Comparing TF-IDF Embedding & Sentence-BERT Embeddings in Text clustering using K-means ++ algorithm

Final Project - EDS 6346: Data Mining for Engineers

Submitted by: Vikram Koti Mourya Vangara, PSID: 2315018
Gayathri Seelam, PSID: 2297215
Neha Reddy Jakka, PSID: 2296660
Aniketh Bharat, PSID: 2381419

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Project Objective

- Compare clustering performance using traditional TF-IDF vs SBERT embeddings
- Use 20 Newsgroups dataset as base
- Apply K-Means++ clustering algorithm to both embeddings
- Evaluate using standard clustering metrics and visualizations

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation & Relevance

- Why Clustering Matters in NLP
  - Helps organize unstructured text into meaningful groups without labeled data.
  - Enables topic discovery, document categorization, and information retrieval.
  - Essential for tasks like:
    - ➢ News aggregation
    - ➢ Customer feedback analysis
    - ➢ Scientific article grouping

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation & Relevance

Importance of Comparing Traditional vs Modern Embeddings

- TF-IDF (Traditional):
  - Captures term frequency, but ignores word context.
  - Struggles with semantic similarity.

- Sentence-BERT (Modern):
  - Embeds whole sentences in semantic space.
  - Captures contextual meaning using transformer-based models.

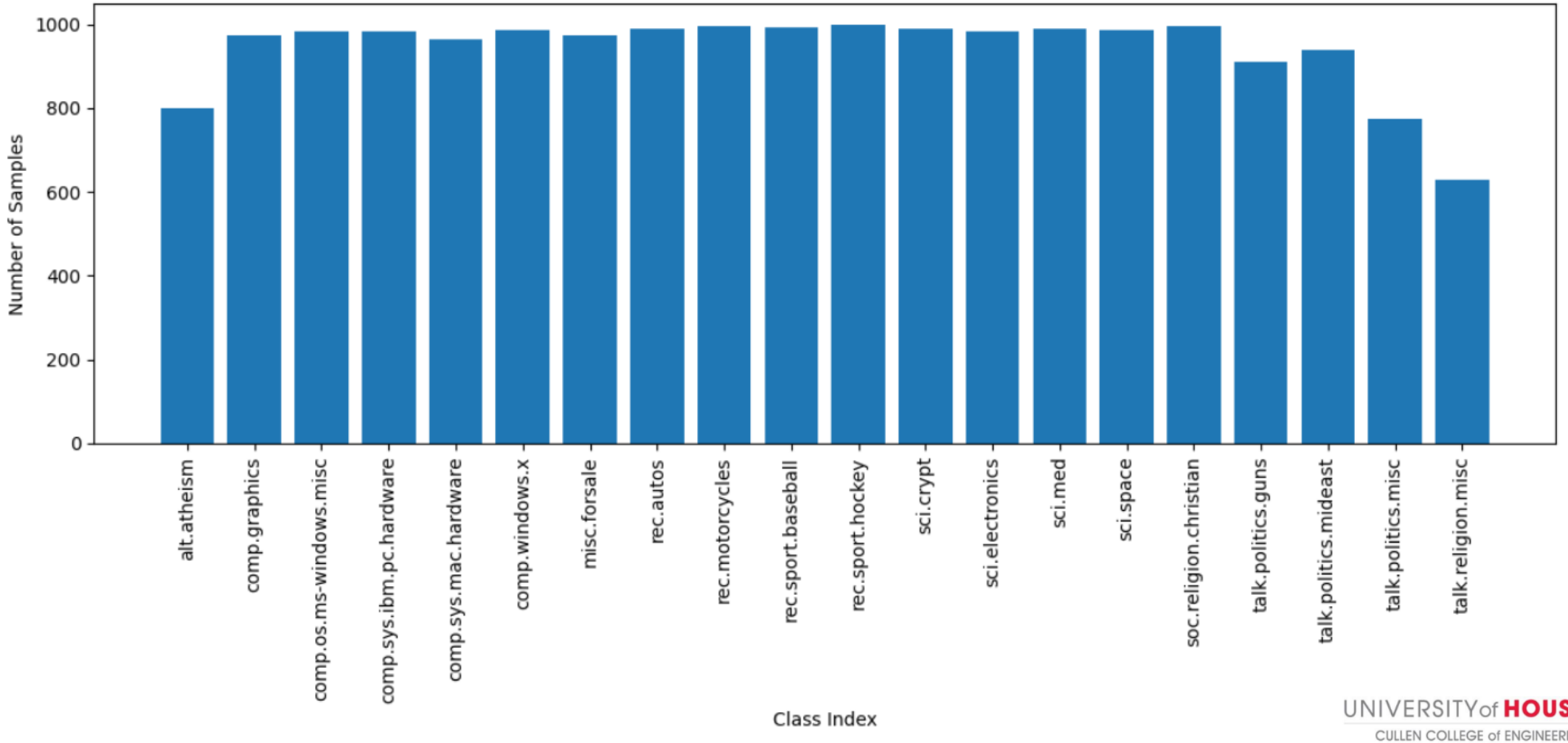
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Dataset Description: 20 Newsgroups

- Consists of approximately 18,000 newsgroup posts across 20 topics
- Covers diverse areas like sports, politics, science, and technology
- Used for benchmarking text classification and clustering algorithms
- Cleaned by removing headers, footers, and quotes

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Class Distribution of the Dataset



UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Text Preprocessing Steps

- Lowercased all text
- Removed headers, footers, and quotes
- Stripped whitespace and punctuation
- Prepared clean inputs for embedding generation

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

TF-IDF Embedding

- Traditional approach for text vectorization
- Captures term frequency-inverse document frequency
- Results in high-dimensional sparse vectors
- Used as baseline in this comparison

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Sentence-BERT (SBERT)

- Transformer-based LLM for dense semantic embeddings
- Model used: all-MiniLM-L6-v2
- Captures sentence-level meaning beyond word co-occurrence
- Improves semantic understanding in clustering

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

K-Means++ Clustering

- Improved initialization over standard K-Means
- Minimizes intra-cluster distance
- Used on both TF-IDF and SBERT embeddings

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Evaluation Metrics

- Silhouette Score: Cluster separation and cohesion
- Adjusted Rand Index (ARI): Agreement with true labels
- Normalized Mutual Info (NMI): Info shared between labels and clusters

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

TF-IDF + KMeans++ Results

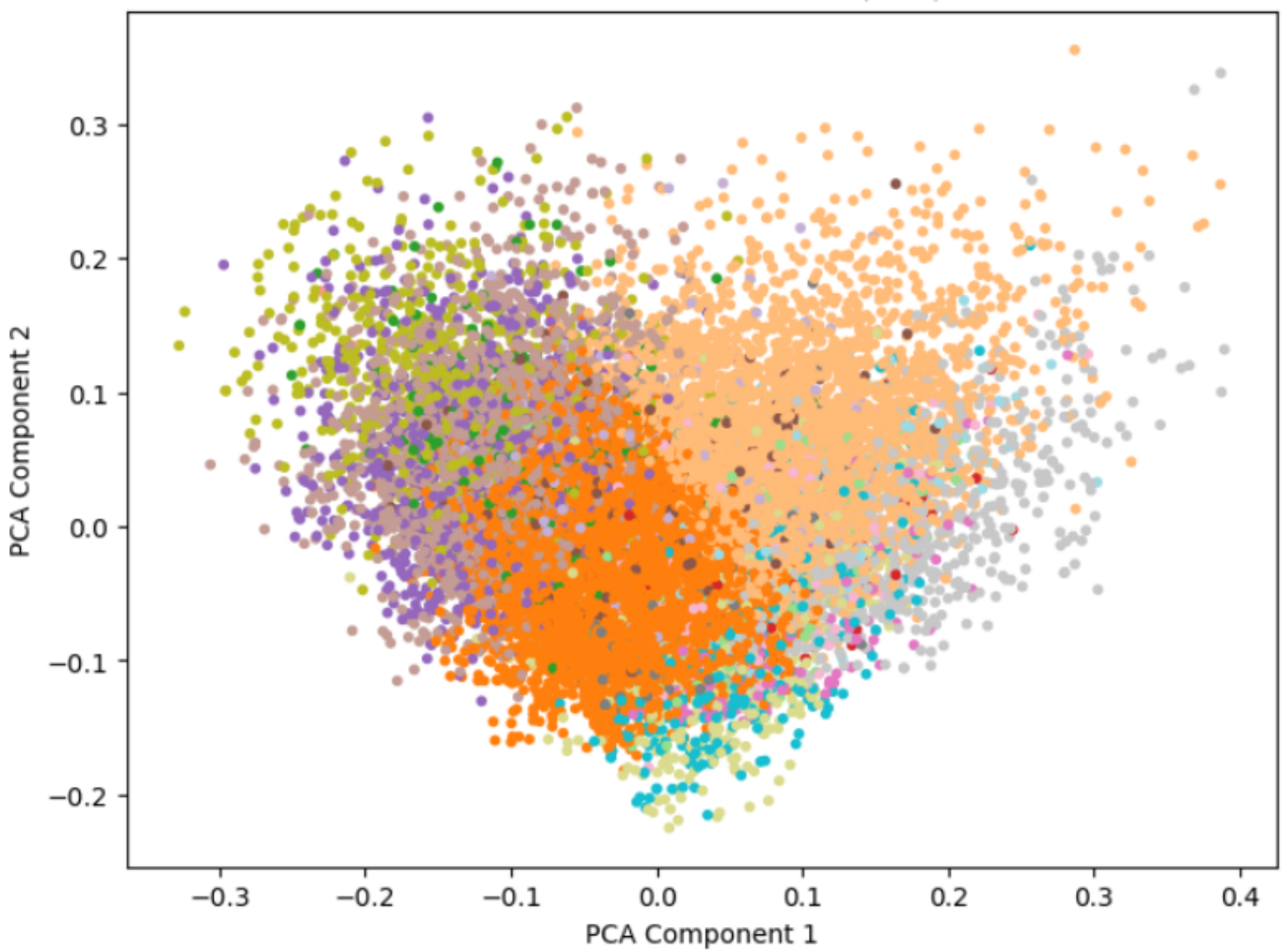
- Silhouette Score: 0.004977
- ARI: 0.026
- NMI: 0.193

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

SBERT + KMeans++ Results

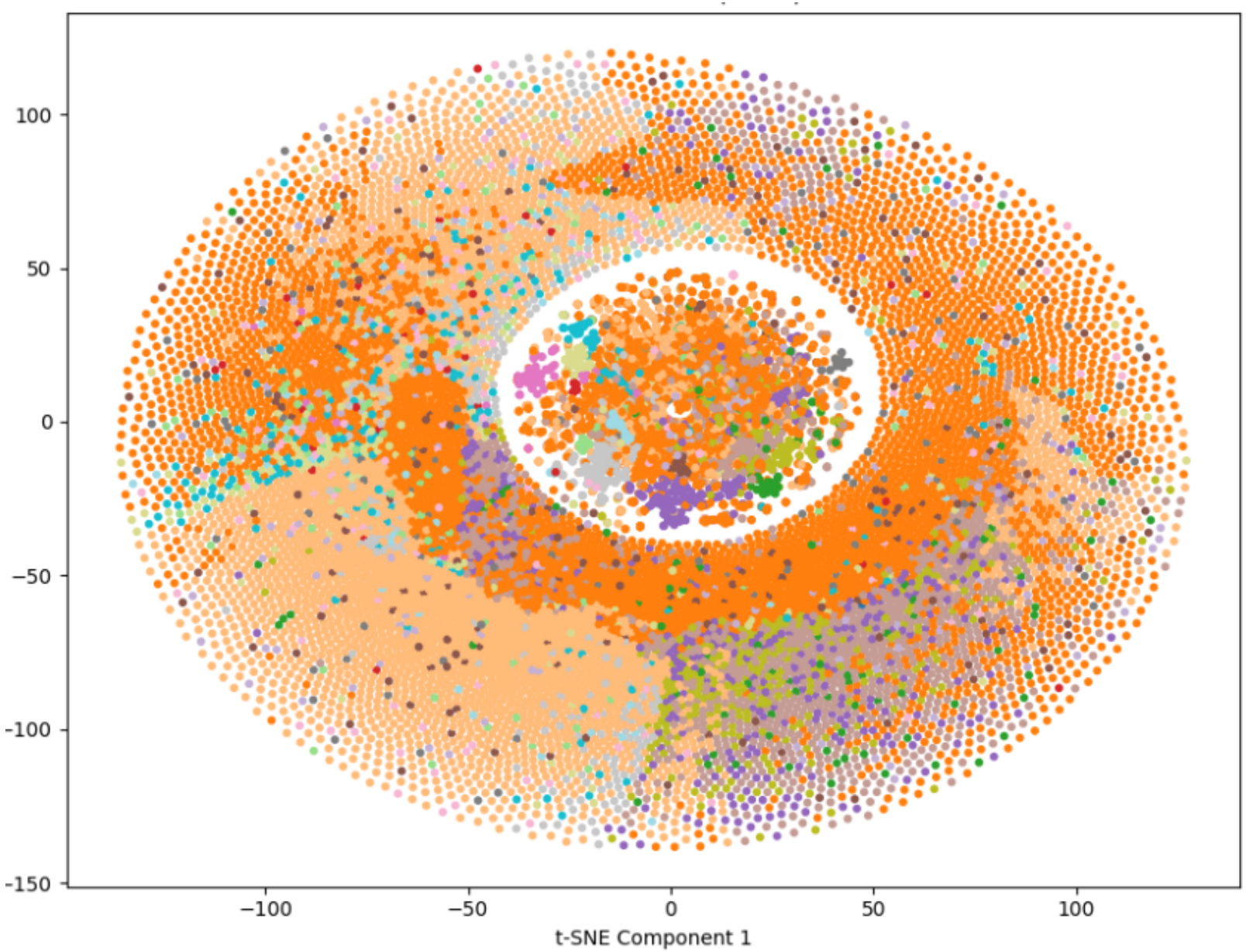
- Silhouette Score: 0.059
- ARI: 0.379
- NMI: 0.528

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING



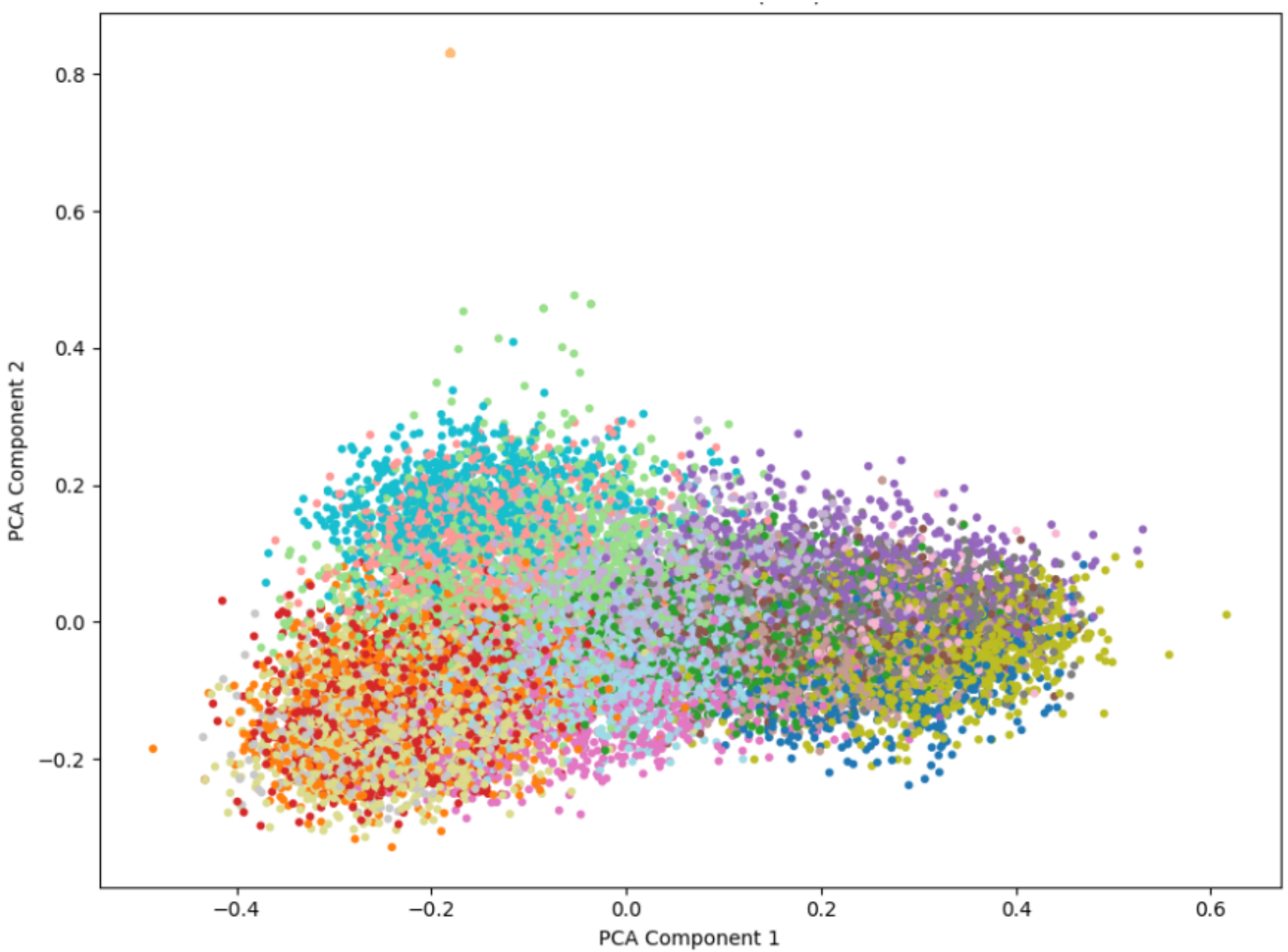
TF-IDF Clusters Visualization - PCA

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING



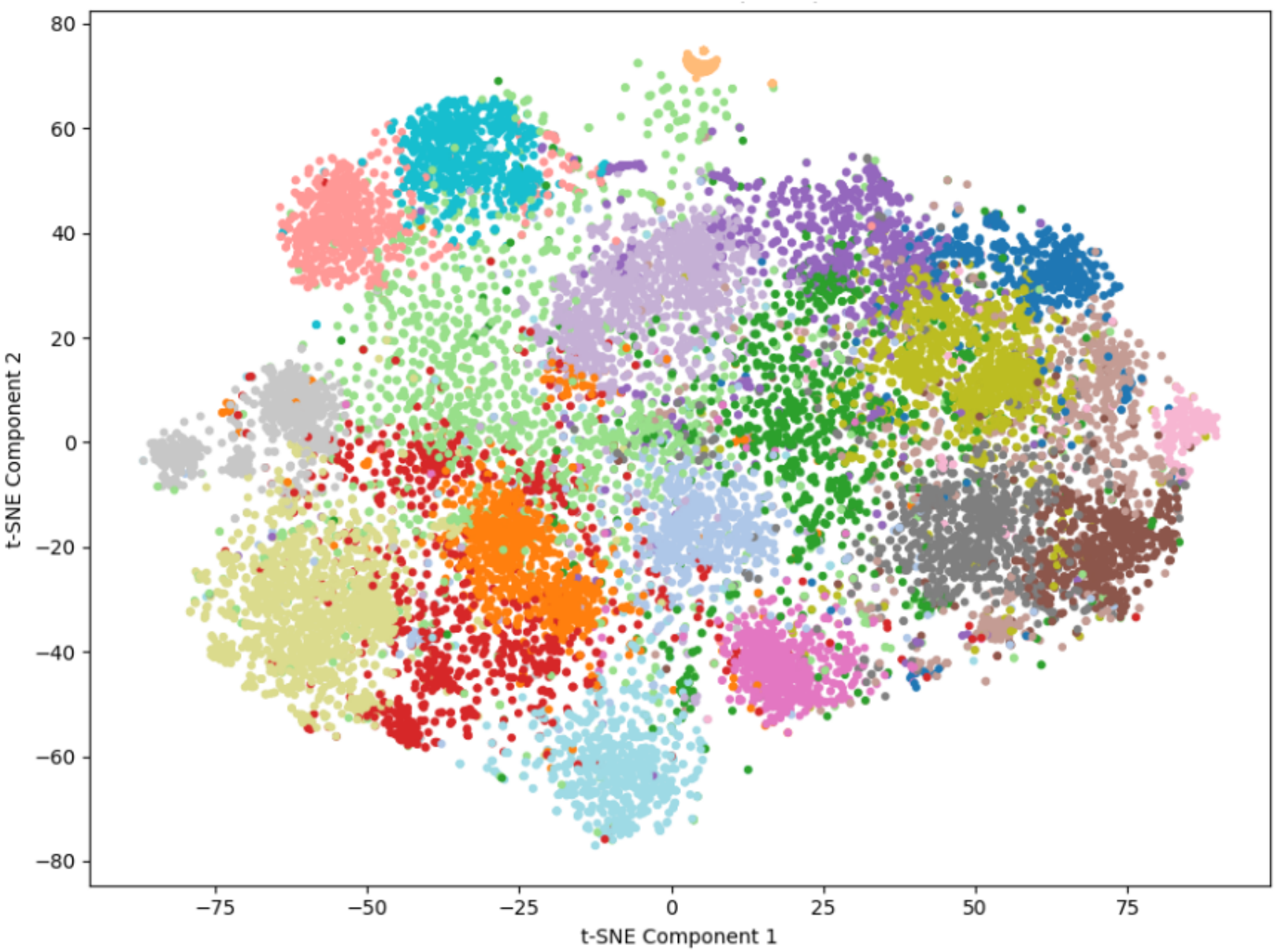
TF-IDF Clusters Visualization – t-SNE

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING



BERT Clusters Visualization - PCA

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING



BERT Clusters Visualization – t-SNE

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Future Work

Exploring Other Embedding Models

- Falcon (TII): Open-source transformer model optimized for inference.
- LLaMA (Meta): Lightweight LLM with strong performance in low-resource scenarios.
- OpenAI Embeddings: Potential for deeper semantic clustering, especially in zero-shot settings.

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Future Work

Applying Alternative Clustering Algorithms

- Agglomerative Hierarchical Clustering (AHC)
  - Builds tree-based clusters to capture nested structure.
- Spectral Clustering
  - Useful for non-convex clusters using graph-based partitioning.
- Fuzzy C-Means
  - Allows soft clustering: a data point can belong to multiple clusters with probabilities.

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Conclusion

- SBERT embeddings outperform TF-IDF for clustering
- Modern LLM-based embeddings improve unsupervised NLP tasks
- Dense representations capture deeper semantic relationships

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Team Contributions

- **Vikram Koti Mourya Vangara**, PSID: 2315018 – Implemented SBERT embedding, and evaluation metrics for model performance
- **Gayathri Seelam**, PSID: 2297215 - Lead on Preprocessing , responsible for PowerPoint Slides
- **Neha Reddy Jakka**, PSID: 2296660 - Implemented TF-IDF Vectorization and clustering using K-Means++ algorithm for both embeddings
- **Aniketh Bharat**, PSID: 2381419 -Implemented t-SNE/PCA Visualisations, performance comparison and overall conclusion, responsible for documenting the Jupyter Notebook

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING