

Penerapan Sistem Pencarian Dokumen Berdasarkan Frasa di Abstrak Perpustakaan Digital Menggunakan Algoritma BM25 dan Word2Vec

Nur Fajar^{*1}, Fahrir Irhamna Rachman², Ida Mulyadi³

^{1,2,3}Informatika, Universitas Muhammadiyah Makassar, Makassar, 90221, Indonesia

e-mail koresponden: 105841101121@student.unismuh.ac.id^{*1}, fachrim141020@unismuh.ac.id², idamulyadi@unismuh.ac.id³

Received: Agustus 01, 2025; Accepted: Agustus 01, 2025; Published: 30 September, 2025

Abstrak

Perkembangan perpustakaan digital menyebabkan meningkatnya volume abstrak dokumen sehingga menuntut metode pencarian yang akurat untuk menemukan buku relevan. Penelitian ini mengusulkan penerapan sistem pencarian berbasis frasa pada abstrak dengan menggabungkan algoritma BM25 dan Word2Vec untuk meningkatkan relevansi hasil. Dataset terdiri dari 500 abstrak skripsi yang dipreproses (lowercasing, tokenisasi, stopwords removal); model Word2Vec dilatih dengan arsitektur skip-gram (vector_size=100, window=5, epochs=50) dan BM25 diinisialisasi pada representasi token dokumen. Skor BM25, Word2Vec (cosine similarity) dan TF-IDF dinormalisasi lalu digabungkan (rata-rata) untuk pemeringkatan akhir. Evaluasi dilakukan menggunakan metrik Precision, Recall dan F1-Score pada beberapa query uji. Hasil menunjukkan peningkatan performa pada banyak query (rata-rata F1 ≈ 0.80) dengan beberapa kasus mencapai nilai sempurna (1.00), meskipun ada variabilitas antar tipe query. Temuan ini menegaskan bahwa penggabungan pencocokan lesikal BM25 dan representasi semantik Word2Vec dapat meningkatkan relevansi pencarian; pengembangan lanjutan direkomendasikan pada metode penggabungan skor dan perluasan korpus.

Kata kunci: Pencarian informasi, bm25, word2vec, tf-idf.

Abstract

The development of digital libraries has led to an increase in the volume of document abstracts, thus demanding accurate search methods to find relevant books. This study proposes the implementation of a phrase-based search system on abstracts by combining the BM25 and Word2Vec algorithms to improve the relevance of the results. The dataset consists of 500 thesis abstracts that were preprocessed (lowercasing, tokenization, stopwords removal); the Word2Vec model was trained with a skip-gram architecture (vector_size=100, window=5, epochs=50) and BM25 was initialized on the document token representation. The BM25, Word2Vec (cosine similarity) and TF-IDF scores were normalized and then combined (averaged) for the final ranking. Evaluation was performed using Precision, Recall and F1-Score metrics on several test queries. The results show improved performance on many queries (average F1 ≈ 0.80) with some cases achieving a perfect score (1.00), although there is variability between query types. These findings confirm that combining BM25 lexical matching and Word2Vec semantic representation can improve search relevance; further development of the score fusion method and corpus expansion is recommended.

Keyword: Information retrieval, bm25, word2vec, tf-idf.

1. Pendahuluan

Perkembangan teknologi informasi telah mendorong digitalisasi koleksi perpustakaan, sehingga perpustakaan digital kini menyimpan ribuan hingga jutaan abstrak buku yang dapat diakses secara elektronik [1], [2]. Namun, pertumbuhan volume data ini menimbulkan tantangan baru dalam proses temu kembali informasi, khususnya dalam membantu pengguna menemukan buku yang paling relevan dengan kebutuhan mereka secara cepat dan akurat [3], [4].

Sistem pencarian konvensional, seperti BM25, mengandalkan pencocokan kata kunci literal dan menghitung relevansi dokumen berdasarkan frekuensi istilah. Meskipun BM25 telah terbukti efektif dan banyak diadopsi dalam berbagai sistem temu kembali informasi, pendekatan ini memiliki

keterbatasan dalam menangkap makna semantik. Akibatnya, hasil pencarian sering kali kurang relevan, terutama untuk query yang kompleks atau mengandung frasa [5].

Sebagai solusi, metode representasi kata berbasis word embedding seperti Word2Vec mulai banyak digunakan. Word2Vec mampu merepresentasikan kata ke dalam vektor numerik yang menangkap hubungan semantik antar kata, sehingga sistem dapat mengenali makna serupa meskipun kata atau frasa yang digunakan berbeda secara literal [6].

Penelitian ini mengusulkan penerapan algoritma BM25 dan Word2Vec dalam sistem pencarian buku berbasis frasa di abstrak. Pendekatan ini diharapkan dapat menggabungkan keunggulan pencocokan literal BM25 dan pemahaman semantik Word2Vec, sehingga meningkatkan relevansi dan akurasi hasil pencarian [7]. Evaluasi sistem pencarian akan dilakukan menggunakan metrik Presisi, Recall dan F1-Score untuk menilai keunggulan integrasi kedua metode [8].

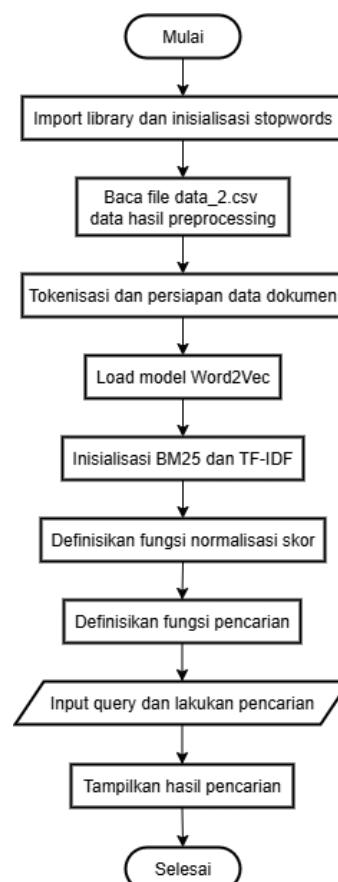
2. Metode Penelitian

2.1. Alat dan Bahan

Alat dan bahan yang digunakan meliputi:

- a. Hardware (Perangkat Keras): Laptop dengan spesifikasi memadai untuk memproses data dan pelatihan model Word2Vec.
- b. Software (Perangkat Lunak):
 - Bahasa pemrograman Python.
 - Google Colab.
 - Database berupa file CSV.
 - Library Python: pandas, gensim, rank_bm25, nltk, re, IPython.display, scikit-learn (TfidfVectorizer), numpy.
- c. Sistem operasi berupa windows.
- d. Dataset: Kumpulan abstrak akademik berupa skripsi.

2.2. Perancangan Sistem

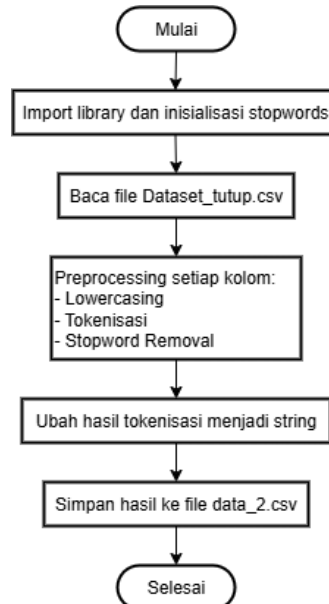


Gambar 1. Flowchart Perancangan Sistem

2.3 Pra-pemrosesan

Pemrosesan dilakukan dilakukan pada abstrak setiap skripsi dan bekerja saat:

- Inisialisasi pencari: Semua abstrak skripsi diproses ketika dijalankan.
 - Pencarian: Query dari pengguna juga diproses sebelum dilakukan pencocokan dengan data.
- a. Pembersihan dataset

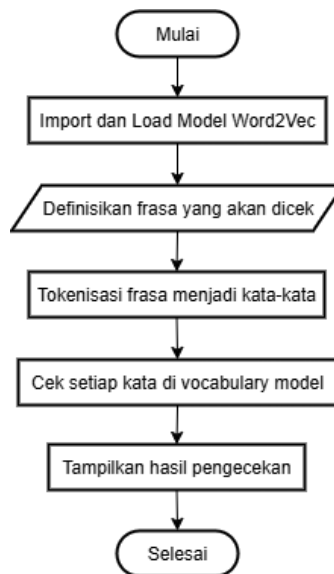


Gambar 2. Pembersihan Dataset

b. Training dataset



Gambar 3. Training Dataset

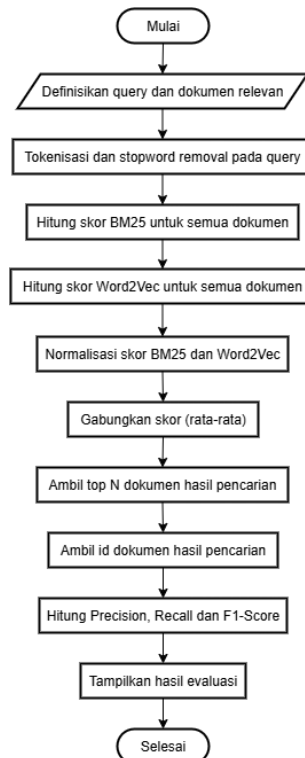


Gambar 4. Pengecekan Vocabulary Model Word2Vec

2.4. Pengujian Sistem

Melakukan pencarian dengan menginput query yang ingin di uji. Untuk mendapatkan daftar id dokumen sebagai acuan pengujian, disajikan proses pencarian pada 'Gambar 1. Flowchart Perancangan Sistem'. Setelah hasil pencarian ditampilkan berdasarkan query yang dicari. Selanjutnya, menginput query serta id yang ditampilkan.

Proses ini bertujuan untuk mengukur efektivitas sistem pencarian dalam menemukan dokumen yang relevan dengan mempertimbangkan baik aspek frekuensi kata maupun kemiripan makna secara semantik. Setiap langkah mulai dari tokenisasi dan penghapusan stopwords pada query, perhitungan skor BM25 dan Word2Vec, normalisasi serta penggabungan skor, hingga penghitungan nilai Precision, Recall dan F1-Score dilakukan secara sistematis agar hasil evaluasi yang diperoleh dapat menggambarkan performa metode gabungan secara objektif.



Gambar 5. Proses Evaluasi Pencarian BM25+Word2Vec

3. Hasil dan Pembahasan

3.1. Dataset

Setiap dokumen hanya mempertahankan judul dan isi abstrak.

Tabel 1. Dataset

No	id	Judul	Abstrak
1	1	PENGARUH PENGGUNAAN MODEL PEMBELAJARAN COURSE REVIEW HORAY TERHADAP HASIL BELAJAR SISWA PADA MATA PELAJARAN PKN DI KELAS III SD NEGERI BONTOMANA KECAMATAN BAROMBONG KABUPATEN GOWA	Muhammad Taufiq. JS. 2019. Pengaruh Penggunaan Model Pembelajaran Course Review Horay Terhadap Hasil Belajar Siswa Pada Mata Pelajaran Pendidikan Kewarganegaraan Kelas III SD Negeri Bontomanai K. Skripsi. Jurusan Pendidikan Guru Sekolah D Muhammad Safril, 105251101518.
2	2	ANALISIS POTENSI GHARAR DAN RIBA DALAM PRAKTIK JUAL BELI BAYAR SETELAH PANEN DI DESA ONGKOE KECAMATAN BELAWA KABUPATEN WAJO	Analisis Potensi Gharar dan Riba dalam Praktik Jual Beli Bayar Setelah Panen di Desa Ongkoe Kecamatan Belawa Kabupaten BENAZIR FIRDAUS. Perancangan
3	500	PERANCANGAN KAMPUS FAKULTAS KEDOKTERAN UNIVERSITAS MUHAMMADIYAH MAKASSAR DENGAN KONSEP BIOPHILIC DESIGN	Kampus Fakultas Kedokteran Universitas Muhammadiyah Makassar dengan Konsep Biophilic Design. (dibimbin)

Tabel 2. Hasil Pembersihan Dataset

No	id	Judul	Abstrak
1	1	pengaruh penggunaan model pembelajaran course review horay hasil belajar siswa mata pelajaran pkn kelas iii sd negeri bontomanaik kecamatan barombong kabupaten gowa	muhammad taufiq. js. 2019. pengaruh penggunaan model pembelajaran course review horay hasil belajar siswa mata pelajaran pendidikan kewarganegaraan kelas iii sd negeri Muhammad Safril, 105251101518.
2	2	analisis potensi gharar riba praktik jual beli bayar panen desa ongkoe kecamatan belawa kabupaten wajo	analisis potensi gharar riba praktik jual beli bayar panen desa ongkoe kecamatan belawa kabupaten
3	500	perancangan kampus fakultas kedokteran universitas muhammadiyah makassar konsep biophilic design	benazir firdaus. perancangan kampus fakultas kedokteran universitas muhammadiyah makassar konsep biophilic design. (dibimbin)

3.2. Pelatihan

pelatihan model Word2Vec menggunakan data abstrak skripsi yang telah melalui tahap preprocessing [9]. Data yang digunakan diambil dari hasil preprocessing, kemudian setiap abstrak diubah menjadi kumpulan kata melalui proses tokenisasi sederhana [10]. Tokenisasi ini bertujuan untuk memecah setiap kalimat abstrak menjadi satuan kata agar dapat diproses lebih lanjut oleh algoritma Word2Vec.

seluruh kumpulan kata dari abstrak skripsi digunakan sebagai input untuk melatih model Word2Vec. Model ini dibangun dengan parameter vektor berukuran 100, window 5 dan menggunakan metode skip-gram agar mampu menangkap hubungan semantik antar kata secara lebih baik [11], [12]. Proses pelatihan dilakukan selama 50 epoch untuk memastikan model dapat mempelajari representasi kata secara optimal dari data yang tersedia [13].

Model ini berfungsi untuk mengukur kemiripan makna antara query dan dokumen, sehingga dapat meningkatkan relevansi hasil pencarian. Dengan adanya model Word2Vec, sistem pencarian dokumen dapat memahami konteks dan hubungan antar kata secara lebih mendalam, tidak hanya berdasarkan kemiripan kata secara literal.

Setelah model Word2Vec dilatih, selanjutnya adalah melakukan pengujian terhadap vocabulary yang telah dipelajari oleh model. Pengujian ini dilakukan dengan cara mendefinisikan beberapa frasa atau query yang relevan dengan topik penelitian, kemudian setiap kata dalam frasa tersebut dicek keberadaannya di dalam vocabulary model Word2Vec [14], [15].

Tabel 3. Pengecekan Vocabulary Model Word2Vec

No	Query	Output
1	pengaruh kualitas pelayanan terhadap kepuasan pelanggan	'pengaruh' ada di vocabulary. 'kualitas' ada di vocabulary. 'pelayanan' ada di vocabulary. 'terhadap' TIDAK ada di vocabulary. 'kepuasan' ada di vocabulary. 'pelanggan' ada di vocabulary.
2	strategi pengembangan usaha kecil menengah di desa	'strategi' ada di vocabulary. 'pengembangan' ada di vocabulary. 'usaha' ada di vocabulary. 'kecil' TIDAK ada di vocabulary. 'menengah' ada di vocabulary. 'di' TIDAK ada di vocabulary. 'desa' ada di vocabulary.
3	analisis kinerja keuangan perusahaan selama pandemi	'analisis' ada di vocabulary. 'kinerja' ada di vocabulary. 'keuangan' ada di vocabulary. 'perusahaan' ada di vocabulary. 'selama' TIDAK ada di vocabulary. 'pandemi' ada di vocabulary.

terdapat beberapa kata yang tidak ditemukan di dalam vocabulary, seperti "dalam", "terhadap", "selama", "kecil" dan "di". Kata-kata ini umumnya merupakan kata hubung atau kata umum yang sering dihapus pada tahap preprocessing sebagai bagian dari stopword removal. Ketidakhadiran kata-kata tersebut di vocabulary tidak terlalu mempengaruhi hasil pencarian, karena kata-kata tersebut memang tidak memiliki makna khusus dalam konteks pencarian dokumen berbasis frasa.

3.3. Pengujian

Sebelum dilakukan pengujian, terlebih dahulu ditentukan query yang akan diuji beserta daftar id dokumen relevan sebagai acuan penilaian.

Tabel 4. Query dan id Relevant BM25+Word2Vec

No	Query	Id Relevant
1	pengaruh kualitas produk dan pelayanan terhadap kepuasan konsumen	65, 67, 53, 66, 450
2	strategi guru dalam meningkatkan kemampuan baca tulis Al-Qur'an	84, 111, 300, 282, 443
3	analisis faktor-faktor yang mempengaruhi produksi bawang merah	158, 230, 362, 72, 76

Pada "Tabel 5. Hasil Pengujian BM25 dan Word2Vec" menampilkan hasil evaluasi sistem pencarian dokumen berdasarkan query. Setiap query diuji dengan sistem pencarian BM25+Word2Vec untuk memperoleh dokumen-dokumen yang paling relevan, yang kemudian dibandingkan dengan daftar id dokumen relevan sebagai acuan penilaian. Nilai Precision, Recall

dan F1-Score dihitung untuk setiap query guna mengukur tingkat akurasi dan efektivitas sistem dalam menemukan dokumen yang sesuai.

Tabel 5. Hasil Pengujian BM25 dan Word2Vec

No	Query	Id Relevant	Precision	Recall	F1-Score
1	pengaruh kualitas produk dan pelayanan terhadap kepuasan konsumen	65, 67, 53, 66, 450	0.80	0.80	0.80
2	strategi guru dalam meningkatkan kemampuan baca tulis Al-Qur'an	84, 111, 300, 282, 443	0.80	0.80	0.80
3	analisis faktor-faktor yang mempengaruhi produksi bawang merah	158, 230, 362, 72, 76	1.00	1.00	1.00

Berdasarkan tabel di atas pencarian dokumen menggunakan gabungan algoritma BM25 dan Word2Vec, terlihat bahwa sistem mampu memberikan performa yang cukup baik pada sebagian besar query. Nilai Precision, Recall dan F1-Score umumnya berada pada angka 0.80, yang menunjukkan bahwa sebagian besar dokumen relevan berhasil ditemukan. Pada query "analisis faktor-faktor yang mempengaruhi produksi bawang merah", sistem mampu mencapai nilai evaluasi sempurna (1.00), menandakan pencarian yang sangat akurat.

4. Kesimpulan

Sistem pencarian frasa pada abstrak yang menggabungkan BM25 (kuat pada kata kunci) dan Word2Vec (menangkap kemiripan makna) terbukti meningkatkan relevansi dan akurasi (Precision, Recall, F1 lebih baik); untuk optimalisasi lanjut disarankan memperluas data (full text/metadata), mengeksplorasi algoritma atau metode penggabungan skor yang lebih baik.

Referensi

- [1] P. S. Kolhe, "The Impact of Digital Libraries on Information Access and Retrieval: A Comprehensive Analysis," *Int. Adv. J. Eng. Sci. Manag.*, vol. 15, no. II, pp. 329–334, 2021.
- [2] D. S. K. Bhat, "The Role of Digital Libraries in Enhancing Information Accessibility: A Study," *Int. J. Multidiscip. Res.*, vol. 6, no. 6, pp. 1–8, 2024, doi: 10.36948/ijfmr.2024.v06i06.32585.
- [3] G. M. Di Nunzio, "Focused Issue on Digital Library Challenges to Support the Open Science Process," *Int. J. Digit. Libr.*, vol. 24, no. 4, pp. 185–189, 2023, doi: 10.1007/s00799-023-00388-9.
- [4] M. Alfitra Haskar, Muhyiddin A M Hayat, S.Kom., M.T, Rizky Yuslian Bakti, S.T., "ANALISIS PENERAPAN ALGORITMA JACCARD SIMILARITY UNTUK REKOMENDASI BUKU BERDASARKAN KATA KUNCI DAN SINONIMNYA," vol. 1, no. 1, pp. 102–107, 2025.
- [5] J. Rayo, R. de la Rosa, and M. Garrido, "A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts," pp. 31–35, 2025, [Online]. Available: <http://arxiv.org/abs/2502.16767>
- [6] D. Jurgens, "Learning about Word Vector Representations and Deep Learning through Implementing Word2vec," *Teach. NLP 2021 - Proc. 5th Work. Teach. Nat. Lang. Process.*, pp. 108–111, 2021, doi: 10.18653/v1/2021.teachingnlp-1.19.
- [7] X. Li, "BM X: Entropy-weighted Similarity and Semantic-enhanced Lexical Search *," no. 1, pp. 1–12, 2024.
- [8] Z. Edrees, "Hybrid Method to Enhance Effectiveness and Efficiency for Information Retrieval in Web Search Engines," 2023.
- [9] S. A. Savittri, A. Amalia, and M. A. Budiman, "A relevant document search system model using word2vec approaches," *J. Phys. Conf. Ser.*, vol. 1898, no. 1, 2021, doi: 10.1088/1742-6596/1898/1/012008.
- [10] P. P. Dokumen, "Integrasi word embeddings dan inverse book frequency dalam pembobotan term untuk peningkatan pencarian dokumen," vol. 9, no. 4, pp. 2529–2537, 2024.
- [11] I. R. Wijayanto, I. Cholissodin, and Y. A. Sari, "Pengaruh Metode Word Embedding dalam Vector Space Model pada Pemerolehan Informasi Materi IPA Siswa SMP," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 3, pp. 950–959, 2021, [Online]. Available: <http://j-ptiik.ub.ac.id>
- [12] G. M. Rosa, R. C. Rodrigues, R. Lotufo, and R. Nogueira, Yes, *BM25 is a Strong Baseline for Legal Case Retrieval*, vol. 1, no. 1. Association for Computing Machinery, 2021. [Online].

-
- Available: <http://arxiv.org/abs/2105.05686>
- [13] P. N. Hutapea, P. P. Adikara, and Y. A. Sari, "Pemerolehan Informasi pada Alkitab Berbahasa Indonesia menggunakan Metode BM25 dan Query Expansion Word2Vec," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, pp. 1844–1852, 2021.
- [14] S. Vilella and G. Ruffo, "(De)-Indexing and the Right to be Forgotten," pp. 1–17, 2025, [Online]. Available: <http://arxiv.org/abs/2501.03989>
- [15] P. M. Mekontchou, A. Fotsoh, B. Batchakui, and E. Ella, "Information Retrieval in long documents: Word clustering approach for improving Semantics," no. 2018, 2023, [Online]. Available: <http://arxiv.org/abs/2302.10150>