

ANALISIS PERBANDINGAN METODE TF-IDF, TF-IDF + *WORDNET*, DAN *SENTENCE-BERT* DALAM PENCARIAN DOKUMEN BERITA MENGGUNAKAN *COSINE SIMILARITY*

Galib Haftha Zuhayir, Wiwik Suharso, Nanda Kurnia Wardati

Program Studi Sistem Informasi, Fakultas Teknik,
Universitas Muhammadiyah Jember

Email: ghafthaz@gmail.com, wiwiksuharso@unmuhjember.ac.id,
nandakurniawardati@unmuhjember.ac.id

ABSTRAK

Penelitian ini bertujuan menganalisis efektivitas dan efisiensi komputasi metode TF-IDF, TF-IDF + *WordNet*, dan *Sentence-BERT* (SBERT) dalam *Information retrieval* dokumen berita menggunakan *Cosine Similarity*. *Dataset* yang digunakan adalah *BBC News Dataset* sebanyak 42.115 dokumen mentah, setelah proses *cleaning* diperoleh 40.111 dokumen bersih dan 14.305 dokumen dengan label *ground truth* hierarkis dari struktur URL (5 Topik L1 dan 53 Subtopik L2) dijadikan korpus final. Representasi dokumen dibatasi pada atribut *title* dan *description*. Relevansi ditentukan berdasarkan kesesuaian Topik dan Subtopik (L2) terhadap target *query*. Evaluasi dilakukan pada 10 *query* eksperimen (5 *query* L1 dan 5 *query* L2) menggunakan *Precision@K*, *Recall@K*, *F1-score@K* ($K = 5, 10, 20$), serta *execution time*. Hasilnya, SBERT mengungguli kedua metode lain pada $K=20$ mencapai *Precision* 0,8850, *Recall* 0,0288, dan *F1-score* 0,0539, dibandingkan TF-IDF (0,8000; 0,0234; 0,0441) dan TF-IDF + *WordNet* (0,7450; 0,0221; 0,0417). Pada fase *retrieval online*, SBERT (0,013 detik/*query*) sebanding dengan TF-IDF, sedangkan TF-IDF + *WordNet* paling lambat (0,036 detik/*query*) tanpa peningkatan akurasi, karena ekspansi sinonim tanpa disambiguasi konteks justru menurunkan presisi. SBERT direkomendasikan untuk sistem yang mengutamakan akurasi semantik, sedangkan TF-IDF sesuai untuk sistem dengan sumber daya terbatas.

Kata kunci: *Information retrieval*, TF-IDF, *WordNet*, *Sentence-BERT*, *Cosine Similarity*.