

BAB I

PENDAHULUAN

1.1 Latar Belakang

Mesin pencari atau *search engine* merupakan sarana penting untuk mengakses informasi di *website*, tempat jutaan orang mencari untuk mendapatkan informasi yang dibutuhkan. Data statistik menunjukkan bahwa sekitar 6 miliar pencarian terjadi setiap hari [1]. Mesin pencari yang paling populer di dunia maya saat ini adalah Google, namun seiring perkembangan teknologi, semakin banyak mesin pencari baru bermunculan untuk memenuhi kebutuhan yang lebih spesifik [2].

Mesin pencari tidak hanya bertugas menampilkan hasil pencarian, tetapi juga harus mampu menilai relevansi antara kata kunci atau query yang dimasukkan pengguna dengan jutaan dokumen yang tersedia. Keakuratan hasil pencarian ini bergantung pada algoritma yang digunakan untuk menilai relevansi dokumen. *Information retrieval* adalah sistem yang secara otomatis mengambil informasi dari kumpulan informasi sesuai dengan kebutuhan pengguna. *Information retrieval* mempunyai keunggulan dalam mendapatkan informasi yang paling akurat sehingga keperluan pengguna dapat tersalurkan secara otomatis [3] [4].

BM25, atau *Best Matching 25*, adalah salah satu algoritma yang digunakan dalam *information retrieval* untuk menghitung relevansi antara query dan dokumen. Algoritma ini bekerja dengan menghitung skor relevansi berdasarkan frekuensi kata dalam dokumen serta panjang dokumen tersebut. BM25 mengadaptasi konsep dari *TF-IDF* namun dengan penambahan faktor tertentu untuk menangani *document length normalization*. Penelitian terbaru menunjukkan bahwa BM25 tetap menjadi salah satu model yang efektif dalam *information retrieval*, bahkan untuk domain khusus seperti pengambilan informasi medis [5]. Penelitian lain yang menggunakan BM25 pada Temu Kembali Informasi Berita Berbahasa Indonesia Mendapatkan hasil rata-rata presisi terbaik = 0,95 pada saat $k = 5$ [6].

Sementara itu, Cosine similarity adalah teknik yang digunakan untuk mengukur kemiripan antara dua dokumen berdasarkan arah vektornya dalam ruang

vektor. Teknik ini banyak digunakan dalam analisis teks karena kesederhanaannya dalam mengukur kemiripan antar dokumen tanpa memperhatikan panjang dokumen. Proses ini melibatkan pemetaan dokumen ke dalam bentuk vektor dengan menggunakan TF-IDF, lalu menghitung sudut antar vektor untuk menentukan tingkat kemiripan. Cosine similarity sangat berguna dalam sistem pencarian informasi dan deteksi plagiarisme [7]. Penelitian lain yang menggunakan Cosine Similarity untuk search engine hadis dengan pengujian sebanyak 50 kali mendapatkan hasil recall 87,83% dan presisi 36,25% [8].

Dalam konteks perkembangan teknologi dan kebutuhan pengguna yang semakin kompleks, sistem pencarian informasi dituntut untuk terus berkembang agar dapat memberikan hasil yang lebih relevan dan tepat waktu. Hal ini menjadi semakin penting mengingat pertumbuhan volume data yang sangat cepat, termasuk data berita, media sosial, dan konten lainnya. Dengan tantangan tersebut, penggunaan algoritma seperti BM25 dan Cosine Similarity tidak hanya menjadi solusi teknis, tetapi juga fondasi untuk pengembangan sistem pencarian informasi modern.

Selain itu, implementasi kedua algoritma ini dalam berbagai domain menunjukkan bahwa relevansi dokumen tidak hanya bergantung pada algoritma, tetapi juga pada pengelolaan data seperti preprocessing, tokenisasi, dan metode representasi teks seperti TF-IDF. Oleh karena itu, pemilihan algoritma yang tepat memerlukan evaluasi mendalam terhadap kebutuhan dan konteks penggunaannya, terutama dalam sistem yang dirancang untuk menyajikan informasi yang akurat dan relevan secara cepat.

Penelitian ini bertujuan untuk membandingkan efektivitas BM25 dan Cosine Similarity dalam menilai relevansi dokumen terhadap query dalam sistem pencarian informasi. BM25, yang dikenal dengan kemampuannya untuk menghitung relevansi berdasarkan frekuensi term dan panjang dokumen, menawarkan keunggulan dalam hal efisiensi dan kecepatan, membuatnya sangat populer dalam aplikasi pencarian besar. Di sisi lain, Cosine Similarity, yang mengukur kesamaan antara vektor dokumen dan query, memberikan keunggulan dalam menangani pencarian berbasis konteks dan semantik, meskipun memerlukan lebih banyak sumber daya komputasi.

Dengan memperbandingkan kedua metode ini, diharapkan dapat ditemukan pendekatan terbaik untuk meningkatkan kualitas dan relevansi hasil pencarian, yang sangat penting dalam mengembangkan sistem pencarian yang lebih canggih, efisien, dan dapat memenuhi kebutuhan pengguna dalam berbagai konteks pencarian informasi. Penelitian ini juga diharapkan dapat memberikan kontribusi terhadap pemahaman yang lebih dalam mengenai cara-cara efektif untuk mengoptimalkan algoritma pencarian.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Dalam sistem pencarian informasi, relevansi hasil pencarian menjadi faktor utama dalam menentukan kepuasan pengguna. Mesin pencari yang tidak dapat memberikan hasil yang relevan dengan query yang dimasukkan dapat mengurangi efektivitasnya. Dua pendekatan yang banyak digunakan dalam menilai relevansi dokumen adalah BM25 dan Cosine Similarity. Meskipun kedua metode ini sudah banyak diterapkan, masih ada ketidakpastian mengenai metode mana yang memberikan hasil yang lebih baik dalam konteks penilaian relevansi. Penelitian ini bertujuan untuk mengeksplorasi perbandingan kedua metode ini untuk menentukan mana yang lebih efektif dalam meningkatkan kualitas hasil pencarian.

1.2.2 Metode Usulan

Penelitian ini mengusulkan untuk membandingkan efektivitas BM25 dan Cosine Similarity dalam menilai relevansi dokumen pada sistem pencarian informasi. Kedua metode ini akan diterapkan pada dataset berita yang telah dikategorikan ke dalam berbagai topik. Dalam penelitian ini, BM25 akan digunakan untuk menghitung skor relevansi berdasarkan frekuensi kata dan panjang dokumen, sementara Cosine Similarity akan digunakan untuk mengukur kemiripan antara query dan dokumen dalam ruang vektor. Kedua metode ini akan dibandingkan berdasarkan tingkat akurasi dan kecepatan pemrosesan dalam menghasilkan hasil pencarian yang relevan.

1.2.3 Pertanyaan Penelitian

Bagaimana perbandingan rata-rata nilai Precision@k yang dihasilkan oleh BM25 dan Cosine Similarity terhadap dokumen yang relevan dalam sistem perolehan informasi yang menggunakan data berita politik?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk membandingkan hasil presisi dari dua metode penilaian, yaitu BM25 dan Cosine Similarity, dalam sistem pencarian informasi. Penelitian ini bertujuan untuk mengetahui metode mana yang lebih unggul dalam menilai presisi dokumen terhadap query dalam konteks dataset berita.

1.3.2 Manfaat Penelitian

Manfaat dari penelitian ini adalah memberikan wawasan yang lebih mendalam mengenai perbandingan BM25 dan Cosine Similarity dalam konteks pencarian informasi. Penelitian ini dapat memberikan kontribusi dalam memilih metode yang paling sesuai untuk aplikasi sistem pencarian, terutama dalam meningkatkan kualitas hasil pencarian berbasis teks.

1.4 Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Penelitian ini menggunakan dataset berita politik indonesia dari laman <https://www.kaggle.com/datasets/sh1zuka/indonesia-news-dataset-2024>.
2. Penelitian ini membandingkan dua metode dalam penilaian presisi dokumen, yaitu BM25 dan Cosine Similarity.
3. Penelitian ini menggunakan *TF-IDF* sebagai metode representasi teks untuk menghitung bobot kata pada dokumen dan query.
4. Metode presisi yang digunakan adalah *precision@K*.

1.5 Sistematika Laporan

Sistematika laporan skripsi merupakan tahapan laporan skripsi dari bab 1, 2, 3, 4, dan 5 yang akan dijelaskan sebagai berikut:

BAB I PENDAHULUAN

Bab ini menjelaskan latar belakang penelitian terkait pentingnya sistem perolehan informasi dalam era digital dan alasan memilih BM25 dan Cosine Similarity sebagai metode yang dibandingkan. Rumusan masalah difokuskan pada perbandingan presisi antara kedua metode tersebut. Selain itu, tujuan penelitian adalah untuk mengevaluasi efektivitas kedua metode, dengan manfaat penelitian mencakup kontribusi teoritis pada ilmu sistem informasi dan aplikasi praktis untuk pengembangan mesin pencari. Dan terdapat batasan masalah.

BAB II KAJIAN PUSTAKA

Bab ini menjelaskan desain eksperimen yang dilakukan untuk membandingkan BM25 dan Cosine Similarity. Selain itu, bab ini juga menjelaskan literatur terkait dengan penelitian ini. Diharapkan semua literatur-literatur ini dapat membantu dalam mempermudah memahami dasar-dasar teori yang digunakan didalam penelitian. Bab ini juga menjelaskan tentang penelitian-penelitian terkait dengan penelitian yang dilakukan.

BAB III METODE USULAN

Bab III memuat penjelasan mengenai tahapan-tahapan mengenai penerapan algoritma *Inverted Index*, *BM25* dan *Cosine Similarity* yang dilakukan serta *dataset* yang digunakan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan mengenai lingkungan uji coba, tahapan pembuatan program, dan hasil uji coba.

BAB V PENUTUP

Bab ini berisi kesimpulan dan saran berdasarkan penelitian yang telah dilakukan.