

**KLASIFIKASI SHORT TEXT MESSAGE SPAM MENGGUNAKAN  
ALGORITMA NAÏVE BAYES CLASSIFIER**

**SKRIPSI**



**Disusun Oleh:**

**JANNIO ALVARES**

**18.11.2373**

**PROGRAM SARJANA  
PROGRAM STUDI INFORMATIKA  
FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA  
2022**

**KLASIFIKASI SHORT TEXT MESSAGE SPAM MENGGUNAKAN  
ALGORITMA NAÏVE BAYES CLASSIFIER**

**SKRIPSI**

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika



diajukan oleh

**JANNIO ALVARES**

**18.11.2373**

Kepada

**PROGRAM SARJANA  
PROGRAM STUDI INFORMATIKA  
FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA  
2022**

## HALAMAN PERSETUJUAN

### SKRIPSI

#### KLASIFIKASI SHORT TEXT MESSAGE SPAM MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER

yang disusun dan diajukan oleh

**Jannio Alvares**

**18.11.2373**

telah disetujui oleh Dosen Pembimbing Skripsi  
pada tanggal 29 September 2021

Dosen Pembimbing,

**Uvoek Anggoro Saputro, M.Kom**

**NIK. 190302419**

# HALAMAN PENGESAHAN

## SKRIPSI

### KLASIFIKASI SHORT TEXT MESSAGE SPAM MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER

yang disusun dan diajukan oleh

**Jannio Alvares**

**18.11.2373**

Telah dipertahankan di depan Dewan Penguji  
pada tanggal 18 Agustus 2022

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Wahyu Sukestyastama Putra, S.T., M.Eng  
NIK. 190302328

Yoga Pristyanto, S.Kom, M.Eng  
NIK. 190302412

Uvoek Anggoro Saputro, M.Kom  
NIK. 190302419

Skripsi ini telah diterima sebagai salah satu persyaratan  
untuk memperoleh gelar Sarjana Komputer  
Tanggal 25 Agustus 2022

DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta, S.Kom., M.Kom.  
NIK. 190302096

## HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Jannio Alvares

NIM : 18.11.2373

Menyatakan bahwa Skripsi dengan judul berikut:

### KLASIFIKASI SHORT TEXT MESSAGE SPAM MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER

Dosen Pembimbing : Uyock Anggoro Saputro, 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, 18 Agustus 2022

Yang Menyatakan,



Jannio Alvares

## KATA PENGANTAR

Puji dan syukur kepada Tuhan Yang Maha Esa atas rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi dengan judul “Klasifikasi Short Text Message Spam Menggunakan Algoritma Naïve Bayes Classifier” sebagai syarat untuk memperoleh gelar sarjana komputer dari program studi Informatika Universitas Amikom Yogyakarta.

Selama proses pengerjaan penelitian serta penulisan tugas akhir ini, penulis mendapatkan dukungan, bantuan, bimbingan dan nasehat dari berbagai pihak sehingga pada kesempatan ini penulis menyampaikan terima kasih kepada:

1. Bapak Uyoek Anggoro Saputro, M.Kom selaku dosen pembimbing skripsi yang telah memberikan bimbingan dan masukan kepada penulis selama menyelesaikan skripsi.
2. Seluruh dosen Informatika Universitas Amikom Yogyakarta yang telah mendidik dan memberikan ilmu kepada penulis.
3. Orang tua secara khusus ibu dan nenek serta seluruh keluarga yang ada di Buntok yang telah memberikan dukungan dan doa selama pengerjaan skripsi.
4. Sahabat dan teman dekat yang selalu memberikan saran serta membantu dalam proses pengerjaan skripsi.
5. Semua pihak yang tidak bisa disebutkan satu per satu yang telah membantu penulis dalam mengerjakan skripsi ini

Penulis menyadari bahwa penulisan tugas akhir ini masih memiliki banyak kekurangan. Untuk itu penulis sangat membutuhkan kritik dan saran untuk perbaikan dimasa yang akan datang. Semoga penulisan tugas akhir ini dapat bermanfaat bagi semua pihak.

Yogyakarta, 18 Agustus 2022

Penulis



Jannio Alvares




## DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PERSETUJUAN.....	iii
HALAMAN PENGESAHAN.....	iv
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	v
KATA PENGANTAR.....	vi
DAFTAR ISI.....	vii
DAFTAR GAMBAR.....	x
DAFTAR TABEL.....	xii
INTISARI.....	xiv
<i>ABSTRACT</i> .....	xv
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.....	3
BAB II TINJAUAN PUSTAKA.....	4
2.1 Studi Literatur.....	4
2.2 Dasar Teori.....	7
2.2.1 <i>Short Message Service</i> .....	7
2.2.2 Spam.....	7
2.2.3 <i>Text Mining</i> .....	7
2.2.4 <i>Natural Language Processing</i> .....	7

2.2.5	<i>Term Frequency-Inverse Document Frequency</i> .....	13
2.2.6	<i>Machine Learning</i> .....	14
2.2.7	Klasifikasi .....	15
2.2.8	<i>Naïve Bayes Classifier</i> .....	16
2.2.9	Cross Validation.....	18
2.2.10	Confusion Matrix .....	19
<b>BAB III METODE PENELITIAN</b> .....		22
3.1	Gambaran Umum Penelitian .....	22
3.1.1	Input Data.....	23
3.1.2	Preprocessing .....	24
3.1.3	Feature Extraction .....	29
3.1.4	Pembagian Data .....	31
3.1.5	Pemodelan Naïve Bayes.....	32
3.1.6	Pengukuran Performa.....	33
3.2	Alat dan Bahan Penelitian .....	35
3.2.1	Data Penelitian .....	35
3.2.2	Alat Penelitian .....	36
3.3	Perancangan Sistem.....	36
3.4	Rancangan Pengujian .....	38
<b>BAB IV HASIL DAN PEMBAHASAN</b> .....		40
4.1	Implementasi <i>Preprocessing</i> .....	40
4.1.1	Cleaning .....	40
4.1.2	<i>Case Folding</i> .....	41
4.1.3	<i>Tokenizing</i> .....	42
4.1.4	<i>Normalization</i> .....	43



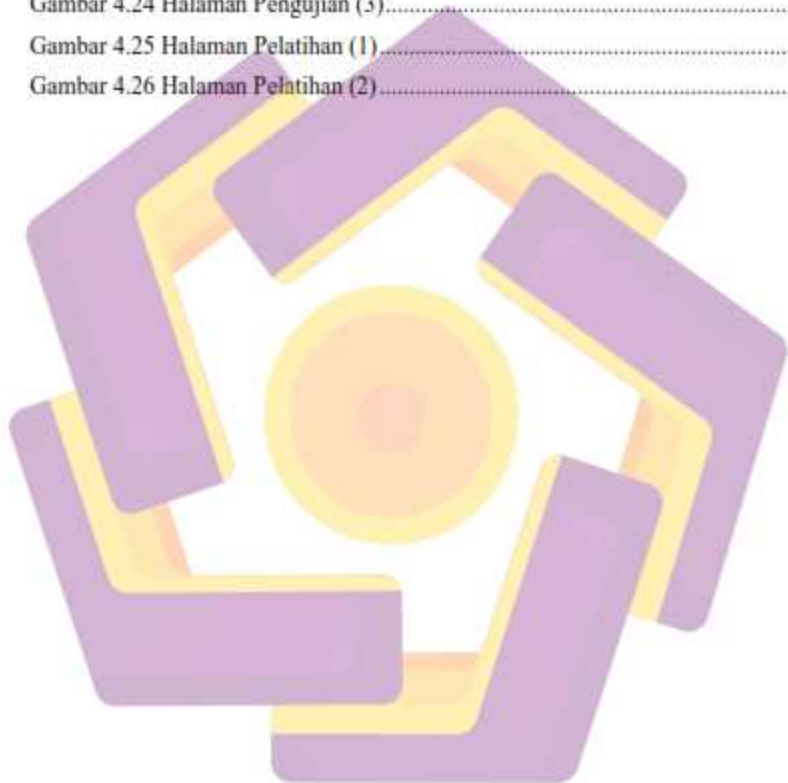


4.1.5	<i>Stopword Removal</i> .....	44
4.1.6	<i>Stemming</i> .....	46
4.2	Implementasi Feature Extraction .....	47
4.3	Implementasi Pembagian Data .....	48
4.4	Implementasi Naïve Bayes .....	48
4.5	Implementasi <i>Confusion Matrix</i> .....	49
4.6	Implementasi Pengujian .....	50
4.6.1	Multinomial Naïve Bayes .....	50
4.6.2	Bernoulli Naïve Bayes .....	54
4.6.3	Gaussian Naïve Bayes .....	57
4.7	Implementasi Sistem .....	61
4.7.1	Halaman Pengujian .....	61
4.7.1	Halaman Pelatihan .....	62
BAB V PENUTUP .....		64
5.1	Kesimpulan .....	64
5.2	Saran .....	64
DAFTAR PUSTAKA .....		65

## DAFTAR GAMBAR

Gambar 2.1 <i>Confusion Matrix Multiclass</i> .....	19
Gambar 3.1 Gambaran Umum Penelitian.....	23
Gambar 3.2 Tahapan <i>Preprocessing</i> .....	25
Gambar 3.3 Tahap <i>Feature Extraction</i> .....	30
Gambar 3.4 Tahap <i>Data Split</i> .....	31
Gambar 3.5 Pemodelan <i>Naïve Bayes Classifier</i> dengan <i>K-Fold</i> .....	33
Gambar 3.6 <i>Confusion Matrix</i> Hasil Pemodelan.....	34
Gambar 3.7 Halaman Pengujian.....	37
Gambar 3.8 Halaman pelatihan sebelum proses pelatihan.....	37
Gambar 3.9 Halaman pelatihan sesudah proses pelatihan.....	38
Gambar 4.1 <i>Source Code Cleaning</i> .....	40
Gambar 4.2 <i>Source Code Case Folding</i> .....	41
Gambar 4.3 <i>Source Code Tokenizing</i> .....	42
Gambar 4.4 <i>Source Code Normalization</i> .....	43
Gambar 4.5 <i>Source Code Stopword Removal</i> .....	45
Gambar 4.6 <i>Source Code Stemming</i> .....	46
Gambar 4.7 <i>Source Code Feature Extraction</i> .....	48
Gambar 4.8 <i>Source Code</i> Pembagian Data.....	48
Gambar 4.9 <i>Source Code Naïve Bayes</i> .....	49
Gambar 4.10 <i>Source Code Confusion Matrix</i> .....	49
Gambar 4.11 Performa Multinomial NB tanpa <i>k-fold</i> .....	50
Gambar 4.12 Performa Multinomial NB <i>fold 3</i> .....	50
Gambar 4.13 Performa Multinomial NB <i>fold 5</i> .....	51
Gambar 4.14 Performa Multinomial NB <i>fold 7</i> .....	51
Gambar 4.15 Performa Multinomial NB <i>fold 10</i> .....	52
Gambar 4.16 <i>Confusion Matrix</i> Multinomial NB tanpa <i>k-fold</i> .....	52
Gambar 4.17 <i>Confusion Matrix</i> Multinomial NB <i>fold 3</i> .....	52
Gambar 4.18 Performa Bernoulli NB.....	54

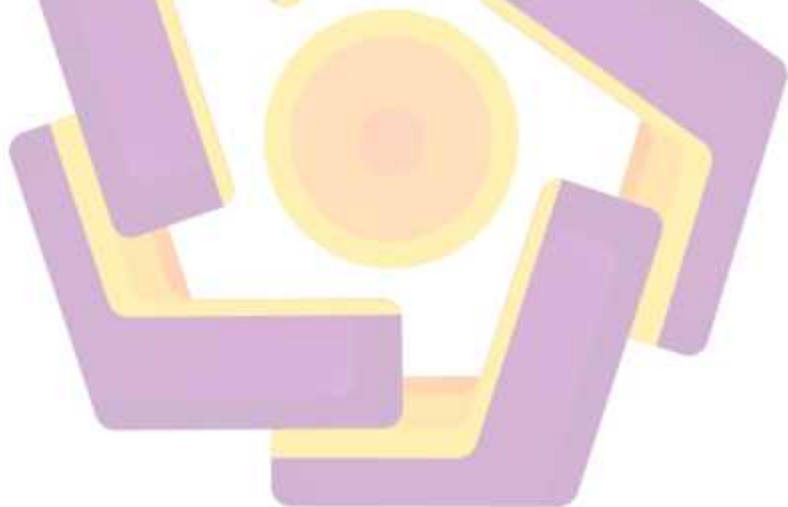
Gambar 4.19 Performa Bernoulli NB <i>fold 3</i> .....	56
Gambar 4.20 Performa Gaussian NB.....	58
Gambar 4.21 Performa Gaussian NB <i>fold 3</i> .....	59
Gambar 4.22 Halaman Pengujian (1).....	61
Gambar 4.23 Halaman Pengujian (2).....	62
Gambar 4.24 Halaman Pengujian (3).....	62
Gambar 4.25 Halaman Pelatihan (1).....	63
Gambar 4.26 Halaman Pelatihan (2).....	63



## DAFTAR TABEL

Tabel 2.1 Perbandingan Dengan Penelitian Lain .....	5
Tabel 2.2 Contoh <i>Cleaning</i> .....	8
Tabel 2.3 Contoh <i>Case Folding</i> .....	9
Tabel 2.4 Contoh <i>Tokenizing</i> .....	9
Tabel 2.5 Contoh <i>Normalization</i> .....	10
Tabel 2.6 Contoh <i>Stopword Removal</i> .....	10
Tabel 2.7 Aturan Pemenggalan Kata .....	11
Tabel 2.8 Contoh <i>Stemming</i> .....	13
Tabel 2.9 <i>Cross Validation</i> .....	18
Tabel 3.1 Contoh <i>Input</i> Tahap Pengujian .....	24
Tabel 3.2 Contoh <i>Input</i> Tahap Pelatihan .....	24
Tabel 3.3 Tahap <i>Cleaning</i> .....	25
Tabel 3.4 Tahap <i>Case Folding</i> .....	26
Tabel 3.5 <i>Tokenizing</i> .....	26
Tabel 3.6 <i>Normalization</i> .....	27
Tabel 3.7 Kamus Normalisasi .....	27
Tabel 3.8 <i>Stopword Removal</i> .....	28
Tabel 3.9 <i>Stemming</i> .....	29
Tabel 3.10 Data Hasil Preprocessing .....	30
Tabel 3.11 Hasil <i>Term Frequency</i> .....	30
Tabel 3.12 Hasil <i>TF-IDF</i> .....	31
Tabel 3.13 Data <i>Train</i> .....	31
Tabel 3.14 Data <i>Test</i> .....	32
Tabel 3.15 Jumlah TP, TN, FP, dan FN setiap kelas .....	34
Tabel 3.16 Rincian Dataset .....	35
Tabel 3.17 Spesifikasi Hardware .....	36
Tabel 3.18 Spesifikasi <i>Software</i> .....	36

Tabel 4.1 Hasil <i>Cleaning</i> .....	40
Tabel 4.2 Tahap <i>Case Folding</i> .....	42
Tabel 4.3 Hasil <i>Tokenizing</i> .....	43
Tabel 4.4 Hasil <i>Normalization</i> .....	44
Tabel 4.5 Hasil <i>Stopword Removal</i> .....	45
Tabel 4.6 Hasil <i>Stemming</i> .....	47
Tabel 4.7 Hasil Pengujian Prediksi Multinomial NB <i>fold 3</i> .....	53
Tabel 4.8 Pengujian Prediksi Multinomial NB.....	54
Tabel 4.9 Pengujian Prediksi Bernoulli NB.....	54
Tabel 4.10 Pengujian Prediksi Bernoulli NB <i>fold 3</i> .....	56
Tabel 4.11 Pengujian Prediksi Gaussian NB.....	58
Tabel 4.12 Pengujian Prediksi Gaussian NB <i>fold 3</i> .....	60



## INTISARI

Short Message Service atau yang lebih dikenal dengan SMS merupakan salah satu komunikasi dengan media teks melalui perangkat telepon seluler atau smartphone. Saat ini perkembangan teknologi informasi memungkinkan pengaksesan data lebih praktis, cepat, dan efisien, oleh karena itu dapat dimanfaatkan sebagian orang untuk melakukan spam untuk melakukan promosi terhadap suatu toko atau jasa. Spam adalah singkatan dari Sending and Posting Advertisement in Mass, yang artinya mengirim pesan secara massal. Selain untuk promosi ada juga yang menyalahgunakannya untuk melakukan tindakan kejahatan seperti penipuan berhadiah. Sebagian orang mungkin dapat mengenali apakah SMS tersebut berupa penipuan atau tidak, namun bagi orang awam yang baru saja mendapat SMS seperti itu tentu kebingungan dan mungkin akan tergiur. Oleh karena itu, tujuan penelitian ini adalah untuk mengklasifikasi jenis-jenis SMS apakah termasuk pesan normal, penipuan, atau promo. Metode yang digunakan dalam penelitian ini adalah Naïve Bayes Classifier. Sebelum data digunakan dilakukan proses preprocessing terlebih dahulu, kemudian dilakukan proses feature extraction menggunakan TF-IDF. Data yang sudah diproses akan dibagi menjadi dua bagian dan akan dilakukan proses cross validation untuk menghasilkan kinerja yang baik. Berdasarkan hasil yang diperoleh, Multinomial Naïve Bayes yang menggunakan proses cross validation mendapat akurasi mencapai 97,7%.

**Kata Kunci:** Klasifikasi, SMS, Spam, Naive Bayes, TF-IDF



## **ABSTRACT**

*Short Message Service or better known as SMS is a form of communication with text media via mobile phones or smartphones. Currently the development of information technology allows access to data more practical, fast, and efficient, therefore some people can use it to do spam to promote a store or service. Spam stands for Sending and Posting Advertisement in Mass, which means sending messages in bulk. In addition to promotion, there are also those who misuse it to commit crimes such as fraud with prizes. Some people may be able to recognize whether the SMS is a fraud or not, but for ordinary people who have just received such an SMS, they are certainly confused and may be tempted. Therefore, the purpose of this study is to classify the types of SMS whether they are normal, fraudulent, or promo messages. The method used in this research is the Naïve Bayes Classifier. Before the data is used, the preprocessing process is carried out first, then the feature extraction process is carried out using TF-IDF. The processed data will be divided into two parts and a cross validation process will be carried out to produce good performance. Based on the results obtained, Multinomial Nave Bayes which uses the cross validation process has an accuracy of 97,7%.*

**Keyword:** Classification, SMS, Spam, Naive Bayes, TF-IDF