

ANALISIS KOMPARATIF JACCARD DAN COSINE SIMILARITY UNTUK PENCARIAN AYAT ALKITAB

Yusuf Imanuel¹, Stanislaus J. Pinastika², Harry Soepandi*³, Katarina Lebai⁴

¹²Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Jember, Indonesia

³⁴Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Jember, Indonesia

*e-mail: harrysupandi@unej.ac.id

Abstrak

Pencarian ayat Alkitab yang relevan berdasarkan tema perasaan merupakan tantangan signifikan bagi pengguna, disebabkan oleh banyaknya jumlah ayat dan variasi bahasa yang kompleks. Sistem pencarian konvensional seringkali tidak mampu menangkap nuansa emosional yang diinginkan pengguna. Penelitian ini bertujuan untuk merancang, mengimplementasikan, dan membandingkan performa sistem temu kembali informasi ayat Alkitab menggunakan metode Jaccard Similarity dan Cosine Similarity. Data yang digunakan terdiri dari 31.102 ayat Alkitab dari website SABDA yang melalui tahap preprocessing teks. Metode Cosine Similarity memanfaatkan pembobotan TF-IDF, sedangkan Jaccard Similarity berbasis himpunan kata. Hasil pencarian kemudian dikelompokkan menggunakan algoritma K-Means Clustering untuk memisahkan hasil relevan dan kurang relevan. Evaluasi dilakukan menggunakan metrik Precision, Recall, dan F1-Score untuk metode similaritas, serta Davies-Bouldin Index (DBI) untuk kualitas clustering. Hasil penelitian menunjukkan bahwa Cosine Similarity secara konsisten mengungguli Jaccard Similarity, dengan nilai macro average F1-Score sebesar 0,46 berbanding 0,34. Kualitas clustering pada Cosine Similarity juga lebih superior dengan skor DBI 0,252. Penelitian ini menyimpulkan bahwa Cosine Similarity merupakan metode yang lebih efektif untuk pencarian ayat Alkitab berbasis sentimen.

Kata kunci: *cosine similarity; jaccard similarity; K-Means; pencarian ayat Alkitab*

Abstract

Searching for relevant Bible verses based on emotional themes is a significant challenge for users due to the large number of verses and complex language variations. This study aims to design, implement, and compare the performance of a Bible verse retrieval system using Jaccard Similarity and Cosine Similarity methods. The data used consists of 31,102 Bible verses from the SABDA website, which underwent a text preprocessing stage. The Cosine Similarity method utilizes TF-IDF weighting, while Jaccard Similarity is based on word sets. The search results were then grouped using the K-Means Clustering algorithm to distinguish between relevant and less relevant results. The evaluation was performed using Precision, Recall, and F1-Score metrics for the similarity methods, and the Davies-Bouldin Index (DBI) for clustering quality. The results show that Cosine Similarity consistently outperformed Jaccard Similarity, with a macro average F1-Score of 0.46 compared to 0.34. The clustering quality for Cosine Similarity was also superior with a DBI score of 0.252. This research concludes that Cosine Similarity is a more effective method for sentiment-based Bible verse retrieval.

Keywords: Bible verse retrieval; cosine similarity; jaccard similarity; K-Means

1. PENDAHULUAN

Alkitab merupakan sumber petunjuk utama bagi umat Kristiani, dengan kekayaan makna yang mendalam. Dalam kehidupan sehari-hari, banyak pengguna mencari panduan hidup melalui ayat-ayat Alkitab yang relevan dengan kondisi emosional yang mereka rasakan[1]. Namun, proses pencarian ini seringkali menjadi tantangan karena jumlah ayat yang sangat banyak dan variasi bahasa yang digunakan. Banyak aplikasi Alkitab digital yang sudah menyediakan fitur pencarian, namun seringkali gagal memberikan hasil yang akurat dan sesuai dengan kebutuhan emosional pengguna, sehingga diperlukan sebuah solusi sistem pencarian yang lebih presisi [2].

Information Retrieval (IR) menawarkan solusi untuk mengatasi masalah ini melalui metode pengukuran kemiripan teks[3]. Beberapa penelitian sebelumnya telah menunjukkan keberhasilan metode similaritas dalam konteks yang beragam. Penelitian oleh O. Nurdiana, dkk. [4] pada aplikasi pencarian terjemahan Al-Qur'an menemukan bahwa Cosine Similarity memberikan hasil yang lebih akurat dibandingkan Jaccard Similarity. Keunggulan Cosine Similarity atas Jacquard Similarity ini juga didukung oleh penelitian lain, seperti pada pendeteksian produk halal oleh Nur Aini Rakhmawati, dkk. [5] dan pencocokan judul artikel ilmiah oleh Rinjeni, T. P, dkk. [6]. Keakuratan hasil pencarian Cosine Similarity juga dirasakan oleh pengguna aplikasi. Dalam sebuah uji coba perangkat lunak pencarian ayat Alquran berbasis Android, sekitar 70% responden menyatakan bahwa aplikasi yang sedang dalam tahap pengembangan tersebut dapat membantu

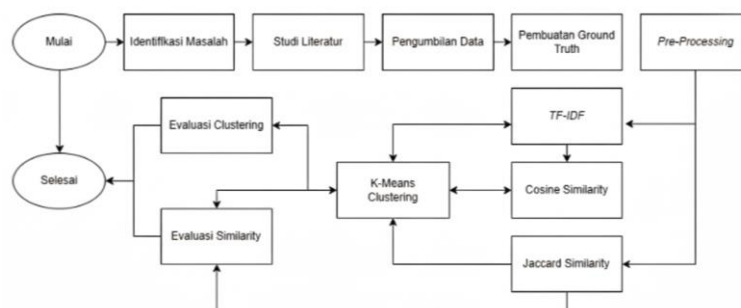
mereka dalam mencari ayat Alquran yang tematik [7]. Oleh karena itu, penelitian ini mengusulkan penerapan dan perbandingan metode Cosine Similarity, yang mengukur kemiripan berdasarkan sudut antar vektor teks, dan Jaccard Similarity, yang menghitung irisan dan gabungan himpunan kata. Hasil pencarian tersebut dikelompokkan menggunakan algoritma K-means Clustering untuk memperkaya analisis.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menganalisis dan membandingkan secara komprehensif efektivitas metode Jaccard Similarity dan Cosine Similarity dalam sistem temu kembali ayat Alkitab berdasarkan tema emosi. Kontribusi utama dari penelitian ini adalah memberikan bukti empiris mengenai metode mana yang lebih unggul untuk konteks pencarian teks religius yang spesifik ini, sekaligus menyajikan kerangka kerja yang dapat meningkatkan akurasi pencarian ayat Alkitab secara tematik dan kontekstual bagi pengguna.

2. METODE

2.1 Tahapan Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan yang sistematis untuk memastikan proses analisis berjalan secara terstruktur dan valid. Tahapan penelitian dimulai dari identifikasi masalah, studi literatur, pengumpulan data, hingga evaluasi hasil dan penarikan kesimpulan. Secara garis besar, alur kerja penelitian diilustrasikan pada Gambar 1. Proses utama meliputi pembuatan ground truth sebagai acuan, preprocessing data teks Alkitab, perhitungan bobot kata menggunakan TF-IDF, penerapan metode Cosine Similarity dan Jaccard Similarity, pengelompokan hasil dengan K-Means Clustering, serta evaluasi komprehensif terhadap performa kedua metode similaritas dan kualitas clustering.



Gambar 1. Tahapan Penelitian

2.2. Pengumpulan Data dan Pembuatan *Ground Truth*

Dataset yang digunakan dalam penelitian ini adalah ayat-ayat Alkitab yang bersumber dari website SABDA, dengan total 31.102 ayat. Setiap ayat merupakan unit data mandiri yang akan diproses dan dianalisis. Proses pengumpulan data dilakukan dengan mengunduh dan mengekstrak teks ayat-ayat tersebut ke dalam format yang dapat diproses lebih lanjut oleh sistem.

Ground truth (kebenaran dasar) diperlukan sebagai patokan dalam evaluasi performa, apakah hasil pencarian sesuai dengan referensi [8]. Pembuatan ground truth ini melibatkan 240 ayat dari Kitab Mazmur sebagai data uji. Dari jumlah tersebut, 40 ayat dipilih secara acak untuk dijadikan query atau kata kunci pencarian. Setiap query kemudian divalidasi oleh tiga ahli (teolog) untuk menentukan ayat-ayat lain yang relevan secara tematik. Konsensus dari para ahli ini digunakan untuk membangun daftar ayat-ayat relevan yang akan menjadi tolok ukur dalam menghitung nilai Precision, Recall, dan F1-Score selama evaluasi sistem[9].

2.3. *Preprocessing* Teks

Tahap preprocessing teks merupakan langkah krusial untuk membersihkan dan mempersiapkan data mentah ayat Alkitab sebelum diolah oleh sistem. Proses ini bertujuan untuk mengurangi noise dan menyeragamkan format teks, sehingga analisis similaritas dapat berjalan lebih akurat [10]. Tahapan preprocessing yang dilakukan dalam penelitian ini meliputi empat langkah utama, yaitu:

- **Case Folding:** Mengubah seluruh huruf dalam teks menjadi huruf kecil (*lowercase*) untuk menghilangkan ambiguitas akibat perbedaan kapitalisasi.
- **Tokenisasi:** Memecah kalimat menjadi unit-unit kata individual yang disebut token. Proses ini juga mencakup penghapusan tanda baca dan karakter selain huruf.
- **Stopword Removal:** Menghapus kata-kata umum yang sering muncul namun tidak memiliki makna signifikan (contoh: 'dan', 'di', 'yang', 'ke') dari daftar token.
- **Stemming:** Mengubah setiap kata ke dalam bentuk dasarnya (kata dasar) untuk menyatukan variasi kata yang memiliki akar makna yang sama (contoh: 'menciptakan' menjadi 'cipta').

Setelah melalui keempat tahapan ini, data teks Alkitab menjadi lebih bersih dan terstruktur, siap untuk proses pembobotan kata dan pengukuran similaritas.

2.4. Pembobotan Kata dan Pengukuran Similaritas

Setelah teks dibersihkan melalui preprocessing, langkah selanjutnya, Pembobotan adalah mengubah data teks menjadi representasi numerik untuk dapat diukur tingkat kemiripannya [11]. Dalam penelitian ini, dua pendekatan pengukuran similaritas dievaluasi. Langkah Awal adalah pendekatan berbasis vektor yang diawali dengan pembobotan kata menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF adalah metode statistik yang digunakan untuk mengukur seberapa penting sebuah kata dalam suatu dokumen di antara kumpulan dokumen lainnya [12]. Bobot TF-IDF akan tinggi jika sebuah kata sering muncul dalam satu dokumen, namun jarang muncul di dokumen-dokumen lain. Perhitungan bobot TF-IDF mengikuti Persamaan (1).

$$W_{d,t} = tf_{d,t} \times idf_{d,t} \quad (1)$$

Vektor TF-IDF dari setiap ayat Alkitab dibandingkan dengan vektor TF-IDF dari query pengguna menggunakan metode Cosine Similarity. Metode ini mengukur kemiripan dengan menghitung nilai kosinus dari sudut antara dua vektor [12]. Nilai yang dihasilkan berkisar antara 0 hingga 1, di mana nilai yang mendekati 1 menandakan tingkat kemiripan yang sangat tinggi. Perhitungan Cosine Similarity dapat dilihat pada Persamaan (2).

$$\frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (2)$$

Pendekatan kedua adalah berbasis himpunan kata menggunakan metode Jaccard Similarity. Metode ini tidak memerlukan pembobotan TF-IDF dan bekerja langsung pada hasil tokenisasi. Jaccard Similarity mengukur kemiripan antara dua dokumen dengan membagi jumlah kata yang sama (irisasi) dengan jumlah total kata unik dari kedua dokumen (gabungan). Rumus perhitungannya disajikan pada Persamaan (3).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

2.5. Pengelompokan dan Evaluasi Hasil

Setelah hasil pencarian diperoleh dari metode Cosine Similarity dan Jaccard Similarity, langkah selanjutnya adalah melakukan pengelompokan dan evaluasi.

2.5.1. Pengelompokan Hasil dengan *K-Means Clustering*

Algoritma K-Means Clustering digunakan untuk mengidentifikasi dan memisahkan hasil pencarian yang relevan dan tidak relevan [13]. K-Means merupakan algoritma clustering non-hierarkis yang efisien untuk mengelompokkan data ke dalam sejumlah cluster yang telah ditentukan [14]. Dalam penelitian ini, data hasil pencarian dikelompokkan menjadi dua cluster ($K=2$) yang merepresentasikan kelompok "ayat relevan" dan "ayat kurang relevan". Pengelompokan ini bertujuan untuk memperjelas pemisahan antara hasil yang benar-benar sesuai dengan kueri dan hasil yang kurang sesuai, sehingga memudahkan analisis dan evaluasi.

2.5.2. Evaluasi Performa Sistem

Evaluasi Kualitas Clustering diukur menggunakan Davies-Bouldin Index (DBI). DBI mengevaluasi seberapa padat cluster dan seberapa baik cluster-cluster tersebut terpisah satu sama lain [15]. Nilai DBI yang lebih rendah menunjukkan kualitas clustering yang lebih baik, artinya cluster lebih padat dan lebih terpisah. Perhitungan DBI mengikuti Persamaan (4).

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right) \quad (4)$$

Evaluasi Performa Similaritas menggunakan metrik Precision, Recall, dan F1-Score, yang merupakan teknik evaluasi performa yang umum digunakan dalam sistem temu kembali informasi [16]. Precision sendiri mengukur proporsi dokumen yang relevan dari seluruh dokumen yang berhasil ditemukan oleh sistem. Artinya, seberapa tepat sistem memberikan hasil yang relevan. Perhitungan Precision disajikan pada Persamaan (5).

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall mengukur proporsi dokumen relevan yang berhasil ditemukan oleh sistem dari seluruh dokumen relevan yang seharusnya ada. Artinya, seberapa lengkap sistem menemukan semua hasil yang relevan. Perhitungan Recall dapat dilihat pada Persamaan (6).

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F1-Score adalah rata-rata harmonik dari Precision dan Recall, memberikan keseimbangan antara keduanya. Metrik ini berguna ketika ada kebutuhan untuk mempertimbangkan baik Precision maupun Recall secara bersamaan. Perhitungan F1-Score disajikan pada Persamaan (7).

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

3. HASIL PENELITIAN

3.1. Hasil Pengukuran Similaritas dan Clustering

Pada tahap ini, disajikan hasil evaluasi kuantitatif dari kedua metode similaritas (Cosine Similarity dan Jaccard Similarity) serta evaluasi kualitas pengelompokannya. Pengujian dilakukan terhadap 10 query emosi yang berbeda, dan hasilnya diukur menggunakan metrik Precision, Recall, F1-Score, dan Davies-Bouldin Index (DBI). Hasil lengkap dari pengujian disajikan pada Tabel 1.

Tabel 1. Hasil Evaluasi Perbandingan Kinerja Metode Similaritas dan Clustering.

Jenis Similaritas	Query	Precision	Recall	F1-Score	DBI
Cosine	Sukacita	0.90	0.45	0.60	0,275
Cosine	Marah	0.20	0.10	0.13	0,237
Cosine	Sedih	0.70	0.35	0.47	0,185
Cosine	Kasih	0.60	0.30	0.40	0,240
Cosine	Macro Average	0.69	0.35	0.46	0,252
Jaccard	Sukacita	0.60	0.30	0.40	0,273
Jaccard	Marah	0.30	0.15	0.20	0,296
Jaccard	Sedih	0.70	0.35	0.47	0,262
Jaccard	Kasih	0.40	0.20	0.27	0,300

Jaccard	Macro Average	0.51	0.26	0.34	0,292
---------	---------------	------	------	------	-------

3.2 Pembahasan

Hasil penelitian yang disajikan pada Tabel 1 secara konsisten menunjukkan bahwa metode Cosine Similarity memiliki kinerja yang lebih unggul dibandingkan dengan Jaccard Similarity dalam konteks pencarian ayat Alkitab berbasis emosi. Keunggulan ini terlihat pada semua metrik evaluasi, baik dari sisi akurasi pencarian (Precision, Recall, F1-Score) maupun dari kualitas pengelompokan hasil (Davies-Bouldin Index).

Perbedaan performa ini dapat dijelaskan dari cara kerja fundamental kedua metode. Jaccard Similarity adalah metrik berbasis himpunan (set-based) yang hanya memperhitungkan keberadaan kata kunci dalam dokumen. Metode ini memperlakukan setiap kata dengan bobot yang sama dan mengabaikan frekuensi kemunculan kata, sehingga kurang mampu menangkap relevansi kontekstual. Sebaliknya, Cosine Similarity bekerja pada representasi vektor teks yang telah diberi bobot melalui TF-IDF. Pendekatan ini tidak hanya mendeteksi keberadaan kata, tetapi juga mengukur tingkat kepentingannya dalam sebuah ayat. Dengan demikian, Cosine Similarity mampu memberikan skor kemiripan yang lebih bernuansa dan akurat, yang pada akhirnya menghasilkan pengelompokan data (clustering) yang lebih optimal, seperti yang ditunjukkan oleh nilai DBI yang lebih rendah.

Keunggulan kerangka kerja berbasis Cosine Similarity ini terletak pada skalabilitasnya untuk menangani query yang lebih kompleks di masa depan. Sementara pencocokan kata kunci sederhana mungkin cukup untuk query literal seperti 'sedih', Cosine Similarity mampu memproses kalimat utuh sebagai satu kesatuan vektor. Hal ini membuka potensi pengembangan sistem untuk dapat memahami query bahasa alami yang lebih panjang, seperti 'ayat untuk menenangkan hati yang sedang gelisah'. Dengan demikian, penelitian ini tidak hanya mengkonfirmasi superioritas Cosine Similarity dalam konteks yang diuji, tetapi juga menyediakan fondasi metodologis yang kuat untuk pengembangan sistem pencarian teks religius yang lebih canggih dan intuitif.

4. DISKUSI

Hasil penelitian yang disajikan pada Tabel 1 secara konsisten menunjukkan bahwa metode Cosine Similarity memiliki kinerja yang lebih unggul dibandingkan dengan Jaccard Similarity dalam konteks pencarian ayat Alkitab berbasis emosi. Keunggulan ini terlihat pada semua metrik evaluasi, baik dari sisi akurasi pencarian (Precision, Recall, F1-Score) maupun dari kualitas pengelompokan hasil (Davies-Bouldin Index).

Perbedaan performa ini dapat dijelaskan dari cara kerja fundamental kedua metode. Jaccard Similarity adalah metrik berbasis himpunan (set-based) yang hanya memperhitungkan keberadaan kata kunci dalam dokumen. Metode ini memperlakukan setiap kata dengan bobot yang sama dan mengabaikan frekuensi kemunculan kata, sehingga kurang mampu menangkap relevansi kontekstual. Sebaliknya, Cosine Similarity bekerja pada representasi vektor teks yang telah diberi bobot melalui TF-IDF. Pendekatan ini tidak hanya mendeteksi keberadaan kata, tetapi juga mengukur tingkat kepentingannya dalam sebuah ayat. Dengan demikian, Cosine Similarity mampu memberikan skor kemiripan yang lebih bernuansa dan akurat, yang pada akhirnya menghasilkan pengelompokan data (clustering) yang lebih optimal, seperti yang ditunjukkan oleh nilai DBI yang lebih rendah. Peneliti menyarankan untuk mengeksplorasi penggunaan metode pembobotan lain seperti Word2Vec atau BERT dalam meningkatkan pemahaman kontekstual serta menggunakan dataset yang lebih besar dan kueri yang lebih variatif untuk penelitian selanjutnya.

5. KESIMPULAN

Berdasarkan hasil dan diskusi penelitian, dapat ditarik kesimpulan bahwa metode Cosine Similarity menunjukkan performa yang secara signifikan lebih unggul dibandingkan Jaccard Similarity untuk sistem temu kembali ayat Alkitab berbasis emosi. Keunggulan ini terbukti dari nilai macro average F1-Score yang lebih tinggi (0.46 vs 0.34) dan skor Davies-Bouldin Index (DBI) yang lebih rendah (0.252 vs 0.292), yang mengindikasikan akurasi pencarian lebih baik dan kualitas cluster yang lebih optimal. Penelitian ini

mengkonfirmasi bahwa pendekatan berbasis vektor dengan pembobotan TF-IDF lebih efektif dalam menangkap relevansi semantik pada teks religius dibandingkan pendekatan berbasis himpunan kata.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada Lembaga Penelitian dan Pengabdian kepada Masyarakat (LP2M) Universitas Jember atas dukungan fasilitas penelitian, serta bapak/ibu peneliti yang telah memberikan kontribusi hingga selesainya penelitian ini.

DAFTAR PUSTAKA

- [1] Muhammad Khalid, "Contributions of Dr. Ibrahim Muhammad Ibrahim to the Development of Arabic and Urdu Literature: A Study of His Life and Works," *Al-Qamar*, vol. 7, pp. 1–14, Jun. 2024.
- [2] A. Angdresey, M. A. Lamongi, and R. Munir, "Information Retrieval System in the Bible," *Cogito Smart Journal* |, vol. 7, no. 1, 2021.
- [3] Abhishek Anand, "Information Retrieval in Computing Model," *International Conference on Intelligent Computing and Control Systems (ICCS)*, vol. 10, Apr. 2020.
- [4] O. Nurdiana and D. Nursantika, "PERBANDINGAN METODE COSINE SIMILARITY DENGAN METODE JACCARD SIMILARITY PADA APLIKASI PENCARIAN TERJEMAH AL-QUR'AN DALAM BAHASA INDONESIA," no. 1, 2016.
- [5] N. Aini Rakhmawati, A. Adi Firmansyah, P. Maulidya Effendi, R. Abdillah, and T. Agung Cahyono, "Auto Halal Detection Products Based on Euclidian Distance and Cosine Similarity," vol. 8, pp. 4–6, 2018, [Online]. Available: <http://halal.addi.is.its.ac.id>;
- [6] T. P. Rinjeni, A. Indriawan, and N. A. Rakhmawati, "Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 553–560. doi: 10.1016/j.procs.2024.03.039.
- [7] C. Hadi, dan Muhammad Rifqi Ma, P. A. Studi Teknik Informatika STMIK Jenderal Yani Yogyakarta, and P. A. Studi Sistem Informasi STMIK Jenderal Yani Yogyakarta, "IMPLEMENTASI COSINE SIMILARITY DALAM APLIKASI PENCARIAN AYAT AL-QUR'AN BERBASIS ANDROID AN IMPLEMENTATION OF COSINE SIMILARITY FOR ANDROID BASED QURAN VERSE SEARCHING APPLICATION."
- [8] M. Wisaksono and E. Pujiyanto, "Pengembangan Sistem Berbasis Pengetahuan untuk Pengendalian Kualitas pada Proses Frontone Produk Infrared (Studi Kasus pada PT Perkin Elmer Optoelectronics Batam)," 2003.
- [9] "STUDI PERBANDINGAN METODE EKSTRAKSI FITUR UNTUK TOPIC MODELING BERBASIS ASPEK DAN SENTIMEN ANALISIS PADA ULASAN PRODUK E-COMMERCE THESIS Oleh: FAHRENDRA KHOIRUL IHTADA NIM. 230605210033 PROGRAM STUDI MAGISTER INFORMATIKA FAKULTAS SAINS DAN TEKNOLOGI UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG 2025."
- [10] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int J Comput Appl*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.
- [11] D. Iskandar and A. Kurniawati, "Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 1, pp. 133–144, Feb. 2025, doi: 10.25126/jtiik.2025129143.
- [12] Dewa Ayu Rai Ariantini dkk, "PENGUKURAN KEMIRIPAN DOKUMEN TEKS BAHASA INDONESIA MENGGUNAKAN METODE COSINE SIMILARITY," *E-Journal Teknik Informatika*, vol. 9, pp. 1–8, Sep. 2019.
- [13] I. A. Nur Afifah and H. Nurdiyanto, "DATA MINING CLUSTERING DALAM PENGELOMPOKAN BUKU PERPUSTAKAAN MENGGUNAKAN ALGORITMA K-MEANS," *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 3, pp. 802–814, Aug. 2023, doi: 10.29100/jupi.v8i3.3891.
- [14] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustering dalam Penentuan Siswa Kelas Unggulan," vol. 15, no. 2.
- [15] R. Gesit Prasasti Alam and Y. Everhard, "Optimasi K-Means Dengan Particle Swarm Optimization Dalam Penentuan Titik Awal Pusat Klaster Data Telekomunikasi K-Means Optimization with Particle Swarm Optimization in Determining the Starting Point of Cluster Centers of Telecommunication Data."
- [16] I. Muslim *et al.*, "Implementasi Text Summarization Pada Review Aplikasi Digital Library System Menggunakan Metode Maximum Marginal Relevance," 2024.