# 

# IMPLEMENTASI SEMANTIC SEARCH DALAM SISTEM REKOMENDASI BUKU ONLINE GRAMEDIA MENGGUNAKAN MODEL TEXT-EMBEDDING-004 DAN COSINE SIMILARITY

# Tugas akhir

# diajukan untuk melengkapi

# persyaratan mencapai

# gelar sarjana

**NAMA : Danang Hapis Fadillah**

**NPM : 202143500557**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK DAN ILMU KOMPUTER**

**UNIVERSITAS INDRAPRASTA PGRI**

**2025**

**LEMBAR PERSETUJUAN UJIAN TUGAS AKHIR**

Nama : Danang Hapis Fadillah

NPM : 202143500557

Fakultas : Teknik dan Ilmu Komputer

Program Studi : Teknik Informatika

Judul Tugas Akhir : Implementasi Semantic Search dalam Sistem

Rekomendasi Buku Online Gramedia Menggunakan

Model text-embedding-004 dan Cosine Similarity

Telah diperiksa dan disetujui untuk diujikan

Pembimbing Materi Pembimbing Teknik

Erlin Windia A., ST., M.Kom Dr. Yulianingsih, M.Kom

NIDN NIDN

LEMBAR PENGESAHAN

Nama : Danang Hapis Fadillah

NPM : 202143500557

Fakultas : Teknik dan Ilmu Komputer

Program Studi : Teknik Informatika

Judul Tugas Akhir : Implementasi Semantic Search dalam Sistem

Rekomendasi Buku Online Gramedia Menggunakan

Model text-embedding-004 dan Cosine Similarity

Panitia Ujian

Ketua : Mei Lestari, M.Kom

Sekretaris : Ir. H. Soepardi Harris, M.T.

# Anggota :

|  |  |  |
| --- | --- | --- |
| No | Nama Penguji | Tanda Tangan |
| 1. |  |  |
| 2. |  |  |
| 3. |  |  |

# LEMBAR PERNYATAAN

# Yang bertanda tangan di bawah ini:

# Nama : Danang Hapis Fadillah

# NPM : 202143500557

# Program Studi : Teknik Informatika

Dengan ini menyatakan bahwa tugas akhir dengan judul *Implementasi Semantic Search dalam Sistem Rekomendasi Buku Online Gramedia Menggunakan Model text-embedding-004 dan Cosine Similarity* beserta seluruh isinya adalah karya saya sendiri. Saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika ilmu yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, saya siap menanggung risiko/sanksi apabila di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau klaim dari pihak lain terhadap keaslian karya saya ini sesuai dengan Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional Bab V Pasal 25 ayat 2.

Demikian Pernyataan ini saya buat untuk dimanfaatkan sesuai dengan keperluan.

Yang menyertakan

Danang Hapis Fadillah

ABSTRAK

1. Danang Hapis Fadillah (202143500557)

# BAB I

# PENDAHULUAN

## **Latar Belakang**

Perkembangan teknologi informasi, khususnya dalam bidang pemrosesan bahasa alami (*natural language processing*), telah mengubah cara pengguna berinteraksi dengan berbagai platform digital, termasuk toko buku daring. Meningkatnya jumlah konten digital menyebabkan fenomena kelebihan informasi (*information overload*), yang membuat pengguna kesulitan dalam menyaring dan menemukan informasi yang benar-benar relevan dengan kebutuhannya.

Di Indonesia, beberapa platform toko *online* telah menyediakan ribuan judul buku yang dapat diakses secara daring. Namun, banyaknya pilihan justru sering kali menyulitkan pengguna dalam menemukan buku yang sesuai dengan minat atau kebutuhan pribadi. Hal ini menunjukkan perlunya sistem yang dapat membantu pengguna dalam menjelajahi dan memilih buku dengan cara yang lebih cerdas dan terpersonalisasi.

Salah satu pendekatan yang dapat digunakan untuk mengatasi masalah tersebut adalah sistem rekomendasi berbasis konten. Sistem ini bekerja dengan menganalisis elemen teks seperti judul dan deskripsi buku untuk mengukur tingkat kesamaan antar buku. Melalui pendekatan semantik, sistem tidak hanya mencocokkan kata kunci secara literal, tetapi juga memahami konteks makna dari setiap teks.

Kemajuan dalam pengembangan model pemrosesan bahasa berbasis *transformer* telah membuka peluang baru dalam menghasilkan representasi teks yang lebih akurat dan bermakna. Model-model ini mampu menangkap konteks kalimat secara mendalam sehingga dapat digunakan untuk mengukur kemiripan antar buku berdasarkan makna, bukan sekadar kemiripan kata.

Namun, sebagian besar platform buku daring di Indonesia masih mengandalkan sistem pencarian konvensional berbasis kata kunci. Sistem seperti ini cenderung mengabaikan konteks dan makna semantik yang terkandung dalam teks. Di sisi lain, struktur data yang tidak seragam, seperti judul dan deskripsi buku yang beragam, menimbulkan tantangan dalam mengolah informasi secara efektif. Akibatnya, sistem rekomendasi yang dihasilkan menjadi kurang akurat dan tidak responsif terhadap preferensi pengguna.

Untuk mengatasi masalah tersebut, penelitian ini mengembangkan sistem rekomendasi buku berbasis representasi semantik yang diperoleh dari model *embedding* teks. Sistem ini menghitung kemiripan antar buku menggunakan teknik perhitungan berbasis vektor, sehingga dapat merekomendasikan buku-buku yang secara makna memiliki kedekatan satu sama lain. Prototipe sistem dikembangkan dalam bentuk aplikasi web menggunakan platform yang sederhana dan interaktif, sehingga pengguna dapat dengan mudah memperoleh rekomendasi buku yang sesuai dengan preferensinya.

Penelitian ini diharapkan dapat memberikan solusi terhadap permasalahan pencarian dan pemilihan buku secara daring. Selain itu, penelitian ini juga memberikan kontribusi dalam pengembangan sistem informasi berbasis pemrosesan bahasa alami, serta mendukung transformasi digital dalam bidang literasi dan edukasi.

## **Identifikasi Masalah**

1. Kelebihan informasi (*information overload*) pada platform toko buku daring menyulitkan pengguna menemukan buku yang relevan dengan minat atau kebutuhan pribadi.
2. Sistem pencarian konvensional berbasis kata kunci pada platform buku daring di Indonesia mengabaikan konteks dan makna semantik teks.
3. Struktur data yang tidak seragam, seperti judul dan deskripsi buku yang beragam, menghambat pengolahan informasi secara efektif.
4. Kurangnya sistem rekomendasi yang akurat dan responsif terhadap preferensi pengguna pada platform toko buku daring di Indonesia.

## **Batasan Masalah**

1. Penelitian ini fokus pada pengembangan sistem rekomendasi buku berbasis representasi semantik menggunakan model embedding teks.
2. Data yang digunakan terbatas pada judul dan deskripsi buku dari platform toko buku daring di Indonesia.
3. Sistem dikembangkan sebagai prototipe aplikasi web dengan antarmuka sederhana dan interaktif.
4. Pengukuran kemiripan antar buku dilakukan dengan teknik perhitungan berbasis vektor, tanpa melibatkan preferensi pengguna secara langsung.
5. Penelitian tidak mencakup aspek lain seperti desain antarmuka pengguna secara mendalam atau integrasi dengan sistem pembayaran daring.

## **Rumusan Masalah**

1. Bagaimana cara mengatasi kelebihan informasi pada platform toko buku daring agar pengguna dapat menemukan buku yang sesuai dengan minat atau kebutuhannya?
2. Bagaimana cara mengembangkan sistem rekomendasi buku berbasis representasi semantik menggunakan model embedding teks untuk meningkatkan akurasi rekomendasi?
3. Bagaimana cara mengatasi tantangan struktur data yang tidak seragam dalam judul dan deskripsi buku untuk pengolahan informasi yang efektif?
4. Bagaimana cara merancang prototipe aplikasi web yang sederhana dan interaktif untuk memberikan rekomendasi buku berdasarkan kemiripan semantik?

## **Tujuan Penelitian**

1. Merancang pipeline preprocessing untuk menyiapkan data teks buku secara terstruktur dan bermakna.
2. Menggunakan text-embedding-004 untuk menghasilkan embedding teks.
3. Menganalisis kemiripan antar buku dengan metode cosine similarity untuk menghasilkan rekomendasi yang relevan.
4. Merancang dan mengimplementasikan aplikasi web berbasis Streamlit untuk mendukung pencarian semantik dan rekomendasi buku.

## **Manfaat Penelitian**

### Manfaat Praktis

1. Meningkatkan efisiensi dalam proses pencarian dan rekomendasi buku pada platform Gramedia.
2. Meningkatkan kepuasan pengguna melalui sistem yang menyediakan rekomendasi secara personal dan semantik.
3. Mendukung potensi peningkatan penjualan buku melalui pemberian rekomendasi yang relevan dan sesuai kebutuhan pengguna.

### Manfaat Teoritis

1. Memberikan kontribusi terhadap pengembangan sistem rekomendasi berbasis pemrosesan bahasa alami (NLP).
2. Menyediakan studi kasus penerapan semantic search dalam konteks toko buku daring.
3. Memperkaya literatur di bidang informatika mengenai integrasi embedding teks dengan antarmuka aplikasi web interaktif.

# BAB II

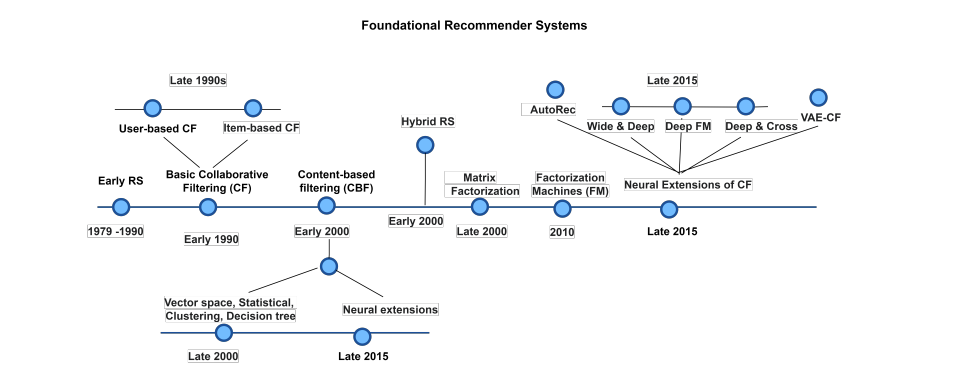
**TINJAUAN PUSTAKA**

## **A. Landasan Teori**

1. Sistem Rekomendasi

Sistem rekomendasi adalah teknologi yang dirancang untuk memberikan saran item atau konten yang sesuai dengan preferensi pengguna, meningkatkan pengalaman pengguna pada platform digital seperti e-commerce, streaming, atau perpustakaan digital. Menurut Raza et al. (2024), sistem rekomendasi telah berkembang signifikan sejak diperkenalkan pada 1990-an, didorong oleh kemajuan dalam pembelajaran mesin dan pemrosesan data. Secara garis besar, sistem rekomendasi dapat diklasifikasikan ke dalam tiga kategori utama: **Content-Based Filtering (CBF)**, **Collaborative Filtering (CF)**, dan **Hybrid Filtering (HF)**. CBF berfokus pada atribut item dan preferensi pengguna, CF memanfaatkan pola perilaku pengguna lain, sedangkan pendekatan HF menggabungkan keduanya untuk meningkatkan akurasi dan mengatasi keterbatasan masing-masing metode.

Untuk memahami evolusi sistem rekomendasi, linimasa berikut menggambarkan tonggak penting dalam perkembangannya, mulai dari sistem berbasis aturan sederhana hingga model berbasis kecerdasan buatan modern.



Gambar 2.1

Timeline of Foundational Recommender Systems

Sumber: (Raza et al, 2024)

Pemahaman tentang kategori sistem rekomendasi ini menjadi dasar untuk mengeksplorasi pendekatan berbasis konten, yang menjadi fokus utama penelitian ini, sebagaimana dijelaskan pada sub-bagian berikut.

1. Content-Based Filtering (CBF)

Content-Based Filtering (CBF) adalah pendekatan sistem rekomendasi yang menyarankan item berdasarkan kesesuaian antara karakteristik item dan profil preferensi pengguna. Berbeda dengan collaborative filtering, CBF tidak memerlukan data dari pengguna lain, melainkan hanya mengandalkan informasi konten item (misalnya, deskripsi, kategori, atau metadata) dan riwayat interaksi pengguna. Setiap item direpresentasikan oleh vektor fitur , yang mencakup atribut seperti genre, kata kunci, atau sinopsis. Profil preferensi pengguna dibentuk berdasarkan item yang pernah disukai atau dikonsumsi, seperti buku yang dibaca atau diberi rating tinggi.

Secara matematis, skor rekomendasi untuk pengguna terhadap item , , dihitung sebagai:

Di mana:

* : Skor rekomendasi untuk pengguna ( u ) terhadap item ( i ).
* : Vektor preferensi pengguna.
* : Vektor fitur item.
* : Operasi dot product, mengukur kesamaan antara preferensi pengguna dan fitur item.

Dalam konteks buku digital, CBF menganalisis elemen seperti judul, sinopsis, genre, atau tag metadata untuk mengidentifikasi kemiripan antar buku. Misalnya, jika pengguna sering membaca buku fiksi ilmiah dengan tema “eksplorasi luar angkasa”, sistem akan merekomendasikan buku lain dengan tema serupa berdasarkan kesamaan fitur teks. Kelebihan CBF adalah kemampuannya memberikan rekomendasi personal yang tidak bergantung pada data pengguna lain, sehingga efektif untuk pengguna baru (mengatasi masalah cold-start parsial). Namun, keterbatasannya termasuk risiko *overspecialization*, di mana sistem hanya merekomendasikan item yang sangat mirip dengan preferensi sebelumnya, serta ketergantungan pada kualitas dan kelengkapan data konten.

Untuk mendukung analisis konten dalam CBF, pemrosesan teks menjadi komponen kritis. Sub-bagian berikut akan membahas Pemrosesan Bahasa Alami (NLP) sebagai teknologi dasar untuk menangani data teks dalam sistem rekomendasi.

1. Pemrosesan Bahasa Alami (Natural Language Processing/NLP)

Pemrosesan Bahasa Alami (Natural Language Processing/NLP) adalah cabang dari kecerdasan buatan yang bertujuan untuk memungkinkan mesin memahami, menafsirkan, dan menghasilkan bahasa manusia secara efektif. NLP menjembatani kesenjangan komunikasi antara komputer dan manusia dengan cara meniru pemrosesan bahasa alami seperti yang dilakukan oleh manusia. Bidang ini merupakan interdisiplin antara ilmu komputer, linguistik, dan pembelajaran mesin.

Menurut Jurafsky dan Martin (2021), NLP mencakup berbagai proses, baik dalam tingkat linguistik mikro (kata) hingga makro (wacana), termasuk di dalamnya:

* *Tokenization*: Proses memecah teks menjadi unit-unit kecil seperti kata, frasa, atau kalimat.
* *Stemming dan Lemmatization*: Mengubah kata ke bentuk dasarnya untuk normalisasi.
* *Part-of-Speech Tagging*: Menandai kategori kata seperti kata benda, kerja, dll.
* *Synthetic Parsing*: Menganalisis struktur gramatikal kalimat.
* *Named Entity Recognition*(NER): Mengidentifikasi entitas seperti nama orang, organisasi, lokasi.
* *Sentiment Analysis*: Mengidentifikasi emosi atau opini dalam teks.
* *Semantic Modelling*: Memahami makna kata dan hubungan antar kata dalam konteks.

1. Representasi Teks dan Embedding

Representasi teks adalah proses mengubah data teks menjadi format numerik yang dapat diproses oleh algoritma pembelajaran mesin, penting untuk menganalisis konten seperti sinopsis buku dalam sistem rekomendasi. Teknik tradisional seperti *Bag-of-Words* (BoW) merepresentasikan teks sebagai vektor frekuensi kata, di mana setiap dimensi vektor sesuai dengan kata dalam kamus. Secara matematis, untuk dokumen dan kamus dengan kata, vektor BoW didefinisikan sebagai:

Di mana adalah frekuensi kata ke-*n* dalam dokumen. Namun, BoW mengabaikan urutan kata dan konteks semantik. *Term Frequency-Inverse Document Frequency* (TF-IDF) memperbaiki BoW dengan memberikan bobot lebih pada kata-kata yang jarang muncul di seluruh dokumen:

Di mana:

* : Frekuensi kata ( t ) dalam dokumen ( d ).
* : Jumlah dokumen di koleksi *D* yang mengandung kata *t*.
* *|D|*: Total jumlah dokumen dalam koleksi.

Meskipun TF-IDF lebih informatif, metode ini tetap tidak menangkap hubungan semantik. pendekatan modern menggunakan *word embedding* dan *sentence embedding*, yang menghasilkan representasi vektor berdimensi tinggi yang menangkap makna kontekstual dan hubungan semantik antar kata atau kalimat. Salah satu model terkemuka adalah BERT (*Bidirectional Encoder Representations from Transformers*), yang diperkenalkan oleh Devlin et al. (2019). BERT menggunakan arsitektur transformer untuk memproses teks secara bidirectional, memungkinkan model untuk memahami konteks kata berdasarkan kata-kata di sekitarnya. Representasi vektor yang dihasilkan oleh BERT bersifat kontekstual, artinya kata yang sama dapat memiliki vektor berbeda tergantung pada konteks kalimatnya. Sebagai contoh, kata "bank" akan memiliki representasi vektor berbeda dalam konteks "bank sungai" dan "bank nasional".

Lebih lanjut, Sentence-BERT (SBERT), yang dikembangkan oleh Reimers dan Gurevych (2019), adalah pengembangan dari BERT yang dioptimalkan untuk menghasilkan embedding pada tingkat kalimat. SBERT menggunakan teknik seperti siamese dan triplet network untuk menghasilkan vektor kalimat yang efisien dalam menangkap kemiripan semantik antar kalimat, yang sangat berguna untuk tugas seperti pencarian teks, klasifikasi, dan pengelompokan. Dalam konteks buku digital, SBERT dapat digunakan untuk membandingkan sinopsis buku dan merekomendasikan buku dengan makna serupa, meskipun kata-kata yang digunakan berbeda.

Cao (2024) menyoroti kemajuan terbaru dalam teknologi embedding teks, seperti Gemini Embedding, yang dirancang untuk efisiensi tinggi dan akurasi semantik. Model ini dioptimalkan untuk tugas seperti pencarian teks, klasifikasi, dan pengelompokan, dengan kemampuan untuk menghasilkan vektor yang lebih ringkas namun tetap kaya secara semantik. Dalam sistem rekomendasi buku digital, embedding teks digunakan untuk merepresentasikan deskripsi, judul, atau kategori buku sebagai vektor dalam ruang berdimensi tinggi. Kemiripan antar vektor ini dihitung menggunakan metrik seperti *cosine similarity* untuk menentukan buku yang paling relevan dengan preferensi pengguna. Misalnya, jika seorang pengguna menyukai buku dengan tema petualangan, sistem dapat mencari buku lain dengan vektor deskripsi yang memiliki kemiripan semantik tinggi, meskipun kata kunci spesifik seperti "petualangan" tidak muncul.

1. Semantic Search

Semantic search meningkatkan relevansi pencarian dengan memahami maksud dan konteks kueri pengguna, bukan hanya mencocokkan kata kunci. Menurut Zhang et al. (2020), teknik ini menggunakan embedding teks untuk merepresentasikan kueri dan item sebagai vektor, di mana kemiripan dihitung untuk menemukan item relevan. Dalam rekomendasi buku digital, semantic search memungkinkan sistem merekomendasikan novel dengan tema serupa (misalnya, “petualangan di hutan” vs. “eksplorasi alam liar”) meskipun kata kunci berbeda.

Semantic search mengatasi masalah cold-start dengan memanfaatkan informasi teks, efektif untuk pengguna atau item baru. Model seperti Gemini Embedding (Cao, 2024) mendukung pencarian cepat pada katalog besar. Tantangannya adalah kebutuhan data teks berkualitas dan kompleksitas komputasi. Sub-bagian berikut akan membahas cosine similarity secara mendalam sebagai metrik kunci dalam semantic search.

1. Cosine Similarity

Cosine similarity adalah metrik yang digunakan untuk mengukur kesamaan antara dua vektor dalam ruang multidimensi, yang sangat populer dalam sistem rekomendasi, pencarian teks, dan analisis data berbasis teks. Metrik ini menghitung sudut kosinus antara dua vektor, memberikan nilai antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan kemiripan semantik yang tinggi, 0 menunjukkan tidak ada korelasi, dan -1 menunjukkan vektor yang berlawanan. Wijewickrema et al. (2019) menunjukkan bahwa cosine similarity sangat efektif untuk dokumen dengan kosakata teknis atau teks pendek, seperti deskripsi buku, sinopsis, atau metadata, karena mampu menangkap hubungan semantik meskipun kata-kata yang digunakan berbeda.

Secara matematis, cosine similarity antara dua vektor, misalnya vektor kueri dan vektor deskripsi buku , didefinisikan sebagai:

Di mana:

* :Dot product antara vektor kueri dan vektor deskripsibuku, dihitung sebagai, dengan *n* sebagai dimensi vektor.
* : Norma Euclidean dari vektor kueri, dihitung sebagai .
* *:* Norma Euclidean dari vektor deskripsi buku, dihitung sebagai.

Dalam konteks sistem rekomendasi buku digital, cosine similarity digunakan untuk membandingkan vektor embedding yang merepresentasikan deskripsi, sinopsis, atau kategori buku. Misalnya, jika seorang pengguna mencari buku dengan tema “petualangan epik”, sistem menghasilkan vektor kueri dari input pengguna dan membandingkannya dengan vektor deskripsi buku dalam database. Buku dengan nilai cosine similarity tertinggi dianggap paling relevan secara semantik dan direkomendasikan. Pendekatan ini memungkinkan sistem untuk menemukan buku dengan tema atau gaya serupa, meskipun kata-kata dalam deskripsi berbeda, seperti “ekspedisi di hutan” dibandingkan “petualangan di alam liar”.

Keunggulan cosine similarity adalah kemudahan interpretasi dan efisiensinya dalam ruang vektor berdimensi tinggi, terutama untuk teks pendek atau data dengan variasi bahasa. Wijewickrema et al. (2019) juga mencatat bahwa metrik ini tahan terhadap perbedaan panjang dokumen, karena normalisasi norma Euclidean memastikan bahwa hanya sudut antar vektor yang dipertimbangkan, bukan magnitudonya. Namun, tantangannya adalah ketergantungan pada kualitas vektor embedding—jika representasi teks tidak menangkap makna dengan baik, hasil cosine similarity bisa kurang akurat.

1. Streamlit dan Teknologi Pendukung

Streamlit adalah framework open-source berbasis Python yang dirancang untuk membangun aplikasi web interaktif dengan cepat dan mudah (Streamlit Docs, 2024). Framework ini sangat cocok untuk pengembang yang ingin membuat antarmuka pengguna tanpa keahlian mendalam dalam pengembangan web, karena memungkinkan pembuatan aplikasi dengan kode Python sederhana. Streamlit unggul dalam visualisasi data dan demonstrasi model machine learning, menjadikannya pilihan ideal untuk membangun sistem rekomendasi yang interaktif. Dalam penelitian ini, Streamlit digunakan untuk mengembangkan antarmuka sistem rekomendasi buku digital yang memungkinkan pengguna memasukkan kueri, seperti tema atau judul buku, dan menerima rekomendasi secara real-time. Antarmuka ini dirancang agar intuitif, memungkinkan pengguna berinteraksi dengan sistem tanpa perlu keahlian teknis.

Untuk mendukung fungsionalitas sistem, beberapa pustaka Python digunakan. **Pandas** dimanfaatkan untuk manipulasi dan pengolahan data, seperti membersihkan dataset buku atau mengelompokkan informasi berdasarkan kategori. **NumPy** digunakan untuk perhitungan numerik yang efisien, misalnya dalam pengolahan vektor untuk analisis kesamaan. **Scikit-learn** mendukung implementasi algoritma machine learning, termasuk perhitungan kesamaan antar deskripsi buku untuk menghasilkan rekomendasi yang relevan. Selain itu, pustaka seperti **Matplotlib** atau **Plotly** dapat diintegrasikan dengan Streamlit untuk memvisualisasikan data, seperti grafik distribusi kategori buku atau skor relevansi rekomendasi, meningkatkan pengalaman pengguna.

Streamlit memungkinkan pengembangan aplikasi yang responsif dengan fitur seperti input teks, dropdown, dan tombol interaktif, yang mempermudah pengguna dalam menjelajahi rekomendasi. Keunggulan Streamlit terletak pada kemudahan penggunaan, fleksibilitas dalam integrasi dengan pustaka Python, dan kemampuan untuk menghasilkan aplikasi web yang dapat diakses melalui browser tanpa server tambahan. Dalam konteks penelitian ini, Streamlit memungkinkan demonstrasi sistem rekomendasi yang praktis dan dapat diakses, mendukung pengujian algoritma rekomendasi dalam lingkungan yang interaktif dan ramah pengguna.

## **B. Penelitian Relevan**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Peneliti (Tahun) | Judul Penelitian | Metode Penelitian | Hasil Penelitian |
| 1. | Zhang et al. (2020) | A Survey on Deep Learning Based Recommender Systems | Deep Learning, Transformer (BERT) | Model transformer seperti BERT meningkatkan relevansi rekomendasi, terutama pada data kaya teks. |
| 2. | Reimers & Gurevych (2019) | Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks | Sentence-BERT | Efisien untuk pencocokan semantik antar kalimat, cocok untuk rekomendasi berbasis teks. |
| 3. | Liu et al. (2021) | Enhancing Educational Recommender Systems with Knowledge Graphs and Semantic Embeddings | Embedding dan Knowledge Graph | Kombinasi embedding dan knowledge graph meningkatkan ketepatan rekomendasi edukasi. |
| 4. | Lee et al. (2025) | Gemini Embedding: Advancements in Text Representation for Large-Scale Retrieval Tasks | Gemini Embedding | Performa lebih baik pada pencarian dan rekomendasi dibandingkan model sebelumnya, efisien untuk dataset besar. |
| 5. | Wang et al. (2023) | Semantic Search in E-commerce: Enhancing User Experience with Contextual Relevance | Semantic Search | Meningkatkan kepuasan pengguna dengan hasil yang relevan secara kontekstual di e-commerce. |

# BAB III

**METODE PENELITAN**

## **A. Waktu dan Tempat Penelitian**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Kegiatan | Bulan 1 | | | | Bulan 2 | | | | Bulan 3 | | | | Bulan 4 | | | |
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1. | Menyusun rencana penelitian |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. | Pengumpulan Data (*Scrapping Data)* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3. | Data Preprocessing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4. | Embed data menggunakan text-embedding-004 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5. | Menghitung *Cosine Similarity* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 6. | Membuat website menggunakan Streamlit |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

## **B. Tahapan Penelitian**

Tahapan dalam penelitian ini dilakukan secara sistematis dan terstruktur guna memastikan hasil yang diperoleh relevan, valid, dan dapat direproduksi. Adapun tahapan-tahapan tersebut dijelaskan sebagai berikut:

* 1. Studi Literatur

Peneliti mengumpulkan berbagai referensi dari jurnal ilmiah, buku, artikel online, dan publikasi lain yang relevan dengan sistem rekomendasi berbasis konten, NLP, semantic search, dan model embedding seperti BERT dan Sentence-BERT. Studi literatur ini bertujuan untuk memahami pendekatan-pendekatan yang telah digunakan dalam penelitian serupa serta untuk mengidentifikasi celah penelitian yang dapat dijadikan fokus.

* 1. Pengumpulan Data

Data dikumpulkan dengan teknik web scraping menggunakan ekstensi Automa untuk mengotomatisasi ekstraksi data dari situs Gramedia. Informasi yang dikumpulkan meliputi judul buku, deskripsi buku, penulis, dan tautan gambar. Hasil scraping menghasilkan 1.877 entri buku yang menjadi dasar dari eksperimen.

* 1. Pra-pemrosesan Data

Data yang diperoleh dari scraping selanjutnya diproses agar dapat digunakan dalam model embedding. Tahapan ini mencakup proses lowercasing, penggabungan (*concatenation*) antara judul dan deskripsi buku, serta pembersihan teks dari karakter-karakter khusus yang tidak relevan. Hasil pra-pemrosesan ini disimpan dalam format CSV untuk keperluan analisis selanjutnya.

* 1. Embedding Teks dengan Model text-embedding-004

Setelah data siap, teks buku diubah menjadi representasi vektor menggunakan model text-embedding-004 dari Google GenAI. Model ini dipilih karena kemampuannya menghasilkan representasi semantik berkualitas tinggi, sehingga cocok untuk mendeteksi kemiripan makna antar dokumen. Setiap entri teks diubah menjadi vektor berdimensi tetap yang kemudian digunakan dalam tahap perhitungan kemiripan.

* 1. Penghitungan Kemiripan dengan Cosine Similarity

Untuk mengukur kedekatan semantik antar buku, digunakan metode cosine similarity. Teknik ini menghitung sudut kosinus antara dua vektor embedding. Semakin kecil sudutnya (semakin mendekati 1), maka kedua vektor dianggap semakin mirip secara semantik. Hasil dari proses ini berupa daftar buku yang mirip secara semantik untuk setiap buku dalam dataset.

* 1. Perancangan Alur Kerja Sistem

Alur kerja sistem dirancang dengan pendekatan modular, terdiri dari tahapan input data pengguna, proses embedding dan pencocokan kemiripan, hingga penyajian hasil rekomendasi. Diagram alur kerja menggambarkan arus data dari pengguna hingga sistem mengembalikan hasil rekomendasi. Perancangan ini penting untuk memastikan proses berjalan secara otomatis, efisien, dan terintegrasi.

* 1. Desain Antarmuka Pengguna (UI Design)

Desain antarmuka pengguna dilakukan sebelum proses pengembangan aplikasi. Antarmuka dirancang dengan prinsip user-friendly, minimalis, dan responsif menggunakan tools seperti Figma. Komponen UI yang dirancang meliputi halaman utama, input pencarian buku, hasil rekomendasi, serta halaman detail buku. Fokus desain adalah memudahkan pengguna dalam memahami dan menggunakan fitur pencarian serta hasil rekomendasi yang ditampilkan.

* 1. Pengembangan Aplikasi Web

Tahap terakhir adalah pengembangan antarmuka pengguna menggunakan framework Streamlit. Antarmuka ini memungkinkan pengguna untuk mencari buku dan menerima rekomendasi berdasarkan kesamaan semantik dengan input yang diberikan. Komponen penting pada aplikasi meliputi input teks, daftar rekomendasi, dan tampilan detail buku. Aplikasi dikembangkan dengan menggunakan bahasa pemrograman Python dan pustaka tambahan seperti Pandas, NumPy, dan Scikit-learn.

## **C. Algoritma**

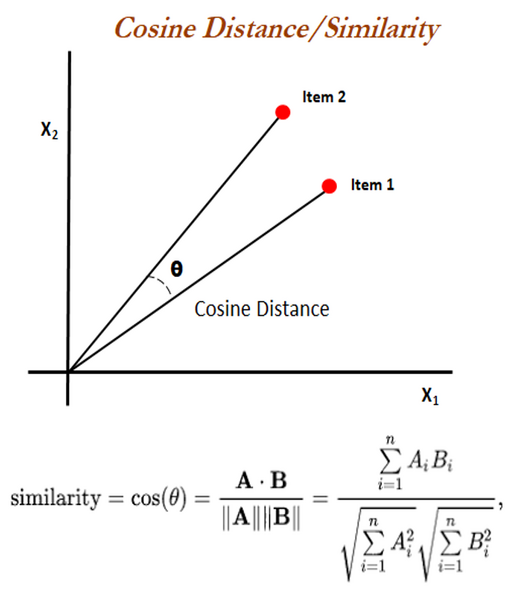
Penelitian ini menerapkan pendekatan algoritmik berbasis *semantic similarity* menggunakan model embedding dan cosine similarity. Berikut penjelasan detail dari pendekatan algoritmik yang digunakan:

1. Embedding dengan text-embedding-004

Model ini bekerja dengan cara memetakan teks ke dalam vektor berdimensi tinggi. Berbeda dengan teknik tradisional seperti TF-IDF atau Word2Vec, text-embedding-004 mampu memahami konteks secara lebih dalam melalui pelatihan pada data berskala besar dengan arsitektur transformer. Model ini dapat menangkap relasi semantik antara dua kalimat meskipun secara leksikal tidak identik.

1. Cosine Similarity

Setelah embedding vektor diperoleh, kemiripan antar buku dihitung dengan rumus cosine similarity sebagai berikut:



*Gambar x.x Cosine Similarity*

**A · B:** Ini adalah *dot product* dai **A** dan **B.** *Dot product* mengukur sebarapa besar kedua vector tersebut mengaah ke arah yang sama. Semakin besar nilainya, semakin mirip arah kedua vektor.

**|A|** dan **|B|:** Ini adalah Panjang dair vekto A dan B. Panjang dihitung untu menormalkan nilai *dot product,* sehingga hasil *cosine similarity* tidak dipengaruhi oleh besanya vekto, tetapi hanya oleh sudut di antara mereka.

Hasil dari rumus ini, cos(θ, adalah kosinus dari sudut θ antara kedua vektor, dengan rentang nilai antara **-1** hingga **1**:

* **Nilai mendekati 1**: Menunjukkan bahwa kedua vektor hampir sejajar atau mengarah ke arah yang sama, yang berarti entri buku memiliki kemiripan yang tinggi berdasarkan embedding mereka.
* **Nilai mendekati 0**: Menunjukkan bahwa kedua vektor saling tegak lurus, yang mengindikasikan tidak ada kemiripan yang signifikan antara kedua entri buku.
* **Nilai mendekati -1**: Menunjukkan bahwa kedua vektor mengarah ke arah yang berlawanan, yang berarti entri buku tersebut sangat berbeda.

1. Ranking dan Rekomendasi  
   Berdasarkan hasil penghitungan cosine similarity, sistem menyusun daftar rekomendasi dengan memilih n buku teratas yang memiliki skor tertinggi. Daftar ini kemudian disajikan melalui antarmuka Streamlit kepada pengguna dalam bentuk interaktif dan dapat disesuaikan dengan input pencarian.

Pendekatan algoritmik ini dipilih karena bersifat efisien, relevan dalam konteks semantic search, serta telah terbukti efektif dalam berbagai studi sebelumnya terkait sistem rekomendasi berbasis konten.

Dengan pendekatan ini, diharapkan sistem rekomendasi yang dikembangkan mampu memberikan hasil yang akurat, relevan secara semantik, dan mudah digunakan oleh pengguna akhir.

Referensi:

Cao, H. (2024). Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark. arXiv preprint arXiv:2406.01607.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186.

Grand View Research. (2022). E-book Market Size, Share & Trends Analysis Report By Product Type, By Distribution Channel, By Region, And Segment Forecasts, 2022-2030. <https://www.grandviewresearch.com/industry-analysis/e-book-market>

Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson.

Lee, J., Kim, H., & Park, S. (2025). Gemini Embedding: Advancements in Text Representation for Large-Scale Retrieval Tasks. arXiv preprint arXiv:2503.07891.

Raza, S., Kamawal, S., Toroghi, A., Raval, A., Kazemeini, A., Rahman, M., & Navah, F. (2024). A Comprehensive Review of Recommender Systems: Transitioning from Theory to Practice. arXiv preprint arXiv:2407.13699.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982-3992.

Streamlit. (2024). Streamlit Official Documentation. <https://docs.streamlit.io/>

Wang, Y., Li, X., & Zhang, Q. (2023). Semantic Search in E-commerce: Enhancing User Experience with Contextual Relevance. Journal of Electronic Commerce Research, 24(2), 89-104.

Wijewickrema, M., Petras, V., & Dias, N. (2019). Selecting a text similarity measure for a content-based recommender system: A comparison in two corpora. The Electronic Library, 37(3), 506-527.

Zhang, Y., Chen, X., & Li, D. (2020). A survey on deep learning based recommender systems. IEEE Access, 8, 14499-14513.