Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Intrinsic Thematic Separability in Sentence-Transformer Embeddings: A Controlled Geometric Study from Synthetic to Real-World Text

Miguel Pavón

mpavon@enaire.es

Aena

Research Article

Keywords: sentence transformers, text clustering, thematic separability, UMAP, HDBSCAN, embedding geometry

Posted Date: March 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8976073/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Intrinsic Thematic Separability in Sentence-Transformer Embeddings: A Controlled Geometric Study from Synthetic to Real-World Text

Miguel Pavón[1*]

[1*]ENAIRE, Madrid, Spain.

Corresponding author(s). E-mail(s): mpavon@enaire.es, mikeentropic@gmail.com;

Abstract

We investigate whether sentence-transformer embeddings encode sufficient topical information to recover imposed thematic structure in corpora of varying origin and difficulty. A synthetic corpus of 8,000 ChatGPT-generated texts across four domains achieves perfect macro-group recovery in 768D ($\boldsymbol{K}$-Means ARI = 1.000), demonstrating that separability is intrinsic to the embeddings and independent of UMAP projection. UMAP + HDBSCAN provides a complementary micro-topic decomposition into 66 thematically pure sub-clusters. Perfect recovery generalises across embedding architectures (BGE-large 1024D) and across LLM sources (ChatGPT + Claude: topic ARI = 0.994, source ARI $\approx \mathbf{0}$). Benchmarking against AG News, 20 Newsgroups, Reuters-21578 and a human-authored Reddit corpus establishes an empirical performance gradient (ARI from 1.000 to 0.40) that tracks embedding geometry, though bootstrap confidence intervals remain wide at $\boldsymbol{n = 6}$ conditions. A TF-IDF + SVD sparse baseline matches the synthetic ceiling but underperforms embeddings by $\mathbf{+0.13}$ to $\mathbf{+0.43}$ ARI on real-world data, confirming that the embedding advantage is specific to naturalistic text. Sub-cluster coherence on AG News is validated via NPMI: all 394 sub-clusters are significantly more coherent than chance (Cohen's $\boldsymbol{d = 6.03}$). A stylistic-vs.-semantic decomposition confirms that nine surface-level features do not explain topical separability; unsupervised clustering does not recover source identity. These results characterise the conditions under which the embedding + clustering pipeline succeeds, positioning synthetic results as an upper bound and providing practitioners with geometric diagnostics for post-hoc interpretation of clustering performance.

Keywords: sentence transformers, text clustering, thematic separability, UMAP, HDBSCAN, embedding geometry

1 Introduction

The rapid adoption of large language models (LLMs) for content generation raises a methodological question for computational text analysis: *can unsupervised clustering pipelines distinguish thematic categories when all texts are produced by the same generative model?* If so, sentence-level semantic representations encode topic structure independently of authorial fingerprint, and standard clustering workflows remain valid on synthetic corpora.

Modern sentence-transformer models (Reimers & Gurevych, 2019) produce normalised embeddings that can be projected via UMAP (McInnes, Healy, & Melville, 2018) and clustered with density-based algorithms such as HDBSCAN (Campello, Moulavi, & Sander, 2013). This combination has been applied successfully in topic modelling (Grootendorst, 2022), social media analysis (Angelov, 2020) and document clustering (Sia, Dalmia, & Mielke, 2020). However, most validation studies assume heterogeneous authorship; the specific question of whether the pipeline separates *same-model* texts by topic alone has received little attention.

This work addresses that gap through a controlled geometric study. A synthetic corpus of 8,000 ChatGPT-generated texts across four thematic domains is used as a controlled baseline, and the pipeline is then applied to progressively harder conditions: a human-authored Reddit corpus, and three standard NLP benchmarks (AG News, 20 Newsgroups, Reuters-21578).

The study differs from prior work in four respects: (i) controlled same-model generation eliminating authorial variance; (ii) explicit disentanglement of intrinsic 768D separability from UMAP projection effects; (iii) hierarchical validation linking macro-topic recoverability with micro-topic coherence; and (iv) systematic benchmarking establishing an empirical gradient and geometric diagnostics for practitioners.

A key concern is that perfect clustering on four semantically distant LLM-generated domains may reflect an easy task rather than a meaningful embedding property. We address this by showing how performance degrades systematically as real-world heterogeneity increases, and by identifying the geometric features of the embedding space that explain this degradation.

2 Related Work

2.1 Sentence embeddings and text clustering

SBERT (Reimers & Gurevych, 2019) introduced Siamese networks for sentence embeddings. The MTEB benchmark (Muennighoff, Tazi, Magne, & Reimers, 2023) evaluates over 100 models, establishing `all-mpnet-base-v2` among the top-performing 768D models. BERTopic (Grootendorst, 2022) operationalises SBERT + UMAP + HDBSCAN as a topic-modelling pipeline but does not study whether the embedding space *itself* separates topics independently of dimensionality reduction. Top2Vec (Angelov, 2020) combines document embeddings with UMAP and HDBSCAN but focuses on topic-vector discovery. Classical comparisons (Steinbach, Karypis, & Kumar, 2000) showed that hierarchical methods outperform partitional

ones on heterogeneous text; our results update this for transformer embeddings, where K-Means and Ward achieve identical performance in 768D.

2.2 Clustering evaluation

Amigó, Gonzalo, Artiles, and Verdejo (2009) compared clustering metrics, showing that ARI, NMI and V-measure capture different aspects of partition agreement. ARI penalises over-segmentation while Cramér's V does not—an asymmetry we exploit to characterise the HDBSCAN vs. K-Means trade-off. Rosenberg and Hirschberg (2007) introduced the V-measure decomposition into homogeneity and completeness, which explains the $V = 1.0$ / ARI < 1 pattern in density-based methods.

2.3 Topic coherence

NPMI (Bouma, 2009) is the standard automated proxy for topic coherence. Röder, Both, and Hinneburg (2015) demonstrated that NPMI variants achieve the highest correlation with human ratings; Lau, Newman, and Baldwin (2014) confirmed this for automated topic model evaluation. We extend NPMI validation from topic models to density-based clusters in embedding space.

2.4 LLM-generated text detection

DetectGPT (Mitchell, Lee, Khazatsky, Manning, & Finn, 2023) and GLTR (Gehrmann, Strobelt, & Rush, 2019) demonstrate that LLM text can be distinguished from human text using generation-probability features. Our finding that sentence-transformer embeddings do *not* encode generation source is complementary: semantic embeddings capture meaning, not statistical generation signatures.

2.5 Synthetic text evaluation and classical topic models

Veselovsky, Ribeiro, and West (2023) document contamination effects in synthetic data; Li, Li, Sun, and Collier (2024) evaluate synthetic corpora for NLP tasks. Our work contributes by quantifying the homogeneity advantage of synthetic text while demonstrating its systematic degradation under real-world conditions. LDA (Blei, Ng, & Jordan, 2003) remains the dominant topic-modelling baseline; recent comparisons (Sia et al., 2020) show that embedding clusters can match or exceed LDA coherence.

3 Materials and Methods

3.1 Corpora

3.1.1 Synthetic corpus (controlled baseline)

8,000 texts generated by ChatGPT across four thematic domains—Motorsport, Politics, Recipes, Sports—mimicking forum-post style (2,000 per category). These four domains were chosen for their clear semantic distinctiveness; this deliberate distance is a feature of the controlled design, not a limitation to hide: it establishes an upper bound under optimal conditions.

3.1.2 Human control corpus

8,000 Reddit comments from four topically matched subreddits (`r/formula1`, `r/PoliticalDiscussion`, `r/Cooking`, `r/soccer`; 2,000 per category), filtered to 50–400 words with deduplication.

3.1.3 AG News (Zhang et al., 2015)

50,000 news articles across 4 categories (Business, Sci/Tech, Sports, World) (Zhang, Zhao, & LeCun, 2015). Moderate difficulty: journalistic prose with genuine semantic overlap between Business and Sci/Tech.

3.1.4 20 Newsgroups (Lang, 1995)

18,331 Usenet posts across 20 fine-grained categories with natural hierarchical structure (Lang, 1995). Also evaluated at 6 macro-groups (prefix-based super-categories).

3.1.5 Reuters-21578 (Apté et al., 1994)

Financial newswire articles filtered to single-label documents in the top-8 categories (7,221 documents) (Apté, Damerau, & Weiss, 1994). Heavily imbalanced (earn: 3,626; ship: 136) with shared financial-domain vocabulary.

3.2 Embedding and clustering pipeline

All texts were encoded with `all-mpnet-base-v2` (Reimers & Gurevych, 2019) (768D, `normalize_embeddings=True`). A second model (`BAAI/bge-large-en-v1.5`, 1024D) (Xiao, Liu, Zhang, & Muennighoff, 2023) was tested for architectural generalisation.

3.2.1 Macro-topic recovery

K-Means (Lloyd, 1982) ($n_{\text{init}} = 10$, seed 42) and Ward agglomerative clustering (Ward, 1963) were applied directly in the native embedding space at k equal to the number of ground-truth categories. To verify stability, each K-Means run was repeated with 10 random seeds (0–9).

3.2.2 Micro-topic discovery

UMAP ($n_{\text{components}} = 3$, $n_{\text{neighbors}} = 15$, min_dist $= 0.1$, cosine metric, seed 42) followed by HDBSCAN (min_cluster_size $= 40$, min_samples $= 5$, `leaf` selection). A grid search over $n_{\text{neighbors}} \in \{5, 10, 15, 30, 50, 100\}$ and $d_{\text{min}} \in \{0.0, 0.05, 0.1, 0.25, 0.5\}$ assessed sensitivity.

3.2.3 Evaluation metrics

ARI (Hubert & Arabie, 1985), NMI, homogeneity, completeness, V-measure (Rosenberg & Hirschberg, 2007), Cramér's V, cosine silhouette with true labels.

3.3 Cross-LLM experiment

580 texts were generated by Claude (Anthropic) matching the same four domains (135–171 per domain), combined with the ChatGPT corpus (8,580 texts total). K-Means $k = 4$ was evaluated against both topic and LLM-source labels.

3.4 Sub-cluster coherence (NPMI)

Applied to the 394 HDBSCAN sub-clusters on AG News. For each cluster, the top-10 TF-IDF keywords were extracted and NPMI computed for all keyword pairs using full-corpus co-occurrence. A random baseline of 199 permutations (same-category random assignment) provided a null distribution. Significance was assessed via Benjamini–Hochberg correction ($\alpha = 0.05$); effect size via Cohen's d. A holdout stability test (50/50 stratified split, nearest-centroid classification) assessed sub-cluster reproducibility.

3.5 Stylistic vs. semantic decomposition

Nine surface-level features (mean sentence length, type–token ratio, lexical entropy, punctuation density, mean word length, capital ratio, digit ratio, function-word ratio, sentence count) were extracted for each AG News text and normalised to $[0, 1]$. K-Means was applied to three feature spaces: semantic (768D), stylistic (9D) and combined (777D). ARI was computed against both topic labels and a temporal-split proxy for source.

3.6 Sparse baselines

LDA (Blei et al., 2003) was applied to AG News ($k \in \{4, 8, 20\}$; 10,000-feature vocabulary; online learning). As a direct sparse counterpart to the embedding pipeline, TF-IDF (10,000 features, min_df = 5, max_df = 0.5, sublinear TF) reduced via truncated SVD to 300 dimensions was also evaluated with K-Means (10 seeds) on all seven corpora.

3.7 Geometric diagnostics

For each benchmark condition, the following metrics were computed in 768D.

Separation ratio. For each pair of ground-truth groups (g_i, g_j) with centroids $\mathbf{c}_i, \mathbf{c}_j$, the pairwise separation ratio is defined as:

$$S_{ij} = \frac{d_{\cos}(\mathbf{c}_i,\, \mathbf{c}_j)}{\max\!\left(\bar{d}_{\cos}^{(i)},\; \bar{d}_{\cos}^{(j)}\right)} \tag{1}$$

where $d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \mathbf{a} \cdot \mathbf{b}$ is the cosine distance and $\bar{d}_{\cos}^{(i)}$ is the mean cosine distance from each point in group g_i to its centroid. The *mean separation ratio* is the average of S_{ij} over all $\binom{k}{2}$ group pairs.

Cosine silhouette with true labels (sample of 5,000 points) and *cosine Davies–Bouldin index* were also computed.

These metrics require ground-truth labels and are therefore *diagnostic* (post-hoc explanatory) rather than predictive in the strict sense: they explain *why* clustering performance varies across conditions, not how to predict it on unlabelled data. Pearson and Spearman correlations were computed between each diagnostic and observed ARI, with 95% bootstrap confidence intervals ($B = 10{,}000$ resamples) to quantify the structural uncertainty inherent in $n = 6$. Leave-one-out cross-validation (LOO-CV) assessed the stability of a linear regression between mean separation ratio and ARI.

4 Results

4.1 Perfect recovery in 768D

Table 1 shows clustering results in the native embedding space.

Table 1 Clustering in the native embedding space (synthetic corpus). H = homogeneity, C = completeness. Cosine silhouette throughout.

Method	k	**Noise**	**Sil.**	H	C	**ARI**	**NMI**
K-Means 768D	4	0%	0.395	1.000	1.000	**1.000**	**1.000**
Ward 768D	4	0%	0.395	1.000	1.000	**1.000**	**1.000**
HDBSCAN EOM 768D	18	0.1%	0.431	1.000	0.673	0.832	0.804
HDBSCAN leaf 768D	49	29.9%	—	1.000	0.522	0.127	0.522
UMAP 3D baselines (Euclidean silhouette):							
K-Means UMAP 3D	4	0%	0.398	—	—	0.802	0.790
Ward UMAP 3D	4	0%	0.355	—	—	0.823	0.838
HDBSCAN leaf UMAP	66	11.3%	0.702	—	—	0.108	0.503

Both K-Means and Ward achieve **perfect recovery** (ARI = 1.000) in 768D, demonstrating that separability is intrinsic to the embeddings. UMAP is not required—it merely provides a low-dimensional visualisation. HDBSCAN EOM in 768D produces 18 clusters with perfect homogeneity ($H = 1.000$) but incomplete grouping ($C = 0.673$): every cluster contains a single thematic group, but each group is split across multiple clusters. ARI penalises this over-segmentation; Cramér's V (which measures association, not partition agreement) remains 1.000.

The advantage of Ward over K-Means (ARI 0.823 vs 0.802) is specific to UMAP 3D, where Ward's variance-minimisation criterion handles non-spherical boundaries better; in 768D, both methods are equivalent.

4.2 UMAP hyperparameter sensitivity

Across 30 UMAP configurations ($n_{\text{neighbors}} \times d_{\text{min}}$), HDBSCAN leaf Cramér's V remained stable (0.904–0.991; mean 0.949 ± 0.021). Ward $k = 4$ ARI ranged from 0.480 ($n_{\text{neighbors}} = 5$, $d_{\text{min}} = 0.25$) to 1.000 ($n_{\text{neighbors}} = 100$, $d_{\text{min}} = 0.1$). **Guideline:** for macro-group recovery, $n_{\text{neighbors}} \geq 50$ preserves sufficient global structure in 3D UMAP. HDBSCAN sub-cluster discovery is robust at any setting.

4.3 Micro-topic structure

HDBSCAN leaf on UMAP 3D discovered 66 sub-clusters (88.7% of points assigned; 11.3% noise). Of these, 64 (97%) are composed entirely of a single thematic group; 2 clusters mix Recipes and Politics (plausibly food-policy language). TF-IDF keywords confirm sub-topical coherence: Recipes sub-clusters map to individual dishes (omelette, beef stew, salmon); Politics sub-clusters separate by policy domain; Sports sub-clusters capture tactical dimensions.

4.4 Generalisation across architectures and LLMs

`BAAI/bge-large-en-v1.5` (1024D) also achieves ARI = 1.000 with higher cosine silhouette (0.491 vs. 0.395), confirming that perfect separability is not model-specific.

In the cross-LLM experiment, the 580 Claude texts are unevenly distributed across domains (Motorsport: 171, Politics: 137, Recipes: 137, Sports: 135) due to generation-time filtering. K-Means $k = 4$ on the combined ChatGPT + Claude corpus achieves ARI = 0.994 against topic labels and ARI = −0.000 against source labels. K-Means $k = 2$ cannot separate the two LLMs (ARI = −0.009). Even $k = 8$ captures no source signal (ARI = −0.001 against source, 0.739 against topic). **Thematic separability dominates over generation source.**

4.5 Empirical performance gradient

Table 2 presents the performance gradient and geometric diagnostics across six topic-clustering conditions.

Table 2 K-Means clustering performance and geometric diagnostics across benchmark conditions (768D). ARI mean ± std over 10 random seeds. Separation ratio, silhouette and Davies–Bouldin (DB) computed with true labels.

Dataset	k	n	ARI (mean ± std)	Sep.	Sep. min	Sil.	DB
Synthetic (ChatGPT)	4	8,000	1.000 ± 0.000	1.519	0.966	0.396	1.524
Human (Reddit)	4	8,000	0.820 ± 0.000	0.985	0.547	0.147	2.739
AG News	4	50,000	0.568 ± 0.000	0.668	0.338	0.058	4.404
Reuters (top-8)	8	7,221	0.469 ± 0.011	0.821	0.393	0.187	3.907
20 Newsgroups (20-cat)	20	18,331	0.404 ± 0.013	0.889	0.092	0.040	7.530
20 Newsgroups (6-macro)	6	18,331	0.403 ± 0.002	0.788	0.422	0.025	3.769

The gradient ARI = 1.000 → 0.820 → 0.568 → 0.469 → 0.404 → 0.403 reflects increasing difficulty from controlled synthetic text to real-world data with genuine semantic overlap. Variance across K-Means initialisations is negligible (std < 0.001) for all datasets except Reuters (std = 0.011) and 20 Newsgroups 20-category (std = 0.013), confirming that the reported values are stable.

Collapsing 20 Newsgroups from 20 to 6 categories does not improve ARI (0.403 vs. 0.404): the nominal prefix-based hierarchy does not match the semantic hierarchy in embedding space.

4.5.1 Negative control: IMDB sentiment

As a boundary test, we applied the same pipeline to IMDB (Maas et al., 2011) (25,000 movie reviews, binary sentiment). K-Means $k = 2$ yields ARI = 0.059—near chance. This confirms that sentence-transformer embeddings do not encode evaluative polarity: both sentiment classes share the same topical domain (cinema) and are geometrically inseparable (separation ratio = 0.085, silhouette = 0.034). IMDB is not included in the topic-clustering gradient because sentiment classification is a fundamentally different task from topical separation.

4.6 Geometric diagnostics

Figure 1 and Table 2 relate embedding geometry to observed ARI across the six topic-clustering conditions.

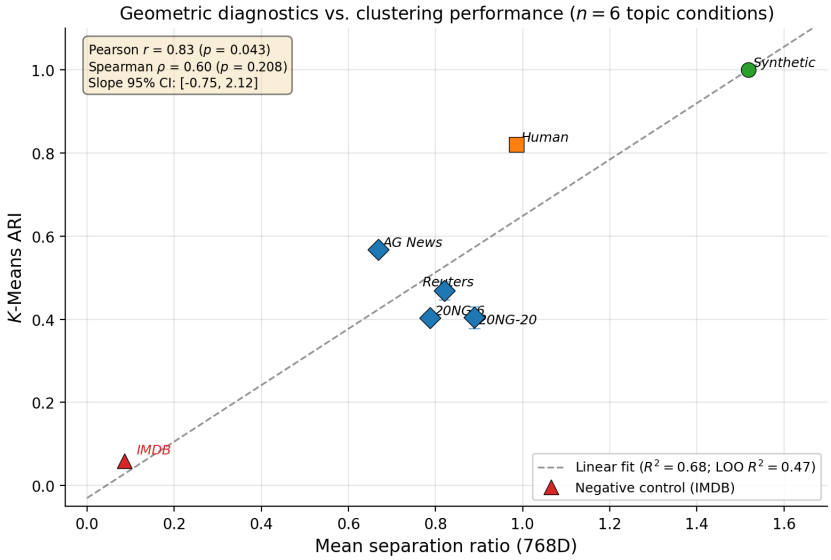


Fig. 1 Mean separation ratio vs. K-Means ARI across benchmark conditions ($n = 6$ topic-clustering conditions). Dashed line: linear regression ($R^2 = 0.68$; slope 95% CI includes zero). IMDB (triangle) is shown as a negative control but excluded from the regression. See Table 3 for full correlation statistics and bootstrap CIs.

Table 3 reports all four correlation combinations with bootstrap 95% CIs.

Several patterns emerge. The cosine silhouette is the most robust diagnostic: both Pearson and Spearman are significant ($p < 0.05$), and its bootstrap CI for Pearson r is the only one that excludes zero. The mean separation ratio achieves a significant Pearson correlation ($r = 0.83$, $p = 0.043$) but its Spearman rank correlation is not significant ($\rho = 0.60$, $p = 0.208$). The rank-order violation is driven by a specific inversion: Reuters (8 categories, sep. ratio = 0.821) has a higher mean separation ratio than AG News (4 categories, ratio = 0.668) yet a lower ARI (0.469 vs. 0.568). This occurs

Table 3 Correlations between geometric diagnostics and K-Means ARI ($n = 6$ topic-clustering conditions). Bootstrap 95% CIs from $B = 10{,}000$ resamples.

Diagnostic	Statistic	Value	p	Bootstrap 95% CI
Sep. ratio	Pearson r	0.83	0.043	[−0.89, 1.00]
Sep. ratio	Spearman ρ	0.60	0.208	[−0.64, 1.00]
Silhouette	Pearson r	0.83	0.042	[0.05, 1.00]
Silhouette	Spearman ρ	0.83	0.042	[0.00, 1.00]

because the mean separation ratio averages over $\binom{k}{2}$ pairs, masking the high heterogeneity within Reuters's 28 pairs (range 0.39–1.14); additionally, clustering difficulty scales with k, a confound that the scalar separation ratio does not capture.

Linear regression of ARI on mean separation ratio yields $R^2 = 0.68$ (slope = 0.68, 95% CI $[-0.75, 2.12]$; intercept = −0.03, CI $[-1.27, 1.06]$); LOO-CV MAE = 0.14, $R^2 = 0.47$. The wide bootstrap CIs for the slope—including zero—confirm that the regression is structurally uncertain at $n = 6$ and should not be interpreted as a calibrated predictive model.

These diagnostics explain *why* the gradient exists: high intra-group heterogeneity and low inter-group separation reduce the geometric diagnostics, which in turn track the observed ARI. Crucially, this relationship requires ground-truth labels to compute—it is a post-hoc explanatory tool, not a method for predicting ARI on unlabelled data.

4.7 Inter-group distances

Table 4 Inter-group cosine distances (768D). Intra-group on diagonal; separation ratio in parentheses. Left: synthetic corpus. Right: human (Reddit) corpus.

	Synthetic (ChatGPT)			
	Motor.	Polit.	Recip.	Sports
Motor.	0.612	0.928 (1.51)	1.006 (1.64)	0.809 (1.32)
Polit.	—	0.615	0.995 (1.62)	0.900 (1.46)
Recip.	—	—	0.417	0.998 (2.26)
Sports	—	—	—	0.442

	Human (Reddit)			
	Motor.	Polit.	Recip.	Sports
Motor.	0.711	0.702 (0.89)	0.925 (1.20)	0.417 (0.55)
Polit.	—	0.793	0.982 (1.24)	0.708 (0.89)
Recip.	—	—	0.773	0.926 (1.20)
Sports	—	—	—	0.758

In the synthetic corpus, all separation ratios exceed 1.0 (range 1.32–2.26). In the human corpus, only three of six pairs exceed 1.0, and the Motorsport–Sports pair collapses to 0.55—explaining why macro-recovery drops from ARI = 1.000 to 0.820. Intra-group distances are uniformly larger in the human corpus (0.71–0.79 vs. 0.42–0.61), confirming that LLM-generated text is more lexically homogeneous within each domain.

4.8 Sub-cluster coherence (NPMI)

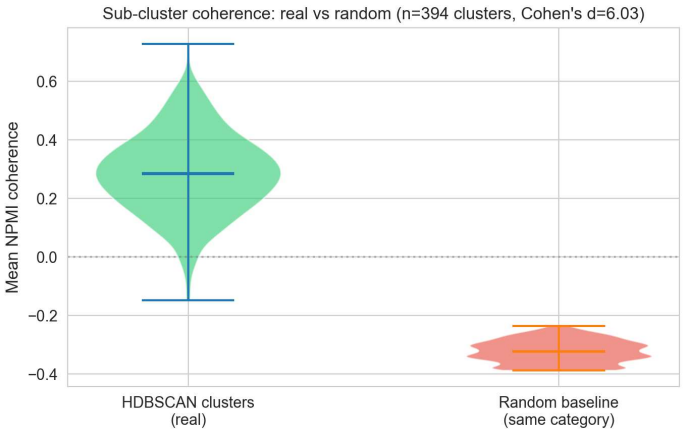


Fig. 2 NPMI coherence of HDBSCAN sub-clusters vs. random baseline on AG News (394 clusters). Real clusters (green) are massively more coherent than same-category random partitions (red). Cohen's $d = 6.03$.

All 394 AG News sub-clusters are significantly more coherent than random partitions (Benjamini–Hochberg corrected $p < 0.05$; 100% significant). Mean NPMI: real $+0.286$ vs. random -0.323 (Cohen's $d = 6.03$). Holdout nearest-centroid accuracy: 91.0% (random baseline: $1/394 = 0.25\%$).

This validates that sub-cluster structure on real-world data—not just synthetic text—reflects genuine lexical-thematic cohesion. An important caveat: NPMI with corpus-internal co-occurrences and TF-IDF keywords validates *internal semantic coherence*, not alignment with an external gold-standard taxonomy of sub-topics. HDBSCAN's tendency to produce small, highly specific clusters can inflate NPMI; the random baseline controls for this by matching cluster size within each thematic category. Nevertheless, the result should be read as "sub-clusters are semantically coherent beyond what same-category random partitions produce," not as "sub-clusters correspond to ground-truth sub-topics."

Table 5 LDA vs. K-Means (768D) on AG News.

Method	k	ARI	NMI
LDA (bag-of-words)	4	0.554	0.534
LDA (bag-of-words)	8	0.375	0.414
K-Means (768D embeddings)	4	0.568	0.572

4.9 Sparse baselines

At $k = 4$, LDA and K-Means on 768D embeddings achieve comparable performance on AG News (ARI = 0.554 vs. 0.568, $\Delta = 0.014$). On a stylistically homogeneous corpus, bag-of-words topic models retain substantial discriminative power.

4.9.1 TF-IDF + SVD baseline across all corpora

Table 6 compares K-Means on 768D sentence-transformer embeddings against K-Means on TF-IDF + SVD(300D) across all benchmark conditions.

Table 6 K-Means ARI: sentence-transformer embeddings (768D) vs. TF-IDF + SVD (300D). Mean ± std over 10 seeds.

Dataset	Emb. 768D	TF-IDF+SVD	Δ
Synthetic	1.000 ± 0.000	1.000 ± 0.000	0.000
Human (Reddit)	0.820 ± 0.000	0.655 ± 0.017	+0.165
AG News	0.568 ± 0.000	0.134 ± 0.123	+0.434
Reuters (top-8)	0.469 ± 0.011	0.341 ± 0.066	+0.128
20NG (20-cat)	0.404 ± 0.013	0.103 ± 0.009	+0.301
20NG (6-macro)	0.403 ± 0.002	0.141 ± 0.015	+0.263
IMDB (sent.)	0.059 ± 0.000	0.134 ± 0.001	−0.075

On the synthetic corpus, both representations achieve perfect recovery (ARI = 1.000), confirming that separating four semantically distant LLM-generated domains is trivial for any reasonable text representation. On all real-world benchmarks, sentence-transformer embeddings substantially outperform the sparse baseline (Δ from +0.128 to +0.434). The gap is largest on AG News, where the sparse representation is highly unstable (std = 0.123) while embeddings are deterministic. Notably, LDA ($k = 4$, ARI = 0.554) far outperforms TF-IDF + SVD + K-Means (0.134) on AG News, indicating that the generative topic-model structure of LDA captures category boundaries that the purely geometric SVD projection misses. The single exception is IMDB, where TF-IDF + SVD achieves a marginally higher (but still near-chance) ARI; neither representation meaningfully separates sentiment.

4.10 Stylistic vs. semantic signal

Semantic embeddings capture topic (ARI = 0.568) while remaining blind to source (ARI $\approx$ 0). The nine surface-level stylistic features tested here capture neither topic

Table 7 Feature-space decomposition on AG News. Source label: temporal-split proxy.

Space	k	ARI (topic)	ARI (source)
Semantic 768D	4	0.568	0.000
Stylistic 9D	4	0.016	0.000
Combined 777D	4	0.572	0.000

nor source; however, "style" is a broad construct, and richer features (POS n-grams, syntactic dependency patterns, cohesion metrics) might reveal dimensions not probed by this set. The conclusion supported by these data is narrower: *these specific surface features do not explain topical separability, and unsupervised clustering does not recover source identity in any of the three feature spaces tested.*

This is corroborated by the cross-LLM experiment (ChatGPT + Claude: topic ARI = 0.994, source ARI = −0.000): whether "source" is defined as temporal position or generative model, unsupervised clustering does not recover it.

4.11 Performance gradient visualisation

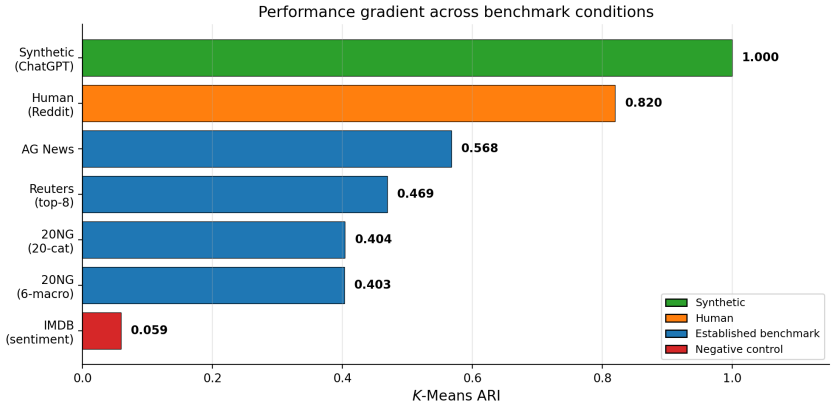


Fig. 3 K-Means ARI across benchmark conditions, ordered by performance. Colour indicates corpus type: green = synthetic, orange = human, blue = established benchmark.

Figure 3 summarises the ARI gradient numerically. Figure 4 provides the corresponding geometric view: UMAP 3D projections for all seven corpora visualise how cluster separability degrades from the synthetic baseline through increasingly difficult real-world conditions. The progression from tight, well-separated clusters (Synthetic ChatGPT) to diffuse, overlapping distributions (20 Newsgroups, Reuters) to complete geometric inseparability (IMDB sentiment) directly mirrors the quantitative ARI gradient in Table 2.

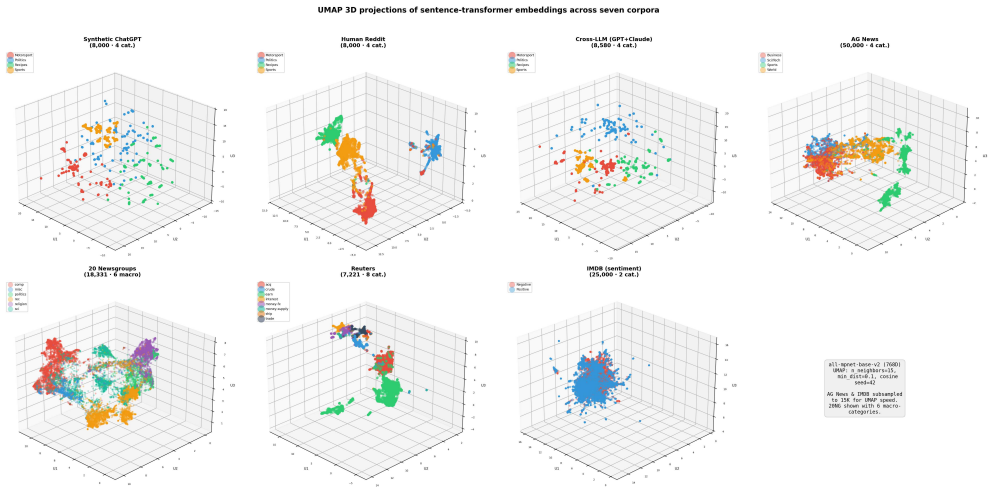


Fig. 4 UMAP 3D projections of sentence-transformer embeddings (`all-mpnet-base-v2`, 768D) across seven corpora, coloured by ground-truth category. Top row: Synthetic ChatGPT (4 cat.), Human Reddit (4 cat.), Cross-LLM ChatGPT+Claude (4 cat.), AG News (4 cat.). Bottom row: 20 Newsgroups (6 macro-cat.), Reuters top-8 (8 cat.), IMDB sentiment (2 cat.). UMAP parameters: $n_{\text{neighbors}} = 15$, min_dist $= 0.1$, cosine metric, seed 42. AG News and IMDB subsampled to 15,000 points. The visual progression mirrors the quantitative ARI gradient (Table 2).

5 Discussion

5.1 What the perfect synthetic result means—and what it does not

Perfect recovery (ARI = 1.000) on four semantically distant LLM-generated domains is not surprising: Motorsport, Politics, Recipes and Sports occupy maximally separated regions of semantic space. The TF-IDF + SVD baseline confirms this directly: it also achieves ARI = 1.000 on the synthetic corpus (Table 6). The synthetic result is therefore not evidence of embedding power—it is a *controlled upper bound* that any reasonable text representation can reach. The interesting question is how performance degrades as conditions become harder, and here sentence-transformer embeddings prove their value: the gap over TF-IDF + SVD ranges from +0.128 (Reuters) to +0.434 (AG News) on real-world benchmarks.

5.2 The degradation gradient as the central finding

The empirical gradient (Table 2, Figure 4) provides this answer. Performance drops from ARI = 1.000 (synthetic) to 0.820 (human Reddit) to 0.568 (AG News) to 0.469 (Reuters) to 0.40 (20 Newsgroups), and the degradation is explained by embedding geometry. The UMAP projections in Figure 4 make this gradient visually explicit: the synthetic corpus shows four compact, non-overlapping clusters; human Reddit preserves four distinguishable groups but with broader intra-cluster dispersion; AG News and Reuters display increasing inter-category overlap; 20 Newsgroups shows

extensive mixing among semantically adjacent macro-groups; and IMDB exhibits no cluster structure whatsoever.

The human corpus (ARI = 0.820) is the most informative comparison point: same four domains, same evaluation pipeline, but human-authored text introduces lexical heterogeneity (intra-group distances 0.71–0.79 vs. 0.42–0.61) that pushes three of six separation ratios below 1.0. This 18-percentage-point ARI drop directly quantifies the "homogeneity advantage" of LLM-generated text.

AG News (ARI = 0.568) and Reuters (ARI = 0.469) introduce genuine inter-category overlap: Business and Sci/Tech share vocabulary; financial-news categories share domain context. 20 Newsgroups pushes further, with adjacent pairs (e.g. `alt.atheism` vs. `talk.religion.misc`, ratio = 0.091) that are effectively inseparable.

5.3 Geometric diagnostics: scope and structural uncertainty

The correlation between geometric diagnostics and ARI (Table 3) may tempt a predictive interpretation. We deliberately avoid this framing for three reasons.

First, separation ratios require ground-truth labels. If labels are available, ARI can be computed directly. The diagnostic value is explanatory: it answers *why* one corpus clusters better than another, not *whether* a new unlabelled corpus will cluster well.

Second, the bootstrap confidence intervals (Table 3) reveal the structural uncertainty of $n = 6$: the slope CI for the regression includes zero ($[-0.75, 2.12]$), and the Pearson r CI for the separation ratio spans $[-0.89, 1.00]$. Even the most robust diagnostic (cosine silhouette, Pearson CI $[0.05, 1.00]$) barely excludes zero. These intervals are not a failure of the analysis—they are the honest consequence of having six data points. They confirm that the observed correlations are suggestive but not statistically robust, and that the regression should be read as a descriptive summary of the gradient, not as a calibrated predictive instrument.

Third, the non-significant Spearman for the separation ratio ($\rho = 0.60$, $p = 0.208$) exposes a specific limitation: the mean separation ratio, being an average over all $\binom{k}{2}$ group pairs, conflates datasets with very different numbers of categories. Reuters ($k = 8$, 28 pairs) includes both well-separated and nearly-overlapping pairs; the mean masks this heterogeneity. A more granular diagnostic—such as the *minimum* pairwise separation ratio, or a metric that accounts for k—might better capture the difficulty of multi-category clustering tasks.

A genuinely predictive approach would require an unsupervised geometric proxy—one that does not depend on ground-truth labels. Candidates include the Gap statistic, silhouette on predicted clusters, or density-based stability measures. Developing such a proxy, and validating it across a larger sample of conditions, is an important direction for future work.

5.4 Sub-cluster coherence is not a synthetic artifact

The NPMI validation on AG News (Cohen's $d = 6.03$, 100% significant clusters, 91% holdout accuracy) demonstrates that HDBSCAN sub-clusters capture genuine lexical-thematic cohesion in real-world text. This addresses the concern that the sub-topical

structure observed in the synthetic corpus might be circular—an artifact of the LLM imposing structure that the embedding trivially recovers.

A clarification on evaluation scope is warranted. The low ARI of HDBSCAN leaf against macro-group labels (0.108 on the synthetic corpus, Table 1) does not indicate failure: HDBSCAN leaf is not *attempting* to recover the four imposed macro-categories. It discovers finer-grained sub-structure, and its performance should be evaluated by internal coherence (NPMI) and stability (holdout accuracy), not by ARI against a macro-level gold standard. The study thus employs two distinct evaluation regimes: *external evaluation against macro-labels* (ARI/NMI for K-Means and Ward) and *semi-external validation of micro-structure* (NPMI coherence and holdout stability for HDBSCAN). This separation is essential to avoid the misleading conclusion that HDBSCAN "fails" when it is optimising a different objective.

5.5 Topical dominance over source signal

Three converging lines of evidence show that unsupervised clustering on sentence-transformer embeddings recovers topic, not source: (i) the cross-LLM experiment (ARI ≈ 0 against source labels); (ii) the AG News stylistic decomposition (ARI ≈ 0 for source in all feature spaces); and (iii) the finding that nine surface-level stylistic features capture neither topic nor source in AG News.

A precise interpretation is important: that K-Means does not recover source does *not* imply that no source signal exists in the embedding space. It implies that, in the dominant geometric structure accessible to centroid-based unsupervised methods at small k, topical variance overwhelms source variance. A supervised classifier or a targeted linear probe might recover a weak source fingerprint, particularly in controlled designs where topic is held constant (same topic, two LLMs). The IMDB negative control illustrates the symmetric argument: ARI ≈ 0 for sentiment does not mean no polarity information exists in the embedding—only that it does not induce unsupervised clusters. Our claim is therefore: *in the geometry relevant to unsupervised clustering, topic dominates and source is not recoverable.*

5.6 Sparse baselines and the embedding advantage

The TF-IDF + SVD baseline (Table 6) clarifies the contribution of sentence-transformer embeddings. On the synthetic corpus, TF-IDF + SVD matches the embedding pipeline perfectly—confirming that the ARI = 1.000 result is not embedding-specific. On real-world data, however, the gap is substantial: embeddings outperform TF-IDF + SVD by +0.13 to +0.43 ARI. The sparse representation is particularly unstable on AG News (std = 0.12), suggesting that the SVD projection does not reliably preserve category boundaries. Interestingly, LDA (ARI = 0.554 on AG News) far outperforms TF-IDF + SVD + K-Means (0.134), highlighting that the generative topic-model structure of LDA captures information that purely geometric sparse methods miss. The embedding pipeline combines the best of both: geometric clustering with semantic representations that encode category boundaries robustly, plus the capacity for micro-topic discovery via HDBSCAN and cross-lingual generality without vocabulary engineering.

5.7 Limitations

1. **Trivially distant synthetic categories.** The four chosen domains are semantically distant. A systematic study with progressively overlapping domains (e.g. *baking* vs. *cooking*) would map the separability boundary within synthetic text—the most informative region of the gradient.
2. **Limited LLM coverage.** Cross-LLM generalisation was tested with Claude only. Replication with open-weight models (Llama, Mistral) would strengthen generality.
3. **Small diagnostic sample and k-variable confound.** The geometric diagnostics are computed over six conditions in which k varies from 4 to 20. Because clustering difficulty scales with k, the number of categories is confounded with semantic overlap in the gradient—the non-significant Spearman ρ for the separation ratio reflects this. Additional datasets with matched k would disentangle the two effects.
4. **Human control confounds.** The Reddit corpus differs from the synthetic corpus in length distribution, register and lexical entropy, not only in authorial heterogeneity. Part of the observed "homogeneity advantage" may reflect regularities induced by the LLM prompt rather than intrinsic properties of human vs. machine authorship. A length- and entropy-matched resampling would isolate the heterogeneity effect and strengthen the causal interpretation.
5. **AG News source proxy.** The temporal-split source label is a weak proxy; the cross-LLM experiment provides the genuine test.

6 Conclusions

Thematic separability in sentence-transformer embeddings is **intrinsic to the 768D space**: K-Means and Ward achieve perfect recovery on a controlled synthetic corpus without UMAP. UMAP + HDBSCAN provides a complementary micro-topic decomposition into semantically coherent sub-clusters, validated on real-world text via NPMI (Cohen's $d = 6.03$). Separability generalises across embedding architectures and across LLM sources; in the geometry relevant to unsupervised clustering, embeddings encode topic and do not expose generation source.

The empirical gradient from ARI = 1.000 (synthetic) to ARI ≈ 0.40 (20 Newsgroups) tracks embedding geometry: the cosine silhouette shows the most robust association (Pearson $r = 0.83$, $p = 0.042$; Spearman $\rho = 0.83$, $p = 0.042$), though bootstrap CIs remain wide at $n = 6$ and the relationship should be treated as descriptive rather than inferentially conclusive. These diagnostics allow practitioners to understand, post-hoc, why clustering succeeds or fails on a given corpus, while acknowledging that the evidence base is small and structurally uncertain.

Practical guidance:

- For macro-group recovery: cluster directly in 768D.
- For micro-topic discovery: UMAP ($n_{\text{neighbors}} \geq 50$) + HDBSCAN leaf.
- Synthetic LLM results represent optimal conditions; expect ARI ≈ 0.4–0.6 on real-world corpora with genuine semantic overlap.

Reproducibility

All experiments were conducted in Python 3.13. Key packages: sentence-transformers 4.1, scikit-learn 1.6, umap-learn 0.5, hdbscan 0.8, scipy 1.15. Random seeds are fixed at 42 throughout (UMAP, HDBSCAN) with multi-seed stability checks (seeds 0–9) for K-Means. Pre-computed embeddings, databases and all analysis scripts are available at the accompanying Zenodo repository.[1]

Acknowledgements

Not applicable.

Declarations

Funding

No funding was received for conducting this study.

Competing interests

The author declares no competing interests.

Ethics approval

Not applicable.

Availability of data and materials

All corpora used (AG News, 20 Newsgroups, Reuters-21578, IMDB) are publicly available benchmarks. The synthetic and Reddit corpora, pre-computed embeddings and analysis scripts are publicly available at Zenodo: https://doi.org/10.5281/zenodo.18782436.

Code availability

All code is publicly available at Zenodo: https://doi.org/10.5281/zenodo.18782436.

Author contribution

M.P. conceived the study, implemented all experiments, performed the analysis and wrote the manuscript.

References

Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, *12*(4), 461–486,

[1] https://doi.org/10.5281/zenodo.18782436

Angelov, D. (2020). Top2Vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, ,

Apté, C., Damerau, F., Weiss, S.M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, *12*(3), 233–251, https://doi.org/10.1145/183422.183423

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022,

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of gscl* (pp. 31–40).

Campello, R.J.G.B., Moulavi, D., Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Pakdd 2013* (pp. 160–172).

Gehrmann, S., Strobelt, H., Rush, A.M. (2019). GLTR: Statistical detection and visualization of generated text. *Proceedings of acl 2019: System demonstrations* (pp. 111–116).

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, ,

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218,

Lang, K. (1995). Newsweeder: Learning to filter netnews. *Icml 1995* (pp. 331–339).

Lau, J.H., Newman, D., Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *Proceedings of eacl 2014* (pp. 530–539).

Li, Y., Li, Y., Sun, L., Collier, N. (2024). Synthetic data generation with large language models for text classification: Potential and limitations. *Proceedings of emnlp 2024* (pp. 10443–10461).

Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137,

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of acl 2011* (pp. 142–150).

McInnes, L., Healy, J., Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, ,

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *Proceedings of icml 2023* (pp. 24950–24962).

Muennighoff, N., Tazi, N., Magne, L., Reimers, N. (2023). MTEB: Massive text embedding benchmark. *Proceedings of eacl 2023* (pp. 2014–2037).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of emnlp-ijcnlp 2019* (pp. 3982–3992).

Röder, M., Both, A., Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of wsdm 2015* (pp. 399–408).

Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of emnlp-conll 2007* (pp. 410–420).

Sia, S., Dalmia, A., Mielke, S.J. (2020). Tired of topic models? Clusters of pretrained word embeddings make fast and good topics too. *Proceedings of emnlp 2020* (pp. 1728–1736).

Steinbach, M., Karypis, G., Kumar, V. (2000). A comparison of document clustering techniques. *Kdd workshop on text mining* (Vol. 400, pp. 525–526).

Veselovsky, V., Ribeiro, M.H., West, R. (2023). Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, ,

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244,

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N. (2023). C-Pack: Packaged resources to advance general Chinese embedding. *arXiv preprint arXiv:2309.07597*, ,

Zhang, X., Zhao, J., LeCun, Y. (2015). Character-level convolutional networks for text classification. *Neurips 2015* (pp. 649–657).