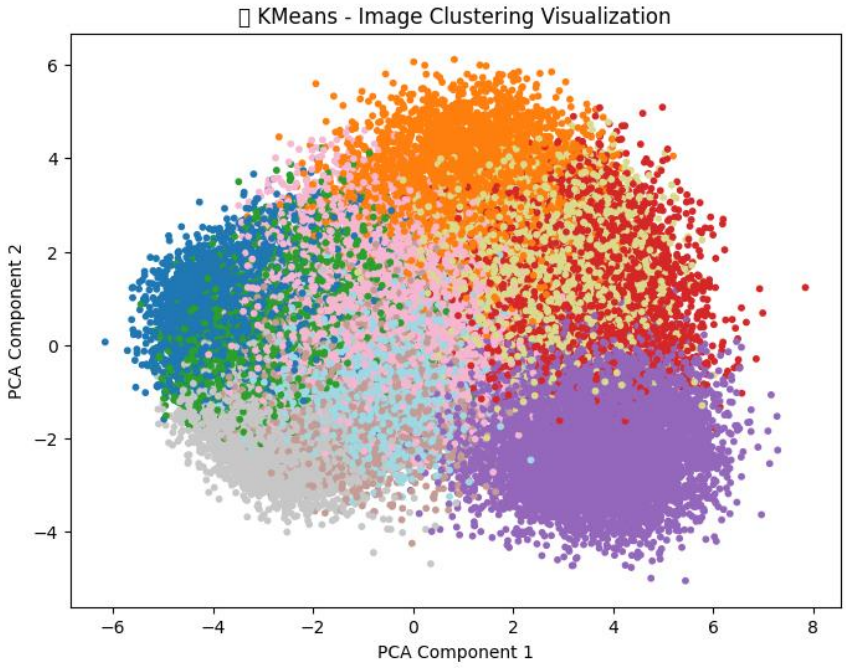

Figure 6.6 KMeans-Image Clustering

Figure 6.6 shows the PCA visualization of K-Means clustering on image features, where visually similar objects are grouped together based on learned deep representations.

3) Observations

- Deep learning-based features significantly improved image clustering quality.
- TF-IDF with POS filtering generated interpretable and meaningful text clusters.
- Multimodal processing enables simultaneous analysis of heterogeneous data.

VII. CONCLUSIONS

This project successfully demonstrates an efficient multimodal clustering framework that integrates classical NLP-based text clustering and deep learning-based image clustering. The use of TF-IDF with advanced preprocessing enables accurate topic discovery from BBC news articles. The application of MobileNetV2 enables efficient and scalable feature extraction from images. The framework operates entirely in an unsupervised manner and produces meaningful results for both modalities. This approach is highly suitable for real-world content organization, digital media indexing, and intelligent information retrieval systems.

VIII. FUTURE WORK

- Integration of Transformer-based text embeddings (Sentence-BERT).
- Joint multimodal embedding using CLIP for cross-modal retrieval.
- Deployment of an interactive web-based dashboard.
- Support for video and audio clustering.
- Large-scale real-time multimodal analytics.

IX. IMPORTANT TAKEAWAYS

- Unsupervised clustering can effectively organize both text and images.
- Deep learning improves feature representation in computer vision tasks.
- PCA visualization helps in understanding high-dimensional data structure.
- Multimodal learning enables unified analysis across heterogeneous data.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[3] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TF-IDF for text classification," in *Proc. 14th Int. Conf. Machine Learning*, 1997, pp. 143–151.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[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] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, 2018.

[7] A. Krizhevsky, "Learning multiple layers of features from tiny images," CIFAR-10 Dataset, 2009.

[8] D. D. Lewis et al., "Reuters-21578 text categorization test collection," 1997.

[9] Kaggle, "BBC News Dataset," [Online]. Available: https://www.kaggle.com/datasets

[10] S. Bengio, "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.