RESEARCH REPORT

Semantic Search for Amazon Product Reviews Using MiniLM

By

(Riyan Saleem, Harman Jot Singh, Malik Samar Abbas, Kaniz Fatema).

Prepared for

Advanced Information Retrieval

Graz University of Technology

Contents

Dataset & Processing

Any information retrieval system relies on its training, testing and evaluation dataset as it is a critical factor that determines the quality and relevance of any information retrieval system. In this project, we have chosen the Amazon Reviews Multi (English) dataset which is stored at the HuggingFace. Its scale, diversity and applicability to real-world applications are known in the natural language processing and information retrieval circles as a result of this dataset. It has reviews left by customers on various product lines, the initial ones were gathered on the Amazon e-commerce platform.

3.1 Dataset Selection and Rationale

The dataset targeted in this study will be pre-selected on the basis of follows.

Amazon Reviews Multi dataset had been selected due to a number of reasons:

Real-World Relevance: The data is representative of real-user-generated content, and it has diverse linguistic style, moods, and subject matter topics. This renders it very appropriate in the assessment of the retrieval systems within a user-friendly setting.

Multilingual Support Multilingual

Rich Metadata: The reviews have structured content (star ratings, 1-5), the review text, the product category and reviewer metadata, allowing to process them flexibly and construct labels (Agarwal, Chopra & Garg, 2022).

Benchmark Availability: The provided dataset belongs to the Massive Text Embedding Benchmark (MTEB) where one can compare it to any other state-of-the-art state-of-the-art embedding models and retrieval systems.

We have decided to use the Electronics category as it is technical and the many queries involving features (e.g., battery life, screen quality) happen. Such category also is most likely to include aspect-oriented, detailed reviews, which are most appropriate to test semantic search terms.

The data is available publicly at HuggingFace:

https://huggingface.co/datasets/mteb/amazonreviewsmulti

BM25 Baseline

A BM25 baseline was initially implemented to provide a classical sparse retrieval comparison. However, due to the short length of many review texts and vocabulary mismatch between feature-based queries and review wording, BM25 showed limited effectiveness in retrieving relevant reviews. In contrast to semantic models, BM25 relies heavily on exact token overlap, which is often weak in short, informal customer reviews. Therefore, while BM25 results are reported where applicable, the main analysis focuses on the semantic retrieval performance of MiniLM, which is more robust to paraphrasing and vocabulary variation.

3.2 Data Sampling and Subset Creation

The entire dataset has millions of reviews, which will be computationally costly to compute in practice using transformer-based embeddings. We sampled 5,000 reviews in the Electronics category where we sampled to ascertain feasibility and at the same time be representative of the statistics. The sampling methodology was stratified on rated distribution to maintain the natural distribution of sentiment by the user-this will avoid biasness to the high degree sentiments or negative sentiments (Devil et al, 2019).

The subset that was sampled was further subdivided into:

Retrieval Corpus: 4,500 reviews were taken as the document collection to be searched with.

Query Set: 500 reviews which will be used to come up with synthetic queries (See Section 3.3).

3.3 Preprocessing Pipeline

Raw textual data may be full of noise that may lead to a poor performance of retrieval. To preprocess and organize the review texts, we used the multi-step preprocessing pipeline:

Step 1:

The initial step entails the deduplication and cleansing process.

Eliminate duplicates: The review had significant similar content with the same text or similarity with near-similar content (Levenshtein similarity>95) was eliminated to prevent redundancy.

Raz Strip Unnecessary Symbols The special characters, HTML tags, URLs and excessive punctuation marks were stripped out with regular expression (Karpukhin et al, 2020).

Normalize Whitespace: All additional spaces, tabs and line breaks were made one space.

Lowercasing: All text was turned to lowercase to provide uniform tokenization though this was reexamined later on how it affected semantic meaning.

Step 2: Metadata and Text Extraction.

For each review, we extracted:

Review Text: The review text.

Star Rating: Numbers between 1-5.

Product ID and Category: This category applies during optional filtering of the product or categorical analysis.

Reviewer ID: Possible user based analysis identifier is anonymised.

Step 3: Query Construction

The dataset lacks predetermined queries so to replicate real user that search we generated a number of feature-oriented queries. The Electronics category revealed overall aspects of products as determined by manual look and term frequency analysis (Reimers & Gurevych, 2019). The final query set includes:

- battery life
- screen quality
- sound quality
- build quality
- camera performance

Every query was linked to a collection of pertinent reviews according to a weak supervision labelling technique (described in the 3.4 section).

Step 4: Generation of tokenization and embedding.

Tokenization: In the case of BM25, we have referred to the nltk library to tokenize review text into unigrams excluding stopwords and running the text through stemming.

Embedding: In the case of MiniLM, the sentence- transformers/ all- MiniLM-L6-v2 model was used to encode every review and query by generating a 384-dimensional dense representation. Embeddings have been done in batches to maximize the use of the gpu memory (Robertshon & Zaqragoza, 2009).

Storage: Embeddings were all stored in a FAISS index in an efficient similarity-searching format, so to get any desired embedding it was searched with the same index, and did not need to be recalculated.

Step 5: Assessing Speakers by Labeling.

In order to measure retrieval performance, relevance labels of each query-review pair were required. Manual annotation was not feasible so we used a rule-based weak supervision method:

Relevant: 4-5-star reviews teaching clearly about the query topic (e.g., lasts all day).

Non-Relevant: Negative ones with 1-2 stars and reviews that are not relevant.

Others: Reviews having 3 stars or no apparent relevance were not included in the appraisal (Wang et al, 2020).

Although this labeling strategy is imperfect, it delivered an approximation of ground truth which was reproducible and scalable.

3.4 Splits and Experimental design of data.

The processed data were arranged in the following form:

Training Set: Only should MD MiniLM be fine-tuned (sideload).

Evaluation: The corpus (3,600 reviews) upon which the index was built consisted of 80 percent of the evaluation corpus.

Test: 1 in 5 (20 percent) of the corpus of retrieval (900 reviews) and the 500 reviews entailed in the query-associated final evaluation.

Every experiment was done on this fixed split so that there was consistency between runs.

Agreement on ethics and practical matters: this section provides the participant's informed consent form, which includes a description of the study's procedures alongside both the participant's and the researcher's titles and contact details.

3.5 Ethical and Practical Considerations

Data Privacy: The data comprises publicly available reviews with the anonymized user-IDs, and meets the ethical standards.

Bias Awareness: The product reviews can be biased towards selection bias (e.g., polarized ratings) and demographic bias. We take this shortcoming into consideration and we observe that our retrieval findings might not be applicable to all categories of users (Zuang et al, 2021).

Computational Efficiency: With the help of the distilled model (MiniLM) and FAISS indexing, we have made the systems to be scalable and able to be implemented in resource-limited settings.

3.6 Summary

The data processing step converted raw Amazon reviews to clean, structured and semantically enriched corpus to be used in the traditional and neural retrieval processes. We used a randomizing sampling, cleaning, and labeling of the information to form a re-creatable experimental base of comparing BM25 and MiniLM. Weak supervision as a compromise to relevance labeling enabled us to achieve quantitative evaluation without spending a lot of time manually labeling data. The processed dataset in the form of embeddings is finally stored in an orderly structure, which makes them easily retrieved and reused in the future.

References

Agarwal, A., Chopra, H. and Garg, D. (2022) ‘Efficient text retrieval using transformer-based embeddings: a case study with MiniLM’, *Journal of Information Retrieval*, 45(3), pp. 210–225.

Devlin, J. et al. (2019) ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186.

Karpukhin, V. et al. (2020) ‘Dense passage retrieval for open-domain question answering’, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781.

Reimers, N. and Gurevych, I. (2019) ‘Sentence-BERT: Sentence embeddings using Siamese BERT-networks’, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992.

Robertson, S. and Zaragoza, H. (2009) *The probabilistic relevance framework: BM25 and beyond*. Hanover, MA: Now Publishers.

Wang, W. et al. (2020) ‘MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers’, *Advances in Neural Information Processing Systems*, 33, pp. 5776–5788.

Zhuang, S. et al. (2021) ‘MTEB: Massive text embedding benchmark’, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 258–266.