JURNAL FASILKOM
Volume 15 No. 3 | Desember 2025: 474-481

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Klasterisasi Topik Khotbah Pendeta Di GBI MPI Palembang Dengan Metode DBSCAN

Kristian Fernando[1], Hafiz Irsyad[2]
[1,2]Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
[1]kristianfernando_2226250070@mhs.mdp.ac.id, [2]hafizirsyad@mdp.ac.id*

Abstract

Comprehensive evaluation of the teaching curriculum proportion at GBI Rayon 15 Musi Palem Indah (MPI) Palembang is a fundamental element in ensuring the doctrinal health of the congregation. However, the current evaluation process is inefficient due to reliance on manual mapping of ever-growing sermon archives. This conventional method carries a high risk of subjectivity bias, making it difficult for church leadership to objectively observe teaching theme trends. This study addresses this issue by developing an automated document clustering system based on Text Mining to process 406 sermon summary documents from the 2023-2025 period. The methodology includes preprocessing, Term Frequency-Inverse Document Frequency (TF-IDF) weighting to highlight distinctive theological terms, and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. DBSCAN was specifically selected for its superiority in handling data with varying densities and its ability to isolate outliers without requiring a static cluster count parameter. Test results indicate an optimal configuration at Epsilon 0.3 and MinPts 3, yielding very high internal validity with a Silhouette Coefficient of 0.8888 and forming 32 core topic clusters. Significant findings reveal a high noise ratio (71%), which effectively separates incidental topics, such as holiday celebrations, from regular material. Practically, these results serve as an early warning system mechanism for the church to detect doctrinal imbalances or material gaps, providing a strategic data-driven foundation for holistic curriculum improvement.

Keywords: DBSCAN, teaching evaluation, text clustering, text mining, TF-IDF.

Abstrak

Evaluasi komprehensif terhadap proporsi kurikulum pengajaran di GBI Rayon 15 Musi Palem Indah (MPI) Palembang adalah elemen fundamental untuk memastikan kesehatan doktrinal jemaat. Kendati demikian, proses evaluasi saat ini berjalan tidak efisien karena bergantung pada pemetaan manual terhadap arsip khotbah yang terus bertambah. Metode konvensional ini memiliki risiko bias subjektivitas tinggi, sehingga menyulitkan pimpinan gereja untuk melihat kecenderungan tema pengajaran secara objektif. Penelitian ini mengatasi masalah tersebut dengan mengembangkan sistem klasterisasi otomatis berbasis *Text Mining* guna mengolah 406 dokumen ringkasan khotbah periode 2023-2025. Metodologi meliputi *preprocessing*, pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk menonjolkan istilah teologis distingtif, dan algoritma *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). DBSCAN dipilih secara spesifik karena keunggulannya menangani data berdensitas beragam dan kemampuan mengisolasi *outlier* tanpa memerlukan parameter jumlah klaster statis. Hasil pengujian menunjukkan konfigurasi optimal pada *Epsilon* 0.3 dan *MinPts* 3, menghasilkan validitas internal sangat tinggi dengan *Silhouette Coefficient* 0.8888 dan membentuk 32 klaster topik inti. Temuan signifikan menunjukkan tingginya rasio *noise* (71%) yang justru efektif memisahkan topik insidental, seperti perayaan hari raya, dari materi reguler. Secara praktis, hasil ini berfungsi sebagai mekanisme peringatan dini (*early warning system*) bagi gereja untuk mendeteksi ketimpangan doktrin atau kekosongan materi, menyediakan landasan strategis berbasis data untuk perbaikan kurikulum yang lebih holistik.

Kata kunci: DBSCAN, evaluasi pengajaran, klasterisasi teks, *text mining*, TF-IDF

1. Pendahuluan

Gereja Bethel Indonesia (GBI) Rayon 15 Musi Palem Indah (MPI) Palembang, sebagai institusi yang aktif menyelenggarakan kegiatan rohani, menghasilkan arsip data tekstual yang masif dari materi khotbah tahunan. Arsip ini bukan sekadar kumpulan teks, melainkan representasi dari arah pengajaran teologis yang membentuk spiritualitas jemaat. Namun, kekayaan data ini belum dimanfaatkan secara optimal karena proses evaluasi tema pengajaran masih bergantung pada metode konvensional. Pimpinan gereja harus membaca, mengingat kembali, dan mengkategorikan ratusan judul serta ringkasan khotbah secara manual. Proses ini memiliki kelemahan fundamental: inefisiensi waktu yang signifikan dan rentan terhadap bias subjektivitas pengamat, di mana persepsi evaluator dapat mendistorsi pemetaan topik yang sebenarnya.

Dalam konteks *Data Science*, teks khotbah merupakan entitas data tidak terstruktur (*unstructured data*)

P-ISSN : 2089-3353
E-ISSN : 2808-9162

dengan kompleksitas tinggi. Suryadi menekankan bahwa khotbah memiliki karakteristik linguistik yang unik dibandingkan teks berita atau akademis [1]. Khotbah sering menggunakan metafora rohani (seperti "Domba", "Gembala", "Buah"), repetisi retoris untuk penekanan, dan kosakata teologis yang khas. Selain itu, teks khotbah sering kali mengandung ambiguitas semantik di mana satu kata dapat memiliki makna ganda tergantung konteks teologisnya. Kompleksitas ini menuntut pendekatan teknologi *Text Mining* yang tidak hanya mampu menghitung frekuensi kata, tetapi juga memetakan kepadatan makna untuk mengekstrak pola tersembunyi secara otomatis dan objektif.

Tantangan utama dalam penelitian ini adalah memilih algoritma klasterisasi yang paling adaptif terhadap karakteristik teks pendek (*short text*) seperti judul dan ringkasan khotbah. Penelitian terdahulu di bidang klasterisasi teks didominasi oleh algoritma *K-Means* dan *Latent Dirichlet Allocation* (LDA). K-Means pada klasterisasi topik komplain dengan hasil *Silhouette Coefficient* 0.70 [2]. Namun, kelemahan fatal K-Means untuk kasus ini adalah kewajiban menentukan jumlah klaster (*k*) di awal. Dalam konteks arsip khotbah tahunan, jumlah topik bersifat dinamis dan tidak dapat diprediksi, sehingga menebak nilai *k* dapat menyebabkan *under-clustering* atau *over-clustering*. Di sisi lain, metode LDA yang digunakan seringkali menghasilkan topik yang tumpang tindih (*overlapping*) dengan nilai validitas moderat (*Silhouette* 0.62) [3], yang menyulitkan pimpinan gereja untuk menarik batas tegas antar tema pengajaran.

Untuk mengatasi keterbatasan tersebut, penelitian ini mengadopsi algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*). Dalam konteks analitika homiletika, kemampuan DBSCAN untuk mengisolasi *noise* memiliki implikasi strategis yang krusial. Berbeda dengan pendekatan klasterisasi konvensional yang memaksa setiap data masuk ke dalam kelompok tertentu, *noise* dalam penelitian ini berfungsi sebagai mekanisme filtrasi untuk memisahkan topik-topik insidental seperti khotbah tamu, perayaan hari raya (Natal/Paskah), atau respons terhadap isu sosial sesaat dari pola pengajaran rutin. Identifikasi ini penting untuk menjaga kemurnian klaster topik utama (*Core Curriculum*), sehingga pimpinan gereja dapat membedakan secara tegas antara doktrin inti yang diajarkan secara konsisten dengan materi suplemen yang bersifat *event-based.*

Penerapan *Text Mining* pada korpus khotbah menghadapi tantangan linguistik unik berupa densitas tinggi metafora rohani (seperti penggunaan kata 'Garam' atau 'Terang' yang bermakna konotatif teologis, bukan denotatif harfiah) serta ambiguitas semantik di mana satu kata dapat memiliki makna ganda tergantung konteksnya. Pendekatan berbasis densitas (*density-based*) seperti DBSCAN dinilai lebih adaptif terhadap karakteristik ini dibandingkan metode *K-Means* atau *Latent Dirichlet Allocation* (LDA).

Penelitian terdahulu menunjukkan bahwa *K-Means* memiliki kelemahan fundamental dalam kasus ini karena mengharuskan penentuan jumlah klaster (*k*) di awal, padahal jumlah topik pengajaran gereja bersifat dinamis dan tidak diketahui (*unknown ground truth*). Sementara itu, LDA seringkali menghasilkan topik yang tumpang tindih (*overlapping*) dengan batas tema yang kabur. Sebaliknya, Vladimir dkk. membuktikan bahwa DBSCAN unggul pada teks pendek dengan nilai *Silhouette* tinggi karena kemampuannya membentuk klaster berdasarkan kerapatan semantik kata-kata teologis yang muncul bersamaan (*co-occurrence*) tanpa terpengaruh variasi panjang dokumen [4].

Meskipun demikian, penerapan DBSCAN tetap memiliki tantangan, terutama sensitivitas terhadap parameter *Epsilon* (ε). Ketidaktepatan dalam menentukan radius densitas dapat menyebabkan *over-segmentation* atau kegagalan mendeteksi topik pada data dengan variasi densitas yang ekstrem (sparse data), sehingga penentuan parameter optimal menjadi kunci keberhasilan implementasi ini.

Penelitian ini bertujuan untuk mengimplementasikan kombinasi metode pembobotan TF-IDF dan algoritma DBSCAN guna menghasilkan pemetaan topik khotbah yang objektif [5]. Lebih dari sekadar klasterisasi teknis, penelitian ini dirancang untuk memberikan kontribusi praktis bagi manajemen gereja, yakni menyediakan landasan data untuk penyusunan jadwal khotbah yang lebih berimbang, mengidentifikasi kesenjangan (*gaps*) dalam kurikulum pengajaran, dan mengevaluasi dominasi topik tertentu. Dengan demikian, sistem ini diharapkan dapat bertransformasi menjadi alat pendukung keputusan (*decision support tool*) yang meningkatkan efektivitas strategi penggembalaan jemaat di GBI MPI Palembang.

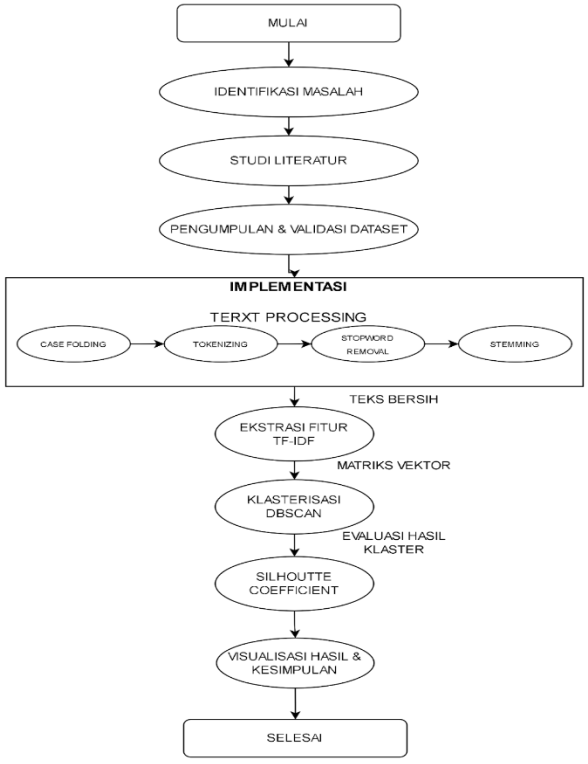
2. Metode Penelitian

Metodologi penelitian ini dirancang dengan pendekatan sistematis yang bertujuan untuk mentransformasi arsip data khotbah yang pada dasarnya merupakan entitas data tidak terstruktur (*unstructured data*) dengan kompleksitas linguistik tinggi menjadi informasi topik yang terorganisir dan bernilai strategis. Kerangka kerja penelitian ini mengadopsi alur kerja *Text Mining* yang komprehensif dan terstruktur, diawali dengan identifikasi masalah fundamental terkait inefisiensi evaluasi manual di lapangan, serta penguatan landasan teori melalui studi literatur. Tahap awal penelitian melibatkan pengumpulan data empiris berupa 406 dokumen ringkasan khotbah. Guna menjamin integritas teologis dan akurasi konten, proses validasi dilakukan melalui mekanisme verifikasi dua tahap (*two-stage verification*). Pertama, data teks diekstraksi secara manual dari arsip digital gereja (slide presentasi dan format PDF) untuk memastikan kelengkapan metadata. Kedua, dataset divalidasi langsung oleh pimpinan gereja *(Subject Matter Experts*) di GBI MPI Palembang untuk mengonfirmasi kesesuaian antara ringkasan teks

P-ISSN : 2089-3353
E-ISSN : 2808-9162

dengan esensi pengajaran yang disampaikan mimbar, sehingga meminimalisir risiko kesalahan interpretasi sebelum data diproses lebih lanjut.

Selanjutnya, proses berlanjut ke tahapan teknis yang meliputi pra-pemrosesan data (*text preprocessing*) untuk membersihkan teks mentah, diikuti oleh ekstraksi fitur menggunakan pembobotan TF-IDF guna mengubah data teks menjadi representasi vektor numerik. Inti dari metodologi ini adalah penerapan algoritma *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), yang dipilih karena kemampuannya dalam mengelompokkan dokumen berdasarkan kerapatan data serta memisahkan *noise* atau *outlier* secara otomatis. Kualitas hasil pengelompokan ini kemudian diukur menggunakan evaluasi *Silhouette Coefficient* untuk memastikan validitas klaster, di mana seluruh alur proses dari awal hingga akhir digambarkan secara visual dalam kerangka kerja pada Gambar 1.



Gambar 1. Kerangka Kerja Penelitian

Berdasarkan Gambar 1, tahapan penelitian diuraikan secara rinci sebagai berikut:

2.1 Identifikasi Masalah dan Pengumpulan Data

Tahap awal penelitian difokuskan secara mendalam pada identifikasi masalah fundamental di GBI Rayon 15 Musi Palem Indah (MPI) Palembang, di mana proses evaluasi pola pengajaran menghadapi kendala efisiensi akibat metode pemetaan topik yang masih dilakukan secara manual. Pertumbuhan jumlah dokumen arsip yang terus meningkat dengan dimensi data yang tinggi menjadi tantangan besar, menyebabkan informasi sulit dikategorikan secara efisien dan cepat tanpa bantuan sistem otomatis [6]. Di era digital saat ini, tantangan dalam mengelola dan mengakses informasi yang relevan dari tumpukan data teks tidak terstruktur seperti arsip khotbah menuntut pendekatan teknologi yang mampu mengelompokkan dokumen berdasarkan kesamaan topik atau tema secara objektif, guna menghindari bias subjektivitas pengamat [7].

Guna mengatasi permasalahan tersebut, studi literatur dilakukan untuk memperkuat landasan teori mengenai karakteristik data teks serta pemilihan algoritma yang tepat. Berbagai penelitian sebelumnya telah membuktikan efektivitas teknik *clustering* dalam mengorganisasi dokumen teks, mulai dari pengelompokan judul skripsi mahasiswa hingga analisis berita *online* [8]. Meskipun algoritma K-Means sering menjadi pilihan populer karena kemudahan implementasinya dalam berbagai aplikasi, penelitian ini mengkaji penerapan algoritma DBSCAN yang dinilai lebih unggul dalam menangani *noise*. Pemahaman mengenai tahapan *text mining*, khususnya *preprocessing*, juga menjadi fokus utama literatur karena data teks mentah yang belum terstruktur harus melalui tahapan pembersihan agar dapat diolah lebih lanjut secara akurat.

Dataset yang digunakan sebagai objek penelitian adalah arsip digital ringkasan khotbah periode tahun 2023 hingga 2025 yang diperoleh langsung dari *database* gereja, dengan total volume data berjumlah 406 dokumen. Data tersebut mencakup atribut-atribut kunci seperti nama file atau judul, nama pengkhotbah, periode waktu, serta ringkasan isi materi. Sebelum memasuki tahap pemrosesan, seluruh dataset divalidasi secara ketat oleh pihak gereja. Proses validasi ini sangat krusial untuk menjamin keakuratan materi dan konsistensi data, sejalan dengan prinsip pengembangan sistem manajemen pengetahuan yang menekankan pentingnya pengecekan data untuk menjaga validitas dan menghindari redudansi informasi.

2.2 Text Preprocessing

Tahapan *Text Preprocessing* memegang peranan vital dalam menjamin akurasi klasterisasi, mengingat data khotbah memiliki karakteristik tidak terstruktur dengan variasi morfologi yang tinggi [9] [10]. Setiap tahapan memiliki kontribusi spesifik terhadap pembentukan fitur, yaitu meliputi :

1. *Case Folding:* Menyeragamkan seluruh karakter menjadi huruf kecil (*lowercase*) untuk menghilangkan perbedaan case-sensitive yang tidak relevan (misalnya, "Tuhan" dan "tuhan") dan memastikan bahwa entitas teologis yang sama dihitung sebagai satu kesatuan frekuensi..

2. *Tokenizing:* Memecah aliran teks ringkasan khotbah menjadi unit-unit analisis diskrit (token). Pada teks khotbah yang seringkali merupakan transkrip lisan, proses ini krusial untuk memisahkan kata dari tanda baca yang dapat mendistorsi makna.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

3. *Stopword Removal:* Mengeliminasi kata-kata penghubung (seperti 'yang', 'dan', 'dari') yang memiliki frekuensi tinggi namun miskin makna semantik. Langkah ini mereduksi dimensi fitur secara signifikan, membiarkan algoritma fokus sepenuhnya pada kata kunci konten (*content words*).

4. *Stemming:* Mengembalikan kata ke bentuk dasarnya (misal: 'mengasihi' menjadi 'kasih'). Dalam konteks teologis, variasi imbuhan seringkali merujuk pada konsep doktrinal yang sama. *Stemming* memastikan bahwa bobot TF-IDF terkonsentrasi pada 'akar kata' topik, sehingga meningkatkan kepadatan (*density*) klaster yang terbentuk.

2.3 Ekstraksi Fitur (TF-IDF)

Dalam konteks teks keagamaan, metode pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*) menawarkan keunggulan distingtif dibandingkan sekadar menghitung frekuensi kata semata. Teks khotbah seringkali dipenuhi oleh kata-kata umum yang repetitif (seperti 'saudara', 'kita', atau 'mari'). Jika menggunakan frekuensi murni, kata-kata ini akan mendominasi dan mengaburkan topik sebenarnya.

TF-IDF bekerja dengan memberikan bobot rendah pada kata-kata umum tersebut (karena nilai *Inverse Document Frequency*-nya rendah) dan sebaliknya memberikan bobot tinggi pada istilah teologis spesifik yang unik pada dokumen tertentu (seperti 'Eskatologi', 'Tawarikh', atau 'Pentakosta'). Mekanisme ini memastikan bahwa fitur yang digunakan oleh algoritma DBSCAN untuk menghitung jarak antar-dokumen benar-benar merepresentasikan inti sari pengajaran, bukan sekadar gaya bahasa penyampaian.

Data teks yang telah bersih ditransformasi menjadi representasi vektor numerik. Pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan. Bobot W_{ij} untuk kata *i* dalam dokumen *j* dihitung dalam persamaan (1):

$$W_{ij} = TF_{ij} x\ IDF_i \tag{1}$$

Dimana TF_{ij} *(Term Frequency)* dihitung dengan persamaan (2) sebagai berikut :

$$TF_{ij} = \frac{F_{ij}}{\sum_k F_{kj}} \tag{2}$$

Dengan F_{ij} adalah jumlah kemunculan kata ke-i dalam dokumen ke-*j*, dan penyebutnya adalah total kata dalam dokumen ke-*j*.

Sementara itu, IDF_i *(Inverse Document Frequency)* dihitung dengan persamaan (3) Sebagai berikut :

$$IDF_i = log\left(\frac{N}{dF_i}\right) \tag{3}$$

Dimana *N* adalah total dokumen dan dF_i adalah jumlah dokumen yang mengandung kata ke-*i* Metode ini memberikan bobot rendah pada kata yang muncul di hampir semua dokumen (kata umum) dan bobot tinggi pada kata yang muncul sering di dokumen tertentu tetapi jarang di dokumen lain. Hal ini efektif untuk menonjolkan kata kunci topik spesifik.

2.4 Klasterisasi DBSCAN

Algoritma DBSCAN diterapkan pada matriks vektor TF-IDF. Berbeda dengan algoritma berbasis *centroid*, DBSCAN bekerja dengan konsep densitas. Dua parameter kunci yang diuji adalah:

1. *Epsilon* (ε) : Jarak maksimum antara dua sampel untuk dianggap bertetangga.

2. *MinPts*: Jumlah minimum sampel dalam lingkungan (ε) agar sebuah titik dianggap sebagai titik inti (*core point*).

Titik yang tidak memenuhi syarat *core point* dan tidak bertetangga dengan *core point* akan diklasifikasikan sebagai *Noise* (-1). Jarak antar vektor dihitung menggunakan Euclidean Distance.

Perhitungan jarak antar dokumen menggunakan *Euclidean Distance* dapat dilihat pada Persamaan (4):

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(pi - qi)^2} \tag{4}$$

dimana *d(p, q)* adalah jarak antara titik data *p* dan *q*, dan *n* adalah jumlah dimensi fitur.

2.5 Evaluasi *Silhouette Coefficient*

Kualitas klaster dievaluasi menggunakan *Silhouette Coefficient (S)* yang mengukur kohesi (kedekatan dalam klaster) dan separasi (jarak antar klaster).

Perhitungan *Silhouette Coefficient* dapat dilihat pada persamaan (5) sebagai berikut :

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

Yang dimana *a(i)* adalah rata-rata jarak dari titik ke semua titik lain di dalam klaster yang sama, dihitung dengan Persamaan (6) sebagai berikut :

$$a(i) = \frac{1}{|Ci| - 1} \sum_{j \in Ci, i \neq j} d(i,j) \tag{6}$$

Untuk *b(i)* adalah rata-rata jarak terkecil dari titik i ke semua titik di klaster lain yang bukan anggotanya, dihitung dengan Persamaan (7) sebagai berikut :

$$b(i) = min_{k \neq i}\left(\frac{1}{|C_k|}\sum_{j \in C_k} d(i,j)\right) \tag{7}$$

Dari hasil perhitungan maka nilai *Silhouette Coefficient* berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan struktur klaster yang baik.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

3. Hasil dan Pembahasan

3.1 *Preprocessing* dan Ekstraksi Fitur

Implementasi penelitian diawali dengan pengolahan dataset yang terdiri dari 406 dokumen ringkasan khotbah. Mengingat data ini merupakan data teks tidak terstruktur dengan variasi linguistik yang tinggi, tahapan pra-pemrosesan (*text preprocessing*) menjadi langkah fundamental yang sangat krusial. Sebagaimana dijelaskan oleh Simanjuntak dkk., data teks mentah seringkali memiliki dimensi yang tinggi dan mengandung banyak *noise*, sehingga sulit untuk dikategorikan secara efisien tanpa melalui proses pembersihan yang sistematis. Tanpa tahapan ini, kualitas klasterisasi dapat menurun secara signifikan karena algoritma akan memproses karakter atau kata yang tidak relevan sebagai fitur pembeda.

Proses *text preprocessing* dalam penelitian ini dilakukan melalui serangkaian tahapan yang meliputi *case folding*, *tokenizing*, *Stopword Removal*, dan *stemming*. Pada tahapan *Stopword Removal*, penelitian ini memanfaatkan pustaka standar Bahasa Indonesia (Sastrawi/NLTK) untuk mengeliminasi kata-kata fungsional yang memiliki frekuensi tinggi namun minim makna distingtif, seperti kata sambung ("yang", "dan") dan kata depan ("di", "dari"). Penghapusan kata-kata ini krusial untuk mereduksi dimensi fitur (*dimensionality reduction*), sehingga algoritma klasterisasi tidak terdistraksi oleh kesamaan struktur kalimat, melainkan fokus pada kesamaan konten teologis.

Tantangan teknis utama ditemukan pada tahap *Tokenizing*. Teks ringkasan khotbah seringkali memiliki struktur non-baku hasil transkripsi lisan atau penggunaan format poin (*bullet points*) pada slide presentasi yang menyatukan kata dengan tanda baca (misalnya "Tuhan," atau "Yesus."). Proses tokenisasi dalam penelitian ini dirancang untuk memisahkan karakter non-alfanumerik tersebut secara presisi. Hal ini memperbaiki akurasi klasterisasi secara signifikan dengan memastikan bahwa sebuah istilah teologis (misalnya "Iman") dihitung sebagai satu entitas fitur yang utuh, terlepas dari tanda baca yang mengikutinya.

Untuk memastikan kesiapan data sebelum masuk ke tahap pemodelan, dilakukan validasi kualitatif terhadap hasil *preprocessing*. Pengecekan dilakukan dengan meninjau sampel *term list* hasil *stemming* guna memastikan tidak terjadi perubahan makna yang ekstrim pada istilah teologis (misalnya memastikan kata "Allah" tidak terpotong secara keliru menjadi "Llah"). Validasi ini menjamin bahwa matriks fitur yang terbentuk benar-benar merepresentasikan esensi pengajaran gereja yang akurat.

Gambaran transformasi data dari teks mentah hingga menjadi token bersih dapat dilihat pada Tabel 1.

Tabel 1. Perbandingan Data Sebelum dan Sesudah *Preprocessing*

Teks Normal	*Case Folding*	*Tokenizing*	*Stopword Removal*	*Stemming*
Khotbah ini menekaenkan pentingnya memiliki gambar diri yang sehat yang berakar pada identitas di dalam Kristus, bukan pada pandangan duniawi. Pemulihan gambar diri yang rusak oleh masa lalu, kegagalan, atau perkataan orang lain dapat terjadi melalui kasih dan pengampunan Kristus.	khotbah ini menekankan pentingnya memiliki gambar diri yang sehat yang berakar pada identitas di dalam kristus bukan pada pandangan duniawi pemulihan gambar diri yang rusak oleh masa lalu kegagalan atau perkataan orang lain dapat terjadi melalui kasih dan pengampunan kristus.	['khotbah', 'ini', 'menekankan', 'pentingnya', 'memiliki', 'gambar', 'diri', 'yang', 'sehat', 'yang', 'berakar', 'pada', 'identitas', 'di', 'dalam', 'kristus', 'bukan', 'pada', 'pandangan', 'duniawi', 'pemulihan', 'gambar', 'diri', 'yang', 'rusak', 'oleh', 'masa', 'lalu', 'kegagalan', 'atau', 'perkataan', 'orang', 'lain', 'dapat', 'terjadi', 'melalui', 'kasih', 'dan', 'pengampunan', 'kristus']	['khotbah', 'menekankan', 'memiliki', 'gambar', 'sehat', 'berakar', 'identitas', 'kristus', 'pandangan', 'duniawi', 'pemulihan', 'gambar', 'rusak', 'kegagalan', 'perkataan', 'orang', 'kasih', 'pengampunan', 'kristus', 'gambar', 'pulih', 'hidup', 'maksimal', 'berbuah', 'menggenapi', 'tujuan', 'tuhan']	['khotbah', 'tekan', 'milik', 'gambar', 'sehat', 'akar', 'identitas', 'kristus', 'pandang', 'duniawi', 'pulih', 'gambar', 'rusak', 'gagal', 'kata', 'orang', 'kasih', 'ampun', 'kristus']

Pada Tabel 1, terlihat transformasi data dari teks mentah menjadi token bersih. Kalimat kompleks yang mengandung banyak kata sambung berhasil direduksi menjadi kata-kata kunci inti.

Setelah data teks dibersihkan, tahap selanjutnya adalah transformasi data menjadi representasi vektor numerik menggunakan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Penerapan TF-IDF terbukti efektif dalam menonjolkan fitur kata yang relevan. Berdasarkan hasil percobaan pada Tabel.2 kata-kata umum memiliki nilai bobot 0.000 atau sangat rendah pada dokumen tertentu. Sebaliknya, kata kunci teologis spesifik seperti 'Kristus' (0.228) dan 'Pulih' (0.259) mendapatkan bobot yang signifikan. Tingginya bobot pada istilah-istilah ini membantu algoritma DBSCAN untuk menarik garis demarkasi yang tegas antar-klaster topik, membedakan misalnya antara topik 'Pemulihan Gambar Diri' dengan topik 'Pertobatan Bangsa' berdasarkan densitas kata kunci unik tersebut.

Hasil dari TF-IDF direpresentasikan sebagai berikut pada Tabel 2.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Tabel 2. Hasil Pembobotan TF-IDF

Index	hidup	kasih	kata	khotbah	kristus	orang	pulih
0	0.074	0.119	0.164	0.048	0.228	0.110	0.259
1	0.000	0.000	0.173	0.051	0.000	0.000	0.000
2	0.000	0.000	0.000	0.045	0.000	0.103	0.122
3	0.168	0.270	0.000	0.054	0.259	0.124	0.000
4	0.083	0.000	0.000	0.107	0.000	0.000	0.145

Matriks bobot yang tersaji pada Tabel 2 di atas menegaskan kesiapan data untuk diproses secara komputasi. Nilai-nilai vektor tersebut bertindak sebagai koordinat spasial yang memungkinkan algoritma DBSCAN mengukur kedekatan (*proximity*) antar-dokumen menggunakan perhitungan jarak. Dengan demikian, dokumen-dokumen yang memiliki kemiripan profil bobot kata akan ditarik mendekat untuk membentuk klaster padat, sementara dokumen dengan profil bobot yang menyimpang secara signifikan akan terisolasi sebagai *noise* pada tahapan selanjutnya.

3.2 Pengujian Parameter DBSCAN

Pengujian dilakukan menggunakan skenario *Grid Search* untuk mencari kombinasi optimal antara parameter *Epsilon* (ε) dan *Minimum Points* (MinPts). Rentang pengujian dilakukan pada (ε) = 0.3 sampai 0.5 dan MinPts = 2 sampai 5. Hasil evaluasi kualitas klaster menggunakan *Silhouette Coefficient* dirangkum dalam Tabel 3.

Tabel 3. Rekapitulasi Hasil Pengujian *Silhouette Coefficient*

Percobaan Ke-	*Epsilon (ε)*	*MinPts*	*Jumlah Klaster Terbentuk*	*Noise*	*Silhouette Score*
1	0.3	2	52	250	0.8792
2	0.3	3	32	290	0.8888
3	0.3	4	12	350	0.8488
4	0.3	5	6	375	0.8438
5	0.4	2	55	236	0.8363
6	0.4	3	35	276	0.8532
7	0.4	4	15	336	0.8415
8	0.4	5	8	364	0.8014
9	0.5	2	61	216	0.7711
10	0.5	3	35	268	0.8018
11	0.5	4	17	322	0.7596
12	0.5	5	9	356	0.7681

Analisis mendalam terhadap Tabel 2 menunjukkan pola perilaku algoritma sebagai berikut:

1. Pengaruh *Epsilon* (ε): Hasil percobaan menunjukkan fenomena di mana perluasan jangkauan *Epsilon* (ε) justru berbanding terbalik dengan kualitas klaster. Pada rentang Epsilon (ε) = 0.3, rata-rata skor *Silhouette Coefficient* sangat tinggi (> 0.84). Namun, ketika Epsilon (ε) dinaikkan menjadi 0.5 (percobaan 9 - 12), skor rata-rata mengalami penurunan drastis ke kisaran 0.75 - 0.80. Secara teoritis, Epsilon merepresentasikan jari-jari maksimum untuk menganggap dua dokumen sebagai 'tetangga' yang memiliki kemiripan semantik. Penurunan kualitas ini terjadi karena pada *Epsilon* (ε) yang lebih besar (0.5), algoritma menjadi terlalu permisif atau toleran terhadap perbedaan fitur kata. Radius yang terlalu luas menyebabkan dokumen-dokumen yang sebenarnya memiliki relevansi topik yang lemah atau samar ikut tertarik masuk ke dalam satu klaster. Hal ini menimbulkan dua dampak negatif pada perhitungan validitas yaitu Penurunan Kohesi (*Cohesion*), di mana jarak rata-rata antar-dokumen di dalam satu klaster menjadi semakin jauh karena anggota klaster semakin heterogen, dan Penurunan Separasi (*Separation*), di mana batas antar-klaster menjadi kabur dan tumpang tindih (*overlapping*), karena dokumen yang berada di pinggiran klaster (*border points*) kini dapat menjangkau dokumen dari topik lain yang berdekatan.

Sebaliknya, penggunaan radius yang lebih ketat (*Epsilon* (ε) = 0.3) terbukti paling efektif untuk korpus khotbah. Radius sempit ini memaksa algoritma untuk hanya mengelompokkan dokumen-dokumen dengan kemiripan leksikal yang sangat tinggi yakni penggunaan kata kunci teologis yang persis sama sehingga menghasilkan struktur klaster yang sangat padat (*dense*) dan terpisah secara tegas satu sama lain.

2. Pengaruh *MinPts*: Pengujian parameter menunjukkan pola signifikan di mana peningkatan nilai *Minimum Points (MinPts)* berbanding lurus dengan peningkatan jumlah *noise* dan penurunan jumlah klaster. Sebagaimana terlihat pada data percobaan, peningkatan *MinPts* dari 2 ke 5 pada *Epsilon* (ε) = 0.3 menyebabkan lonjakan jumlah *noise* dari 250 menjadi 375 dokumen. Secara teknis, *MinPts* berfungsi sebagai ambang batas densitas minimum; ketika nilai ini dinaikkan, algoritma memperketat syarat pembentukan sebuah kelompok. Akibatnya, topik-topik spesifik yang memiliki frekuensi kemunculan rendah (misalnya topik yang hanya dibahas 3-4 kali dalam setahun) yang sebelumnya terdeteksi sebagai klaster kecil pada *MinPts* rendah, kini gagal memenuhi syarat ambang batas baru tersebut. Fenomena ini menyebabkan kelompok-kelompok kecil tersebut terurai dan dokumen-dokumen di dalamnya terklasifikasi sebagai *noise* atau *outlier*. Hal ini mengonfirmasi karakteristik data khotbah yang memiliki distribusi 'ekor panjang' (*long tail distribution*), di mana terdapat banyak variasi topik mikro yang jarang berulang namun tetap ada.

Tingginya persentase data yang teridentifikasi sebagai *Noise* (mencapai 71% atau 290 dokumen pada konfigurasi optimal) bukanlah indikator kegagalan, melainkan temuan analitis penting. Analisis manual menunjukkan bahwa dokumen *noise* ini berisi topik-topik *niche* (ceruk) atau insidental seperti khotbah tamu, perayaan hari raya (Natal/Paskah), atau respons isu sosial sesaat. Di sinilah letak keunggulan DBSCAN dibandingkan metode *centroid-based* seperti K-Means. Jika K-Means akan memaksa topik-topik unik ini masuk ke dalam klaster besar (yang menurunkan nilai kohesi), DBSCAN secara agresif mengisolasinya. Implikasinya, 32 klaster yang

P-ISSN : 2089-3353
E-ISSN : 2808-9162

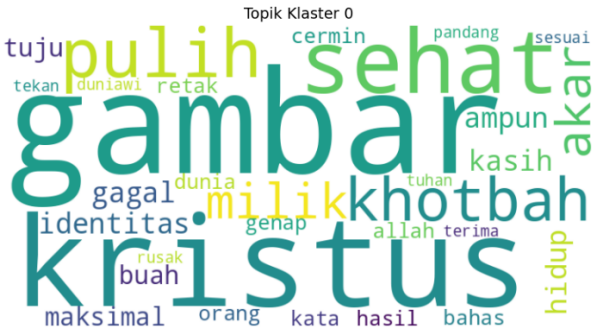
terbentuk benar-benar merupakan 'Sari Pati' pengajaran dengan kepadatan tema sangat tinggi (*Strong Structure*, Silhouette 0.8888), sementara *noise* berfungsi sebagai indikator variasi pengayaan materi.

Namun, kondisi ini membawa tantangan berupa potensi *blind spot* evaluasi. Jika pengambil keputusan hanya berfokus pada 32 klaster utama, gereja berisiko kehilangan sinyal awal terhadap tren pengajaran baru yang sedang muncul. Oleh karena itu, secara praktis, hasil ini memberikan peta navigasi ganda bagi pimpinan gereja yaitu menggunakan 32 klaster utama (seperti 'Keluarga', 'Iman', 'Misi') untuk memvisualisasikan proporsi pengajaran rutin secara kuantitatif guna menyeimbangkan doktrin, dan mengategorikan ulang data *noise* secara berkala sebagai 'Materi Suplemen' untuk memastikan tidak ada pergeseran teologis krusial yang luput dari pengamatan.

3.3 Visualisasi dan Interpretasi Topik

Setelah validasi matematis melalui *Silhouette Coefficient*, tahap selanjutnya adalah interpretasi semantik untuk memahami makna teologis di balik setiap klaster yang terbentuk. Sebagaimana dijelaskan oleh Defriani,dkk. visualisasi merupakan langkah krusial untuk memverifikasi bahwa setiap klaster yang terbentuk memiliki topik spesifik, di mana kata-kata di dalamnya saling terkait secara semantik dengan topik tersebut [11]. Dalam penelitian ini, teknik *Word Cloud* digunakan untuk merepresentasikan distribusi kata kunci, di mana ukuran huruf mencerminkan frekuensi kemunculan kata dalam klaster; semakin sering sebuah kata muncul, semakin besar ukurannya, sehingga memudahkan identifikasi tema dominan secara visual.

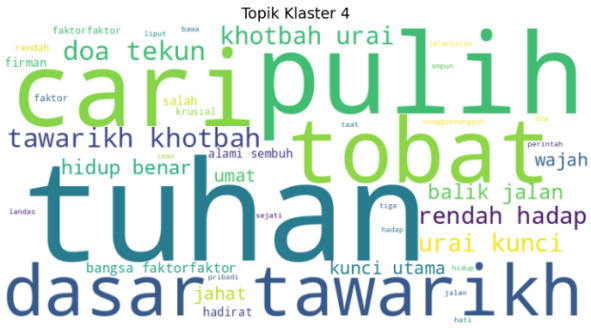
Pada Gambar 2 (klaster 0), kata kunci dominan yang muncul adalah "Gambar", "Pulih", "Kristus", dan "Identitas".



Gambar 2. Visualisasi *Word Cloud* Topik Klaster 0

Dalam konteks teologi Kristen, kemunculan bersamaan dari kata-kata ini bukan sekadar kebetulan statistik, melainkan merujuk secara spesifik pada konsep doktrinal *Imago Dei* (Manusia diciptakan segambar dengan Allah). Algoritma DBSCAN berhasil mengidentifikasi pola bahwa topik pemulihan gambar diri yang rusak akibat dosa selalu dikaitkan dengan identitas di dalam Kristus. Hal ini menunjukkan bahwa sistem mampu mengelompokkan khotbah-khotbah yang bersifat doktrinal-psikologis ke dalam satu kelompok kohesif yang terpisah secara tegas dari topik lain.

Visualisasi pada Gambar 3 (Klaster 4) menunjukkan dominasi kata "Tawarikh", "Tobat", "Cari", dan "Pulih".



Gambar 3. Visualisasi *Word Cloud* Topik Klaster 4

Ini merujuk pada prinsip teologis kausalitas dari nats 2 Tawarikh 7:14, di mana pemulihan keadaan ("sembuh/pulih") adalah hasil dari tindakan merendahkan diri dan pertobatan ("balik jalan"). Perbedaan nuansa antara "Pemulihan Identitas" (Klaster 0) dan "Pemulihan Kondisi Bangsa/Umat" (Klaster 4) berhasil dipisahkan dengan sangat tajam oleh algoritma.

3.4 Analisis Komparatif dengan Penelitian Terdahulu

Untuk memvalidasi keunggulan metode yang diusulkan, dilakukan perbandingan hasil dengan penelitian-penelitian sebelumnya yang menggunakan metode berbeda (K-Means dan LDA) pada domain teks serupa. Perbandingan disajikan dalam Tabel 4.

Tabel 4. Perbandingan Performa Metode Klasterisasi Teks

Peneliti	Metode	Objek	Hasil (Silhouette)
Fitriyani (2024)	K-Means	Toko Obat	0.52
Siringoringo (2020)	LDA dan K-Means	Berita Online	0.62
Simanjuntak (2023)	K-Means dan Word2Vec	Berita Online	0.73
Penelitian Ini (2025)	DBSCAN + TF-IDF	Ringkasan Khotbah	0.888

Berdasarkan Tabel 3, terlihat bahwa metode DBSCAN yang diusulkan dalam penelitian ini menghasilkan *Silhouette Coefficient* yang paling tinggi (0.88) dibandingkan seluruh metode berbasis K-Means (0.62 - 0.73). Hal ini mengonfirmasi temuan Vladimir bahwa untuk teks pendek dengan variasi tinggi, pendekatan berbasis densitas (DBSCAN) lebih unggul daripada pendekatan berbasis *centroid* (K-Means) yang memaksakan setiap data masuk ke dalam klaster (sferis).

3.5 Analisis Fenomena *Noise*

Salah satu temuan penting dalam penelitian ini adalah tingginya jumlah dokumen yang dikategorikan sebagai *noise* (290 dokumen atau sekitar 71% dari dataset).

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Dalam konteks algoritma klasterisasi konvensional, angka ini mungkin dianggap sebagai kelemahan. Namun, dalam konteks evaluasi pengajaran gereja, ini merupakan temuan analitis yang berharga.

Noise merepresentasikan khotbah-khotbah yang memiliki topik unik, insidental, atau sangat spesifik sehingga tidak memiliki "tetangga" yang cukup untuk membentuk sebuah klaster padat. Contohnya adalah khotbah tamu, khotbah hari raya (Natal/Paskah), atau topik yang merespons isu sosial sesaat. Bagi manajemen gereja, data ini mengindikasikan bahwa materi pengajaran di GBI MPI sangat variatif dan kaya, tidak monoton hanya pada 32 topik utama yang terklaster. Hal ini memberikan wawasan bahwa pengajaran gereja memiliki keseimbangan antara topik doktrinal inti (32 klaster yang terbentuk) dan topik-topik pengayaan (yang terdeteksi sebagai *noise*).

4. Kesimpulan

Penelitian ini berhasil mengembangkan dan mengimplementasikan sistem klasterisasi otomatis untuk pemetaan topik khotbah di GBI MPI Palembang. Integrasi metode *Text Mining* dengan algoritma DBSCAN dan pembobotan TF-IDF terbukti efektif dalam menjawab tantangan inefisiensi dan subjektivitas metode manual.

Kesimpulan spesifik yang dapat ditarik meliputi:

1. Efektivitas Model: Konfigurasi parameter Epsilon (ε) = 0.3 dan MinPts = 3 merupakan model paling optimal yang menghasilkan 32 klaster topik dengan nilai *Silhouette Coefficient* 0.8888. Angka ini menunjukkan struktur klaster yang sangat kuat (*Strong Structure*) dan validitas pemisahan topik yang tegas.

2. Keunggulan Metodologis: Dibandingkan dengan metode K-Means dan LDA pada penelitian sejenis, DBSCAN terbukti lebih superior dalam menangani teks pendek berbahasa Indonesia, terutama karena kemampuannya menangani *noise* dan tidak memerlukan penentuan jumlah klaster awal.

3. Implikasi Praktis: Sistem ini berhasil memisahkan "Topik Inti" (seperti Iman, Keluarga, Pertobatan) dari "Materi Pengayaan" (*Noise*). Hasil ini memberikan landasan data kuantitatif bagi pimpinan gereja untuk menyusun jadwal khotbah yang lebih berimbang dan strategis di masa depan.

4. Penelitian selanjutnya disarankan untuk mengeksplorasi penggunaan metode representasi kata yang lebih canggih berbasis semantik, seperti *Word2Vec* atau *BERT Embeddings*, untuk menangkap konteks teologis yang lebih mendalam yang mungkin terlewatkan oleh pendekatan statistik TF-IDF [12].

5. Ucapan Terimakasih

Penulis mengucapkan terima kasih kepada GBI Rayon 15 Musi Palem Indah (MPI) Palembang yang telah memberikan akses terhadap arsip data khotbah serta dukungan selama proses penelitian ini berlangsung.

Daftar Rujukan

[1] Suryadi, R. 2022. Pengaruh Khotbah Alkitabiah Dari Pengkhotbah Terhadap Intensitas Beribadah. *JURNAL TABGHA,* 3(1), pp.26-38.

[2] G. H. Setiawan, M. D. A. Pranata, I. B. A. Arimbawa, I. W. P. Giri, and N. P. L. Carisa Dayani, 2025. "*Topic Clustering of Student Complaints Based on Semantic Meaning Using the indoBERT and K-Means Models*", *JAIC*, vol. 9, no. 4, pp. 1715–1721.

[3] Siringoringo, R., Jamaluddin, & Perangin-Angin, R., 2020. Pemodelan Topik Berita Menggunakan Latent Dirichlet Allocation dan K-Means Clustering. *Jurnal Informatika Kaputama (JIK)*, 4(2), pp.216-222.

[4] Vladimir, Z. V., Alamsyah, D., & Widhiarso, W., 2022. Klasterisasi Topik Skripsi Informatika dengan Metode DBSCAN. *Jurnal Algoritme*, 3(1), pp.82

[5] Rahman, A., Waskitho, R. B., Nuha, M. F. A. U., & Rakhmawati, N. A., 2021. Klasterisasi Topik Konten Channel Youtube Gaming Indonesia Menggunakan Latent Dirichlet Allocation. *Journal Information Engineering and Educational Technology (JIEET)*, 5(2), pp.78-83.

[6] Simanjuntak, H. T. A., Silaban, P. E. P., Manurung, J. K. S., & Sormin, V. H., 2023. Klasterisasi Berita Bahasa Indonesia dengan Menggunakan K-Means dan Word Embedding. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 10(3), pp.641-652.

[7] Adhitama, A., Hidayatullah, S., & Rahman, M., 2024. Klusterisasi Judul Berita Pada Website Detik Menggunakan Algoritma Kmeans. *Indonesian Journal of Innovation Science and Knowledge*, 1(3), pp.194-207.

[8] Defriani, M. R., & Muttaqin, M. R., 2020. Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa. *ILKOM Jurnal Ilmiah*, 12(2), pp.121-129.

[9] Kurniawan, E., & Hendrastuty, N., 2024. Penerapan Algoritma K-Means untuk Melakukan Klasterisasi pada Peringkasan Teks. *Jurnal Informatika Teknologi dan Sains (JINTEKS)*, 6(3), pp.514-520.

[10] Fitriyani, D., Jajuli, M., & Garno, G., 2024. Implementasi Algoritma K-Means untuk Klasterisasi dalam Pengelolaan Persediaan Obat (Studi Kasus Apotek Naza). *Jurnal Informatika dan Teknik Elektro Terapan*, 12(3), pp.1652-1658.

[11] Ariyanti, D., & Iswardani, K., 2020. Teks mining untuk klasifikasi keluhan masyarakat pada pemkot Probolinggo menggunakan algoritma Naïve Bayes. *Jurnal IKRA-ITH Informatika*, 4(3), pp.125-132.

[12] Asnada, M. I., Rahayudi, B., & Ridok, A., 2022. Pengelompokan Topik Skripsi Mahasiswa Fakultas Ilmu Komputer Universitas Brawijaya berdasarkan Judul pada Periode 2015-2019 menggunakan Metode Semi Supervised K-Means. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 6(1), pp.58-65.