Metrics for Evaluating BERTopic Models

J. van der Pol

2025-09-11

Bertopic evaluation metrics

This note summarizes the main diagnostics you're computing for BERTopic models—what each metric measures, how it's defined, and how to interpret it in practice. Throughout, remember that BERTopic (via HDBSCAN/UMAP) assigns `-1` to outliers; exclude `-1` from cluster-based metrics to avoid skewing results.

Topic Coherence

C_V coherence

Idea: Measures the semantic consistency of the top words in each topic using **word co-occurrence** (within a sliding window) and **Pointwise mutual information**-based word vectors; coherence is the average cosine similarity among these vectors.

Sketch of formulation.
1. Build word–context vectors with entries given by NPMI between a word and all other words (estimated from the corpus).
2. For a topic t with top words $W_t = w_1, \dots, w_m$, compute mean pairwise cosine similarities:

$$\mathrm{C}_V(t) = \frac{2}{m(m-1)} \sum_{i<j} \cos(\phi(w_i), \phi(w_j))$$

where $\phi(w)$ is the NPMI-based vector of word w. 3. Aggregate across topics: $\mathrm{C}_V = \frac{1}{T} \sum_t \mathrm{C}_V(t)$.

Range & interpretation. 0 – 1 (higher is better). Rough guide: > 0.5 = moderate, > 0.6 = good. This is very sensitive to preprocessing (stopwords, n-grams).

UMass coherence

Idea: Probability-based coherence using **document co-occurrence** counts. Typically more conservative (often negative values).

Formulation: For ordered word pairs (w_i, w_j) from a topic (with w_j treated as a "conditioning" word), let:
- $D(w_j)$ = number of documents containing w_j
- $D(w_i, w_j)$ = number of documents containing both

Then for topic t with word list $w_1, \ldots, w_m$:

$$\mathrm{C}_{\mathrm{UMass}}(t) = \frac{2}{m(m-1)} \sum_{i>j} \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)}$$

with small ϵ (e.g., 1) for smoothing. Average across topics for the model score.

Range & interpretation. Typically $(-\infty, 0]$. **Closer to 0 is better.** Very negative = weak co-occurrence.

Silhouette Score (clustering quality)

Idea: How well documents fit their assigned topic vs. the nearest alternative topic, based on distances in **embedding space**.

Formulation. For document i: - $a(i)$ = mean distance to all other points in the **same** cluster - $b(i)$ = lowest mean distance to points in **any other** cluster

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \in [-1, 1].$$

Model score is the mean $s(i)$ over all non-outlier documents.

Distance choice. With sentence embeddings, use **cosine distance**:

$$d_{\cos}(x, y) = 1 - \frac{x \cdot y}{\|x\|\,\|y\|}.$$

Interpretation: > 0.5 strong structure, 0.2–0.5 weak/moderate, ≤ 0 poor separation. In high-dimensional text embeddings, expect modest values; compare **relatively** across parameter/model choices.

Topic Diversity (unique top-words ratio)

Idea: Measures redundancy across topics' top-n words.

Formulation. If there are T topics (excluding -1) and you take the top n words per topic,

$$\text{Diversity} = \frac{\left|\bigcup_{t=1}^{T} \text{Top}_n(t)\right|}{T \times n} \in [0, 1].$$

Interpretation. Higher = less reuse of the same words across topics (more distinct vocabularies). Very high diversity with **very small** topics can indicate over-fragmentation; very low diversity suggests overlapping or generic topics.

Cluster Size Statistics

Idea. Describes how many documents fall into each topic.

Formulation. Let $c_1, \ldots, c_T$ be the sizes of non-outlier clusters. Report:

- **Mean size:** $\bar{c} = \frac{1}{T} \sum_{t=1}^{T} c_t$
- **Min size:** $\min_t c_t$
- **Max size:** $\max_t c_t$

Interpretation. - Many tiny clusters → potential over-clustering (e.g., too small `min_cluster_size`, too few neighbors). - One or two huge clusters → under-clustering or overly aggressive parameters (e.g., very large `min_cluster_size`, too many neighbors).

Use these alongside silhouette and coherence to tune parameters such as `min_cluster_size`, `min_samples`, `n_neighbors`, and UMAP dimensionality.

Topic Entropy (normalized)

Idea. How evenly documents are distributed across topics.

Formulation. With cluster counts $c_1, \ldots, c_T$ and $p_t = c_t / \sum_{k=1}^{T} c_k$:

$$H = -\sum_{t=1}^{T} p_t \log_2 p_t, \qquad H_{\text{norm}} = \frac{H}{\log_2 T} \in [0, 1].$$

Interpretation.
- ≈ 0: one/few topics dominate (imbalanced).
- ≈ 1: documents spread evenly (balanced).

Balance is not always "better"—aim for a distribution that matches domain expectations.

Practical Reading Guide

- **Compare relatively.** Use these scores to **rank** model/parameter settings trained on the **same corpus** and preprocessing.
- **Trade-offs are normal.** Increasing semantic tightness (↑ coherence) can reduce separation (↓ silhouette) or increase overlap (↓ diversity).
- **Preprocessing matters.** Stopword handling, n-grams, lemmatization, and rare-word pruning strongly affect coherence and diversity.
- **Mind outliers (-1).** Exclude them for silhouette, entropy, and size stats; report the **outlier rate** separately when useful.

Typical Ranges (rule-of-thumb)

- **C_V:** 0–1; higher is better (often 0.5–0.7+ after good preprocessing).
- **UMass:** ≤ 0; closer to 0 is better (very negative = poor co-occurrence).
- **Silhouette:** $[-1, 1]$; > 0.2 modest, > 0.5 strong (text often yields modest values).
- **Diversity:** $[0, 1]$; > 0.85 suggests distinct vocabularies, but check for over-fragmentation.
- **Entropy (norm):** $[0, 1]$; closer to 1 = balanced topic sizes; closer to 0 = dominance by few topics.

Use these together—no single metric tells the whole story.