

KLASIFIKASI SPAM PADA EMAIL BERBAHASA INDONESIA MENGUNAKAN FASTTEXT DAN BERNOULLI NAÏVE BAYES

*Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 di Jurusan Teknik Informatika*



Oleh:

Zatun Aulia Putri
NIM: 09021182025029

**JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2025**

KLASIFIKASI SPAM PADA EMAIL BERBAHASA INDONESIA MENGUNAKAN FASTTEXT DAN BERNOULLI NAÏVE BAYES

*Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 di Jurusan Teknik Informatika*



Oleh:

Zatun Aulia Putri
NIM: 09021182025029

**JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2025**

HALAMAN PENGESAHAN

SKRIPSI

Klasifikasi Spam Pada Email Berbahasa Indonesia Menggunakan FastText dan Bernoulli Naïve Bayes

Sebagai salah satu syarat untuk penyelesaian studi di
Program Studi S1 Teknik Informatika

Oleh:

ZATUN AULIA PUTRI

09021182025029

Pembimbing 1	: <u>Rizki Kurniati, M.T</u> NIP. 199107122019032016
Pembimbing 2	: <u>Muhammad Naufal Rachmatullah, M.T</u> NIP. 199212012022031008

Mengetahui

Ketua Jurusan Teknik Informatika



Hadipurnawan Satria, Ph.D
198004182020121001

TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI

Pada hari Jum'at tanggal 25 Juli 2025 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Zatun Aulia Putri

NIM : 09021182025029

Judul : Klasifikasi Spam Pada Email Berbahasa Indonesia Menggunakan FastText dan Bernoulli Naïve Bayes

dan dinyatakan **LULUS**.

1. Ketua Penguji

Dian Palupi Rini, M.Kom., Ph.D
NIP. 197802232006042002



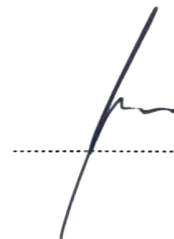
2. Penguji

Alvi Syahrini Utami, M.Kom.
NIP. 197812222006042003



3. Pembimbing I

Rizki Kurniati, M. T
NIP. 199107122019032016



4. Pembimbing II

M. Naufal Rachmatullah, M. T
NIP. 199212012022031008



Mengetahui,
Ketua Jurusan Teknik Informatika



Hadipurnawan Satria, Ph. D.
NIP. 198004182020121001

HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : Zatun Aulia Putri

NIM : 09021182025029

Program Studi : Teknik Informatika

Judul : Klasifikasi Spam Pada Email Berbahasa Indonesia Menggunakan
FastText dan Bernoulli Naïve Bayes

Hasil Pengecekan Software iThenticate/Turnitin: 6%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan skripsi ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapa pun.



Palembang, 28 Juli 2025



Zatun Aulia Putri
NIM. 09021182025029

MOTTO DAN PERSEMBAHAN

“However, this is your life. And only you can choose how to live it, with all its challenges and imperfections.”

Kupersembahkan karya tulis ini kepada:

- Orang Tua, dan Adik yang kusayangi
- Teman-teman seperjuangan TI'20
- Universitas Sriwijaya

ABSTRACT

Indonesia ranks sixth globally in terms of the number of spam senders. Numerous studies have been conducted on spam detection and filtering, with Bayesian algorithms being among the most commonly used approaches. This study aims to classify Indonesian-language email messages into spam and non-spam categories. A secondary dataset consisting of 2,604 messages was used, comprising 1,362 spam messages and 1,242 non-spam messages. Word representation was performed using FastText with an n-gram approach to capture sub-word level information, while classification was carried out using the Bernoulli Naïve Bayes algorithm based on binary values. The experiments compared the performance of the Bernoulli Naïve Bayes algorithm with and without the use of FastText. Evaluation was conducted using accuracy, confusion matrix, and classification report metrics, with a 70:30 data split. The results showed that both models, with and without FastText, achieved 95% accuracy. However, the model incorporating FastText demonstrated more balanced performance across classes and higher recall in detecting spam. In contrast, the model without FastText achieved perfect precision and recall for spam but showed decreased performance for non-spam. Therefore, the use of FastText contributes to improving the sensitivity and balance of spam email classification in the Indonesian language.

Keywords: *Bernoulli Naïve Bayes, FastText, Indonesian language, Spam Email, Text Classification.*

Supervisor I



Rizki Kurniati, M. T
NIP.199107122019032016

Supervisor II



M. Naufal Rachmatullah, M. T
NIP.199212012022031008

Approved,

Head of Informatics Department



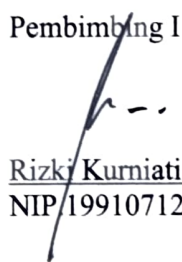
Hadipurnawan Satria, Ph.D.
NIP.198004182020121001

ABSTRAK

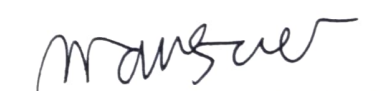
Indonesia menempati peringkat keenam dalam jumlah pengirim spam secara global. Penelitian terkait deteksi dan penyaringan spam telah banyak dilakukan, dan algoritma yang sering digunakan dalam penelitian tersebut adalah Bayesian. Penelitian ini bertujuan untuk melakukan klasifikasi pesan email dalam bahasa Indonesia ke dalam kategori spam dan *non-spam*. Dataset sekunder digunakan dalam penelitian ini berjumlah 2604 data dengan 1362 pesan spam, dan 1242 pesan *non-spam*. Representasi kata dilakukan menggunakan FastText dengan pendekatan n-gram untuk menangkap informasi sub-kata, sedangkan klasifikasi dilakukan menggunakan algoritma Bernoulli Naïve Bayes berbasis nilai biner. Pengujian dilakukan untuk membandingkan performa algoritma Bernoulli Naïve Bayes tanpa dan dengan menggunakan FastText. Evaluasi dilakukan menggunakan metrik akurasi, *confusion matrix* dan *classification report* dengan pembagian data 70:30. Hasil pengujian menunjukkan bahwa model Bernoulli Naïve Bayes dengan maupun tanpa menggunakan FastText, keduanya mencapai akurasi 95%. Namun model dengan menggunakan FastText menunjukkan keseimbangan performa antar kelas serta recall yang lebih tinggi untuk deteksi spam. Sebaliknya, model tanpa FastText memiliki precision dan recall sempurna untuk spam, tetapi performa sedikit menurun pada *non-spam*. Dengan demikian, penggunaan FastText membantu meningkatkan sensitivitas dan keseimbangan klasifikasi terhadap email berbahasa Indonesia.

Kata Kunci: Bahasa Indonesia, Bernoulli Naïve Bayes, FastText, Email Spam, Klasifikasi Teks.

Pembimbing I


Rizki Kurniati, M. T
NIP.199107122019032016

Pembimbing II


M. Naufal Rachmatullah, M. T
NIP.199212012022031008

Mengetahui,

Ketua Jurusan Teknik Informatika


Hadipurnawan Satria, Ph.D.
NIP.198004182020121001

KATA PENGANTAR

Puji syukur penulis ucapkan kepada Allah Subhanallahu wa ta'ala atas limpahan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul “Klasifikasi Spam Pada Email Berbahasa Indonesia Menggunakan FastText dan Bernoulli Naïve Bayes” dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk menyelesaikan pendidikan Strata-1 pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Sriwijaya.

Dalam proses penyusunan skripsi ini, penulis menyadari bahwa keberhasilan ini tidak terlepas dari bantuan, bimbingan dan dukungan dari berbagai pihak. Karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih kepada:

1. Orang Tua dan adik, yang telah memberikan doa dan dukungan selama masa perkuliahan dan penyusunan skripsi.
2. Bapak Hadipurnawan Satria, Ph.D. selaku Ketua Jurusan Strata-1 Teknik Informatika.
3. Ibu Rizki Kurniati, M.T. selaku Dosen Pembimbing I, dan Bapak M. Naufal Rachmatullah, M.T. selaku Dosen Pembimbing II, atas bimbingan, arahan dan masukan terkait proses penyusunan skripsi.
4. Seluruh Dosen Jurusan Teknik Informatika, yang telah memberikan ilmu selama masa perkuliahan.
5. Seluruh Staf Jurusan Teknik Informatika, Fakultas Ilmu Komputer, dan Universitas Sriwijaya, yang telah membantu dalam urusan administrasi dan akademik selama masa perkuliahan.
6. Rekan-rekan seperjuangan yang telah memberikan dukungan, baik moral maupun praktikal selama masa perkuliahan dan penyusunan skripsi.

Penulis menyadari bahwa skripsi ini masih belum sempurna, Oleh karena itu, penulis menerima kritik dan saran yang membangun dari semua pihak demi perbaikan skripsi ini di masa depan. Semoga skripsi ini dapat memberikan manfaat dan menjadi referensi bagi pengembangan penelitian selanjutnya

Palembang, 27 Juli 2025

Zatun Aulia Putri

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGESAHAN SKRIPSI	ii
TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI	iii
HALAMAN PERNYATAAN	iv
MOTTO DAN PERSEMBAHAN	v
ABSTRACT	vi
ABSTRAK	vii
KATA PENGANTAR	viii
DAFTAR ISI	ix
DAFTAR TABEL	xii
DAFTAR GAMBAR	xiv
BAB I PENDAHULUAN	I-1
1.1 Pendahuluan	I-1
1.2 Latar Belakang	I-1
1.3 Rumusan Masalah	I-3
1.4 Tujuan Penelitian	I-3
1.5 Manfaat Penelitian	I-3
1.6 Batasan Masalah	I-3
1.7 Sistematika Penulisan	I-3
1.8 Kesimpulan	I-5
BAB II KAJIAN LITERATUR	II-1
2.1 Pendahuluan	II-1
2.2 Landasan Teori	II-1
2.2.1 Email Spam	II-1
2.2.2 Bernoulli Naïve Bayes	II-2
2.2.3 FastText	II-3
2.2.4 Confusion Matrix	II-3
2.2.5 Klasifikasi	II-4
2.2.6 Pre-processing	II-5
2.2.7 Rational Unified Process	II-7
2.3 Penelitian lain yang Relevan	II-8
2.4 Kesimpulan	II-10
BAB III METODOLOGI PENELITIAN	III-1
3.1 Pendahuluan	III-1

3.2 Pengumpulan Data	III-1
3.2.1. Jenis dan Sumber Data	III-1
3.2.2. Metode Pengumpulan Data	III-3
3.3 Tahapan Penelitian	III-3
3.3.1. Menentukan Kerangka Kerja Penelitian	III-4
3.3.2. Menentukan Kriteria Pengujian	III-6
3.3.3. Menentukan Format Data Pengujian	III-6
3.3.4. Menentukan Alat Bantu Penelitian	III-7
3.3.5. Menentukan Pengujian Penelitian	III-7
3.3.6. Menentukan Analisis Penelitian	III-8
3.4 Metode Pengembangan Perangkat Lunak	III-8
3.4.1. Fase Insepsi	III-8
3.4.2. Fase Elaborasi	III-9
3.4.3. Fase Konstruksi	III-9
3.4.4. Fase Transisi	III-9
3.5 Manajemen Proyek Penelitian	III-9
3.6 Kesimpulan	III-12
BAB IV PENGEMBANGAN PERANGKAT LUNAK	IV-1
4.1 Pendahuluan	IV-1
4.2 Fase Insepsi	IV-1
4.2.1 Pemodelan Bisnis	IV-1
4.2.2 Kebutuhan Perangkat Lunak	IV-2
4.2.3 Analisis dan Desain	IV-3
4.2.4.1. Analisis Kebutuhan Perangkat Lunak	IV-3
4.2.4.2. Analisis Data	IV-3
4.2.4.3. Analisis Pra-pemrosesan	IV-3
4.2.4.4. Analisis FastText	IV-10
4.2.4.5. Analisis Bernoulli Naïve Bayes	IV-13
4.2.4 Implementasi	IV-16
4.2.4.1 Use Case	IV-16
4.2.4.2 Definisi Aktor	IV-16
4.2.4.3 Definisi Use Case	IV-17
4.2.4.4 Skenario Use Case	IV-17
4.3 Fase Elaborasi	IV-20
4.3.1 Perancangan Antarmuka	IV-20
4.3.2 Kebutuhan Sistem	IV-20
4.3.3 Activity Diagram	IV-21
4.3.4 Sequence Diagram	IV-22
4.4 Fase Konstruksi	IV-24
4.4.1 Kebutuhan Sistem	IV-24
4.4.2 Class Diagram	IV-24
4.4.3 Implementasi	IV-24
4.4.3.1. Implementasi Kelas	IV-24
4.4.3.2. Implementasi Antarmuka	IV-25
4.5 Fase Transisi	IV-26

4.5.1	Pemodelan Bisnis	IV-26
4.5.2	Rencana Pengujian	IV-26
4.5.3	Implementasi	IV-26
4.6	Kesimpulan	IV-27
BAB V HASIL DAN PEMBAHASAN		V-1
5.1	Pendahuluan	V-1
5.2	Data Hasil Penelitian	V-1
5.2.1	Konfigurasi Percobaan	V-1
5.2.2	Data Hasil Konfigurasi I	V-2
5.2.3	Data Hasil Konfigurasi II	V-3
5.2.4	Data Hasil Konfigurasi III	V-5
5.3	Analisis Hasil Penelitian	V-6
5.4	Kesimpulan	V-7
BAB VI KESIMPULAN DAN SARAN		VI-1
6.1	Pendahuluan	VI-1
6.2	Kesimpulan	VI-1
6.3	Saran	VI-2
DAFTAR PUSTAKA		xv

DAFTAR TABEL

Tabel II-1. <i>Confussion Matrix</i>	II-4
Tabel II-2. Contoh kalimat <i>Cleaning</i>	II-5
Tabel II-3. Contoh kalimat <i>Case Folding</i>	II-6
Tabel II-4. Contoh kalimat <i>Tokenizing</i>	II-6
Tabel II-5. Contoh kalimat <i>Stopword Removal</i>	II-6
Tabel II-6. Contoh kalimat <i>Stemming</i>	II-7
Tabel III-1. Contoh Isi Dataset.....	III-1
Tabel III-2. Rancangan Tabel <i>Confusion Matrix</i>	III-6
Tabel III-3. Tabel Hasil Analisis Klasifikasi.....	III-8
Tabel III-4. Rencana Aktivitas Penelitian.....	III-10
Tabel IV-1. Kebutuhan Fungsional dan Non-Fungsional.....	IV-2
Tabel IV-2. Contoh Teks.....	IV-4
Tabel IV-3. <i>Cleaning</i>	IV-5
Tabel IV-4. <i>Case Folding</i>	IV-6
Tabel IV-5. <i>Tokenizing</i>	IV-7
Tabel IV-6. <i>Stopword Removal</i>	IV-8
Tabel IV-7. <i>Stemming</i>	IV-9
Tabel IV-8. Analisis perhitungan FastText.....	IV-10
Tabel IV-9. Binarisasi vektor kata “dapat”.....	IV-12
Tabel IV-10. Tabel Fitur Dokumen.....	IV-13
Tabel IV-11. Nilai Probabilitas Kondisional.....	IV-14
Tabel IV-12. Contoh Dokumen Uji.....	IV-15
Tabel IV-13. Tabel Definisi Aktor.....	IV-16
Tabel IV-14. Tabel Definisi Use Case.....	IV-17
Tabel IV-15. Skenario Memasukkan file.....	IV-17
Tabel IV-16. Skenario Melakukan Klasifikasi.....	IV-18
Tabel IV-17. Skenario Preprocessing.....	IV-19

Tabel IV-18. Implementasi Kelas	IV-24
Tabel IV-19. Rencana Pengujian Memasukkan File	IV-26
Tabel IV-20. Rencana Pengujian Melakukan Klasifikasi	IV-26
Tabel IV-21. Implementasi Pengujian Memasukkan File	IV-27
Tabel IV-22. Implementasi Pengujian Melakukan Klasifikasi	IV-27
Tabel V-1. Data Hasil Skenario Pembagian Data	V-2
Tabel V-2. Data Hasil Evaluasi Skenario Pembagian Data	V-2
Tabel V-3. Data Hasil Pengujian Performa	V-3
Tabel V-4. Data Hasil Confusion Matrix	V-3
Tabel V-5. Data Hasil Pengujian Perbandingan	V-5
Tabel V-6. Data Hasil Confusion Matrix Perbandingan	V-5

DAFTAR GAMBAR

Gambar II-1. Fase Rational Unified Process	II-8
Gambar III-1 Tahapan Penelitian	III-3
Gambar III-2. Rancangan Kerangka Kerja	III-4
Gambar IV-1. Diagram Use Case	IV-16
Gambar IV-2. Perancangan Antarmuka	IV-20
Gambar IV-3. Diagram Aktivitas Memasukkan File	IV-21
Gambar IV-4. Diagram Aktivitas Melakukan Klasifikasi	IV-21
Gambar IV-5. Diagram Aktivitas Preprocessing	IV-22
Gambar IV-6. Diagram Sekuensial Memasukkan File	IV-22
Gambar IV-7. Diagram Sekuensial Melakukan Klasifikasi	IV-23
Gambar IV-8. Diagram Sekuensial Preprocessing	IV-23
Gambar IV-9. Diagram Kelas	IV-24
Gambar IV-10. Implementasi Antarmuka	IV-25
Gambar V-1. Visualisasi Kurva ROC	V-4
Gambar V-2. Visualisasi Kurva Precision-Recall	V-4

BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada bab pendahuluan ini diuraikan mengenai latar belakang penelitian, rumusan masalah yang diteliti, penetapan tujuan penelitian, manfaat dari penelitian, ruang lingkup batasan masalah dan sistematika penulisan yang digunakan dalam penelitian.

1.2 Latar Belakang Masalah

Spam pada email merupakan pesan yang dikirim secara massal tanpa persetujuan penerima yang umumnya berisi iklan, penipuan atau konten berbahaya seperti tautan (URL) dan lampiran yang berpotensi mengandung phishing atau malware (Putra, 2008). Sejak awal 1990-an, penggunaan spam telah menjadi masalah yang terus berkembang dan menjadi tantangan bagi sebagian besar pengguna email di seluruh dunia. Berdasarkan data statistik spam terbaru dari AV-test, Indonesia menempati peringkat keenam sebagai negara dengan jumlah pengirim spam (*spammer*) terbanyak secara global¹⁾.

Berbagai penelitian telah dilakukan untuk mengembangkan metode deteksi dan penyaringan spam. Salah satu pendekatan yang paling umum digunakan adalah metode klasifikasi berbasis algoritma Bayesian. Meskipun telah banyak teknik dikembangkan dan dibandingkan, algoritma Bayesian tetap menjadi

¹⁾ <https://www.av-test.org/en/statistics/spam/>

pilihan populer karena kesederhanaan model serta efisiensi komputasinya (Vernanda et al., 2020).

Salah satu varian dari algoritma tersebut adalah Bernoulli Naïve Bayes, yang mengandalkan kehadiran kata dalam dokumen tanpa memperhitungkan frekuensinya. Algoritma ini dinilai cocok untuk klasifikasi teks, terutama saat jumlah data yang digunakan terbatas (Singh, Gurinder, et al., 2019). Penelitian sebelumnya menunjukkan algoritma Bernoulli Naïve Bayes memberikan hasil yang lebih baik dibandingkan dengan algoritma Gaussian Naïve Bayes dan Multinomial Naïve Bayes dalam melakukan klasifikasi (Kenang Candra Alivian Pratama et al., 2022).

Untuk meningkatkan akurasi klasifikasi, diperlukan pendekatan representasi kata berbasis *word embedding* (Mikolov et al., 2013). Dalam penelitian ini, digunakan model FastText yang dikembangkan oleh Facebook AI Research. FastText memiliki keunggulan dalam menangani kata-kata yang tidak muncul dalam data pelatihan (*out-of-vocabulary*) dengan memanfaatkan informasi sub-kata melalui karakter n-gram (Yao et al., 2020). Penelitian sebelumnya menunjukkan bahwa FastText mampu memberikan representasi kata yang baik, bahkan untuk bahasa dengan kompleksitas morfologi tinggi seperti bahasa Indonesia (Bojanowski et al., 2017; Nurdin et al., 2020).

Berdasarkan hal tersebut, penelitian ini menggunakan FastText sebagai representasi kata dan Bernoulli Naïve Bayes sebagai algoritma klasifikasi untuk mengkategorisasi spam pada email berbahasa Indonesia. Diharapkan kombinasi ini dapat menghasilkan performa klasifikasi yang akurat dan efisien.

1.3 Rumusan Masalah

1. Bagaimana penerapan FastText dan Bernoulli Naïve Bayes untuk klasifikasi spam pada email berbahasa Indonesia?
2. Bagaimana kinerja FastText dan Bernoulli Naïve Bayes dalam mengklasifikasi email sebagai spam atau non-spam?

1.4 Tujuan Penelitian

1. Implementasi FastText sebagai representasi kata, dan Bernoulli Naïve Bayes sebagai algoritma klasifikasi pada teks email berbahasa Indonesia.
2. Mengevaluasi performa FastText dan Bernoulli Naïve Bayes dalam klasifikasi email spam.

1.5 Manfaat Penelitian

1. Memberikan referensi tentang penerapan FastText dan Bernoulli Naïve Bayes dalam melakukan klasifikasi terhadap teks berbahasa Indonesia.
2. Menjadi salah satu metode klasifikasi spam pada email, khususnya berbahasa Indonesia.

1.6 Batasan Masalah

1. Representasi kata FastText hanya dapat menggunakan versi numpy dibawah 2.0.
2. Bernoulli Naïve Bayes hanya dapat mengolah data biner, sehingga representasi kata dari FastText yang menghasilkan vektor float diubah menjadi bentuk biner (0/1).

1.7 Sistematika Penulisan

Sistematika penulisan proposal skripsi ini adalah sebagai berikut.

BAB I. PENDAHULUAN

Pada bab I akan diuraikan mengenai latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, sistematika penulisan dan kesimpulan pada bab ini.

BAB II. KAJIAN LITERATUR

Pada bab II akan diuraikan mengenai dasar teori yang digunakan dalam penelitian, seperti definisi Bernoulli Naïve Bayes, Confussion Matrix, Email Spam, FastText dan Pre-processing.

BAB III. METODOLOGI PENELITIAN

Pada bab III akan diuraikan mengenai tahapan penelitian yang terdeskripsi secara rinci dengan mengacu pada suatu kerangka kerja. Pada akhir bab ini, akan ada perancangan manajemen proyek pada pelaksanaan penelitian.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Pada bab IV akan diuraikan mengenai tahapan metode *Rational Unified Process* yang digunakan untuk pengembangan perangkat lunak yang dibuat untuk melakukan proses klasifikasi teks terhadap spam pada email berbahasa Indonesia.

BAB V. HASIL DAN PEMBAHASAN

Pada bab V akan diuraikan mengenai hasil dan pembahasan terhadap pengembangan perangkat lunak yang dikembangkan dalam penelitian ini. *Classification report*, *confusion matrix*, kurva ROC dan kurva *Precision-Recall* digunakan untuk mengukur performa metode klasifikasi yang dikembangkan.

BAB VI. KESIMPULAN DAN SARAN

Pada bab VI akan diuraikan mengenai kesimpulan dan saran terhadap penelitian yang telah dilakukan.

1.8 Kesimpulan

Berdasarkan latar belakang yang telah diuraikan, penelitian ini bertujuan untuk melakukan analisis dan klasifikasi terhadap email spam berbahasa Indonesia dengan kategori spam dan non-spam menggunakan FastText sebagai representasi teks dan Bernoulli Naïve Bayes (BNB) sebagai algoritma klasifikasi. Diharapkan, kombinasi dari keduanya akan menghasilkan metode klasifikasi yang akurat dan efektif, serta menjadi referensi untuk meningkatkan sistem deteksi spam pada email berbahasa Indonesia.

DAFTAR PUSTAKA

- Artur, M. (2021). Review the performance of the Bernoulli Naïve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features. *Procedia Computer Science*, 190. <https://doi.org/10.1016/j.procs.2021.06.066>
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5. https://doi.org/10.1162/tacl_a_00051
- Cormack, G. v. (2006). Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4). <https://doi.org/10.1561/15000000006>
- Ha, J., Kambe, M., & Pe, J. (2011). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. <https://doi.org/10.1016/C2009-0-61819-5>
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 2. <https://doi.org/10.18653/v1/e17-2068>
- Kenang Candra Alivian Pratama, H., Suharso, W., & A'yun, Q. (2022). Pengklasifikasian Kanker Payudara Dan Kanker Paru-Paru Dengan Metode Gaussian Naïve Bayes, Multinomial Naïve Bayes, Dan Bernoulli Naïve Bayes. *Jurnal Smart Teknologi*, 3(4). <http://103.242.78.149/index.php/JST/article/view/7592>
- Khasanah, I. N. (2021). Sentiment Classification Using fastText Embedding and Deep Learning Model. *Procedia CIRP*, 189. <https://doi.org/10.1016/j.procs.2021.05.103>
- Kruchten, P. (2004). *THE RATIONAL UNIFIED PROCESS AN INTRODUCTION THIRD EDITION*.
- Larabi-Marie-Sainte, S., Ghouzali, S., Saba, T., Aburahmah, L., & Almohaini, R. (2022). Improving spam email detection using deep recurrent neural network. *Indonesian Journal of Electrical Engineering and Computer Science*, 25(3). <https://doi.org/10.11591/ijeecs.v25.i3.pp1625-1633>
- Nurdin, A., Anggo Seno Aji, B., Bustamin, A., & Abidin, Z. (2020). PERBANDINGAN KINERJA WORD EMBEDDING WORD2VEC, GLOVE, DAN FASTTEXT PADA KLASIFIKASI TEKS. *Jurnal Tekno Kompak*, 14(2). <https://doi.org/10.33365/jtk.v14i2.732>

- Putra, E. N. (2016). PENGIRIMAN E-MAIL SPAM SEBAGAI KEJAHATAN CYBER DI INDONESIA. *Jurnal Cakrawala Hukum*, 7(2). <https://doi.org/10.26905/idjch.v7i2.1906>
- Putri, N. B., & Wijayanto, A. W. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Komputika : Jurnal Sistem Komputer*, 11(1). <https://doi.org/10.34010/komputika.v11i1.4350>
- Rahman, M. D., Djunaidy, A., & Mahananto, F. (2021). Penerapan Weighted Word Embedding pada Pengklasifikasian Teks Berbasis Recurrent Neural Network untuk Layanan Pengaduan Perusahaan Transportasi. *Jurnal Sains Dan Seni ITS*, 10(1). <https://doi.org/10.12962/j23373520.v10i1.56145>
- Rahman, R., & Fauzi Abdulloh, F. (2023). Performance of Various Naïve Bayes Using GridSearch Approach In Phishing Email Dataset. *Sinkron*, 8(4). <https://doi.org/10.33395/sinkron.v8i4.12958>
- Rodrigues, A. P., Fernandes, R., Aakash, A., Abhishek, B., Shetty, A., Atul, K., Lakshmana, K., & Shafi, R. M. (2022). Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/5211949>
- Sani, R. R., Pratiwi, Y. A., Winarno, S., Udayanti, E. D., & Alzami, F. (2022). Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia. *Jurnal Masyarakat Informatika*, 13(2). <https://doi.org/10.14710/jmasif.13.2.47983>
- Setiawan, E. I., & Lestari, I. (2021). Stance Classification Pada Berita Berbahasa Indonesia Berbasis Bidirectional LSTM. *Journal of Intelligent System and Computation*, 3(1). <https://doi.org/10.52985/insyst.v3i1.148>
- Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*. <https://doi.org/10.1109/ICACTM.2019.8776800>
- Siti Khomsah, Rima Dias Ramadhani, & Sena Wijaya. (2022). The Accuracy Comparison Between Word2Vec and FastText On Sentiment Analysis of Hotel Reviews. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 6(3). <https://doi.org/10.29207/resti.v6i3.3711>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4). <https://doi.org/10.1016/j.ipm.2009.03.002>

- Vernanda, Y., Hansun, S., & Kristanda, M. B. (2020). Indonesian language email spam detection using n-gram and naïve bayes algorithm. *Bulletin of Electrical Engineering and Informatics*, 9(5). <https://doi.org/10.11591/eei.v9i5.2444>
- Wibisono, A. D., Dadi Rizkiono, S., & Wantoro, A. (2020). FILTERING SPAM EMAIL MENGGUNAKAN METODE NAIVE BAYES. *TELEFORTECH: Journal of Telematics and Information Technology*, 1(1). <https://doi.org/10.33365/tft.v1i1.685>
- Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1). <https://doi.org/10.1177/0165551516677946>
- Yao, T., Zhai, Z., & Gao, B. (2020). Text Classification Model Based on fastText. *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Information Systems, ICAIIS 2020*. <https://doi.org/10.1109/ICAIS49377.2020.9194939>