

**PERBANDINGAN EFEKTIVITAS MODEL TF-IDF DAN
WORD2VEC DALAM SISTEM REKOMENDASI BUKU
BERBASIS CONTENT-BASED FILTERING**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Sistem Informasi



disusun oleh

MUHAMMAD HADZIQ FAUZAN

22.12.2655

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

**PERBANDINGAN EFEKTIVITAS MODEL TF-IDF DAN
WORD2VEC DALAM SISTEM REKOMENDASI BUKU
BERBASIS CONTENT-BASED FILTERING**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Sistem Informasi



disusun oleh

MUHAMMAD HADZIQ FAUZAN

22.12.2655

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**PERBANDINGAN EFEKTIVITAS MODEL TF-IDF DAN WORD2VEC
DALAM SISTEM REKOMENDASI BUKU BERBASIS CONTENT-BASED
FILTERING**

yang disusun dan diajukan oleh

Muhammad Hadziq Fauzan

22.12.2655

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 21 April 2026

Dosen Pembimbing,



Donni Prabowo, S.Kom., M.Kom

NIK. 190302253

HALAMAN PENGESAHAN
SKRIPSI
PERBANDINGAN EFEKTIVITAS MODEL TF-IDF DAN WORD2VEC
DALAM SISTEM REKOMENDASI BUKU BERBASIS CONTENT-BASED
FILTERING

yang disusun dan diajukan oleh

Muhammad Hadziq Fauzan

22.12.2655

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 April 2026

Susunan Dewan Penguji

Nama Penguji

Krisnawati, S.Si., M.T.
NIK. 190302038

Irma Rofni Wulandari, S.Pd., M.Eng.
NIK. 190302329

Donni Prabowo, S.Kom., M.Kom.
NIK. 190302253

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 April 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muhammad Hadziq Fauzan
NIM : 22.12.2655

Menyatakan bahwa Skripsi dengan judul berikut:

PERBANDINGAN EFEKTIVITAS MODEL TF-IDF DAN WORD2VEC DALAM SISTEM REKOMENDASI BUKU BERBASIS CONTENT-BASED FILTERING

Dosen Pembimbing : Donni Prabowo, S.Kom., M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 21 April 2026

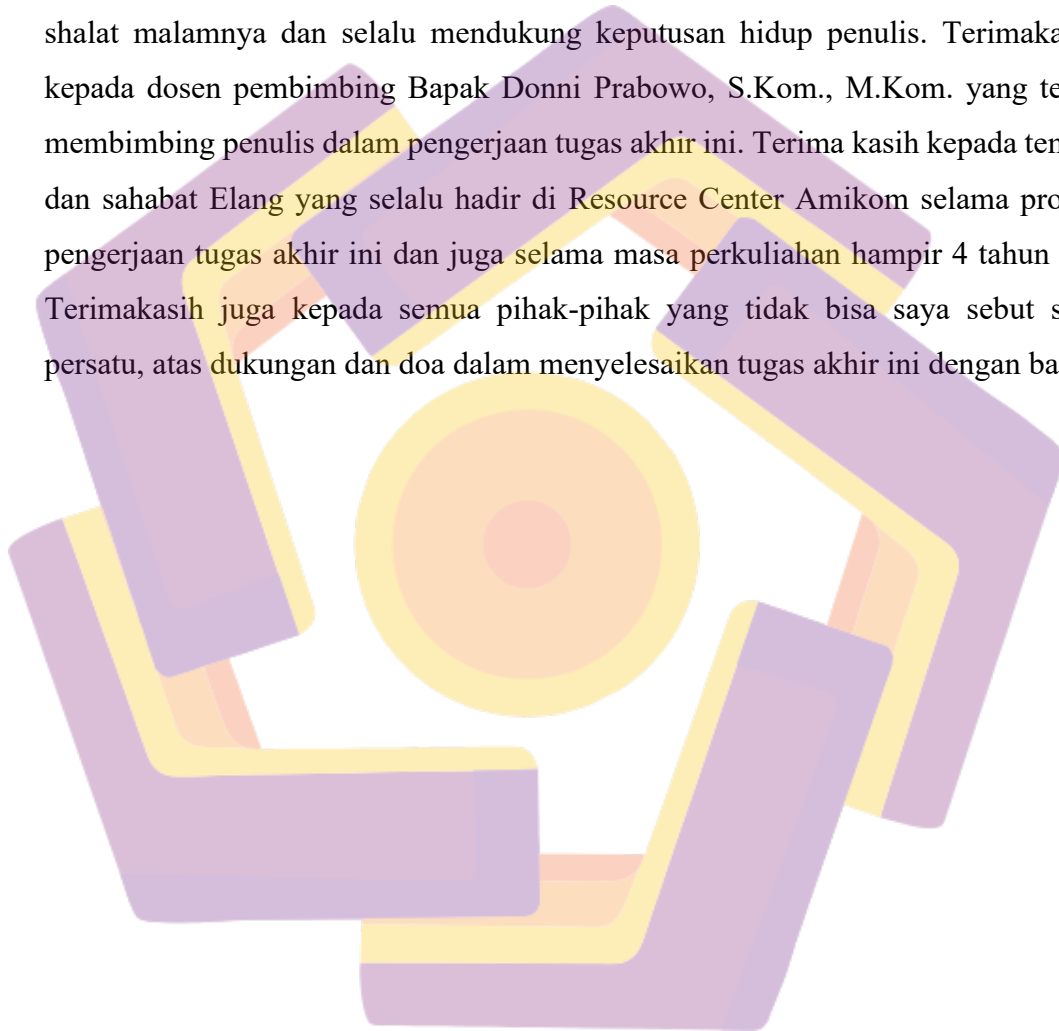
Yang Menyatakan,



Muhammad Hadziq Fauzan

HALAMAN PERSEMBAHAN

Puji syukur atas kehadiran Allah SWT atas segala rahmat dan karunia-Nya sehingga saya dapat menyelesaikan tugas akhir berikut dengan baik. Tugas akhir ini dipersembahkan kepada semua yang mencintai dan menyayangi penulis. Terimakasih kepada kedua orang tua yang selalu menyebut nama penulis dalam shalat malamnya dan selalu mendukung keputusan hidup penulis. Terimakasih kepada dosen pembimbing Bapak Donni Prabowo, S.Kom., M.Kom. yang telah membimbing penulis dalam pengerjaan tugas akhir ini. Terima kasih kepada teman dan sahabat Elang yang selalu hadir di Resource Center Amikom selama proses pengerjaan tugas akhir ini dan juga selama masa perkuliahan hampir 4 tahun ini. Terimakasih juga kepada semua pihak-pihak yang tidak bisa saya sebut satu persatu, atas dukungan dan doa dalam menyelesaikan tugas akhir ini dengan baik.



KATA PENGANTAR

Puji syukur atas kehadiran Allah SWT atas segala rahmat dan karunia-Nya sehingga penulisan tugas akhir skripsi dengan judul **“PERBANDINGAN EFEKTIVITAS MODEL TF-IDF DAN WORD2VEC DALAM SISTEM REKOMENDASI BUKU BERBASIS CONTENT-BASED FILTERING”** dapat diselesaikan dengan lancar. Tugas akhir ini disusun untuk memenuhi salah satu syarat kelulusan Program Sarjana Fakultas Ilmu Komputer Universitas Amikom Yogyakarta pada Program Studi Sistem Informasi.

Penulis juga menyampaikan rasa terima kasih kepada semua pihak yang telah membantu dan berkontribusi serta memberikan dukungan moral untuk menyelesaikan tugas akhir skripsi ini. Penulis mengucapkan banyak terimakasih kepada :

1. Bapak Prof. Dr. M. Suyanto, MM. selaku Rektor Universitas Amikom Yogyakarta
2. Prof. Dr. Kusri, M.Kom. selaku Dekan Fakultas Ilmu Komputer
3. Bapak Anggit Dwi Hartanto, M.Kom selaku Ketua Program Studi Sistem Informasi
4. Bapak Donni Prabowo, S.Kom., M.Kom. selaku Dosen Pembimbing yang telah berkenan meluangkan waktu, arahan, dan saran penulisan pada penyusunan skripsi ini.
5. Kedua orang tua penulis, Bapak dan Ibu yang telah memberikan dukungan, memberikan arahan dan nasihat dan selalu memotivasi dalam keadaan apapun.
6. Kepada saudara kandung penulis, yang dimanapun berada selalu meluangkan waktu memberikan motivasi, semangat, saran, dan doa serta memberikan uang saku tambahan kepada penulis selama perkuliahan.

Yogyakarta, 12 April 2026

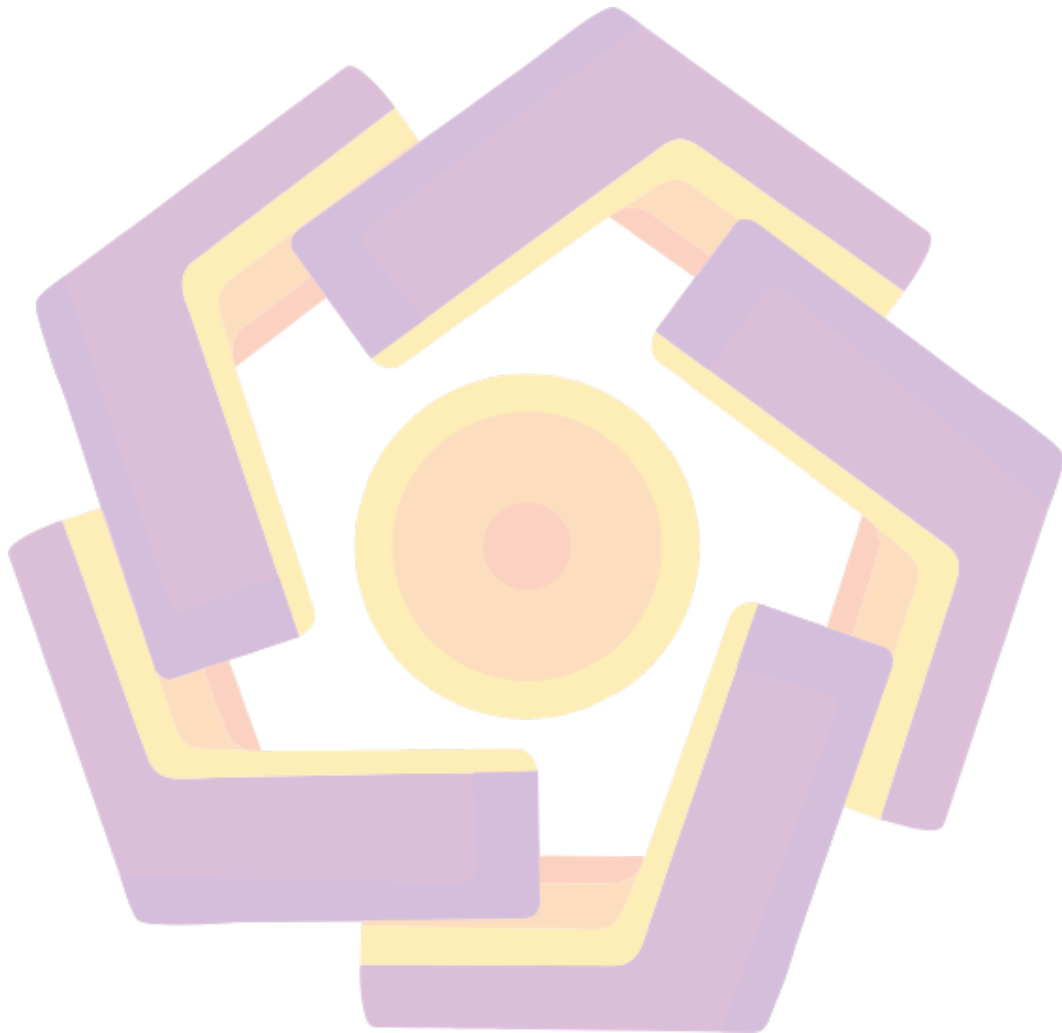
Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	x
DAFTAR GAMBAR.....	xi
DAFTAR LAMBANG DAN SINGKATAN	xiii
DAFTAR ISTILAH	xiv
INTISARI	xvi
<i>ABSTRACT</i>	xvii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	4
1.6 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	5
2.1 Studi Literatur	5
2.2 Dasar Teori.....	15

2.2.1.	Buku	15
2.2.2.	Bahasa Pemrograman Python	16
2.2.3.	Sistem Rekomendasi	16
2.2.4.	Content-Based Filtering	18
2.2.5.	Cosine Similarity	18
2.2.6.	Algoritma TF-IDF	19
2.2.7.	Algoritma Word2Vec	20
2.2.8.	Preprocessing	21
BAB III METODE PENELITIAN		23
3.1	Objek Penelitian	23
3.2	Alur Penelitian	23
3.3	Alat dan Bahan	27
BAB IV HASIL DAN PEMBAHASAN		29
4.1	Pengumpulan dan Persiapan Data	29
4.1.1.	Import Library	29
4.1.2.	Pengumpulan Data	30
4.1.3.	Persiapan Data	32
4.1.4.	Preprocessing Data	34
4.2	Implementasi	39
4.2.1.	Dataset Kaggle (Goodreads)	39
4.2.2.	Dataset Kemdikbudristek	54
4.3	Pengujian dan Evaluasi	61
4.3.1.	Pengujian Efektivitas	61
4.3.2.	Analisis Hasil Pengujian	64
BAB V PENUTUP		68

5.1	Kesimpulan	68
5.2	Saran	69
REFERENSI		70



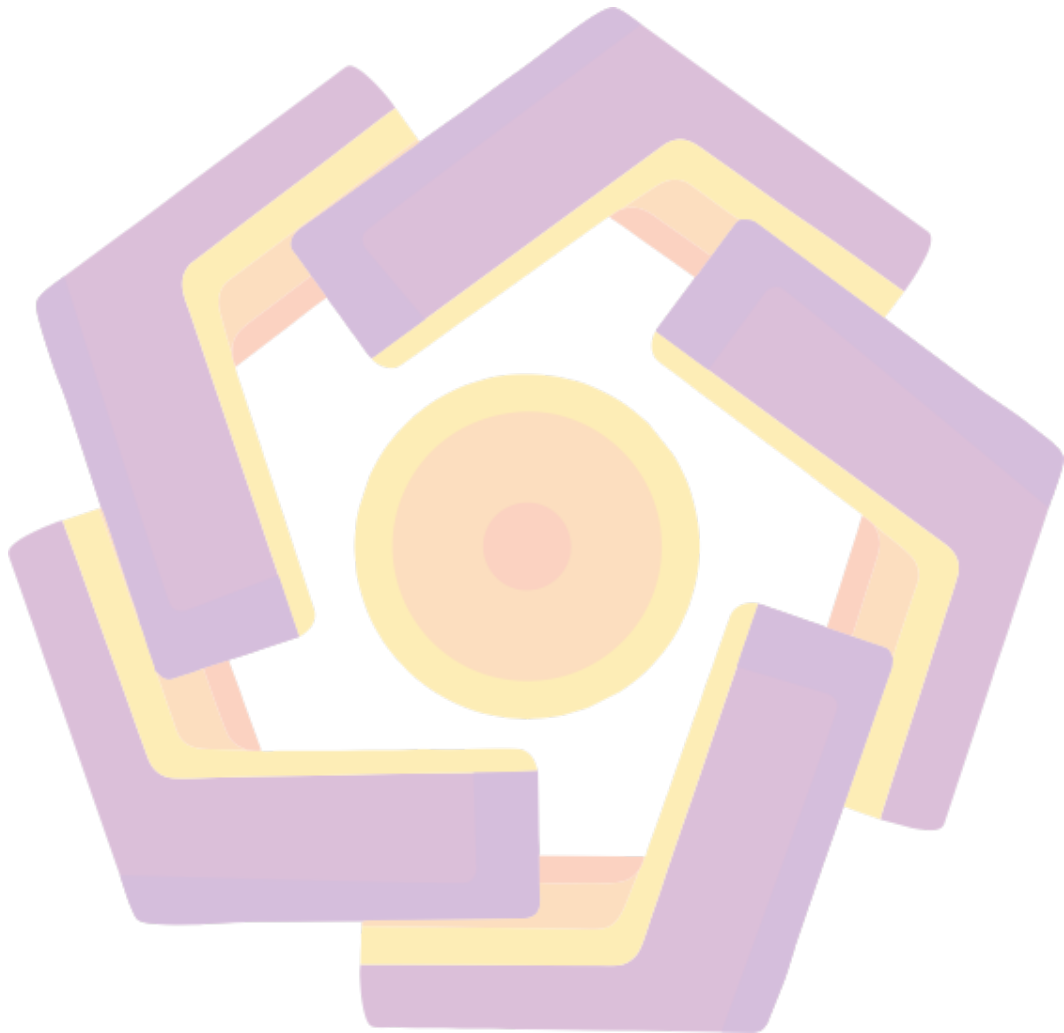
DAFTAR TABEL

Tabel 2. 1 Keaslian Penelitian	8
Tabel 4. 1 Source Code Import Library	29
Tabel 4. 2 Perintah Mounting Google Drive	31
Tabel 4. 3 Tabel Atribut pada Dataset	33
Tabel 4. 4 Script Preprocessing Untuk Algoritma Word2Vec	35
Tabel 4. 5 Script Tambahan Untuk Preprocessing Algoritma TF-IDF	37
Tabel 4. 6. Kode Script Implementasi TF-IDF Pada Dataset Kaggle.....	40
Tabel 4. 7. Tabel Data Sampel Buku	42
Tabel 4. 8. Tabel Pengujian Perhitungan TF-IDF Dataset Kaggle (Menampilkan 20 Token Teratas)	44
Tabel 4. 9, Kode Script Implementasi Cosine Similarity Pada TF-IDF Untuk Dataset Kaggle	46
Tabel 4. 10 Kode Script Melatih Model Algoritma Word2Vec Dengan Arsitektur CBOW.....	50
Tabel 4. 11. Kode Script Implementasi Word2Vec Pada Dataset Kaggle.....	51
Tabel 4. 12. Kode Script Implementasi Cosine Similarity Pada TF-IDF Untuk Dataset Kaggle	52
Tabel 4. 13. Tabel Perhitungan TF-IDF Dataset Kemdikbudristek (Menampilkan 20 Token Teratas)	55
Tabel 4. 14. Kode Script Implementasi Cosine Similarity Pada Dataset Kemdikbudristek.....	56
Tabel 4. 15. Kode Script Training Dataset Dengan Arsitektur Skip-Gram	60
Tabel 4. 16 Evaluasi Skenario Dataset Kaggle Dengan Algoritma TF-IDF.....	62
Tabel 4. 17 Evaluasi Skenario Dataset Kaggle Dengan Algoritma Word2Vec ...	62
Tabel 4. 18 Evaluasi Skenario Dataset Kemdikbudristek Dengan Algoritma TF-IDF	63
Tabel 4. 19 Evaluasi Skenario Dataset Kemdikbudristek Dengan Algoritma Word2Vec	63
Tabel 4. 20 Tabel Nilai Evaluasi Matriks MAP.....	67

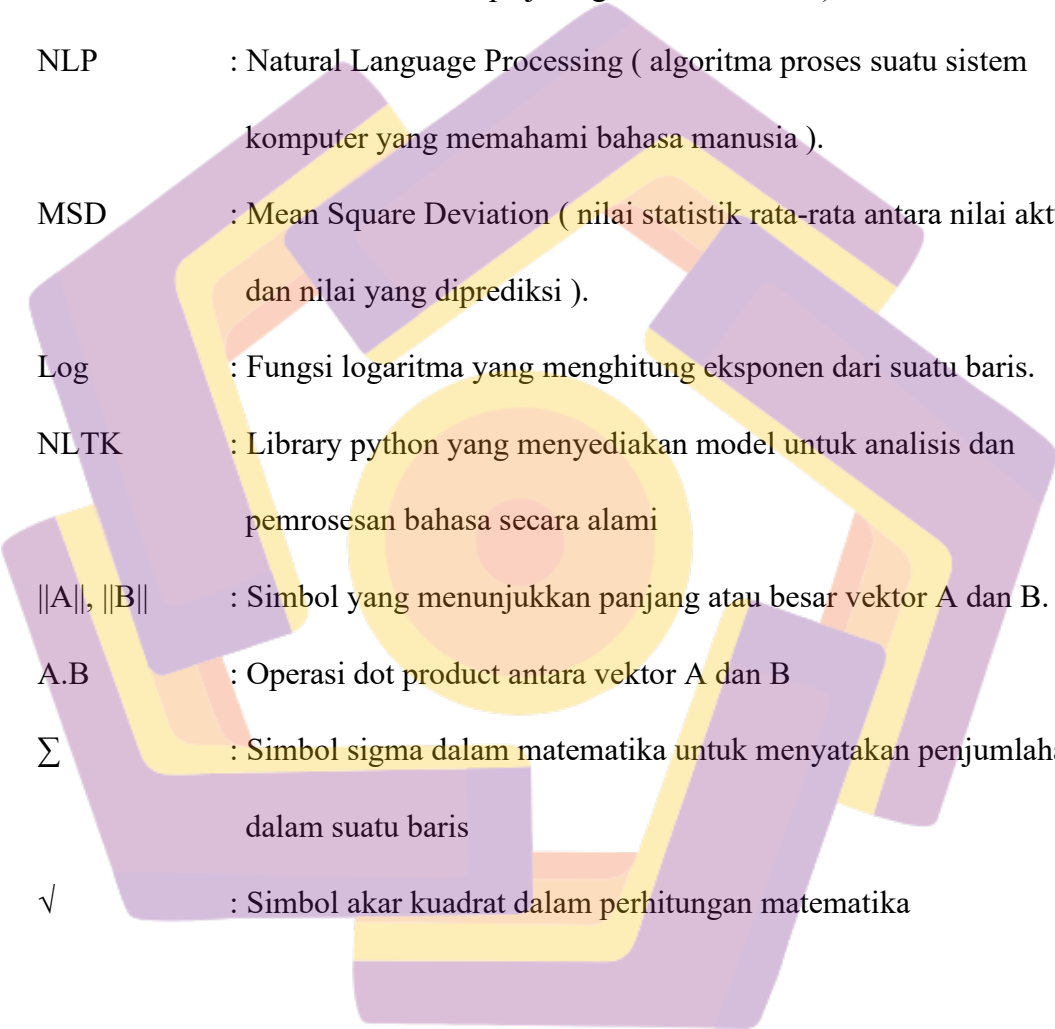
DAFTAR GAMBAR

Gambar 2. 1. Alur Kerja Algoritma TF-IDF [24]	19
Gambar 3. 1. Alur Penelitian	24
Gambar 4. 1. Dataset Buku Goodreads.....	31
Gambar 4. 2 Dataset Buku Lokal Kemdikbudristek.....	32
Gambar 4. 3 Tabel Dataset Setelah di Filter	34
Gambar 4. 4. Tabel Hasil Preprocessing Pada Dataset Kaggle	39
Gambar 4. 5. Hasil Pengujian Rekomendasi Buku Berjudul “The Hunger Games”	48
Gambar 4. 6. Hasil Pengujian Rekomendasi Buku Berjudul “Wuthering Heights”	48
Gambar 4. 7. Hasil Pengujian Rekomendasi Buku Berjudul “The Intelligent Investor ”	48
Gambar 4. 8. Arsitektur dari CBOW dan Skip-Gram pada Algoritma Word2Vec	50
Gambar 4. 9. Hasil Pengujian Rekomendasi Word2Vec Buku “The Hunger Games”	53
Gambar 4. 10. Hasil Pengujian Rekomendasi Word2Vec Buku “Wuthering Heights”	54
Gambar 4. 11. Hasil Pengujian Rekomendasi Word2Vec Buku “The Intelligent Investor”	54
Gambar 4. 12 Hasil Pengujian Rekomendasi Buku Berjudul “Ensiklopedia Sains Spektakuler - Energi dan Gerak”	58
Gambar 4. 13. Hasil Pengujian Rekomendasi Buku Berjudul “Ensiklopedia Geografi L-R”	59
Gambar 4. 14. Hasil Pengujian Rekomendasi Buku Berjudul “Seri Ensiklopedia IPA: Flora Khas Indonesia”	59
Gambar 4. 15. . Arsitektur Skip-Gram pada Algoritma Word2Vec	59
Gambar 4. 16 Hasil Pengujian Rekomendasi Buku “Ensiklopedia Sains Spektakuler - Energi dan Gerak”	60

Gambar 4. 17 Hasil Pengujian Rekomendasi Buku “Ensiklopedia Tematik Pengetahuan-Geografi-5”.....	61
Gambar 4. 18 Hasil Pengujian Rekomendasi Buku “Seri Ensiklopedia IPA: Flora Khas Indonesia”	61



DAFTAR LAMBANG DAN SINGKATAN



TF	: Term Frequency (Frekuensi kemunculan suatu kata).
IDF	: Inverse Document Frequency (Pembobotan pada kata berdasarkan seberapa jarang kata itu muncul).
NLP	: Natural Language Processing (algoritma proses suatu sistem komputer yang memahami bahasa manusia).
MSD	: Mean Square Deviation (nilai statistik rata-rata antara nilai aktual dan nilai yang diprediksi).
Log	: Fungsi logaritma yang menghitung eksponen dari suatu baris.
NLTK	: Library python yang menyediakan model untuk analisis dan pemrosesan bahasa secara alami
$\ A\ , \ B\ $	: Simbol yang menunjukkan panjang atau besar vektor A dan B.
$A.B$	: Operasi dot product antara vektor A dan B
Σ	: Simbol sigma dalam matematika untuk menyatakan penjumlahan dalam suatu baris
$\sqrt{}$	: Simbol akar kuadrat dalam perhitungan matematika

DAFTAR ISTILAH

- Fluktuatif : Suatu kondisi atau fenomena yang bersifat berubah ubah. secara tidak stabil baik kenaikan maupun penurunan sebuah statistik.
- Metadata : Suatu bentuk data berisikan informasi yang memberikan informasi tambahan dari data lain.
- Sparsity : Kondisi matriks atau data yang memiliki nilai nol atau kosong.
- Output : Hasil akhir dari suatu proses.
- Word embedding : Teknik representasi kata ke dalam bentuk vektor numerik (angka) untuk menemukan representasi vektor yang serupa.
- Knowledge-Based Filtering : Metode rekomendasi yang menggunakan parameter pengetahuan tentang item dan preferensi pengguna.
- Cosine Similarity : Metode pengukuran kemiripan antara dua vektor berdasarkan sudut kosinus di dalam suatu vektor.
- Collaborative Filtering : Metode rekomendasi yang menggunakan parameter berdasarkan preferensi pengguna atau kesamaan karakteristik antar item.
- Word2Vec : Algoritma pembelajaran representasi kata yang menghasilkan nilai vektor berdasarkan kata konteks disekitarnya.
- Doc2Vec : Algoritma representasi dokumen yang mengubah teks

menjadi nilai vektor.

FastText : Algoritma yang dibuat oleh Facebook yang merupakan bentuk lebih baik dibanding Word2Vec dengan keunggulan dapat menangani kata asing dan kata baku.

BERT : Bidirectional Encoder Representations for Transformers, merupakan model yang memahami konteks kata secara dua arah dalam kalimat.

Open-Source : Konsep yang menggambarkan suatu software atau data yang dibuat tersedia secara publik untuk dimodifikasi siapapun.

Preprocessing : Merupakan proses persiapan data mentah agar siap digunakan dalam proses analisis, pemodelan machine learning, ataupun perancangan sistem rekomendasi

INTISARI

Secara global, industri penerbitan buku terus mengalami pertumbuhan. Berdasarkan data UNESCO dan *International Publishers Association* (IPA), setiap tahun diterbitkan lebih dari dua juta judul buku baru di seluruh dunia. Di Indonesia sendiri, menurut data Perpustakaan Nasional (PERPUSNAS) tahun 2020 dan 2021, tercatat hampir 150.000 buku diterbitkan dalam satu tahun lebih tinggi dibanding 2019 berada di angka 120.000 buku. Peningkatan jumlah publikasi tersebut menunjukkan berkembangnya aktivitas penerbitan dan ketersediaan bahan bacaan, namun juga menimbulkan tantangan dalam membantu pembaca menemukan buku yang relevan dengan minat dan kebutuhannya. Dengan adanya sebuah sistem rekomendasi yang menerapkan berbagai macam algoritma dengan kebutuhan yang bermacam-macam tentu juga menjadi tantangan lagi bagi para pengelola untuk menghasilkan rekomendasi yang dapat disesuaikan dengan para pembaca.

Dengan adanya kedua masalah tersebut penelitian ini bertujuan untuk membandingkan kinerja antara dua pendekatan algoritma dalam sistem rekomendasi buku, yaitu Word2Vec dan *Cosine Similarity* dengan metode *Content-Based Filtering*. Kedua metode ini memiliki karakteristik yang berbeda dalam menghasilkan sebuah rekomendasi, namun keduanya tetaplah memiliki kebutuhan data yang minim dengan hanya memerlukan data dari konten atau buku itu sendiri. Maka dari itu pengujian ini dilakukan dengan menggunakan dataset buku yang mencakup informasi seperti judul, penulis, kategori, dan rating pembaca. Melalui perbandingan ini diharapkan dapat diperoleh pemahaman mengenai algoritma yang lebih efektif dalam meningkatkan relevansi rekomendasi serta memberikan kontribusi terhadap pengembangan sistem rekomendasi yang cocok dengan pembaca maupun pengguna sistem hingga penelitian kedepannya.

Kata kunci: Rekomendasi, Buku, Perbandingan, Efektivitas, Analisis

ABSTRACT

Based data on global, book publishing industry continues to grow. According to data from UNESCO and the International Publishers Association (IPA), more than two million new book titles are published worldwide each year. In Indonesia, based on data from the National Library (PERPUSNAS) for 2020 and 2021, nearly 150,000 books were published annually, a figure higher than in 2019, which recorded around 120,000 titles. This increase in publications reflects the expansion of publishing activities and the availability of reading materials, yet it also presents challenges in assisting readers to identify books that are relevant to their interests and needs. The implementation of recommendation systems that employ diverse algorithms tailored to varying requirements further adds complexity for publishers and system managers in generating recommendations that can be effectively personalized for readers.

Considering the aforementioned issues, this study aims to conduct a comparative analysis of the performance between two algorithmic approaches within a book recommendation system, specifically, the study focuses on Word2Vec and Cosine Similarity with Content-Based Filtering. These two methods possess distinct characteristics in generating recommendations; however, both require relatively minimal data, relying primarily on the attributes of the items or books themselves. Therefore, the evaluation in this research utilizes a book dataset containing information such as title, author, category, and reader ratings. Through this comparative analysis, the study aims to obtain insights into which algorithm demonstrates higher effectiveness in enhancing recommendation relevance and to contribute to the advancement of recommendation systems that better align with readers' preferences and future research development.

Keyword: Recommendation, Book, Comparative, Effectiveness, Analysis