ABSTRAK

Peningkatan jumlah artikel ilmiah di bidang medis selama beberapa dekade terakhir, terutama pasca-pandemi COVID-19, telah menciptakan tantangan bagi peneliti dan praktisi kesehatan dalam menemukan artikel yang relevan secara efisien. Kendala tersebut terdiri dari kompleksitas istilah medis, variasi terminologi, serta keterbatasan waktu. Penelitian ini bertujuan untuk mengembangkan sistem rekomendasi berbasis Content-Based Filtering dengan memanfaatkan penggabungan metode TF-IDF dan fastText. TF-IDF digunakan untuk mengekstrak informasi relevan dari artikel, sementara fastText membantu menangkap kemiripan semantik antar istilah medis.

Dataset yang digunakan berjumlah 4500 yang diperoleh dari situs Elsevier berisikan data informasi karya ilmiah medis yang meliputi, judul, *Digital Object Identifier*(DOI), abstrak, link artikel, tanggal publikasi, nama publikasi, dan isi artikel dengan lingkup topik yang dibatasi pada bidang tertentu. Proses pra-pemrosesan data dilakukan melalui enam tahap, yaitu penghapusan duplikat, *case folding* menjadi *lowercase*, tokenisasi, filter karakter non-alfabet, penghapusan *stopwords*, dan lemmatisasi menggunakan library SpaCy. Setelah melalui tahapan pra-pemrosesan data, data tersebut digunakan sebagai input training model fastText dengan parameter tertentu menggunakan library *gensim*, sementara skor TF-IDF dihitung untuk menentukan bobot setiap kata dalam dokumen. Hasil penggabungan kedua metode tersebut menghasilkan representasi dokumen berbasis fastText dengan pembobotan TF-IDF.

Sistem ini diimplementasikan menggunakan framework Flask dan dilakukan sebuah skenario pengujian dengan lima query yang dirancang untuk mencakup berbagai topik medis. Evaluasi dilakukan menggunakan metrik Normalized Discounted Cumulative Gain (NDCG), menghasilkan rata-rata nilai NDCG sebesar 0,955. Nilai tertinggi sebesar 1.0 diperoleh pada query ketiga dan keempat, sementara query pertama dan kelima memperoleh nilai mendekati sempurna, masing-masing 0,968 dan 0,961. Query kedua menunjukkan nilai terendah, yaitu 0,847, mengindikasikan adanya artikel teratas yang kurang sesuai dengan ekspektasi partisipan. Survei terhadap partisipan yang memiliki latar belakang di bidang kesehatan menunjukkan bahwa 25% partisipan merasa sistem berhasil mengatasi variasi istilah medis sepenuhnya, sementara 75% merasa masih ada istilah yang tidak teratasi. Selanjutnya, seluruh partisipan menilai sistem cukup membantu memahami topik medis meskipun terdapat variasi istilah. Terakhir, sebagian besar partisipan menilai sistem cukup efektif dalam menyederhanakan pencarian artikel medis, meskipun 25% responden merasa efektivitas sistem masih perlu ditingkatkan.

Kata Kunci: fastText, TF-IDF, NDCG, Content-Based Filtering, semantik, istilah medis, artikel ilmiah.

ABSTRACT

The increase in the number of scientific articles in the medical field over the past few decades, especially post-COVID-19 pandemic, has created challenges for researchers and healthcare practitioners in efficiently finding relevant articles. These challenges include the complexity of medical terminology, terminological variations, and time constraints. This research aims to develop a recommendation system based on Content-Based Filtering by integrating the TF-IDF and fastText methods. TF-IDF is used to extract relevant information from articles, while fastText aids in capturing semantic similarities among medical terms.

The dataset used consists of 4,500 records obtained from the Elsevier website, containing information about medical scientific works, including titles, Digital Object Identifiers (DOIs), abstracts, article links, publication dates, publication names, and article content, with a focus on specific topics. The data preprocessing stage involves six steps: duplicate removal, case folding to lowercase, tokenization, filtering non-alphabetic characters, stopword removal, and lemmatization using the SpaCy library. After completing the preprocessing data, the data is used to train the fastText model with specific parameters using the Gensim library, while the TF-IDF scores are calculated to determine the weight of each word in the document. The integration of these two methods results in fastText-based document representations with TF-IDF weighting.

The system is implemented using the Flask framework, and a testing scenario is conducted with five queries designed to cover various medical topics. Evaluation using the Normalized Discounted Cumulative Gain (NDCG) metric yields an average NDCG score of 0.955. The highest score, 1.0, is achieved for the third and fourth queries, while the first and fifth queries achieve near-perfect scores of 0.968 and 0.961, respectively. The second query shows the lowest score, 0.847, indicating some top-ranked articles did not meet participant expectations. A survey conducted with participants from healthcare backgrounds revealed that 25% felt the system fully addressed medical terminological variation, while 75% indicated that some terms were still unresolved. Furthermore, all participants agreed that the system was helpful in understanding medical topics despite the presence of terminological variation. Lastly, most participants considered the system effective in simplifying the search for medical articles, although 25% suggested that the system's effectiveness could still be improved.

Keywords: fastText, TF-IDF, NDCG, Content-Based Filtering, semantic, medical terms, scientific articles.