Identifikasi Otomatis Pertanyaan Duplikat pada Forum Kesehatan Berbahasa Indonesia dengan Memanfaatkan Learning-to-Rank = Automatic Identification of Duplicate Questions in Indonesian Consumer Health Forums Using Learning-to-Rank

Febi Imanuela, author

Abstrak

Perkembangan teknologi pada bidang kesehatan di Indonesia telah menghadirkan layanan konsultasi dengan dokter melalui forum tanya jawab kesehatan. Seiring dengan berjalannya waktu, muncul permasalahan pertanyaan duplikat pada forum. Permasalahan ini perlu ditangani agar dapat mempercepat proses pengembalian jawaban untuk keluhan yang serupa dan menjaga jumlah pertanyaan agar tetap scalable dengan kapasitas dokter penjawab. Namun, pertanyaan duplikat merupakan suatu tantangan tersendiri karena kompleksitas bahasa natural. Penelitian ini memanfaatkan pendekatan Information Retrieval untuk mengidentifikasi pasangan pertanyaan duplikat pada domain ini sebagai suatu pasangan query dan dokumen yang relevan. Setelah melakukan ranking awal menggunakan BM25 sebagai model baseline, performa hasil ranking ditingkatkan melalui proses re-ranking menggunakan model learning-to-rank LambdaMART yang berbasis fitur. Penelitian ini memanfaatkan fitur perhitungan jarak dan similaritas antara pasangan vektor representasi query dan dokumen, yang diperoleh dari model word embeddings dan transformer. Selain itu, diusulkan fitur scoring yang diperoleh dari model Cross Encoder, serta model BM25 yang menjadi model baseline. Penelitian ini juga mengusulkan fitur-fitur yang mempertimbangkan jumlah keywords gagasan utama query yang dikandung dokumen. Evaluasi eksperimen dilakukan menggunakan cross validation dan error analysis, dengan MRR sebagai metrik utama. Performa tertinggi yang dicapai eksperimen adalah MRR senilai 0,951 dengan p value senilai 0,016 yang signifikan terhadap baseline. Dengan demikian, penelitian ini menunjukkan dukungan empiris terhadap peningkatan efektivitas model re-ranking yang diusulkan untuk melakukan identifikasi otomatis terhadap karakteristik query dan dokumen yang relevan, yakni pasangan pertanyaan duplikat dalam konteks ini.

.....The development of technology in the healthcare sector in Indonesia has introduced consultation services with doctors through consumer health forums. Over time, the issue of duplicate questions on these forums emerged. This problem needs to be addressed to accelerate the response process for similar questions and to keep the number of questions scalable with the capacity of the responding doctors. However, duplicate questions present their own challenge due to the complexity of natural language. This study utilizes Information Retrieval approach to identify pairs of duplicate questions in this domain as query and relevant document pairs. After initial ranking using BM25 as the baseline model, the ranking performance is improved through a re-ranking process using the feature-based LambdaMART model. This study leverages features that calculate the distance and similarity between vector representations of the query and document, obtained from word embedding and transformer models. Additionally, scoring features derived from the Cross Encoder model and the BM25 baseline model are proposed. The study also suggests features that consider the number of main idea keywords from the query that is also contained within the document. Experiment evaluation is conducted using cross validation and error analysis, with Mean Reciprocal Rank (MRR) as the primary metric. The highest performance achieved in the experiments is an MRR of 0.951

with a p-value of 0.016, which is significant to the baseline. Thus, this study provides empirical support for the effectiveness of the proposed re-ranking model for automatic identification of the query and relevant document, specifically duplicate question pairs in this context.