MIDI 2024 Conference

Pushing the Boundaries of Turkish Semantic Search: A Fine-Tuned NV-Embed-v2 Approach

Abstract

This study presents the development of a semantic search model optimized for Turkish-language data, using NV-Embed-v2 as the base model. The model was fine-tuned on a custom dataset created from Camlica Basim Yayin publications, encompassing rich question-answer pairs and document titles. This dataset preparation, along with advanced hyperparameters and training techniques, enabled the model to capture semantic nuances and provide relevant query results. Our fine-tuned model demonstrated significant improvements over the pretrained version, achieving 80% in Top-1 accuracy and 92% in Top-5 accuracy, highlighting its effectiveness for Turkish semantic search.

Due to the extended training time required for additional epochs, we plan to implement a feedback system to gather user insights during live deployment. This feedback will allow us to refine the model based on real-world user behavior, which may differ from controlled test scenarios. Future work involves the integration of tag generation and a hybrid search engine, combining semantic search with full-text retrieval, to further enhance the accuracy and user experience of the system. Overall, this research demonstrates the potential of fine-tuning large language models for specialized semantic search tasks in Turkish and lays the groundwork for future advancements.