

**SISTEM REKOMENDASI CAFÉ MENGGUNAKAN
ALGORITMA *BERT* DAN *COSINE SIMILARITY* (STUDI
KASUS: DATASET *YELP*)**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Informatika



disusun oleh

MAXIMILA BENA VANIA LAKSMIRA

22.11.5022

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**SISTEM REKOMENDASI CAFÉ MENGGUNAKAN
ALGORITMA *BERT* DAN *COSINE SIMILARITY* (STUDI
KASUS: DATASET *YELP*)**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Informatika



disusun oleh

MAXIMILA BENA VANIA LAKSMIRA
22.11.5022

Kepada

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

HALAMAN PERSETUJUAN

SKRIPSI

**SISTEM REKOMENDASI CAFÉ MENGGUNAKAN ALGORITMA *BERT*
DAN *COSINE SIMILARITY* (STUDI KASUS: DATASET *YELP*)**

yang disusun dan diajukan oleh

Maximila Benavanla Laksmira

22.11.5022

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 22 Januari 2026

Dosen Pembimbing,



Anna Baita, S.Kom., M.Kom.
NIK. 190302290

HALAMAN PENGESAHAN

SKRIPSI

**SISTEM REKOMENDASI CAFÉ MENGGUNAKAN ALGORITMA *BERT*
DAN *COSINE SIMILARITY* (STUDI KASUS: DATASET *YELP*)**

yang disusun dan diajukan oleh

Maximila Benavania Laksmira

22.11.5022

Telah dipertahankan di depan Dewan Penguji
pada tanggal 22 Januari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Windha Mega P. D., S.Kom., M.Kom.,
NIK. 190302185

Supriatin, A.Md., S.Kom., M.Kom.
NIK. 190302239

Anna Baita, S.Kom., M.Kom.
NIK. 190302290



Skrripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 22 Januari 2026

DEKAN FAKULTAS ILMU KOMPUTER



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

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Maximila Benavania Laksmira
NIM : 22.11.5022

Menyatakan bahwa Skripsi dengan judul berikut:

Sistem Rekomendasi Café Menggunakan Algoritma BERT dan Cosine Similarity

Dosen Pembimbing : Anna Balta, 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, 22 Januari 2026

Yang Menyatakan,



Maximila Benavania Laksmira

HALAMAN PERSEMBAHAN

Dengan mengucap rasa Syukur yang mendalam atas kehadiran Tuhan YME, yang telah memberikan kasih, kekuatan, dan kesehatan penulis, sehingga skripsi ini dapat terselesaikan dengan baik dan tanpa terkendala. Oleh karena itu, penulis ingin mempersembahkan kepada:

1. Tuhan Yang Maha Esa, atas segala rahmat, karunia, kesehatan, dan kemudahan yang diberikan selama proses penyusunan hingga penyelesaian skripsi ini. Terima kasih karena selalu menjadi tempat penulis berserah dan berkeluh kesah dalam setiap proses yang dilalui.
2. Bapak dan Ibu tercinta, terima kasih atas doa, semangat, nasihat, dukungan, kesabaran, serta kasih sayang yang selalu diberikan kepada penulis hingga saat ini.
3. Keluarga besar tercinta, yang senantiasa memberikan semangat, dukungan, serta doa dalam setiap langkah perjalanan penulis.
4. Teman-teman dekat saya tercinta yang tidak bisa saya sebutkan namanya satu persatu. Terima kasih telah menjadi bagian dari perjalanan ini, berbagi pengalaman dan pengetahuan, serta selalu memberikan dukungan dan semangat sehingga penulis dapat menyelesaikan skripsi ini.
5. Kepada diri sendiri, terima kasih karena telah mampu melewati setiap proses, belajar dari setiap kesalahan, mencintai setiap perjuangan yang membentuk perjalanan, serta tetap percaya pada kemampuan diri hingga akhirnya skripsi ini dapat diselesaikan dengan baik.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat, karunia, serta penyertaan-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "**Sistem Rekomendasi Café Menggunakan Algoritma *BERT* dan *Cosine Similarity***" dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana pada program studi yang penulis tempuh.

Dalam proses penyusunan skripsi ini, penulis menyadari bahwa banyak pihak yang telah memberikan bantuan, dukungan, bimbingan, serta arahan sehingga skripsi ini dapat terselesaikan. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih kepada:

1. Bapak Prof. Dr. M. Suyanto, selaku Rektor Universitas AMIKOM Yogyakarta.
2. Ibu Prof. Dr. Kusriani, M., Dekan Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta.
3. Ibu Eli Pujastuti, M.Kom, selaku Ketua Program Studi Informatika Universitas AMIKOM Yogyakarta
4. Ibu Anna Baita, M.Kom, selaku dosen pembimbing yang telah memberikan bimbingan, arahan, serta motivasi kepada penulis selama proses penyusunan skripsi ini.
5. Tim Dosen Penguji yang telah memberikan saran dan masukan yang sangat bermanfaat bagi penyempurnaan skripsi ini.
6. Seluruh pihak yang telah mendukung penulisan skripsi ini yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun demi penyempurnaan karya ini di masa yang akan datang. Semoga skripsi ini dapat memberikan manfaat bagi pembaca dan bagi pengembangan ilmu pengetahuan.

Yogyakarta, 22 Januari 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.....	xi
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN.....	xiii
DAFTAR LAMBANG DAN SINGKATAN.....	xiv
DAFTAR ISTILAH.....	xvii
INTISARI.....	xx
ABSTRACT.....	xxi
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah.....	2
1.4 Tujuan Penelitian.....	3
1.5 Manfaat Penelitian.....	3
1.6 Sistematika Penulisan.....	4

BAB II TINJAUAN PUSTAKA.....	6
2.1 Studi Literatur.....	6
2.2 Dasar Teori.....	14
2.2.1 Ulasan Konsumen.....	14
2.2.2 Sistem Rekomendasi.....	15
2.2.3 <i>Natural Language Processing</i>	16
2.2.4 Analisis Sentimen.....	18
2.2.5 <i>Deep Learning</i>	19
2.2.6 <i>Preprocessing</i>	20
2.2.7 <i>Stratified Sampling</i>	22
2.2.8 <i>Transformer</i>	23
2.2.9 <i>BERT</i>	24
2.2.10 <i>Cosine Similarity</i>	28
2.2.11 Metrik Evaluasi.....	29
BAB III METODE PENELITIAN.....	36
3.1 Objek Penelitian.....	36
3.2 Alur Penelitian.....	36
3.2.1 Pengumpulan Dataset.....	39
3.2.2 <i>Preprocessing Data</i>	39
3.2.3 <i>Stratified Sampling</i>	40
3.2.4 Modeling.....	40
3.2.5 <i>Embedding Ulasan</i>	41

3.2.6	<i>Profile Café</i>	42
3.2.7	Sistem Rekomendasi.....	43
3.2.8	Evaluasi Rekomendasi.....	43
3.2.9	<i>Deployment</i>	44
3.3	Alat dan Bahan.....	44
3.3.1	Data Penelitian.....	44
3.3.2	Alat dan Instrumen Penelitian.....	45
BAB IV HASIL DAN PEMBAHASAN.....		47
4.1	Deskripsi Data.....	47
4.2	<i>Preprocessing Data</i>	49
4.2.1	<i>Data Cleaning</i>	49
4.2.2	<i>Exploratory Data Analysis (EDA)</i>	54
4.2.3	<i>Handling Missing Value</i>	60
4.2.4	<i>Handling Outlier</i>	60
4.2.5	Pelabelan Sentimen.....	61
4.3	<i>Data Split (Stratified Sampling)</i>	62
4.4	Pelatihan Model Menggunakan <i>BERT</i>	62
4.4.1	Persiapan <i>Input</i> dan Tokenisasi.....	62
4.4.2	Konfigurasi Pelatihan Model.....	63
4.4.3	Proses Pelatihan dan <i>Early Stopping</i>	64
4.5	Evaluasi Model Klasifikasi Sentimen.....	65
4.6	Pembentukan <i>Profile Café</i>	65

4.6.1	Pembentukan <i>Embedding</i> Ulasan.....	66
4.6.2	Agregasi <i>Embedding</i> untuk Pembentukan <i>Embedding</i> Café.....	66
4.6.3	<i>Profile</i> Café.....	73
4.6.4	Perhitungan <i>Similarity Score</i>	75
4.7	Hasil Sistem Rekomendasi Café.....	77
4.7.1	<i>Output</i> Sistem Rekomendasi.....	77
4.7.2	Pengujian Sistem Rekomendasi.....	80
4.7.3	Analisis Kesesuaian Hasil Rekomendasi.....	107
4.8	Evaluasi Sistem Rekomendasi.....	108
4.8.1	Evaluasi Seluruh Dataset.....	108
4.8.2	Evaluasi <i>Hold-out</i>	109
4.9	<i>Deployment</i>	111
BAB V PENUTUP.....		113
5.1	Kesimpulan.....	113
5.2	Saran.....	114
REFERENSI.....		115
LAMPIRAN.....		120

DAFTAR TABEL

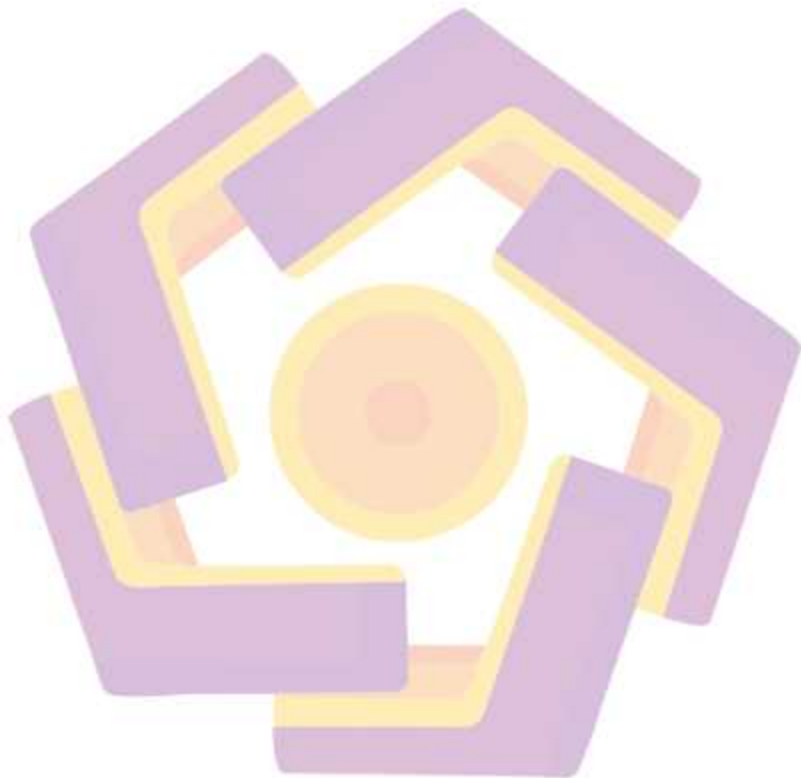
Tabel 2. 1 Keaslian Penelitian	8
Tabel 3. 1 Ilustrasi <i>Embedding</i> Ulasan Cafe	42
Tabel 4. 1 Atribut Dataset	48
Tabel 4. 2 Data <i>Cleaning</i>	50
Tabel 4. 3 Data <i>Penanganan Outlier</i>	60
Tabel 4. 4 Hasil Label Sentimen	61
Tabel 4. 5 Pembagian Data	62
Tabel 4. 6 Konfigurasi Pelatihan Model <i>BERT</i>	63
Tabel 4. 7 Evaluasi Model Klasifikasi Sentimen	65
Tabel 4. 8 Contoh <i>Embedding</i> Ulasan	68
Tabel 4. 9 Contoh Hasil <i>Sentimen Score</i> Perulisan	73
Tabel 4. 10 Hasil Perhitungan <i>Sentimen Score</i>	74
Tabel 4. 11 <i>Similarity Score</i> dan <i>Sentimen Score</i> Cafe	76
Tabel 4. 12 Hasil Sitem Rekomendasi Cafe	78
Tabel 4. 13 Verifikasi Manual Hasil Rekomendasi Cafe	81
Tabel 4. 14 Ulasan Representatif <i>Starbucks</i>	82
Tabel 4. 15 Ulasan Representatif Starbucks	85
Tabel 4. 16 Ulasan Representatif The Coffee Room	91
Tabel 4. 17 Ulasan Representatif Bagel Junction Deli	95
Tabel 4. 18 Ulasan Representatif Barbie's Soft Pretzels	99
Tabel 4. 19 Hasil Evaluasi Sistem Rekomendasi	108
Tabel 4. 20 Hasil Evaluasi <i>Hold-out</i>	110

DAFTAR GAMBAR

Gambar 2. 1 Boxplot untuk Deteksi Outlier	21
Gambar 2. 2 <i>BERT</i> input <i>embedding</i>	25
Gambar 2. 3 Proses <i>pre-training</i> dan <i>fine-tuning BERT</i>	25
Gambar 2. 4 Proses <i>fine-tuning</i> pada tingkat token	26
Gambar 2. 5 <i>Confusion Matrix</i>	30
 Gambar 3. 1 Alur Penelitian A	 37
Gambar 3. 2 Alur Penelitian B	38
 Gambar 4. 1 <i>Histogram</i> Distribusi <i>Rating Review (Stars)</i>	 54
Gambar 4. 2 <i>Word Cloud</i> Hasil Ulasan	55
Gambar 4. 3 <i>Boxplot</i> Panjang Ulasan dan <i>Rating Review (Stars)</i>	56
Gambar 4. 4 <i>Scatter Plot</i> Jumlah Ulasan dan <i>Rating Rata-Rata Cafe</i>	57
Gambar 4. 5 <i>Histogram</i> Panjang Ulasan A	58
Gambar 4. 6 <i>Boxplot</i> Panjang Ulasan B	59
Gambar 4. 7 Distribusi <i>Average Similarity</i> per <i>Cafe</i>	109
Gambar 4. 8 Distribusi <i>Average Similarity Top-K Hold-out</i>	110
Gambar 4. 9 Antar muka Sistem Rekomendasi	111

DAFTAR LAMPIRAN

Lampiran 1 Hasil Keseluruhan Sistem Rekomendasi.....	120
--	-----



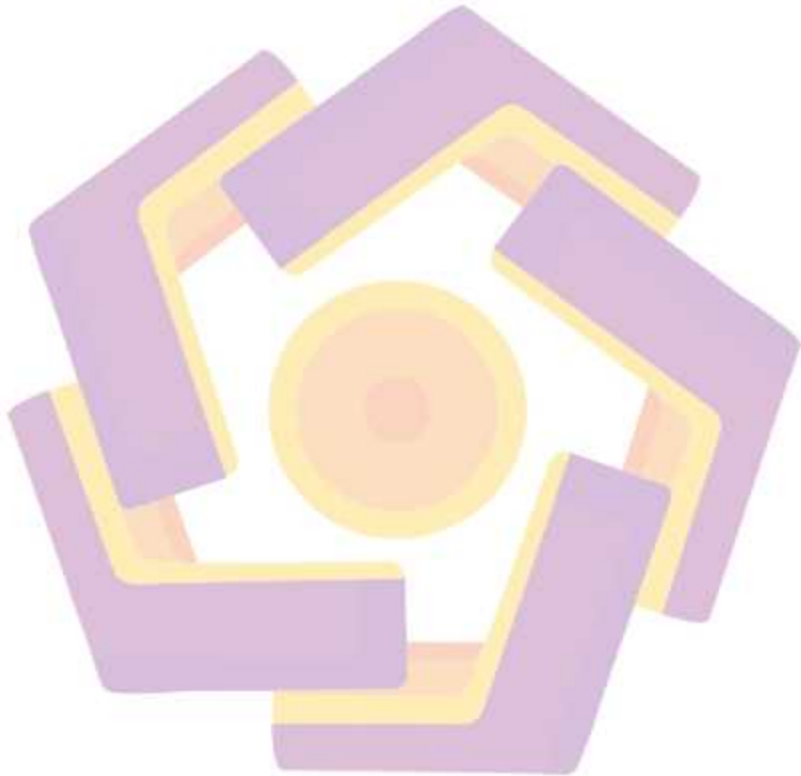
DAFTAR LAMBANG DAN SINGKATAN

α	Parameter pengendali kontribusi komponen
i	Item
sim	Similarity
IQR	Interquartile Range
n_h	Jumlah sampel dari strata ke- h
n	Total sampel
N_h	Ukuran populasi strata ke- h
S_h	Simpangan baku pada strata ke- h
L	Jumlah strata
Q	Query
K	Key
V	Value
T	Transform
d_k	Dimensi dari Key
E	Input embedding
T_l	Token ke- i
[CLS]	Simbol khusus keluaran klasifikasi
[SEP]	Simbol pemisah khusus
θ	Theta, digunakan sebagai parameter lapisan
l	Learning rate
η_l	Learning rate lapisan ke- l
ξ	Decay factor
e	Embedding
e_i	Embedding ulasan ke- i
TP	True positif
FP	False positif
FN	False negatif
TN	True negatif

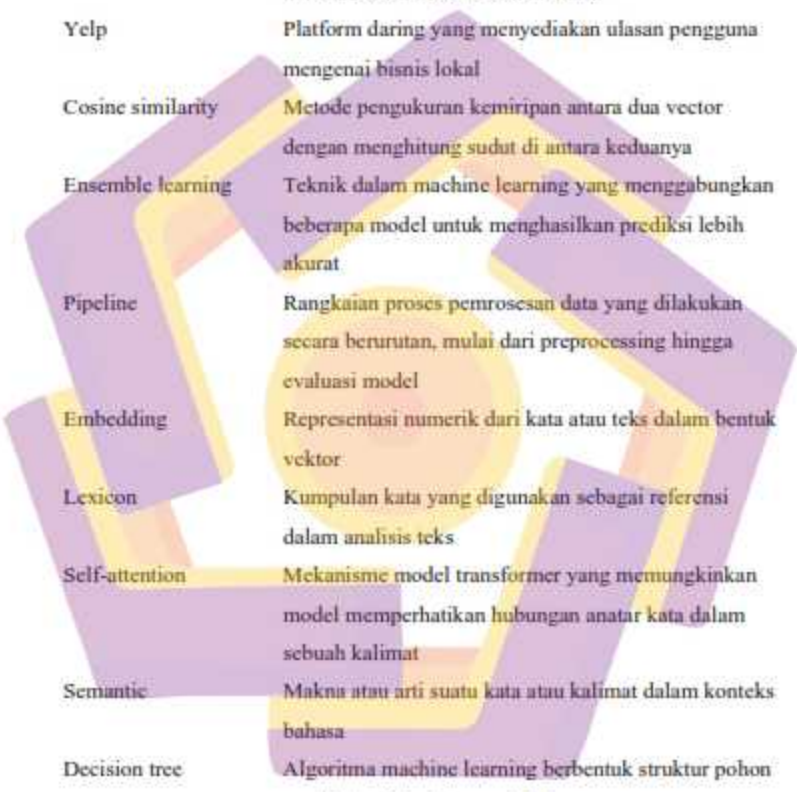


q	Item acuan atau item kueri
D	Discount
f	Fungsi
log	Logaritma operasi matematika untuk menentukan pangkat suatu bilangan terhadap basis tertentu
BERT	Bidirectional Encoder Representations from Transformers
LSTM	Long Short Term Memory
SVM	Support Vector Machine
NLP	Natural Language Preprocessing
EDA	Exploratory Data Analysis
Bi-LSTM	Bidirectional Long Short Term Memory
MLP	Multilayer Perceptron
KNN	K-Nearest Neighbors
CNN	Convolutional Neural Network
DL	Deep Learning
GPU	Graphics Processing Unit
CPU	Computer Processing Unit
RNN	Recurrent Neural Network
MLM	Masked Language Modeling
NSP	Next Sentence Prediction
NER	Name Entity Recognition
DCG	Discounted Cumulative Gain
nDCG	Normalized Discounted Cumulative Gain)
HTML	HyperText Markup Language
URL	Uniform Resource Locator
RAM	Random Access Memory
VRAM	Video Random Acces Memory
JSON	JavaScript Object Notation
CSV	Comma-Separated Values
CUDA	Compute Unified Device Architecture

float32	Tipe data numerik ntuk merepresentasikan bilangan pecahan dengan presisi 32-bit
Wi-Fi	Teknologi jaringan nirkabel yang memungkinkan perangkat elektronik terhubung ke jaringan internet tanpa kabel
TF-IDF	Term Frequency-Inverse Document Frequency



DAFTAR ISTILAH



Vektor	Besaran yang mempunyai arah
Hybrid	Pendekatan sistem rekomendasi yang menggabungkan dua atau lebih metode rekomendasi
Yelp	Platform daring yang menyediakan ulasan pengguna mengenai bisnis lokal
Cosine similarity	Metode pengukuran kemiripan antara dua vector dengan menghitung sudut di antara keduanya
Ensemble learning	Teknik dalam machine learning yang menggabungkan beberapa model untuk menghasilkan prediksi lebih akurat
Pipeline	Rangkaian proses pemrosesan data yang dilakukan secara berurutan, mulai dari preprocessing hingga evaluasi model
Embedding	Representasi numerik dari kata atau teks dalam bentuk vektor
Lexicon	Kumpulan kata yang digunakan sebagai referensi dalam analisis teks
Self-attention	Mekanisme model transformer yang memungkinkan model memperhatikan hubungan antar kata dalam sebuah kalimat
Semantic	Makna atau arti suatu kata atau kalimat dalam konteks bahasa
Decision tree	Algoritma machine learning berbentuk struktur pohon untuk klasifikasi atau prediksi
Logic regression	Metode statistic untuk memprediksi probabilitas suatu kelas dalam masalah klasifikasi
Token	Unit terkecil dari teks berupa kata atau sub-kata
Corpus	Kumpulan dokumen teks yang digunakan sebagai sumber data

Count vectorization	Metode representasi teks dengan menghitung jumlah kemunculan kata dalam dokumen
Vocabulary	Kumpulan seluruh kata unik yang terdapat dalam suatu corpus
Pre-trained	Model yang telah dilatih sebelumnya menggunakan dataset besar
Fine-tuning	Proses melatih ulang model pre-trained pada dataset yang lebih spesifik
Supervised	Metode pembelajaran menggunakan data yang memiliki label
Semi-supervised	Menggunakan kombinasi data berlabel dan tidak berlabel
Unsupervised	Metode pembelajaran tanpa label
Attention	Mekanisme yang memungkinkan model fokus pada bagian tertentu dari input
Noise	Data yang tidak relevan atau mengandung kesalahan
Encoder	Bagian model yang mengubah input menjadi representasi vector
Decoder	Bagian model yang menghasilkan output dari representasi encoder
Softmax	Fungsi aktivasi yang mengubah nilai menjadi distribusi probabilitas
Classifier	Model yang digunakan untuk mengklasifikasikan data ke dalam kategori tertentu
Optimizer	Algoritma yang digunakan untuk memperbarui parameter model
Learning rate	Ukuran Langkah pembaruan parameter model
Overfitting	Terjadi ketika model terlalu menyesuaikan diri dengan data pelatihan
Python	Bahasa pemrograman dalam data science dan machine learning



Google colab	Lingkungan komputasi berbasis cloud untuk code Python
Pandas	Pustaka Python untuk manipulasi dan analisis data
NumPy	Pustaka Python untuk komputasi numerik
Matplotlib	Pustaka Python untuk visualisasi data
Seaborn	Pustaka visualisasi data berbasis Matplotlib
Scikit-learn	Pustaka machine learning pada Python
PyTorch	Framework deep learning untuk membangun dan melatih model neural network
Cloud	Layanan komputasi internet yang menyediakan penyimpanan dan komputasi jarak jauh
Word cloud	Visualisasi frekuensi kata dalam bentuk ukuran teks
Truncated	Proses pemotongan teks yang melebihi panjang maksimum input model
Batch size	Jumlah data yang diproses dalam satu iterasi pelatihan
Query	Representasi untuk mencari informasi relevan

INTISARI

Banyaknya ulasan café di platform digital sering kali menyulitkan pengguna dalam menemukan café yang sesuai dengan preferensi suasana, seperti tempat yang nyaman dan tenang untuk belajar. Informasi *rating* bintang (*stars*) saja tidak cukup merepresentasikan pengalaman pengguna secara menyeluruh karena tidak menangkap konteks semantik, sentimen, dan suasana (*mood*) yang terkandung dalam ulasan teks.

Penelitian ini mengembangkan sistem rekomendasi café berbasis *content-based filtering* dengan memanfaatkan model *BERT* untuk mengekstraksi *embedding* semantik, sentimen, dan suasana dari ulasan pengguna. Profile café dibentuk melalui agregasi *embedding* menggunakan *mean pooling*, kemudian tingkat kesesuaiannya terhadap preferensi pengguna dihitung menggunakan *cosine similarity*. Selain itu, *sentimen score* digunakan sebagai komponen tambahan dalam pendekatan *hybrid* untuk menghasilkan pemeringkatan rekomendasi yang lebih kontekstual. Evaluasi sistem dilakukan menggunakan metrik *Average Similarity* dan *nDCG* pada skenario seluruh dataset dan *hold-out item*.

Hasil penelitian menunjukkan sistem rekomendasi berbasis *BERT* dan *cosine similarity* mampu memberikan rekomendasi yang relevan dan kontekstual. Evaluasi menunjukkan *Average Similarity Top-5*: 98% dan *nDCG*: 97%, menandakan kualitas rekomendasi yang tinggi. Sistem ini efektif membantu pengguna memilih café sesuai kebutuhan dan menjadi acuan untuk pengembangan sistem rekomendasi berbasis teks di masa depan.

Kata kunci: sistem rekomendasi café, *BERT*, *cosine similarity*, analisis sentimen, *content-based filtering*.

ABSTRACT

The large number of café reviews on digital platforms often makes it difficult for users to find cafés that match their preferred atmosphere, such as comfortable and quiet places for studying. Star rating information alone is insufficient to fully represent user experiences because it fails to capture the semantic context, sentiment, and mood embedded in textual reviews.

This study develops a café recommendation system based on content-based filtering by leveraging the BERT model to extract semantic, sentiment, and mood embeddings from user reviews. Café profiles are constructed through embedding aggregation using mean pooling, and their relevance to user preferences is calculated using cosine similarity. In addition, sentiment scores are incorporated as an additional component in a hybrid approach to produce more contextual recommendation rankings. System evaluation is conducted using the Average Similarity and nDCG metrics under full-dataset and hold-out item scenarios.

The study shows that the recommendation system based on BERT and cosine similarity is capable of providing relevant and contextual recommendations. Evaluation results indicate Average Similarity Top-5: 98% and nDCG: 97% reflecting high recommendation quality. This system effectively assist users in selecting caafes according to their need and serves as a reference for the futre developmeny of txt-based recommendation systems.

Keyword: *café recommendation system, BERT, cosine similarity, sentiment analysis, content-based filtering.*