Semantic

Yi-Chun (Rimi) Chen

November 24, 2025

1 Semantic Similarity for the Dataset

Our experiments explored three clause representations: **TF-IDF** (surface lexical overlap within the dataset), **LSA** (latent factors also derived only from the dataset), and **SBERT** (pretrained sentence embeddings). The first two capture dataset-specific statistics rather than true semantics, while SBERT provides a general semantic space learned from large external corpora.

The correct task is to represent each clause in a *general semantic space*, where paraphrases and meaning-equivalent expressions are close even without direct word overlap. This is only achievable with pretrained contextual embeddings (e.g., SBERT, SimCSE, Clinical-BERT).

Table 1: Similarity metrics versus representations: what counts as "semantics" and whether it is dataset dependent.

Representation	Metric	Semantic?	Dataset dependent?
TF-IDF bag of words	Cosine	No (lexical overlap)	Yes
LSA (from same data)	Cosine	Minimal (latent factors)	Yes
Pretrained sentence embeddings	Cosine	Yes (contextual semantics)	No

Detailed interpretation:

- **TF–IDF + cosine.** Representation is built from the training corpus vocabulary. Result: similarity is dataset specific and mostly lexical. Cosine here does not make it "general semantics."
- **LSA + cosine.** Representation comes from latent factors induced by the same corpus. Result: still dataset-specific smoothing, not general semantics.
- **SBERT + cosine.** Representation is pretrained on large external corpora. Result: clause similarity reflects general semantics much better and is not tied to dataset idf or LSA space.

In short, cosine is only a *metric*; the semantics come from the representation. TF-IDF and LSA induce corpus-specific spaces, while pretrained embeddings enable dataset-independent semantic comparison of clauses.