



SKRIPSI

**MODEL *STACKING ENSEMBLE* DENGAN
MAXIMUM ENTROPY SEBAGAI *META-
LEARNER* UNTUK KLASIFIKASI OPINI ULASAN
BUKU PADA *PLATFORM GOODREADS***

SESILLIA AGUSTIN
NPM 22083010038

DOSEN PEMBIMBING

Wahyu Syaifullah J. S, S.Kom., M.Kom.
Alfan Rizaldy Pratama, S.Tr.T., M.Tr.Kom.

**KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI SAINS DATA
SURABAYA
2026**

LEMBAR PENGESAHAN

MODEL *STACKING ENSEMBLE* DENGAN *MAXIMUM ENTROPY* SEBAGAI *META-LEARNER* UNTUK KLASIFIKASI OPINI ULASAN BUKU PADA *PLATFORM GOODREADS*

Oleh:
SESILLIA AGUSTIN
NPM. 22083010038

Telah dipertahankan di hadapan dan diterima oleh Tim Penguji Sidang Skripsi Program Studi Sains Data Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur pada Tanggal 26 Agustus 2026:

Menyetujui,

Wahyu Syaifullah J. S., S.Kom.,
M.Kom.

NIP. 19860825 202121 1 003


.....

(Pembimbing I)

Alfan Rizaldy Pratama,
S.Tr.T., M.Tr.Kom.

NIP. 19990606 202406 1 001


.....

(Pembimbing II)

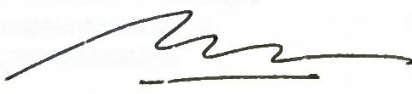
Aviolla Terza Damaliana, S.Si.,
M.Stat.

NIP. 19940802 202203 2 015


.....

(Ketua Penguji)

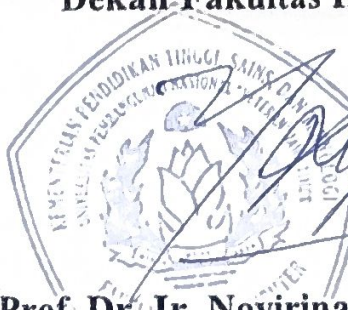
Andri Fauzan Adziima, M. Si.
NIP. 19950512 202406 1 001


.....

(Penguji I)

Mengetahui,

Dekan Fakultas Ilmu Komputer



Prof. Dr. Ir. Novirina Hendrasarie, MT.
NIP. 19681126 199403 2 001

LEMBAR PERSETUJUAN

**MODEL *STACKING ENSEMBLE* DENGAN *MAXIMUM ENTROPY*
SEBAGAI *META-LEARNER* UNTUK KLASIFIKASI OPINI ULASAN
BUKU PADA *PLATFORM GOODREADS***

Oleh:
SESILLIA AGUSTIN
NPM. 22083010038

Telah disetujui untuk mengikuti Ujian Skripsi

Menyetujui,

**Koordinator Program Studi Sains Data
Fakultas Ilmu Komputer**



Amri Muhaimin, S.Stat., M.Stat., M.S.
NIP. 19950723 202406 1 002

SURAT PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan di bawah ini:

Nama Mahasiswa : Sesillia Agustin
NPM : 22083010038
Program : Sarjana (S1)
Program Studi : Sains Data
Fakultas : Fakultas Ilmu Komputer

Menyatakan bahwa dalam dokumen ilmiah Skripsi ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan secara lengkap dalam daftar pustaka.

Dan saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi. Apabila di kemudian hari ditemukan indikasi plagiat pada Skripsi ini, saya bersedia menerima sanksi sesuai dengan peraturan perundang-undangan yang berlaku.

Demikian surat pernyataan ini saya buat dengan sesungguhnya tanpa ada paksaan dari siapapun juga dan untuk dipergunakan sebagaimana mestinya.



Surabaya, 4 September 2026
Yang Membuat Pernyataan,



SESILLIA AGUSTIN
NPM. 22083010038

ABSTRAK

Nama Mahasiswa / NPM : Sesillia Agustin / 22083010038
Judul Skripsi : Model *Stacking ensemble* dengan *Maximum Entropy* sebagai Meta-Learner untuk Klasifikasi Opini Ulasan Buku pada Platform *Goodreads*
Dosen Pembimbing : 1. Wahyu Syaifullah J. S, S.Kom., M.Kom.
2. Alfian Rizaldy Pratama, S.Tr.T., M.Tr.Kom.

Ulasan buku pada platform *Goodreads* menjadi salah satu sumber informasi penting bagi calon pembaca untuk menentukan apakah suatu buku layak dibaca. Namun, teks ulasan yang seringkali panjang membuat opini dari penulis ulasan sulit untuk dipahami. Penelitian ini bertujuan membangun model analisis sentimen untuk mengklasifikasikan ulasan buku dari *Goodreads* ke dalam dua kelas, yaitu *recommended* dan *not recommended*, menggunakan pendekatan *Stacking ensemble* dengan *Naive Bayes* dan *XGBoost* sebagai *Base learner*, serta *Maximum Entropy* sebagai *meta-learner*. Data penelitian terdiri dari 900 ulasan hasil *scraping* dari enam judul buku dengan genre berbeda, yang melalui tahap *preprocessing* (*text cleaning, case folding, tokenizing, stopword removal, dan lemmatization*) sebelum direpresentasikan secara numerik menggunakan metode BM25. Penelitian ini turut membandingkan dua metode pelabelan data, yaitu VADER (otomatis) dan anotasi manual, serta menerapkan teknik *Class Weighting* untuk menangani ketidakseimbangan kelas. Model diuji pada tiga rasio pembagian data (60:40, 70:30, 80:20) untuk kedua *dataset*. Hasil terbaik dicapai pada *dataset* hasil anotasi manual dengan rasio 70:30, menghasilkan *macro F1-score* sebesar 0,76 dan *recall* kelas minoritas (*not recommended*) sebesar 0,73. *Dataset* hasil anotasi manual secara konsisten mengungguli *dataset* berbasis VADER pada seluruh rasio yang diuji, menunjukkan pengaruh signifikan kualitas metode pelabelan terhadap performa model. Hasil penelitian ini diimplementasikan ke dalam sebuah *Dashboard* interaktif berbasis web bernama "*The Reading Room*", yang menampilkan fitur analisis sentimen pada ulasan buku secara *real-time* bagi pengguna.

Kata Kunci: Analisis Sentimen, *Goodreads*, *Stacking ensemble*, *Naive Bayes*, *XGBoost*, *Maximum Entropy*.

ABSTRACT

Student Name / NPM : Sesillia Agustin / 22083010038
Undergraduate thesis title : *Stacking ensemble Model With Maximum Entropy As Meta-Learner For Opinion Classification Of Book Reviews On The Goodreads Platform*
Advisors : 1. Wahyu Syaifullah J. S, S.Kom., M.Kom.
2. Alfan Rizaldy Pratama, S.Tr.T., M.Tr.Kom.

Book reviews on the Goodreads platform serve as an important source of information for prospective readers in deciding whether a book is worth reading. However, the long formed reviews making it harder for reader to understand the initial meaning of said review, creating a need for an automated sentiment analysis system. This study aims to build a sentiment analysis model to classify Goodreads book reviews into two classes, recommended and not recommended, using a Stacking ensemble approach with Naive Bayes and XGBoost as Base learners, and Maximum Entropy (Logistic Regression) as the meta-learner. The research data consists of 900 reviews scraped from six book titles across different genres, which underwent a preprocessing stage (text cleaning, case folding, tokenizing, stopword removal, and lemmatization) before being numerically represented using the BM25 method. This study also compares two data labeling methods, namely VADER (automatic) and manual annotation, and applies Class Weighting techniques to address class imbalance. The model was tested on three data-splitting ratios (60:40, 70:30, 80:20) for both datasets. The best result was achieved on the manually annotated dataset with a 70:30 ratio, yielding a macro F1-score of 0.76 and a minority-class (not recommended) recall of 0.73. The manually annotated dataset consistently outperformed the VADER-based dataset across all tested ratios, indicating the significant influence of labeling method quality on model performance. The results of this research were implemented into an interactive web-based Dashboard named "The Reading Room," which provides real-time sentiment analysis features on book reviews for users.

Keywords: *Sentiment Analysis, Goodreads, Stacking ensemble, Naive Bayes, XGBoost, Maximum Entropy.*

KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas segala rahmat, hidayah dan karunia-Nya kepada penulis sehingga proposal skripsi dengan judul **“Model *Stacking Ensemble* dengan *Maximum Entropy* sebagai *Meta-Learner* untuk Klasifikasi Opini Ulasan Buku pada Platform Goodreads”** dapat terselesaikan dengan baik.

Penulis mengucapkan terima kasih kepada Bapak Wahyu Syaifullah J. S, S.Kom., M.Kom., selaku Dosen Pembimbing I dan Bapak Alfian Rizaldy Pratama, S.Tr.T., M.Tr.Kom., selaku Dosen Pembimbing 2 yang telah meluangkan waktunya untuk memberikan bimbingan, nasehat serta motivasi kepada penulis. Selain itu, selama penyusunan proposal skripsi penulis juga banyak menerima bantuan dari berbagai pihak. Untuk itu penulis mengucapkan terima kasih kepada:

1. Ibu Prof. Dr. Ir. Novirina Hendrasarie, MT., selaku Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
2. Bapak Amri Muhaimin, S.Stat., M.Stat., M.S., selaku Koordinator Program Studi Sains Data Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur.
3. Bapak Trimono, S.Si., M.Si., selaku penanggung jawab skripsi Program Studi Sains Data.
4. Seluruh Dosen dan Staff Program Studi Sains Data.
5. Kedua Orang Tua Penulis yang telah mengorbankan begitu banyak hal untuk mendukung karir akademik penulis, dan sentiasa mendukung penulis dalam segala hal tanpa terkecuali.
6. Chelsea Dwi Septiani, selaku adik, teman, partner, dan segala hal yang baik di dunia ini, yang selalu mendukung penulis dengan cara uniknya sendiri.
7. Teman-teman kuliah Penulis, Desi, Dinda, Aldea, Ardina, Ayu, Emilia, dan begitu banyak teman-teman lain yang tidak bisa penulis sebutkan satu-persatu namanya, karena telah menjadi teman untuk sandaran bahu penulis yang sendirian jauh dari rumah.
8. Teman-teman penulis yang jauh jaraknya, Ummi, Pania, Ana, Era, Ayu, dan teman lainnya yang tidak mungkin penulis sebutkan semua namanya, yang

selalu mendukung dan menyemangati penulis meski tidak bertemu secara langsung.

9. Madeline Miller, Emily Wilson, Mary Shelley, Jane Austen, The Bronte Sisters, Luisa May Alcott, Dee Lestari, Ziggy, Virginia Woolf, Sappho, serta semua penulis perempuan lain yang menjadi inspirasi awal dari topik yang diambil dalam penelitian ini, yang secara langsung dan tidak langsung terus berusaha memperjuangkan suara, pendapat, dan pemikiran seluruh perempuan di dunia dengan mengabadikannya dalam karya mereka.

Penulis menyadari bahwa di dalam penyusunan skripsi ini banyak terdapat kekurangan. Untuk itu kritik dan saran yang membangun dari semua pihak sangat diharapkan demi kesempurnaan penulisan skripsi ini. Akhirnya, dengan segala keterbatasan yang penulis miliki semoga laporan ini dapat bermanfaat bagi semua pihak umumnya dan penulis pada khususnya.

Surabaya, September 2026

Penulis

DAFTAR ISI

HALAMAN SAMPUL.....	i
LEMBAR PENGESAHAN	ii
LEMBAR PERSETUJUAN	iii
SURAT PERNYATAAN BEBAS PLAGIASI.....	iv
ABSTRAK	v
ABSTRACT	vi
KATA PENGANTAR.....	vii
DAFTAR ISI.....	ix
DAFTAR GAMBAR.....	xi
DAFTAR TABEL	xiv
DAFTAR LAMPIRAN	xv
DAFTAR NOTASI.....	xvi
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	6
1.3. Batasan Masalah	7
1.4. Tujuan Penelitian	8
1.5. Manfaat Penelitian	8
BAB II TINJAUAN PUSTAKA	11
2.1. Penelitian Terdahulu	11
2.2. Kerangka Teori	15
2.2.1. Klasifikasi Opini atau Analisis Sentimen	15
2.2.2. Preprocessing	16

2.2.3. <i>Feature selection</i>	18
2.2.4. <i>Labelling</i>	19
2.2.5. <i>Imbalance Handling</i>	20
2.2.6. <i>Naive Bayes</i>	21
2.2.7. <i>XGBoost</i>	23
2.2.8. <i>Maximum Entropy</i>	25
2.2.9. <i>Stacking ensemble</i>	27
2.2.10. <i>Matriks Evaluasi</i>	29
2.2.11. <i>Goodreads</i>	31
BAB III METODOLOGI PENELITIAN.....	33
3.1. Variabel Penelitian dan Sumber Data.....	33
3.2. Variabel Penelitian dan Sumber Data.....	36
3.2.1. <i>Scrapping Data</i>	36
3.2.2. <i>Preprocessing Data</i>	37
3.2.3. <i>Labelling</i>	37
3.2.4. <i>Feature selection (BM25)</i>	38
3.2.5. <i>Class Weighting</i>	38
3.2.6. <i>Model Building</i>	39
3.2.7. <i>Evaluation</i>	40
3.3. <i>Desain Sistem</i>	40
BAB IV HASIL DAN PEMBAHASAN	45
4.1. Hasil Penelitian.....	45
4.1.1. <i>Pengumpulan Data</i>	45
4.1.2. <i>Text Preprocessing</i>	50
4.1.3. <i>Labelling</i>	57
4.1.4. <i>Model Building</i>	62
4.1.5. <i>Matriks Evaluasi dan Hasil</i>	83
4.2. <i>Tampilan Antar Muka (Graphical User Interface)</i>	95
BAB V PENUTUP	103

5.1. Kesimpulan	103
5.2. Saran Pengembangan	105
DAFTAR PUSTAKA.....	107
LAMPIRAN.....	111

DAFTAR GAMBAR

Gambar 2. 1 <i>Decision tree</i> dalam <i>XGBoost</i>	24
Gambar 2. 2 <i>Confusion Matrix</i>	29
Gambar 3. 1 Proses analisis.....	36
Gambar 3. 2 Alur <i>Preprocessing</i>	37
Gambar 3. 3 Alur Model	39
Gambar 3. 4 Halaman Utama Sistem.....	41
Gambar 3. 5 Halaman Buku.....	41
Gambar 3. 6 Halaman Buku dan Analisis.....	42
Gambar 3. 7 Halaman Model dan data	42
Gambar 3. 8 Sketsa <i>Dashboard</i> ketika analisis belum ditampilkan	43
Gambar 3. 9 Sketsa <i>Dashboard</i> ketika analisis belum ditampilkan	43
Gambar 3. 10 Alur Sistem	43
Gambar 4. 1 Tampilan Halaman <i>Goodreads</i>	45
Gambar 4. 2 Rasio Kelas Labeling Manual	59
Gambar 4. 3 Rasio Hasil Labeling Otomatis	62
Gambar 4. 4 Tree Pertama Perhitungan <i>XGBoost</i> Manual	77
Gambar 4. 5 <i>Confusion Matrix</i> Model VADER Rasio 60:40	83
Gambar 4. 6 <i>Confusion Matrix</i> Model VADER Rasio 70:30	84
Gambar 4. 7 <i>Confusion Matrix</i> Model VADER Rasio 80:20	84
Gambar 4. 8 <i>Confusion Matrix</i> Model Manual Rasio 60:40.....	85
Gambar 4. 9 <i>Confusion Matrix</i> Model Manual Rasio 70:30.....	86
Gambar 4. 10 <i>Confusion Matrix</i> Model Manual Rasio 80:20.....	86
Gambar 4. 11 AUC Model-model <i>Standalone</i> pada <i>Dataset</i> Manual	91
Gambar 4. 12 AUC Model-model <i>Standalone</i> pada <i>Dataset</i> VADER	91
Gambar 4. 13 ROC dari Model <i>Stacking</i> pada <i>Dataset</i> Manual.....	92
Gambar 4. 14 ROC dari Model <i>Stacking</i> pada <i>Dataset</i> VADER	93
Gambar 4. 15 Tampilan Halaman Utama pada Sistem.....	96
Gambar 4. 16 Tampilan Halaman Buku pada Sistem.....	96
Gambar 4. 17 Tampilan Ulasan Pada Sistem.....	97

Gambar 4. 18 Tampilan Ulasan dan Hasil Analisisnya pada Sistem	97
Gambar 4. 19 Halaman Login pada Sistem	98
Gambar 4. 20 Halaman <i>Request</i> Buku pada Sistem	98
Gambar 4. 21 Tampilan Bagian Atas Halaman <i>Behind the Screen</i> pada Sistem ..	99
Gambar 4. 22 Tampilan Bagian Bawah Halaman <i>Behind the Screen</i> pada Sistem	99
Gambar 4. 23 Tampilan <i>Header</i> Admin pada Sistem	100
Gambar 4. 24 Tampilan Halaman <i>Upload</i> Buku pada Sistem	100
Gambar 4. 25 Tampilan Halaman <i>Upload Dataset</i> pada Sistem.....	101
Gambar 4. 26 Tampilan Halaman <i>Request</i> Buku dari Sisi Admin pada Sistem .	102
Gambar 4. 27 Tampilan Halaman Utama setelah <i>Upload</i> Buku Baru	102

DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu	11
Tabel 3. 1 Contoh ulasan buku dari <i>Platform Goodreads</i>	33
Tabel 3. 2 Variabel dan Deskripsinya	35
Tabel 3. 3 Struktur Data	35
Tabel 4. 1 Judul dan Genre Buku	46
Tabel 4. 2 Contoh Ulasan Hasil <i>Scraping</i>	46
Tabel 4. 3 Hasil Akhir <i>Dataset</i> Ulasan	49
Tabel 4. 4 <i>Dataset</i> Setelah Proses <i>Text Cleaning</i>	51
Tabel 4. 5 <i>Dataset</i> Setelah Proses <i>Case Folding</i>	52
Tabel 4. 6 <i>Dataset</i> Setelah Proses <i>Text Tokenizing</i>	54
Tabel 4. 7 <i>Dataset</i> Setelah Proses <i>Stopwords Removal</i>	55
Tabel 4. 8 <i>Dataset</i> Akhir Setelah <i>Text Preprocessing</i>	56
Tabel 4. 9 <i>Dataset</i> Hasil Labeling Manual	59
Tabel 4. 10 <i>Dataset</i> Hasil Labeling Otomatis	61
Tabel 4. 11 Rasio Stratification pada Data <i>training</i> 80%	63
Tabel 4. 12 Jumlah Data setiap Rasio Split	64
Tabel 4. 13 Data Sebelum dan Sesudah <i>Preprocessing</i>	65
Tabel 4. 14 Indeks dan Nilai BM25 Kata	67
Tabel 4. 15 Daftar Kata dengan Kontribusi Terbesar	70
Tabel 4. 16 Lima Data yang Digunakan Sebagai Contoh	71
Tabel 4. 17 Nilai BM25 dari 10 Kata pada Setiap Baris Data yang Digunakan...	73
Tabel 4. 18 Nilai <i>Likelihood</i> Setiap Fitur Dalam Setiap Kelas	74
Tabel 4. 19 Hasil Klasifikasi Manual <i>Naive Bayes</i>	75
Tabel 4. 20 Label Data Menggunakan Angka 0 dan 1	76
Tabel 4. 21 Nilai Probabilitas Awal	76
Tabel 4. 22 Nilai BM25 Fitur ' <i>story</i> ' pada Setiap Data	77
Tabel 4. 23 Nilai Kontribusi dan Probabilitas Setiap Data dari <i>Tree</i> Pertama	78
Tabel 4. 24 Nilai gradien dan <i>Hessian</i> untuk <i>Tree</i> Kedua	79
Tabel 4. 25 Nilai BM25 Fitur ' <i>like</i> ' pada Setiap Data dalam Sampel	79

Tabel 4. 26 Klasifikasi Akhir XGBoost Manual	79
Tabel 4. 27 Probabilitas Kedua Kelas dari Kedua <i>Base Learners</i>	80
Tabel 4. 28 <i>Output Maximum Entropy</i> sebagai <i>Meta Learner</i>	82
Tabel 4. 29 <i>Classification Report</i> Seluruh Eksperimen	87
Tabel 4. 30 Selisih Nilai <i>Macro F1</i> Kedua <i>Dataset</i>	88
Tabel 4. 31 Nilai ROC AUC Model <i>Stacking Ensemble</i> pada Kedua <i>Dataset</i>	89
Tabel 4. 32 Perbandingan Hasil Model <i>Stacking</i> dan <i>Standalone</i> pada Rasio Data 80:20	90

DAFTAR LAMPIRAN

Lampiran 1. LoA publikasi artikel jurnal

Lampiran 2. Qr Code dataset, notebook, dan GUI

Lampiran 3. Hasil uji plagiasi

DAFTAR NOTASI

$score(D, Q)$	:	Fungsi BM25
w_i	:	Fungsi <i>Class Weighting</i>
N	:	Total sampel
k	:	Jumlah kelas
n_i	:	Jumlah sampel dalam kelas I
$p(C_k X)$	:	Probabilitas kondisional posterior teorema <i>Bayes</i>
Y_{nb}	:	Rumus parameter klasifikasi <i>Naïve Bayes</i>
L	:	Fungsi objektif <i>XGBoost</i>
$l(y_i, y_i)$	:	Fungsi <i>loss XGBoost</i>
$\Omega(f)$	:	Fungsi regulasi <i>XGBoost</i>
y_i	:	Rumus parameter klasifikasi <i>XGBoost</i>
K	:	Jumlah pohon keputusan
f_k	:	Fungsi pohon ke-k
x_i	:	Data ke-i
D	:	Fitur dalam data <i>training</i>
x	:	Jumlah dokumen
y	:	Kelas dalam dokumen
$p(x y)$	:	Probabilitas bersyarat <i>Maximum Entropy</i>
$Z(x)$	:	Nilai normalisasi
TP	:	<i>True Positive</i>
TN	:	<i>True Negative</i>
FP	:	<i>False Positive</i>
FN	:	<i>False Negative</i>
R_i	:	Dokumen teks dari ulasan
t_j	:	<i>Term</i> yang didapatkan dari proses <i>tokenizing</i>
V_{ij}	:	Nilai matriks <i>term</i> ke-j dalam ulasan ke-i
S	:	Sentimen