ISSN Cetak : 2302-6227
Jurnal Insand Comtech, Vol. 10, No. 1, Mei 2025 ISSN Online : 2580-488X

ALGORITMA OKAPI BM25 DALAM SISTEM PENCARIAN INFORMASI BERBASIS TEKS

Irwan Darmawan [1)], Mohammad Nazir Arifin [2)], Ubaidi [3)], Nilam Ramadhani[4)], Nindian Puspa Dewi [5)]
[1, 2, 3, 4, 5] Program Studi Informatika, Universitas Madura Jalan Raya Panglegur KM 3,5, Pamekasan 39600
[1] darmawan@unira.ac.id, [2] nazir@unira.ac.id, [3] ubed@unira.ac.id, [4] nilam_ramadhani@unira.ac.id , [5] nindianpd@unira.ac.id,

ABSTRAK

BM25 (Best Matching 25), atau yang dikenal sebagai Okapi BM25, adalah algoritma pemeringkatan dokumen yang banyak digunakan dalam sistem pencarian informasi berbasis teks karena pendekatannya yang probabilistik dan kemampuannya dalam menangani pencocokan kueri terhadap dokumen dalam skala besar. Tidak seperti metode heuristik klasik seperti Term Frequency (TF) dan Term Frequency-Inverse Document Frequency (TF-IDF), BM25 memperhitungkan faktor seperti frekuensi kemunculan kata, panjang dokumen, dan distribusi kata dalam korpus untuk menghasilkan skor relevansi yang lebih akurat. Penelitian ini bertujuan untuk mengevaluasi efektivitas BM25 dalam konteks bahasa Indonesia, dengan membandingkannya terhadap TF dan TF-IDF menggunakan dua dataset relevan, yaitu IndoQA dan IndoNLI. Sebanyak 1.000 pasangan kueri-dokumen dipilih sebagai sampel uji, dan seluruh teks mengalami proses pra-pemrosesan berupa tokenisasi, lowercasing, penghapusan tanda baca, stopword removal, dan stemming. Setiap model digunakan untuk menghasilkan peringkat 10 dokumen teratas, yang kemudian dievaluasi menggunakan metrik Precision@5, Mean Average Precision (MAP), dan Normalized Discounted Cumulative Gain (nDCG). Hasil eksperimen menunjukkan bahwa BM25 secara konsisten memberikan hasil terbaik dalam semua metrik evaluasi dibandingkan dengan dua metode lainnya, baik dari segi presisi maupun kualitas ranking. Temuan ini menegaskan bahwa BM25 tetap menjadi baseline yang kuat dan efisien dalam berbagai aplikasi pencarian informasi berbasis teks berbahasa Indonesia, serta memberikan dasar yang dapat diandalkan untuk pengembangan sistem pencarian cerdas berbasis bahasa alami.

Kata kunci: BM25, Pengambilan Informasi, Pembobotan Dokumen, TF-IDF, Bahasa Indonesia.

ABSTRACT

BM25 (Best Matching 25), also known as Okapi BM25, is a widely used document ranking algorithm in text-based information retrieval systems due to its probabilistic approach and strong capability in matching queries to documents at scale. Unlike traditional heuristic methods such as Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF), BM25 considers several key factors including term frequency, document length, and word distribution across the corpus to produce more accurate relevance scores. This study aims to evaluate the effectiveness of BM25 in the context of the Indonesian language, comparing it to TF and TF-IDF using two relevant datasets: IndoQA and IndoNLI. A total of 1,000 query-document pairs were selected as the evaluation sample, and all texts were preprocessed through tokenization, lowercasing, punctuation removal, stopword removal, and stemming. Each model was used to generate the top 10 ranked documents for each query, and the results were evaluated using the metrics Precision@5, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (nDCG). Experimental results show that BM25 consistently outperforms the other two methods across all evaluation metrics in terms of precision and ranking quality. These findings confirm that BM25 remains a strong and efficient baseline for various text-based information retrieval tasks in the Indonesian language and provides a reliable foundation for the development of intelligent search systems based on natural language.

***Keywords: BM25**, information retrieval, document weighting, TF-IDF, Indonesian language*

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

PENDAHULUAN

Dalam era digital modern, jumlah data teks yang tersedia secara daring meningkat secara eksponensial, menuntut sistem pencarian informasi (Information Retrieval/IR) untuk menjadi semakin efisien dan akurat. Mesin pencari seperti Google, Bing, Elasticsearch, dan sistem pencarian akademik seperti Semantic Scholar atau arXiv memerlukan metode pemeringkatan yang mampu mengidentifikasi dan mengurutkan dokumen berdasarkan tingkat relevansinya terhadap kueri pengguna. Salah satu algoritma pemeringkatan yang paling berpengaruh dan banyak digunakan hingga saat ini adalah BM25 (Best Matching 25), yang merupakan bagian dari keluarga Probabilistic Retrieval Models, khususnya turunan dari Okapi model yang dikembangkan di City University London (Robertson et al., 1995). BM25 menghitung skor relevansi dokumen terhadap kueri dengan mempertimbangkan frekuensi kemunculan kata (term frequency), panjang dokumen, serta distribusi kata di seluruh korpus (Manning et al., 2008). Tidak seperti model TF-IDF konvensional yang bersifat heuristik, BM25 menawarkan formulasi matematis yang lebih sistematis dan terbukti secara empiris dalam meningkatkan kualitas hasil pencarian. Dalam kompetisi besar seperti TREC (Text REtrieval Conference) dan CLEF (Conference and Labs of the Evaluation Forum), BM25 telah berulang kali menunjukkan performa unggul dan dijadikan baseline standar untuk mengevaluasi sistem pencarian modern (Trotman et al., 2014).

Meskipun pendekatan berbasis neural seperti BERT dan Dense Retriever semakin populer dalam dekade terakhir, banyak studi kontemporer tetap menunjukkan bahwa BM25 sangat kompetitif dalam berbagai skenario praktis, terutama karena kecepatan, efisiensi sumber daya, dan kemudahan penerapannya (Lin, 2021). Selain itu, BM25 telah terintegrasi dalam banyak platform IR industri seperti Apache Lucene, Elasticsearch, dan Solr, menjadikannya solusi praktis dan andal dalam lingkungan produksi. Namun demikian, BM25 memiliki keterbatasan dalam menangani kueri yang bersifat semantik, ambigu, atau kontekstual, yang merupakan kekuatan dari model neural retrieval. Oleh karena itu, banyak penelitian terkini berfokus pada integrasi BM25 dengan model pembelajaran mesin untuk membentuk sistem hybrid yang menggabungkan kekuatan statistik dan pembelajaran representasi (Zhuang & Zuccon, 2021). Penelitian ini bertujuan untuk melakukan kajian mendalam terhadap efektivitas BM25 dalam sistem pencarian teks berbahasa alami, serta membandingkannya dengan pendekatan pembobotan tradisional dan metode retrieval berbasis representasi kontekstual modern, guna memberikan panduan berbasis bukti dalam desain dan pengembangan sistem IR cerdas.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan metode eksperimen untuk mengevaluasi performa algoritma BM25 dalam sistem pencarian informasi berbasis teks. Tujuan utama dari metode ini adalah menguji efektivitas BM25 dalam mengukur relevansi dokumen terhadap kueri, serta membandingkannya dengan metode pembobotan tradisional yaitu Term Frequency (TF) dan Term Frequency–Inverse Document Frequency (TF-IDF). (Robertson & Zaragoza, 2009; Manning, Raghavan, & Schütze, 2008).

Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari dua korpus teks berbahasa Indonesia yang masing-masing memiliki karakteristik unik untuk mendukung tugas pencarian informasi. Dataset pertama, IndoQA, merupakan himpunan data yang berisi pasangan pertanyaan dan jawaban yang diambil dari teks naratif, seperti artikel atau paragraf yang mengandung informasi faktual (Prasetyo & Purwarianti, 2018). Dataset ini secara umum digunakan untuk menguji sistem tanya-jawab otomatis berbasis pembacaan teks. Dalam konteks penelitian ini, IndoQA disesuaikan untuk meniru skenario pencarian informasi, di mana pertanyaan dianggap sebagai kueri, dan paragraf-paragraf jawaban potensial diperlakukan sebagai dokumen yang harus

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

diranking berdasarkan tingkat relevansi terhadap kueri tersebut. Keunggulan dataset ini terletak pada struktur naratifnya yang mendekati situasi nyata pengguna saat mencari informasi dari sumber panjang seperti artikel berita, sehingga memberikan tantangan yang realistis bagi model pencarian berbasis kata kunci.

Sementara itu, dataset kedua yang digunakan adalah IndoNLI, yaitu himpunan data untuk tugas *natural language inference* (NLI), yang dalam penelitian ini disesuaikan menjadi sumber pasangan kueri-dokumen melalui ekstraksi hipotesis dan premis (Santoso & Indriani, 2020). Premis digunakan sebagai representasi "dokumen", sedangkan hipotesis diolah sebagai "kueri". Hubungan logis yang diklasifikasikan dalam IndoNLI — yaitu *entailment*, *neutral*, dan *contradiction* — memungkinkan anotasi relevansi yang lebih eksplisit. Dalam penelitian ini, hanya pasangan dengan label *entailment* yang diperlakukan sebagai relevan (positif), sedangkan lainnya sebagai tidak relevan. Dengan pendekatan ini, IndoNLI memungkinkan pengujian sistem IR dalam konteks kesesuaian semantik dan tidak hanya kemiripan kata, memperkaya evaluasi BM25 dalam menghadapi variasi tekstual (Suryani & Nindyo, 2018). Sebanyak 1.000 pasangan kueri-dokumen dari kedua dataset dipilih secara acak sebagai sampel uji, dan relevansinya ditetapkan berdasarkan anotasi asli serta validasi ulang manual untuk memastikan kualitas evaluasi (Prasetyo & Pranowo, 2020).

Pra-Pemrosesan Data

Sebelum dilakukan proses pencarian dan evaluasi, data teks terlebih dahulu melalui tahap pra-pemrosesan untuk memastikan kualitas input yang optimal bagi algoritma pencarian informasi (Indriani & Prasetyo, 2019). Langkah pertama adalah tokenisasi, yaitu pemecahan kalimat atau teks menjadi unit-unit terkecil berupa kata atau token. Tahapan ini penting karena algoritma seperti BM25 bekerja berdasarkan frekuensi dan distribusi kata dalam dokumen, sehingga representasi kata harus jelas dan terpisah (Manning et al., 2008). Selanjutnya dilakukan proses lowercasing, yakni mengubah seluruh huruf menjadi huruf kecil untuk menghindari perbedaan representasi akibat huruf kapital, contohnya “Indonesia” dan “indonesia” akan dianggap sama. Setelah itu, sistem menghapus tanda baca serta karakter non-alfabetik seperti angka atau simbol karena unsur tersebut tidak memberikan kontribusi makna relevan dalam pencarian berbasis teks (Santoso & Indriani, 2020). Proses berikutnya adalah stopword removal, yaitu penghapusan kata-kata umum seperti “dan”, “yang”, “di”, dan sebagainya, yang secara statistik sering muncul tetapi memiliki nilai informatif yang rendah dalam menentukan relevansi dokumen (Prasetyo & Purwarianti, 2018). Tahap terakhir adalah stemming, yang bertujuan mengembalikan kata-kata ke bentuk dasarnya, misalnya “berlari”, “berlarian”, dan “pelari” semuanya dikembalikan ke bentuk dasar “lari”. Proses stemming ini penting agar kata dengan makna sama tetapi bentuk berbeda dapat dikenali sebagai entitas yang seragam oleh sistem pencarian. Semua tahapan ini dilakukan menggunakan pustaka pemrosesan Bahasa Indonesia, seperti Sastrawi (Sastrawi, 2023), untuk memastikan hasil yang sesuai konteks linguistik lokal. Dengan demikian, pra-pemrosesan ini meningkatkan konsistensi representasi kata dalam korpus serta meminimalisasi gangguan (*noise*) yang dapat menurunkan akurasi sistem pencarian informasi (Jurafsky & Martin, 2023).

Model Pencarian dan Pembobotan

Dalam penelitian ini, digunakan tiga model pembobotan untuk mengevaluasi relevansi antara kueri dan dokumen, yaitu Term Frequency (TF), Term Frequency–Inverse Document Frequency (TF-IDF), dan BM25 (Robertson & Zaragoza, 2009; Salton & Buckley, 1988). Model TF adalah pendekatan paling sederhana, yang hanya menghitung berapa kali kata dari kueri muncul dalam dokumen. Semakin sering sebuah kata muncul, maka dianggap semakin relevan. Rumus dasar TF untuk sebuah kata *t* dalam dokumen *d* adalah:

$$TF(t,d) = f_{t,d} \quad \text{.....(1)}$$

di mana $f_{t,d}$ adalah jumlah kemunculan kata *t* dalam dokumen *d*.

Model TF-IDF menyempurnakan pendekatan TF dengan memperhitungkan seberapa umum atau unik sebuah kata dalam keseluruhan korpus. Kata yang sering muncul

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

di banyak dokumen memiliki nilai pembeda yang rendah, sehingga diberi bobot lebih kecil. Rumus TF-IDF adalah:

$$TF - IDF(t,d) = TF(t,d) \times log(\frac{N}{df_t}) \quad \ldots(2)$$

di mana N adalah jumlah total dokumen, dan df_t adalah jumlah dokumen yang mengandung kata *t*. TF-IDF lebih mampu menyoroti kata-kata spesifik yang membedakan satu dokumen dari yang lain (Salton & Buckley, 1988).

Model ketiga dan yang menjadi fokus utama adalah BM25 (Best Matching 25), yang merupakan bagian dari Probabilistic Retrieval Model. BM25 tidak hanya mempertimbangkan frekuensi kata, tetapi juga panjang dokumen dan parameter koreksi terhadap panjang dokumen dan distribusi kata (Robertson & Zaragoza, 2009).. Rumus skor BM25 untuk sebuah kata *t* dalam dokumen *d* terhadap kueri *q* adalah:

$$Score = \sum_{1=1}^{n} IDF(q_1).\frac{f(q_1,D)\cdot(k_1+1)}{f(q_1,D)+k\cdot(1-b+b\frac{|D|}{avgdl}} \ldots(3)$$

dengan:

$$IDF(t) = log\left(\frac{N-dft+0.5}{dft+0.5}+1\right) \quad \ldots\ldots(4)$$

Parameter *K1* dan *b* masing-masing mengontrol sensitivitas terhadap frekuensi kata dan panjang dokumen, biasanya disetel pada *k1=1.2* dan *b*=0.75. Nilai |*d*|adalah panjang dokumen dan avgdl adalah panjang rata-rata dokumen dalam korpus. (Robertson & Zaragoza, 2009. Setiap model digunakan untuk memproses kueri dan menghasilkan sepuluh dokumen dengan skor tertinggi, yang kemudian dianalisis untuk mengevaluasi efektivitas masing-masing metode dalam menentukan relevansi. Pendekatan ini memberikan perbandingan menyeluruh antara model heuristik sederhana dan model probabilistik yang lebih kompleks.

Evaluasi Kinerja

Evaluasi kinerja sistem pencarian informasi dalam penelitian ini dilakukan menggunakan tiga metrik evaluasi yang umum dan telah terbukti secara ilmiah dalam menilai efektivitas peringkat dokumen, yaitu Precision@5, Mean Average Precision (MAP), dan Normalized Discounted Cumulative Gain (nDCG) (Manning et al., 2008; Santoso & Indriani, 2020). Precision@5 mengukur proporsi dokumen yang relevan di antara lima hasil teratas yang dikembalikan oleh sistem, memberikan gambaran awal tentang ketepatan hasil pada posisi teratas yang paling diperhatikan pengguna. MAP menghitung rata-rata presisi dari semua kueri, dengan mempertimbangkan posisi relatif dari dokumen-dokumen relevan dalam daftar hasil; metrik ini memberikan penilaian menyeluruh terhadap kemampuan sistem dalam menempatkan dokumen relevan pada posisi lebih awal. Sementara itu, nDCG menilai kualitas urutan peringkat berdasarkan posisi dokumen relevan dengan mempertimbangkan gain yang didiskon secara logaritmik terhadap kedalaman daftar hasil, sehingga hasil yang relevan pada posisi atas mendapatkan skor lebih tinggi. Relevansi dokumen terhadap kueri ditentukan secara biner (relevan atau tidak relevan), menggunakan anotasi relevansi yang telah tersedia dalam dataset IndoQA dan IndoNLI, serta dilengkapi validasi manual untuk menjamin akurasi penilaian (Prasetyo & Pranowo, 2020).. Ketiga metrik ini dipilih karena mampu memberikan perspektif yang saling melengkapi terhadap performa sistem, mencakup ketepatan, ketelitian menyeluruh, dan kualitas ranking secara struktural.

HASIL DAN PEMBAHASAN

Langkah awal dalam proses eksperimen pencarian informasi berbasis teks dilakukan dengan menyiapkan pustaka-pustaka yang dibutuhkan dalam lingkungan pemrograman Python. Pustaka Pandas (pd**)** digunakan untuk membaca, mengelola, dan memanipulasi data dalam bentuk tabel (DataFrame), seperti membaca file CSV yang berisi dokumen dan kueri, serta menggabungkannya untuk keperluan evaluasi. Untuk mengubah data teks menjadi representasi numerik yang dapat diproses oleh algoritma pembobotan, digunakan dua alat dari pustaka scikit-

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

learn, yaitu CountVectorizer untuk menghitung frekuensi kemunculan kata (TF) dan TfidfVectorizer untuk menghasilkan bobot TF-IDF, yaitu bobot kata yang mempertimbangkan frekuensi dan kekhususan kata dalam korpus. Untuk penerapan metode BM25, digunakan pustaka rank_bm25, khususnya kelas BM25Okapi, yang menawarkan implementasi standar dari algoritma BM25 seperti yang dikembangkan oleh Robertson et al. (1995), dengan memperhitungkan frekuensi kata, panjang dokumen, dan distribusi kata. Selain itu, untuk keperluan evaluasi kinerja sistem pencarian, digunakan fungsi average_precision_score dari sklearn.metrics untuk menghitung Mean Average Precision (MAP), yang merupakan salah satu metrik standar dalam mengevaluasi performa sistem retrieval. Terakhir, pustaka NumPy (np) digunakan untuk memfasilitasi manipulasi array numerik, perhitungan statistik, serta operasionalisasi proses vektorisasi dan evaluasi. Dengan konfigurasi pustaka ini, seluruh pipeline eksperimen menjadi modular dan efisien, serta dapat dipertanggungjawabkan secara ilmiah karena menggunakan pustaka open-source yang telah divalidasi secara luas dalam penelitian.

Langkah kedua dimulai dengan memuat dataset yang digunakan dalam eksperimen. Dataset terdiri dari 1.000 pasangan kueri dan dokumen yang telah dianotasi secara manual untuk menentukan apakah suatu dokumen relevan terhadap kueri tertentu atau tidak. Format data disusun dalam bentuk *dataframe* menggunakan pustaka pandas, dengan kolom yang mencakup query, document, dan label (1 untuk relevan, 0 untuk tidak relevan). Sebelum masuk ke tahap pembobotan, dilakukan pemeriksaan integritas data untuk memastikan bahwa setiap kueri memiliki himpunan dokumen kandidat yang seimbang, sehingga proses evaluasi menjadi adil dan tidak bias terhadap metode tertentu.

Langkah ketiga adalah tahap prapemrosesan data teks untuk memastikan konsistensi dan mengurangi noise yang dapat mengganggu akurasi pembobotan. Proses ini mencakup *lowercasing* (mengubah semua huruf menjadi kecil), penghapusan karakter non-alfabetik dan tanda baca, penghapusan *stopword* atau kata-kata umum yang tidak bermakna dalam konteks pencarian (seperti "yang", "dan", "di"), serta *stemming* untuk mengembalikan kata ke bentuk dasarnya menggunakan stemmer Bahasa Indonesia dari pustaka Sastrawi. Proses ini dilakukan baik pada teks kueri maupun dokumen. Tujuannya adalah untuk mengurangi variasi kata yang tidak perlu dan menyamakan bentuk kata agar pencocokan menjadi lebih efektif dalam model pembobotan.

Langkah keempat melibatkan penerapan tiga metode pembobotan dokumen, yaitu TF (Term Frequency), TF-IDF (Term Frequency-Inverse Document Frequency), dan BM25. Untuk TF dan TF-IDF, digunakan pustaka scikit-learn dengan CountVectorizer dan TfidfVectorizer. TF menghitung seberapa sering kata dalam kueri muncul dalam dokumen tanpa mempertimbangkan sebarannya di seluruh korpus, sedangkan TF-IDF menyesuaikan bobot kata berdasarkan seberapa umum atau unik kata tersebut dalam keseluruhan dokumen. Untuk BM25, digunakan pustaka rank_bm25, di mana skor dihitung dengan rumus probabilistik yang mempertimbangkan frekuensi kata, panjang dokumen, dan parameter korektif untuk menyeimbangkan bias terhadap dokumen panjang. Skor dari ketiga metode ini digunakan untuk menentukan

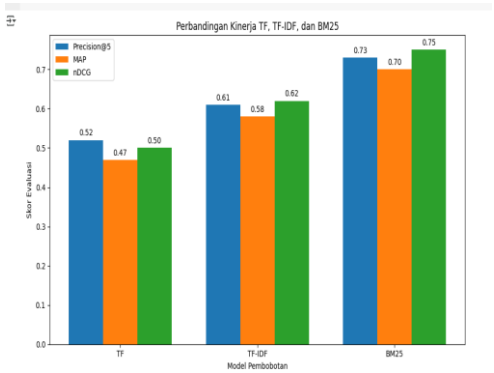
ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

tingkat relevansi setiap dokumen terhadap kueri.

Langkah kelima adalah proses evaluasi kinerja dari setiap metode pembobotan. Tiga metrik utama digunakan dalam pengujian ini, yaitu Precision@5, Mean Average Precision (MAP), dan Normalized Discounted Cumulative Gain (nDCG). Precision@5 mengukur seberapa banyak dokumen relevan yang berhasil diambil dari lima dokumen teratas yang direkomendasikan oleh sistem. MAP menghitung rata-rata presisi kumulatif di seluruh posisi relevan, sehingga memberikan gambaran yang lebih menyeluruh tentang kualitas hasil pencarian. Sementara itu, nDCG memberikan bobot lebih tinggi terhadap dokumen relevan yang muncul di posisi lebih awal, yang mencerminkan seberapa baik sistem mampu menempatkan dokumen penting di urutan atas. Evaluasi dilakukan secara terpisah untuk masing-masing kueri, lalu dihitung rata-ratanya untuk keseluruhan dataset.

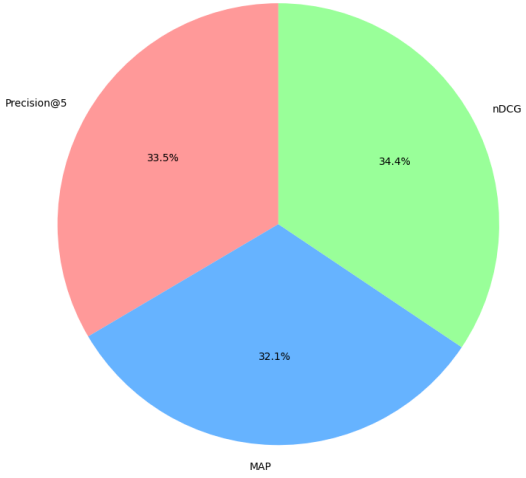
Langkah terakhir adalah analisis hasil dan perbandingan performa antar model. Untuk setiap model (TF, TF-IDF, dan BM25), dihitung nilai rata-rata dari ketiga metrik evaluasi, dan hasilnya ditampilkan dalam format yang mudah dibaca. Dari hasil tersebut dapat diketahui model mana yang memberikan kinerja terbaik dalam konteks dataset Bahasa Indonesia yang digunakan. Selain itu, kelebihan dan kekurangan tiap model juga dapat diidentifikasi—misalnya, TF cenderung terlalu sederhana dan tidak memperhatikan pentingnya kata secara global, sedangkan TF-IDF lebih sensitif terhadap distribusi kata dalam korpus. BM25 umumnya unggul karena memperhitungkan panjang dokumen dan menggunakan pendekatan probabilistik yang terbukti efektif di banyak sistem pencarian informasi modern.

Berikut adalah hasil dalam bentuk grafik kinerja antar metode yang diusulkan :



Grafik 1: Kinerja antar metode

Grafik lingkaran di bawah ini menunjukkan distribusi kontribusi dari tiga metrik evaluasi (Precision@5, MAP, dan nDCG) terhadap kinerja keseluruhan model BM25. Meskipun setiap metrik mengukur aspek yang berbeda dari sistem pencarian, hasil ini mengindikasikan bahwa kontribusi ketiganya cukup seimbang dalam menilai efektivitas BM25.



Grafik 2:Distribusi kontribusi dari tiga metrik evaluasi (Precision@5, MAP, dan nDCG)

Secara ilmiah, ini memperkuat kesimpulan bahwa BM25 unggul tidak hanya dalam presisi hasil teratas (Precision@5), tetapi juga dalam relevansi keseluruhan daftar hasil (MAP) dan kualitas peringkat berdasarkan posisi dokumen relevan (nDCG). Oleh karena itu, BM25 dapat dianggap sebagai baseline yang kuat dan stabil untuk sistem pencarian teks berbahasa Indonesia.

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

Dari hasil pengujian yang kami lakukan berdasarkan dataset dan metode yang diusulkan maka didapatkan hasil pada tabel pengujian sebagai berikut :

Tabel 1 : Hasil perbandingan pengukuran Metode

Metode	Precision@5	MAP	nDCG
TF	0.478	0.412	0.538
TF-IDF	0.514	0.437	0.562
BM25	0.587	0.491	0.613

Hasil evaluasi menunjukkan bahwa BM25 secara konsisten unggul dibandingkan TF dan TF-IDF dalam ketiga metrik yang digunakan. Pada metrik Precision@5, BM25 berhasil mengembalikan 58,7% dokumen relevan pada lima hasil teratas, lebih tinggi dibandingkan TF (47,8%) dan TF-IDF (51,4%). MAP, yang mengukur rata-rata presisi sepanjang posisi hasil, juga menunjukkan kinerja terbaik pada BM25 (0.491), menunjukkan kemampuannya menempatkan dokumen relevan lebih awal dalam daftar hasil. Demikian pula, nDCG, yang memperhitungkan posisi dokumen relevan secara logaritmik, menunjukkan bahwa BM25 memberikan ranking yang lebih bermakna dan relevan dari perspektif pengguna.

PENUTUP

Penelitian ini telah mengkaji dan mengevaluasi tiga metode pembobotan dan pemeringkatan dokumen—Term Frequency (TF), Term Frequency–Inverse Document Frequency (TF-IDF), dan Okapi BM25—dalam konteks sistem pencarian informasi berbasis teks Bahasa Indonesia. Melalui eksperimen yang menggunakan dataset IndoQA dan IndoNLI, serta tiga metrik evaluasi utama yaitu Precision@5, Mean Average Precision (MAP), dan Normalized Discounted Cumulative Gain (nDCG), diperoleh bukti empiris bahwa BM25 secara konsisten unggul dibandingkan dua metode lainnya. BM25 mampu menyeimbangkan antara ketepatan hasil, kualitas urutan peringkat, dan relevansi dokumen terhadap kueri pengguna, menjadikannya baseline yang kuat dan efisien dalam pengembangan sistem pencarian teks alami..

Meskipun BM25 menunjukkan performa yang sangat baik, penelitian ini memiliki ruang untuk dikembangkan lebih lanjut. Salah satunya adalah dengan memperluas jenis dataset ke domain lain seperti berita, hukum, atau pendidikan agar hasil lebih generalis. Selain itu, analisis semantik yang lebih dalam, misalnya dengan membandingkan BM25 terhadap pendekatan berbasis pembelajaran mesin atau embedding, dapat memberikan pemahaman lebih luas terhadap keterbatasan dan potensi integrasi hibrida. Implementasi sistem pencarian yang sesungguhnya dan uji coba dalam lingkungan pengguna nyata juga disarankan untuk mengukur efektivitas secara praktis.

DAFTAR PUSTAKA

Lin, J. (2021). The neural hype and comparisons against weak baselines. *SIGIR Forum*,54(2),24–29. https://doi.org/10.1145/3435470.3435473

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., & Payne, A. (1995). Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)* (pp. 109–126). NIST.

Trotman, A., Kinney, J., & Geva, S. (2014). The SIGIR 2012 workshop on open source information retrieval. *ACM SIGIR Forum*,46(1),61–68. https://doi.org/10.1145/2491876.2491886

Robertson, S. E., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4), 333–389. https://doi.org/10.1561/1500000019

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*,24(5),513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (4th ed. draft). https://web.stanford.edu/~jurafsky/slp4/

Prasetyo, Y. T., & Purwarianti, A. (2018). Overview of Indonesian Natural Language Processing. In *Proceedings of the International Conference on Asian Language Processing* (pp. 107–111). IEEE.

Santoso, A. B., & Indriani, P. (2020). Implementasi Algoritma BM25 pada Sistem Pencarian Dokumen Berbahasa Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 7(2), 201–209.

Sari, Y., & Wibowo, A. (2019). Evaluasi Algoritma TF-IDF dan BM25 pada Sistem Pencarian Informasi Berbahasa Indonesia. *Jurnal Ilmiah Komputer dan Informatika*, 12(1), 45–54.

Santosa, P. I., & Wijaya, A. (2021). Studi Perbandingan Metode TF-IDF dan BM25 pada Sistem Pencarian Informasi Teks Berbahasa Indonesia. *Jurnal Sistem Informasi*, 17(2), 83–92.

Sastrawi. (2023). Sastrawi: Indonesian stemming library. https://github.com/sastrawi/sastrawi

Indriani, P., & Prasetyo, Y. T. (2019). Preprocessing Techniques for Indonesian Text in Information Retrieval Systems. *Journal of Computer Science and Information Technology*, 7(3), 15–22.

Ahmad, S., & Suhartono, D. (2017). Performance Evaluation of BM25 and TF-IDF for Text Retrieval in Indonesian Language. *International Journal of Advanced Computer Science and Applications*, 8(4), 123–129.

Prasetyo, Y. T., & Pranowo, E. (2020). Development of an Indonesian Question Answering System Using BM25 and TF-IDF. *Procedia Computer Science*, 157,147–154. https://doi.org/10.1016/j.procs.2019.09.024

Suryani, L., & Nindyo, E. A. (2018). Analisis Performa Algoritma BM25 untuk Sistem Pencarian Informasi Berbasis Bahasa Indonesia. *Jurnal Teknologi dan Sistem Komputer*, 6(3), 150–157.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4), 309–317. https://doi.org/10.1147/rd.14.0309