

ABSTRAK

Volume artikel berita bahasa Indonesia yang terus bertambah sering kali membuat pencarian informasi konvensional menjadi kurang efisien secara komputasi dan kerap menghasilkan dokumen yang tidak relevan dengan intensi pengguna. Penelitian ini bertujuan untuk meningkatkan efisiensi waktu komputasi dan relevansi hasil pada sistem pencarian artikel berita bahasa Indonesia. Teknik pengambilan data dilakukan melalui *web scraping* dari portal berita CNN Indonesia, dilanjutkan dengan ekstraksi kata kunci menggunakan algoritma YAKE untuk menyimulasikan kueri pengguna. Metode penelitian yang diusulkan memanfaatkan model IndoBERT sebagai tahap penyaringan awal untuk mengklasifikasikan kueri pencarian ke dalam lima kategori spesifik, sehingga ruang pencarian dokumen menjadi lebih sempit dan terarah. Selanjutnya, analisis data dilakukan dengan mengukur tingkat kemiripan semantik antara kueri dan artikel berita pada kategori terpilih menggunakan representasi BERT dan *Sentence-BERT* (SBERT) dan kalkulasi *Cosine Similarity*. Kinerja sistem kemudian dievaluasi dengan membandingkan skenario pencarian menyeluruh secara murni dan pencarian berfilter kategori menggunakan parameter metrik *Mean Average Precision* (MAP), *Normalized Discounted Cumulative Gain* (nDCG), *Precision*, *Recall*, serta perbandingan efisiensi waktu komputasi.

Kata Kunci: Pencarian Berita, Klasifikasi Kueri, IndoBERT, Sentence-BERT, *Cosine Similarity*.

ABSTRACT

The continuously growing volume of Indonesian news articles often makes conventional information retrieval computationally inefficient and frequently yields documents that are irrelevant to the user's intent. This study aims to improve the computational time efficiency and result relevance of an Indonesian news article search system. Data collection was conducted through web scraping from the CNN Indonesia news portal, followed by keyword extraction using the YAKE algorithm to simulate user queries. The proposed research method utilizes the IndoBERT model as an initial filtering stage to classify search queries into five specific categories, thereby narrowing and targeting the document search space. Furthermore, data analysis was performed by measuring the semantic similarity between the queries and the news articles within the selected categories using BERT and Sentence-BERT (SBERT) representations, along with Cosine Similarity calculations. The system's performance was then evaluated by comparing a purely exhaustive search scenario with a category-filtered search scenario using Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (nDCG), Precision, and Recall metrics, as well as a comparison of computational time efficiency.

Keywords: *News Search, Query Classification, IndoBERT, Sentence-BERT, Cosine Similarity.*

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	ii
HALAMAN PERNYATAAN ORISINALITAS	iii
ABSTRAK.....	iv
<i>ABSTRACT</i>	v
PENGANTAR.....	vi
DAFTAR ISI	vii
DAFTAR GAMBAR	x
DAFTAR TABEL.....	xi
DAFTAR KODE PROGRAM	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.2.1 Permasalahan.....	3
1.2.2 Metode Usulan	4
1.2.3 Pertanyaan Penelitian	4
1.3 Tujuan dan Manfaat Penelitian	4
1.3.1 Tujuan Penelitian.....	4
1.3.2 Manfaat Penelitian	4
1.4 Batasan Masalah.....	5
1.5 Sistematika Penulisan	5
BAB II KAJIAN PUSTAKA	7
2.1 Pencarian Berita	7
2.2 YAKE (Yet Another Keyword Extraction).....	7
2.3 Preprocessing	8
2.4 Transformer.....	8
2.4.1 Input Embedding.....	9
2.4.2 Encoder dan Decoder	10
2.4.3 Mekanisme Self-Attention	12
2.5 BERT (Bidirectional Encoder Representation from Transformer)	12
2.5.1 Representasi Input BERT	13
2.5.2 Proses Pelatihan BERT	14

2.6	Representasi Vektor Matriks	15
2.6.1	Representasi Vektor Matriks BERT	15
2.6.2	Representasi Vektor Matriks <i>Sentence</i> -BERT	15
2.7	IndoBERT-BASE	16
2.8	Similarity Ranking	17
2.9	Evaluasi Similarity Ranking	17
2.10	Penelitian Terdahulu	19
BAB III METODOLOGI PENELITIAN		24
3.1	Kerangka Penelitian	24
3.2	Teknik Pengumpulan Data	25
3.1.1	Scraping Data Berita Indonesia	25
3.1.2	Keyword Extraction YAKE	26
3.3	Metode / Alat yang Digunakan	30
3.2.1	Klasifikasi Model IndoBERT	30
3.2.2	Representasi Vektor Matriks	38
3.2.3	Cosine Similarity Ranking	40
3.2.4	Evaluasi Similarity	44
3.2.5	Skenario Uji Coba	46
BAB IV HASIL DAN PEMBAHASAN		47
4.1	Pengumpulan dan Pengolahan Data	47
4.1.1	Pengumpulan Berita Indonesia	47
4.1.2	Data Understanding	51
4.1.3	Preprocessing Data Berita Indonesia	52
4.1.4	Keyword Extraction Data Berita Indonesia	55
4.1.5	<i>Preprocessing</i> Data <i>Keyword</i> Berita Indonesia	57
4.1.6	Pembagian Data <i>Keyword</i> Berita Indonesia	58
4.2	Pembahasan dan Analisis Hasil Penelitian	59
4.2.1	Model BERT Klasifikasi	60
4.2.2	Representasi Vektor Matriks	62
4.2.3	Similarity Ranking	64
4.2.4	Evaluasi Similarity Ranking	65
4.2.5	Skenario Pengujian	67
4.2.6	Implementasi Sistem Berbasis Flask	74
BAB V KESIMPULAN DAN SARAN		76
5.1	Kesimpulan	76

5.2	Saran.....	76
REFERENSI		78
DAFTAR RIWAYAT HIDUP		81

DAFTAR GAMBAR

Gambar 2.1 Arsitektur Transformer	9
Gambar 2.2 Arsitektur <i>Encoder</i>	10
Gambar 2.3 Arsitektur <i>Decoder</i>	11
Gambar 2.4 Representasi Input BERT	13
Gambar 2.5 Training BERT	14
Gambar 3.1 Alur Metodologi untuk Sistem Mesin Pencarian Artikel Berita	24
Gambar 3.2 Alur Sistem BERT	31
Gambar 3.3 Representasi <i>Input Output</i>	34
Gambar 4.1 Halaman Utama Portal Berita CNN	47
Gambar 4.2 Data Artikel Berita CNN Indonesia	51
Gambar 4.3 Distribusi Jumlah Kata Data Berita	53
Gambar 4.4 Distribusi Kategori Data Berita	54
Gambar 4.5 Hasil Tokenisasi Menggunakan BertTokenizer	59
Gambar 4.6 Hasil Klasifikasi Kueri Pencarian	64
Gambar 4.7 Grafik Evaluasi Skenario 1	70
Gambar 4.8 Grafik Evaluasi Skenario 2	71
Gambar 4.9 Grafik Evaluasi Skenario 3	71
Gambar 4.10 Grafik Evaluasi Skenario 4	72
Gambar 4.11 Grafik Perbandingan Waktu Komputasi Skenario	72
Gambar 4.12 <i>Interface</i> Halaman Utama	74
Gambar 4.13 <i>Halaman Hasil Pencarian</i>	75

DAFTAR TABEL

Tabel 2.1 Penelitian Terkait.....	21
Tabel 3.1 Data Artikel Berita	25
Tabel 3.2 Data Kueri Pencarian	30
Tabel 3.3 Token Embedding.....	31
Tabel 3.4 Token Embedding.....	32
Tabel 3.5 Konversi Token ke ID Numerik	32
Tabel 3.6 Input Mask.....	33
Tabel 3.7 Positional Embedding	34
Tabel 3.8 <i>Embedding Vector</i>	35
Tabel 3.9 Segmentasi Artikel	39
Tabel 3.10 Tokenisasi SBERT.....	39
Tabel 3.11 Contoh Kueri dan Dokumen.....	41
Tabel 3.12 Frekuensi Kata.....	42
Tabel 3.13 Vektor Frekuensi Kata.....	42
Tabel 3.14 Perhitungan <i>Cosine Similarity</i>	43
Tabel 3.15 Skenario Uji Coba	46
Tabel 4.1 Jumlah Kolom dengan Nilai Kosong	52
Tabel 4.2 Data Kueri Hasil Ekstraksi.....	56
Tabel 4.3 Data <i>Keyword</i> Setelah <i>Preprocessing</i>	57
Tabel 4.4 Hasil Evaluasi Model Klasifikasi Dengan Data Tes.....	62
Tabel 4.5 Hasil Evaluasi Similarity Ranking	66
Tabel 4.6 Hasil Similarity Ranking Skenario 1	67
Tabel 4.7 Hasil Similarity Ranking Skenario 2.....	68
Tabel 4.8 Hasil Similarity Ranking Skenario 3.....	69
Tabel 4.9 Hasil Similarity Ranking Skenario 4.....	69
Tabel 4.10 Hasil Evaluasi Skenario Pengujian	72

DAFTAR KODE PROGRAM

Kode Program 4.1 <i>Scraping</i> Data	49
Kode Program 4.2 Penggabungan Semua Kategori Artikel Berita	52
Kode Program 4.3 Penghapusan Baris Data dengan Nilai Kosong	53
Kode Program 4.4 <i>Sampling</i> Data Berita Berdasarkan Distribusi Kata.....	53
Kode Program 4.5 <i>Sampling</i> Data Berita Berdasarkan Distribusi Kategori	54
Kode Program 4.6 Normalisasi Data Berita Indonesia	55
Kode Program 4.7 Ekstraksi Kata Kunci	56
Kode Program 4.8 <i>Cleaning</i> Data <i>Keyword</i>	57
Kode Program 4.9 Tokenisasi BertTokenizer	58
Kode Program 4.10 Pembagian Data <i>Keyword</i>	59
Kode Program 4.11 Training Model Klasifikasi IndoBERT	61
Kode Program 4.12 Representasi SBERT.....	63
Kode Program 4.13 Proses <i>Similarity Ranking</i>	65
Kode Program 4.14 Proses Evaluasi Similarity	67