Unsupervised Clustering of News Articles Using BERT Embeddings

1. Introduction & Problem Statement

In today's media-saturated landscape, the ability to distill and categorize large volumes of unstructured news content is essential. Manual labeling is costly and time-consuming, especially when dealing with dynamic or domain-specific corpora. Hence, we explore **unsupervised clustering** as a means to organize short news articles without prior labels.

Use Case
We aim to cluster articles from the **AG News dataset** to:

- Help editors rapidly identify trending topics.
- Assist analysts in discovering new or shifting themes over time.
- Enable topic-aware filtering or recommendation systems.

2. Dataset Overview

Dataset: AG News Corpus (via HuggingFace `ag_news` loader)

- **Composition:** English-language news headlines and short descriptions.
- **Classes:** 4 labeled categories (World, Sports, Business, Sci/Tech) — not used in clustering.
- **Used Subset:** 1000+ entries sampled for efficiency in experimentation.

The dataset is pre-tokenized and well-suited for semantic analysis, given the richness of short, information-dense entries.

3. Methodology

3.1 Text Preprocessing

A comprehensive preprocessing pipeline was implemented:

- Lowercasing
- Removal of punctuation, digits, and URLs
- Stopword filtering (NLTK)
- Lemmatization

This normalization ensures consistent embedding quality and semantic focus.

3.2 Text Embedding with BERT

We used **SentenceTransformer's all-MiniLM-L6-v2** model to convert each sentence into a 384-dimensional vector:

- Captures nuanced semantics
- More expressive than TF-IDF or word2vec
- Suitable for downstream distance-based clustering

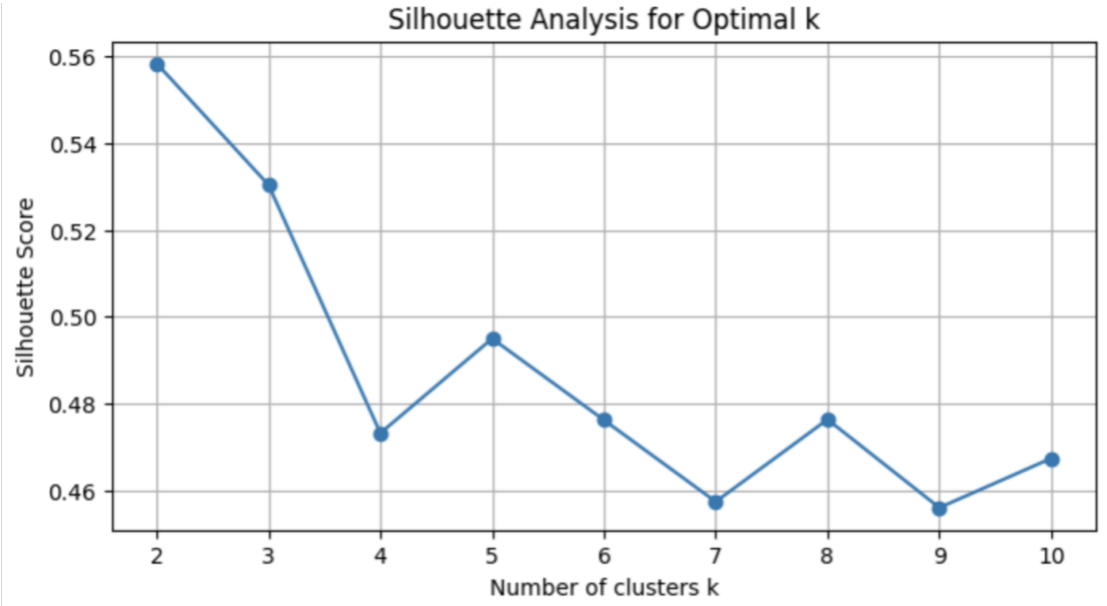
3.3 Dimensionality Reduction via UMAP

To prepare data for clustering and visualization:

- UMAP reduced vectors to **10D** (for clustering) and **2D** (for plotting).
- Parameters: n_neighbors=15, min_dist=0.0, metric='cosine'
- Preserves local neighborhood structure better than PCA or t-SNE

3.4 Clustering (KMeans)

- KMeans was applied on the 10D embeddings.
- **Optimal cluster count (best_k)** selected via **Silhouette Score** across k=2 to k=10.
- Final chosen value of k achieved silhouette score ≈ *0.5582* , indicating well-separated clusters.

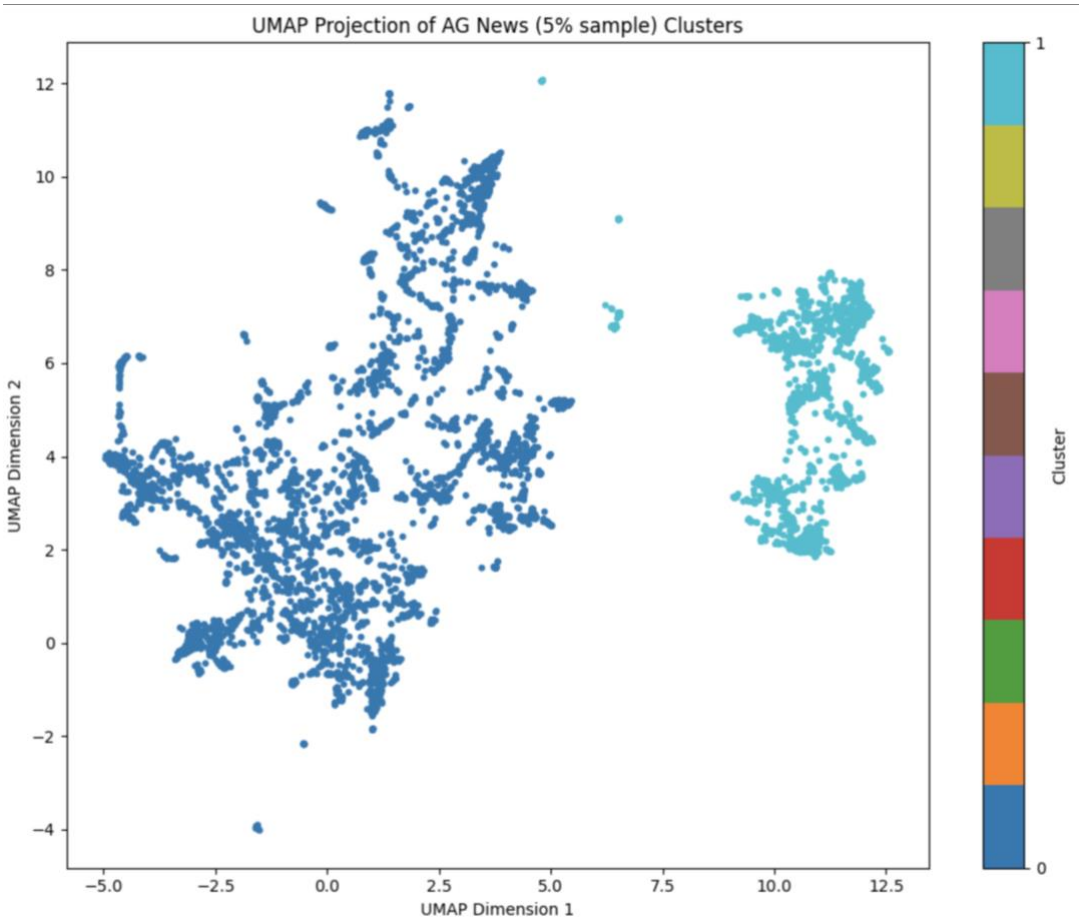



```
Best k by max silhouette score: 2
```

4. Evaluation & Visualization

- **Silhouette Score:** Quantified cohesion/separation; peak observed at k= 2

- **UMAP Scatter Plot:** Visualized in 2D with distinct cluster colors



Clusters were coherent and reflected topical similarity in the news texts.

5. Results

Representative entries in each cluster showed that:

- Clusters grouped semantically similar articles
- Redundancy was reduced; hidden themes were surfaced.

This supports the utility of transformer-based embeddings for unsupervised topic discovery.

6. Conclusion

This project demonstrates the power of BERT embeddings and unsupervised learning to uncover latent topics in text data. By reducing dimensionality and clustering semantically, we've shown that it's feasible to organize complex corpora like news datasets without human labels.