

**ANALISIS KLASIFIKASI EMAIL SPAM
MENGUNAKAN ALGORITMA NAÏVE BAYES**

SKRIPSI



**Oleh
Azan Rahman
210210102**

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK DAN KOMPUTER
UNIVERSITAS PUTERA BATAM
TAHUN 2025**

ANALISIS KLASIFIKASI EMAIL SPAM MENGUNAKAN ALGORITMA NAÏVE BAYES

SKRIPSI

**Untuk memenuhi salah satu syarat
Memperoleh gelar sarjana**



**Oleh
Azan Rahman
210210102**

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK DAN KOMPUTER
UNIVERSITAS PUTERA BATAM
TAHUN 2025**

SURAT PERNYATAAN ORISINALITAS

Yang bertanda tangan di bawah ini saya:

Nama : Azan Rahman

NPM : 210210102

Fakultas : Teknik dan Komputer

Program studi : Teknik Informatika

Menyatakan bahwa “**SKRIPSI**” yang saya buat dengan judul:

ANALISIS KLASIFIKASI SPAM EMAIL MENGGUNAKAN ALGORITMA
NAÏVE BAYES

Adalah hasil karya sendiri dan bukan “duplikasi” dari karya orang lain. Sepengetahuan saya, di dalam naskah Skripsi ini tidak terdapat karya ilmiah atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis dikutip didalam naskah ini dan disebutkan dalam sumber kutipan dan daftar pustaka.

Apabila ternyata di dalam naskah Skripsi ini dapat dibuktikan terdapat unsur-unsur Plagiasi, saya bersedia naskah Skripsi ini digugurkan dan gelar akademik yang saya peroleh dibatalkan, serta diproses sesuai dengan peraturan perundang-undangan yang berlaku. Demikian pernyataan ini saya buat dengan sebenarnya tanpa ada paksaan dari siapapun

Batam, 7 Februari 2025


2FF7CAKX459332959

Azan Rahman

**ANALISIS KLASIFIKASI SPAM EMAIL MENGGUNAKAN
ALGORITMA NAÏVE BAYES**

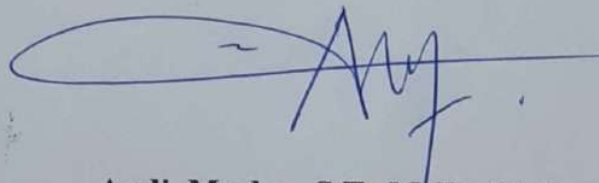
SKRIPSI

**Untuk memenuhi salah satu syarat
Memperoleh gelar Sarjana**

**Oleh
Azan Rahman
210210102**

**Telah disetujui oleh Pembimbing pada tanggal
seperti tertera dibawah ini**

Batam, 7 Februari 2025



**Andi Maslan, S.T., M.SI., Ph.D.
Pembimbing**

ABSTRAK

Perkembangan teknologi informasi dan komunikasi telah membawa perubahan signifikan dalam cara manusia berkomunikasi, termasuk dalam penggunaan surat elektronik (email). Namun, meningkatnya penggunaan email juga diikuti oleh maraknya penyebaran email spam, yang dapat mengganggu pengguna dan menimbulkan risiko keamanan. Penelitian ini bertujuan untuk menganalisis klasifikasi email spam dengan menerapkan algoritma Naïve Bayes. Proses penelitian diawali dengan tahap *preprocessing* data guna meningkatkan kualitas teks sebelum dilakukan klasifikasi dimulai dari tokenisasi, lemmatisasi, penghapusan kata tidak penting (stopword), serta konversi teks menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) agar dapat direpresentasikan dalam bentuk numerik. Selanjutnya, algoritma Naïve Bayes digunakan untuk mengklasifikasikan email ke dalam kategori spam atau non-spam. Kinerja model dianalisis menggunakan confusion matrix serta sejumlah metrik evaluasi, termasuk akurasi, presisi, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa model awal memperoleh akurasi sebesar 88%, namun memiliki nilai *False Negative* yang cukup tinggi akibat ketidakseimbangan kelas pada dataset, di mana jumlah email non-spam lebih banyak dibandingkan email spam. Untuk mengatasi permasalahan ini, diterapkan teknik penyeimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Setelah penerapan teknik ini, model mengalami peningkatan performa dengan akurasi mencapai 98% serta penurunan signifikan pada nilai *False Negative*. Hasil penelitian ini menunjukkan bahwa algoritma Naïve Bayes memiliki kinerja yang baik dalam mengklasifikasikan email spam setelah dilakukan penyesuaian terhadap ketidakseimbangan data.

Kata Kunci: *Naïve Bayes, klasifikasi spam, email, preprocessing, evaluasi model*

ABSTRACT

The advancement of information and communication technology has significantly transformed the way people communicate, including the use of electronic mail (email). However, the increasing use of email has also led to the widespread distribution of spam emails, which can disrupt users and pose security risks. This study aims to analyze spam email classification by implementing the Naïve Bayes algorithm. The research process begins with data preprocessing to enhance text quality before classification. This includes tokenization, lemmatization, stopword removal, and text conversion using Term Frequency-Inverse Document Frequency (TF-IDF) to represent text in numerical form. Subsequently, the Naïve Bayes algorithm is utilized to classify emails into spam or non-spam categories. The model's performance is evaluated using a confusion matrix and several evaluation metrics, including accuracy, precision, recall, and F1-score. Experimental results indicate that the initial model achieved an accuracy of 88%, but exhibited a high False Negative rate due to class imbalance in the dataset, where the number of non-spam emails exceeded spam emails. To address this issue, a class balancing technique using the Synthetic Minority Over-sampling Technique (SMOTE) was applied. After implementing this technique, the model demonstrated improved performance, achieving an accuracy of 98% and a significant reduction in False Negative values. The findings of this study indicate that the Naïve Bayes algorithm performs well in classifying spam emails after addressing data imbalance issues.

Keywords: *Naïve Bayes, spam classification, email, preprocessing, model evaluation*

KATA PENGANTAR

Segala puji bagi Allah yang telah melimpahkan segala rahmat dan karuniaNya, sehingga penulis dapat menyelesaikan laporan tugas akhir yang merupakan salah satu persyaratan untuk menyelesaikan program studi strata satu (S1) pada Program Studi Teknik Informatika Universitas Putera Batam.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Karena itu, kritik dan saran akan senantiasa penulis terima dengan senang hati. Dengan segala keterbatasan, penulis menyadari pula bahwa skripsi ini takkan terwujud tanpa bantuan, bimbingan, dan dorongan dari berbagai pihak. Untuk itu, dengan segala kerendahan hati, penulis menyampaikan ucapan terima kasih kepada:

1. Rektor Universitas Putera Batam;
2. Dekan Fakultas Teknik dan Komputer;
3. Ketua Program Studi Teknik Informatika, Andi Maslan S.T., M.SI., Ph.D.;
4. Bapak Andi Maslan. S.T., M.SI.,Ph.D. selaku pembimbing skripsi pada Program Studi Teknik Informatika Universitas Putera Batam;
5. Dosen dan Staff Universitas Putera Batam;

Semoga Allah membalas kebaikan dan selalu mencurahkan hidayah serta taufik-Nya, Amin.

Batam, 7 Februari 2025



Penulis (Azan rahman)

DAFTAR ISI

SURAT PERNYATAAN ORISINALITAS	i
ABSTRAK	iii
ABSTRACT	iv
KATA PENGANTAR	v
DAFTAR ISI	vi
DAFTAR GAMBAR	viii
DAFTAR TABEL	ix
DAFTAR RUMUS	x
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Identifikasi Masalah	3
1.3 Batasan Masalah.....	3
1.4 Rumusan Masalah	4
1.5 Tujuan Penelitian.....	4
1.6 Manfaat Penelitian.....	4
1.6.1 Manfaat teoritis	4
1.6.2 Manfaat praktis.....	5
BAB II TINJAUAN PUSTAKA	6
2.1 Teori Dasar	6
2.1.1 Keamanan Jaringan	6
2.1.2 Email	7
2.1.3 Spam.....	8
2.1.4 Machine Learning	8
2.1.5 Algoritma machine learning.....	9
2.2 Software Pendukung	15
2.2.1 Python.....	15
2.2.2 Pycharm.....	16
2.2.3 Jupyter Notebook	17
2.3 Penelitian Terdahulu.....	18
2.4 Kerangka Pemikiran	20

BAB III METODE PENELITIAN	22
3.1 Desain Penelitian.....	22
3.2 Metode pengumpulan data	23
3.3 Metode Perancangan	24
3.3.1 Preprocessing data.....	25
3.3.2 TF-IDF	27
3.3.3 Training dan Testing.....	30
3.3.4 Naïve Bayes.....	30
3.3.5 Evaluasi Model.....	32
3.4 Lokasi dan jadwal.....	32
BAB IV HASIL DAN PEMBAHASAN	34
4.1 Hasil Penelitian	34
4.1.1 Dataset email.....	34
4.1.2 Hasil preprocessing	35
4.1.3 TF-IDF	38
4.1.4 Traning dan testing.....	40
4.1.5 Algoritma naïve bayes.....	41
4.1.6 Hasil Evaluasi.....	43
4.2 Pembahasan	47
4.2.1 Analisis Hasil	47
4.2.2 Pengujian.....	48
BAB V KESIMPULAN DAN SARAN	50
5.1 Kesimpulan.....	50
5.2 Saran.....	51
DAFTAR PUSTAKA.....	52

DAFTAR GAMBAR

Gambar 2. 1 Kerangka Pemikiran	20
Gambar 3. 1 Desain Penelitian	22
Gambar 3. 2 Metode Perancangan.....	25
Gambar 3. 3 Cara Kerja Algoritma Naïve Bayes	31
Gambar 4. 1 perbedaan distribusi data sebelum dan sesudah smote	46

DAFTAR TABEL

Tabel 2. 1 Tabel Confusion Matrix	18
Tabel 2. 2 Penelitian Terdahulu	13
Tabel 3. 1 Contoh Teks Email.....	28
Tabel 3. 2 Perhitungan TF-IDF.....	29
Tabel 3. 4 Lokasi dan Jadwal Penelitian.....	33
Tabel 4. 1 Dataset Email.....	34
Tabel 4. 2 Hasil Tokenisasi	36
Tabel 4. 3 Hasil Lemmatization.....	37
Tabel 4. 4 Hasil Stopword Removal.....	38
Tabel 4. 5 Contoh Email	38
Tabel 4. 6 Hasil TF-IDF	39
Tabel 4. 7 Perbandingan Training dan Testing	40
Tabel 4. 8 Dataset Email.....	41
Tabel 4. 9 Perhitungan Naïve Bayes.....	42
Tabel 4. 10 Tabel Confusion Matrix	43
Tabel 4. 11 Tabel Confusion Matrix	45
Tabel 4. 12 Hasil Pengujian.....	48

DAFTAR RUMUS

Rumus 1	12
Rumus 2	12
Rumus 3	12
Rumus 4	13
Rumus 5	14
Rumus 6	14
Rumus 7	14
Rumus 8	15
Rumus 1	128
Rumus 2	128
Rumus 3	128