



Implementation of Preprocessing and Synonym Expansion for Indonesian News Retrieval System

Adrian Suparto¹, Michael Joy Clement², Abdul Rahman³, Hafiz Irsyad⁴

^{1,2,4}Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang

³Teknik Elektro, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang

¹adriansuparto_2226250027@mhs.mdp.ac.id, ²michaeljoyclement_2226250041@mhs.mdp.ac.id, ³arahman@mdp.ac.id,
⁴hafizirsyad@mdp.ac.id

ABSTRACT

In Indonesian information retrieval systems, vocabulary differences between user queries and target documents are often a major obstacle in obtaining relevant search results. This research examines the effectiveness of applying synonym-based query expansion techniques to improve search relevance in IR systems. The system is designed using TF-IDF weighting and Cosine Similarity technique to calculate the closeness between query and document. A total of 10 queries were tested against a collection of news documents, with a manual approach in expanding keywords based on synonyms referred from KBBI. The evaluation was conducted using Precision@20 as the main metric. The results showed that the precision increased significantly from an average of 0.51 without query expansion to 0.725 after synonyms were added to the query. This shows that query meaning expansion can improve search accuracy in the context of a rich natural language such as Indonesian. This research indicates that the integration of semantic-based expansion techniques has great potential in optimizing the performance of IR systems. In the future, automated approaches such as semantic embedding or digital synonym mapping can be an alternative for more extensive and efficient development.

Keywords: cosine similarity, Indonesian language, information retrieval, query expansion, synonym, TF-IDF

ABSTRAK

Dalam sistem temu kembali informasi berbahasa Indonesia, perbedaan kosakata antara *query* pengguna dan dokumen target sering kali menjadi hambatan utama dalam memperoleh hasil pencarian yang relevan. Penelitian ini mengkaji efektivitas penerapan teknik *query expansion* berbasis sinonim untuk meningkatkan relevansi pencarian dalam sistem IR. Sistem dirancang menggunakan pembobotan TF-IDF serta teknik *Cosine Similarity* untuk menghitung kedekatan antara *query* dan dokumen. Sebanyak 10 *query* diuji terhadap koleksi dokumen berita, dengan pendekatan manual dalam memperluas kata kunci berdasarkan sinonim yang dirujuk dari KBBI. Evaluasi dilakukan menggunakan Precision@20 sebagai metrik utama. Hasil menunjukkan bahwa precision meningkat secara signifikan dari rata-rata 0,51 tanpa *query expansion* menjadi 0,725 setelah sinonim ditambahkan ke dalam *query*. Hal ini menunjukkan bahwa perluasan makna *query* mampu meningkatkan akurasi pencarian dalam konteks bahasa alami yang kaya variasi seperti Bahasa Indonesia. Penelitian ini mengindikasikan bahwa integrasi teknik ekspansi berbasis semantik memiliki potensi besar dalam mengoptimalkan performa sistem IR. Ke depan, pendekatan otomatis seperti *embedding* semantik atau pemetaan sinonim digital dapat menjadi alternatif untuk pengembangan yang lebih luas dan efisien.

Kata kunci: Bahasa Indonesia, cosine similarity, ekspansi kueri, Sistem Temu Kembali, sinonim, TF-IDF

1. PENDAHULUAN

Dalam era digital saat ini, perkembangan teknologi informasi telah menyebabkan ledakan data, terutama dalam bentuk teks yang tersebar luas di internet [1]. Dari media sosial, artikel berita, forum diskusi, hingga publikasi ilmiah, volume data teks terus meningkat secara eksponensial. Salah satu kontributor terbesar dari data teks tersebut adalah berita digital, yang diproduksi dan diperbarui setiap detik oleh ribuan portal berita daring di seluruh dunia, termasuk Indonesia [2]. Konten berita yang terus bertambah tidak hanya menciptakan tantangan dalam hal penyimpanan dan manajemen, tetapi juga dalam hal mengelola dan mengekstraksi informasi yang relevan dengan cepat dan akurat.

Berita digital juga memiliki karakteristik unik yang membedakannya dari jenis teks lainnya. Umumnya, berita berisi narasi kompleks, penggunaan istilah yang beragam, dan konteks yang berubah-ubah tergantung isu yang sedang hangat. Hal ini menjadikan pencarian informasi dalam kumpulan teks berita lebih menantang dibandingkan jenis teks lainnya [3]. Untuk menjawab tantangan tersebut, salah satu pendekatan yang telah digunakan secara luas adalah sistem *Information Retrieval (IR)*, yaitu proses menemukan informasi relevan dari sekumpulan dokumen berdasarkan permintaan pengguna [4]. Efektivitas sistem *IR* kerap terkendala oleh kompleksitas bahasa alami, terutama dalam Bahasa Indonesia yang memiliki struktur morfologi yang kaya serta keberagaman padanan sinonim [5]. Salah satu pendekatan yang terbukti mampu meningkatkan performa sistem *IR* adalah melalui *query expansion*, yaitu perluasan istilah pencarian dengan menambahkan kata-kata yang memiliki makna serupa [6].

Penelitian yang dilakukan oleh [7] menunjukkan bahwa penggunaan *thesaurus* dalam proses ekspansi *query* dapat meningkatkan relevansi hasil pencarian dokumen. Penelitian lain yang dilakukan oleh [6] mengimplementasikan *query expansion* menggunakan *WordNet* pada sistem *IR* Bahasa Indonesia dan berhasil meningkatkan presisi pencarian dari 52% menjadi 70%. Keberhasilan *IR* juga sangat bergantung pada tahapan *preprocessing*, seperti *stemming* dan penghapusan *stopword* juga memainkan peran penting dalam sistem *IR* [8]. Untuk Bahasa Indonesia, library *Sastrawi* telah banyak digunakan karena kemampuannya dalam mengembalikan kata berimbuhan ke bentuk dasarnya, berdasarkan algoritma *Nazief* dan *Adriani* [9]. Penggunaan *Sastrawi* dalam proses *preprocessing* dapat meningkatkan akurasi sistem *IR* dengan mengurangi variasi kata yang tidak perlu [2].

Penelitian sebelumnya telah menunjukkan bahwa metode *query expansion* dan *preprocessing* dapat meningkatkan akurasi pencarian informasi. Namun, sebagian besar metode tersebut masih bergantung pada kamus sinonim asing seperti *WordNet*, yang tidak selalu mencerminkan cara penggunaan dan makna kata dalam konteks Bahasa Indonesia. Dalam penelitian ini, digunakan pendekatan berbasis kamus sinonim manual yang disusun dari Kamus Besar Bahasa Indonesia (KBBI), yang lebih sesuai dengan struktur dan nuansa makna dalam bahasa Indonesia. Berbeda dari pendekatan sebelumnya, sistem ini dirancang untuk menangkap kekhasan bahasa Indonesia dalam konteks lokal. Dengan pendekatan ini, sistem pencarian diharapkan dapat menghasilkan keluaran yang lebih relevan secara linguistik dan kontekstual dibandingkan dengan metode yang menggunakan sumber sinonim asing.

2. METODE PENELITIAN

Pada penelitian ini peneliti melakukan beberapa tahap penelitian yaitu studi literatur, pengumpulan data, perancangan algoritma dan teknik pengujian dan analisis.

2.1. Studi Literatur

Studi literatur dalam penelitian ini dilakukan dengan menelusuri berbagai sumber digital yang relevan, seperti jurnal-jurnal ilmiah yang diakses melalui internet, artikel penelitian dari *Google Scholar*, serta publikasi dari repositori universitas. Literatur yang dikaji mencakup topik-topik seperti sistem *Information Retrieval (IR)*, teknik *query expansion*, dan tantangan pemrosesan Bahasa Indonesia.

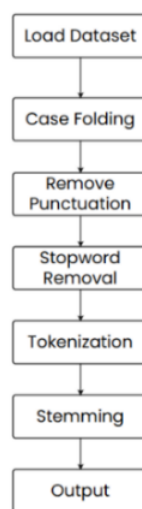
2.2. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini berupa kumpulan judul berita berbahasa Indonesia yang disusun dalam format *CSV* dengan jumlah data sebanyak 32.651 judul berita. *Dataset* ini diperoleh dari sumber publik di *Kaggle* dengan judul “*Indonesian News Dataset*” dari 7 situs berita Indonesia yaitu *Tempo*, *CNN Indonesia*, *CNBC Indonesia*, *Okezone*, *Suara*, *Kumparan*, and *JawaPos* yang dikembangkan dengan teknik *Web Scraping* [10]. Selain itu, penelitian ini menggunakan kamus sinonim yang dibuat secara manual sebagai komponen utama dalam proses *query expansion*. Kamus ini disusun berdasarkan rujukan dari Kamus Besar Bahasa Indonesia (KBBI), dengan memperhatikan kesetaraan makna antar kata dalam konteks penggunaannya di teks berita. Pemilihan pendekatan manual dilakukan untuk memastikan bahwa sinonim yang digunakan tetap relevan secara semantik dan kontekstual terhadap domain berita, sekaligus menghindari perluasan makna yang bersifat ambigu atau terlalu umum.

2.3. Perancangan Algoritma

Preprocessing merupakan tahap awal yang krusial dalam sistem *Information Retrieval (IR)* [11]. Proses ini bertujuan untuk membersihkan dan menstandarkan data teks agar menjadi representasi yang lebih sederhana dan relevan untuk diproses secara komputasional. Dengan menerapkan *preprocessing*, sistem dapat mengurangi redundansi kata, menyamakan bentuk kata, dan membuang informasi yang tidak signifikan, sehingga meningkatkan akurasi dan efisiensi dalam pengolahan selanjutnya [12].

Dalam penelitian ini, proses *preprocessing* diterapkan untuk mengolah judul-judul berita berbahasa Indonesia guna meningkatkan kualitas pencarian informasi. Tahapan dimulai dengan *Load Dataset*, yaitu membaca file *CSV* yang berisi data mentah. Selanjutnya, dilakukan *Case Folding* untuk mengubah seluruh huruf menjadi huruf kecil, memastikan konsistensi dalam analisis. Proses *Remove Punctuation* menghapus tanda baca yang tidak memberikan kontribusi signifikan dalam pencocokan teks. *Stopword Removal* kemudian menghilangkan kata-kata umum yang kurang bermakna, seperti "yang", "dan", "di", menggunakan pustaka *Sastrawi*. Setelah itu, *Tokenization* memecah kalimat menjadi token per kata, diikuti oleh *Stemming* yang mengubah kata menjadi bentuk dasarnya, contohnya "menjalankan" menjadi "jalan", juga menggunakan library *Sastrawi*. Hasil akhir dari proses ini adalah teks yang telah dibersihkan dan siap digunakan dalam tahap selanjutnya seperti pembentukan vektor *TF-IDF* dan penghitungan *cosine similarity* untuk menentukan relevansi dokumen. Tahapan proses secara lengkap dapat dilihat pada Gambar 1.

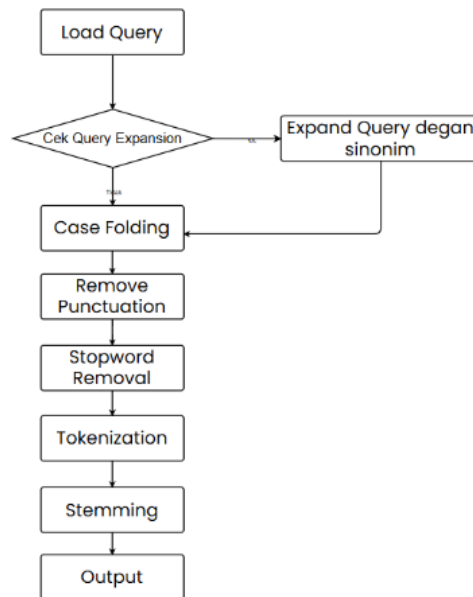


Gambar 1. *Preprocessing Dataset*

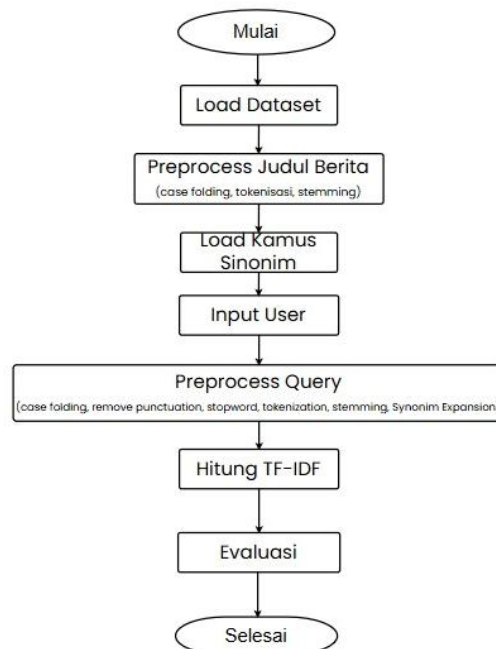
Ekspansi *query* adalah teknik dalam sistem *Information Retrieval (IR)* yang bertujuan untuk memperluas *query* pengguna dengan menambahkan kata-kata yang relevan, seperti sinonim atau istilah terkait. Langkah ini membantu dalam mengatasi masalah perbedaan kosakata antara *query* dan dokumen yang dicari [13]. Gambar 2 merupakan alur proses *preprocessing query* dalam sistem *Information Retrieval (IR)* memiliki peran penting dalam menyiapkan input pencarian agar lebih bersih, terstandar, dan siap untuk dibandingkan dengan dokumen dalam koleksi data. Tahapan ini diawali dengan memuat input dari pengguna (*query*), kemudian dilanjutkan dengan sejumlah proses pembersihan seperti *case folding*, penghapusan tanda baca, eliminasi *stopword*, *tokenisasi*, hingga *stemming*. Salah satu fitur penting yang membedakan *preprocessing query* dengan dokumen adalah adanya opsi untuk dilakukan *query expansion*, yaitu penambahan sinonim untuk memperluas jangkauan hasil pencarian. Ekspansi ini bersifat opsional dan dapat diaktifkan berdasarkan pengaturan sistem. Setelah semua proses dijalankan, hasil akhir berupa *query* yang telah diproses akan digunakan dalam tahap pencocokan dengan representasi dokumen.

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode statistik yang digunakan dalam pengolahan teks dan sistem temu kembali informasi untuk mengevaluasi seberapa penting sebuah kata (*term*) dalam suatu dokumen relatif terhadap seluruh kumpulan dokumen [14]. *Term Frequency (TF)* akan mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen. Semakin sering kata tersebut muncul, semakin tinggi nilai *TF*-nya [15]. *Inverse Document Frequency (IDF)* akan mengukur seberapa umum atau jarangny sebuah kata muncul di seluruh dokumen. Kata yang muncul di banyak dokumen memiliki nilai *IDF* yang rendah, sedangkan kata yang jarang muncul memiliki nilai *IDF* yang tinggi [15].

Cosine similarity dalam sistem pencarian informasi digunakan untuk mengukur kesamaan antara *query* pengguna dan dokumen-dokumen dalam basis data. Setiap dokumen diberi skor berdasarkan nilai *cosine similarity*-nya terhadap *query*, dan kemudian dokumen-dokumen tersebut diurutkan dari yang paling relevan hingga yang paling tidak relevan [16]. Dengan demikian, pengguna dapat melihat hasil pencarian yang paling sesuai dengan *query* mereka di posisi teratas.



Gambar 2. Preprocessing Query



Gambar 3. Perancangan Algoritma

Alur keseluruhan sistem yang dikembangkan dalam penelitian ini, yang berfokus pada penerapan *preprocessing* dan *query expansion* dalam sistem *Information Retrieval (IR)* untuk dokumen berita berbahasa Indonesia. Alur ini

menggambarkan integrasi dua jalur utama, yaitu *preprocessing* dokumen (judul berita) dan *preprocessing query* pengguna, yang keduanya diproses lebih lanjut menggunakan metode *TF-IDF* untuk menghasilkan pencocokan informasi yang relevan. Proses dimulai dengan tahapan Mulai, dilanjutkan dengan *Load Dataset* untuk memuat data berita dalam format *CSV*. Pada tahap *Preprocess* Judul Berita, dilakukan normalisasi teks melalui *case folding*, *tokenisasi*, dan *stemming* agar struktur teks menjadi seragam. Setelah itu, sistem memuat Kamus Sinonim yang akan digunakan dalam proses ekspansi *query*. Setelah pengguna melakukan *Input Query*, sistem menjalankan tahap *Preprocess Query*, yang mencakup *query expansion* berbasis kamus sinonim, *case folding*, penghapusan tanda baca, penghilangan *stopword*, *tokenisasi*, dan *stemming*. Proses ini memastikan bahwa *query* pengguna berada dalam format yang sepadan dengan dokumen yang telah diproses sebelumnya, serta diperluas secara semantik melalui penambahan sinonim. Selanjutnya, sistem melakukan perhitungan *TF-IDF* untuk mengukur bobot kemunculan istilah pada dokumen relatif terhadap keseluruhan koleksi. Akhirnya, dilakukan Evaluasi terhadap hasil pencarian menggunakan metrik performa seperti *Precision@20* untuk menilai relevansi pencocokan sebagai akhir dari seluruh proses. Tahapan proses secara lengkap dapat dilihat pada Gambar 3.

2.4. Teknik Pengujian dan Analisis

Tahapan evaluasi performa sistem *Information Retrieval (IR)* diterapkan dalam penelitian ini. Evaluasi dilakukan dengan membandingkan dua skenario sistem: Model A (tanpa *synonym expansion*) dan Model B (dengan *query expansion* berbasis sinonim). Masing-masing model diuji menggunakan 10 *query* berbeda yang dimasukkan oleh 10 responden berbeda, yang mewakili rentang usia 20-60 tahun. Setiap responden melakukan penilaian terhadap 20 dokumen teratas yang dikembalikan oleh sistem, berdasarkan nilai tertinggi dari *cosine similarity*.

Penilaian relevansi dilakukan secara manual, di mana setiap dokumen yang ditampilkan diberi label relevan atau tidak relevan oleh pengguna [17]. Hasil pelabelan ini kemudian digunakan untuk menghitung nilai *Precision@20*, yaitu rasio antara jumlah dokumen relevan yang ditemukan dalam 20 dokumen teratas terhadap total 20 dokumen yang dikembalikan. Persamaan 1 digunakan untuk mengukur seberapa tepat sistem *Information Retrieval (IR)* dalam menampilkan dokumen yang relevan pada posisi teratas hasil pencarian. Dalam hal ini, *Precision@n* menunjukkan rasio antara jumlah dokumen relevan yang berhasil ditemukan dalam *n* dokumen teratas terhadap total *n* dokumen yang ditampilkan [6].

$$\text{Precision@n} = \frac{\text{Jumlah Dokumen Relevan pada Top n}}{n} \quad (1)$$

Misalnya, jika dari 20 dokumen teratas ($n = 20$) yang dikembalikan oleh sistem, terdapat 12 dokumen yang benar-benar relevan, maka:

$$\text{Precision@20} = \frac{12}{20} = 0.6$$

Artinya, 60% dari hasil yang ditampilkan di posisi 1–20 adalah relevan. Metrik ini sangat berguna untuk mengevaluasi kualitas sistem pencarian, terutama dari perspektif pengguna yang hanya melihat beberapa hasil teratas saja. Setelah menghitung *Precision@n* untuk masing-masing *query* secara individu (menggunakan Persamaan 1), kita perlu melihat kinerja sistem secara keseluruhan. Oleh karena itu, digunakan *Average Precision@n* yang dirumuskan dalam Persamaan 2 [6]:

$$\text{Average Precision@n} = \frac{1}{n} \sum_{i=1}^n \text{Precision@20i} \quad (2)$$

dengan *N* adalah jumlah total *query* (dalam hal ini 10). Nilai rata-rata *precision* dihitung terpisah untuk masing-masing model, sehingga diperoleh dua nilai akhir yaitu *Avg. Precision@20* (dengan *synonym expansion*) untuk Model A. *Avg. Precision@20* (tanpa *synonym expansion*) untuk Model B. Evaluasi kedua nilai ini menjadi dasar untuk menilai efektivitas penggunaan *synonym expansion* dalam meningkatkan performa sistem *IR* terhadap dokumen berita berbahasa Indonesia.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dari serangkaian tahapan penelitian yang telah dilakukan, meliputi studi literatur, pengumpulan data, perancangan algoritma, serta teknik pengujian dan analisis. Fokus utama pada pembahasan ini adalah mengevaluasi kinerja sistem *Information Retrieval (IR)* berdasarkan nilai *Precision@K*, baik tanpa maupun dengan penerapan *synonym expansion*. Pengujian dilakukan untuk mengukur sejauh mana penambahan sinonim dapat meningkatkan relevansi hasil pencarian.

3.1. Evaluasi dengan *Precision@K*

Precision@K digunakan untuk menilai tingkat akurasi sistem dengan menguji relevansi *query* terhadap dokumen yang dikembalikan dalam urutan *top K*. Pengujian ini bertujuan untuk memeriksa apakah *query* yang diberikan sesuai dengan dokumen yang ditemukan dalam urutan *top K*, kemudian mengukur akurasi. Sebanyak 10 dokumen akan diuji, dengan 10 *query* yang digunakan. Daftar *query* dan *query* yang telah diperluas dengan ekspansi sinonim tercantum pada Tabel 1. Top *K* yang diuji untuk setiap *query* adalah 20. Sebagai contoh, ID Query Q-01 pada Tabel 1 menunjukkan kata kunci utama "korupsi pejabat". *Query* ini mewakili topik pencarian yang ingin mengidentifikasi dokumen-dokumen terkait isu korupsi yang melibatkan pejabat publik. Untuk meningkatkan relevansi hasil pencarian, *query* ini kemudian diperluas menggunakan *synonym expansion* menjadi: "gratifikasi, pejabat, suap, pemerasan, penyelewengan, korupsi". Ekspansi ini bertujuan agar sistem *IR* tidak hanya mencari kata "korupsi pejabat" secara literal, tetapi juga mampu mengenali istilah lain yang sering muncul dalam konteks yang sama. Dengan demikian, peluang sistem untuk mengembalikan dokumen relevan dalam top 20 hasil pencarian menjadi lebih besar.

Tabel 1. Daftar Query dan Synonym Expansion

ID Query	Query	Synonym Expansion
Q-01	korupsi pejabat	gratifikasi, pejabat, suap, pemerasan, penyelewengan, korupsi
Q-02	bencana alam	petaka, malapetaka, bencana, tragedi, alam, musibah
Q-03	demonstrasi	protes, unjuk, rasa, demo, demonstrasi, aksi, massa
Q-04	vaksin	mengebalkan, menyuntikkan, menyuntik, melindungi, merawat
Q-05	PHK	dirumahkan, pengangguran, pemberhentian, phk, pemutusan kerja
Q-06	peristiwa kebakaran	api, meluas, insiden api, peristiwa, terbakar, musibah kebakaran, kebakaran, peristiwa
Q-07	harga minyak	bahan, bakar, komoditas minyak, biaya BBM, tarif, nilai, harga, minyak, nominal, bensin, ongkos, pengeluaran
Q-08	kondisi lingkungan	lingkungan, kondisi, kerusakan alam, ekologi, pelestarian, lingkungan, alam
Q-09	fasilitas kesehatan	Fasilitas, layanan, medis, fasilitas, kesehatan, penanganan, kesehatan, pelayanan, publik
Q-10	konflik agraria	perselisihan, tanah, pertanahan, konflik, masalah, perebutan, sengketa, agraria, lahan

3.2. Pengujian Tanpa Menggunakan *Synonym Expansion* & Menggunakan *Synonym Expansion*

Pengujian dengan menggunakan *synonym expansion* menunjukkan adanya relevansi antara dokumen dan *query*. Proses *synonym expansion* dilakukan secara manual dalam 20 sesi penilaian, di mana setiap penilai memasukkan *query* dan mengevaluasi dokumen yang dianggap relevan hingga posisi ke-20. Sebagai pembandingan, pengujian tanpa *synonym expansion* juga dilakukan dengan cara yang sama, tanpa perluasan *query* secara manual terhadap 20 sesi penilaian, di mana masing-masing penilai menilai relevansi dokumen hingga urutan ke 20. Tabel 2 menampilkan hasil evaluasi sistem *Information Retrieval (IR)* menggunakan metrik *Precision@20 (P@20)*, baik dengan *synonym expansion* maupun tanpa *synonym expansion*. Nilai *P@20* menunjukkan seberapa banyak dokumen yang relevan dari 20 hasil teratas yang dikembalikan oleh sistem untuk setiap *query*.

Tabel 2. Hasil Pengujian dengan *Precision @K*

Query	Dengan Query Expansion berbasis Sinonim	Tanpa Query Expansion berbasis Sinonim
Q-01	P@20=15/20= 0.75	P@20=15/20= 0.75
Q-02	P@20=14/20= 0.70	P@20=15/20= 0.75
Q-03	P@20=18/20= 0.90	P@20=10/20= 0.50
Q-04	P@20=13/20= 0.65	P@20= 8/20 = 0.40
Q-05	P@20=18/20= 0.90	P@20=16/20= 0.80
Q-06	P@20=15/20= 0.75	P@20= 2/20 = 0.10
Q-07	P@20=19/20= 0.95	P@20=15/20= 0.75
Q-08	P@20=12/20= 0.60	P@20=10/20= 0.50
Q-09	P@20=11/20= 0.55	P@20= 9/20 = 0.45
Q-10	P@20=10/20= 0.50	P@20= 2/20 = 0.10
Rata-Rata	0.725 = 72.5%	0.51 = 51%

Query Q-03, Q-04, Q-06, Q-10 menunjukkan peningkatan sangat signifikan, bahkan hingga 40-65 poin. Misalnya, *Q-06* meningkat dari 0.10 menjadi 0.75, yang berarti sistem awal tidak mampu mengenali konteks "peristiwa

kebakaran" tanpa bantuan sinonim. Sementara itu, *Q-01* dan *Q-02* menunjukkan nilai yang sama atau sedikit menurun. Ini menandakan bahwa pada beberapa kasus, query sudah cukup representatif atau justru penambahan sinonim menyebabkan pencarian menjadi kurang fokus.

4. KESIMPULAN

Berdasarkan hasil evaluasi yang dilakukan terhadap sistem temu kembali informasi berbahasa Indonesia menggunakan 10 *query* uji dan metode pembobotan *TF-IDF* serta perhitungan kemiripan dengan *Cosine Similarity*, diperoleh bahwa penerapan *query expansion* berbasis sinonim memberikan peningkatan signifikan terhadap relevansi hasil pencarian. *Precision@20* meningkat dari rata-rata 0,51 tanpa *synonym expansion* menjadi 0,725 dengan *synonym expansion*. Artinya, probabilitas sistem dalam mengembalikan dokumen yang relevan meningkat sebesar 21,5% setelah dilakukan perluasan makna pada *query*. Temuan ini menunjukkan bahwa penggunaan perluasan kata kunci yang sesuai konteks, khususnya dalam Bahasa Indonesia, dapat meningkatkan performa sistem temu kembali secara terukur. Implikasi dari penelitian ini mengarah pada pentingnya integrasi pendekatan semantik dalam sistem *IR* untuk mendukung pencarian informasi yang lebih efektif, terutama pada dokumen berita yang memiliki variasi kosakata tinggi. Namun demikian, penggunaan ekspansi sinonim secara manual masih memiliki keterbatasan dalam hal skalabilitas dan objektivitas pemilihan sinonim, karena sangat bergantung pada subjektivitas penyusun dan cakupan kamus yang digunakan. Oleh karena itu, pada pengembangan selanjutnya disarankan untuk mengeksplorasi metode ekspansi *query* yang bersifat otomatis, seperti pemanfaatan kamus digital atau model *embedding* semantik, agar sistem lebih efisien dan dapat diterapkan pada skala data yang lebih besar.

DAFTAR PUSTAKA

- [1] F. Wu *et al.*, "MIND: A large-scale dataset for news recommendation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 3597–3606, 2020, doi: 10.18653/v1/2020.acl-main.331.
- [2] I. S. Wibowo, A. Witanti, and I. Susilawati, "Keyword Extraction Judul Berita Online Di Indonesia Menggunakan Metode TF-IDF," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 1, pp. 99–111, 2024, [Online]. Available: <http://jurnal.mdp.ac.id>
- [3] S. N. Mohanty, J. M. Chatterjee, S. Jain, and A. Elngar, "Recommender System with Machine Learning and Artificial Intelligence," *Recomm. Syst. with Mach. Learn. Artif. Intell.*, no. July, 2020, doi: 10.1002/9781119711582.
- [4] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods," *Build. Informatics, Technol. Sci.*, vol. 4, no. 1, pp. 277–284, 2022, doi: 10.47065/bits.v4i1.1670.
- [5] Sudirman, Hermansyah, and Mansyur, "Jurnal Ilmu Pendidikan," *J. Ilmu Pendidik.*, vol. 4, no. 2, pp. 318–333, 2020.
- [6] M. Dwi Laxmi and M. Ali Fauzi, "Query Expansion Pada Sistem Temu Kembali Informasi Berbahasa Indonesia Dengan Metode Pembobotan TF-IDF Dan Algoritme Cosine Similarity Berbasis Wordnet," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 1, pp. 2548–964, 2019, [Online]. Available: <http://j-ptiik.ub.ac.id>
- [7] Y. Supendi, E. Yulianto, D. A. Dewi, and K. M. S. Haryana, "Thesaurus-Based Query Expansion on Information Retrieval To Improve the Quality of Document Searching Result," *J. Theor. Appl. Inf. Technol.*, vol. 101, no. 21, pp. 6822–6833, 2023.
- [8] A. Sinaga and S. P. Nainggolan, "Analisis Perbandingan Akurasi Dan Waktu Proses Algoritma Stemming Arifin-Setiono Dan Nazief-Adriani Pada Dokumen Teks Bahasa Indonesia," *Sebatik*, vol. 27, no. 1, pp. 63–69, 2023, doi: 10.46984/sebatik.v27i1.2072.
- [9] R. Ahmad and R. R. O. Sasue, "Sistem Penilaian Esai Otomatis Menggunakan Algoritma Stemming Nazief Dan Adriani," *J. Teknol. Transp. dan Logistik Pus. Penelit. dan Pengabd. Masy. Politek. Transpostrasi Barat Bali*, vol. 1, no. 2, pp. 101–108, 2020.
- [10] I. Maulana, "Indonesian News Dataset," 2023. <https://www.kaggle.com/datasets/iqbalmaulana/indonesian-news-dataset/data> (accessed May 09, 2025).
- [11] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, no. July 2023, p. 102342, 2024, doi: 10.1016/j.is.2023.102342.
- [12] P. S. Reddy, "Natural language processing and understanding," *CARING Newsl.*, vol. 21, no. 4, p. 8, 2025, doi: 10.15662/IJAREEIE.2025.1402022.

-
- [13] K. Djoudi, Z. Alimazighi, and B. Hedjazi, "Information retrieval with query expansion and re-ranking: a survey," *IET Conf. Proc.*, pp. 114–119, 2024, doi: 10.1049/icp.2025.0071.
- [14] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (TF-IDF) Dalam Temu Kembali Informasi Pada Dokumen Teks," *SINTESIA J. Sist. dan Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 81–88, 2023.
- [15] V. M. Hersianty *et al.*, "PENERAPAN ALGORITMA TF-IDF DAN COSINE SIMILARITY," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 1619–1625, 2025.
- [16] Y. Nuri and E. Senyurek, "Research Abstracts Similarity Implementation By Using TF-IDF Algorithm," vol. 27, no. 1, pp. 4–10, 2025, doi: 10.9790/0661-2701040410.
- [17] S. Upadhyay *et al.*, *A Large-Scale Study of Relevance Assessments with Large Language Models: An Initial Look*, vol. 1, no. 1. Association for Computing Machinery, 2024. [Online]. Available: <http://arxiv.org/abs/2411.08275>