Jurnal Informatika Terpadu

https://journal.nurulfikri.ac.id/index.php/JIT
ISSN ONLINE : 2460-8998

EVALUASI PENERAPAN BM25 PADA KAMUS BAHASA MANADO-INDONESIA MENGGUNAKAN METRIK MAP DAN NDCG

Geovani Hendrik Marinu[1], Gladly Caren Rorimpandey[2], Audy Aldrin Kenap[3]

[1,2,3] Teknik Informatika, Fakultas Teknik, Universitas Negeri Manado
Tondano, Sulawesi Utara, Indonesia 95618
geovanimarinu1@gmail.com, gladlycrorimpandey@unima.ac.id, gagaken8@gmail.com

Abstract

This study evaluates the performance of the Okapi BM25 algorithm in a Manado–Indonesian digital dictionary search system and measures the effect of lexical weighting on result-ranking quality. An experimental quantitative approach was used with a corpus of 3,513 lexical entries from the Balai Bahasa (Language Center) of North Sulawesi Province. The text underwent preprocessing consisting of punctuation removal, case folding, and tokenization, and was then indexed under two schemes: pure BM25 and BM25 with lexical weighting. Evaluation was conducted on 50 single-word test queries using a graded-relevance ground truth (0–3) with Top-5 retrieval, measured by Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG). The results show that pure BM25 achieved a MAP@5 of 0.8063 and an NDCG@5 of 0.8874, while lexical weighting improved both metrics to 0.9299 (a 15.33% increase) and 0.9433 (a 6.30% increase), with the largest improvement observed on common-word queries (MAP increased by 27.0%). The study concludes that BM25 is a viable foundation for search systems for low-resource regional language dictionaries, and that effectively integrating lexical weights improves early precision. Future research should explore adaptive string matching and semantic re-ranking to handle spelling variation in the Manado language.

Keywords: *Digital Dictionary, Information Retrieval, Lexical Weighting, Manado Language, Okapi BM25*

Abstrak

Penelitian ini bertujuan mengevaluasi kinerja algoritma Okapi BM25 pada sistem pencarian kamus digital dwibahasa Manado–Indonesia serta mengukur pengaruh pemberian bobot leksikal terhadap kualitas pemeringkatan hasil. Penelitian menggunakan pendekatan eksperimental kuantitatif dengan korpus 3.513 entri leksikon yang diperoleh dari Balai Bahasa Provinsi Sulawesi Utara. Teks diproses melalui tahap pra-pemrosesan berupa penghapusan tanda baca, penyeragaman huruf (*case folding*), dan tokenisasi, kemudian di indeks dengan dua skema, yaitu BM25 murni dan BM25 dengan bobot leksikal. Evaluasi dilakukan terhadap 50 kueri uji kata tunggal menggunakan *ground truth* relevansi bertingkat (0–3) pada pengambilan Top-5, yang diukur dengan *Mean Average Precision* (MAP) dan *Normalized Discounted Cumulative* Gain (NDCG). Hasil evaluasi menunjukkan bahwa BM25 murni mencapai MAP@5 sebesar 0,8063 dan NDCG@5 sebesar 0,8874, sedangkan pemberian bobot leksikal meningkatkan kedua metrik menjadi 0,9299 (kenaikan 15,33%) dan 0,9433 (kenaikan 6,30%), dengan perbaikan terbesar pada kueri kata umum (MAP naik 27,0%). Penelitian ini menyimpulkan bahwa BM25 layak menjadi fondasi sistem pencarian kamus bahasa daerah berstatus *low-resource*, dan integrasi bobot leksikal efektif memperbaiki presisi awal. Penelitian lanjutan disarankan mengeksplorasi pencocokan *string* adaptif dan *re-ranking* semantik untuk menangani variasi ejaan bahasa Manado.

Kata kunci: Bahasa Manado, Bobot Leksikal, *Information Retrieval*, Kamus Digital Bahasa Daerah, Okapi BM25

1. PENDAHULUAN

Pelestarian bahasa daerah di era digital menghadapi tantangan teknis yang mendasar, terutama terkait ketersediaan teknologi yang mampu memproses karakteristik linguistik yang beragam. Bahasa Manado (Manado-Malay), sebagai bahasa komunikasi utama di Sulawesi Utara, memiliki variasi leksikal dan ejaan yang tinggi. Tantangan serupa, mencakup variasi morfologi yang tinggi, keterbatasan data digital (*low-resource*), serta minimnya standardisasi linguistik yang menghambat pengembangan sistem pemrosesan bahasa alami yang efektif [1].

Upaya konsolidasi sumber daya bahasa daerah Indonesia telah dilakukan melalui sejumlah inisiatif berskala besar. NusaCrowd berhasil mengonsolidasikan 137 *dataset* bahasa Indonesia dan bahasa daerahnya, namun menemukan bahwa

Diterima 26 Oktober 2026, Direvisi 14 November 2026, Diterima untuk publikasi 24 November 2026

sebagian besar sumber daya NLP yang ada masih didominasi oleh bahasa Indonesia standar, sementara bahasa-bahasa daerah seperti bahasa Manado tetap sangat terbatas representasinya [2], [3], [4]. Inisiatif NusaX membangun sumber daya paralel pertama untuk 10 bahasa lokal Indonesia berstatus *low-resource*, termasuk leksikon bilingual dan *dataset* sentimen multibahasa, namun bahasa Manado belum termasuk di antara bahasa-bahasa yang dicakup [5]. Kekosongan representasi ini menegaskan bahwa pengembangan dan evaluasi sumber daya pencarian untuk bahasa Manado masih merupakan area yang relevan untuk dikaji.

Dalam konteks kamus digital bahasa daerah, ketersediaan sistem pencarian yang efektif menjadi faktor penentu keberhasilan adopsi teknologi oleh pengguna akhir. NLP untuk bahasa sumber daya rendah merupakan bidang riset yang strategis, karena pelestarian keanekaragaman linguistik digital sangat bergantung pada tersedianya alat pencarian dan pengolahan teks yang efektif [6], [7], [8], [9]. Pada praktiknya, mayoritas kamus bahasa daerah di Indonesia masih berbentuk cetak dan belum dilengkapi dengan fungsi pencarian yang adaptif terhadap variasi ejaan penggunanya, sehingga aksesibilitas digital tetap menjadi tantangan nyata bagi komunitas penutur bahasa daerah [10].

Upaya digitalisasi kamus bahasa daerah yang sudah ada cenderung mengandalkan pencocokan *string* secara eksak [11]. Pengembangan kamus dwibahasa Sahu Tala'i–Indonesia, misalnya, menggunakan algoritma Aho-Corasick yang efektif untuk pencarian kata yang tepat, tetapi tidak menyediakan mekanisme peringkat relevansi antar entri sehingga tidak dapat menangani variasi ejaan yang umum dijumpai pada bahasa-bahasa daerah [12]. Pendekatan berbasis kecocokan karakter *literal* semacam ini sering gagal menemukan entri yang relevan ketika pengguna memasukkan kata kunci dengan variasi ejaan, dialek, atau bentuk morfologis yang berbeda dari notasi baku dalam basis data kondisi yang lazim terjadi pada bahasa Manado mengingat ketiadaan ortografi standarnya.

Information retrieval berbasis algoritma probabilitas menawarkan solusi atas keterbatasan tersebut. Algoritma BM25 memeringkat hasil pencarian berdasarkan frekuensi kemunculan istilah, tingkat kelangkaan istilah, dan normalisasi panjang dokumen, sehingga relevansi tidak lagi bergantung pada kecocokan *string* yang *exact* [13]. Karakteristik ini menjadikan BM25 sesuai untuk koleksi teks pendek seperti entri kamus digital, dan algoritma ini tidak memerlukan data latih dalam jumlah besar sehingga ideal diterapkan pada konteks bahasa daerah berstatus *low-resource*. Dalam evaluasi terhadap *dataset* IndoQA dan IndoNLI, BM25 secara konsisten mengungguli TF dan TF-IDF pada seluruh metrik evaluasi yang diuji, termasuk Precision@5, *Mean Average Precision*, dan NDCG [14]. Pada cakupan yang lebih luas, evaluasi lintas 18 bahasa menunjukkan bahwa BM25 tetap menjadi *baseline* yang tangguh, termasuk pada bahasa-bahasa Asia Tenggara bersumber daya terbatas [15]. Relevansi BM25 juga terjaga dalam arsitektur *retrieval* modern: penyuntikan skor BM25 sebagai fitur teks ke dalam *cross-encoder* BERT terbukti meningkatkan performa *re-ranking* pada koleksi berskala besar [16], dan integrasinya ke dalam kerangka *retrieval-augmented generation* terbukti mengurangi halusinasi model bahasa besar pada domain spesifik [17]. Bukti-bukti ini menunjukkan bahwa BM25 bukan sekadar metode klasik, melainkan komponen fondasi yang tetap relevan dan dapat dipasangkan dengan teknik *information retrieval* mutakhir.

Beberapa penelitian terdahulu sudah mengkaji pembangunan kamus bahasa atau penerapan BM25 pada sistem pencarian, seperti Rantung dkk yang membangun kamus bahasa Manado dengan menggunakan metode RAD dan menerapkan TF-IDF dengan *Cosine Similarity* sebagai metode pencarian kata yang kemudian di evaluasi menggunakan *precision*, *recall* dan F1-Score [18]. Penelitian oleh Wikarsa dkk membangun kamus bahasa Sahu Tala`i dengan menggunakan algoritma Aho Corasik [12]. Penelitian yang dilakukan oleh Tariq dengan menggunakan algoritma *Levenshtein Distance* sebagai algoritma pembantu dalam menangani ejaan yang salah [19] Penelitian yang mencakup perbandingan TF-IDF dan BM25 dilakukan oleh Syarif dkk sebagai sistem pencarian artikel pada Temujurnal.com. Hasil yang di dapatkan penelitian tersebut adalah BM25 secara konsisten mengungguli TF-IDF [20].

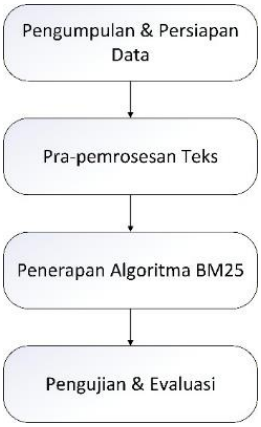
Meskipun penerapan BM25 pada konteks pencarian berbahasa Indonesia telah banyak dikaji, evaluasi yang ada secara umum masih berfokus pada korpus dokumen formal berbahasa Indonesia standar. Belum ditemukan penelitian yang secara khusus mengevaluasi performa BM25 pada korpus kamus bahasa daerah berstatus *low-resource*, terutama bahasa Manado yang memiliki karakteristik variasi ejaan dan struktur dwibahasa yang berbeda dari korpus berbahasa Indonesia standar. Kekosongan inilah yang menjadi *novelty* utama penelitian ini.

Evaluasi efektivitas sistem pencarian memerlukan metrik yang tepat [21]. BM25 berperan sebagai fungsi pemeringkatan berbasis probabilitas yang menghitung skor relevansi setiap dokumen terhadap kueri, sehingga kualitas urutan hasil yang dihasilkannya perlu diukur secara terukur. Untuk keperluan tersebut, penelitian ini menggunakan *Normalized Discounted Cumulative* Gain (NDCG) dan *Mean Average Precision* (MAP). NDCG menilai kualitas urutan hasil pencarian dengan memberikan bobot lebih besar pada dokumen relevan yang berada di posisi atas, sehingga sistem yang menempatkan dokumen paling relevan di peringkat awal memperoleh nilai lebih tinggi. MAP, di sisi lain, mengukur ketepatan sistem dengan merata-ratakan nilai *precision* pada setiap posisi ditemukannya dokumen relevan untuk satu kueri, kemudian merata-ratakannya lagi untuk seluruh kueri uji, sehingga mencerminkan konsistensi performa sistem secara

keseluruhan. Berdasarkan urgensi dan *research gap* yang ditemukan, belum ada penelitian yang menerapkan algoritma BM25 pada kamus bahasa Manado serta mengevaluasinya menggunakan metrik MAP dan NDCG. Oleh karena itu, penelitian ini berfokus pada evaluasi algoritma BM25 pada kamus bahasa Manado menggunakan kedua metrik tersebut.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental kuantitatif untuk menerapkan dan mengevaluasi algoritma BM25 pada korpus kamus digital dwibahasa Manado-Indonesia. Tahapan penelitian terdiri atas persiapan data, pra-pemrosesan teks, implementasi algoritma, serta pengujian dan evaluasi, seperti diilustrasikan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1 Metode pengumpulan data, instrumen penelitian, dan metode pengujian

Data yang digunakan diperoleh dari Balai Bahasa Provinsi Sulawesi Utara dalam format *Comma-Separated Values* (CSV), terdiri atas 3.787 entri leksikon bahasa Manado beserta padanannya dalam bahasa Indonesia. Setiap entri memiliki lima atribut: Kata sebagai leksikon utama, Ejaan sebagai notasi fonetik pengucapan, Arti sebagai padanan bahasa Indonesia, *Contoh_Manado* sebagai contoh kalimat berbahasa Manado, dan *Contoh_Indonesia* sebagai terjemahannya. Proses pemuatan data menyaring entri yang memiliki kolom Kata kosong atau hanya berisi spasi, sehingga korpus yang digunakan dalam pengindeksan dan pengujian berjumlah 3.513 entri valid, dengan selisih 274 entri yang tidak diikutsertakan karena tidak memiliki leksikon utama yang dapat di indeks.

Atribut Ejaan tidak diikutsertakan dalam proses pengindeksan karena notasi fonetiknya tidak lazim digunakan sebagai kueri oleh pengguna dan berpotensi menghasilkan skor relevansi yang tidak akurat; atribut ini tetap disimpan dan ditampilkan pada antarmuka hasil pencarian. Keempat atribut lainnya di indeks dengan bobot leksikal berbeda sesuai tingkat kepentingannya terhadap relevansi pencarian, sebagaimana dirangkum pada Tabel 1.

Tabel 1. Skema Pembobotan Atribut Korpus

Atribut	Bobot	Keterangan
Kata	+ 5,0	Leksikon utama
Arti	+2,0	Padanan Indonesia
Contoh_Manado	+1,5	Konteks penggunaan kata
Contoh_Indonesia	+0,5	Perluasan pencocokan kata
Ejaan	-	Tidak di indeks

2.2 Tahapan penelitian

Tahap pertama mencakup pra-pemrosesan teks: Sebelum dicocokkan, teks gabungan dari setiap entri kamus dan kueri pengguna diproses melalui empat langkah pra-pemrosesan yang bertujuan menyeragamkan variasi bahasa sekaligus memperbaiki ketepatan pencarian. Langkah pertama ialah *case folding*, yakni mengubah semua huruf menjadi huruf kecil, sehingga penulisan "Makan", "makan", maupun "MAKAN" dianggap sebagai satu token yang sama. Langkah kedua menangani tanda hubung dengan menggantinya menjadi spasi. Hal ini diperlukan karena bahasa Manado lazim memakai kata ulang yang disertai klitik, misalnya "*banya-banya*", yang dengan cara ini dipecah menjadi dua token tersendiri sehingga tetap dapat ditemukan oleh kueri berupa satu kata. Langkah ketiga adalah membuang tanda baca dan karakter non-alfanumerik dengan bantuan pola ekspresi reguler [^\w\s]. Langkah terakhir berupa tokenisasi, yaitu memecah seluruh kueri menjadi kata-kata yang berdiri sendiri [22].

Tahap kedua mencakup penerapan algoritma BM25: Skor relevansi BM25 untuk dokumen D terhadap kueri Q yang terdiri atas term q_1, q_2, ..., q_n dihitung menggunakan persamaan berikut [23]:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

dengan $f(q_i,D)$ adalah frekuensi kemunculan term q_i dalam dokumen D, |D| adalah panjang dokumen D dalam jumlah token, avgdl adalah rata-rata panjang seluruh dokumen dalam koleksi, k_1 adalah parameter saturasi frekuensi term, dan b adalah parameter normalisasi panjang dokumen. Nilai $\text{IDF}(q_i)$ dihitung sebagai:

$$\text{IDF}(q_i) = ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

dengan N adalah jumlah total dokumen dalam koleksi, $n(q_i)$ adalah jumlah dokumen yang mengandung term q_i, dan konstanta 0,5 berfungsi sebagai *smoothing* untuk menghindari pembagian dengan nol. Penelitian ini

menetapkan k_1 = 1,2 dan b = 0,75 sesuai nilai standar yang direkomendasikan untuk berbagai domain aplikasi [23].

Algoritma diimplementasikan menggunakan bahasa pemrograman *Python* dan dijalankan dalam lingkungan pengujian (*testing environment*) berbasis *framework Flask* yang dapat diakses melalui peramban secara lokal. Lingkungan ini dirancang khusus untuk mengeksekusi kueri uji dan mengalkulasi metrik evaluasi secara interaktif, dan bukan merupakan produk kamus digital yang siap digunakan; ruang lingkup penelitian ini terbatas pada evaluasi kinerja algoritma, bukan pada perancangan antarmuka pengguna. Korpus disimpan dan diakses dalam format CSV menggunakan pustaka *pandas*.

Inisialisasi model hanya dilakukan satu kali ketika aplikasi pertama kali dijalankan atau ketika data korpus diperbarui; seluruh permintaan pencarian selanjutnya menggunakan indeks yang telah terbangun di memori, sehingga waktu respons pencarian tetap cepat meskipun korpus berukuran ribuan entri.

Tahap ketiga mencakup pengujian dan evaluasi: Evaluasi kinerja sistem dilakukan dengan membandingkan daftar entri yang dikembalikan oleh sistem terhadap *ground truth* yang dikonstruksi untuk 50 kueri uji berupa kata tunggal. Setiap entri pada *ground truth* dianotasi menggunakan skor relevansi bertingkat (*graded relevance*) dalam empat tingkatan: 0 (tidak relevan), 1 (terkait), 2 (sinonim), dan 3 (*exact match*). Kinerja sistem dinilai berdasarkan peringkat entri yang dikembalikan dalam Top-5 *retrieval* yang mengukur kemampuan sistem dalam menempatkan entri di posisi teratas.

NDCG memanfaatkan informasi skor relevansi bertingkat serta memberikan diskon pada entri yang muncul pada posisi peringkat yang lebih bawah. Nilai NDCG merupakan hasil normalisasi DCG terhadap DCG pada pengurutan ideal (IDCG), sehingga berkisar antara 0 dan 1. Skor akhir diperoleh dengan merata-ratakan nilai NDCG seluruh kueri uji pada nilai *k* tertentu:

$$DCG@k = \Sigma(2^{r}el(i) - 1)/log_2(i+1)$$

$$NDCG@k = DCG@k/IDCG@k$$

dengan rel(i) merupakan skor relevansi entri pada posisi ke-i, dan IDCG@k adalah nilai DCG@k pada pengurutan ideal, yaitu kondisi ketika seluruh entri diurutkan menurun berdasarkan skor relevansinya. Fungsi *gain* eksponensial diterapkan agar entri *exact match* (*gain* $2^3 - 1 = 7$) dibedakan secara tegas dari entri sinonim (*gain* 3) maupun entri terkait (*gain* 1). Penelitian ini menetapkan k = 5, konsisten dengan praktik penggunaan NDCG@5 pada evaluasi sistem pencarian berbasis BM25 [24] .

MAP**,** Berbeda dengan NDCG, MAP memerlukan penilaian relevansi dalam bentuk biner, sehingga skor *ground truth* ditransformasikan terlebih dahulu: entri dengan skor 1–3 dikategorikan sebagai relevan, sedangkan entri dengan skor 0 dikategorikan sebagai tidak relevan. *Average Precision* (AP) untuk satu kueri dihitung sebagai rata-rata nilai presisi pada setiap posisi yang memuat entri relevan. Selanjutnya, MAP diperoleh dengan merata-ratakan nilai AP dari seluruh kueri uji:

$$AP(q) = (1/R) \times \Sigma P(k) \times rel_{\beta}(k)$$

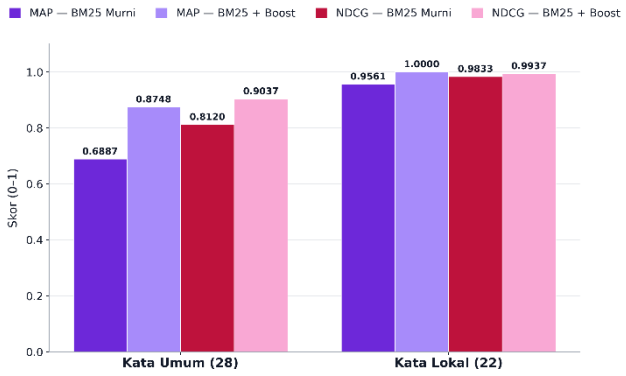
$$MAP = (1/|Q|) \times \Sigma AP(q)$$

dengan R merupakan jumlah entri relevan untuk kueri q; $P(k)$ merupakan nilai presisi pada posisi ke-k; $rel_{\beta}(k)$ bernilai 1 apabila entri pada posisi ke-k memiliki skor *ground truth* minimal 1, dan bernilai 0 untuk selainnya; serta $|Q|$ merupakan banyaknya kueri uji [25], [26].

Penggunaan kedua metrik secara bersamaan memungkinkan evaluasi yang saling melengkapi: NDCG menangkap kualitas pengurutan dengan tetap mempertimbangkan tingkat relevansi antar entri, sedangkan MAP menekankan kemampuan sistem menempatkan entri relevan pada posisi teratas secara keseluruhan.

3. HASIL DAN PEMBAHASAN

Evaluasi sistem dilakukan terhadap 50 kueri uji kata tunggal dengan pengambilan Top-5 entri, menggunakan dua skema pembobotan: (1) BM25 murni dan (2) BM25 dengan pemberian bobot leksikal. Kinerja kedua skema dibandingkan berdasarkan nilai MAP@5 dan NDCG@5 sebagaimana diuraikan pada sub-bab evaluasi. Gambar 2 menyajikan rata-rata perbandingan skor MAP dan NDCG untuk setiap kategori kata dan Tabel 2 menyajikan ringkasan hasil evaluasi kueri.



Gambar 2. Rata-rata MAP & NDCG per kategori kueri

Tabel 2. Hasil evaluasi per query (k_1=1,2; b=0,75; Top-N=5)

Kategori	BM25 Murni	BM25 Dengan Bobot Leksikal
MAP@5	0,8063	0,9299
NDCG@5	0,8874	0,9433

Berdasarkan Tabel 2, penerapan bobot leksikal meningkatkan MAP@5 dari 0,8063 menjadi 0,9299 (kenaikan relatif 15,33%) dan NDCG@5 dari 0,8874 menjadi 0,9433 (kenaikan relatif 6,30%). Peningkatan MAP yang lebih besar dibandingkan NDCG mengindikasikan bahwa bobot leksikal terutama memperbaiki *presisi awal* (*early precision*), yaitu kemampuan sistem menempatkan entri relevan khususnya entri *exact match* pada peringkat teratas. Hal ini merupakan konsekuensi langsung dari bobot +5,0 pada pencocokan eksak terhadap nilai kata entri, yang secara efektif menjamin entri kepala leksikon menempati posisi pertama sebelum skor BM25 diperhitungkan. Pada tingkat per kueri, jumlah kueri yang mencapai AP sempurna (1,0) meningkat dari 28 menjadi 39 kueri, dan kueri dengan NDCG sempurna (1,0) meningkat dari 25 menjadi 34 kueri dari total 50 kueri uji.

Pengelompokan hasil berdasarkan kategori kueri pada Gambar 2 memperlihatkan pola yang informatif. Pada kueri kata lokal, BM25 murni sebenarnya telah mencapai kinerja mendekati sempurna (MAP 0,9561; NDCG 0,9833) karena kata lokal umumnya bersifat langka sehingga memiliki nilai IDF tinggi dan tidak bersaing dengan entri lain. Penerapan bobot leksikal menyempurnakannya: seluruh 22 kueri kata lokal mencapai AP = 1,0 dan hanya satu kueri ("*vor*") yang memiliki skor NDCG di bawah 1,0. Sebaliknya, peningkatan paling substansial terjadi pada kueri kata umum, yaitu MAP naik dari 0,6887 menjadi 0,8748 (+27,0%) dan NDCG dari 0,8120 menjadi 0,9037 (+11,3%). Temuan ini menegaskan bahwa kelemahan utama BM25 murni pada leksikon dua bahasa terletak pada kueri kata umum: term berfrekuensi tinggi muncul berulang kali pada *field* arti dari banyak entri (misalnya dalam kalimat definisi), sehingga kekuatan diskriminatif IDF melemah dan entri kepala yang sesuai tergeser oleh entri lain yang hanya menyebut term tersebut di dalam definisinya. Bobot leksikal mengatasi masalah ini dengan memasukkan *prior* struktural, yakni informasi bahwa pencocokan pada *field* nilai kata secara linguistik lebih penting daripada pencocokan di dalam *field* arti.

Kueri "*spanggal*", "*fugado*", "*sambunyi*", dan "*dor*" meraih skor tertinggi pada kedua skema, dengan MAP dan NDCG sama-sama bernilai 1,0. Keempatnya adalah kata lokal yang jarang muncul dalam koleksi, sehingga nilai IDF-nya tinggi. Akibatnya, entri kepala yang tepat sudah unggul pada skor BM25 tanpa memerlukan bobot tambahan. Bobot leksikal pada kueri ini hanya mempertegas urutan yang sudah benar dan tidak mengubah hasil, sehingga skema pembobotan tambahan terbukti tidak merugikan kueri yang sudah optimal.

Sebaliknya, kueri "rusak" adalah skor terendah pada BM25 murni, dengan MAP dan NDCG sebesar 0,0000 karena seluruh entri pada Top-5 tidak relevan. Kueri "rusak" banyak tersebar pada *field* contoh kalimat entri lain, sehingga entri sinonim gagal di peringkatkan pada Top-5. Setelah bobot leksikal diterapkan, skor kueri "rusak" naik menjadi MAP 0,6667 dan NDCG 0,7003. Peningkatan serupa terlihat pada kueri "minum", dari MAP 0,0833 dan NDCG 0,0530 menjadi 0,5556 dan 0,5534, serta pada kueri "*kukis"*, dari MAP 0,2778 dan NDCG 0,8537 menjadi 1,0000 untuk keduanya. Ketiga kasus ini menunjukkan bahwa bobot +5,0 bagi kecocokan eksak dengan nilai kata entri mampu menempatkan kueri relevan di peringkat teratas, saat statistik term pada BM25 murni tidak mampu memeringkatkan kueri relevan di peringkat teratas.

Meski demikian, tidak semua kueri mengalami peningkatan skor. Kueri "panas" tetap berkinerja paling rendah setelah pembobotan, dengan MAP hanya naik dari 0,1250 ke 0,1667 dan NDCG dari 0,3558 ke 0,4131. Kueri "anak" juga meningkat, dari MAP 0,2500 dan NDCG 0,5991 menjadi 0,5000 dan 0,7611, tetapi hasilnya masih jauh dari sempurna. Pada kasus semacam ini, keterkaitan antara kueri dan entri relevan hanya muncul lewat kecocokan sebagian (*substring*) pada teks arti, yang hanya mendapat bobot +0,5. Kontribusi sekecil itu kalah oleh skor BM25 entri lain yang memuat term kueri lebih sering. Kelompok kueri ini memperlihatkan batas pendekatan pencocokan leksikal murni, yang tidak mampu menangkap keterkaitan yang tidak tercermin dalam kemiripan bentuk kata.

Urutan antar-entri relevan juga menjadi perhatian pada beberapa kueri. Pada kueri "teman", entri relevan yang terambil bertambah dari tiga menjadi lima, sehingga MAP naik dari 0,6000 ke 1,0000 dan NDCG dari 0,5802 ke 0,8404. Namun, NDCG yang belum mencapai 1,0 menandakan urutan entri menurut tingkat relevansinya belum ideal. Pola yang sama muncul pada kueri "lihat", "diam", "lambat", dan "*nimau*", yang NDCG-nya tetap 0,8340 pada kedua skema meski MAP sudah 1,0. Nilai 0,8340 sesuai dengan kondisi dua entri relevan (satu kecocokan eksak dan satu sinonim) yang posisinya tertukar. Urutan itu tidak berubah pada skema berbobot, karena entri dengan relevansi tertinggi hanya memperoleh bobot kecocokan eksak pada *field* arti (+1,5), sedangkan entri sinonim mendapat bobot kecocokan prefiks (+2,0). Ketidakselarasan ini juga menjelaskan satu-satunya penurunan kinerja, yaitu pada kueri "*vor*", dengan NDCG turun dari 1,0000 ke 0,8623 sementara MAP tetap 1,0000. Kemungkinan penyebabnya, bobot prefiks mengangkat entri yang awalannya mirip ke posisi di atas entri dengan relevansi lebih tinggi.

Secara keseluruhan, analisis kasus menunjukkan bahwa bobot leksikal paling efektif untuk kueri yang kinerjanya terganggu oleh sebaran term yang merata pada *field* arti, umumnya kata umum berfrekuensi tinggi, dan kurang diperlukan untuk kata lokal yang langka. Selain itu, hierarki bobot (+5,0; +2,0; +1,5; +0,5) belum sepenuhnya selaras dengan gradasi relevansi pada *ground truth*. Karena itu, sebagian kecil kueri masih menghasilkan urutan yang kurang optimal di antara entri relevan, walaupun presisi dan *recall* pada Top-5 sudah tercapai.

4. KESIMPULAN

Penelitian ini telah mengevaluasi kinerja algoritma BM25 pada korpus kamus digital dwibahasa Manado–Indonesia yang terdiri atas 3.513 entri valid, dengan membandingkan dua skema pembobotan, yaitu BM25 murni dan BM25 yang dipadukan dengan bobot leksikal. Evaluasi dilakukan terhadap 50 kueri uji kata tunggal menggunakan *ground truth* beranotasi relevansi bertingkat dan diukur melalui MAP@5 serta NDCG@5 pada pengambilan Top-5. Hasil evaluasi menunjukkan bahwa BM25 murni telah memberikan kinerja yang baik dengan MAP@5 sebesar 0,8063 dan NDCG@5 sebesar 0,8874. Temuan ini menegaskan bahwa BM25 layak dijadikan fondasi sistem pencarian untuk leksikon bahasa daerah berstatus *low-resource*, karena algoritma tersebut tidak memerlukan data latih dalam jumlah besar namun tetap mampu memeringkat entri secara relevan pada koleksi teks pendek.

Penerapan bobot leksikal terbukti meningkatkan kinerja secara konsisten, yaitu MAP@5 naik menjadi 0,9299 (kenaikan 15,33%) dan NDCG@5 naik menjadi 0,9433 (kenaikan 6,30%). Perbaikan paling substansial terjadi pada kueri kata umum dengan kenaikan MAP sebesar 27,0%, yang mengonfirmasi bahwa informasi struktural leksikon mampu mengompensasi melemahnya daya diskriminasi statistik *term* ketika kueri berupa kata berfrekuensi tinggi yang tersebar luas pada *field* arti. Sementara itu, kueri kata lokal mencapai kinerja mendekati sempurna (MAP 1,0000; NDCG 0,9937) karena kelangkaan term menghasilkan nilai IDF yang tinggi. Dari 50 kueri uji, jumlah kueri dengan AP sempurna meningkat dari 28 menjadi 39 kueri, dan satu kasus kegagalan total pada BM25 murni (kueri "rusak") berhasil dipulihkan setelah pemberian bobot leksikal.

Meskipun demikian, hasil penelitian juga mengidentifikasi keterbatasan pendekatan pencocokan leksikal murni. Beberapa kueri seperti "panas" dan "anak" masih menghasilkan skor rendah karena keterkaitan entri relevannya hanya terwakili oleh kecocokan sebagian pada teks arti, dan hierarki nilai bobot belum sepenuhnya sejajar dengan gradasi relevansi *ground truth* sehingga pemeringkatan antar-entri relevan pada sebagian kecil kueri belum optimal. Oleh karena itu, penelitian selanjutnya disarankan untuk memperluas jumlah dan ragam kueri uji termasuk kueri multi-kata dan kueri dengan variasi ejaan khas bahasa Manado, mengeksplorasi teknik pencocokan *string* adaptif seperti *fuzzy matching* atau *distance-based retrieval* untuk menangani ketiadaan ortografi baku, serta mengintegrasikan pendekatan semantik sebagai lapisan *re-ranking*. Selain itu, penelitian lanjutan yang akan dilakukan mencakup optimasi bobot atribut dan parameter k_1 dan b melalui eksperimen sistematis agar sesuai dengan karakteristik korpus kamus, perbandingan BM25 dengan metode pemeringkatan lain seperti TF-IDF dan BM25F.

DAFTAR PUSTAKA

[1] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, May 2022, pp. 7226–7249. doi: 10.18653/v1/2022.acl-long.500.

[2] S. Cahyawijaya *et al.*, "NusaCrowd: Open Source Initiative for Indonesian NLP Resources," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2023, pp. 13745–13818. doi: 10.18653/v1/2023.findings-acl.868.

[3] F. Iman, N. Rahmawati, and D. A. Pratama, "Bahasa Lokal dalam Ekosistem AI: Tantangan Komunikasi, Representasi, dan Deteksi Disinformasi Digital di Indonesia," *Journal of Communication, Media, and Social Transformation*, vol. 1, no. 2, pp. 66–77, Oct. 2026, doi: 10.71094/jcmst.v1i2.94.

[4] L. O. Sahidin *et al.*, "Bahasa Daerah Berbasis Literasi Digital Bagi Masyarakat Desa," *Amal Ilmiah: Jurnal Pengabdian Kepada Masyarakat*, vol. 6, no. 1, pp. 116–124, Nov. 2024, doi: 10.36709/amalilmiah.v6i1.293.

[5] G. I. Winata *et al.*, "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, May 2023, pp. 815–834. doi: 10.18653/v1/2023.eacl-main.57.

[6] P. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural language processing applications for low-resource languages," *Cambridge University Press*, vol. 31, no. 2, pp. 183–197, Mar. 2025, doi: 10.1017/nlp.2024.33.

[7] A. A. Ghirrid, R. T. K. Sari, and R. T. Aldisa, "Algoritma Natural Language Processing Untuk Aplikasi Penerjemah (Indonesia – Jawa) Menggunakan Metode Speech Processing," *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 3, pp. 746–759, Jul. 2024, doi: 10.35870/jtik.v8i3.2244.

[8] N. D. Saputro and A. Khusniyah, "Design and Evaluation of an AI-Based Indonesian–Malay Bilingual Dictionary Using Machine Learning Techniques," *Scientific Journal of Informatics*, vol. 13, no. 2, pp. 339–350, May 2026, doi: 10.15294/sji.v13i2.40232.

[9] M. B., Haryono, and Marsin, "Integration of Artificial Intelligence in the Preservation of Regional Arts and Languages: A Digital Humanities Approach in Indonesia," *ARMADA : Jurnal Penelitian Multidisiplin*, vol. 3, no. 9, pp. 303–308, Sep. 2025, doi: 10.55681/armada.v2i6.1756.

[10] D. A. Sulistyo, A. P. Wibawa, D. D. Prasetya, F. A. Ahda, I. N. G. Arya Astawa, and F. Andika Dwiyanto, "Multilingual Parallel Corpus for Indonesian Low-Resource Languages," *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 5, p. 2176, Sep. 2025, doi: 10.62527/joiv.9.5.3412.

[11] N. F. Hasan and M. S. A. Iribaram, "Digitalisasi Kamus Bahasa Daerah Papua Menggunakan Metode Rapid Application Development," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 710–720, Jul. 2022, doi: 10.30812/matrik.v21i3.1688.

[12] L. Wikarsa, T. Suwanto, and C. H. Loha, "Development Of A Bilingual Dictionary Of Sahu Tala'i – Indonesia Using The Aho-Corasick Algorithm," *Jurnal Pekommas*, vol. 9, no. 1, pp. 81–91, Jun. 2024, doi: 10.56873/jpkm.v9i1.5281.

[13] G. D. P. Maramis, M. C. P. Ranti, and K. Santa, "BM25 Algorithm for Improving Academic Website Search Accuracy," *Indonesian Journal of Innovation Studies*, vol. 27, no. 1, Nov. 2025, doi: 10.21070/ijins.v27i1.1799.

[14] I. Darmawan, M. Nazir Arifin, N. Ramadhani, and N. Puspa Dewi, "Algoritma Okapi BM25 Dalam Sistem Pencarian Informasi Berbasis Teks," *Jurnal Insand Comtech*, vol. 10, no. 1, pp. 30–37, May 2025, doi: doi.org/10.53712/jic.v10i1.2649.

[15] X. Zhang *et al.*, "MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1114–1131, Sep. 2023, doi: 10.1162/tacl_a_00595.

[16] A. Askari, A. Abolghasemi, G. Pasi, W. Kraaij, and S. Verberne, "Injecting the BM25 Score as Text Improves BERT-Based Re-rankers," *arXiv preprint arXiv:2301.09728*, 2023. doi: 10.48550/arXiv.2301.09728.

[17] Q. Chen, W. Zhou, J. Cheng, and J. Yang, "An Enhanced Retrieval Scheme for a Large Language Model with a Joint Strategy of Probabilistic Relevance and Semantic Association in the Vertical Domain," *Applied Sciences*, vol. 14, no. 24, p. 11529, Dec. 2024, doi: 10.3390/app142411529.

[18] V. P. Rantung, G. H. Marinu, and J. Y. Mambu, "RAD Development of a Manado–Indonesian Digital Dictionary Using TF–IDF and Cosine Similarity," *CogITo Smart Journal*, vol. 12, no. 1, pp. 26–42, Jun. 2026, doi: 10.31154/cogito.v12i1.1047.26-42.

[19] A. Thariq, "Aplikasi Kamus Bahasa Daerah dengan Algoritma Levenshtein Distance Berbasis Android," *TEKNIMEDIA: Teknologi Informasi dan Multimedia*, vol. 5, no. 2, pp. 152–160, Dec. 2024, doi: 10.46764/teknimedia.v5i2.204.

[20] M. Syarif *et al.*, "Analisis Perbandingan Algoritma BM25 dan TF-IDF untuk Temu Kembali Metadata Jurnal Indonesia pada Temujurnal.com," *Jurnal ICT: Information Communication & Technology*, vol. 25, no. 2, pp. 173–180, Dec. 2025, doi: 10.36054/jict-ikmi.v25i2.339.

[21] G. Agnihotram and J. Sarkar, "Evaluating Precision and Recall at Retrieval Time in Retrieval-Augmented Generation (RAG) Systems," *American Journal of Computer Science and Technology*, vol. 8, no. 4, pp. 174–180, Oct. 2025, doi: 10.11648/j.ajcst.20250804.11.

[22] A. B. Aditya, S. Samsudin, W. P. Rizki, M. Mahendra, and A. Setiawan, "Analisis Sentimen Ulasan Aplikasi Gojek Menggunakan Support Vector Machine Dan Random Forest," *Jurnal Informatika Terpadu*, vol. 11, no. 2, pp. 134–143, Nov. 2025, doi: 10.54914/jit.v11i2.1884.

[23] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends® in Information Retrieval*, vol. 4, no. 1–2, pp. 1–174, Sep. 2009, doi: 10.1561/1500000019.

[24] B. M. Mulyo, R. Purnama, R. F. Zainal, and L. Febrianti, "Pengembangan Sistem Tanya Jawab Hukum Indonesia Menggunakan Pendekatan NLP dan Model BM25," *Scan: Jurnal Teknologi Informasi dan Komunikasi*, vol. 20, no. 2, pp. 82–89, Jun. 2025, doi: 10.33005/scan.v20i1.5662.

[25] S. P. Azzahra, A. Azzahra, Y. A. Apriyanto, and A. Wijaya, "Analisis Ulasan Produk Amazon Menggunakan Random Forest Sentimen dan Probabilistic Retrieval Model," *Journal Of Informatics And Busisnes*, vol. 2, no. 4, pp. 519–528, Jan. 2025, doi: 10.47233/jibs.v2i4.2070.

[26] M. Luqman, K. Ardaffa Fahrezi, and A. Setiawan, "Analisis Perbandingan Kinerja Sistem Inventori dengan Algoritma Pendeteksi Barang Berbasis YOLOv11 dan YOLOv8," *PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer*, vol. 13, no. 3, pp. 55–62, Nov. 2026, doi: 10.30656/prosisko.v13i3.12220.