

ABSTRAK

Penelitian ini berfokus pada perancangan dan implementasi mesin pencari berita menggunakan model TF-IDF dan *Cosine similarity* dengan dataset dari Detik.com. Dalam era informasi yang berkembang pesat, pengguna seringkali menghadapi masalah kelebihan informasi, sehingga sulit menemukan berita yang relevan dengan minat mereka. Oleh karena itu, sistem pencarian yang akurat diperlukan untuk membantu pengguna dalam menemukan berita yang sesuai. Metode yang diusulkan meliputi pengumpulan data melalui *web crawling*, diikuti dengan tahapan *Preprocessing* yang mencakup *case folding*, tokenisasi, penghapusan *stopword*, dan lemmatisasi. Setelah itu, fitur diekstraksi menggunakan TF-IDF untuk menghitung bobot kata, dan tingkat kemiripan antara query pengguna dan dokumen berita diukur menggunakan *Cosine similarity*. Evaluasi sistem dilakukan dengan metrik presisi untuk menilai relevansi hasil pencarian berdasarkan skenario satu, dua, dan tiga kata kunci. Hasil pengujian menunjukkan bahwa sistem berhasil melakukan perankingan dokumen dari nilai kemiripan tertinggi hingga terendah. Pengujian dengan satu kata kunci "banjir" menghasilkan nilai presisi sebesar 60%. Namun, pengujian dengan dua kata kunci "banjir Jakarta" dan tiga kata kunci "banjir Jakarta Februari" menghasilkan nilai presisi 0%. Hal ini terjadi karena keterbatasan variasi konten pada dataset sebanyak 1.260 artikel yang dilakukan *crawling*, di mana tidak terdapat dokumen yang memuat konten spesifik terkait banjir di wilayah Jakarta pada waktu tersebut. Penelitian ini menyimpulkan bahwa performa mesin pencari berbasis konten sangat bergantung pada kekayaan dan cakupan data di dalam dataset.

Kata Kunci: TF-IDF, *Cosine similarity*, mesin pencari berita, *web crawling*, presisi.

ABSTRACT

This research focuses on the design and implementation of a news search engine using TF-IDF and Cosine similarity models with a dataset sourced from Detik.com. In an era of rapid information growth, users often face the challenge of information overload, making it difficult to find news relevant to their interests. Therefore, an accurate search system is required to assist users in discovering appropriate news content. The proposed method involves data collection through web crawling, followed by Preprocessing stages including case folding, tokenization, stopword removal, and stemming. Subsequently, features are extracted using TF-IDF to calculate word weights, and the degree of similarity between user queries and news documents is measured using Cosine similarity. The system is evaluated using precision metrics across one, two, and three-keyword scenarios. The test results indicate that the system successfully ranks documents from the highest to the lowest similarity scores. Testing with a single keyword "banjir" yielded a precision value of 60%. Conversely, tests using two keywords "banjir Jakarta" and three keywords "banjir Jakarta Februari" resulted in a precision score of 0%. This occurred due to the limited variation of content within the 1,260 crawled articles, which lacked specific documents concerning floods in Jakarta during that period. This study concludes that the performance of a content-based search engine is heavily dependent on the richness and comprehensiveness of the data within the dataset.

Keywords: *TF-IDF, Cosine similarity, news search engine, web crawling, precision.*