

PENERAPAN VECTOR SPACE MODEL PADA REKOMENDASI LAGU ROHANI KRISTEN BERBASIS LIRIK LAGU

Fernando Sugianto Putra, Yoannita

Informatika, Universitas Multi Data Palembang

Jl Rajawali No.14, 9 Ilir, Kec. Ilir Tim. II, Kota Palembang, Sumatera Selatan

fernandosugiantoputra_2226250106@mhs.mdp.ac.id

ABSTRAK

Perkembangan teknologi musik digital yang pesat saat ini telah memudahkan akses terhadap ribuan lagu religi, termasuk lagu rohani Kristen yang digunakan untuk mendukung peribadatan dan kebutuhan spiritual sehari-hari. Namun, banyaknya pilihan lagu seringkali membuat pengguna kesulitan dalam menemukan lagu yang benar-benar relevan dengan konteks atau suasana hati spiritual tertentu secara cepat dan akurat. Penelitian ini bertujuan untuk mengimplementasikan *Vector Space Model* (VSM) dalam rekomendasi lagu rohani Kristen berbasis lirik untuk memberikan hasil evaluasi dari model yang telah dibangun. Metode yang digunakan meliputi *pre-processing* teks secara menyeluruh (termasuk *stemming* Sastrawi), pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF), dan penghitungan kemiripan menggunakan *Cosine Similarity*. Dataset yang digunakan mencakup 1.068 lirik lagu bersih dengan pembagian proporsi 80:20 untuk data latih dan data uji. Hasil pengujian menunjukkan performa sistem yang sangat impresif dengan tingkat Akurasi mencapai 97,65% dan *Precision@10* sebesar 78,22%. Hal ini membuktikan bahwa penerapan VSM efektif dalam mentransformasi lirik lagu menjadi representasi matematis yang akurat. Secara keseluruhan, penelitian ini memberikan kontribusi dalam bidang *information retrieval* dengan menunjukkan bahwa pendekatan berbasis vektor mampu menghasilkan rekomendasi yang relevan secara kontekstual guna mendukung kebutuhan spiritual pengguna.

Kata kunci : cosine similrity, lagu rohani kristen, lirik lagu, TF-IDF, vector space model

1. PENDAHULUAN

Perkembangan teknologi informasi dalam beberapa tahun terakhir telah mendorong peningkatan signifikan dalam ketersediaan informasi digital, termasuk pada bidang musik religi. Berbagai platform digital kini menyediakan ribuan lagu rohani yang dapat diakses secara instan oleh masyarakat. Namun tantangan utama muncul pada efektivitas temu kembali informasi (*Information Retrieval*). Jumlah data lirik yang sangat besar membuat pencarian manual menjadi kurang efisien karena banyaknya kesamaan terminologi teks namun dengan kedalaman makna yang berbeda. Kondisi ini menimbulkan permasalahan dalam menemukan lagu rohani yang memiliki konteks spiritual serupa secara cepat dan akurat, sehingga diperlukan suatu pemodelan matematis yang mampu memetakan kedekatan konteks antar teks secara objektif [1].

Dalam konteks musik Kristiani, lirik lagu memegang peran yang sangat penting karena mengandung pesan spiritual, ungkapan penyembahan, dan refleksi iman yang mendalam. Potongan lirik sering kali menjadi indikator utama bagi seseorang untuk menemukan lagu lain yang memiliki tema atau makna serupa. Karakteristik teks lirik yang seringkali memiliki repetisi kata dan tema serupa memerlukan pendekatan *Content-Based Filtering* yang kuat untuk menentukan tingkat kemiripan dokumen. Penelitian terbaru menunjukkan bahwa pendekatan ini sangat efektif dalam menyelesaikan masalah pengenalan pola teks dan kategorisasi dokumen berdasarkan kesamaan konte [2].

Meskipun berbagai metode temu kembali informasi telah banyak diterapkan, penelitian yang secara spesifik berfokus pada domain lagu rohani Kristen masih relatif terbatas. Padahal, lagu rohani memiliki karakteristik unik dari sisi bahasa dan terminologi teologis yang sering kali merujuk pada ayat Alkitab tertentu, sehingga memiliki nilai kontekstual yang sangat tinggi. Keterbatasan penelitian di domain ini menunjukkan adanya celah untuk melakukan evaluasi performa model secara lebih komprehensif dalam konteks lirik lagu rohani

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* dalam kerangka kerja *Vector Space Model* (VSM) merupakan pendekatan yang sangat efektif untuk mengukur tingkat kemiripan antar dokumen teks. TF-IDF bekerja dengan memberikan bobot pada kata yang dianggap penting, sementara *Cosine Similarity* menghitung tingkat kedekatan antar dokumen berdasarkan orientasi vektornya [3]. Berbagai studi dalam lima tahun terakhir menunjukkan bahwa kombinasi kedua metode ini telah terbukti efektifitas yang relevan di berbagai bidang. Sebagai contoh, metode ini telah terbukti efektif dalam memberikan rekomendasi resep masakan berdasarkan kesesuaian bahan dengan hasil yang akurat [4]. Selain itu, efektivitasnya juga terlihat pada sistem rekomendasi lagu berdasarkan jenis mood menggunakan TF-IDF dan *Cosine Similarity* [5], serta pada sistem rekomendasi artikel literatur akademik [3].

Berdasarkan kondisi tersebut, penelitian ini berfokus pada analisis penerapan *Vector Space Model* (VSM) menggunakan metode TF-IDF dan *Cosine*

Similarity pada data lirik lagu rohani Kristen. Penelitian ini bertujuan untuk membangun model representasi teks dan menguji performanya melalui evaluasi kuantitatif menggunakan metrik Akurasi, *Precision*, *Recall*, dan *F1-Score*. Dengan menggunakan dataset sebanyak 1.068 lirik lagu rohani Kristen berbahasa Indonesia yang telah divalidasi, diharapkan penelitian ini dapat memberikan kontribusi ilmiah mengenai efektivitas pemodelan ruang vektor dalam sistem temu kembali informasi, khususnya pada domain musik rohani yang memiliki kekhasan bahasa teologis.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Berbagai penelitian telah membahas penerapan *Vector Space Model* (VSM) dan metode pembobotan TF-IDF dalam sistem temu kembali informasi maupun sistem rekomendasi berbasis teks. Dalam konteks rekomendasi lagu, C. Virginia et al. [5] mengembangkan sistem rekomendasi lagu berdasarkan jenis *mood* menggunakan metode TF-IDF dan *Cosine Similarity*. Hasil penelitian menunjukkan bahwa pendekatan tersebut mampu memberikan rekomendasi lagu yang relevan sesuai dengan preferensi suasana hati pengguna. Sejalan dengan itu, A. Y. Timur dan A. N. Rohman [6] menerapkan metode *Content-Based Filtering* dan *Cosine Similarity* pada sistem rekomendasi lagu Indonesia, yang terbukti efektif dalam menghitung tingkat kemiripan antar lagu berdasarkan fitur teks sehingga menghasilkan rekomendasi yang akurat.

Pada penerapan *Vector Space Model* juga menunjukkan performa yang baik. A. T. Adiyanto et al. [7] menerapkan VSM pada sistem pencarian dokumen kearsipan dan membuktikan bahwa metode ini mampu meningkatkan ketepatan hasil pencarian berdasarkan nilai kemiripan dokumen terhadap query. Penelitian oleh W. O. W. Makmun et al. [8] menerapkan VSM pada sistem pencarian artikel arkeologi dan membuktikan bahwa metode ini mampu memberikan hasil pencarian yang akurat dengan nilai rata-rata *precision* tertinggi mencapai 98,125% untuk kata kunci tunggal. Selain itu, S. Panambunan et al. [9] mengimplementasikan VSM pada aplikasi pengarsipan berbasis web di lingkungan kepolisian, yang menunjukkan bahwa algoritma ini efektif dalam mempercepat proses penemuan dokumen serta memberikan urutan relevansi yang tepat berdasarkan nilai kemiripan (*similarity*).

Selain pada domain musik, juga banyak diterapkan pada bidang lain. C. Y. Fuady et al. [10] mengimplementasikan metode *Content-Based Filtering* untuk rekomendasi film berdasarkan genre dan sinopsis. Penelitian tersebut menunjukkan bahwa pembobotan TF-IDF dan pengukuran kemiripan menggunakan *Cosine Similarity* mampu meningkatkan relevansi hasil rekomendasi film. Sementara itu, A. Fitria et al. [11] mengembangkan sistem rekomendasi lowongan pekerjaan berbasis

Content-Based Filtering, yang memanfaatkan kesesuaian kata kunci antara profil pengguna dan deskripsi pekerjaan untuk menghasilkan rekomendasi yang lebih personal dan tepat sasaran.

Berdasarkan penelitian-penelitian tersebut, dapat disimpulkan bahwa metode TF-IDF dan *Cosine Similarity* dalam kerangka kerja *Vector Space Model* terbukti efektif dan fleksibel dalam berbagai domain, baik pada sistem rekomendasi musik, film, pekerjaan, maupun sistem pencarian dokumen. Namun demikian, penerapan metode ini secara khusus pada domain rekomendasi lagu rohani Kristen berbasis lirik masih sangat terbatas. Oleh karena itu, penelitian ini berfokus pada implementasi *Vector Space Model* untuk mengukur kemiripan lirik lagu rohani Kristen guna menghasilkan rekomendasi yang relevan berdasarkan konteks spiritual dan teologis yang terkandung dalam teks lagu.

2.2. Lagu Rohani

Lagu didefinisikan sebagai gubahan musik yang memadukan alunan instrumen dengan syair atau lirik untuk dinyanyikan. Lirik merupakan rangkaian kata yang berfungsi menyampaikan pesan, emosi, atau makna tertentu melalui paduan nada dan irama [6]. Sebagai bentuk ekspresi pengalaman pribadi, penulisan lirik lagu memiliki kompleksitas tersendiri karena lahir dari berbagai sumber inspirasi yang berbeda dibandingkan dengan karya sastra lainnya seperti puisi.

Dalam konteks spiritual, lirik lagu rohani memiliki peran krusial sebagai media penyampaian pesan teologis sekaligus sarana untuk membangun kedekatan antara manusia dengan Tuhan. Melalui lirik yang mengandung nilai-nilai motivasi dan inspiratif, pendengar dapat merasakan kedamaian batin serta memperoleh penguatan iman dalam menjalani kehidupan sehari-hari [12]. Dengan demikian, lagu rohani tidak hanya berfungsi sebagai sarana peribadatan, tetapi juga sebagai instrumen penguatan spiritual yang memotivasi individu untuk melakukan kebaikan sesuai dengan nilai-nilai religius.

2.3. Information Retrieval (IR)

Information Retrieval (IR) adalah bidang ilmu yang mempelajari proses pencarian, pengambilan, dan penyajian informasi atau dokumen yang relevan dari sekumpulan data besar berdasarkan permintaan (*query*) pengguna. Tujuan utama dari sistem ini adalah memfasilitasi pengguna untuk menemukan informasi yang dibutuhkan secara cepat dan efisien, meskipun hanya sebagian dokumen yang memiliki tingkat relevansi dengan permintaan mereka. Dalam praktiknya, IR melibatkan komponen utama seperti *preprocessing* teks, ekstraksi fitur, pemeringkatan dokumen, hingga evaluasi kinerja untuk memastikan ketepatan informasi yang disajikan [13].

Dalam pengembangannya, mekanisme IR beroperasi dengan melakukan komparasi antar dokumen guna mengukur derajat kedekatan atau

kemiripan konten di antara keduanya. Implementasi bidang ini telah mencakup berbagai domain luas, seperti pengelompokan teks secara otomatis, optimasi pencarian arsip digital, hingga pengembangan sistem rekomendasi. Sebagai contoh, dalam ekosistem platform lowongan pekerjaan, IR dimanfaatkan untuk mencocokkan kualifikasi pada *curriculum vitae* (CV) pelamar dengan deskripsi pekerjaan yang tersedia, sehingga proses seleksi kandidat menjadi lebih efektif dibandingkan pencarian konvensional [11]. Hingga saat ini, arsitektur klasik seperti *Vector Space Model* (VSM) yang dipadukan dengan teknik pembobotan TF-IDF tetap menjadi pilihan utama sebagai *baseline* dalam berbagai riset *Information Retrieval*. Hal ini dikarenakan karakteristik metodenya yang efisien namun tetap mampu menghasilkan tingkat akurasi yang tinggi dalam pengolahan data tekstual.

2.4. Pre-Processing text

Preprocessing text merupakan tahapan fundamental dalam *Natural Language Processing* (NLP) yang berfungsi untuk mentransformasi data teks mentah menjadi format yang lebih terstruktur dan konsisten. Proses ini bertujuan untuk membersihkan dataset dari elemen-elemen yang tidak relevan, sehingga data siap digunakan untuk proses analisis atau pelatihan model secara lebih efektif [2]. Dengan melakukan pembersihan di awal, sistem dapat meminimalkan gangguan (*noise*) dalam data yang dapat memengaruhi akurasi hasil evaluasi pada tahap pencarian informasi atau klasifikasi.

Secara teknis, tahapan *preprocessing* melibatkan beberapa langkah krusial, dimulai dari *case folding* untuk menyeragamkan huruf, penghapusan tanda baca, hingga tokenisasi untuk memecah kalimat menjadi satuan kata (*term index*). Selain itu, dilakukan pula pembersihan kata-kata umum yang kurang memiliki nilai informasi melalui *stopword removal*, serta penerapan teknik *stemming* atau *lemmatization* untuk mengubah kata berimbuhan menjadi kata dasar. Melalui serangkaian tahapan ini, dataset lirik lagu rohani yang diolah akan menghasilkan fitur-fitur kata yang optimal untuk diproses lebih lanjut oleh *Vector Space Model*.

2.5. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF merupakan sebuah teknik pembobotan kata yang diaplikasikan untuk mengevaluasi seberapa penting kedudukan suatu istilah di dalam sebuah dokumen dalam hubungannya dengan keseluruhan korpus [14]. Mekanisme ini mengombinasikan dua nilai perhitungan utama, yaitu *Term Frequency* (TF) yang menunjukkan intensitas kemunculan sebuah kata dalam dokumen tertentu, serta *Inverse Document Frequency* (IDF) yang berfungsi untuk mengukur tingkat keunikan atau kelangkaan kata tersebut di seluruh koleksi dokumen.

a. Term Frequency (TF)

TF mengukur seberapa sering suatu istilah t muncul dalam dokumen d . Nilainya dihitung menggunakan Persamaan (1).

$$tf(t, d) = 1 + \log tf \quad (1)$$

Keterangan :

tf = term frekuensi atau banyaknya kata pada dokumen.

$tf(t, d)$ = term frekuensi atau banyaknya kata t pada dokumen d atau pembobotan lokal

b. Inverse Document Frequency (IDF)

IDF digunakan untuk mengukur tingkat kepentingan suatu kata dalam keseluruhan dokumen. Perhitungannya ditunjukkan pada Persamaan (2).

$$idf(t) = \log \left(\frac{N}{df(t)} \right) \quad (2)$$

Keterangan :

$idf(t)$ = inverse document frequency

n = Jumlah keseluruhan dokumen

$df(t)$ = jumlah dokumen yang mengandung *term* t

c. Pembobotan TF-IDF

Nilai bobot akhir kata dihitung dengan mengalikan TF dan IDF seperti ditunjukkan pada Persamaan (3).

$$W(t, d) = tf(t, d) \times idf(t) \quad (3)$$

Keterangan :

$tf(t, d)$ = jumlah kata dalam d

$idf(t)$ = jumlah dokumen yang memuat *term*

Teknik pembobotan TF-IDF telah menunjukkan efektivitas yang signifikan dalam mengekstraksi fitur informasi krusial dari sebuah teks. Implementasinya sangat luas, terutama dalam mengoptimalkan performa klasifikasi dokumen, proses temu kembali informasi, hingga pengembangan sistem rekomendasi berbasis konten [10].

2.6. Vector Space Model (VSM)

Vector Space Model (VSM) adalah sebuah model aljabar atau kerangka kerja matematika yang digunakan untuk merepresentasikan dokumen teks ke dalam bentuk vektor dalam ruang multidimensi. Dalam penelitian ini, lirik lagu yang merupakan data teks diubah menjadi serangkaian angka koordinat yang dapat diukur kemiripannya secara geometris. Konsep di balik VSM adalah dengan menempatkan istilah (*term*), dokumen, dan pertanyaan (*query*) dalam satu ruang vektor sehingga memungkinkan sistem untuk menghitung persamaan atau kemiripan di antara mereka [7]. Model ini bekerja dengan cara memetakan kata-kata ke dalam ruang tersebut, di mana dokumen yang memiliki kemiripan isi akan memiliki posisi vektor yang berdekatan [15]. Dalam VSM, setiap kata unik yang terdapat dalam seluruh koleksi lirik lagu dianggap sebagai dimensi atau sumbu koordinat. Jika sebuah korpus memiliki n kata unik, maka setiap lagu direpresentasikan sebagai vektor d dalam ruang berdimensi n :

$$dj = (w1j, w2j, w3j, \dots, wnj) \quad (4)$$

Nilai bobot pada setiap dimensi tersebut ditentukan oleh hasil perhitungan TF-IDF. Dengan representasi ini, lirik lagu diproses sebagai titik-titik koordinat yang memiliki arah dan panjang tertentu, sehingga sistem dapat menghitung tingkat relevansi antar dokumen secara otomatis [16].

2.7. Cosine Similarity

Cosine Similarity adalah teknik yang dipakai untuk mengukur sejauh mana dua objek memiliki tingkat kesamaan [17]. Metode ini digunakan untuk mengukur tingkat kemiripan antara dua objek dengan menghitung sudut kosinus di antara kedua vektor yang merepresentasikan objek tersebut. Semakin kecil sudut yang terbentuk, maka semakin tinggi tingkat kesamaan keduanya, dapat dilihat dari persamaan (5).

$$\cos \alpha = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^n A_i \cdot B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (5)$$

Keterangan :

A = Vektor A, yang akan dibandingkan kemiripannya

B = Vektor B, yang akan dibandingkan kemiripannya

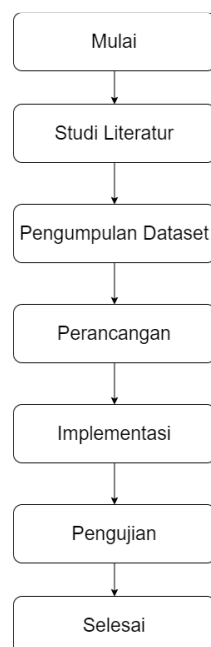
A X B = dot product antara vektor A dan vektor B

|A| = panjang vektor A

|B| = panjang vektor B

|A||B| = cross product antara |A| dan |B|

3. METODE PENELITIAN



Gambar 1. Tahapan metode penelitian

Alur penelitian ini disusun secara sistematis berdasarkan tahapan yang diilustrasikan pada Gambar 1. Rangkaian proses dimulai dari studi literatur dan pengumpulan dataset, yang kemudian dilanjutkan dengan perancangan sistem, tahap implementasi, serta pengujian akhir. Metodologi yang terstruktur ini

diterapkan untuk menjamin validitas pembangunan model dan memungkinkan dilakukannya evaluasi secara kuantitatif.

3.1. Studi Literatur

Tahapan studi literatur dilaksanakan guna mendalami teori serta meninjau penelitian-penelitian sebelumnya yang relevan dengan *Information Retrieval*, implementasi VSM, metode pembobotan TF-IDF, dan algoritma *Cosine Similarity*. Referensi dikumpulkan dari berbagai jurnal ilmiah, prosiding, serta repositori penelitian. Proses ini berperan sebagai landasan teoretis utama dalam merumuskan metodologi dan merancang arsitektur sistem.

3.2. Pengumpulan Dataset

Dataset penelitian ini terdiri dari 1.068 lirik lagu rohani Kristiani berbahasa Indonesia yang dikumpulkan melalui metode *web scraping* dari situs *GSN Soeki* dan *Kingdom Salvation*, serta penambahan data secara manual. Data yang telah diekstraksi disimpan dalam format Excel dengan atribut judul (*title*) dan lirik (*lyrics*). Untuk menjamin relevansi dan akurasi, dataset telah divalidasi oleh Ketua Departemen Musik GBI MPI Palembang sebelum digunakan dalam proses pemodelan VSM dan evaluasi sistem.

3.3. Perancangan Sistem

Perancangan sistem dalam penelitian ini difokuskan pada pengembangan alur kerja otomatis untuk mentransformasi lirik lagu rohani menjadi representasi numerik. Tahapan dimulai dengan inspeksi kualitas data melalui pengecekan nilai yang hilang (*missing value*) dan identifikasi data duplikat untuk menjamin integritas dataset. Setelah data divalidasi, tahapan berlanjut pada pembersihan data melalui rangkaian proses *pre-processing* teks yang dirancang secara sistematis menggunakan pustaka *Natural Language Toolkit* (NLTK) dan Sastrawi guna memastikan format data yang seragam. Proses diawali dengan *case folding* untuk mengonversi seluruh karakter menjadi huruf kecil, diikuti dengan tahap *cleaning* menggunakan teknik *Regular Expression* (RegEx) untuk menghapus angka, tanda baca, serta karakter khusus lainnya sehingga hanya menyisakan karakter alfabet dan spasi tunggal. Selanjutnya, dilakukan proses tokenisasi untuk memecah teks menjadi unit kata individu, serta *stopword removal* untuk menghapus kata-kata umum bahasa Indonesia yang tidak memiliki bobot informatif. Tahap akhir dari pembersihan teks adalah *stemming* menggunakan algoritma Sastrawi yang berfungsi mengembalikan setiap kata berimbuhan ke dalam bentuk dasarnya guna menyeragamkan variasi morfologi kata. Setelah data lirik berada dalam kondisi bersih, dirancang model *Vector Space Model* (VSM) dengan menerapkan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Dalam perancangan ini, dilakukan optimasi fitur dengan

menetapkan parameter *Minimum Document Frequency* ($\text{min_df}=2$) dan *Maximum Document Frequency* ($\text{max_df}=0.7$) untuk menyaring kata-kata yang terlalu jarang atau terlalu umum muncul dalam korpus. Sebagai tahap akhir perancangan, algoritma *Cosine Similarity* diintegrasikan sebagai instrumen utama untuk mengukur derajat kemiripan melalui perhitungan sudut kosinus antar vektor, yang memungkinkan sistem menentukan tingkat relevansi antara *query* pengguna dengan database lirik lagu.

3.4. Implementasi

Implementasi sistem direalisasikan menggunakan bahasa pemrograman *Python* di lingkungan *Google Colab* dengan memanfaatkan pustaka utama seperti *pandas* untuk manajemen dataset dan *scikit-learn* untuk pemrosesan tingkat lanjut. Pada tahap ini, fungsi *TfidfVectorizer* diimplementasikan untuk melakukan *fit-transform* terhadap data lirik, yang menghasilkan sebuah matriks ruang vektor multidimensi di mana setiap baris merepresentasikan lagu dan setiap kolom merepresentasikan bobot kata unik. Alur pemrosesan dirancang untuk mampu mengambil sampel potongan lirik dari data uji secara acak, mengubahnya menjadi vektor *query*, dan melakukan komputasi skor kemiripan menggunakan fungsi *cosine_similarity*. Hasil akhir dari proses implementasi ini adalah sebuah sistem yang dapat menyajikan daftar lagu secaraurut berdasarkan skor kemiripan tertinggi dalam rentang nilai 0 hingga 1.

3.5. Pengujian

Tahap pengujian dilakukan untuk mengevaluasi performa dan keandalan metode yang telah dibangun melalui skenario eksperimen yang objektif. Dataset lirik lagu rohani terlebih dahulu dibagi menjadi data latih dan data uji dengan proporsi 80:20 untuk menjamin validitas pengujian pada data yang belum pernah dipelajari oleh model. Evaluasi dilakukan secara kuantitatif dengan mengukur tingkat keberhasilan sistem melalui lima metrik utama, yaitu *Precision@K*, *Recall@K*, *F1-Score*, *Accuracy*, serta *Average Cosine Similarity*. Pengujian dijalankan dengan memberikan masukan berupa potongan lirik dari data uji dan memverifikasi apakah sistem mampu mengidentifikasi lagu yang tepat dalam daftar rekomendasi teratas berdasarkan ambang batas *threshold* yang telah ditetapkan sebesar 0.25. Melalui pengujian ini, efektivitas *Vector Space Model* kombinasi *TF-IDF* dan *Cosine Similarity* dalam menangani pencarian lirik lagu rohani diukur secara komprehensif.

4. HASIL DAN PEMBAHASAN

Gambar 2 menunjukkan rangkaian proses *preprocessing* terhadap 1068 data lirik lagu rohani Kristen yang mencakup tahapan *case folding*, *cleaning*, tokenisasi, dan *stopword removal*. Penggunaan algoritma Sastrawi dalam tahap *stemming*

berfungsi untuk menormalisasi istilah guna mereduksi variasi morfologi kata ke bentuk dasarnya. Selain itu, dilakukan deduplikasi data yang mengeliminasi 6 baris duplikat sehingga menghasilkan 1062 dataset bersih yang konsisten. Seluruh tahapan ini krusial untuk memastikan kualitas input sebelum data direpresentasikan ke dalam *Vector Space Model* (VSM), di mana setiap lirik akan dikonversi menjadi vektor numerik berdasarkan bobot TF-IDF untuk kemudian dihitung tingkat kemiripannya menggunakan *Cosine Similarity*.

```
Total data awal: 1068 lirik lagu

Memulai proses Pre-processing lirik...
(Proses ini memakan waktu beberapa menit karena komputasi Stemming Sastrawi)

--- PROSES PRE-PROCESSING SELESAI ---
✓ Case Folding & Cleaning selesai
✓ Tokenisasi & Stopword Removal selesai
✓ Stemming Sastrawi selesai

Menghapus duplikat: 6 baris dihapus.
Ukuran dataset: 1068 -> 1062
Total data bersih: 1062
```

Gambar 2. Proses *preprocessing* lirik lagu

Tabel 1 Hasil deduplikasi data

Keterangan data	Jumlah Lagu
Jumlah Lagu awal	1068
Jumlah Lagu duplikat Terdeteksi	6
Jumlah akhir	1062

Berdasarkan data yang disajikan pada Tabel 1, terlihat bahwa proses inspeksi data berhasil mengidentifikasi dan mengeliminasi 6 baris lagu duplikat dari total 1068 lagu awal. Tahapan deduplikasi ini sangat krusial karena menghasilkan dataset final sebanyak 1062 lagu yang unik dan konsisten. Dengan dataset yang telah terbebas dari redundansi, *Vector Space Model* (VSM) dapat membentuk ruang vektor yang lebih efisien tanpa adanya bias dari data yang berulang. Hal ini memungkinkan proses ekstraksi fitur melalui TF-IDF dan perhitungan sudut kosinus pada tahap *Cosine Similarity* menjadi lebih akurat, sehingga sistem dapat memberikan rekomendasi lirik lagu rohani yang benar-benar representatif dan relevan terhadap *query* atau masukan pengguna.

Tabel 2 Pembagian dataset

Data Latih	849
Data Uji	213

Tahap berikutnya adalah pembagian dataset *data splitting* yang dilakukan untuk memisahkan data latih dengan data pengujian. Berdasarkan tabel 2, dataset lirik lagu rohani yang berjumlah 1062 baris dibagi dengan proporsi 80% sebagai Data Latih dan 20% sebagai Data Uji. Pembagian ini menghasilkan 849 baris data latih yang akan dipetakan ke dalam *Vector Space Model* (VSM) sebagai basis pengetahuan sistem, serta 213 baris data uji yang bertindak sebagai *query* untuk mengukur performa sistem.

Penggunaan rasio 80:20 ini bertujuan untuk menyediakan basis data yang cukup besar bagi sistem agar dapat mengekstraksi fitur TF-IDF secara optimal,

sekaligus menyisakan jumlah data uji yang representatif untuk dievaluasi tingkat kemiripannya menggunakan *Cosine Similarity*. Dengan skema ini, validasi sistem dapat dilakukan secara objektif karena data yang diuji merupakan bagian yang terpisah dari basis data referensi utama.

```
--- TAHAP 1: MEMBANGUN MODEL VSM ---
Jumlah Data Latih (Database) : 849 baris
Jumlah Kata Unik (Dimensi)   : 1734
Keterangan: Fitur diseleksi otomatis dengan min_df=2 & max_df=0.7
```

Gambar 3 Proses membangun model vsm

Pada Gambar 3 bagian Tahap 1, diperlihatkan hasil dari konstruksi arsitektur *Vector Space Model* (VSM) yang menjadi fondasi utama model ini. Sebanyak 849 baris data latih yang telah melalui proses *preprocessing* ditransformasi menjadi representasi numerik menggunakan teknik pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Dalam proses ini, dilakukan optimasi fitur secara otomatis menggunakan parameter *Minimum Document Frequency min_df=2* untuk mengeliminasi kata-kata yang terlalu jarang muncul, serta *Maximum Document Frequency max_df=0.7* untuk menyaring istilah yang terlalu umum yang dapat mengurangi daya pembeda antar dokumen. Penerapan filter fitur tersebut menghasilkan ruang vektor dengan dimensi sebanyak 1.734 kata unik yang terkandung di dalam seluruh korpus lirik lagu rohani. Dengan jumlah vektor sebanyak 849 dan dimensi fitur sebesar 1.734, model VSM ini memiliki kerapatan data yang optimal untuk mendukung perhitungan kemiripan yang akurat, sehingga meminimalisir beban komputasi tanpa mengurangi kualitas representasi teks dari setiap lirik lagu yang ada di dalam database.

```
--- TAHAP 2: PENGUJIAN SAMPEL QUERY HASIL REKOMENDASI ---
Lagu Uji (Query): Sukacita - NDC Worship
-----
- Sgala Puji Syuku - Niko Njotorahardjo (Skor: 0.5346)
- Lagu tentang Kasih yang Manis (Skor: 0.5058)
- Kunyanyi Haleluya - Syamphony Worship (Skor: 0.4520)
- SUKACITA SURGA (Skor: 0.4450)
- Tertawa - GMS (Skor: 0.3984)
```

Gambar 4 Hasil rekomendasi

Selanjutnya, pada Tahap 2 dilakukan pengujian sampel *query* menggunakan salah satu data dari dataset uji, yaitu lagu berjudul "Sukacita - NDC Worship". Sistem melakukan transformasi lirik tersebut ke dalam vektor yang sama dengan model VSM, kemudian menghitung derajat kemiripan menggunakan algoritma *Cosine Similarity*. Hasilnya, sistem berhasil memberikan daftar rekomendasi lirik dengan skor kemiripan tertinggi secara *real-time*, seperti lagu "Sgala Puji Syukur" oleh Niko Njotorahardjo dengan skor 0.5346. Hal ini membuktikan bahwa model mampu mengenali kedekatan konteks lirik lagu meskipun judul atau penyanyinya berbeda, selama terdapat kemiripan distribusi kata di dalam ruang vektor.

Evaluasi performa sistem dilakukan menggunakan beberapa metrik standar untuk mengukur efektivitas model memberikan hasil rekomendasi lirik lagu. Hasil evaluasi menyeluruh terhadap 213 baris data uji ditunjukkan pada Gambar 5.

```
--- TAHAP 3: MEMULAI EVALUASI SISTEM (Threshold > 0.25) ---
--- REKAPITULASI HASIL EVALUASI AKHIR ---
Total Data Latih (Database)   : 849 baris
Total Data Uji (Target)       : 213 baris
Top-K Rekomendasi@           : 10
Rata-rata Akurasi (Success Rate) : 0.9765 (97.65%)
Rata-rata Precision           : 0.7822 (78.22%)
Rata-rata Recall               : 0.5393 (53.93%)
F1-Score                       : 0.6384
Rata-rata Cosine Sim           : 0.3750
```

Gambar 5 Hasil evaluasi

Berdasarkan Gambar 5, dapat diketahui bahwa nilai *Precision@10* yang mencapai 0.7822 mengindikasikan bahwa penerapan *Vector Space Model* sangat akurat dalam memetakan lirik lagu ke dalam ruang vektor multidimensi. Nilai *Recall@10* sebesar 0.5393 menunjukkan bahwa meskipun terdapat variasi diksi yang luas dalam lagu rohani, sistem telah mampu mengidentifikasi sebagian besar lagu yang relevan melalui perhitungan jarak antar vektor. Walaupun nilai *Average Cosine Similarity* berada pada angka 0.3750, tingginya nilai *F1-score* dan akurasi yang mencapai 97.65% menegaskan bahwa kerangka kerja VSM memiliki performa yang sangat stabil dalam mengelola kemiripan semantik lirik secara otomatis.

Secara keseluruhan, hasil penelitian menunjukkan bahwa penerapan *Vector Space Model* pada rekomendasi lagu rohani Kristen berbasis lirik lagu mampu menghasilkan informasi yang sangat relevan. Penggunaan VSM yang didukung oleh pembobotan TF-IDF dan kalkulasi *Cosine Similarity* terbukti efektif dalam mentransformasi teks lirik menjadi representasi matematis yang presisi. Hal ini memungkinkan sistem memetakan kedekatan konteks spiritual antar lagu secara akurat, sehingga memberikan pengalaman pencarian yang personal dan relevan bagi pengguna dalam skenario penggunaan nyata.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengimplementasikan *Vector Space Model* (VSM) pada sistem rekomendasi lagu rohani Kristen berbasis lirik lagu yang mampu memberikan rekomendasi secara relevan berdasarkan kemiripan lirik lagu. Hasil evaluasi terhadap 213 data uji menunjukkan bahwa penerapan VSM dengan dukungan pembobotan TF-IDF dan *Cosine Similarity* memperoleh nilai Akurasi sebesar 97,65%, *Precision@10* sebesar 0,7822, dan *Recall@10* sebesar 0,5393. Penggunaan parameter *min_df* dan *max_df* serta proses *stemming* Sastrawi terbukti efektif dalam meningkatkan ketepatan hasil rekomendasi sehingga dapat membantu pengguna menemukan lagu rohani

dengan konteks serupa secara cepat. Meskipun demikian, nilai *recall* dan rata-rata *similarity* masih dipengaruhi oleh tingginya variasi diksi pada lirik lagu, sehingga pada penelitian berikutnya disarankan untuk menerapkan teknik ekspansi *query*, memperluas skala dataset dari berbagai genre musik rohani, serta mengintegrasikan metode representasi teks yang lebih modern agar kualitas rekomendasi menjadi semakin personal, luas, dan akurat.

DAFTAR PUSTAKA

- [1] S. W. Aji, T. S. Wicaksono, M. Yosep, and T. Sulistyono, "Sistem Pencarian dan Rekomendasi Musik Dengan Analisis Lirik Lagu Untuk Pengalaman Mendengarkan Yang Lebih Personal Menggunakan Content Based-Filtering," *Jurnal Mnemonic*, vol. 8, no. 2, pp. 169–174, 2025, doi: 10.36040/mnemonic.v8i2.14879.
- [2] J. S. Situmorang, P. P. Adikara, and D. E. Ratnawati, "Rekomendasi Lagu Berdasarkan Lirik Lagu Menggunakan Metode N-Gram dan Cosine Similarity," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 6, pp. 2279–2288, 2021, doi: [Online]. Available: <http://j-ptiik.ub.ac.id>.
- [3] A. P. Muriaa, H. Sujainia, and H. S. Pratiwia, "Sistem Rekomendasi Artikel sebagai Acuan Studi Literatur Article Recommendation System as a Reference for Literature Studies Using the N-Gram Method," *JURISTIK (Jurnal Ris. Sains dan Teknol. Inform.)*, vol. 01, no. 1, pp. 69–84, 2023, doi: 10.26418/juristi.v1i1.61170.
- [4] N. N. K. Sari, R. Priskila, and P. B. A. A. Putra, "Implementasi Content-Based Filtering Menggunakan TF-IDF dan Cosine Similarity Untuk Sistem Rekomendasi Resep Makanan," *J. Teknol. Inf.*, vol. 18, no. 1, pp. 43–51, 2024, doi: 10.47111/JTI.
- [5] C. Virginia, Christy, and H. Irsyad, "Rekomendasi Lagu Berdasarkan Jenis Mood Menggunakan Tf-Idf Dan Cosine Similarity," *J. Insa. Comtech*, vol. 10, no. 1, pp. 22–29, 2025 [Online]. Available: <http://ejournal.unira.ac.id>.
- [6] A. Y. Timur and A. N. Rohman, "Sistem Rekomendasi Lagu Indonesia Menggunakan Metode Content-Based Filtering Dan Cosine Similarity," *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, vol. 13, no. 1, pp. 1415–1423, 2025 doi: 10.23960/jitet.v13i1.5949.
- [7] A. T. Adiyanto, P. Studi, T. Informatika, and U. S. Semarang, "Information Retrieval Sistem Kearsipan Pencarian Dokumen Di Dinas Pemberdayaan Perempuan Dan Perlindungan Anak Kota Semarang Menggunakan Metode Vector Space Model," vol. 7, no. 1, 2022, doi: 10.51544/jurnalmi.v7i1.2538.
- [8] W. O. W. Makmun, I. P. Ningrum, and A. M. Sajiah, "Penerapan Vector Space Model (Vsm) Pada Sistem Pencarian Artikel Arkeologi," *semantik*, vol. 8, no. 1, p. 69, 2022, doi: 10.55679/semantik.v8i1.15346.
- [9] S. Panambunan, A. Hasibuan, and K. Santa, "Implementasi Vector Space Model pada Aplikasi Pengarsipan Berbasis Web di POLDA Sulut," *J. Minfo Polgan*, vol. 14, no. 2, pp. 2131–2137, 2025, doi: 10.33395/jmp.v14i2.15203.
- [10] C. Y. Fuady, V. C. Mawardi, and T. Sutrisno, "Implementasi Metode Content Based Filtering Untuk Rekomendasi Film Berdasarkan Genre Dan Sinopsis," *J. Serina Sains, Tek. dan Kedokt.*, vol. 1, no. 2, pp. 227–236, 2023, doi: 10.24912/jsstk.v1i2.28509.
- [11] A. Fitria, S. Zaman, and M. Y. Ainul, "Sistem Rekomendasi Lowongan Pekerjaan Menggunakan Content-based filtering," *JEPIN (Jurnal Edukasi dan Penelit. Inform.)*, vol. 10, no. 3, pp. 421–427, 2024 [Online]. Available: [https://jurnal.untan.ac.id/..](https://jurnal.untan.ac.id/)
- [12] F. E. H. Trianto, "Pengaruh menyanyikan lagu rohani terhadap spiritualitas mahasiswa teologi di STTIAA Mojokerto," *Geneva: Jurnal Teologi dan Pendidikan Agama Kristen*, vol. 5, no. 2, pp. 73–83, Des. 2023. [Online]. Available <https://www.stttaa.ac.id/e-journal>.
- [13] W. Hersh, "Search still matters: Information retrieval in the era of generative AI," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 2159–2161, 2024, doi: 10.1093/jamia/ocae014.
- [14] J. Wong, I. Sanu, and H. Irsyad, "Implementasi TF-IDF, Cosine Similarity, dan Logistic Regression Pada Rekomendasi Buku Berdasarkan Mood Pembaca Dengan Data Oversampling," *Device J. Inf. Syst. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, pp. 142–154, 2025, doi: 10.46576/device.v6i1.6499.
- [15] R. Al Rasyid and D. H. U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 1, pp. 170–178, 2024, doi: 10.35870/jtik.v8i1.1416.
- [16] S. G. Rumagit, D. Paseru, and T. C. Suwanto, "Implementasi Algoritma Vector Space Model untuk Pencarian Dokumen pada Aplikasi Monitoring Uji Material (Studi Kasus: Lab. BPJN Jayapura)," *J. Ilm. Sains*, vol. 24, no. April, pp. 57–69, 2024, doi: 10.35799/jis.v24i1.52967.
- [17] J. M. Alyza, F. S. Utomo, Y. Purwati, B. A. Kusuma, and M. S. Azmi, "Music Recommendation System Based on Cosine Similarity and Supervised Genre Classification," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 9, no. 1, pp. 77–80, 2023, doi: 10.33480/jitk.v9i1.4324.