**Overview**

This project explores the comparative performance of two Sentence-BERT (SBERT) models in a Retrieval-Augmented Generation (RAG) pipeline using Qdrant as the vector database. The objective is to evaluate how each model handles semantic search queries over a dataset over three different datasets related to food and service experiences.

**Models Used**

The following pre-trained SBERT models were evaluated:

* all-MiniLM-L6-v2 (sbert1): A lightweight general-purpose embedding model optimized for sentence-level semantic similarity.
* multi-qa-MiniLM-L6-cos-v1 (sbert2): Fine-tuned specifically for multi-question answering tasks, designed to improve relevance in QA retrieval use cases.

Both models produce 384-dimensional vectors, which were stored in separate Qdrant collections for comparison.

**Dataset Summary**

* Dataset1:
  + Small dataset of products reviews
* Dataset2:
  + Larger dataset of car reviews
* Dataset3:
  + Largest dataset of restaurant reviews

**Search Evaluation**

Example from dataset1:

* Query: “terrible”
* **--- sbert1 Results ---**
* It’s okay, not too bad but not great either. (?)
* Worst purchase I've ever made. (✓)
* Terrible experience. The item broke after one use. sbert1 Results: Emphasized flavor, steak, and desserts; consistent theme around meal quality. (✓)
* **--- sbert2 Results ---**
* Worst purchase I've ever made. (✓)
* It’s okay, not too bad but not great either. (?)
* Absolutely amazing customer service! ⮽

Example from dataset2:

* Query: “good”
  + **--- sbert1 Results ---**
  + Good but not great. It's a reliable car but lacks excitement. (✓)
  + Good car, but not as fast as I expected. (✓)
  + I’m not satisfied, it keeps having electrical issues. ⮽
  + **--- sbert2 Results ---**
  + Good but not great. It's a reliable car but lacks excitement. (✓)
  + Good car, but not as fast as I expected. (✓)
  + I’m very pleased with this car. Excellent value for the money. (✓)

**Final Results**

Overall all-MiniLM-L6-v2 (sbert1) did a better job than the other model. It had more hits from each query and less outliers in its results.

**Observations**

* Model Similarity: In queries with overlapping language to training data (e.g., "delicious", "wonderful"), both models performed similarly.
* Sensitivity to Vocabulary: Rare or expressive adjectives not present in the dataset led to noisy results, even for very close terms.
* Speed: Embedding and search speed was performant for both models, completing searches in under a second per query on local hardware.

**Tools & Environment**

* Vector Store: Qdrant (localhost, default port)
* Embedding Framework: sentence-transformers
* Environment: Python 3.11, local machine execution

**Conclusion**

This evaluation confirms that both all-MiniLM-L6-v2 and multi-qa-MiniLM-L6-cos-v1 are viable for short-text RAG tasks. While general-purpose models like all-MiniLM-L6-v2 offer fast and broad coverage, domain-tuned models like multi-qa-MiniLM-L6-cos-v1 may slightly outperform in tasks involving nuanced question-answering.

Future steps could involve:

* Adding more diverse queries including negatives or domain-specific jargon.
* Incorporating a scoring metric
* Expanding to a larger, more realistic dataset