E-ISSN : 2988-1986
https://ejournal.cahayailmubangsa.institute/index.php/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 11 No 1 Tahun 2026

ANALISIS PERFORMA SPIMI PADA SISTEM TEMU KEMBALI INFORMASI UNTUK TEKS INFORMAL MEDIA SOSIAL BERBAHASA INDONESIA

Ayu Lutfiah [1], Siti Samriah [2], Valerina Azzahra [3] , Adha Mashur Sajiah[4]

[1,2,3,4] Department of Informatics Engineering, Faculty of Engineering, Halu Oleo University, Kendari city, Southeast Sulawesi, Indonesia

ayulutfiah.001@gmail.com, samriahsiti@gmail.com, valerinaazzahra22@gmail.com, adha.m.sajiah2@uho.ac.id

ABSTRAK

Penelitian ini menganalisis pengaruh karakteristik linguistik teks informal media sosial (Twitter/X) dan teks formal berita berbahasa Indonesia terhadap performa algoritma *Single-Pass In-Memory Indexing* (SPIMI) dan kualitas retrieval. Sebanyak 8.802 dokumen seimbang dari dua korpus dieksperimenkan dengan variasi ukuran buffer 64 MB, 128 MB, dan 256 MB menggunakan desain faktorial penuh. Analisis statistik menggunakan uji *Kruskal-Wallis* dan *Mann-Whitney U* menunjukkan bahwa konsumsi memori puncak teks informal secara signifikan lebih tinggi (2,62 MB vs. 2,46 MB, $p < 0,0001$), sementara waktu build tidak berbeda signifikan ($p = 0,4426$). Analisis regresi linier berganda mengidentifikasi kepadatan slang (*doc_len*) dan kepadatan slang (*slang_density*) dan rasio type-token sebagai prediktor dominan terhadap metrik performa SPIMI. Evaluasi retrieval pada semua konfigurasi buffer menunjukkan bahwa ukuran buffer tidak memengaruhi kualitas retrieval; korpus formal unggul pada MAP (0,135 vs. 0,071) dan NDCG@10 (0,377 vs. 0,311) dengan latensi pencarian yang lebih rendah. Temuan ini menegaskan perlunya konfigurasi buffer adaptif dan penguatan *pipeline* prapemrosesan dalam pengembangan sistem temu kembali informasi berbahasa Indonesia.

Kata Kunci: SPIMI, Teks Informal Bahasa Indonesia, Indeks Terbalik

ABSTRACT

This study analyzes the influence of linguistic characteristics of informal social media text (Twitter) and formal Indonesian news text on the performance of the Single-Pass In-Memory Indexing (SPIMI) algorithm and retrieval quality. A balanced dataset of 8,802 documents from two corpora was experimented with three buffer sizes (64 MB, 128 MB, 256

Article History
Received: Juni 2026
Reviewed:Juni 2026
Published: Juni 2026

Plagiarism Checker No 237.u873247.fh8374

Copyright : Author
Publish by : Kohesi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

MB) using a full factorial design. Statistical analyses via Mann-Whitney U tests revealed that peak memory consumption of informal text is significantly higher (2.62 MB vs. 2.46 MB, p < 0.0001), while build time showed no significant difference (p = 0.4426). Multiple linear regression identified slang density (slang_density) and type-token ratio (ttr) as dominant predictors of SPIMI performance metrics. Retrieval evaluation across all buffer sizes showed that buffer size does not affect retrieval quality; the formal corpus outperformed in Mean Average Precision (MAP: 0.135 vs. 0.071) and NDCG@10 (0.377 vs. 0.311) with lower retrieval latency. These findings underscore the necessity of adaptive buffer configuration and enhanced preprocessing pipelines in developing Indonesian-language information retrieval systems.

***Keywords**: SPIMI, Indonesian Informal Text, Inverted Index*

1. PENDAHULUAN

Pertumbuhan media sosial di Indonesia dalam satu dekade terakhir telah berlangsung dengan laju yang luar biasa. Berdasarkan laporan *We Are Social* dan *Hootsuite*, Indonesia tercatat sebagai salah satu negara dengan pengguna media sosial aktif terbesar di dunia, dengan lebih dari 167 juta pengguna aktif per tahun 2023 [1]. Platform seperti Twitter/X, Instagram, TikTok, dan berbagai forum daring seperti Kaskus menjadi ruang ekspresi utama masyarakat Indonesia, menghasilkan jutaan unggahan teks setiap harinya. Fenomena ini menciptakan volume data teks yang sangat masif dan terus bertumbuh secara eksponensial, sehingga menuntut kehadiran sistem temu kembali informasi (TKI) yang andal, efisien, dan mampu memproses data dalam skala besar.

Namun demikian, teks yang dihasilkan oleh pengguna media sosial Indonesia memiliki karakteristik linguistik yang sangat berbeda dari teks formal. Teks media sosial berbahasa Indonesia umumnya dipenuhi oleh penggunaan bahasa informal, singkatan, slang, campur kode antara bahasa Indonesia dan bahasa Inggris atau bahasa daerah, serta penulisan yang tidak mengikuti kaidah ejaan yang disempurnakan (EYD). Contoh nyata dari fenomena ini antara lain penggunaan kata seperti "gak" menggantikan "tidak", "yg" untuk "yang", hingga ekspresi campur kode seperti "update dong bro". Kondisi ini menyebabkan kosakata yang harus ditangani oleh sistem TKI menjadi sangat besar dan tidak terprediksi, karena setiap pengguna dapat menciptakan variasi bentuk kata yang berbeda-beda [2].

Sistem temu kembali informasi merupakan infrastruktur fundamental yang memungkinkan pengguna menemukan dokumen yang relevan dari sekumpulan dokumen besar berdasarkan kueri yang diberikan. Salah satu komponen terpenting dalam sistem TKI adalah proses index construction, yaitu pembangunan struktur data indeks terbalik (inverted index) yang memetakan setiap kata ke daftar dokumen yang mengandung kata tersebut. Kualitas dan efisiensi proses pembangunan indeks ini secara langsung menentukan performa keseluruhan sistem TKI, baik dari sisi kecepatan pemrosesan, konsumsi memori, maupun ukuran penyimpanan indeks yang dihasilkan [3].

Di antara berbagai algoritma pembangunan indeks yang telah dikembangkan, *Single-Pass In-Memory Indexing* (SPIMI) menjadi salah satu pendekatan yang paling banyak dikaji dalam literatur sistem TKI. SPIMI bekerja dengan cara membaca aliran token secara sekuensial dan langsung menyimpan pasangan (*term*, docID) ke dalam struktur kamus di memori utama tanpa melakukan pengurutan eksternal. Ketika memori penuh, SPIMI akan menuliskan blok sementara ke disk, kemudian menggabungkan seluruh blok di akhir proses menjadi indeks final. Pendekatan ini dikenal lebih efisien dibandingkan metode berbasis pengurutan seperti *Blocked Sort-Based Indexing* (BSBI) untuk korpus berukuran sedang yang dapat ditampung di memori utama [3]. Namun demikian, efisiensi SPIMI sangat bergantung pada karakteristik korpus yang diproses, khususnya ukuran kosakata dan tingkat keberagaman term dalam dokumen.

Sebagian besar penelitian yang mengkaji sistem TKI dan proses pembangunan indeks selama ini menggunakan korpus teks formal atau teks berbahasa Inggris sebagai bahan pengujian. Penelitian tentang performa sistem TKI dalam konteks teks informal berbahasa Indonesia masih sangat terbatas. Padahal, teks informal media sosial berbahasa Indonesia memiliki fenomena linguistik yang unik dan kompleks, di antaranya adalah tingginya tingkat *out-of-vocabulary* (OOV), ledakan kosakata akibat variasi ejaan dan singkatan, serta campur kode yang tidak konsisten. Fenomena-fenomena ini berpotensi memberikan dampak signifikan terhadap performa SPIMI, terutama dalam hal penggunaan memori, waktu pembangunan indeks, dan ukuran indeks yang dihasilkan [4].

Beberapa penelitian terdahulu di Indonesia telah menyentuh permasalahan teks informal dalam konteks pengolahan bahasa alami. Prasetyo dan Purwarianti (2015) [5] mengkaji tantangan normalisasi teks informal Twitter berbahasa Indonesia dan menemukan bahwa tingkat kosakata di luar kamus standar pada teks Twitter dapat mencapai lebih dari 40% dari total token, jauh lebih tinggi dibandingkan teks berita formal yang hanya berkisar 3–5%. Temuan ini mengimplikasikan bahwa sistem TKI yang dirancang untuk teks formal tidak dapat diterapkan

secara langsung pada teks informal tanpa modifikasi yang signifikan pada lapisan preprocessing maupun strategi indexing-nya.

Mengingat besarnya volume data teks informal media sosial berbahasa Indonesia dan terbatasnya kajian mendalam tentang performa algoritma SPIMI pada jenis teks tersebut, terdapat kebutuhan penelitian yang nyata untuk menganalisis secara sistematis bagaimana karakteristik linguistik teks informal memengaruhi efisiensi SPIMI dalam proses pembangunan indeks. Penelitian ini dirancang untuk mengisi kesenjangan tersebut dengan mengukur dan menganalisis performa SPIMI, mencakup waktu pembangunan indeks, konsumsi memori puncak, biaya I/O, dan ukuran indeks pada tiga lapisan register teks: informal tinggi, semi-formal, dan formal sebagai pembanding.

Hasil penelitian ini diharapkan dapat memberikan kontribusi ilmiah berupa profil performa SPIMI pada kondisi teks informal berbahasa Indonesia yang belum pernah dilaporkan sebelumnya dalam literatur, sekaligus menghasilkan panduan konfigurasi SPIMI yang optimal untuk pengembang sistem TKI yang bekerja dengan data media sosial berbahasa Indonesia. Selain itu, penelitian ini akan menyediakan dataset *benchmark* berlapis yang dapat dimanfaatkan oleh peneliti lain dalam kajian serupa di masa mendatang [6].

2. KAJIAN PUSTAKA

2.1 Sistem Temu Kembali Infromasi dan Inverted Index

Sistem Temu Kembali Informasi (TKI) adalah infrastrustur fundamental yang memungkinkan pengguna menemukan dokumen yang relevan dari suatu koleksi besar berdasarkan queri yang diberikan. Bagian inti dalam system TKI modern disebut *inverted index* yang berupa struktur data yang memetakan setiap term ke dalam daftar dokumen yang mengandung *term* tersebut (*posting list*). Efisiensi pembangunan dan pemanggilan *inverted index* secara langsung menentukan performa keseluruhan sistem TKI, baik dari sisi kecepatan, konsumsi memori, maupun ukuran penyimpanan.

Büttcher, Clarke, dan Cormack mendefinisikan *inverted index* sebagai sebuah struktur yang terdiri dari dua komponen utama yaitu kamus term yang menyimpan seluruh term unik dalam koleksi dan posting list yang berisi daftar docID beserta frekuensinya untuk setiap *term* [7]. Pendekatan pembobotan term menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency)* merupakan teknik standar yang digunakan untuk menilai bobot terkait seberapa penting sebuah term dalam konteks dokumen dan koleksi secara keseluruhan. Pada penelitian yang dilakukan oleh Septiani dan Isabela mengonfirmasi bahwa pembobotan TF-IDF pada teks

berbahasa Indonesia menghasilkan peningkatan akurasi retrieval yang signifikan dibandingkan pendekatan *boolean* sederhana [8].

2.2 Algoritma SPIMI (*Single-Pass In-Memory Indexing*)

SPIMI merupakan sebuah algoritma pembangunan indeks yang dirancang untuk menangani koleksi dokumen sangat besar yang tidak dapat dimuat seluruhnya di memori utama. Berbeda dari metode *Blocked Sort-Based Indexing* (BSBI) yang memerlukan pemetan term ke termID, SPIMI bekerja langsung dengan term string, sehingga menghilangkan kebutuhan struktur data pemetaan eksternal. Algoritma ini beroperasi dalam tiga fase pertama build yang merupakan proses membangun kamus term dan *posting list* di memori, kedua fase flush yang berupa proses menulis blok sementara ke disk saat memori penuh, dan fase terakhir adalah merge yang berupa penggabungan seluruh blok menjadi satu *inverted index* final.

Keunggulan utama SPIMI dibandingkan BSBI yaitu tidak diperlukannya pengurutan token eksternal sebelum flush, karena pengurutan dilakukan per blok secara terpisah. Alasan tersebutlah yang menjadikan SPIMI lebih efisien dalam hal waktu CPU, meskipun memerlukan manajemen memori yang cermat, terutama saat menangani corpus dengan *vocabulary* yang sangat besar [7]. Pada konteks teks informal media sosial berbahasa Indonesia, ukuran *vocabulary* dapat meningkat secara dramatis akibat variasi ejaan, penggunaan slang, dan fenomena campur kode, sehingga memberikan tekanan ekstra pada fase *flush* dan konsumsi memori total SPIMI.

2.3 Karakteristik Teks Informal Berbahasa Indonesia

Teks media sosial berbahasa Indonesia menampilkan karakteristik linguistik yang sangat berbeda dari teks formal. Dalam sebuah penelitian terkait normalisasi slang berbasis kamus, menemukan bahwa tingkat *out-of-vocabulary* (OOV) pada tweet berbahasa Indonesia dapat mencapai lebih dari 40% dari total token, jauh melampaui teks berita formal yang hanya berkisar 3–5%[9]. Fenomena ini disebabkan oleh beberapa faktor, diantaranya singkatan ('gak', 'udah', 'bgt'), variasi ejaan fonetis, afiksasi non-standar, serta penggunaan ungkapan ekspresif seperti 'wkwk' atau 'hehe'.

Selain tingginya OOV *rate*, teks Twitter berbahasa Indonesia juga ditandai oleh fenomena *code-mixing* yang intensif antara bahasa Indonesia dan bahasa Inggris, serta sesekali dengan bahasa daerah. Dalam sebuah jurnal yang diterbitkan pada tahun 2024 menyatakan bahwa kata-kata slang dalam teks Twitter berbahasa Indonesia tidak hanya mempersulit proses identifikasi semantik, tetapi juga mengakibatkan ledakan kosakata yang tidak terprediksi dalam sistem NLP[10]. Hal ini dapat berdampak secara langsung pada ukuran *vocabulary* dan kinerja sistem

indeksasi berbasis *inverted index*.

2.4 Prapemrosesan Teks Bahasa Indonesia untuk NLP

Prapemrosesan teks merupakan tahapan krusial dalam *pipeline* NLP berbahasa Indonesia. Tahapan standar dalam proses ini meliputi pembersihan teks, tokenisasi, normalisasi slang, penghapusan *stopword*, dan *stemming*. Untuk normalisasi slang, pendekatan berbasis kamus (*dictionary-based*) terbukti efektif untuk teks pendek seperti tweet [11]. Sementara itu, untuk stemming, pustaka Sastrawi yang mengimplementasikan algoritma *Enhanced Confix Stripping* (ECS) telah menjadi standar de facto untuk teks berbahasa Indonesia, dengan presisi mencapai 70,3% [12].

Dalam sebuah studi yang dilakukan oleh Tala [4] mengenai stemming bahasa Indonesia menunjukkan bahwa proses stemming dapat meningkatkan recall sistem TKI secara konsisten hingga 27,4%, meskipun terkadang dapat mengorbankan presisi akibat over-stemming. Temuan ini membuktikan pentingnya kalibrasi *pipeline* prapemrosesan, khususnya pada teks informal yang memiliki distribusi morfologis tidak terprediksi. Rosid et al.[13] dalam jurnalnya ikut serta memperkuat temuan ini dengan menunjukkan bahwa penggunaan Sastrawi dalam preprocessing dokumen teks meningkatkan akurasi klasifikasi sebesar 4,7% dibandingkan tanpa stemming.

2.5 Metrik Evaluasi Sistem Temu Kembali Informasi

Evaluasi kualitas sistem TKI menerapkan metrik standar berupa pengukuran akurasi dan kualitas peringkat dokumen hasil retrieval. Baeza-Yates dan Ribeiro-Neto [14] mendefinisikan tiga metrik utama yaitu pertaama berupa *Precision@K* yang merupakan proporsi dokumen relevan di antara K dokumen teratas, kedua *Mean Average Precision (MAP)* yang berupa rata-rata presisi pada setiap posisi di mana dokumen relevan ditemukan, memberikan gambaran menyeluruh tentang kualitas ranking, dan ketiga adalah *Normalized Discounted Cumulative Gain (NDCG@K)* yang mengukur kualitas ranking dengan memberikan bobot lebih pada dokumen relevan yang muncul di posisi lebih tinggi.

Septiani dan Isabela.[8] dalam penelitiannya menegaskan bahwa MAP dan NDCG merupakan metrik yang paling komprehensif untuk mengevaluasi sistem TKI berbahasa Indonesia, terutama saat berhadapan dengan *query* yang mengandung term informal. Kedua metrik tersebut terbukti mampu menangkap perbedaan kualitas retrieval yang tidak dapat dideteksi oleh metrik precision sederhana.

2.6 Analisis Statistik dalam Penelitian Sistem TKI

Dalam menguji signifikansi perbedaan performa antar kondisi eksperimen, penelitian sistem TKI umumnya menggunakan metode statistik parametrik dan non-parametrik. Uji *one-way ANOVA* digunakan ketika data terdistribusi normal, sedangkan uji *Kruskal-Wallis* digunakan sebagai

alternatif non-parametrik. Untuk perbandingan berpasangan antar kelompok, *uji Mann-Whitney U* merupakan pilihan yang tepat karena tidak mengasumsikan normalitas distribusi [14]. Regresi linier berganda digunakan untuk memodelkan hubungan antara fitur linguistik dan metrik performa, memungkinkan identifikasi prediktor dominan yang memengaruhi efisiensi SPIMI.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental komputasional untuk menganalisis performa algoritma *Single-Pass In-Memory Indexing (SPIMI)* pada dua jenis korpus teks berbahasa Indonesia, yaitu teks informal media sosial dan teks formal berita. Metodologi dirancang secara sistematis mencakup akuisisi data, pra-pemrosesan linguistik, implementasi algoritma, desain eksperimen terkontrol, analisis statistik, serta evaluasi kualitas *retrieval*.

3.1 Akuisisi dan Persiapan Data

Data penelitian bersumber dari dua berkas CSV yang merepresentasikan dua register bahasa yang kontras: korpus informal dari platform Twitter/X dan korpus formal dari kumpulan berita daring. Kedua berkas dibaca menggunakan fungsi ***load_corpus_csv()*** yang mengonversi setiap baris teks menjadi representasi dokumen terstandar dengan atribut *doc_id*, *text*, *corpus*, dan *created_at*.

Korpus	Sumber Data	Total Awal	Setelah Balancing	Karakteristik Utama
A_informal	Twitter Emotion Dataset (CSV)	4.401 tweet	4.401 tweet	OOV 82,70% avg 8,9 token/doc
B_formal	News Dataset (CSV)	20.951 artikel	4.401 sampel acak	OOV 89,01% avg 8,5 token/doc
Total	—	25.352 dokumen	8.802 dokumen	Rasio informal:formal = 1:1

Tabel 1. Ringkasan Dataset Penelitian

3.2 Prapemrosesan Data dan Ekstraksi Linguistic Feature Matrix (LFM)

Prapemrosesan teks dilakukan melalui *pipeline* berurutan menggunakan pustaka Sastrawi, NLTK, dan kamus slang kustom. Setiap tweet mentah dikonversi menjadi daftar token yang bersih dan terstandar melalui lima tahapan berikut:

1. Pembersihan Teks *(Text Cleaning)*: Teks dikonversi ke huruf kecil; URL, @mention, dan simbol tanda baca dihapus; angka dieliminasi; dan spasi berlebih dinormalisasi menggunakan ekspresi reguler.
2. Tokenisasi: Teks bersih dipecah menjadi token individual. Token dengan panjang kurang dari dua karakter diabaikan untuk menyaring karakter tunggal yang tidak bermakna.
3. Normalisasi Slang: Token slang dipetakan ke bentuk baku menggunakan kamus *SLANG_DICT* (contoh: 'gak'→'tidak', 'udah'→'sudah', 'bgt'→'banget') guna mengurangi inflasi *vocabulary* yang tidak informatif.
4. Penghapusan *Stopword*: Token frekuensi tinggi yang tidak bermakna secara semantik dihapus menggunakan daftar *stopword* berbasis KBBI (contoh: 'yang', 'dan', 'di', 'ke').
5. *Stemming*: Token direduksi ke bentuk kata dasar menggunakan pustaka Sastrawi yang mengimplementasikan algoritma *Enhanced Confix Stripping (ECS)*. Tahap ini meningkatkan *recall* sistem TKI dengan mengelompokkan kata-kata berakar sama [12].

Setelah pra-pemrosesan, tujuh fitur linguistik diekstrak per dokumen untuk membentuk ***Linguistic Feature Matrix (LFM)***. LFM berfungsi sebagai instrumen kuantifikasi tingkat informalitas yang kemudian dikaitkan dengan metrik performa SPIMI melalui analisis regresi. Ketujuh fitur LFM beserta deskripsinya disajikan pada Tabel 1.

Fitur LFM	Deskripsi
Out-Of-Vocabulary Rate (OOV)	Proporsi token yang tidak ditemukan dalam kamus baku (simulasi KBBI). OOV tinggi mengindikasikan tingkat informalitas yang lebih tinggi. Pada korpus informal, OOV rata-rata mencapai 82,83%, sedangkan formal 88,95%.
Slang Density	Proporsi token yang teridentifikasi sebagai kata slang berdasarkan *SLANG_DICT*. Indikator langsung frekuensi penggunaan slang seperti 'gak', 'udah', 'bgt' dalam teks.
Type-Token Ratio (TTR)	Rasio antara jumlah kata unik (*types*) dengan total kata (*tokens*). TTR tinggi dalam teks informal mengindikasikan variasi ejaan atau penggunaan kata non-standar yang luas.
Code-Mixing Index	Proporsi token berbahasa Inggris menggunakan identifikasi heuristis. CMI tinggi menunjukkan fenomena

(CMI)	campur kode (*code-mixing*) yang umum ditemukan di Twitter berbahasa Indonesia.
Document Length (doc_len)	Jumlah total token setelah pembersihan awal. Panjang dokumen memengaruhi kompleksitas proses indeksasi. Rata-rata informal: 7,18 token/dokumen; formal: 7,64 token/dokumen (setelah truncation hingga 9 token/dokumen).
Hapax Rate	Proporsi token yang hanya muncul satu kali dalam dokumen. *Hapax rate* tinggi mengindikasikan kosakata yang sangat beragam atau banyaknya variasi ejaan dan tipografi.
Vocabulary Growth	Jumlah term unik yang tersisa setelah seluruh tahap prapemrosesan. Menggambarkan ukuran *vocabulary* yang akan diindeks oleh SPIMI.

Tabel 2. Fitur-Fitur dalam Linguistic Feature Matrix (LFM)

Seluruh hasil LFM tersimpan ke berkas *results/lfm_matrix.csv* (8.802 baris) untuk keperluan analisis statistik selanjutnya.

3.3 Implementasi Algoritma SPIMI

Algoritma SPIMI diimplementasikan sebagai kelas Python *SPIMI* berdasarkan spesifikasi Büttcher, Clarke, dan Cormack [7], dengan tiga fase operasional yang berjalan secara sekuensial:

1. Fase *Build*: Token dari setiap dokumen yang telah diproses ditambahkan satu per satu ke *dictionary* in-memory menggunakan fungsi *add_token()*. Setiap entri menyimpan pasangan (*docID, term frequency*). Konsumsi memori dipantau secara kontinu menggunakan pustaka *tracemalloc*.
2. Fase *Flush*: Saat estimasi penggunaan memori mencapai ambang batas *buffer_size_mb*, isi *dictionary* diurutkan secara leksikografis dan diserialisasi ke disk dalam format berkas *.pkl* sebagai blok sementara. *Dictionary* kemudian dikosongkan untuk menampung token-token berikutnya.
3. Fase *Merge*: Setelah seluruh dokumen diproses, semua blok sementara di disk digabungkan menjadi satu *inverted index* final. *Posting list* setiap term diurutkan berdasarkan docID dan nilai TF yang duplikat dijumlahkan. Berkas blok sementara

dihapus setelah penggabungan berhasil.

Metrik kinerja yang diukur meliputi: waktu eksekusi fase *build*, *flush*, *merge*, dan total; konsumsi memori puncak (RAM peak dalam MB); jumlah operasi *flush*; ukuran *vocabulary*; jumlah total posting; ukuran indeks di disk (KB); serta throughput (dokumen per detik). Fungsi *search()* mengimplementasikan *Boolean AND retrieval* dengan pembobotan TF-IDF menggunakan formula *idf = log((N+1)/(df+1)) + 1*.

3.4 Desain Eksperimen

Eksperimen dirancang dengan desain faktorial penuh untuk mengevaluasi pengaruh dua faktor utama secara sistematis: (1) tingkat informalitas teks, terdiri dari korpus informal (*A_informal*) dan korpus formal (*B_formal*); dan (2) ukuran buffer memori yang dialokasikan, dengan tiga level yaitu 64 MB, 128 MB, dan 256 MB. Setiap kombinasi dijalankan sebanyak 5 kali repetisi, menghasilkan total 30 run eksperimen (2 korpus × 3 ukuran buffer × 5 repetisi).

Repetisi pertama setiap kombinasi dikategorikan sebagai *warm-up run* dan diabaikan untuk mengeliminasi efek *cold-start* sistem operasi. Data dari empat repetisi valid kemudian diagregasi (rata-rata dan standar deviasi) dan disimpan dalam berkas *results/experiment_results_raw.csv* dan *results/experiment_results_aggregated.csv*. Reproduksibilitas eksperimen dijamin melalui penetapan seed acak (*random.seed(42)* dan *numpy.random.seed(42)*).

3.5 Analisis Statistik

Analisis statistik dilakukan untuk menguji signifikansi perbedaan performa SPIMI antar kondisi eksperimen dan memodelkan hubungan antara fitur LFM dengan metrik kinerja. Empat metode analisis diterapkan secara hierarkis.

1. Pertama, uji normalitas Shapiro-Wilk digunakan untuk menentukan apakah distribusi data bersifat normal ($\alpha = 0{,}05$). Hasilnya menentukan pemilihan uji komparatif pada langkah berikutnya.
2. Uji one-way ANOVA digunakan jika data terdistribusi normal, atau Kruskal-Wallis sebagai alternatif non-parametrik jika tidak normal. Kedua uji diterapkan pada metrik *total_time_ms*, *ram_peak_mb*, *vocab_size*, dan *index_size_kb* dengan hipotesis nol H_0: tidak ada perbedaan signifikan antar kelompok ($\alpha = 0{,}05$).
3. Uji post-hoc Mann-Whitney U dua sisi diterapkan untuk perbandingan berpasangan antara korpus informal dan formal guna mengidentifikasi pasangan dengan perbedaan signifikan secara statistik.
4. Regresi linier berganda digunakan untuk memodelkan hubungan antara enam fitur LFM

(variabel independen: *oov_rate, slang_density, ttr, cmi, doc_len, hapax_rate*) dan tiga metrik performa SPIMI (variabel dependen: *total_time_ms, ram_peak_mb, vocab_size*). Fitur distandarisasi menggunakan *StandardScaler* sebelum dimasukkan ke model agar koefisien dapat dibandingkan langsung. Nilai R^2 digunakan sebagai ukuran kualitas model. Hipotesis formal yang diuji adalah H_0: tingkat informalitas teks TIDAK berpengaruh signifikan terhadap performa SPIMI, dan H_1: tingkat informalitas teks BERPENGARUH SIGNIFIKAN terhadap performa SPIMI ($\alpha = 0,05$).

3.6 Evaluasi Kualitas Retrieval

Kualitas retrieval dievaluasi menggunakan *inverted index* yang dibangun pada ketiga ukuran buffer (64 MB, 128 MB, 256 MB) untuk mengevaluasi konsistensi kualitas retrieval terhadap variasi konfigurasi buffer. Set kueri benchmark terdiri dari 15 kueri yang terbagi dalam dua kategori: kueri informal (Q01–Q08) yang mengandung ekspresi slang dan bahasa sehari-hari (contoh: *'gak ngerti apa maksudnya'*), serta kueri formal (Q09–Q15) yang menggunakan diksi formal dan terminologi berita (contoh: *'pertumbuhan ekonomi nasional'*).

Setiap kueri diproses melalui *pipeline* pra-pemrosesan yang identik dengan dokumen sebelum digunakan dalam fungsi *search()*. Karena tidak tersedia *ground truth* yang tervalidasi untuk data Twitter informal, penilaian relevansi dilakukan secara simulatif: dokumen yang mengandung minimal satu token kueri dianggap relevan, kemudian diambil sepertiga dari total dokumen relevan tersebut sebagai set *ground truth*. Metrik evaluasi yang digunakan meliputi

(1) Precision@5 (P@5): proporsi dokumen relevan di antara 5 hasil teratas;

(2) Mean Average Precision (MAP): rata-rata presisi pada setiap posisi ditemukannya dokumen relevan;

(3) NDCG@10: kualitas ranking dengan bobot lebih pada posisi lebih tinggi; dan

(4) latensi kueri (ms): waktu rata-rata pemrosesan satu kueri.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Eksperimen SPIMI

Eksperimen dilakukan menggunakan desain faktorial penuh 2 × 3 × 5 (2 korpus × 3 ukuran buffer × 5 repetisi) menghasilkan total 30 run, dengan repetisi pertama setiap kombinasi dikecualikan sebagai warm-up run. Hasil agregasi dari empat repetisi valid pada setiap kombinasi disajikan pada Tabel 3.

Korpus	Buffer (MB)	P@5	MAP	NDCG@10	Lat(ms)
A_informal	64	0,320	0,071	0,311	0,169
A_informal	128	0,320	0,071	0,311	0,224
A_informal	256	0,320	0,071	0,311	0,164
B_formal	64	0,320	0,135	0,377	0,092
B_formal	128	0,320	0,135	0,377	0,080
B_formal	256	0,320	0,135	0,377	0,081

Tabel 3. Hasil Evaluasi Kualitas Retrieval per Korpus dan Ukuran Buffer

Dari Tabel 3 terlihat bahwa kualitas retrieval (P@5, MAP, NDCG@10) bersifat konsisten di semua ukuran buffer untuk kedua korpus, mengkonfirmasi bahwa ukuran buffer SPIMI lebih memengaruhi efisiensi build daripada kualitas pencarian. Kedua korpus memperoleh nilai Precision@5 (P@5) yang identik yaitu 0,320 pada semua konfigurasi buffer. Namun, korpus formal unggul pada metrik ranking: MAP formal (0,135) hampir dua kali lebih tinggi dibandingkan MAP informal (0,071), dan NDCG@10 formal (0,377) lebih tinggi dari NDCG@10 informal (0,311). Latensi retrieval korpus formal (0,080–0,092 ms) lebih rendah dibandingkan korpus informal (0,164–0,224 ms), menunjukkan bahwa indeks teks formal lebih efisien untuk ditelusuri.

Nilai MAP yang rendah pada korpus informal (0,071) mengindikasikan bahwa meskipun sistem berhasil menemukan dokumen relevan di antara 5 hasil teratas (P@5 = 0,320), peringkat keseluruhan dokumen relevan dalam daftar hasil masih jauh dari ideal. Fenomena ini dapat dijelaskan oleh rendahnya konsistensi term matching pada teks informal: variasi ejaan slang yang tidak dinormalisasi sepenuhnya menyebabkan banyak dokumen relevan tidak muncul pada posisi awal hasil retrieval. Temuan ini sejalan dengan hasil penelitian Jalilifard et al. (2020) [15] yang menegaskan bahwa performa TF-IDF pada teks informal secara inheren lebih rendah karena tingginya ambiguitas semantik term non-standar.

Di sisi lain, hasil ini berbanding terbalik dengan temuan Catena et al. (2014) [16] pada korpus formal berskala besar (GOV2, ClueWeb09) yang menunjukkan bahwa peningkatan kompresi indeks tidak mengakibatkan penurunan kualitas retrieval yang signifikan. Dalam penelitian ini,

perbedaan ukuran indeks antara korpus informal (506,7 KB) dan formal (545,0 KB) ternyata tidak menurunkan P@5, namun secara nyata memengaruhi latensi dan kualitas ranking yang diukur oleh MAP dan NDCG@10.

4.4 Pembahasan

4.4 Pengaruh Karakteristik Linguistik terhadap Efisiensi SPIMI

Hasil penelitian ini secara keseluruhan membuktikan bahwa karakteristik linguistik teks, khususnya panjang dokumen, kepadatan slang, dan keberagaman kosakata memiliki dampak yang signifikan terhadap performa SPIMI. Korpus informal menghasilkan ukuran vocabulary sebesar 6.452 term, sekitar 56% lebih besar dibandingkan korpus formal (4.135 term), dengan konsumsi RAM puncak yang secara statistik signifikan lebih tinggi (2,62 MB vs. 2,46 MB, $p < 0,0001$). Temuan ini mengonfirmasi bahwa vocabulary explosion pada teks informal berbahasa Indonesia, meskipun lebih moderat setelah truncation ke 9 token/dokumen, tetap memberikan tekanan ekstra pada konsumsi memori SPIMI.

Secara teoritis, kompleksitas waktu SPIMI adalah O(T) di mana T adalah jumlah total token dalam koleksi [7]. Namun dalam praktiknya, waktu aktual juga dipengaruhi oleh ukuran kamus yang harus dikelola. Setelah truncation dokumen ke rata-rata 9 token/dokumen, perbedaan waktu build antara kedua korpus tidak lagi signifikan secara statistik ($p = 0,4426$), meskipun tren informal lebih lambat pada buffer kecil (64 MB: 232,3 ms vs. 145,0 ms) masih terlihat.

Temuan ini sejalan dengan penelitian Catena et al. (2014) [16] yang menunjukkan bahwa ukuran vocabulary dan karakteristik distribusi term adalah faktor penentu utama efisiensi indeks, serta dengan penelitian Jalilifard et al. (2020) [15] yang mengidentifikasi bahwa teks informal dari media sosial secara konsisten menimbulkan tantangan lebih besar dalam pembobotan dan pengindeksan term dibandingkan teks formal.

4.4.2 Pengaruh Ukuran Buffer terhadap Performa SPIMI

Pola pengaruh ukuran buffer terhadap throughput menunjukkan dinamika yang menarik dan berbeda antara kedua korpus. Pada korpus informal, peningkatan buffer dari 64 MB ke 256 MB meningkatkan throughput secara substansial dari 19.834,2 menjadi 35.500,4 dokumen/detik (peningkatan 79,0%). Hal ini terjadi karena buffer yang lebih besar memungkinkan SPIMI mempertahankan lebih banyak token di memori tanpa overhead realokasi, sangat menguntungkan saat mengelola kamus vocabulary yang lebih besar.

Sebaliknya, pada korpus formal, throughput relatif stabil (32.465,1 – 33.017,2 – 32.890,5 dokumen/detik) di seluruh konfigurasi buffer. Ini menunjukkan bahwa untuk vocabulary yang lebih kecil, ukuran buffer tidak memberikan perbedaan berarti. Implikasi praktisnya adalah

bahwa konfigurasi buffer optimal bersifat corpus-dependent: untuk teks informal, buffer 256 MB memberikan throughput maksimal, sementara untuk teks formal, buffer 64 MB sudah memadai.

Temuan tentang pengaruh ukuran buffer ini memperkaya pemahaman tentang perilaku SPIMI: untuk korpus dengan vocabulary kecil (teks formal), ukuran buffer tidak memberikan perbedaan berarti terhadap throughput, sementara untuk korpus informal dengan vocabulary yang lebih besar, buffer yang lebih besar memberikan keuntungan signifikan [7]. Hasil ini memperkaya pemahaman tentang perilaku SPIMI dalam kondisi corpus yang beragam.

4.4.3 Kualitas Retrieval dan Implikasi Prapemrosesan

Perbedaan kualitas retrieval antara kedua korpus, terutama pada metrik MAP (0,071 vs. 0,135) dan NDCG@10 (0,311 vs. 0,377) menunjukkan bahwa *pipeline* prapemrosesan yang diterapkan dalam penelitian ini, meskipun mencakup normalisasi slang, penghapusan *stopword*, dan *stemming* ECS, belum sepenuhnya mampu mengatasi kompleksitas linguistik teks informal. Nilai MAP yang rendah (0,071) pada korpus informal mengindikasikan bahwa sistem retrieval seringkali menemukan dokumen yang mengandung token kueri, namun gagal menempatkan dokumen yang paling relevan di posisi teratas ranking.

Fenomena ini dapat dikaitkan dengan dua faktor. Pertama, normalisasi slang berbasis kamus (*SLANG_DICT*) yang digunakan hanya mampu menangani slang yang terdaftar secara eksplisit, sehingga variasi ejaan baru atau slang yang belum masuk kamus (novel slang) tidak dapat dinormalisasi. Kedua, pembobotan TF-IDF standar tidak mampu menangkap hubungan semantik antar term informal yang berbeda bentuk namun bermakna sama. Hal ini konsisten dengan temuan Jalilifard et al. (2020) [15] yang mengusulkan pendekatan *Semantic Sensitive* TF-IDF (STF-IDF) sebagai solusi untuk permasalahan serupa pada teks informal medis, di mana TF-IDF standar menghasilkan error rate 27,2% sementara STF-IDF berhasil mereduksinya menjadi 13,7%.

Temuan latensi retrieval yang lebih tinggi pada korpus informal (0,164–0,224 ms vs. 0,080–0,092 ms pada korpus formal) juga memiliki implikasi penting bagi pengembang sistem pencarian berbasis media sosial: indeks yang dibangun dari teks informal membutuhkan waktu akses yang lebih lama, yang pada skala produksi dengan jutaan dokumen akan berdampak signifikan terhadap pengalaman pengguna. Oleh karena itu, strategi kompresi indeks atau penggunaan struktur data yang lebih efisien seperti yang dibahas dalam Pibiri dan Venturini (2020) [17] perlu dipertimbangkan untuk deployment sistem TKI pada data media sosial berbahasa Indonesia skala besar.

5. SIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini berhasil menganalisis secara sistematis pengaruh karakteristik linguistik teks berbahasa Indonesia, khususnya tingkat informalitas terhadap efisiensi algoritma SPIMI dalam pembangunan inverted index. Berdasarkan 30 run eksperimen dan analisis statistik yang komprehensif, lima kesimpulan utama dapat ditarik:

1. Informalitas teks berpengaruh signifikan terhadap konsumsi RAM dan ukuran vocabulary SPIMI, namun tidak pada waktu build. Korpus informal memiliki konsumsi RAM puncak yang lebih tinggi (2,62 MB vs. 2,46 MB, $p < 0,0001$) dan vocabulary yang lebih besar (6.452 vs. 4.135 term). Waktu build tidak berbeda signifikan secara statistik setelah penyeimbangan panjang dokumen ($p = 0,4426$), sehingga H_1 diterima secara parsial: tingkat informalitas teks berpengaruh signifikan terhadap konsumsi RAM dan vocabulary SPIMI, tetapi tidak terhadap waktu build.
2. Vocabulary explosion tetap menjadi faktor pembeda utama. Setelah truncation dokumen ke rata-rata 9 token/dokumen, korpus informal masih menghasilkan 6.452 term unik (56% lebih besar dari korpus formal dengan 4.135 term) akibat variasi ejaan slang, singkatan, dan fenomena *code-mixing* yang khas pada teks media sosial berbahasa Indonesia. Ekspansi kosakata ini secara langsung meningkatkan beban komputasi pada fase build SPIMI.
3. Pengaruh ukuran buffer bersifat *corpus-dependent*. Pada korpus informal, buffer 256 MB menghasilkan throughput tertinggi (35.500,4 dok/s), sementara pada korpus formal throughput relatif stabil di semua ukuran buffer (≈32.000–33.000 dok/s). Temuan ini menambahkan dimensi baru pada pemahaman tentang SPIMI yang sebelumnya mengasumsikan bahwa buffer yang lebih besar selalu menghasilkan performa lebih baik.
4. Panjang dokumen (*doc_len*) dan kepadatan slang (*slang_density*) adalah prediktor linguistik paling dominan. Model regresi linier berganda ($R^2 = 0,015–0,284$) menunjukkan bahwa slang_density berkorelasi positif kuat dengan konsumsi RAM dan ukuran vocabulary, sementara doc_len berpengaruh positif moderat pada semua metrik, mengkonfirmasi bahwa karakteristik linguistik dokumen individual dapat digunakan untuk memprediksikan beban komputasi indexing.
5. Kualitas retrieval teks formal secara konsisten lebih baik pada metrik ranking dan konsisten di semua ukuran buffer. Meskipun P@5 kedua korpus identik (0,320), MAP formal (0,135) hampir dua kali lebih tinggi dari MAP informal (0,071) dan latensi retrieval informal lebih lambat (0,164–0,224 ms vs. 0,080–0,092 ms). Hal ini mengindikasikan

bahwa *pipeline* prapemrosesan yang ada belum cukup optimal untuk menangani heterogenitas term pada teks informal berbahasa Indonesia.

5.2 Saran

Berdasarkan temuan penelitian ini, beberapa saran dikemukakan untuk pengembangan sistem TKI dan penelitian lanjutan:

1. Konfigurasi buffer adaptif berbasis profil korpus. Pengembang sistem TKI yang menangani data media sosial berbahasa Indonesia disarankan menggunakan buffer minimal 256 MB, sementara untuk corpus formal 64 MB sudah memadai. Ke depan, mekanisme pemilihan ukuran buffer secara otomatis berdasarkan estimasi awal ukuran *vocabulary* (misalnya, dari sampel 1–5% dokumen pertama) dapat dikembangkan untuk mengoptimalkan alokasi memori secara dinamis.
2. Penguatan *pipeline* pra-pemrosesan teks informal. Normalisasi slang berbasis kamus (*SLANG_DICT*) perlu diperkuat dengan pendekatan berbasis model bahasa kontekstual (misalnya, IndoBERT atau IndoNLG) untuk menangani novel slang dan variasi ejaan yang tidak terdaftar. Hal ini krusial mengingat OOV rate rata-rata 82,83% pada korpus Twitter yang dianalisis.
3. Adopsi pembobotan *term* yang lebih canggih. Penggantian TF-IDF standar dengan pendekatan yang lebih sensitif terhadap konteks *semantic* seperti BM25+ atau Semantic Sensitive TF-IDF (Jalilifard et al., 2020) direkomendasikan untuk meningkatkan kualitas ranking pada kueri informal, yang saat ini menghasilkan MAP hanya 0,071.
4. Perluasan fitur *Linguistic Feature Matrix*. LFM dalam penelitian ini menggunakan tujuh fitur; penelitian lanjutan dapat menambahkan fitur seperti kompleksitas sintaksis, kepadatan emoji, atau pola afiksasi non-standar untuk meningkatkan akurasi model prediktif performa SPIMI (target $R^2 > 0{,}85$).
5. Studi komparatif dengan algoritma indexing lain. Perbandingan sistematis antara SPIMI dan BSBI (*Blocked Sort-Based Indexing*) pada corpus informal berbahasa Indonesia perlu dilakukan untuk mengidentifikasi algoritma yang paling robust, terutama pada skala korpus yang lebih besar (>100.000 dokumen) di mana perbedaan strategi pengelolaan memori akan lebih kentara.
6. Pengembangan ground truth anotasi manual. Evaluasi retrieval dalam penelitian ini menggunakan ground truth simulatif; penelitian lanjutan perlu mengembangkan dataset benchmark dengan anotasi relevansi manual oleh ahli bahasa Indonesia untuk menghasilkan evaluasi MAP dan NDCG yang lebih sahih dan dapat dibandingkan dengan

standar internasional.

7. Pengujian skalabilitas pada corpus produksi. Penelitian ini menggunakan 8.802 dokumen; eksperimen pada skala jutaan dokumen (misalnya, menggunakan dataset Twitter Indonesia publik) perlu dilakukan untuk mengevaluasi skalabilitas SPIMI dan mengidentifikasi *bottleneck* yang mungkin muncul pada lingkungan produksi nyata.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada semua pihak yang telah berkontribusi dalam penyelesaian penelitian ini. Ucapan terima kasih secara khusus disampaikan kepada Dosen pengampu mata kuliah sistem temu kembali yang telah memberikan arahan, masukan, dan bimbingan yang sangat berharga selama proses penelitian berlangsung. Tidak lupa, penulis menyampaikan apresiasi kepada rekan-rekan penelitian yang telah memberikan diskusi, masukan, dan semangat selama proses pengerjaan. Kolaborasi dan dukungan yang diberikan turut memperkaya perspektif dan kualitas penelitian ini.

Penelitian ini diharapkan dapat memberikan kontribusi yang berarti bagi pengembangan sistem temu kembali informasi berbahasa Indonesia, khususnya dalam konteks pemrosesan teks informal media sosial.

DAFTAR PUSTAKA

[1] Kemp, S. (2023). Digital 2023: Indonesia.

[2] Kusumaningrum, R., & Adhy, S. (2017). Analisis sentimen opini publik mengenai sarana transportasi menggunakan metode naïve Bayes dan SVM pada media sosial Twitter. Jurnal Teknologi dan Sistem Komputer, 5(1), 1–8.

[3] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

[4] Tala, F. Z. (2003). A study of stemming effects on information retrieval in Bahasa Indonesia [Master's thesis, Universiteit van Amsterdam]. ILK Research Group

[5] Prasetyo, E., & Purwarianti, A. (2015). Normalisasi teks informal Twitter bahasa Indonesia menggunakan pendekatan berbasis kamus dan aturan. Jurnal Ilmu Komputer dan Informasi, 8(2), 93–101

[6] Gunawan, D., Wijaya, A., & Rahardja, U. (2020). Sistem temu kembali informasi berbasis pembobotan TF-IDF pada dokumen teks berbahasa Indonesia. Jurnal Nasional Teknologi dan Sistem Informasi, 6(2), 78–86

[7] S. Büttcher, C. L. A. Clarke, and G. V. Cormack, *Information Retrieval: Implementing*

and Evaluating Search Engines. Cambridge, MA: MIT Press, 2010.

[8] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks," *Sintesia*, vol. 1, pp. 81–88, 2022.

[9] R. Riyaddulloh and A. Romadhony, "Normalisasi Teks Bahasa Indonesia Berbasis Kamus Slang Studi Kasus: Tweet Produk Gadget Pada Twitter," *e-Proceeding Eng.*, vol. 8, no. 4, p. 4216, 2021.

[10] P. M. S. Ardinata, A. A. J. Permana, and I. N. S. W. Wijaya, "Identifikasi Dan Normalisasi Teks Slang Dengan FASTTEXT PADA TWITTER DALAM BAHASA INDONESIA," vol. 21, no. 1, 2024.

[11] E. Yulianti, A. Kurnia, M. Adriani, and Y. S. Duto, "Normalisation of Indonesian-English Code-Mixed Text and its Effect on Emotion Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 674–685, 2021, doi: 10.14569/IJACSA.2021.0121177.

[12] J. Pardede, "Perbandingan Algoritma Stemming Porter , Sastrawi , Idris , Comparison of Stemming Algorithms Porter , Sastrawi , Idris , and Arifin Setiono on Indonesian Text Documents," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 12, no. 1, 2025, doi: 10.25126/jtiik.2025128860.

[13] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.

[14] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, 2nd ed. New York: ACM Press/Addison-Wesley, 2011.

[15] Jalilifard, A., Caridá, V. F., Mansano, A. F., & Cristo, R. (2020). Semantic Sensitive TF-IDF to Determine Word Relevance in Documents. https://arxiv.org/pdf/2001.09896 3.

[16] Catena, M., Macdonald, C., & Ounis, I. (2014). On Inverted Index Compression for Search Engine Efficiency. In European Conference on Information Retrieval (ECIR 2014).

[17] Pibiri, G. E., & Venturini, R. (2020). Techniques for Inverted Index Compression. ACM Computing Surveys (CSUR). https://doi.org/10.1145/3415148. 2.