Clustering Textual Data

Salveen Singh Dutt
Karina Tiurina
Patryk Prusak

What is the Goal?

- Similar comments will be in one cluster
- With all clusters it’s easy to drive business decisions: What % of comments say A, compared to B?

	review
0	One of the other reviewers has mentioned that ...
1	A wonderful little production. The...
2	I thought this was a wonderful way to spend ti...
3	Basically there's a family where a little boy ...

Literature:

- Paper 1 - Text Clustering with LLM Embeddings – Nov 2024
- Paper 2 - Revolutionary text clustering – July 2024

Text Clustering with LLM Embeddings

Text Clustering with Large Language Model Embeddings

Alina Petukhova[a,*], João P. Matos-Carvalho[a,b], Nuno Fachada[a,b]

[a] *COPELABS, Lusófona University, Campo Grande, 376, Lisbon, 1700-921, Portugal*
[b] *Center of Technology and Systems (UNINOVA-CTS) and Associated Lab of Intelligent Systems (LASI), Caparica, 2829-516, Portugal*

Abstract

Text clustering is an important method for organising the increasing volume of digital content, aiding in the structuring and discovery of hidden patterns in uncategorised data. The effectiveness of text clustering largely depends

Methodology

- **Embeddings Used:**
  - TF-IDF, BERT, OpenAI (GPT), Falcon, LLaMA-2.
- **Clustering Algorithms:**
  - K-Means, Agglomerative Hierarchical Clustering, Spectral, FuzzyCM.
- **Datasets:** CSTR, Reuters, SyskillWebert, MN-DS, 20 Newsgroups.
- **Metrics:** F1, ARI, HS, SS, CHI.

Results

- **OpenAI embeddings + K-Means** achieved the best overall results.
- **BERT embeddings** excelled among open-source models.
- **Falcon embeddings** surpassed LLaMA-2 due to mixed training data.

- **K-Means:** Robust and consistent across datasets.
- **FuzzyCM:** Best for overlapping categories.
- **AHC:** Effective for nested structures.

Revolutionary text clustering

Revolutionary text clustering: Investigating transfer learning capacity of SBERT models through pooling techniques

Yasin Ortakci

Department of Computer Engineering, Karabuk University, Balıklarkayası Mevkii, Merkez, 78050, Karabuk, Turkiye

A R T I C L E I N F O

Keywords:
SBERT
Large language models
Sentence embeddings
Text clustering
Pooling techniques

A B S T R A C T

Large Language Models (LLMs), one of the most advanced representatives of neural networks, have revolutionized the field of natural language processing. Among the many applications of these models, text clustering is gaining increasing interest. In particular, the fact that LLMs digitize text more semantically and contextually than existing methods in the literature has led LLMs to produce more successful results with clustering algorithms. However, since these models are not specifically designed for text clustering, they can lead to processing times that exceed acceptable runtime thresholds. To address this challenge, the Sentence BERT

Methodology

- **Models:** DistilBERT, DistilRoBERTa, ALBERT, MPNET.
- **Pooling Techniques:** CLS, Mean, Max.
- **Algorithm:** K-Means for clustering sentence embeddings.
- **Datasets:** Yahoo Answers, DBpedia, AG News, UCI News Aggregator.
- **Evaluation Metrics:** ARI, Completeness, HS, NMI

Pooling techniques

Pooling Technique	Description	Strengths	Weaknesses
CLS Pooling	Use `[CLS]` token embedding.	Fast and efficient.	May not capture entire sentence context.
Mean Pooling	Average embeddings of all tokens.	Captures overall sentence meaning.	Sensitive to noisy tokens.
Max Pooling	Take maximum value across dimensions.	Highlights dominant features.	May lose nuanced information.

Results

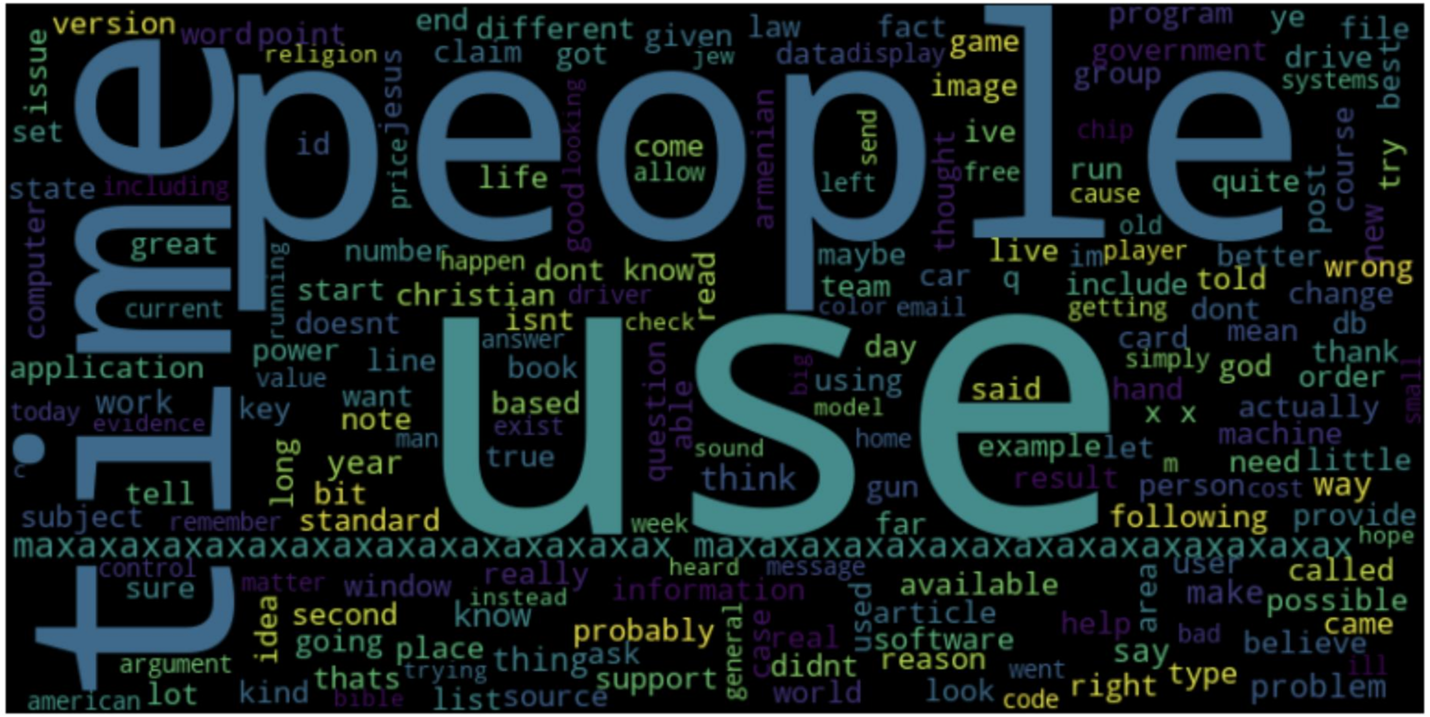
- **Pooling Techniques:**
  - Mean pooling outperformed CLS and Max.
  - Consistent performance across datasets.
- **Model Comparison:**
  - DistilRoBERTa had slightly better average performance.
  - All models performed competitively in clustering tasks.
- SBERT models showed superior or comparable results to advanced clustering methods.

POC – Clustering Textual Data

- **Embeddings Used:**
  - SBERT (DistilRoBERTa), Word2Vec.
- **Dimensionality Reduction Techniques:**
  - UMAP, PCA, None
- **Clustering Algorithms:**
  - K-Means, Agglomerative Hierarchical Clustering, Spectral, HDBScan, DBScan, Gaussian Mixture
- **Datasets:** 20 Newsgroups, Amazon Reviews
- **Metrics:** Silhouette Score

POC Dataset EDA

- 20 News Group
  - 18846 entries: 11K train, 7.5K test
  - 20 clusters
  - Each cluster holds about 4-5% of total data

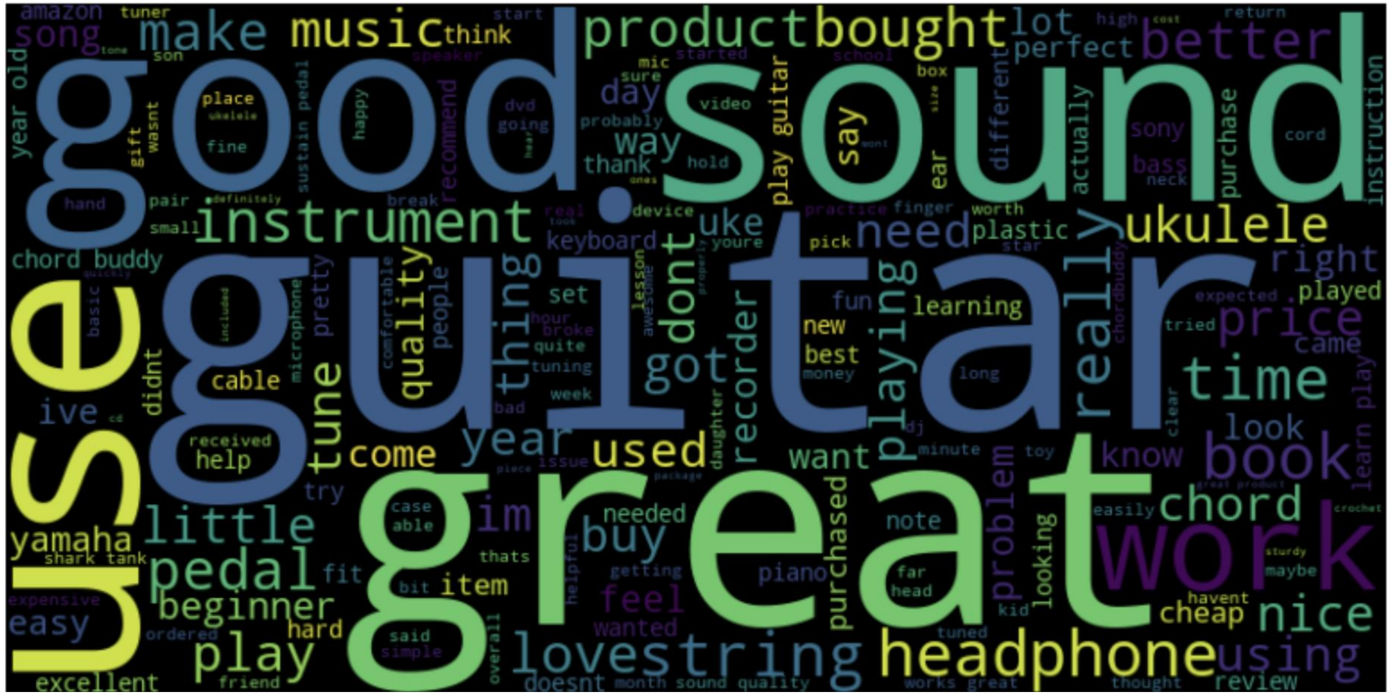


POC Dataset EDA

- Amzon Reviews (Musical Instruments)
  - 1.5M reviews
  - Unlabeled Data

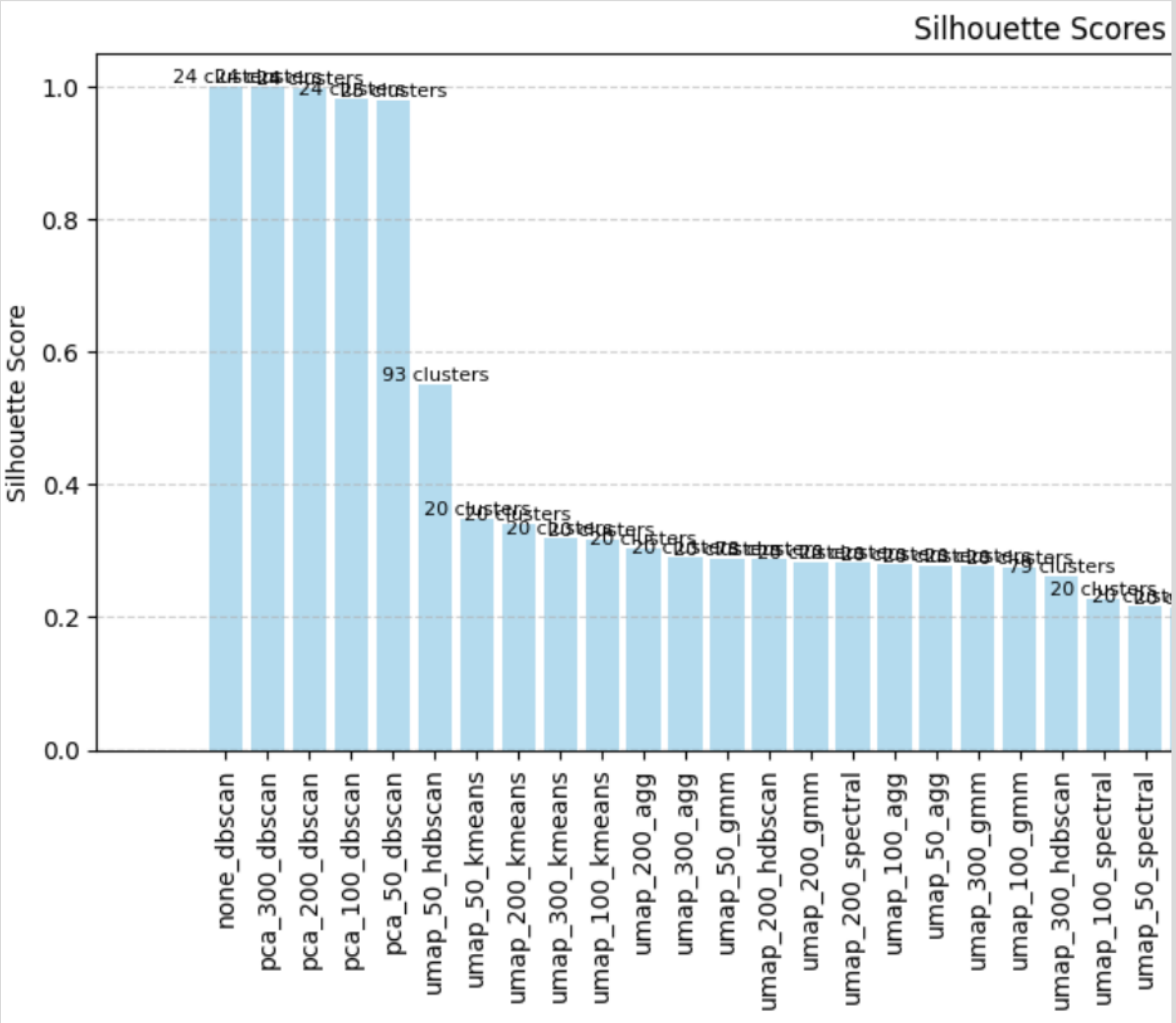
```
Sentiment:

count    100000.000000
mean          0.556390
std           0.411669
min          -0.996100
25%           0.421500
50%           0.659700
75%           0.867300
max           0.999900
```



POC Results

- Amazon Reviews (sample of 15000 reviews)
- Embedding – SBERT
- UMAP, PCA, None
- K-Means, Agglomerative Hierarchical Clustering, Spectral, HDBScan, DBScan, Gaussian Mixture



POC Results

- 20 News Group sample of 5000)
- Embedding – SBERT
- UMAP
- HDBScan

- Silhouette Score – 0.73

POC Results

- 20 News Groups
- Embedding – word2vec
- UMAP, None
- KMeans, HDBScan, Agg, Spectral, GMM, DBScan

	Number of clusters	Silhouette Score
kmeans (UMAP)	5.0	0.554122
kmeans (Word2Vec)	5.0	0.470782
hdbscan (UMAP)	2.0	0.674647
hdbscan (Word2Vec)	2.0	0.697429
agg (UMAP)	5.0	0.491545
agg (Word2Vec)	5.0	0.466261
spectral (UMAP)	5.0	0.342129
spectral (Word2Vec)	5.0	0.449555
gmm (UMAP)	5.0	0.470693
gmm (Word2Vec)	5.0	0.069928
dbscan (UMAP)	15.0	0.147319
dbscan (Word2Vec)	1.0	-1.000000

POC Results

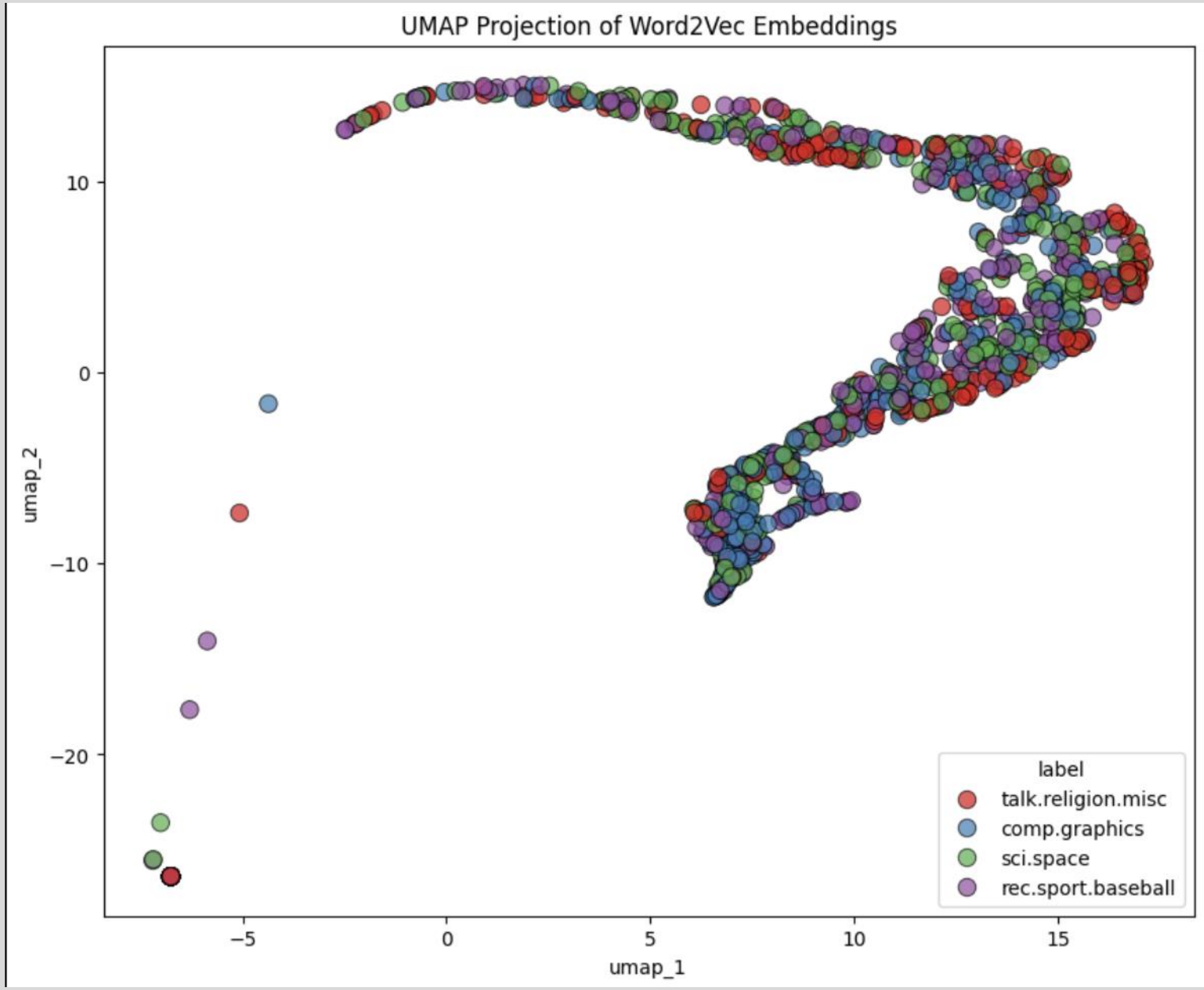
- Additional
- 20 News Group
  sample of 5000)
- Embedding – SBERT
- Dim Reduction - None
- SVM

F1 Score – 60%

The score decreases dramatically if dim reduction is used.

POC Results

- 20 News Group (4 groups only)
  Embedding – word2vec
- UMAP, None
- HDBScan



Future Work

- Use other embedding techniques such as BERT, comparing against SBERT and LLM embeddings (from open source models such as Gemma)
- More metrics for understanding how well clustering went
- Partial labelling of Amazon Reviews for improvement of accuracy in supervised learning.
- Labeling subset of Amazon Reviews using LLM

References

- **Petukhova, A., Matos-Carvalho, J. P., & Fachada, N. (2024)**
  - *Text clustering with large language model embeddings*.
  - Published in the *International Journal of Cognitive Computing in Engineering*.
  - Advance online publication: https://doi.org/10.1016/j.ijcce.2024.11.004.
- **Yasin Ortakci (2024)**
  - *Revolutionary text clustering: Investigating transfer learning capacity of SBERT models through pooling techniques.*
  - Published in the *Journal of Advanced Text Clustering*.
- **Tom B. Brown, Benjamin Mann, and Nick Ryder et al. (2020)**
  - *Language models are few-shot learners*.
  - Published in *Advances in Neural Information Processing Systems (NeurIPS), Vol. 33*.
- **Tadeusz Caliński and Jerzy Harabasz (1974)**
  - *A dendrite method for cluster analysis.*
  - Published in *Communications in Statistics - Theory and Methods, Vol. 3(1):1–27*.
- **Thomas N. Kipf and Max Welling (2017)**
  - *Semi-supervised classification with graph convolutional networks.*
  - Presented at the International Conference on Learning Representations (ICLR).

References

- **Adrien Bougouin, Florian Boudin, and Béatrice Daille (2021)**
  - *Keyphrase extraction for clustering purposes: Combining textual and semantic features.*
  - Published in the *Journal of Information Retrieval*.
- **Andrew Rosenberg and Julia Hirschberg (2007)**
  - *V-measure: A conditional entropy-based external cluster evaluation measure.*
  - Published in the Proceedings of the joint EMNLP-CoNLL conference.
- **Peter J. Rousseeuw (1987)**
  - *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*.
  - Published in the *Journal of Computational and Applied Mathematics*.
- **Douglas Steinley (2004)**
  - *Properties of the Hubert-Arable adjusted Rand index*.
  - Published in *Psychological Methods*.
- **Jianlong Xie, Ross Girshick, and Ali Farhadi (2016)**
  - *Unsupervised deep embedding for clustering analysis*.
  - Presented at the International Conference on Machine Learning (ICML).