ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

SISTEM EKSTRAKSI INFORMASI LIRIK LAGU DAERAH MENGGUNAKAN *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* (TF-IDF) DAN *COSINE SIMILARITY*

Nur Ifdayanti✉, Fajar Ratnawati
Jurusan Teknik Informatika, Politeknik Negeri Bengkalis, Riau, Indonesia
Email: nurifdayanti710@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol10No2.pp575-581

ABSTRACT

This study develops an information extraction system for Indonesian regional folk-song lyrics by combining Term Frequency-Inverse Document Frequency (TF-IDF), Cosine Similarity, and Part-of-Speech (POS) Tagging with rule-based extraction. A preliminary survey of 33 respondents indicated limited familiarity with regional songs and difficulty understanding lyric meanings, motivating a web-based system developed using Rapid Application Development (RAD). The system processes lyric queries through case folding, tokenization, stopword removal, stemming, TF-IDF weighting, and Cosine Similarity ranking to retrieve the most relevant song. The selected lyrics are then analyzed to extract nouns, verbs, adjectives, pronouns, predicates, actions, subjects, objects, and related semantic elements. Black-box testing showed that all main functions operated successfully. Retrieval evaluation on ten test queries achieved 100% Top-1 Accuracy, while linguistic extraction against manual annotation produced Precision 0.74, Recall 0.76, and F1-Score 0.75. These results indicate that integrating statistical weighting and linguistic analysis can effectively support cultural documentation and regional-language learning.
Keyword: TF-IDF, Cosine Similarity, POS Tagging, Information Extraction, Regional Folk Song.

ABSTRAK

Penelitian ini mengembangkan sistem ekstraksi informasi lirik lagu daerah Indonesia dengan mengombinasikan Term Frequency-Inverse Document Frequency (TF-IDF), Cosine Similarity, dan Part-of-Speech (POS) Tagging berbasis ekstraksi aturan. Survei awal terhadap 33 responden menunjukkan tingkat pengenalan lagu daerah yang masih terbatas serta kesulitan memahami makna lirik, sehingga diperlukan sistem berbasis web yang dikembangkan menggunakan metode Rapid Application Development (RAD). Sistem memproses query lirik melalui case folding, tokenizing, stopword removal, stemming, pembobotan TF-IDF, dan pemeringkatan Cosine Similarity untuk mengambil lagu yang paling relevan. Lirik terpilih kemudian dianalisis untuk mengekstraksi nomina, verba, adjektiva, pronomina, predikat, aksi, subjek, objek, dan elemen semantik terkait. Pengujian blackbox menunjukkan seluruh fungsi utama berjalan sesuai harapan. Evaluasi retrieval terhadap sepuluh query uji menghasilkan Top-1 Accuracy 100%, sedangkan evaluasi ekstraksi linguistik terhadap anotasi manual menghasilkan Precision 0,74, Recall 0,76, dan F1-Score 0,75. Hasil ini menunjukkan bahwa integrasi pembobotan statistik dan analisis linguistik dapat mendukung dokumentasi budaya dan pembelajaran bahasa daerah.
Kata Kunci: TF-IDF, Cosine Similarity, POS Tagging, Ekstraksi Informasi, Lagu Daerah.

PENDAHULUAN

Perkembangan teknologi informasi membawa perubahan besar dalam upaya pelestarian budaya, termasuk pelestarian lagu daerah yang mengandung nilai sejarah, kearifan lokal, dan identitas masyarakat. Indonesia memiliki lagu daerah yang beragam dari berbagai provinsi, seperti Bubuy Bulan dan Es Lilin dari Jawa Barat serta Butet dan Anju Ahu dari Sumatera Utara. Namun, sebagian lagu daerah masih belum terdigitalisasi secara baik dan hanya tersimpan dalam bentuk fisik atau diwariskan secara lisan, sehingga berisiko sulit ditemukan, dianalisis, atau dipelajari oleh generasi muda.

Survei awal terhadap 33 responden menunjukkan bahwa tingkat pengenalan masyarakat terhadap lagu daerah masih terbatas. Sebanyak 54,5% responden hanya mengenal sebagian lagu daerah, 45,5% kurang mengenal atau sekadar mengetahui judulnya, 93,9% jarang atau hampir tidak pernah mendengarkan lagu daerah, dan 78,8% mengalami kesulitan memahami makna lirik. Temuan ini menunjukkan adanya kebutuhan nyata terhadap sistem yang tidak hanya menyimpan lirik, tetapi juga membantu pengguna memperoleh informasi penting, tema, serta makna yang terkandung di dalamnya.

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

Ekstraksi informasi merupakan proses mengidentifikasi elemen penting dari teks tidak terstruktur sehingga informasi yang diperoleh dapat disajikan secara ringkas dan bermakna. Dalam konteks lirik lagu daerah, ekstraksi informasi dapat digunakan untuk menemukan kata kunci, kategori linguistik, tema, subjek, aksi, serta ringkasan makna yang berkaitan dengan nilai budaya. Pemrosesan tersebut dapat mendukung kegiatan dokumentasi budaya, pendidikan bahasa daerah, dan pencarian informasi lirik secara lebih terstruktur.

Beberapa penelitian terdahulu menunjukkan bahwa TF-IDF efektif untuk pembobotan kata dan ekstraksi kata kunci pada berbagai domain teks. Wibowo et al. (2024) menerapkan TF-IDF pada judul berita online dan memperoleh kinerja yang sangat tinggi, namun keluaran sistem terbatas pada daftar kata kunci tanpa analisis peran gramatikal kata. Timur (2025) menggunakan TF-IDF dan *Cosine Similarity* pada rekomendasi lagu berbasis lirik, sedangkan Al Rasyid et al. (2024) menerapkan kombinasi kedua metode tersebut pada pencarian destinasi wisata; kedua penelitian tersebut menghentikan proses pada tahap perankingan dan rekomendasi dokumen tanpa mengurai lebih lanjut struktur linguistik dari teks yang diperoleh. Dengan kata lain, penelitian-penelitian terdahulu umumnya memposisikan TF-IDF dan *Cosine Similarity* semata-mata sebagai mesin pencarian atau perekomendasi dokumen, sehingga keluaran yang dihasilkan berhenti pada skor kemiripan atau daftar dokumen relevan. Integrasi TF-IDF dan Cosine Similarity dengan POS Tagging serta rule-based extraction untuk mengekstraksi elemen linguistik pada lirik lagu daerah Indonesia masih belum banyak dikaji, sehingga celah inilah yang menjadi landasan kebaruan penelitian ini.

Penelitian ini bertujuan mengembangkan sistem ekstraksi informasi lirik lagu daerah menggunakan TF-IDF dan Cosine Similarity, dilengkapi POS Tagging dan pendekatan rule-based untuk mengidentifikasi elemen linguistik. Berbeda dengan penelitian terdahulu yang umumnya memanfaatkan TF-IDF dan *Cosine Similarity* hanya sebagai mekanisme pencarian atau perekomendasi dokumen, kebaruan penelitian ini terletak pada integrasi empat komponen sekaligus, yaitu pembobotan statistik (TF-IDF), pengukuran kemiripan (*Cosine Similarity*), identifikasi kelas kata (POS Tagging), dan ekstraksi berbasis aturan (*rule-based extraction*), sehingga sistem tidak berhenti pada tahap penemuan dokumen yang relevan, melainkan dilanjutkan dengan pembongkaran struktur linguistik lirik secara otomatis menjadi elemen subjek, predikat, objek, aksi, negasi, dan sasaran. Integrasi tersebut menghasilkan keluaran yang lebih rinci daripada sekadar skor kemiripan atau daftar kata kunci sebagaimana dihasilkan penelitian-penelitian sebelumnya, sehingga sistem dapat membantu pengguna memahami isi lirik dan nilai budaya lagu daerah secara lebih mudah, sekaligus menjawab celah penelitian pada ranah ekstraksi informasi linguistik untuk lirik lagu daerah Indonesia yang masih jarang dikaji.

KAJIAN LITERATUR

Ekstraksi Informasi dan Lagu Daerah

Ekstraksi informasi adalah proses menemukan informasi terstruktur dari dokumen tidak terstruktur atau semi-terstruktur dan menjadi bagian dari pemrosesan bahasa alami. Pada lirik lagu daerah, proses ini tidak hanya berkaitan dengan pengambilan kata kunci, tetapi juga dengan identifikasi peran kata, tema, dan makna semantik yang dapat mewakili pesan budaya dalam lirik. Lagu daerah sendiri umumnya menggunakan bahasa daerah, memuat nilai kehidupan masyarakat, serta sering kali menjadi media pewarisan tradisi lokal.

TF-IDF dan *Cosine Similarity*

Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk menilai pentingnya suatu term dalam dokumen berdasarkan frekuensi kemunculan term dan tingkat kelangkaannya pada seluruh koleksi dokumen. Term yang sering muncul pada satu dokumen tetapi jarang muncul pada dokumen lain akan memperoleh bobot lebih tinggi. Bobot TF-IDF dalam penelitian ini dihitung menggunakan Persamaan (1).

$$W(t,d) = \log(1 + tf(t,d)) \times \log10(N / df(t)) \quad ....(1)$$

Pada Persamaan (1), tf(t,d) adalah frekuensi kemunculan term t pada dokumen d, df(t) adalah jumlah dokumen yang memuat term t, dan N adalah jumlah seluruh dokumen dalam koleksi.

Cosine Similarity digunakan untuk mengukur kemiripan antara vektor query dan vektor dokumen. Nilai kemiripan berada pada rentang 0 sampai 1, dengan nilai mendekati 1 menunjukkan dokumen semakin relevan terhadap query. Rumus Cosine Similarity ditunjukkan pada Persamaan (2).

$$cos(theta) = (A.B) / (||A|| \times ||B||) .....(2)$$

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

POS *Tagging* dan *Rule-Based Extraction*

POS *Tagging* digunakan untuk mengidentifikasi kelas kata seperti nomina, verba, adjektiva, pronomina, partikel, dan kategori linguistik lainnya. Karena lirik lagu daerah memiliki kosakata lokal yang tidak selalu dikenali model bahasa umum, penelitian ini menambahkan rule-based extraction sebagai mekanisme pendukung untuk menentukan elemen seperti subjek, objek, predikat, aksi, negasi, dan sasaran. Kombinasi tersebut diharapkan dapat meningkatkan kelengkapan informasi linguistik yang dihasilkan sistem.

Perbandingan dengan Metode Alternatif

Selain TF-IDF dan *Cosine Similarity*, beberapa metode lain seperti BM25, Word2Vec, dan Sentence-BERT juga umum digunakan pada tugas pencarian dan pengukuran kemiripan teks. BM25 pada dasarnya merupakan penyempurnaan probabilistik dari TF-IDF dengan menambahkan faktor normalisasi panjang dokumen, sehingga berpotensi memberikan peringkat yang lebih baik pada koleksi dokumen dengan variasi panjang yang besar. Word2Vec dan Sentence-BERT merupakan pendekatan berbasis embedding yang mampu menangkap kemiripan makna antarkata atau antarkalimat meskipun kata yang digunakan berbeda, namun performanya sangat bergantung pada ketersediaan korpus pelatihan berskala besar. Untuk bahasa daerah Indonesia, korpus berskala besar maupun model embedding terlatih (*pretrained*) pada bahasa-bahasa tersebut masih sangat terbatas, sehingga penerapan Word2Vec atau Sentence-BERT berisiko menghasilkan representasi yang kurang optimal atau memerlukan pelatihan ulang model dari awal dengan data yang sangat terbatas. TF-IDF dan Cosine Similarity dipilih pada penelitian ini karena tidak memerlukan korpus pelatihan berskala besar, ringan secara komputasi, mudah diinterpretasikan, dan telah terbukti pada penelitian-penelitian sebelumnya efektif untuk pencarian dan pengukuran kemiripan dokumen berbasis teks. Meskipun demikian, perbandingan kinerja secara langsung dan terukur antara TF-IDF/Cosine Similarity dengan BM25, Word2Vec, dan Sentence-BERT belum dilakukan pada penelitian ini dan menjadi salah satu arah penting bagi penelitian lanjutan agar keunggulan metode yang digunakan dapat dibuktikan secara lebih objektif.

Penelitian Terdahulu

Kajian terdahulu menunjukkan bahwa TF-IDF dan *Cosine Similarity* konsisten digunakan pada pencarian informasi, rekomendasi, dan pengukuran kemiripan dokumen. Situmorang et al. (2021) mengombinasikan N-Gram dan Cosine Similarity untuk rekomendasi lagu berdasarkan lirik. Pribadi dan Munir (2021) menggunakan TF-IDF untuk mendeteksi kemiripan judul tugas akhir. Djunaidi et al. (2024) menerapkan TF-IDF dan Cosine Similarity pada pencarian informasi obat. Namun, penelitian-penelitian tersebut umumnya berhenti pada tahap pencarian, perankingan, atau rekomendasi dokumen tanpa mengurai lebih lanjut struktur gramatikal maupun peran semantik kata di dalam teks yang diperoleh, dan belum secara khusus membahas ekstraksi elemen linguistik dari lirik lagu daerah Indonesia. Penelitian ini mengisi celah tersebut dengan mengintegrasikan TF-IDF dan *Cosine Similarity* sebagai mesin pencarian dengan POS Tagging dan *rule-based extraction* sebagai lapisan analisis lanjutan, sehingga keluaran sistem berupa elemen linguistik terstruktur (subjek, predikat, objek, aksi, negasi, dan sasaran) yang belum ditawarkan oleh penelitian-penelitian pembanding di atas.

METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan metode *Rapid Application Development* (RAD) karena kebutuhan sistem telah dapat diidentifikasi sejak awal dan pengembangan membutuhkan proses perancangan prototipe yang cepat. Tahapan penelitian meliputi identifikasi masalah, requirement planning, user design, construction, cutover, pengujian sistem, dan penyusunan laporan. Metode ini sesuai untuk pengembangan sistem berbasis web yang membutuhkan iterasi rancangan antarmuka dan evaluasi fungsi sistem secara langsung.

Alat dan Data Penelitian

Sistem dikembangkan menggunakan perangkat keras dengan prosesor Intel Core i3, RAM 12 GB, dan penyimpanan 1 TB HDD. Perangkat lunak yang digunakan meliputi Windows 11, Microsoft Word, Visual Studio Code, Microsoft Visio, Figma, Apache, PHP, dan Laravel. Data penelitian berupa lirik lagu daerah dari berbagai wilayah Indonesia yang dihimpun melalui studi literatur terhadap buku kumpulan lagu daerah, artikel budaya, portal lagu daerah, dan sumber daring pendukung. Data pendukung juga diperoleh melalui kuesioner terhadap 33 responden untuk mengidentifikasi kebutuhan pengguna.

Analisis Kebutuhan Sistem

Kebutuhan fungsional sistem meliputi kemampuan menampilkan teks lengkap lirik lagu daerah, menampilkan hasil ekstraksi kata penting,

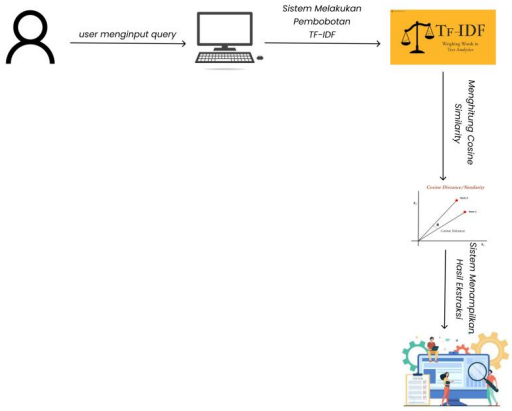
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

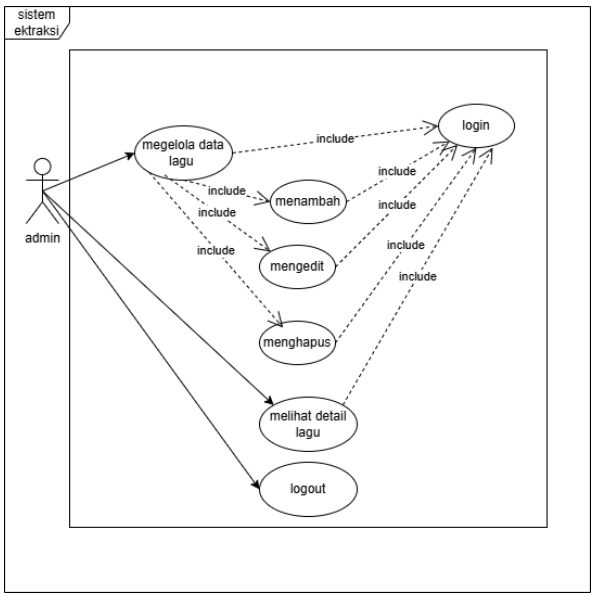
menampilkan skor kemiripan antarlirik, menampilkan tema atau topik utama berdasarkan kata kunci, menampilkan ringkasan makna lagu, serta menampilkan informasi budaya seperti daerah asal, bahasa, dan nilai budaya yang terkandung. Kebutuhan non-fungsional dirangkum pada Tabel 1.

Tabel 1. Kebutuhan non-Fungsional Sistem

Parameter	Kebutuhan
Usability	Antarmuka sederhana, responsif, dan mudah digunakan oleh pengguna umum tanpa pelatihan teknis.
Efficiency	Sistem mampu memproses dan menampilkan hasil pencarian dalam waktu kurang dari 3 detik per permintaan.
Reliability	Sistem mampu beroperasi dengan ketersediaan 95% tanpa crash serta menangani minimal 100 pencarian per hari.



Gambar 1. Alur Sistem yang Diusulkan



Gambar 2. Use Case Diagram Sistem

Text Preprocessing

Sebelum pembobotan TF-IDF, setiap dokumen lirik melewati tahap text preprocessing. Tahapan tersebut meliputi case folding untuk menyeragamkan huruf menjadi huruf kecil, tokenizing untuk memecah teks menjadi token kata, stopword removal untuk menghapus kata umum yang tidak berkontribusi besar terhadap makna, dan stemming untuk mengembalikan kata ke bentuk dasarnya. Hasil preprocessing menjadi dasar pembentukan vektor dokumen dan vektor query.

Pembobotan dan Pencarian Dokumen

Setiap term hasil preprocessing diberi bobot menggunakan TF-IDF. Ketika pengguna memasukkan query berupa potongan lirik atau kata kunci, sistem melakukan preprocessing yang sama terhadap query, kemudian menghitung nilai Cosine Similarity antara vektor query dan seluruh vektor dokumen lirik dalam dataset. Lagu dengan nilai kemiripan tertinggi ditampilkan sebagai hasil utama atau Top 1.

Ekstraksi Elemen Linguistik

Setelah lagu paling relevan diperoleh, sistem mengekstraksi elemen linguistik menggunakan POS Tagging berbasis Stanza dan aturan linguistik berbasis kaidah. Elemen yang diekstraksi meliputi nomina, verba, adjektiva, pronomina, predikat, aksi, subjek, objek, negasi, partikel, sasaran, dan makna akhir berdasarkan input pengguna. Hasil ekstraksi kemudian ditampilkan dalam halaman hasil agar pengguna dapat memahami struktur dan makna lirik secara lebih jelas.

Pengujian

Pengujian dilakukan melalui tiga skema. Pertama, blackbox testing digunakan untuk memastikan seluruh fungsi utama berjalan sesuai output yang diharapkan. Kedua, Top 1 Accuracy digunakan untuk mengukur ketepatan sistem dalam menampilkan lagu paling relevan terhadap 10 query uji. Ketiga, evaluasi ekstraksi linguistik dilakukan dengan membandingkan hasil sistem terhadap anotasi manual menggunakan Precision, Recall, dan F1-Score.

Top 1 Accuracy = (jumlah query benar / total query) x 100% … (3)

Precision = TP / (TP + FP) …. (4)

Recall = TP / (TP + FN) … (5)

F1-Score = (2 x Precision x Recall) / (Precision + Recall) … (6)

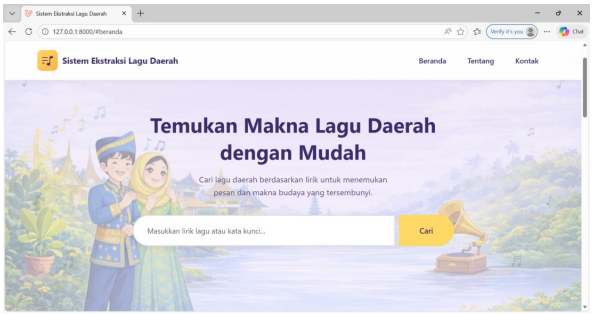
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

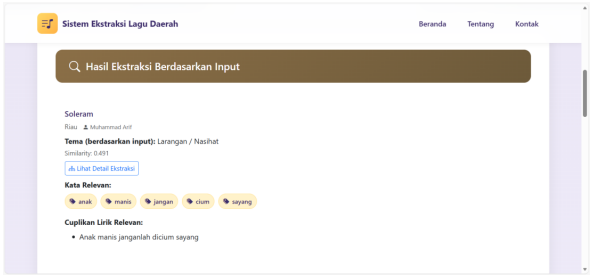
HASIL DAN PEMBAHASAN

Antarmuka Sistem

Sistem ekstraksi informasi lirik lagu daerah berhasil dibangun sebagai aplikasi web. Halaman beranda menyediakan kolom pencarian bagi pengguna untuk memasukkan lirik lagu atau kata kunci. Setelah query diproses, halaman hasil menampilkan judul lagu, daerah asal, tema berdasarkan input, skor Cosine Similarity, kata relevan hasil ekstraksi, cuplikan lirik, serta detail elemen linguistik yang dapat dibuka oleh pengguna.



Gambar 3. Tampilan Halaman Beranda Sistem



Gambar 4. Tampilan Halaman Hasil Ekstraksi

Hasil Pengujian Blackbox

Hasil pengujian blackbox menunjukkan bahwa enam skenario utama, yaitu input query lirik, input kosong, proses TF-IDF, proses Cosine Similarity, ekstraksi linguistik, dan penampilan hasil, seluruhnya berjalan sesuai output yang diharapkan. Ringkasan hasil pengujian ditunjukkan pada Tabel 2.

Tabel 2. Hasil Pengujian *Blackbox*

No	Skenario Uji	Output yang Diharapkan	Status
1	Input query lirik	Query diproses dan hasil ditampilkan	Berhasil
2	Input kosong	Sistem menolak input dan tidak menampilkan hasil	Berhasil
3	Proses pembobotan TF-IDF	Bobot TF-IDF terbentuk	Berhasil
4	Proses perhitungan Cosine Similarity	Nilai similarity muncul	Berhasil
5	Ekstraksi linguistik	Elemen linguistik tampil	Berhasil
6	Tampilkan hasil	Output tampil pada halaman hasil	Berhasil

Hasil Pengujian Top 1 Accuracy

Pengujian retrieval dilakukan terhadap 10 query berupa potongan lirik dari 10 lagu daerah. Seluruh query berhasil menampilkan lagu referensi yang benar pada peringkat pertama. Nilai Cosine Similarity berkisar antara 0,241 sampai 0,508, sehingga diperoleh Top 1 Accuracy sebesar 100%. Hasil ini menunjukkan bahwa representasi TF-IDF dan perhitungan Cosine Similarity mampu membedakan lirik lagu daerah pada dataset uji yang digunakan.

Tabel 3. Hasil Pengujian Top 1 Accuracy

No	Query	Lagu Referensi	Nilai CS	Status
1	Anak manis janganlah dicium sayang	Soleram (Riau)	0,501	Benar
2	Pisangku belum masak	Ampar-ampar Pisang (Kalimantan Selatan)	0,281	Benar
3	Bungong teuleubeh teuleubeh indah lagoi ma	Bungong Juempo (Aceh)	0,407	Benar
4	Mana dimana anak kambing saya	Anak Kambing Saya (Timor)	0,302	Benar
5	Angin mamiri ku pasang	Angin Mamiri (Sulawesi Selatan)	0,455	Benar
6	Teemi nokibe kubano ko bombe ko yuma no bungo aweade	Yamko Rambe Yamko (Papua Barat)	0,474	Benar
7	Luruihlah jalan Payakumbuah	Ayam Den Lapeh (Sumatra Barat)	0,241	Benar
8	Ayam hitam telurnya putih mencari makan dipinggir kali	Ayo Mama (Maluku)	0,437	Benar
9	Burung kakak tua hinggap di jendela	Burung Kakak Tua (Maluku)	0,269	Benar
10	Desaku yang kucintai lewo inak jadi go	Desaku (Nusa Tenggara Timur)	0,508	Benar

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

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Tabel 4. Rekapitulasi Akurasi Top 1

Total Query	Benar	Salah	Top 1 Accuracy
10	10	0	100%

Hasil Evaluasi Ekstraksi Linguistik

Evaluasi ekstraksi linguistik dilakukan dengan membandingkan hasil POS Tagging dan rule-based extraction terhadap anotasi manual pada 10 query uji. Elemen yang dievaluasi mencakup nomina, verba, adjektiva, pronomina, preposisi, partikel, subjek, objek, predikat, aksi, negasi, dan sasaran. Rekapitulasi confusion matrix menunjukkan 31 True Positive, 10 False Negative, dan 11 False Positive.

Tabel 5. Rekapitulasi Confusion Matrix

Metrik	Total
True Positive (TP)	31
False Negative (FN)	10
False Positive (FP)	11

Tabel 6. Hasil Akhir Evaluasi Ekstraksi Linguistik

Metrik	Nilai
Precision	0,74 (74%)
Recall	0,76 (76%)
F1-Score	0,75 (75%)

Berdasarkan hasil perhitungan, nilai Precision 0,74 menunjukkan bahwa sebagian besar elemen yang diekstraksi sistem sesuai dengan anotasi manual. Nilai Recall 0,76 menunjukkan bahwa sistem mampu menangkap sebagian besar elemen yang seharusnya terdeteksi. F1-Score 0,75 menunjukkan keseimbangan yang cukup baik antara ketepatan dan kelengkapan ekstraksi, meskipun ketiga nilai tersebut belum mendekati rentang tinggi (di atas 0,85) yang umum dijadikan acuan pada tugas ekstraksi informasi berbasis teks umum. Penelusuran lebih lanjut terhadap kasus yang keliru menunjukkan bahwa akar penyebab menurunnya ketiga metrik tersebut secara konsisten kembali pada satu faktor utama, yaitu keterbatasan cakupan kosakata bahasa daerah pada model POS Tagging berbasis Stanza yang dilatih terutama dengan korpus bahasa Indonesia baku. Kesalahan false positive umumnya terjadi ketika sistem menandai kata sebagai elemen tertentu yang tidak ada pada ground truth; pola ini banyak dijumpai pada kata daerah yang secara morfologis mirip dengan kelas kata dalam bahasa Indonesia baku tetapi memiliki makna atau fungsi gramatikal berbeda dalam konteks lirik, sehingga model salah mengklasifikasikannya. Sebaliknya, false negative banyak terjadi pada verba atau pronomina yang berasal dari kosakata daerah yang tidak umum dan tidak dikenali sama sekali oleh model (out-of-vocabulary), sehingga kata tersebut tidak diberi label kelas kata apa pun. Sebagian kesalahan lain juga bersumber dari keterbatasan pola rule-based extraction yang disusun berdasarkan struktur kalimat Subjek-Predikat-Objek baku, sedangkan sejumlah lirik lagu daerah menggunakan susunan kata yang lebih bebas atau elipsis (penghilangan subjek/predikat) sesuai kaidah bahasa daerah masing-masing, sehingga aturan tersebut tidak selalu berhasil mengenali batas antar-elemen secara tepat. Dengan demikian, nilai Precision, Recall, dan F1-Score yang bertahan pada kisaran 0,75 dapat dipandang sebagai representasi langsung dari kesenjangan cakupan kosakata dan kaidah gramatikal antara bahasa daerah dan model bahasa Indonesia baku yang digunakan sebagai basis POS Tagging, dan bukan semata-mata kelemahan pada TF-IDF maupun Cosine Similarity yang keduanya justru terbukti sangat andal pada tahap retrieval (Top 1 Accuracy 100%). Temuan ini menegaskan perlunya pengembangan kamus atau korpus anotasi bahasa daerah yang lebih luas serta penyesuaian (fine-tuning) model POS Tagging pada data kebahasaan daerah untuk meningkatkan performa ekstraksi pada penelitian lanjutan.

Secara keseluruhan, hasil pengujian menunjukkan bahwa kombinasi TF-IDF, Cosine Similarity, POS Tagging, dan rule-based extraction dapat digunakan untuk membangun sistem ekstraksi informasi lirik lagu daerah. Pengujian retrieval pada penelitian ini telah mencakup lagu dari sepuluh provinsi yang tersebar di Sumatra, Kalimantan, Sulawesi, Nusa Tenggara, Maluku, dan Papua, sehingga memberikan gambaran awal kemampuan sistem pada keragaman rumpun bahasa daerah di Indonesia. Meskipun demikian, cakupan satu query per lagu pada sepuluh provinsi tersebut masih jauh dari cukup untuk mewakili ratusan bahasa dan dialek daerah yang ada di Indonesia, sehingga capaian Top-1 Accuracy 100% pada penelitian ini perlu dimaknai sebagai hasil pada dataset uji berskala kecil, bukan sebagai generalisasi kinerja sistem pada seluruh ragam lirik lagu daerah. Perluasan dataset uji secara sistematis, misalnya dengan menambah jumlah query per lagu, menambah jumlah lagu per provinsi, serta memasukkan lagu daerah dari provinsi yang belum terwakili, diperlukan pada penelitian lanjutan untuk menguji kemampuan sistem secara lebih menyeluruh pada beragam bahasa dan lirik lagu daerah di Indonesia. Selain perluasan dataset, pengembangan kamus linguistik lokal, evaluasi dengan metode pembanding seperti BM25, Word2Vec, atau Sentence-

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

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

BERT, serta evaluasi yang lebih komprehensif juga menjadi arah penting bagi penelitian selanjutnya.

KESIMPULAN

Penelitian ini berhasil mengembangkan sistem ekstraksi informasi lirik lagu daerah menggunakan TF-IDF dan Cosine Similarity sebagai metode pembobotan dan pengukuran kemiripan teks, serta POS Tagging dan rule-based extraction untuk mengidentifikasi elemen linguistik. Sistem mampu menampilkan hasil pencarian lagu paling relevan, skor kemiripan, kata relevan, cuplikan lirik, serta elemen linguistik seperti nomina, verba, adjektiva, pronomina, predikat, aksi, subjek, objek, negasi, dan makna akhir.

Hasil pengujian menunjukkan seluruh fungsi utama sistem berjalan sesuai skenario blackbox. Evaluasi retrieval terhadap 10 query uji yang mencakup sepuluh provinsi menghasilkan Top-1 Accuracy sebesar 100%, sedangkan evaluasi ekstraksi linguistik memperoleh Precision 0,74, Recall 0,76, dan F1-Score 0,75 yang utamanya dipengaruhi oleh keterbatasan cakupan kosakata bahasa daerah pada model POS Tagging. Dengan demikian, integrasi pendekatan statistik dan linguistik dapat digunakan secara efektif untuk mendukung ekstraksi informasi lirik lagu daerah, dokumentasi budaya, dan pembelajaran bahasa daerah, meskipun hasil ini masih bersifat awal dan belum dapat digeneralisasikan pada seluruh ragam bahasa daerah di Indonesia. Penelitian selanjutnya disarankan memperluas dataset dan jumlah query uji secara sistematis pada lebih banyak bahasa dan lirik lagu daerah, menambahkan lexicon bahasa daerah, membandingkan kinerja TF-IDF dan Cosine Similarity secara terukur dengan metode lain seperti BM25, Word2Vec, dan Sentence-BERT, serta menguji model berbasis machine learning atau deep learning untuk meningkatkan akurasi ekstraksi.

DAFTAR PUSTAKA

Al Rasyid, R., Handayani, D., & Ningsih, U. (2024). Penerapan algoritma TF-IDF dan Cosine Similarity untuk query pencarian pada dataset destinasi wisata. *Jurnal Teknologi Informasi dan Komunikasi, 8*(1). https://doi.org/10.47111/jti.v20i1.23986

Bahri, S. (2020). Aplikasi pencarian bahan pustaka di perpustakaan menggunakan metode Vector Space Model. *JIMP: Jurnal Informatika Merdeka Pasuruan, 5*.

Bakiyev, B. (2022). Method for determining the similarity of text documents for the Kazakh language, taking into account synonyms: Extension to TF-IDF. In *2022 International Conference on Smart Information Systems and Technologies (SIST)*. IEEE. https://doi.org/10.1109/SIST54437.2022.9945747

Djunaidi, K., et al. (2024). Pencarian fungsional obat menggunakan algoritma TF-IDF dan Cosine Similarity. *PETIR, 17*(1). https://doi.org/10.33322/petir.v17i1.2341

Fajriansyah, M., Adikara, P. P., & Widodo, A. W. (2021). Sistem Rekomendasi Film Menggunakan Content Based Filtering. *Jurnal Pengembangan Teknologi dan Ilmu Komputer, 5*(6), 2188-2199.

Findawati, O. Y., Muhammad, M. M., Rosid, A., & Kom, S. (2020). *Buku ajar text mining*. UMSIDA Press.

Kristianto, F. S. (2020). Ekstraksi Informasi Meme-Internet Berbahasa Indonesia Dengan Mesin Pencarian. *Jurnal Indonesia Sosial Teknologi, 1*(5).

Mulyani, S. (2016). *Metode analisis dan perancangan sistem*. Abdi Sistematika.

Pribadi, F., & Munir, R. (2021). Sistem pencarian similaritas judul tugas akhir menggunakan metode TF-IDF. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi, 7*(2), 58. https://doi.org/10.24014/coreit.v7i2.15338

Rosa, A. S., & Shalahuddin, M. (2016). *Rekayasa perangkat lunak*. Informatika.

Santoso, G., Marcellia, L., Ramadhani, D. N., & Zabidi, S. F. (2023). Lagu daerah dan lagu wajib nasional untuk meningkatkan gairah mengisi kemerdekaan Indonesia, *Jurnal Pendidikan, 2*(02), 286-293.

Situmorang, J. S., Adikara, P. P., & Ratnawati, D. E. (2021). Rekomendasi lagu berdasarkan lirik lagu menggunakan metode N-Gram dan Cosine Similarity. *Jurnal Pengembangan Teknologi dan Ilmu Komputer, 5*(6), 2279-2288.

Supiyanto, S., & Sriyono, S. (2023). Metode Cosine Similarity Untuk Mendeteksi Kemiripan Pada Dokumen Teks. *SAINS J. MIPA Dan Pengajarannya, 1*(1), 001-007.

Timur, A. Y. (2025). Sistem rekomendasi lagu Indonesia menggunakan metode content-based filtering dan Cosine Similarity. *Jurnal Informatika dan Teknik Elektro Terapan, 13*(1). https://doi.org/10.23960/jitet.v13i1.5949

Wibowo, I. S., Witanti, A., & Susilawati, I. (2024). Keyword extraction judul berita online di Indonesia menggunakan metode TF-IDF. *Jurnal Teknik Informatika dan Sistem Informasi, 11*(1).

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