Vol. 6, No. 2, Agustus 2025, hal. 319-328

CoSciTech Computer Science and Information Technology

p-ISSN: 2723-567X
e-ISSN: 2723-5661

Jurnal Computer Science and Information Technology (CoSciTech)

http://ejurnal.umri.ac.id/index.php/coscitech/index

Optimasi algoritma deteksi spam email dengan BERT-MI dan jaringan dense

Florentina Yuni Arini [*1], Syahrindra Rafli Santosa[2], Bilqis Winy Aqsa Dewi[3], Muhammad Dzaky[4], Muhammad Zakariyya[5], Muhammad Rizky Albani[6], Nafisa Salsabila[7]
Email: [1*]floyuna@mail.unnes.ac.id, [2]raflisantosa28@students.unnes.ac.id, [3]bilqiswiny022@students.unnes.ac.id, [4]dzaky2355@students.unnes.ac.id, [5]zakariyyam20@students.unnes.ac.id, [6]rzkyalbani@students.unnes.ac.id, [7]nafisasalsabila45@students.unnes.ac.id

[1]Teknik Informatika, Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
[2]Teknik Informatika, Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
[3]Teknik Informatika, Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
[4]Teknik Informatika, Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
[5]Teknik Informatika, Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
[6]Teknik Informatika, Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
[7]Teknik Informatika, Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang

Diterima: 18 Juni 2025 | Direvisi: - | Disetujui: 31 Agustus 2025

Abstrak

Deteksi email spam merupakan tantangan penting dalam menjaga keamanan dan efisiensi komunikasi digital. Penelitian ini mengusulkan dan mengevaluasi sebuah *pipeline* yang dioptimalkan untuk deteksi email spam dengan mengintegrasikan *Bidirectional Encoder Representations from Transformers* (BERT) untuk ekstraksi fitur, seleksi fitur menggunakan *Mutual Information* (MI) untuk mereduksi dimensi, dan jaringan saraf *dense* untuk klasifikasi. Dataset Lingspam, yang terdiri dari 2893 email (2412 *ham* dan 481 spam), digunakan dalam eksperimen dengan pembagian 80% data pelatihan dan 20% data pengujian. Fitur teks diekstraksi menggunakan BERT (*bert-base-uncased*), menghasilkan *embedding* 768 dimensi yang kemudian direduksi menjadi 200 fitur paling relevan menggunakan MI. Model jaringan saraf *dense* dengan arsitektur 256-128-64-32-1 neuron dilatih menggunakan *optimizer* Adam, fungsi *loss binary cross-entropy*, dan teknik *early stopping* serta *class weights* untuk menangani ketidakseimbangan kelas. Hasil evaluasi pada data pengujian menunjukkan performa yang sangat tinggi dengan akurasi 99.14%, presisi 0.9596, *recall* 0.9896, F1-*score* 0.9744, dan ROC-AUC 0.9995. Pendekatan ini menunjukkan bahwa kombinasi BERT-MI dengan jaringan *dense* mampu mencapai akurasi yang sebanding dengan metode yang lebih kompleks, namun dengan potensi arsitektur yang lebih sederhana dan efisien.

Kata kunci: *deteksi email spam, BERT, mutual information, jaringan saraf dense, lingspam*

Optimization of email spam detection algorithm using BERT-MI and dense network

Abstract

Email spam detection is a critical challenge in maintaining the security and efficiency of digital communication. This research proposes and evaluates an optimized pipeline for email spam detection by integrating Bidirectional Encoder Representations from Transformers (BERT) for feature extraction, Mutual Information (MI) for feature selection to reduce dimensionality, and a dense neural network for classification. The Lingspam dataset, consisting of 2893 emails (2412 ham and 481 spam), was used in the experiments with an 80% training and 20% testing data split. Text features were extracted using BERT (bert-base-uncased), resulting in a 768-dimensional embedding, which was then reduced to the 200 most relevant features using MI. A dense neural network model with a 256-128-64-32-1 neuron architecture was trained using the Adam optimizer, binary cross-entropy loss function, and techniques such as early stopping and class weights to handle class imbalance. Evaluation results on

the test data demonstrated very high performance, achieving an accuracy of 99.14%, precision of 0.9596, recall of 0.9896, F1-score of 0.9744, and ROC-AUC of 0.9995. This approach indicates that the combination of BERT-MI with a dense network can achieve accuracy comparable to more complex methods, but with the potential for a simpler and more efficient architecture.

Keywords: *email spam detection, BERT, mutual information, dense neural network, lingspam*

1. PENDAHULUAN

Email telah lama berdiri sebagai benteng dalam dunia komunikasi, menghubungkan individu dan organisasi lintas benua [1]. Dengan fleksibilitas dan jangkauannya, email muncul sebagai alat yang sangat diperlukan untuk menyampaikan informasi, membangun hubungan, dan menjalankan bisnis [1]. Email merupakan suatu entitas penting yang digunakan untuk berkomunikasi digital melalui internet, selain itu digunakan untuk transfer informasi berupa file bahkan dapat digunakan untuk media iklan [2]. Namun, seiring dengan meningkatnya volume penggunaan email, muncul pula berbagai tantangan keamanan, salah satunya adalah email spam. Email spam dapat diartikan sebagai tindakan mendistribusikan pesan yang tidak diminta, seringkali dikirim secara massal menggunakan email [2]. Email spam, juga disebut sebagai *unsolicited commercial email* atau *unsolicited bulk email*, telah menyebabkan beberapa masalah komunikasi dalam kehidupan sehari-hari kita [2]. Pada tahun 2010, diperkirakan spam mencapai hampir 90% dari seluruh email yang terkirim, menghabiskan sumber daya yang signifikan [3]. Kerugian yang disebabkan karena spam antara lain spam menempati sumber daya yang besar (termasuk *bandwidth* jaringan dan ruang penyimpanan), dan contoh kasus spam bisa berupa iklan perjudian maupun pornografi [2]. Selain itu, spam tidak hanya membanjiri kotak masuk tetapi juga membawa risiko seperti distribusi *malware* dan upaya *phishing* [4]. Oleh karena itu, pengembangan sistem deteksi email spam yang efektif dan efisien menjadi sangat krusial untuk menjaga integritas dan keamanan komunikasi email global [4].

Berbagai pendekatan telah dikembangkan untuk mengatasi masalah email spam. Penyaringan berbasis aturan melibatkan aturan atau pola yang telah ditentukan untuk menandai spam berdasarkan kriteria spesifik seperti kata kunci atau informasi pengirim [4]. Algoritma Naïve Bayes, misalnya, merupakan fungsi yang banyak digunakan oleh pengembang filter spam karena sederhana dan mudah diimplementasikan untuk klasifikasi teks [2],[5]. Meskipun efektif untuk spam yang jelas, filter berbasis aturan kesulitan beradaptasi dengan taktik spam yang berkembang. Metode pembelajaran mesin tradisional lainnya seperti *Support Vector Machine* (SVM) juga telah diterapkan. Namun, seiring dengan semakin canggihnya teknik spamming, terdapat kebutuhan kritis akan teknik lanjutan untuk melengkapi metode tradisional [4]. Kemunculan era *deep learning* dan kemajuan dalam Pemrosesan Bahasa Alami (NLP) telah menawarkan jalan yang menjanjikan untuk meningkatkan kemampuan deteksi spam [4]. Model seperti *Recurrent Neural Networks* (RNN), *Long Short-Term Memory* (LSTM), dan *Convolutional Neural Networks* (CNN) [6] mampu mempelajari fitur secara otomatis dari data teks. Lebih lanjut, model bahasa canggih seperti GPT-4, BERT (*Bidirectional Encoder Representations from Transformers*), dan RoBERTa telah menunjukkan kemampuan yang luar biasa dalam memahami teks mirip manusia dan sedang dieksplorasi untuk tugas klasifikasi spam [4].

Meskipun model *deep learning* canggih seperti BERT yang dikombinasikan dengan arsitektur sekuensial (misalnya BiLSTM) menawarkan akurasi yang tinggi, implementasinya seringkali dihadapkan pada tantangan komputasi yang intensif, baik dari segi waktu pelatihan maupun kebutuhan sumber daya perangkat keras. Selain itu, model sekuensial seperti BiLSTM mungkin kurang optimal untuk jenis *input* yang bersifat statis, seperti vektor *embedding* keseluruhan teks email yang dihasilkan oleh token [CLS] pada BERT, karena kekuatan utama BiLSTM terletak pada pemrosesan urutan data temporal atau sekuensial. Hal ini membuka peluang untuk mencari arsitektur yang lebih sederhana namun tetap mampu mempertahankan performa tinggi. Penelitian ini termotivasi oleh kebutuhan untuk mengembangkan sebuah *pipeline* deteksi email spam yang tidak hanya akurat tetapi juga lebih efisien secara komputasi dengan menyederhanakan arsitektur model klasifikasi. Studi oleh [1] juga menyoroti pentingnya keterkiriman email dan bagaimana faktor-faktor seperti konten dan penggunaan tindakan pencegahan (teks biasa, subjek jelas, penghindaran elemen mencurigakan) dapat memitigasi pemicu spam, meskipun fokus mereka adalah pada konten yang dihasilkan AI. Penggunaan email untuk mengirim permintaan manuskrip yang tidak diminta secara massal juga merupakan salah satu bentuk spam yang mengganggu produktivitas akademisi [7].

Berdasarkan latar belakang dan tinjauan di atas, penelitian ini bertujuan untuk melakukan optimasi pada algoritma deteksi spam email dengan mengusulkan sebuah *pipeline* yang efisien. *Pipeline* ini mengintegrasikan kekuatan representasi kontekstual dari *embedding* BERT, efektivitas seleksi fitur menggunakan *Mutual Information* (MI) untuk mereduksi dimensi fitur secara signifikan, dan kesederhanaan arsitektur jaringan saraf *dense* untuk tugas klasifikasi akhir. Pertanyaan utama yang ingin dijawab oleh penelitian ini adalah:

1. Bagaimana performa *pipeline* deteksi email spam yang terdiri dari *embedding* BERT, seleksi fitur *Mutual Information*, dan jaringan saraf *dense* dalam hal akurasi, presisi, *recall*, F1-*score*, dan ROC-AUC pada dataset standar Lingspam?

2. Apakah pendekatan yang diusulkan dengan arsitektur jaringan saraf *dense* yang lebih sederhana mampu mencapai performa yang sebanding atau bahkan lebih baik dibandingkan dengan pendekatan yang menggunakan arsitektur *deep*

learning yang lebih kompleks (seperti BiLSTM) setelah dilakukan seleksi fitur, sekaligus menawarkan potensi efisiensi komputasi yang lebih baik?

3. Sejauh mana seleksi fitur menggunakan *Mutual Information* dapat mengurangi dimensi *embedding* BERT tanpa mengorbankan informasi penting yang dibutuhkan untuk klasifikasi spam yang akurat?

Kontribusi utama dari penelitian ini adalah menyajikan sebuah metode deteksi email spam yang mengoptimalkan keseimbangan antara akurasi tinggi dan efisiensi komputasi melalui penerapan seleksi fitur MI yang cepat dan penggunaan jaringan saraf *dense* yang lebih ramping, sehingga menawarkan solusi yang lebih praktis untuk aplikasi dunia nyata.

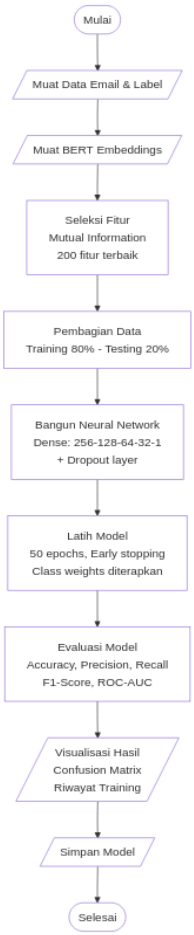
2. METODE PENELITIAN

2.1. Analisis Masalah

Penelitian ini didorong oleh kebutuhan untuk mengeksplorasi pendekatan baru dalam deteksi email spam yang tidak hanya akurat tetapi juga mempertimbangkan efisiensi komputasi. Sebagaimana diketahui, email spam terus menjadi tantangan signifikan dalam komunikasi digital, menyebabkan kerugian waktu, sumber daya, dan potensi risiko keamanan seperti penyebaran *malware* dan *phishing*. Meskipun model *deep learning* canggih seperti GWO-BERT-BiLSTM telah menunjukkan akurasi yang sangat tinggi (99.14%) dalam tugas ini [8], arsitektur tersebut seringkali melibatkan kompleksitas yang tinggi dan beban komputasi yang besar [9]. Penelitian ini bertujuan untuk menyelidiki apakah kombinasi *embedding* BERT, seleksi fitur *Mutual Information* untuk reduksi dimensi, dan arsitektur jaringan saraf *dense* yang relatif lebih sederhana dapat mencapai tingkat akurasi yang kompetitif, sekaligus menawarkan potensi penyederhanaan model dan pengurangan waktu komputasi. Fokus utama adalah pada eksplorasi dan evaluasi performa pendekatan baru ini dalam upaya mengoptimalkan algoritma deteksi spam email.

2.2. Desain Pemecahan Masalah

Desain pemecahan masalah dalam penelitian ini mengikuti alur sistematis yang diilustrasikan pada Gambar 1. Tahapan-tahapan utama dalam desain ini dijelaskan sebagai berikut:



Gambar 1. Flowchart Deteksi Email Spam Menggunakan BERT-MI dan Jaringan Dense

2.2.1 Dataset

Penelitian ini menggunakan dataset publik **Lingspam**, khususnya varian lemm_stop. Dataset ini terdiri dari email yang telah melalui proses lemmatisasi dan penghapusan *stop words*. Dataset Lingspam berisi total 2893 email, yang terbagi menjadi 2412 email *ham* (non-spam) dan 481 email spam. Setiap email dalam dataset ini diberi label biner, yaitu 0 untuk email *ham* dan 1 untuk email spam.

2.2.2. Ekstraksi Email dan Label Spam/Non-Spam

Tahap awal adalah memuat dataset. Proses ini melibatkan pembacaan setiap file email dari struktur direktori dataset Lingspam. Selama proses pembacaan, seluruh teks email dikonversi menjadi huruf kecil (*lowercasing*) untuk memastikan konsistensi data. Label untuk setiap email (spam atau *ham*) ditentukan berdasarkan awalan nama file; file yang dimulai dengan spmsg diberi label 1 (spam), dan sisanya diberi label 0 (*ham*).

2.2.3. Ekstraksi *Embedding* BERT (Token CLS)

Untuk merepresentasikan konten email dalam bentuk vektor numerik, penelitian ini memanfaatkan *Bidirectional Encoder Representations from Transformers* (BERT) [10]. BERT dirancang untuk melakukan pra-pelatihan representasi dua arah yang mendalam dari teks tanpa label dengan menggabungkan konteks kiri dan kanan di semua lapisan . Model pra-terlatih **bert-base-uncased** dari *library* Transformers (Hugging Face) digunakan. Arsitektur model BERT adalah *multi-layer bidirectional Transformer encoder* berdasarkan implementasi asli yang dijelaskan dalam [11]. Mekanisme inti dari BERT adalah *Transformer encoder* yang menggunakan *self-attention*. Fungsi *attention* dapat dideskripsikan sebagai pemetaan sebuah *query* dan sekumpulan pasangan *key-value* ke sebuah *output*, di mana *query*, *keys*, *values*, dan *output* semuanya adalah vektor . *Output* dihitung sebagai jumlah tertimbang dari *values*, di mana bobot yang ditetapkan untuk setiap *value* dihitung oleh fungsi

kompatibilitas dari *query* dengan *key* yang sesuai . *Scaled Dot-Product Attention* yang digunakan dalam Transformer dan BERT dihitung sebagai berikut:

$$A(Q,\ K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

Fungsi *attention* ini memetakan sebuah *query* dan sekumpulan pasangan *key-value* ke sebuah *output*, di mana semua entitas tersebut berbentuk vektor.

- Q **(Query)**: Representasi vektor dari kata yang sedang menjadi fokus pemrosesan.
- K **(Key)**: Representasi vektor dari seluruh kata dalam sekuens teks, yang berfungsi sebagai pembanding terhadap *Query*.
- V **(Value)**: Representasi vektor dari seluruh kata dalam sekuens, yang nilainya akan dibobotkan untuk membentuk *output*.
- QK^T: Operasi perkalian matriks ini menghasilkan skor atensi yang mengkuantifikasi relevansi antara *Query* dan setiap *Key*.
- $\sqrt{d_k}$: Faktor penskalaan, dengan d_k sebagai dimensi vektor *Key* dan *Query*. Penskalaan ini krusial untuk menjaga stabilitas gradien selama proses pelatihan.
- Hasil dari operasi tersebut kemudian dinormalisasi menggunakan fungsi *softmax* untuk menghasilkan bobot atensi, yang selanjutnya dikalikan dengan matriks *Value* (V) untuk menghasilkan representasi kontekstual dari *output*.

Fungsi *softmax* digunakan dalam mekanisme *attention* untuk mentransformasi skor atensi mentah menjadi sebuah distribusi probabilitas.

$$soft\max(Z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \qquad (2)$$

Fungsi ini mengambil sebuah vektor input Z dan mengubah setiap komponen z_i menjadi nilai probabilitas. Setiap nilai output berada dalam rentang [0, 1] dan jumlah keseluruhan nilai output adalah 1, sehingga membentuk distribusi probabilitas yang valid.

2.2.4. Seleksi Fitur dengan *Mutual Information* (MI)

Untuk mereduksi dimensi *embedding* BERT (768 dimensi) dan meningkatkan efisiensi, teknik seleksi fitur **Mutual Information (MI)** diterapkan. MI mengukur ketergantungan non-linear antara masing-masing fitur X (dimensi *embedding*) dan variabel target Y (label spam/ham) **[11]**. MI antara dua variabel diskrit X dan Y didefinisikan sebagai:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (3)$$

di mana $p(x,y)$ adalah probabilitas bersama $X = x$ dan $Y = y$, dan $p(x)$ serta $p(y)$ adalah probabilitas marginal dari X dan Y. Skor MI dihitung menggunakan mutual_info_classif dari Scikit-learn. Sebanyak **200 fitur dengan skor MI tertinggi** dipilih.

2.2.5. Pembagian Data dan Penanganan Ketidakseimbangan Kelas

Matriks fitur yang telah direduksi (2893 email × 200 fitur) dibagi menjadi **80% data pelatihan** dan **20% data pengujian** menggunakan train_test_split dari Scikit-learn secara bertingkat (*stratified*) dengan random_state=42. Untuk menangani ketidakseimbangan kelas dalam data pelatihan, **pembobotan kelas (*class weights*)** dihitung menggunakan compute_class_weight dari Scikit-learn dengan parameter class_weight='balanced'.

2.2.6. Arsitektur Model Jaringan Saraf *Dense*

Model klasifikasi menggunakan jaringan saraf *dense* (DNN) dibangun dengan TensorFlow/Keras. Arsitekturnya adalah sebagai berikut:

- **Lapisan Masukan**: Menerima 200 fitur.
- **Lapisan Tersembunyi**: Terdiri dari empat lapisan *Dense* dengan neuron masing-masing **256, 128, 64, dan 32**. Semua lapisan tersembunyi menggunakan fungsi aktivasi **ReLU (*Rectified Linear Unit*)**. Fungsi ReLU didefinisikan sebagai:

$$ReLU(x) = \max(0,x) \qquad (4)$$

Fungsi ReLU merupakan fungsi aktivasi non-linear yang akan menghasilkan output berupa input itu sendiri jika nilainya positif, dan menghasilkan nol jika nilainya negatif. Penggunaan ReLU membantu mengatasi masalah vanishing gradient dan efisien secara komputasi.

- **Lapisan Keluaran**: Satu neuron dengan fungsi aktivasi **Sigmoid**, yang memetakan output ke rentang [0, 1] untuk probabilitas kelas spam. Fungsi Sigmoid didefinisikan sebagai:

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (5)$$

Fungsi Sigmoid mentransformasi nilai input skalar z ke rentang [0, 1], yang dapat diinterpretasikan sebagai probabilitas keanggotaan kelas. Dalam konteks klasifikasi biner ini, outputnya merepresentasikan probabilitas sebuah email diklasifikasikan sebagai spam.

2.2.7. Pelatihan Model

Kompilasi Model: Model dikompilasi dengan *optimizer* **Adam** (*learning rate* 0.0008), fungsi *loss* **binary_crossentropy**, dan metrik **accuracy**. Fungsi *binary cross-entropy loss* untuk satu sampel adalah:

$$L = -(y \log(p) + (1-y) \log(1-p)) \quad (6)$$

Di mana:

- L : Nilai *loss* untuk satu sampel data.
- y : Label sebenarnya (ground truth), yaitu 1 untuk spam dan 0 untuk ham.
- p : Prediksi probabilitas dari model (output dari fungsi Sigmoid) bahwa sampel tersebut adalah spam.
- Jika $y = 1$ (spam): Rumus menjadi −log(p). *Loss* akan kecil jika p mendekati 1, dan besar jika p mendekati 0.
- Jika $y = 0$ (ham): Rumus menjadi −log(1−p). *Loss* akan kecil jika p mendekati 0, dan besar jika p mendekati 1.

Fungsi ini secara efektif memberikan "hukuman" kepada model ketika prediksinya jauh dari kenyataan.

2.2.8. Evaluasi Model

Performa model dievaluasi pada data pengujian menggunakan metrik berikut, yang dihitung berdasarkan komponen *confusion matrix*: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) [12].

- Akurasi

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

- Presisi

$$Presisi = \frac{TP}{TP+FP} \quad (8)$$

- *Recall*

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

- F1-Score

$$F1-Score = 2 \cdot \frac{Presisi \cdot Recall}{Presisi+Recall} \quad (10)$$

- ROC-AUC: Mengukur kemampuan model membedakan antar kelas. Kurva ROC diplot dengan *True Positive Rate* (TPR, sama dengan Recall) terhadap *False Positive Rate* (FPR).

$$FPR = \frac{FP}{FP+TN} \quad (11)$$

2.2.9. Visualisasi Hasil dan Penyimpanan Model

Confusion matrix dan plot riwayat pelatihan divisualisasikan menggunakan Seaborn dan Matplotlib. Model yang telah dilatih disimpan.

2.3. Lingkungan Implementasi

Eksperimen dijalankan pada Google Colaboratory. *Library* utama meliputi: TensorFlow/Keras, Scikit-learn, Transformers (Hugging Face), Numpy, Matplotlib, dan Seaborn.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan dan membahas hasil eksperimen dari pendekatan deteksi email spam yang diusulkan, yang mengintegrasikan *embedding* BERT, seleksi fitur *Mutual Information* (MI), dan jaringan saraf *dense*. Evaluasi dilakukan pada data pengujian dari dataset Lingspam untuk menilai efektivitas dan potensi efisiensi dari model yang dikembangkan.

3.1. Performa Model Deteksi Spam yang Diusulkan (BERT-MI-Dense)

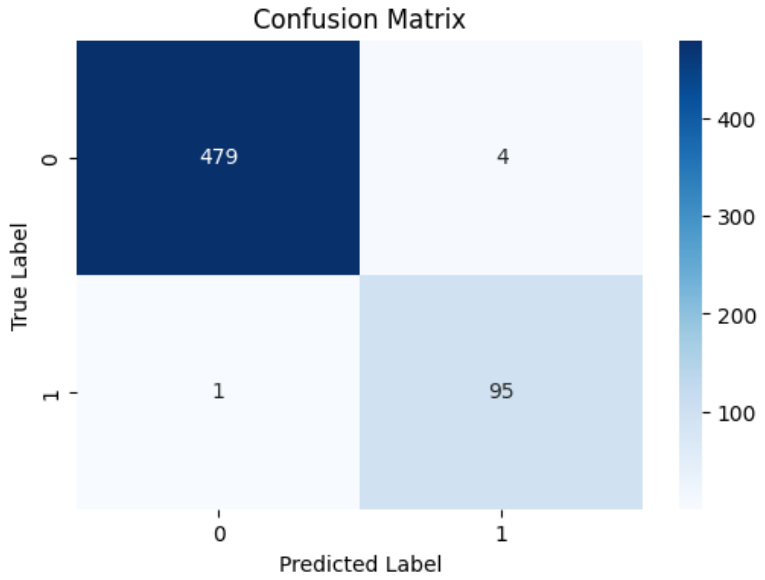
Model jaringan saraf *dense* yang diusulkan, setelah dilatih dengan fitur *embedding* BERT yang telah direduksi menggunakan MI, dievaluasi pada data pengujian yang terdiri dari 579 email (482 *ham* dan 97 spam). Metrik performa yang dicapai oleh model disajikan pada Tabel 1.

Tabel 1. Metrik Performa Model BERT-MI-Dense pada Data Pengujian

METRIK	NILAI
AKURASI	0.9914
PRESISI	0.9596
RECALL	0.9896
F1-*SCORE*	0.9744
ROC-AUC	0.9995

Hasil pada Tabel 1 menunjukkan bahwa model yang diusulkan mencapai performa yang sangat tinggi di semua metrik evaluasi. Akurasi sebesar 99.14% mengindikasikan bahwa model mampu mengklasifikasikan email spam dan *ham* dengan benar pada sebagian besar kasus. Nilai presisi 0.9596 menunjukkan bahwa dari semua email yang diprediksi sebagai spam, sekitar 95.96% di antaranya adalah benar-benar spam. *Recall* sebesar 0.9896 menandakan bahwa model berhasil mengidentifikasi 98.96% dari seluruh email spam aktual yang ada dalam data pengujian. F1-*Score* sebesar 0.9744, yang merupakan rata-rata harmonik dari presisi dan *recall*, juga menunjukkan keseimbangan yang sangat baik antara kedua metrik tersebut, terutama penting untuk dataset yang memiliki ketidakseimbangan kelas seperti Lingspam. Selain itu, nilai ROC-AUC yang sangat tinggi, yaitu 0.9995, mengindikasikan kemampuan model yang luar biasa dalam membedakan antara kelas spam dan *ham* pada berbagai ambang batas probabilitas.

Untuk analisis lebih lanjut mengenai kesalahan klasifikasi, *confusion matrix* dari hasil prediksi model pada data pengujian disajikan pada Gambar 2.



Gambar 2. Confusion Matrix Model BERT-MI-Dense pada Data Pengujian

Dari Gambar 2, dapat diamati bahwa model membuat sangat sedikit kesalahan. Terdapat:

- ***True Negatives* (TN)**: 481 email *ham* yang diprediksi dengan benar sebagai *ham*.

- ***False Positives* (FP)**: 1 email *ham* yang salah diprediksi sebagai spam.
- ***False Negatives* (FN)**: 4 email spam yang salah diprediksi sebagai *ham*.
- ***True Positives* (TP)**: 95 email spam yang diprediksi dengan benar sebagai spam.

RJumlah *false positives* yang rendah (hanya 1) sangat penting karena meminimalkan risiko email penting yang bukan spam masuk ke folder spam. Sementara itu, jumlah *false negatives* yang juga rendah (4) menunjukkan bahwa model efektif dalam menangkap sebagian besar email spam.

Konvergensi model selama proses pelatihan dapat diamati dari plot riwayat pelatihan pada Gambar 3.

Gambar 3. Plot Riwayat Pelatihan dan Validasi Model BERT-MI-Dense ((a) Loss, (b) Akurasi)

Grafik *loss* pelatihan dan validasi (Gambar 3a) menunjukkan bahwa *training loss* menurun secara signifikan pada *epoch-epoch* awal dan kemudian stabil pada nilai yang sangat rendah. *Validation loss* juga menunjukkan pola serupa, menurun dan stabil di sekitar nilai yang rendah, meskipun dengan sedikit fluktuasi. Tidak ada peningkatan signifikan pada *validation loss* yang mengindikasikan *overfitting* parah, terutama karena penggunaan mekanisme *early stopping* yang menghentikan pelatihan ketika tidak ada perbaikan lebih lanjut pada *validation loss* (terlihat pelatihan berhenti sebelum 50 *epochs* maksimal). Grafik akurasi pelatihan dan validasi (Gambar 3b) menunjukkan bahwa *training accuracy* dengan cepat mencapai nilai yang sangat tinggi, mendekati 100%. *Validation accuracy* juga meningkat pesat dan stabil pada tingkat yang tinggi, menunjukkan bahwa model mampu melakukan generalisasi dengan baik pada data yang tidak terlihat selama proses pelatihan.

3.2. Analisis Pengaruh Seleksi Fitur dan Arsitektur Model

Salah satu aspek penting dari penelitian ini adalah penggunaan seleksi fitur *Mutual Information* (MI) untuk mereduksi dimensi *embedding* BERT dari 768 menjadi 200 fitur. Pengurangan dimensi ini bertujuan untuk mengurangi kompleksitas model dan potensi waktu komputasi tanpa mengorbankan performa secara signifikan. Hasil akurasi 99.14% yang dicapai oleh model jaringan *dense* dengan hanya 200 fitur menunjukkan bahwa MI berhasil mempertahankan informasi yang paling relevan untuk tugas klasifikasi spam.

Untuk memberikan konteks lebih lanjut mengenai efektivitas arsitektur jaringan *dense* yang diusulkan, dilakukan perbandingan dengan model BiLSTM yang juga menggunakan fitur hasil seleksi MI (BiLSTM-MI) yang dilatih pada dataset dan dengan pembagian yang sama. Model BiLSTM-MI mencapai akurasi sebesar 0.9534, presisi 0.8350, *recall* 0.8958, F1-*Score* 0.8643, dan ROC-AUC 0.9936. Meskipun performa BiLSTM-MI ini tergolong baik, model BERT-MI-Dense menunjukkan keunggulan di semua metrik evaluasi. Hal ini mengindikasikan bahwa untuk representasi fitur statis seperti *embedding* [CLS] dari BERT yang telah direduksi dimensinya, arsitektur jaringan *dense* yang lebih sederhana mungkin lebih efektif atau lebih mudah untuk dioptimalkan dibandingkan arsitektur sekuensial seperti BiLSTM. Model BiLSTM dirancang untuk menangkap dependensi sekuensial dalam data, yang mungkin kurang relevan ketika inputnya adalah satu vektor fitur tunggal per email.

Sebagai catatan tambahan, eksperimen awal dengan model BiLSTM tanpa seleksi fitur MI menunjukkan performa yang jauh lebih rendah (akurasi di kisaran 0.5630 - 0.6356, berdasarkan output notebook Anda untuk dua model BiLSTM awal). Ini

semakin memperkuat argumen bahwa penanganan fitur, baik melalui seleksi fitur maupun pemilihan arsitektur yang tepat, sangat krusial ketika bekerja dengan *embedding* berdimensi tinggi seperti BERT.

3.3. Pembahasan Terhadap Pertanyaan Penelitian dan Implikasi

Berdasarkan hasil yang diperoleh, penelitian ini dapat menjawab pertanyaan-pertanyaan penelitian yang telah dirumuskan sebelumnya:

1. **Bagaimana performa *pipeline* deteksi email spam yang terdiri dari *embedding* BERT, seleksi fitur *Mutual Information*, dan jaringan saraf *dense*?** Hasil eksperimen menunjukkan bahwa *pipeline* yang diusulkan memiliki performa yang sangat tinggi, dengan akurasi mencapai 99.14% dan nilai ROC-AUC 0.9995 pada data pengujian dataset Lingspam. Ini menunjukkan kemampuan klasifikasi yang sangat baik antara email spam dan *ham*.

2. **Apakah pendekatan yang diusulkan dengan arsitektur jaringan saraf *dense* yang lebih sederhana mampu mencapai performa yang sebanding atau bahkan lebih baik dibandingkan dengan pendekatan yang menggunakan arsitektur *deep learning* yang lebih kompleks (seperti BiLSTM atau referensi GWO-BERT-BiLSTM) setelah dilakukan seleksi fitur, sekaligus menawarkan potensi efisiensi komputasi yang lebih baik?** Ya, pendekatan BERT-MI-Dense mencapai akurasi 99.14%, yang sebanding dengan hasil penelitian Nasreen et al. (2024) yang menggunakan GWO-BERT-BiLSTM (juga 99.14% akurasi pada dataset Lingspam). Hal ini menarik karena arsitektur jaringan *dense* umumnya dianggap lebih sederhana dan memiliki parameter yang lebih sedikit dibandingkan BiLSTM, terutama setelah dimensi input direduksi secara signifikan oleh MI. Dari sisi waktu komputasi, sel notebook yang menjalankan keseluruhan proses pelatihan dan evaluasi model *Dense Network* dengan MI dieksekusi dalam 2892.029 detik (sekitar 48.2 menit) di lingkungan Google Colab dengan GPU T4. Meskipun perbandingan langsung waktu komputasi dengan GWO-BERT-BiLSTM memerlukan implementasi dan pengujian pada platform yang sama, penggunaan MI untuk reduksi fitur dan arsitektur *dense* yang lebih ramping secara inheren berpotensi menawarkan efisiensi yang lebih baik.

3. **Sejauh mana seleksi fitur menggunakan *Mutual Information* dapat mengurangi dimensi *embedding* BERT tanpa mengorbankan informasi penting yang dibutuhkan untuk klasifikasi spam yang akurat?** MI terbukti sangat efektif dalam mengurangi dimensi fitur dari 768 menjadi 200 (pengurangan sekitar 74%) sambil tetap memungkinkan model *dense* mencapai akurasi yang sangat tinggi. Ini menunjukkan bahwa 200 fitur yang dipilih oleh MI berhasil menangkap sebagian besar informasi relevan yang terkandung dalam *embedding* BERT asli untuk tugas klasifikasi spam pada dataset ini.

Kontribusi utama dari penelitian ini adalah demonstrasi bahwa *pipeline* yang relatif lebih sederhana dan efisien, yaitu dengan menggabungkan *embedding* BERT, seleksi fitur MI, dan jaringan saraf *dense*, mampu menghasilkan performa deteksi spam yang sangat kompetitif. Hasil ini setara dengan pendekatan yang mungkin lebih kompleks, seperti yang dilaporkan oleh Nasreen et al. (2024). Pendekatan ini menawarkan solusi praktis yang dapat diimplementasikan dengan sumber daya komputasi yang mungkin lebih terbatas, tanpa mengorbankan akurasi deteksi secara signifikan.

Implikasi praktis dari penelitian ini adalah bahwa untuk tugas klasifikasi teks dengan *embedding* BERT yang statis (seperti penggunaan token [CLS]), penggunaan seleksi fitur yang cerdas seperti MI diikuti oleh arsitektur jaringan saraf *dense* yang dioptimalkan dapat menjadi alternatif yang sangat efektif dan efisien dibandingkan dengan model sekuensial yang lebih berat.

DAFTAR PUSTAKA

[1] N. Bouchareb and I. Morad, "ANALYZING THE IMPACT OF AI-GENERATED EMAIL MARKETING CONTENT ON EMAIL DELIVERABILITY IN SPAM FOLDER PLACEMENT," *HOLISTICA J. Bus. Public Adm.*, vol. 15, no. 1, pp. 96–106, 2024, doi: 10.2478/hjbpa-2024-0006.

[2] J. Al Amien, H. Mukhtar, and M. A. Rucyat, "Filtering spam email menggunakan algoritma naïve bayes," *J. Comput. Sci. Inf. Technol.*, vol. 3, no. 1, pp. 9–19, 2022.

[3] A. G. West and I. Lee, "Towards the Effective Temporal Association Mining of Spam Blacklists," 2011.

[4] K. I. Roumeliotis, N. D. Tselikas, and N. K. Dimitrios, "Next-Generation Spam Filtering : Comparative Fine-Tuning of LLMs , NLPs , and CNN Models for Email Spam Classification," *Electronics*, vol. 13, no. 11, pp. 1–24, 2024.

[5] F. Apri Wenando, R. Hayami, S. Soni, A. Fitria, and D. Shifana, "Sentimen Analisis Masyarakt terhadap Kasus Penembakan Brigadir J Menggunakan Algoritma Naïve Bayes Classifier," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 4, no. 2, 2023, doi: 10.37859/coscitech.v4i2.5686.

[6] Gina Purnama Insany, Indra Yustiana, and Sri Rahmawati, "Penerapan KNN dan ANN pada klasifikasi status gizi balita berdasarkan indeks antropometri," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 4, no. 2, pp. 385–393, 2023, doi: 10.37859/coscitech.v4i2.5079.

[7] J. Sureda-negre, A. Calvo-sastre, and R. Comas-forgas, "Predatory journals and publishers : Characteristics and impact

of academic spam to researchers in educational sciences," *Learn. Publ.*, vol. 35, no. January, pp. 441–447, 2022, doi: 10.1002/leap.1450.

[8] G. Nasreen, M. Murad Khan, M. Younus, B. Zafar, and M. Kashif Hanif, "Email spam detection by deep learning models using novel feature selection technique and BERT," *Egypt. Informatics J.*, vol. 26, no. January, p. 100473, 2024, doi: 10.1016/j.eij.2024.100473.

[9] T. Um, G. Kim, W. Lee, B. Oh, B. Seo, and M. Kweun, "FastFlow : Accelerating Deep Learning Model Training with Smart Offloading of Input Data Pipeline," *PVLDB (Proceedings VLDB Endowment)*, vol. 16, no. 5, pp. 1086–1099, 2023, doi: 10.14778/3579075.3579083.

[10] J. Devlin, M. C. Kenton, L. Kristina, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 2019.

[11] A. Vaswani *et al.*, "Attention Is All You Need," *31st Conf. Neural Inf. Process. Syst. (NIPS 2017)*, 2017.

[12] F. Salim *et al.*, "Klasifikasi Berita Palsu Menggunakan Pendekatan Hybrid CNN-LSTM," *J. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, pp. 55–59, 2025.