BAB I

PENDAHULUAN

1.1 Latar Belakang

Repositori skripsi dan tesis merupakan salah satu aset penting dalam dunia pendidikan tinggi. Dokumen-dokumen ini memuat hasil penelitian mahasiswa yang tidak hanya mencerminkan pemahaman akademik, tetapi juga dapat dijadikan rujukan dan inspirasi untuk penelitian lanjutan. Dalam era digital saat ini, mayoritas dokumen tersebut disimpan dalam format digital, khususnya PDF, dan tersedia melalui sistem repositori online milik universitas. Namun, semakin banyaknya jumlah dokumen yang diunggah setiap tahun juga menimbulkan tantangan baru, yaitu bagaimana cara menelusuri dan menemukan informasi akademik yang relevan secara cepat dan akurat.

Selama ini, pencarian dokumen digital masih mengandalkan metode berbasis kata kunci. Pendekatan ini memiliki kelemahan mendasar, yaitu hanya mencocokkan kata secara literal tanpa mempertimbangkan makna atau hubungan semantik antar kata [6]. Akibatnya, query yang diberikan pengguna sering kali menghasilkan hasil pencarian yang kurang relevan. Hal ini sejalan dengan penelitian Rahman et al. (2015), yang menunjukkan bahwa sistem pencarian berbasis kata kunci sering kali tidak dapat menangkap makna sebenarnya dari sebuah teks, sehingga mengurangi efektivitas pencarian [1]. Selain itu, penelitian terbaru oleh Susanto et al. (2018) juga menyoroti pentingnya representasi semantik dalam meningkatkan akurasi pencarian informasi dalam dokumen digital [2].

Sebagai solusi, teknologi pencarian semantik menawarkan pendekatan yang lebih canggih. pencarian semantik memungkinkan sistem pencarian memahami makna query secara lebih mendalam dengan menangkap hubungan semantik antara query dan dokumen. Menurut penelitian Reimers dan Gurevych (2019), penggunaan model Sentence-BERT terbukti mampu

meningkatkan relevansi hasil pencarian dengan mengubah teks menjadi representasi embedding yang dapat diproses secara komputasi [3]. Selain itu, penelitian Amien (2023) menjelaskan bagaimana perkembangan NLP dalam Bahasa Indonesia, termasuk penerapan model transformer, mampu meningkatkan kemampuan sistem pencarian berbasis semantik di berbagai konteks [4].

Sementara itu, perkembangan NLP di Indonesia juga menunjukkan hal positif dalam mendukung pencarian semantik. Menurut Amien (2023), penerapan model transformer seperti SBERT dalam Bahasa Indonesia semakin relevan karena tersedianya dataset lokal dan pretrained model multilingual yang mendukung pemrosesan bahasa Indonesia [4]. Penelitian Wibawa dan Anggraeni (2023) juga menunjukkan bahwa integrasi NLP dalam *preprocessing* dokumen akademik, seperti segmentasi teks dan penghilangan elemen non-informasi, mampu meningkatkan hasil ekstraksi informasi pada dokumen PDF [1]. Bahkan, pada konteks non-akademik seperti analisis media sosial, penggunaan pendekatan semantik juga mulai diterapkan untuk memahami makna kompleks dalam teks seperti ditunjukkan oleh penelitian Nur Oktavia et al. (2024) yang menganalisis tweet buzzer menggunakan NLP [2].

Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk mengembangkan sistem pencarian semantik berbasis Natural Language Processing (NLP) yang dirancang untuk dokumen skripsi dalam format PDF. Sistem akan memanfaatkan model Sentence-BERT untuk mengubah kalimat dalam dokumen menjadi vektor embedding yang menangkap makna semantik. Proses pencarian dilakukan dengan membandingkan vektor embedding dari query pengguna dengan seluruh embedding dalam korpus dokumen menggunakan cosine similarity. Dengan pendekatan ini, diharapkan sistem dapat memberikan hasil pencarian yang lebih relevan dan kontekstual dibandingkan pencarian berbasis kata kunci. Sistem ini akan dibangun menggunakan bahasa pemrograman Python dan dievaluasi menggunakan

metrik evaluasi pencarian seperti Mean Reciprocal Rank (MRR), Precision, dan Recall untuk mengukur efektivitas hasil pencarian.

Dengan pendekatan ini, diharapkan sistem dapat memberikan hasil pencarian yang lebih relevan dan akurat, sehingga memudahkan pengguna, terutama mahasiswa dan peneliti, dalam menemukan dokumen akademik yang sesuai dengan kebutuhan mereka

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana efektivitas metode Cosine Similarity dalam menentukan peringkat relevansi dokumen skripsi berdasarkan representasi vektor yang dihasilkan oleh Sentence-BERT?
2. Bagaimana performa sistem pencarian semantik tersebut ketika diuji dengan berbagai variasi query (kata, kalimat, paragraf, dan abstrak)?
3. Seberapa efektif sistem pencarian semantik dalam meningkatkan relevansi hasil pencarian dibandingkan dengan pencarian berbasis kata kunci?

1.3 Tujuan Penelitian

Agar penelitian ini terfokus dan dapat dilaksanakan secara terarah, beberapa tujuan penelitian ditetapkan sebagai berikut:

1. Mengembangkan sistem pencarian semantik menggunakan SBERT untuk dokumen skripsi.
2. Mengukur efektivitas metode cosine similarity dalam pencarian berbasis semantic.
3. Menyediakan solusi pencarian dokumen skripsi yang lebih efisien dan responsif.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat, baik secara teoritis maupun praktis:

1. Secara Teoritis

   Penelitian ini memberikan kontribusi terhadap pengembangan ilmu pengetahuan di bidang Natural Language Processing (NLP), khususnya dalam hal penerapan model Sentence-BERT untuk representasi semantik dokumen berbahasa Indonesia. Selain itu, penelitian ini juga menambah referensi akademik terkait penggunaan cosine similarity sebagai metode pengukuran kedekatan semantik antar dokumen.

2. Secara Praktis

   Sistem yang dikembangkan pada penelitian ini dapat digunakan sebagai alat bantu pencarian dokumen skripsi berbasis makna, tidak hanya berdasarkan kata kunci. Hal ini dapat membantu pengguna (seperti mahasiswa atau dosen) dalam menemukan dokumen yang relevan secara lebih efektif dan efisien.

1.5 Batasan Masalah

Untuk menjaga fokus dan ruang lingkup penelitian, beberapa batasan masalah ditetapkan sebagai berikut:

1. Dataset yang digunakan berasal dari dokumen akademik skripsi dalam bahasa Indonesia.
2. Dokumen hanya berupa PDF yang dapat diambil teksnya.
3. Sistem diuji menggunakan 5 jenis variasi query: 1 kata, 2 kata, 1 kalimat, 1 paragraf, dan 1 abstrak untuk mengukur performa pencarian semantik dalam berbagai panjang input.
4. Database vektor Sentence-Bert berdimensi 384