Journal of Computer Science and Information Technology (JCSIT)
Volume 3, No 3 – Juni 2026
e-ISSN : 3031-8467

IMPLEMENTASI MODEL TRANSFORMER INDOBERT BERBASIS DEEP LEARNING UNTUK DETEKSI SPAM PESAN TEKS BERBAHASA INDONESIA

Annisa Alfrini[1], Desi Anggreani[2], Fahrim Irhamna Rachman[3]
[1,2,3]Universitas Muhammadiyah Makassar, Indonesia
[1]105841112122@student.unismuh.ac.id
[2]desianggreani@unismuh.ac.id
[3]fachrim141020@unismuh.ac.id
* Coresponding Author

Received: 03-06- 2026	Revised: 20-06-2026	Approved: 27-06-2026

ABSTRAK

Perkembangan teknologi informasi dan komunikasi dalam beberapa tahun terakhir telah mendorong peningkatan penggunaan media digital secara masif, salah satunya adalah aplikasi Telegram. Namun, fenomena ini memunculkan permasalahan baru, yaitu ancaman penyebaran pesan spam yang tidak terkendali, berpotensi mengganggu kenyamanan, menyebabkan kerugian finansial, serta mengancam keamanan data pengguna akibat skema penipuan. Penelitian ini bertujuan untuk mengimplementasikan model deep learning berbasis arsitektur Transformer, secara spesifik menggunakan model IndoBERT, untuk mendeteksi dan mengklasifikasikan pesan teks Telegram berbahasa Indonesia ke dalam kategori spam dan non-spam (ham). Penelitian ini memanfaatkan dataset seimbang yang berjumlah 5.971 pesan teks (2.979 spam dan 2.992 non-spam), yang dikumpulkan secara langsung melalui metode scraping dari berbagai grup Telegram publik. Tahapan penelitian dirancang secara sistematis, meliputi prapemrosesan teks (case folding, cleaning, dan tokenisasi awal), tokenisasi lanjutan menggunakan metode WordPiece, penyesuaian parameter, serta proses fine-tuning pada model pralatih IndoBERT. Evaluasi performa model diukur secara komprehensif menggunakan confusion matrix. Hasil penelitian menunjukkan bahwa model IndoBERT berhasil mengidentifikasi pola linguistik spam, termasuk penggunaan bahasa informal, tautan eksternal, hingga manipulasi karakter. Model menghasilkan kinerja klasifikasi yang sangat optimal dengan perolehan nilai akurasi sebesar 97,32%, presisi 98,62%, recall 95,97%, dan F1-score 97,28%. Selain itu, pengujian langsung terhadap 10 sampel data baru (unseen data) menunjukkan tingkat keberhasilan prediksi sebesar 100% secara spesifik pada skala pengujian tersebut. Kesimpulannya, pendekatan model IndoBERT terbukti sangat efektif, tangguh, dan adaptif untuk diimplementasikan sebagai sistem penyaringan spam otomatis secara real-time pada teks berbahasa Indonesia.

Kata Kunci: *deteksi spam, Deep Learning, Transformer, IndoBERT, klasifikasi teks.*

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi dalam beberapa tahun terakhir telah mendorong peningkatan penggunaan media digital dan platform pesan instan secara masif. Transformasi ini memfasilitasi pertukaran informasi yang efisien tanpa batas geografis, namun sekaligus memunculkan ancaman keamanan siber yang signifikan berupa penyebaran pesan spam [1]. Spam yang didistribusikan secara massal umumnya mengandung muatan promosi agresif, penipuan digital (*scam*), hingga tautan *phishing* yang berpotensi menyebabkan kerugian finansial dan pencurian data pribadi pengguna. Seiring dengan peningkatan volume data teks harian, teknik penyamaran spam (*obfuscation*) menjadi semakin kompleks. Oleh karena itu, dibutuhkan sistem klasifikasi teks yang mampu mendeteksi anomali pesan secara otomatis, akurat, dan adaptif [2].

Secara historis, metode deteksi spam bertumpu pada pendekatan berbasis aturan (*rule-based*) dan algoritma *machine learning* tradisional seperti *naïve bayes* maupun *support vector machine* (SVM). Namun, metode probabilistik tersebut memiliki keterbatasan fundamental dalam memahami konteks semantik yang kompleks karena hanya bergantung pada frekuensi kemunculan fitur kata secara manual [2]. Paradigma

klasifikasi teks kemudian bergeser drastis dengan hadirnya pendekatan *deep learning*, khususnya melalui arsitektur Transformer yang diperkenalkan oleh Vaswani dkk. [3]. Arsitektur ini mengandalkan mekanisme *self-attention* yang memungkinkan model untuk menangkap hubungan relasional antar kata secara global dan paralel [4]. Pengembangan lebih lanjut dari arsitektur ini menghasilkan model *bidirectional encoder representations from transformers* (BERT), yang menetapkan standar baru (*state of the art*) dalam *natural language processing* (NLP) berkat kemampuannya merepresentasikan konteks kalimat secara dua arah (*bidirectional*).

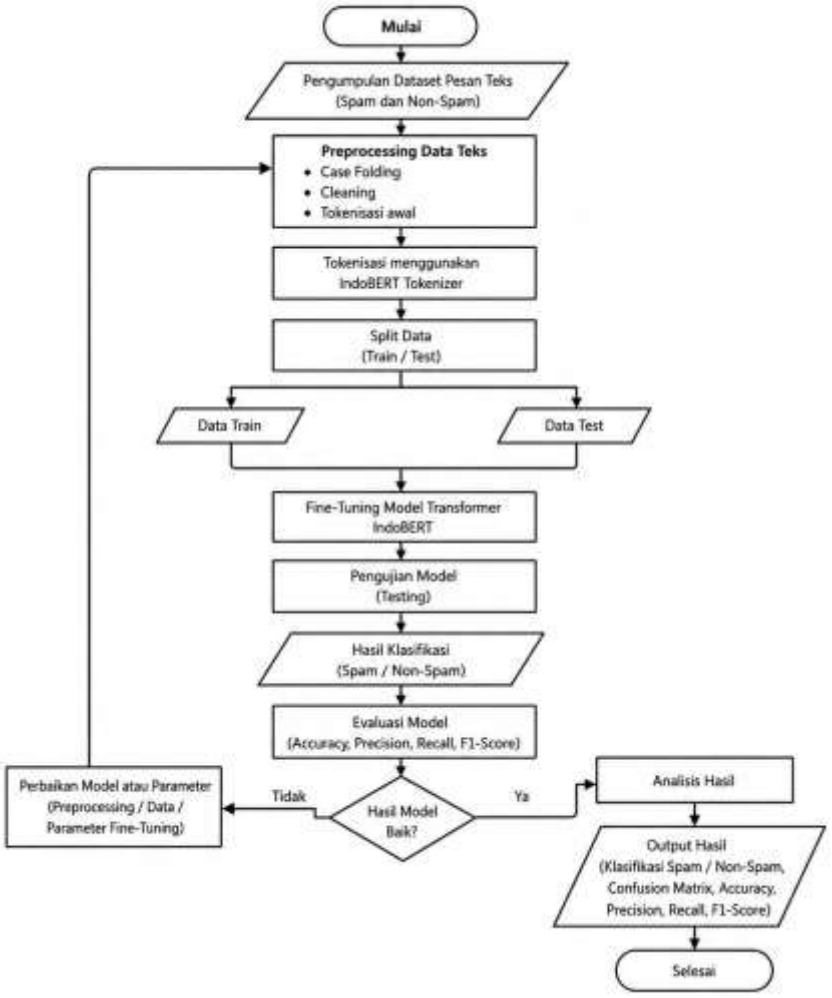
Untuk pemrosesan teks spesifik berbahasa Indonesia, arsitektur ini diadaptasi menjadi model bahasa pralatih IndoBERT. Model yang dilatih menggunakan korpus teks bahasa Indonesia berskala raksasa terbukti memiliki performa kompetitif dalam memahami struktur linguistik dan kosakata lokal [5], [6]. Berbagai studi terdahulu telah mengonfirmasi superioritas IndoBERT dalam klasifikasi teks. Amin dkk. [1] membuktikan bahwa implementasi IndoBERT mampu mengungguli algoritma klasik dalam mendeteksi spam berbahasa Indonesia. Optimalisasi performa juga terus dieksplorasi, mulai dari penerapan teknik augmentasi data untuk mengatasi keterbatasan sampel [2], ekstraksi fitur tambahan seperti emoji dan relasi konteks [7], [8], hingga integrasi model menggunakan *framework Hugging Face* untuk menghasilkan antarmuka klasifikasi prediktif [9].

Meskipun model berbasis Transformer menunjukkan kinerja impresif, sejumlah tantangan teknis masih belum sepenuhnya terselesaikan. Secara teoretis, model berbasis BERT bergantung pada representasi token yang diperoleh dari data pralatih. Meskipun menggunakan tokenisasi WordPiece, BERT dapat mengalami penurunan kualitas representasi ketika menghadapi kata slang ekstrem, singkatan tidak baku, dan variasi penulisan yang jarang muncul dalam korpus pelatihan. Aji dkk. [10] menjelaskan bahwa keberagaman bahasa, dialek, dan fenomena *code-switching* di Indonesia menjadi tantangan bagi sistem NLP karena meningkatkan variasi linguistik yang harus diproses model. Dalam konteks media sosial, variasi tersebut sering muncul dalam bentuk kata slang yang dapat mengurangi konsistensi representasi semantik. Oleh karena itu, tahapan prapemrosesan berupa normalisasi teks informal menjadi sangat krusial untuk menekan tingkat *out of vocabulary* (OoV) sehingga model mampu menghasilkan representasi fitur kontekstual yang jauh lebih optimal sebelum memasuki lapisan klasifikasi spam.

Berdasarkan celah penelitian tersebut, penelitian ini bertujuan mengimplementasikan arsitektur Transformer (IndoBERT) untuk mendeteksi pesan spam berbahasa Indonesia di Telegram. Telegram dipilih secara spesifik karena eksistensi grup publik berskala masif menjadikannya target utama infiltrasi spam otomatis. Berbeda dengan platform konvensional seperti SMS [2], spam di Telegram memiliki pola yang jauh lebih kompleks, memadukan narasi manipulatif, penyamaran karakter tingkat tinggi (seperti "PR0M0"), dan tautan beruntun untuk mengecoh filter keamanan. Sebagai solusi, penelitian ini memfokuskan kontribusi pada optimalisasi prapemrosesan teks untuk mengatasi anomali *slang* dan *obfuscation*, serta penerapan *fine-tuning* berbasis *early stopping*. Kinerja model dievaluasi secara komprehensif menggunakan *accuracy, precision, recall*, dan *F1-score* [11]. Hasil penelitian ini diharapkan membuktikan ketangguhan IndoBERT terhadap teks informal, sekaligus menghasilkan sistem klasifikasi otomatis yang aplikatif secara *real-time* pada ekosistem komunikasi digital.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental dengan merancang sistem deteksi spam untuk mengklasifikasikan pesan teks berbahasa Indonesia ke dalam kategori spam dan non-spam. Sistem ini dibangun berbasis *deep learning* menggunakan arsitektur Transformer, secara spesifik memanfaatkan model pralatih IndoBERT yang mampu memahami karakteristik linguistik bahasa lokal. Secara visual, tahapan-tahapan dalam penelitian ini diilustrasikan pada Gambar 1.



Gambar 1. Flowchart Sistem Deteksi Spam Teks

Berdasarkan Gambar 1, tahap pertama dimulai dengan pengumpulan dataset utama berupa sekumpulan pesan teks berbahasa Indonesia yang telah dianotasi ke dalam dua kelas, yaitu spam dan non-spam. Sebelum diolah oleh arsitektur *deep learning*, data mentah wajib melalui tahapan prapemrosesan teks untuk menghilangkan derau (*noise*) linguistik dan meningkatkan kualitas data. Proses ini difokuskan pada tiga tahap, yaitu penyeragaman seluruh huruf menjadi huruf kecil (*case folding*), pembersihan teks dari karakter yang tidak relevan seperti tanda baca, simbol, dan tautan (*cleaning*), serta pemecahan kalimat menjadi satuan kata tunggal (*tokenisasi awal*). Dataset yang telah bersih ini kemudian dibagi menggunakan metode *stratified split* dengan rasio 80% sebagai data latih (*training data*) dan 20% sebagai data uji (*testing data*). Dalam fase pelatihan, proporsi 20% data uji tersebut difungsikan secara bersamaan sebagai parameter validasi (*validation data*) untuk mengevaluasi *validation accuracy* pada akhir setiap iterasi *epoch*. Hal ini dilakukan untuk mekanisme

pencegahan *overfitting* (*early stopping*), sehingga model dapat dilatih secara proporsional dan dievaluasi secara objektif.

Data teks yang telah diproses kemudian dimasukkan ke dalam sistem dan diubah menggunakan *tokenizer* bawaan IndoBERT. Metode tokenisasi lanjutan yang digunakan adalah *WordPiece* (*subword tokenization*), yang secara efektif mampu menangani kata-kata di luar kosakata (*out of vocabulary*) dengan memecahnya menjadi representasi numerik (ID token). Setiap representasi numerik ini kemudian diproses melalui *embedding layer* yang terdiri atas integrasi *token embedding* dan *positional embedding* untuk memetakan nilai matematis sekaligus menangkap urutan posisi kata secara akurat di dalam kalimat.

Proses ekstraksi fitur linguistik yang menjadi inti dari sistem ini terjadi di dalam *Transformer encoder* IndoBERT yang memanfaatkan mekanisme *multi-head attention* dan *feed forward neural network* untuk memahami konteks kalimat secara dua arah (*bidirectional*). Representasi makna dari keseluruhan kalimat diekstraksi menggunakan token khusus [CLS]. Nilai representasi [CLS] ini kemudian diteruskan ke *fully connected layer* yang diakhiri dengan fungsi aktivasi *Softmax* untuk menghasilkan skor probabilitas akhir terkait prediksi pesan ke dalam kelas spam atau non-spam. Selama proses ini, tahapan *fine-tuning* dilakukan secara berulang untuk menyesuaikan bobot parameter model pralatih IndoBERT agar dapat secara spesifik mempelajari dan mengenali pola unik dari dataset lokal Telegram yang digunakan. Merujuk pada alur Gambar 1, sistem dilengkapi dengan mekanisme "Perbaikan Model atau Parameter". Kriteria kuantitatif yang memicu alur perbaikan ini dieksekusi berdasarkan ambang batas akurasi (*threshold accuracy*) minimum yang ditetapkan sebesar 85% (0,85). Apabila hasil evaluasi akhir model tidak mampu melampaui persentase ambang batas tersebut, maka alur sistem menolak menyimpan model dan memicu proses perbaikan ulang dengan cara memodifikasi konfigurasi *hyperparameter* (seperti *learning rate, max length*, maupun jumlah *epoch*) hingga performa optimal tercapai.

Pengujian sistem dilakukan terhadap keluaran model menggunakan data uji yang tidak pernah dilibatkan dalam proses pelatihan sebelumnya. Hal ini bertujuan untuk mengukur tingkat generalisasi model secara nyata terhadap data baru. Evaluasi kinerja klasifikasi dianalisis menggunakan *confusion matrix* yang memetakan jumlah prediksi ke dalam empat kategori, yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Berdasarkan distribusi matriks tersebut, performa sistem diukur melalui empat metrik standar, yaitu akurasi untuk melihat persentase ketepatan sistem secara keseluruhan, presisi untuk mengukur ketepatan tebakan pada kelas spam, *recall* untuk mengetahui sensitivitas sistem dalam mendeteksi seluruh pesan spam aktual, dan *F1-score* sebagai ukuran rata-rata harmonik yang menyeimbangkan nilai presisi dan *recall*.

HASIL PENELITIAN DAN PEMBAHASAN

Pengumpulan, Prapemrosesan, dan Pembagian Data

Tahapan awal penelitian ini melibatkan pengumpulan data tekstual bersumber dari platform media sosial Telegram melalui metode *scraping* secara otomatis, menghasilkan total data mentah sebanyak 5.971 baris pesan. Tahapan berikutnya adalah pelabelan data yang dilakukan secara manual dengan mengacu pada indikator pesan spam. Penentuan label didasarkan pada penggunaan kata-kata promosi agresif, ajakan persuasif bertransaksi, penyertaan tautan (*hyperlink*) eksternal, serta penggunaan teknik penyamaran kata. Perbandingan karakteristik pola linguistik antara

pesan spam dan non-spam ditunjukkan pada Tabel 1.

Tabel 1. Karakteristik Pesan Spam dan Non-Spam

Label	Contoh Pesan
Spam	"🔥PROMO TERBATAS! Diskon hingga 70% hari ini saja. Buruan cek sebelum kehabisan: https://tokopedia.com"
Spam	"Di promo T.G.I.F kali ini, Sahabat [#Alfamart](https://www.instagram.com/explore/tags/alfamart/) dapat menukarkan 100 A-Poin untuk voucher buy 1 get 1 di Domino's loo! Nikmati 1 medium super value pizza secara GRATIS!🍕 Periode 8 - 14 Desember 2023 S&K Berlaku!"
Spam	🚨 BRO!!! Lihat produk LUAR BIASA ini biar hidupmu LEBIH BAIK! 💊🔥 👉 KLIK SEKARANG: http://www.gretan.com/ss/ 👈 (BURUAN SEBELUM DIHAPUS!!!) 🏃♂️💨 Bypass: Pikiran ceria itu kuat. Perang = cara mengajar geografi Amerika. Waktu+sabar = kekuatan. ⚠️ WARNING: Jika terbunuh, Anda kehilangan hidup Anda. 💀🛑
Spam	🚨 Rencana Kesehatan Kristen Kami! 🏥✨ Cover Visi, Gigi, Medis & banyak lagi bersama orang-orang baik di Christian Health Center! Nilai kami membedakan kami. 🙏💯 👉 KLIK DI SINI: http://www.[Link_Dihapus].com 👈 Finish Solutions | 9600 La Cienega, Inglewood, CA 90301 | Pesan ini adalah iklan/ajakan.
Non-Spam	Tp belum pasti ya. Kalau ada temen mah mau. Soalnya ga ada halangan untuk hdr kyknya hihi
Non-Spam	Mau minta kirim foto2 selempang yg kita waktu itu dong. Makasihhh
Non-Spam	Ada semua di panduan dan sudah dishare di grup ini dan fb
Non-Spam	[Info] bagi teman2 yg mengumpulkan data asisten lab/matakuliah bukti berupa surat keterangannya sudah bisa diambil ke pak andr
Non-Spam	Ada yg tdk bisa jam tertentu? Kemungkinan pagi jam 8, tapi nanti diumumkan.

Berdasarkan proses pelabelan, dataset menunjukkan distribusi kelas yang sangat seimbang. Kondisi ideal ini sangat mendukung proses pelatihan karena dapat mengurangi kecenderungan model untuk hanya mengenali satu kelas dominan saja, sehingga pembelajaran karakteristik linguistik teks dapat berlangsung secara proporsional. Rincian hasil pelabelan data yang merepresentasikan tingkat keseimbangan kelas disajikan pada Tabel 2.

Tabel 2. Hasil Pelabelan Data Teks Telegram

Kategori	Jumlah Data	Persentase
Ham (*Non-Spam*)	2.992	50,11%
Spam	2.979	49,89%
Total	5.971	100%

Keseimbangan proporsi ini menjadi poin krusial untuk meminimalkan risiko terjadinya bias mayoritas pada algoritma klasifikasi. Sebelum digunakan pada tahap *fine-tuning*, keseluruhan data melalui prapemrosesan teks untuk menghilangkan derau (*noise*). Proses ini terdiri dari *case folding*, *cleaning* (penghilangan tanda baca, simbol,

emotikon, dan tautan URL), serta tokenisasi awal. Dataset bersih ini kemudian dibagi menggunakan metode pembagian data secara *stratified* (*stratified split*) dengan rasio 80% data latih (4.776 baris) dan 20% data uji (1.195 baris) guna menjaga representasi karakteristik distribusi data yang objektif.

Fine-Tuning Model IndoBERT

Model yang digunakan dalam penelitian ini adalah model pralatih *indobenchmark/indobert-base-p1* yang diimplementasikan melalui kelas *BertForSequenceClassification*. Teks diproses menggunakan *WordPiece Tokenizer* bawaan IndoBERT dengan parameter *max_length* sebesar 128 token guna menyeragamkan dimensi masukan. Token dikonversi menjadi indeks numerik (*input IDs*) dan dipetakan menjadi representasi vektor berdimensi 768 melalui *Embedding Layer* yang merupakan integrasi dari *token, segment,* dan *position embedding*.

Pemrosesan ekstraksi fitur kontekstual secara dua arah terjadi di dalam *Transformer encoder* (12 lapisan). Dari seluruh urutan token, arsitektur ini memanfaatkan vektor dari token khusus [CLS] sebagai representasi menyeluruh yang mengintegrasikan informasi semantik isi kalimat. Vektor [CLS] diteruskan menuju lapisan *fully connected* untuk menghasilkan nilai *logits* (skor mentah). Nilai *logits* selanjutnya diproses menggunakan fungsi aktivasi *Softmax* untuk diubah menjadi probabilitas pasti pada rentang 0 hingga 1. Penentuan label akhir dilakukan menggunakan fungsi *argmax* yang memilih kelas dengan probabilitas tertinggi.

Tahapan yang telah dijelaskan tersebut menggambarkan alur pemrosesan data pada satu kali *forward pass*, dimulai dari masukan berupa teks, proses tokenisasi, pembentukan *embedding*, ekstraksi representasi kontekstual oleh *Transformer Encoder*, hingga menghasilkan probabilitas kelas melalui lapisan klasifikasi. Pada fase *fine-tuning*, rangkaian proses tersebut tidak hanya dilakukan satu kali, melainkan diulang secara berulang terhadap seluruh data latih selama beberapa *epoch*. Setelah setiap *forward pass* selesai, model menghitung nilai kesalahan prediksi (*loss*) berdasarkan perbedaan antara hasil prediksi dan label sebenarnya. Nilai *loss* tersebut kemudian digunakan dalam proses *backpropagation* untuk memperbarui bobot internal model sehingga kemampuan klasifikasi meningkat secara bertahap pada setiap iterasi pelatihan. Proses *fine-tuning* pada penelitian ini dikendalikan menggunakan beberapa *hyperparameter* yang berperan dalam mengatur mekanisme pelatihan model, sebagaimana disajikan pada Tabel 3.

Tabel 3. Parameter dan Konfigurasi *Fine-Tuning* IndoBERT

Parameter	Rincian Nilai dan Detail Konfigurasi
Batch Size	16
Epochs	20
Learning Rate	2e-5
Warmup Ratio	0.1 (10% dari total langkah untuk pemanasan
Optimizer	AdamW (dengan nilai eps = 1e-8)
Early Stopping Patience	5 *epochs* (toleransi jika akurasi validasi stagnan

Berdasarkan Tabel 3, proses *fine-tuning* menggunakan *batch size* sebesar 16, sehingga pada setiap iterasi model mempelajari 16 data secara bersamaan sebelum dilakukan pembaruan bobot. Proses pelatihan dirancang berlangsung hingga maksimum 20 *epoch*, yaitu ketika seluruh data latih telah diproses sebanyak 20 kali. Selama proses tersebut, pembaruan bobot dilakukan menggunakan *optimizer* AdamW dengan *learning rate* sebesar 2×10^{-5}. Nilai *learning rate* yang relatif kecil dipilih untuk menjaga agar perubahan bobot berlangsung secara bertahap sehingga pengetahuan

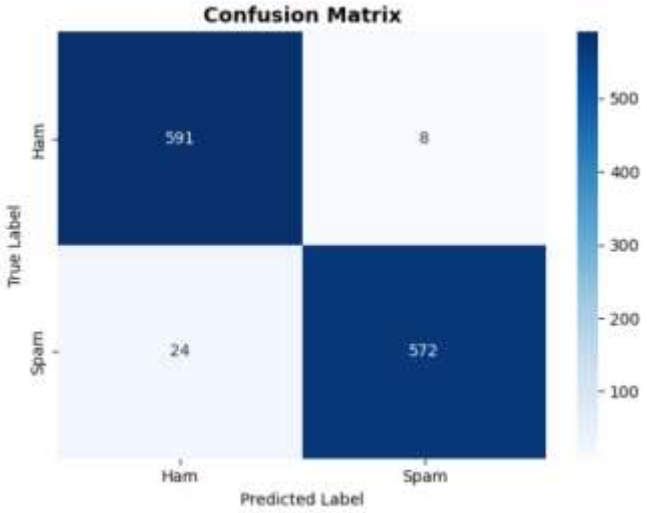
bahasa yang telah dipelajari pada tahap *pre-training* tetap dapat dipertahankan, sekaligus memungkinkan model beradaptasi terhadap karakteristik pesan spam pada dataset penelitian.

Selain itu, penelitian ini menerapkan *warmup ratio* sebesar 0,1, yang berarti nilai *learning rate* dinaikkan secara bertahap selama 10% dari total langkah pelatihan sebelum mencapai nilai yang telah ditentukan. Strategi ini bertujuan menjaga kestabilan proses optimasi pada tahap awal pelatihan serta mengurangi risiko perubahan bobot yang terlalu besar. Untuk mencegah terjadinya *overfitting*, proses *fine-tuning* juga dilengkapi dengan mekanisme *early stopping* dengan nilai *patience* sebesar lima *epoch*. Apabila performa model pada data validasi tidak menunjukkan peningkatan selama lima *epoch* berturut-turut, proses pelatihan akan dihentikan secara otomatis dan bobot model dengan performa terbaik akan dipertahankan sebagai model akhir. Model hasil *fine-tuning* tersebut selanjutnya digunakan pada tahap evaluasi untuk mengukur kemampuan klasifikasi terhadap data uji yang belum pernah dipelajari sebelumnya.

Hasil Klasifikasi dan Evaluasi Model

Pengujian performa model IndoBERT dilakukan menggunakan 1.195 data uji, yang secara proporsional terdiri atas 599 pesan non-spam (ham) dan 596 pesan spam. Penggunaan data uji yang sama sekali tidak dilibatkan dalam proses pelatihan ini bertujuan untuk mengukur secara objektif kemampuan model dalam menggeneralisasi dan mengklasifikasikan data baru. Kemampuan generalisasi ini sangat penting untuk memastikan bahwa sistem tidak hanya sekadar menghafal data latih (*overfitting*), melainkan benar-benar tangguh dalam mengenali variasi penyamaran pesan spam di platform Telegram. Hasil prediksi dari pengujian tersebut kemudian dievaluasi dan divisualisasikan menggunakan *confusion matrix* (Gambar 2). Visualisasi ini memberikan pemetaan yang terstruktur mengenai hubungan antara kelas aktual (*ground truth*) dan kelas hasil prediksi model, sehingga jumlah prediksi yang benar maupun yang mengalami kesalahan klasifikasi dapat diidentifikasi dengan jelas.

Pemetaan matriks tersebut selanjutnya menjadi instrumen utama untuk menganalisis risiko salah sasaran (*false alarm*).



Gambar 2. Confusion Matrix Pengujian Model IndoBERT

Model mendominasi nilai pada diagonal utama dengan jumlah prediksi benar yang sangat tinggi, yaitu 591 data *true negative* (TN) dan 572 data *true positive* (TP).

Kesalahan klasifikasi sangat minimal, ditunjukkan oleh 8 data *false positive* (FP) dan 24 data *false negative* (FN). Analisis kesalahan (*error analysis*) menunjukkan bahwa nilai FP (pesan *ham* diprediksi spam) umumnya terjadi akibat penggunaan diksi kasual yang identik dengan *template* pemasaran. Sebagai contoh, pesan komunikasi sah seperti "Ica, besok jangan lupa bawa brosur promo kampus, kita bagi-bagi bareng" rentan diklasifikasikan sebagai spam akibat dominansi token "brosur", "promo", dan "bagi-bagi". Sebaliknya, kasus FN (pesan spam diprediksi *ham*) disebabkan oleh teknik penyamaran promosi tanpa tautan eksternal yang didesain menyerupai percakapan natural, contohnya: "Hai kak, maaf ganggu, kalau minat nambah penghasilan sampingan bisa langsung chat nomor ini ya, makasih". Keterbatasan ini menunjukkan bahwa meskipun model sangat tangguh, teknik manipulasi semantik tingkat tinggi pada platform komunikasi instan terkadang masih mampu mengecoh klasifikasi.

Evaluasi kuantitatif dari kinerja model tersebut diukur menggunakan empat metrik standar yang dirangkum pada Tabel 4.

Tabel 4. Hasil Pengujian Kinerja Klasifikasi IndoBERT

Metrik Evaluasi	Rumus Dasar	Hasil Kinerja (%)
Accuracy	(TP + TN) / Keseluruhan Data	97,32%
Precision	TP / (TP + FP)	98,62%
Recall	TP / (TP + FN)	95,97%
F1-Score	2 * (Precision * Recall) / (Precision + Recall)	97,28%

Nilai presisi 98,62% menegaskan bahwa risiko pemblokiran pesan sah dapat ditekan pada tingkat yang sangat rendah, sedangkan *F1-Score* 97,28% merepresentasikan keseimbangan performa model yang tangguh dan stabil. Untuk memperkuat signifikansi performa arsitektur yang diajukan, penelitian ini menetapkan eksperimen pembanding menggunakan algoritma *machine learning* konvensional, yaitu *multinomial naive bayes* (MultinomialNB) berbasis ekstraksi fitur TF-IDF sebagai *baseline*. Pengujian pada dataset yang sama menunjukkan bahwa model konvensional tersebut mampu mencapai tingkat akurasi 89,45% dan *F1-score* 88,12%. Perbandingan ini membuktikan bahwa kemampuan Transformer (IndoBERT) dalam memahami relasi kontekstual kalimat secara dua arah jauh mengungguli metode probabilistik konvensional yang hanya bergantung pada frekuensi kemunculan kata.

Implementasi Antarmuka dan Pengujian Data Baru

Keunggulan dan keunikan utama dari penelitian ini terletak pada pengujian kelenturan model terhadap data baru (*unseen data*) serta integrasinya ke dalam layanan bot interaktif di platform Telegram secara *real-time*.

Gambar 3. Antarmuka Bot Telegram pada Pengujian Pesan

Pada pengujian data baru yang merepresentasikan variasi teks riil, model diuji secara spesifik menggunakan 10 sampel data (terdiri atas 5 pesan spam dan 5 pesan *ham*) yang sepenuhnya baru dan tidak berafiliasi dengan dataset awal. Pengujian tersebut mencatatkan tingkat keberhasilan prediksi yang direpresentasikan pada Tabel 5.

Tabel 5. Hasil Pengujian Model Menggunakan Sampel Data Baru

Pesan	Label	Deteksi	Akurasi
Selamat! Anda terpilih mendapatkan hadiah senilai Rp500.000! Klik link ini sekarang: bit.ly/klaimhadiah	Spam	Spam	99,93%
Hei, kamu udah ngerjain tugas kelompok buat besok belum?	Non-Spam	Non-Spam	100,00%
Meetingnya jadi jam 3 sore ya, tolong kabarin yang lain juga.	Non-Spam	Non-Spam	100,00%
PR0M0 TERBATAS! Diskon 90% untuk semua produk hari ini aja. Buruan sebelum kehabisan! wa.me/628xxx	Spam	Spam	99,96%
Makasih ya udah bantu kemarin, aku beneran tertolong banget.	Non-Spam	Non-Spam	100,00%
GR4TIS ongkir + cashback 50%! Daftar sekarang dan dapatkan bonus eksklusif. Link: tokobagus.id/promo	Spam	Spam	99,37%
Jadwal rapat bulan ini sudah saya kirim ke email masing-masing.	Non-Spam	Non-Spam	100,00%
Nanti makan siang di mana? Aku bebas kok, ikut aja.	Non-Spam	Non-Spam	100,00%
⚡ FLASH SALE! Beli 1 gratis 1 hanya sampai malam ini. Jangan sampai nyesel! Hubungi: 0812-XXXX-XXXX	Spam	Spam	99,65%
🔥 PROMO TERBATAS! Diskon hingga 70% hari ini saja. Buruan cek sebelum kehabisan: https://hadiah.com	Spam	Spam	99,98%

Berdasarkan Tabel 5, model menunjukkan tingkat *robustness* yang sangat tinggi

dalam mengenali manipulasi karakter pada pesan spam, seperti penggunaan istilah "PR0M0" atau "GR4TIS". Karakteristik manipulatif berhasil diidentifikasi dengan probabilitas kepastian (*confidence score*) mendekati mutlak. Mekanisme *chat-driven* melalui antarmuka Telegram membuktikan bahwa pendekatan arsitektur IndoBERT tidak hanya akurat secara teoretis, tetapi juga sangat adaptif, praktis, dan aplikatif sebagai sistem penyaringan spam otomatis langsung pada ekosistem distribusinya.

KESIMPULAN

Penelitian ini menyimpulkan bahwa implementasi arsitektur Transformer berbasis model pralatih IndoBERT terbukti efektif dan adaptif dalam mengklasifikasikan pesan spam berbahasa Indonesia di platform Telegram. Kemampuan model dalam mengekstraksi fitur semantik dari bahasa informal, tautan eksternal, hingga manipulasi karakter penyamaran menghasilkan kinerja klasifikasi yang sangat optimal. Hal ini secara kuantitatif dibuktikan melalui perolehan tingkat akurasi sebesar 97,32%, presisi 98,62%, *recall* 95,97%, dan *F1-score* 97,28%. Pengujian lebih lanjut terhadap sampel data mandiri (*unseen data*) juga menegaskan kapabilitas generalisasi model yang stabil dan tangguh dalam menangani variasi teks riil di luar proses pelatihan. Untuk menyempurnakan batasan penelitian ini, pengembangan selanjutnya sangat disarankan untuk memperluas skala dan keragaman dataset, menerapkan teknik augmentasi data, serta melakukan studi komparatif menggunakan arsitektur Transformer lain maupun algoritma *deep learning* (seperti LSTM atau CNN) untuk menemukan efisiensi komputasi tertinggi. Sebagai rekomendasi praktis operasional, model IndoBERT yang dihasilkan dari penelitian ini sangat direkomendasikan untuk diimplementasikan dan diintegrasikan secara penuh ke dalam sistem penyaringan spam otomatis berbasis *real-time* pada ekosistem platform komunikasi digital.

DAFTAR PUSTAKA

[1] M. B. M. Amin *et al.*, “Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model Bert,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 6, pp. 1291–1302, Dec. 2024, doi: 10.25126/jtiik.2024118121.

[2] N. Latifah, R. Dwiyansaputra, and G. S. Nugraha, “Multiclass Text Classification of Indonesian Short Message Service (SMS) Spam using Deep Learning Method and Easy Data Augmentation,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 3, pp. 663–676, Jun. 2024, doi: 10.30812/matrik.v23i3.3835.

[3] A. Vaswani *et al.*, “Attention Is All You Need,” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008. doi: https://doi.org/10.48550/arXiv.1706.03762.

[4] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, “Efficient Transformers: A Survey,” *ACM Comput. Surv.*, vol. 55, no. 6, pp. 1–28, Jun. 2022, doi: 10.1145/3530811.

[5] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857. doi: https://doi.org/10.18653/v1/2020.aacl-main.85.

[6] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” in

Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 757–770. doi: https://doi.org/10.18653/v1/2020.coling-main.66.

[7] A. R. Chrismanto, A. Kartika Sari, and Y. Suyanto, "SPAMID-PAIR: A Novel Indonesian Post-Comment Pairs Dataset Containing Emoji," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 11, pp. 92–100, 2022, doi: https://dx.doi.org/10.14569/IJACSA.2022.0131110.

[8] A. R. Chrismanto, A. K. Sari, and Y. Suyanto, "Enhancing Spam Comment Detection on Social Media With Emoji Feature and Post-Comment Pairs Approach Using Ensemble Methods of Machine Learning," *IEEE Access*, vol. 11, pp. 80246–80265, 2023, doi: 10.1109/ACCESS.2023.3299853.

[9] R. A. Supono and M. I. Imani, "Implementasi Machine Learning untuk Klasifikasi Email Spam Menggunakan Indobert, Hugging Face Transfomers dan Streamlit," *Jurnal Sosial dan Teknologi (SOSTECH)*, vol. 6, no. 1, pp. 420–440, 2026, doi: https://doi.org/10.59188/jurnalsostech.v6i1.32659.

[10] A. Fikri Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Long Papers, 2022, pp. 7226–7249. doi: https://doi.org/10.18653/v1/2022.acl-long.500.

[11] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 820–836, 2024, doi: https://doi.org/10.1162/tacl_a_00675.