Data Science and Machine Learning

Clustering Documents with Partial Similarity

Instructor: Shynar Akhmetzhanova
Done by: Dina Kantayeva

Astana, 2025.

Introduction

Nowadays identifying semantically similar documents has become a main task for plagiarism detection applications. Traditional similarity-based techniques often fail to detect documents that are partially paraphrased. Recent developments in natural language field using transformer-based models (Sentence-BERT (SBERT), allow to find semantic relationships between documents.

The objective of this project is to develop an unsupervised pipeline to cluster documents with partial semantic similarity using **SBERT** embeddings and **density-based clustering** algorithms.

The aims are:

- Preprocess and vectorize a collection of potentially similar documents (3000+ samples)
- Apply clustering algorithms (DBSCAN, KMeans) to group related documents
- Evaluate the quality of clustering by internal metrics (Silhouette Score)
- Explore semantic patterns within discovered clusters

Database

The dataset consists of a collection of suspicious documents (3000+) that are paraphrased, summarized, or directly copied variations, generated by AI tools. Each row in the dataset represents a single document. The goal is to identify and cluster documents that exhibit **partial semantic similarity**.

The dataset is derived from the PlagBench corpus (2024), an open-source benchmark developed to evaluate the robustness of text similarity detection models. GitHub repository: https://github.com/Brit7777/plagbench

I extracted and used only the *susp_doc* column from the original *plagbench_evaluation_set.csv*, which contains suspicious (potentially plagiarized or paraphrased) documents.

Before clustering, the following preprocessing steps were applied:

1. Duplicate and empty documents were removed.
2. All text was converted to lowercase to reduce lexical variance.

3. All non-alphanumeric characters (except whitespace) were removed.
4. Multiple spaces and line breaks were replaced with a single space.

Methodology

To address the problem of clustering documents with partial semantic similarity, I used the following methods:

1. **Sentence-BERT (SBERT)** is a transformer-based model that maps sentences to dense vector embeddings which capture deep semantic meaning. I used *all-MiniLM-L6-v2*. Unlike traditional TF-IDF, SBERT handles paraphrasing, rewording, and changes in sentence structure better.
2. **KMeans Clustering** - initially used to explore the effect of different cluster counts. It provides straightforward partitioning but requires *n_clusters* to be specified manually.
3. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** is chosen as the primary clustering method because it does not require predefining the number of clusters and identifies and excludes outlier points.

Data preprocessing/processing/visualising:

1. Cleaned and normalized texts;
2. Removed duplicates;
3. Loaded into a DataFrame for vectorization;
4. Text embedding via SBERT;
5. Clustering approaches – **Kmean** tested with *n_clusters* ranging from 2 to 10, **DBSCAN** applied with *eps*=1.0 and *min_samples*=5;
6. Used **Silhouette Score** to evaluate cluster cohesion and separation;
7. Visualized results using **t-SNE** to reduce dimensionality to 2D.

Used libraries:

Library	Purpose
pandas	Data handling and processing
re (regex)	Text cleaning
sentence-transformers	SBERT embeddings for semantic similarity
scikit-learn	KMeans, DBSCAN, Silhouette Score, t-SNE
matplotlib	Cluster visualization
numpy	Numerical operations

Results

To assess the quality of the document clustering, I used the **Silhouette Score** - a standard internal metric for unsupervised clustering tasks.The Silhouette Score measures how similar an object is to its own cluster compared to other clusters. It ranges from -1 (incorrect clustering) to +1 (well-clustered).

- **susp_docs_cleaned.csv**(text/csv) - 1217924 bytes, last modified: 5/18/2025 - 100% done

```
Saving susp_docs_cleaned.csv to susp_docs_cleaned.csv
```

	susp_doc	cleaned_text
0	We examined a multi-armed bandit with a limite...	we examined a multiarmed bandit with a limited...
1	In an attempt to realize its ambitious goal, W...	in an attempt to realize its ambitious goal wo...
2	Ellen was shopping with her parents at the gro...	ellen was shopping with her parents at the gro...
3	We show that, for a large class of piecewise ...	we show that for a large class of piecewise sm...
4	Cruise, the self-driving car subsidiary of Gen...	cruise the selfdriving car subsidiary of gener...

Algorithm	Parameters	Clusters Found	Silhouette Score
KMeans	n_clusters = 3	3	**0.0525**
DBSCAN	eps = 1.0, min_samples = 5	40 (auto)	**0.136**

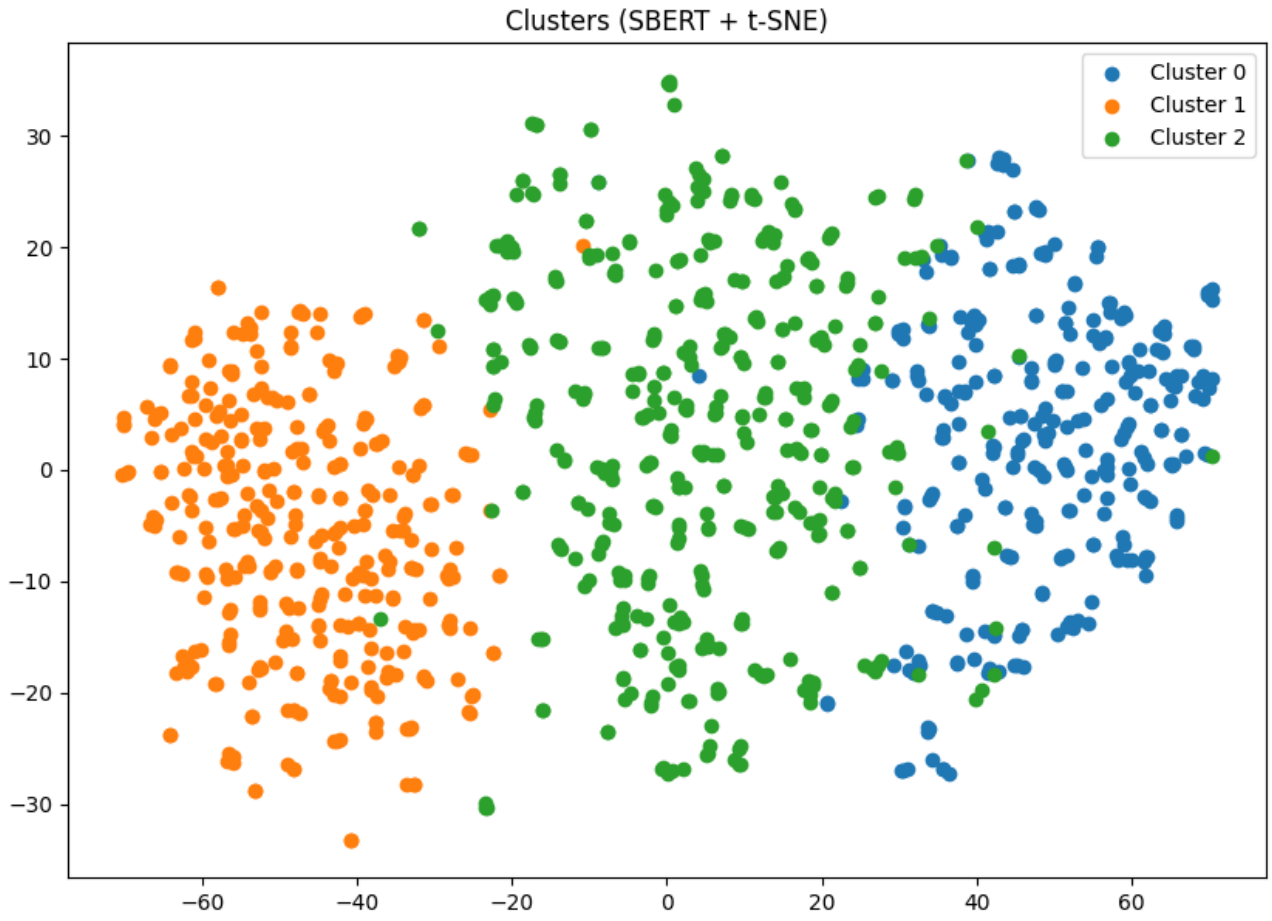

Figure: t-SNE projection of clustered (3) documents using Kmeans

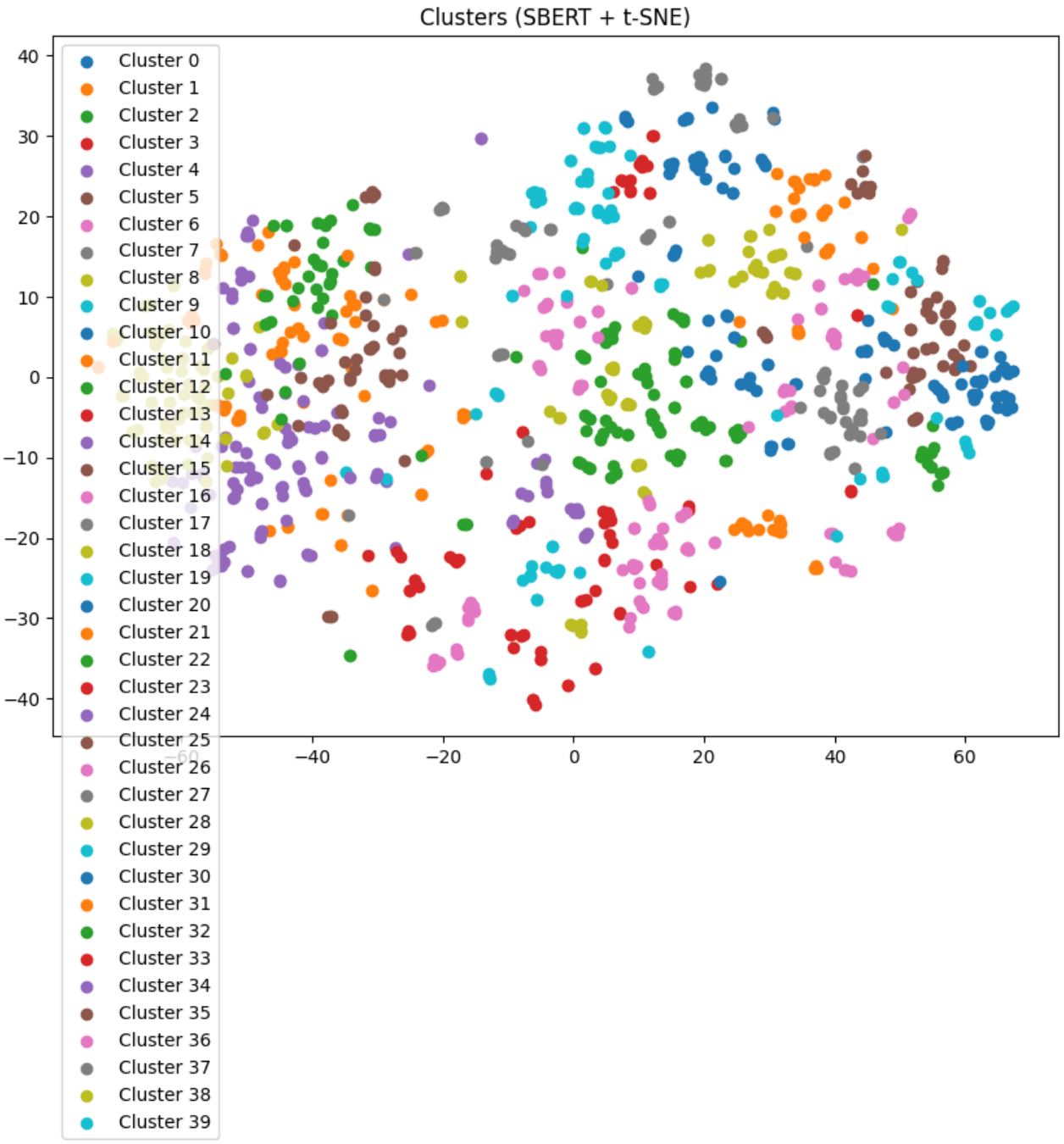

Figure: t-SNE projection of clustered (40) documents using DBSCAN

DBSCAN outperformed KMeans, providing more detailed separation between documents. Also it avoided the need to predefine the number of clusters.

To reduce the 384-dimensional SBERT embeddings to 2D for visualization, I used t-SNE (t-distributed Stochastic Neighbor Embedding). Each point in the plot represents a document, colored by its assigned cluster.

Discussion

The results demonstrate that clustering documents based on partial semantic similarity is feasible using SBERT embeddings combined with density-based clustering algorithms.

DBSCAN proved effective in discovering meaningful semantic clusters without requiring the number of clusters to be specified in the beginning. A silhouette score of **0.136** suggests modest but significant internal cohesion among the clustered documents since the high semantic variability of the text data. Visualizations confirmed that many clusters formed around common narrative structures or themes (personal experiences, moral lessons, academic stories).

Examples of documents from the same cluster (cluster #18):

This cluster includes texts with a pronounced narrative structure describing personal situations, inner experiences, and overcoming difficulties.

"Ellen was shopping with her parents at the grocery store... letting go of her mother's hand she wandered off toward the apples..."- a story about a child who got lost in a store.

"One day Kerry learned the hard way that spreading rumors about other people could have serious consequences..."- a moral story about the consequences of rumors at school.

"Her family couldn't afford the tuition fees... but Nina was determined to find a way to make her dream a reality..."- a story about overcoming financial difficulties for the sake of studying.

Based on the analysis of cluster No. 18, the following types of similarity can be distinguished:

Type of similarity	Description
Subject Matter	All texts are related to personal stories, experiences
Vocabulary	There are words from everyday life: *store, parents, dream, reality*

Type of similarity	Description
Structure	Consistent narrative style, often in the past tense

Challenges faced:

1. SBERT provides rich semantic features, but clustering in high-dimensional space may require dimensionality reduction for visualization.
2. DBSCAN performance heavily depends on `eps` and `min_samples`, which were tuned manually. A poor choice may lead to merging of dissimilar groups.
3. There were no true cluster labels for external validation. Only internal metrics like Silhouette Score and qualitative inspection were available.

Conclusion

This project explored unsupervised clustering of documents with partial semantic similarity using clustering algorithms.

SBERT was effective in converting documents into semantically meaningful vector embeddings, enabling clustering based on meaning rather than surface-level features. DBSCAN algorithm outperformed KMeans in flexibility and clustering quality, achieving a higher silhouette score (0.136) and discovering 40 naturally-occurring clusters without requiring prior knowledge of the number of groups. t-SNE confirmed that semantically related documents were successfully grouped together. Qualitative analysis showed that documents within the same cluster often shared common themes**,** moral structures, or experiential narratives - even when lexical choices and syntax differed.

References

1. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084.https://arxiv.org/abs/1908.10084
2. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), 226–231.

3. PlagBench Dataset (2024). Benchmark for Paraphrase-based Plagiarism Detection. https://github.com/Brit7777/plagbench
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
5. Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.
6. Hugging Face – sentence-transformers library. https://www.sbert.net