

**IMPLEMENTASI *ENSEMBLE VOTING* (*NAIVE BAYES* DAN
SVM) UNTUK DETEKSI EMAIL SPAM**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

FITROTUL FAIZAH

22.11.5021

Kepada

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

**IMPLEMENTASI *ENSEMBLE VOTING* (*NAIVE BAYES* DAN
SVM) UNTUK DETEKSI EMAIL SPAM**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

FITROTUL FAIZAH

22.11.5021

Kepada

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

HALAMAN PERSETUJUAN
SKRIPSI
IMPLEMENTASI *ENSEMBLE VOTING* (*NAIVE BAYES* DAN *SVM*)
UNTUK DETEKSI EMAIL SPAM

yang disusun dan diajukan oleh

Fitrotul Faizah

22.11.5021

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 23 Januari 2026

Dosen Pembimbing,



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

HALAMAN PENGESAHAN

SKRIPSI

IMPLEMENTASI *ENSEMBLE VOTING (NAIVE BAYES DAN SVM)* UNTUK DETEKSI EMAIL SPAM

yang disusun dan diajukan oleh

Fitrotul Faizah

22.11.5021

Telah dipertabangkan di depan Dewan Penguji
pada tanggal 23 Januari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda/Tangan

Ike Verawati, S.Kom., M.Kom.
NIK. 190302237

Rifda Faticha Alfa Aziza S.Kom., M.Kom.
NIK. 190302392

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

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 23 Januari 2026

DEKAN FAKULTAS ILMU KOMPUTER



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

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Fitrotul Faizah

NIM : 22.11.5021

Menyatakan bahwa Skripsi dengan judul berikut:

Implementasi Ensemble Voting (Naive Bayes dan SVM) untuk Deteksi Email Spam

Dosen Pembimbing : Anna Baita, 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, 23 Januari 2026

Yang Menyatakan,



Fitrotul Faizah

HALAMAN PERSEMBAHAN

Segala puji dan ucapan rasa syukur panjatkan kepada Allah SWT, atas berkat rahmat serta hidayah-Nya sehingga saya dapat menyelesaikan skripsi ini. Dengan segenap hati dan ketulusan serta rasa syukur, dan bahagia telah sampai pada titik ini, tentunya bukan suatu hal yang mudah, tetapi dengan niat, dukungan, dan juga doa dari orang-orang baik di sekitar saya, pada akhirnya skripsi ini terselesaikan dengan baik. Saya persembahkan skripsi ini untuk :

1. Kedua orang tua saya tercinta, Ayah dan Ibu, yang senantiasa memberikan doa, kasih sayang, dukungan, pengorbanan, serta motivasi yang tiada henti-hentinya demi tercapainya semua cita-cita dan keinginan anak-anaknya. Terimakasih Ayah dan Ibu sehingga saya dapat mencapai titik ini.
2. Teman-teman seperjuangan saya, yang selalu kebersamaai dalam proses skripsi ini serta memberikan semangat, dukungan, dan kebersamaan. Terimakasih untuk cerita, pelukan hangat, dan semangat yang tak pernah padam. Kehadiran dan kebersamaan kalian menjadi bagian penting dalam perjalanan saya selama di bangku perkuliahan.
3. Untuk diri saya sendiri, Fitrotul Faizah. Terimakasih sudah bertahan sejauh ini atas segala perjuangan, air mata, dan ketidakpastian perjalanan panjang ini. Dan yang paling penting, terimakasih sudah berani memilih untuk mencoba, belajar, dan kuat untuk menyelesaikan apa yang telah dimulai dengan baik.

Penulis menyadari bahwa skripsi ini masih terdapat kekurangan dan keterbatasan. Oleh karena itu, kritik dan saran yang membangun demi perbaikan di masa mendatang. Semoga karya ini dapat bermanfaat dan menjadi salah satu langkah awal untuk mencapai cita-cita serta memberikan kontribusi yang baik ke depannya.

KATA PENGANTAR

Puji syukur penulis panjatkan kehadiran Allah SWT, yang telah memberikan kesempatan dan kekuatan sehingga penulis dapat menyelesaikan skripsi ini. Sholawat serta salam penulis sampaikan kepada junjungan kita nabi besar Muhammad SAW, yang telah menuntun penulis ke jalan yang semestinya. Karya ini tidak akan terwujud tanpa dukungan dan bantuan dari berbagai individu yang telah memberikan inspirasi, petunjuk dan motivasi. Oleh karena itu, penulis ingin mengucapkan terimakasih kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M., selaku Rektor Universitas Amikom Yogyakarta.
2. Ibu Prof. Dr. Kusri, S.Kom., M.Kom., selaku Dekan Fakultas Ilmu Komputer Amikom Yogyakarta
3. Ibu Eli Pujastuti, S.Kom., M.Kom., selaku Ketua Program Studi Informatika Universitas Amikom Yogyakarta
4. Ibu Anna Baita, S.Kom., M.Kom, selaku Dosen Pembimbing yang telah memberikan bimbingan, arahan, serta motivasi selama penyusunan skripsi.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi perbaikan di masa mendatang. Semoga skripsi ini dapat memberikan manfaat.

Yogyakarta, 23 Januari 2026

Penulis

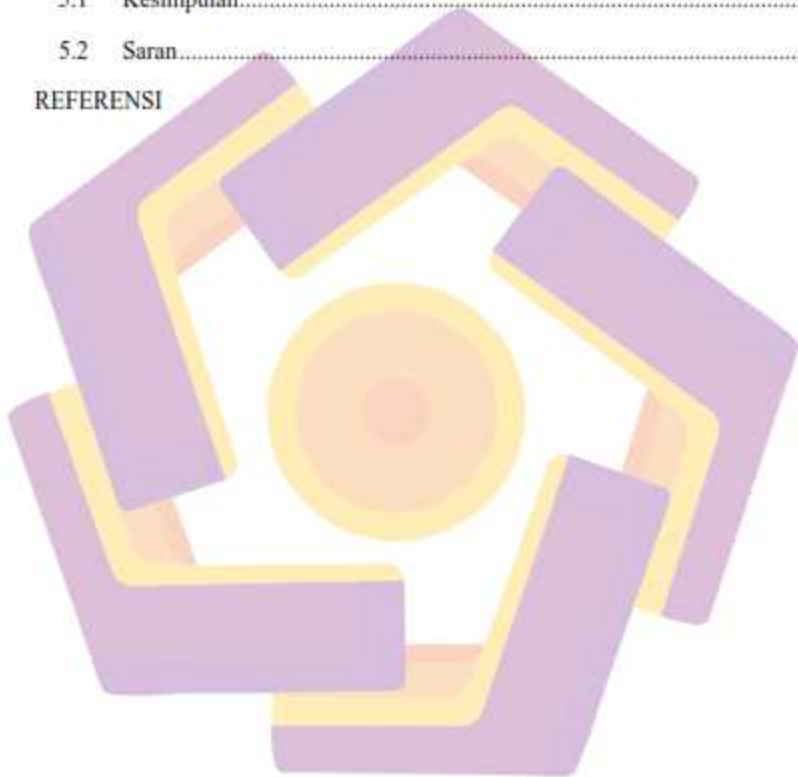
DAFTAR ISI

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 LAMBANG DAN SINGKATAN	xiii
DAFTAR ISTILAH.....	xiv
INTISARI	xv
ABSTRACT.....	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah.....	2
1.4 Tujuan Penelitian.....	2
1.5 Manfaat Penelitian.....	3
1.6 Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA	6
2.1 Studi Literatur	6
2.2 Dasar Teori.....	12

2.2.1	Email	12
2.2.2	Spam	12
2.2.3	Machine Learning	12
2.2.4	Natural Language Processing (NLP)	14
2.2.5	Preprocessing	14
2.2.6	Exploratory Data Analysis (EDA)	17
2.2.7	Feature Extraction	17
2.2.8	Naive Bayes	17
2.2.9	Support Vector Machine	20
2.2.10	Hyperparameter Tuning	25
2.2.11	Ensemble Learning	26
2.2.13	Evaluation Model	28
2.2.14	K-Fold Cross-validation	32
2.2.15	Log Loss	33
BAB III METODE PENELITIAN		34
3.1	Objek Penelitian	34
3.2	Alur Penelitian	34
3.3	Alat dan Bahan	39
3.3.1	Data Penelitian	39
BAB IV HASIL DAN PEMBAHASAN		42
4.1	Data Acquisition	42
4.2	Exploratory Data Analysis (Data Mentah)	43
4.2.1	Distribusi Kelas	43
4.2.2	Distribusi Email Berdasarkan Rentang Panjang	44
4.2.3	Distribusi Panjang Email Berdasarkan Kelas	46

4.3	<i>Data Split</i>	47
4.4	<i>Preprocessing Data</i>	48
4.4.1	Penghapusan Data Duplikat	48
4.4.2	<i>Case Folding</i>	49
4.4.3	<i>Data Cleaning</i>	50
4.4.4	<i>Tokenization</i>	51
4.4.5	Normalisasi Kata.....	53
4.4.6	<i>Stemming</i>	55
4.4.7	<i>Stopword Removal</i>	56
4.4.8	<i>Length Filtering</i>	58
4.5	<i>Exploratory Data Analisis (Data Bersih)</i>	59
4.5.1	Distribusi Kelas.....	59
4.5.2	Distribusi Email Berdasarkan Rentang Panjang.....	60
4.5.3	Distribusi Panjang Email Berdasarkan Kelas.....	61
4.5.4	<i>WordCloud</i> Email Spam dan Email Ham.....	62
4.6	<i>Feature Extraction</i>	64
4.6.1	Konfigurasi <i>TF-IDF</i>	64
4.7	<i>Modelling</i>	67
4.7.1	Pemodelan <i>Baseline</i>	68
4.7.2	Pemodelan dengan Pengaturan <i>Hyperparameter Tuning</i>	70
4.7.3	Pemodelan <i>Ensemble Voting</i>	74
4.8	<i>Evaluasi Model</i>	76
4.8.1	Kompilasi Hasil	77
4.8.2	Analisis Perbandingan Kinerja Model.....	79
4.8.3	Analisis Stabilitas.....	80

4.8.4	<i>Ensemble Voting</i> dalam Mengatasi Kelemahan Algoritma	82
4.8.5	Uji Data Eksternal.....	86
4.8.6	<i>Deployment</i>	88
BAB V PENUTUP		90
5.1	Kesimpulan.....	90
5.2	Saran.....	91
REFERENSI		92



DAFTAR TABEL

Tabel 2. 1 Keaslian Penelitian	8
Tabel 4. 1 Struktur Dataset	42
Tabel 4. 2 Distribusi Label	43
Tabel 4. 3 Rentang Panjang Email	45
Tabel 4. 4 Distribusi Label Dataset Setelah Penghapusan Data Duplikat	48
Tabel 4. 5 Contoh Hasil <i>Case Folding</i>	49
Tabel 4. 6 Contoh Hasil <i>Data Cleaning</i>	50
Tabel 4. 7 Contoh Hasil <i>Tokenization</i>	52
Tabel 4. 8 Contoh Hasil Normalisasi Kata	53
Tabel 4. 9 Contoh Sample Kamus Normalisasi	55
Tabel 4. 10 Contoh Hasil <i>Stemming</i>	55
Tabel 4. 11 Contoh Hasil <i>Stopword Removal</i>	57
Tabel 4. 12 Contoh Hasil <i>Length Filtering</i>	58
Tabel 4. 13 Konfigurasi <i>TF-IDF</i>	64
Tabel 4. 14 Hasil <i>TF-IDF</i>	65
Tabel 4. 15 Hasil Eksperimen Parameter <i>TF-IDF</i>	66
Tabel 4. 16 Hasil Pengujian Pemodelan <i>Baseline</i>	68
Tabel 4. 17 <i>Search Space</i> Pemodelan	70
Tabel 4. 18 Hasil <i>Best Parameter</i>	71
Tabel 4. 19 Hasil Pengujian Pemodelan <i>Hyperparameter Tuning</i>	71
Tabel 4. 20 Hasil Pengujian Pemodelan <i>Ensemble Baseline</i>	74
Tabel 4. 21 Hasil Pengujian Pemodelan <i>Ensemble</i> dengan Pengaturan <i>Hyperparameter Tuning</i>	75
Tabel 4. 22 Daftar Model dan Deskripsi Konfigurasi	77
Tabel 4. 23 Kompilasi Hasil Pengujian Model	78
Tabel 4. 24 Hasil Analisis Stabilitas	81
Tabel 4. 25 Hasil <i>Log Loss</i>	83
Tabel 4. 26 Pengaruh Parameter dan Waktu Komputasi	84
Tabel 4. 27 Hasil Uji Eksternal	87

DAFTAR GAMBAR

Gambar 2. 1 <i>Hard Margin SVM</i>	21
Gambar 2. 2 <i>Soft Margin SVM</i>	22
Gambar 2. 3 <i>Hard Voting</i>	27
Gambar 2. 4 <i>Soft voting</i>	28
Gambar 2. 5 <i>Confusion Matrix</i>	29
Gambar 2. 6 <i>ROC curve</i>	31
Gambar 3. 1 Alur Penelitian A	34
Gambar 3. 2 Alur Penelitian B	35
Gambar 4. 1 Distribusi Kelas Data Mentah	44
Gambar 4. 2 Distribusi Email Berdasarkan Rentang Panjang	45
Gambar 4. 3 Distribusi Panjang Email Berdasarkan Kelas	46
Gambar 4. 4 Distribusi kelas Data Bersih	59
Gambar 4. 5 Jumlah Email Berdasarkan Rentang Panjang Data Bersih	60
Gambar 4. 6 Distribusi Panjang Email Berdasarkan Kelas Data Bersih	61
Gambar 4. 7 <i>WordCloud</i> Email Spam Data Bersih	62
Gambar 4. 8 <i>WordCloud</i> Email <i>Ham</i> Data Bersih	63
Gambar 4. 9 Kurva <i>ROC Baseline</i>	69
Gambar 4. 10 Kurva <i>ROC-AUC Model Tuning 3 Fold</i>	72
Gambar 4. 11 Kurva <i>ROC-AUC Model Tuning 5 Fold</i>	73
Gambar 4. 12 Kurva <i>ROC-AUC Model Ensemble Soft voting</i>	76
Gambar 4. 13 Hasil <i>Deployment</i>	89

DAFTAR LAMBANG DAN SINGKATAN

W_{ij}	Jumlah kata ke- i yang muncul pada dokumen ke- j
D_L	Jumlah dokumen atau data
df_i	Jumlah dokumen yang mengandung kata ke- i
<i>SVM</i>	<i>Support Vector Machines</i>
<i>NB</i>	<i>Naive Bayes</i>
$P(h)$	Probabilitas hipotesis (<i>prior probability</i>)
$P(D)$	Probabilitas h berdasarkan kondisi D (<i>Posterior probability</i>)
h	Hipotesis data dengan suatu <i>class</i> tertentu
D	Data yang belum memiliki <i>class</i>
y_i	kelas data ke- i
w	nilai bobot <i>support vector</i> yang tegak lurus dengan <i>hyperplane</i>
x_j	data ke- j
b	nilai bias
k	Jumlah <i>fold</i> yang digunakan pada proses <i>K-Fold Cross-validation</i>
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
NLP	Natural Language Processing
EDA	Exploratory Data Analysis
TF-IDF	Term Frequency-Inverse Document Frequency
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
PSO	Particle Swarm Optimization
BoW	Bag of Words
RBF	Radial Basis Function
QP	Quadratic Programming

DAFTAR ISTILAH

<i>Email Spam</i>	Email tidak diinginkan yang mengandung promosi
<i>Email Ham</i>	Email normal/ bukan spam
<i>Machine Learning</i>	Cabang <i>AI</i> yang memungkinkan sistem belajar dari data
<i>Supervised Learning</i>	Metode <i>Machine Learning</i> dengan data berlabel
<i>Preprocessing</i>	Tahap pembersihan dan persiapan data
<i>Data Cleaning</i>	Membersihkan data dari kesalahan seperti duplikasi
<i>Case Folding</i>	Mengubah teks menjadi huruf kecil
<i>Tokenization</i>	Memecah teks menjadi token/kata
<i>Stemming</i>	Mengubah kata menjadi bentuk dasar
<i>Stopword Removal</i>	Menghapus kata umum yang tidak informatif
<i>Length Filtering</i>	Menghapus token terlalu pendek
<i>Feature Extraction</i>	Mengubah teks menjadi representasi numerik
<i>Ensemble Learning</i>	Penggabungan beberapa model/algoritma
<i>Hard Voting</i>	Prediksi berdasarkan suara mayoritas
<i>Soft Voting</i>	Prediksi berdasarkan rata-rata probabilitas
<i>Grid Search</i>	Metode tuning mencoba semua kombinasi parameter
<i>Cross-Validation</i>	Teknik validasi model dengan pembagian data
<i>Log Loss</i>	Metrik evaluasi berbasis probabilitas
<i>Confusion Matrik</i>	Tabel evaluasi klasifikasi
<i>Precision</i>	Ketepatan prediksi positif
<i>Recall</i>	Kemampuan mendeteksi seluruh data positif
<i>F1-Score</i>	Rata-rata harmonis <i>precision</i> dan <i>recall</i>
<i>Margin</i>	Jarak antara <i>hyperplane</i> dan data terdekat pada <i>SVM</i>
<i>Kernel Trick</i>	Teknik transformasi data pada <i>SVM</i>
<i>Hyperplane</i>	Garis pemisah untuk membedakan dua kelas data
<i>Hyperparameter Tuning</i>	Proses mencari parameter terbaik

INTISARI

Perkembangan komunikasi digital menjadikan email sebagai media utama dalam bisnis, pendidikan, dan personal. Namun, meningkatnya penggunaan email diikuti dengan maraknya spam yang mengganggu produktivitas serta menimbulkan risiko keamanan seperti penipuan, *phishing*, dan penyebaran *malware*. Oleh karena itu, diperlukan sistem deteksi email spam yang akurat dan stabil.

Algoritma *machine learning* seperti *Naive Bayes (NB)* dan *Support Vector Machine (SVM)* telah banyak digunakan untuk deteksi spam. meskipun demikian, *Naive Bayes* memiliki keterbatasan akibat asumsi independensi fitur, sedangkan *SVM* sensitif terhadap parameter dan komputasi tinggi. Penelitian ini menerapkan metode *ensemble voting* dengan mengkombinasikan *NB* dan *SVM* guna meningkatkan *precision* dan stabilitas dibandingkan algoritma tunggal. Tahapan penelitian meliputi data acquisition, *preprocessing* data menggunakan teknik *Natural Language Processing*, *feature extraction* dengan *TF-IDF*, pemodelan *NB* dan *SVM*, *hyperparameter tuning*, serta penerapan *ensemble voting* menggunakan *hard voting* dan *soft voting*. Evaluasi performa model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-Score*, serta analisis stabilitas model berdasarkan nilai standar deviasi hasil *K-Fold Cross-validation*.

Hasil penelitian menunjukkan bahwa *ensemble voting* meningkatkan *precision* dan stabilitas deteksi spam dibanding *NB* dan *SVM* tunggal. *Soft voting* mencapai *precision* 0,9712 pada *Baseline* dan 0,9729 pada *tuning* mendekati *NB tuning* 0,9825 dengan standar deviasi *F1-Score* 0,0028 hingga 0,0047. *Ensemble* juga menurunkan *Log Loss NB* hingga 33% dan meningkatkan *precision SVM* hingga 0,9889 pada *hard voting* serta 0,9747 pada *soft voting* meskipun waktu komputasi lebih tinggi, sehingga prediksi spam menjadi lebih stabil dan akurat.

Kata kunci: *Naive Bayes, Support Vector Machine, Ensemble Voting, Deteksi Spam, Machine Learning*

ABSTRACT

The rapid development of digital communication has made email a primary medium in business, education, and personal activities. However, the increasing number of email users has been followed by the rise of spam, which disrupts productivity and poses security risks such as fraud, phishing, and malware distribution.

Machine learning algorithms such as Naive Bayes (NB) and Support Vector Machine (SVM) have been widely used for spam detection. Nevertheless, Naive Bayes has limitations due to the feature independence assumption, while SVM is sensitive to parameter selection and computationally intensive. This study applies an ensemble voting method by combining NB and SVM to improve precision and stability compared to individual algorithms. The research process includes data acquisition, data preprocessing using Natural Language Processing techniques, feature extraction with TF-IDF, modeling with NB and SVM, hyperparameter tuning, and the application of ensemble voting using hard voting and soft voting. Model performance was evaluated using accuracy, precision, recall, and F1-Score metrics, along with stability analysis based on the standard deviation from K-Fold Cross-Validation.

The results show that ensemble voting improves both precision and stability in spam detection compared to NB and SVM individually. Soft voting achieved a precision of 0.9712 in the Baseline scenario and 0.9729 after tuning, approaching NB tuning at 0.9825, with F1-Score standard deviation ranging from 0.0028 to 0.0047. The ensemble also reduced NB Log Loss by up to 33% and increased SVM precision to 0.9889 in hard voting and 0.9747 in soft voting, despite higher computational time, resulting in more stable and accurate spam predictions.

Keyword: *Naive Bayes, Support Vector Machine, Ensemble Voting, Spam Detection, Machine Learning*