 ISSN:2302-8149
e-ISSN:2540-9719

Optimasi *Fine-Tuning* IndoBERT-Lite untuk Deteksi Spam pada Layanan Pelanggan Digital

Optimization of IndoBERT-Lite Fine-Tuning for Spam Detection in Digital Customer Services

[1]Farouq Mulya Al Simabua*, [2]Lathifah Alfat
[1,2]Program Studi Informatika, Fakultas Teknologi dan Desain, Universitas Pembangunan Jaya
[1,2]Jl. Boulevard UPJ, Bintaro Jaya, Kec. Ciputat, Kota Tangerang Selatan, Banten 15413, Indonesia
*e-mail: farouqsimabua@gmail.com

(***received*:** 7 May 2026, ***revised*:** 15 May 2026, ***accepted*:** 16 May 2026)

Abstrak

Sistem moderasi teks otomatis pada platform layanan publik sering kali dieksploitasi oleh teks spam calo manipulatif yang menawarkan jasa keuangan ilegal. Penelitian klasifikasi teks terdahulu sering memprioritaskan metrik akurasi tinggi namun mengabaikan dampak kebocoran data (*data leakage*) akibat template spam yang berulang, sebuah kecacatan metodologi yang memicu *overfitting* parah pada model. Penelitian ini bertujuan merancang dan mengoptimalkan model *Natural Language Processing* (NLP) menggunakan arsitektur IndoBERT-Lite untuk membedakan keluhan organik pengguna dan komentar manipulatif calo. Metodologi yang diusulkan berfokus pada deduplikasi data ekstrem, menyaring 55.156 rekaman mentah menjadi dataset seimbang berisi 4.626 sampel unik (57,1% organik, 42,9% spam). Proses pelatihan dioptimalkan menggunakan *Gradient Accumulation* dan Early Stopping guna memastikan kemampuan generalisasi yang sesungguhnya. Hasil evaluasi menunjukkan bahwa model yang dioptimalkan berhasil memitigasi overfitting awal, mencapai tingkat akurasi dan F1-score sebesar 98% terhadap himpunan data pengujian baru (*unseen data*). Hasil riset tersebut memberikan solusi moderasi otomatis yang andal dan bebas dari kebocoran data untuk sistem layanan pelanggan digital internal.

Kata kunci: klasifikasi teks, kebocoran data, IndoBERT-Lite, moderasi otomatis, spam calo

Abstract

Automated text moderation systems on public service platforms are often exploited by manipulative spam messages from brokers offering illegal financial services. Previous text classification studies have frequently prioritized high accuracy metrics while overlooking the impact of data leakage caused by repetitive spam templates, a methodological flaw that can lead to severe model overfitting. This study aims to design and optimize a Natural Language Processing (NLP) model using the IndoBERT-Lite architecture to distinguish between organic user complaints and manipulative broker-generated comments. The proposed methodology focuses on extreme data deduplication, reducing 55,156 raw records into a balanced dataset containing 4,626 unique samples (57.1% organic and 42.9% spam). The training process was optimized using Gradient Accumulation and Early Stopping to ensure genuine model generalization capability. The evaluation results demonstrate that the optimized model successfully mitigated the initial overfitting problem, achieving both accuracy and F1-score values of 98% on unseen test data. These findings provide a reliable and data leakage–free automated moderation solution for internal digital customer service systems.

Keywords: *broker spam, data leakage, early stopping, IndoBERT-Lite, Text classification*

1 Pendahuluan

Eksploitasi ruang diskusi publik pada platform layanan pelanggan digital oleh oknum calo (broker spam) kini telah berkembang menjadi ancaman serius terhadap privasi masyarakat dan kredibilitas instansi jaminan sosial seperti BPJS Ketenagakerjaan. Kanal resmi yang sejatinya diperuntukkan bagi keluhan organik peserta sering kali dibanjiri oleh penawaran jasa ilegal. Teks manipulatif ini dirancang dengan pola bahasa dinamis, memanfaatkan singkatan, dan memodifikasi

 ISSN:2302-8149
e-ISSN:2540-9719

karakter untuk secara sengaja menembus filter moderasi konvensional berbasis aturan (*rule-based*), sehingga penanganan otomatis menggunakan algoritma cerdas menjadi sangat krusial guna menjaga kebersihan data interaksi teks [1], [2]. Penelitian ini bertujuan memformulasikan masalah kebocoran data pada teks repetitif, merancang tahapan pembersihan ekstrem, serta membangun model klasifikasi berbasis IndoBERT-Lite yang sangat akurat, efisien, dan siap diimplementasikan sebagai sistem penyaring (*filtering*) otomatis di ruang lingkup layanan pelanggan institusi.

2 Tinjauan Literatur

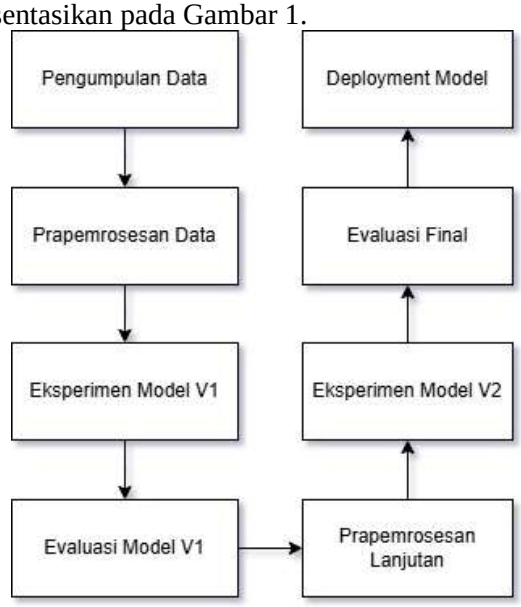
Pendekatan Natural Language Processing (NLP) dengan arsitektur berbasis Transformer seperti BERT telah hadir dan menjadi standar utama untuk mengatasi kompleksitas klasifikasi teks yang tidak mampu ditangani oleh metode konvensional [3], [4]. Namun, terdapat kesenjangan (*research gap*) yang signifikan pada penelitian-penelitian terdahulu: mayoritas riset hanya berfokus pada pencapaian metrik akurasi tinggi tanpa memprioritaskan penyelesaian isu krusial terkait data leakage (kebocoran data). Pada kasus spam di media sosial, kebocoran data sering kali dipicu oleh tingginya duplikasi teks akibat perulangan template (salin-tempel) dari akun calo, yang justru akan menjebak model ke dalam penghafalan (*memorization*) alih-alih generalisasi pola semantik yang sesungguhnya [5], [6].

Penelitian ini bermaksud untuk menambal celah penelitian tersebut, studi ini mengusulkan kebaruan (*novelty*) berupa optimasi fine-tuning menggunakan arsitektur IndoBERT-Lite yang dipadukan dengan strategi mitigasi data leakage melalui tahapan deduplikasi ekstrem. Varian arsitektur ini dipilih karena kemampuannya memangkas parameter komputasi tanpa kehilangan kapabilitas dalam memahami konteks semantik bahasa Indonesia informal [7], [8].

3 Metode Penelitian

3.1 Pendekatan Penelitian

Metodologi dalam riset ini dijalankan melalui skema eksperimental iteratif demi memastikan terbentuknya model klasifikasi teks yang andal. Rangkaian prosesnya disederhanakan ke dalam beberapa tahap esensial yang menitikberatkan pada mitigasi kebocoran data dan peningkatan kinerja model, seperti yang direpresentasikan pada Gambar 1.



Gambar 1 Diagram alir tahapan penelitian

3.2 Teknik Pengumpulan Data

Data dikumpulkan dari kanal media sosial resmi institusi (BPJS Ketenagakerjaan). Korpus data ini sarat dengan penggunaan bahasa Indonesia informal, modifikasi teks, dan singkatan karena bersumber dari interaksi daring publik [9]. Teks keluhan organik dari peserta memiliki pola yang sangat kontras jika dibandingkan dengan teks manipulatif (spam) calo [10]. Contoh variasi teks yang ditemukan di lapangan dapat dilihat pada Tabel 1.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1 Contoh variasi teks pada korpus data media sosial

Label / Kelas	Contoh Teks Komentar
Organik	"Malam kak kalo mau cek no jamsosteknya gimana ya @bpjs.ketenagakerjaan"
Organik	"BPJS Ketenagakerjaan apakah bisa dalam satu keluarga satu orang yang daftar?"
Organik	"Selamat pagi bpk/ibu mau bertanya prihal masalah login di bpjsku email sya salah trus cara untuk merubah email bagai mananya ya"
Calo (Spam)	"Yukkk kak yang mau cairin BPJS ketenagakerjaan nya, siap bantu Sampek cair ya kak, data aman terpercaya fee dijamin murah, dan pastinya adm di akhir setelah cair. Chat wa aja https://wa.me/6281233870797"
Calo (Spam)	"Yang bermasalah seputar BPJS ketenagakerjaan bisa saya bantu proses cepat aman amanah adm murah"
Calo (Spam)	"JASA PENCAIRAN BPJS ,ADM SETELAH CAIR !!! Wa : 089516150381"

Pengumpulan data ini secara keseluruhan menghasilkan populasi masif sebanyak 55.156 rekaman data mentah. Populasi tersebut menunjukkan ketidakseimbangan kelas (*class imbalance*) alami yang sangat signifikan, di mana 53.072 rekaman merupakan teks organik dan hanya 2.084 data merupakan komentar calo (*spam*). Penelitian ini menerapkan strategi penyeimbangan distribusi kelas (*balanced class*) [11] dengan teknik *undersampling* guna mencegah model mengalami bias terhadap kelas mayoritas. Penelitian ini selanjutnya menggunakan seluruh 2.088 data *spam* yang tersedia dan memadukannya dengan sampel acak sebanyak 3.083 data organik untuk membentuk dataset eksperimen awal yang lebih proporsional.

3.3 Prapemrosesan Data

Tahap prapemrosesan pada penelitian ini dilakukan secara minimal (lightweight preprocessing), yakni terbatas pada pembersihan karakter struktural bawaan sistem seperti tanda kutip ganda ("), spasi berlebih, dan baris baru (newline). Penelitian ini secara sengaja tidak menerapkan prapemrosesan konvensional tingkat lanjut (seperti case folding, stemming, atau penghapusan stopword) guna mempertahankan struktur linguistik asli dari teks informal dan template calo.

Tahapan ini memfokuskan transformasi skema atribut untuk menyatukan dua sumber data mentah di samping pembersihan teks dasar. Sumber pertama adalah basis data produksi yang telah diauto-label dengan Regex dan divalidasi tiga adjudikator internal untuk meminimalisasi bias serta tingkat kesalahan prediksi (*false positive/false negative*) dari aturan Regex [10]. Sumber kedua adalah dataset acuan (ground truth) dari instansi. Kedua sumber distandarisasi menjadi format final dengan dua atribut utama: text dan label (0 untuk organik, 1 untuk spam/calo), seperti disajikan pada Tabel 2.

Tabel 2 Transformasi atribut data

No	Format Data Mentah	Format Final
1	JASA PENCAIRAN BPJS ,ADM SETELAH CAIR !!! Wa : 089516150381; 01/06/2025 16.48; 90; 1 (Sumber: Basis data produksi)	JASA PENCAIRAN BPJS ,ADM SETELAH CAIR !!! Wa : 089516150381; 1
2	Narsih Cucu Suarsih saya bisa bantu proses pencarian BPJS Ya hubungi 081350371651; 21/08/2024 06.56; 90; 1 (Sumber: Basis data produksi)	Narsih Cucu Suarsih saya bisa bantu proses pencarian BPJS Ya hubungi 081350371651; 1
3	@paingat_sipayung Yukkk kak yang mau cairin BPJS ketenagakerjaan nya, siap bantu Sampek cair ya kak; 20/01/2024 02.49; 100; 1 (Sumber: Basis data produksi)	@paingat_sipayung Yukkk kak yang mau cairin BPJS ketenagakerjaan nya, siap bantu Sampek cair ya kak; 1
4	"apakah kartu saya msh aktif" (Sumber: Data instansi - Kolom Non SPAM)	apakah kartu saya msh aktif; 0
5	"Apakah bisa d cairkan klo GK ada paklaring" (Sumber: Data instansi - Kolom Non SPAM)	Apakah bisa d cairkan klo GK ada paklaring; 0
6	"Dimas Cahya Putra siap bantu proses klaim bpjsnya ka Punya kendala? Yuk konsultasiin dulu sama Queen jasa	Dimas Cahya Putra siap bantu proses klaim bpjsnya ka Punya kendala? Yuk konsultasiin

 ISSN:2302-8149
e-ISSN:2540-9719

Super aman dan amanah” (Sumber: Data instansi - Kolom SPAM)	dulu sama Queen jasa Super aman dan amanah; 1

3.4 Penanganan Duplikasi Data

Penelitian ini melakukan eksperimen tahap awal (Model v1) menggunakan dataset terstandarisasi untuk mengukur performa dasar (*baseline*) sebelum pengembangan model akhir. Pelatihan model menghasilkan metrik yang anomali pada eksperimen awal tersebut, di mana tingkat akurasi mencapai angka yang nyaris sempurna (>99%) dengan hanya 7 kesalahan prediksi dari total 1,034 data pengujian. Namun, evaluasi kritis terhadap riwayat pelatihan (*training history*) menunjukkan indikasi yang meragukan. Analisis pada kurva *loss* memperlihatkan bahwa *Training Loss* terus menurun mendekati nol, sedangkan *Validation Loss* mengalami stagnasi di angka ~0.03 sejak langkah (*step*) ke-500.

Fenomena stagnasi validasi yang diiringi akurasi tinggi tersebut merupakan indikator kuat terjadinya *overfitting*. Investigasi lanjutan terhadap dataset latih dan uji mengungkap adanya kebocoran data (*data leakage*) berskala signifikan. Kebocoran ini dipicu oleh karakteristik komentar calo yang didominasi oleh teks *template* hasil salin-tempel (*copy-paste*) secara masif [12]. Tingkat kemiripan (*similarity*) antar teks pada kelas spam mencapai 98%, di mana perbedaannya sering kali hanya terletak pada nama akun di awal kalimat (misalnya: variasi antara "[Nama Pengguna] JASA PENCAIRAN..." dan "[Nama Pengguna] JASA PENCAIRAN..."). Kehadiran data identik ini menyebabkan model berhenti mengekstraksi fitur semantik bahasa, dan sebaliknya beralih "menghafal" pola teks yang muncul berulang kali di data latih maupun data uji.

Penelitian ini menerapkan prapemrosesan lanjutan melalui teknik deduplikasi ekstrem sebagai langkah mitigasi untuk mencegah akurasi semu. Pembersihan data dilakukan menggunakan skrip Python untuk mengevaluasi tingkat kemiripan teks dan mengeliminasi data yang tumpang tindih [13]. Proses ini berhasil menghapus sebanyak 545 baris data (sekitar 10.5% dari dataset awal) yang teridentifikasi sebagai duplikat eksak..

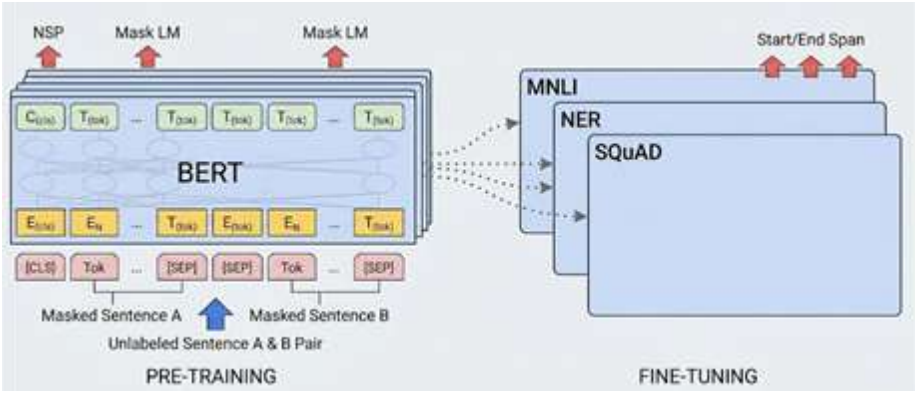
Tahap pembersihan ekstrem ini menghasilkan dataset bersih (Dataset v2) yang secara eksklusif berisi 4.626 sampel teks unik dan saling terasing (*unseen*). Proses pemangkasan duplikat ini membentuk distribusi data yang representatif terhadap kondisi natural di lapangan, menghasilkan rasio kelas yang seimbang (*balanced class*) [14] yaitu 2,641 data organik (57.1%) dan 1.985 data spam (42.9%). Dataset unik inilah yang kemudian difungsikan sebagai korpus final untuk melatih dan menguji model secara objektif.

3.5 Arsitektur Model IndoBERT-Lite

Model klasifikasi teks dalam penelitian ini dibangun menggunakan arsitektur IndoBERT-Lite (khususnya varian indobenchmark/indobert-lite-base-p1). IndoBERT-Lite mengadopsi prinsip arsitektur ALBERT (A Lite BERT), sebuah pendekatan yang berbeda dengan arsitektur BERT standar pada penelitian terdahulu [15]. Keunggulan utamanya terletak pada teknik *factorized embedding parameterization* dan *cross-parameter sharing* [16]. Teknik *cross-parameter sharing* memungkinkan model untuk menggunakan kembali parameter yang sama di seluruh lapisan *Transformer*, sehingga secara drastis mengurangi jumlah parameter total tanpa mengorbankan performa secara signifikan. Pengurangan jumlah parameter ini sangat krusial agar model dapat dijalankan secara efisien pada unit pemroses grafis (*Graphics Processing Unit*/GPU) kelas menengah dengan memori video (*VRAM*) terbatas.

IndoBERT-Lite secara fungsional beroperasi melalui dua fase utama, yakni prapelatihan (*Pre-training*) dan penyesuaian akhir (*Fine-Tuning*), sebagaimana diilustrasikan pada Gambar 2.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 2 Arsitektur dasar dan tahapan pre-training serta fine-tuning pada model berbasis BERT[17]

Gambar 2, pada fase *Pre-training*, model bahasa telah lebih dulu dilatih menggunakan korpus teks berskala masif melalui tugas *Masked Language Modeling* (Mask LM) dan *Next Sentence Prediction* (NSP). Pembelajaran dua arah (*bidirectional*) pada fase ini memungkinkan model untuk menangkap struktur sintaksis dan semantik bahasa secara utuh [17]. Arsitektur dasar tersebut selanjutnya diadaptasi untuk tugas klasifikasi teks biner pada fase *Fine-Tuning*. Bobot parameter model disesuaikan kembali pada tahap ini menggunakan dataset komentar terdeduplikasi milik instansi, sehingga model dapat membedakan konteks makna kata yang identik antara keluhan organik dan tawaran manipulatif calo secara spesifik [18].

Kemampuan generalisasi model ini juga telah teruji secara komprehensif dalam berbagai tugas NLP bahasa Indonesia, di mana arsitektur *Lite* terbukti secara empiris mampu menyeimbangkan antara efisiensi parameter komputasi dan tingginya tingkat presisi klasifikasi [19], termasuk pada studi komparasi langsung yang membuktikan bahwa varian Lite mampu menghasilkan kinerja yang setara dengan arsitektur Base dalam menangani teks ulasan informal [20]. Pemanfaatan varian IndoBERT ini juga didukung oleh keberhasilan implementasi serupa pada deteksi teks negatif di media sosial TikTok yang membuktikan skalabilitas dan konsistensi model dalam melakukan moderasi konten berbahasa Indonesia secara otomatis [21]. Hal ini menjadikannya pilihan paling optimal untuk implementasi pada sistem moderasi otomatis internal. Rincian konfigurasi model yang digunakan disajikan pada Tabel 3.

Tabel 3 Spesifikasi model IndoBERT-Lite

Parameter	Deskripsi / Nilai
Model Name	indobenchmark/indobert-lite-base-p1
Architecture	ALBERT-based (Lite BERT)
Embedding Size	128
Hidden Groups	1
Hidden Layers	12
Max Sequence Length	128 Token

3.6 Parameter dan Skenario Pelatihan

Proses pelatihan model dilakukan dengan mendistribusikan dataset ke dalam porsi 80% untuk data latih (*training set*) sebanyak 3,700 rekaman dan 20% untuk data uji (*test set*) sebanyak 926 rekaman. Pembagian ini dilakukan menggunakan teknik pengambilan sampel terstratifikasi (*stratified sampling*) untuk menjaga proporsi distribusi label yang seimbang pada kedua set data. Seluruh proses eksperimen dijalankan pada lingkungan komputasi dengan spesifikasi perangkat keras GPU NVIDIA RTX 3060 6GB guna mendukung pemrosesan paralel pada arsitektur *Transformer*.

Skenario pelatihan pada model final (v2) dirancang secara spesifik untuk mengatasi kendala memori video (VRAM) sekaligus memitigasi risiko overfitting. Pengaturan learning rate ditetapkan sebesar 2e-5, yang disesuaikan lebih rendah dari eksperimen awal guna menjaga kestabilan konvergensi model, sebagaimana direkomendasikan dalam evaluasi optimasi hyperparameter model Transformer berbahasa Indonesia [22]. Pelatihan ini juga menerapkan teknik *Gradient Accumulation*

 ISSN:2302-8149
e-ISSN:2540-9719

sebanyak 2 langkah untuk mensimulasikan ukuran *batch* efektif sebesar 16, meskipun kapasitas fisik memori hanya menampung *batch* sebesar 8. Mekanisme Early Stopping dengan tingkat kesabaran (patience) sebanyak 3 kali tahapan evaluasi diterapkan guna memastikan model memiliki kemampuan generalisasi yang optimal. Mekanisme ini berfungsi untuk menghentikan proses iterasi secara otomatis apabila nilai kerugian validasi (*validation loss*) tidak lagi menunjukkan penurunan yang signifikan. Konfigurasi lengkap *hyperparameter* yang digunakan dalam pelatihan disajikan pada Tabel 4.

Tabel 4 Konfigurasi hyperparameter pelatihan model

Hyperparameter	Nilai / Pengaturan
Learning Rate	2e-5
Training Batch Size	8
Gradient Accumulation	2
Optimizer	AdamW (fused)
Weight Decay	0.01
Precision	Brain Floating Point 16 (BF16)
Evaluation Strategy	per 50 steps
Early Stopping Patience	3

3.7 Metrik Evaluasi Kinerja

Penelitian ini menggunakan pendekatan *Confusion Matrix* untuk mengukur efektivitas dan performa model IndoBERT-Lite dalam mengklasifikasikan teks. Matriks ini memetakan prediksi model ke dalam empat kuadran evaluasi: *True Positive* (TP) ketika teks calo diprediksi benar sebagai calo, *True Negative* (TN) ketika keluhan organik diprediksi benar sebagai organik, *False Positive* (FP) ketika keluhan organik salah diprediksi sebagai calo, dan *False Negative* (FN) ketika teks calo salah diprediksi sebagai keluhan organik [23], [24].

Kinerja model dievaluasi secara matematis berdasarkan keempat nilai tersebut menggunakan empat metrik utama. Metrik akurasi digunakan untuk mengukur persentase total prediksi benar dari keseluruhan data uji, yang perhitungannya disajikan pada persamaan (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Selanjutnya, tingkat ketepatan antara data yang diminta dengan hasil prediksi yang diberikan oleh model diukur melalui presisi (*precision)*, seperti ditunjukkan pada persamaan (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Kemampuan model dalam menemukan kembali informasi atau mengenali seluruh data pada kelas spam calo diukur menggunakan metrik *recall*, yang formulasinya dapat dilihat pada persamaan (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Terakhir, penelitian ini menggunakan F1-Score sebagai acuan validasi utama guna memberikan gambaran performa model secara seimbang melalui rata-rata harmonik antara presisi dan recall, sebagaimana didefinisikan pada persamaan (4) [19]:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

 ISSN:2302-8149
e-ISSN:2540-9719

4 Hasil dan Pembahasan

4.1 Hasil

Bab ini menguraikan secara komprehensif hasil eksperimen klasifikasi teks yang telah dilakukan beserta analisis teknis terhadap metrik evaluasi model. Rangkaian pembahasan pada bab ini disusun berdasarkan alur kronologis penelitian yang terbagi ke dalam empat tahapan utama: (1) pemaparan hasil pelatihan model awal (*baseline*) sebelum penanganan duplikasi data, (2) analisis mendalam terkait anomali *overfitting* dan temuan kebocoran data (*data leakage*), (3) evaluasi performa model akhir yang telah dilatih ulang menggunakan dataset terdeduplikasi, serta (4) unjuk kerja fungsionalitas sistem klasifikasi pada tahap implementasi akhir.

4.1.1 Karakteristik Dataset Baseline (Model v1)

Dataset awal (v1) yang digunakan pada tahap eksperimen dasar ini berjumlah 5.171 rekaman. Analisis distribusi label dilakukan untuk memastikan keseimbangan proporsi data antar kelas target sebelum masuk ke proses pelatihan. Hasil analisis menunjukkan bahwa dataset memiliki komposisi yang cukup berimbang, yakni terdiri dari 3,083 data (59.6%) pada kelas keluhan organik ("AMAN") dan 2,088 data (40.4%) pada kelas teks manipulatif ("CALO"). Sebaran kelas yang proporsional ini menjadi landasan awal yang krusial untuk memitigasi risiko bias model terhadap salah satu kelas mayoritas pada tahap pelatihan dasar. Karakteristik linguistik pada kelas target divisualisasikan menggunakan Word Cloud pada Gambar 3. Visualisasi ini memperlihatkan dominasi frekuensi kata-kata tertentu pada kelas calo.

Gambar 3 Word cloud model v1

Gambar 3 menunjukkan bahwa kata-kata pendekatan personal seperti "bantu", "saya", "bisa", dan "kak" tampak mendominasi secara absolut, berdampingan dengan kata kunci manipulatif dan transaksional seperti "inbox", "cair", "akun", "data", "jasa", dan "WA". Kepadatan (*density*) kata-kata spesifik ini menjadi indikasi awal tingginya kemunculan teks berulang (*template*) penawaran jasa dalam korpus data.

4.1.2 Evaluasi Performa Model v1

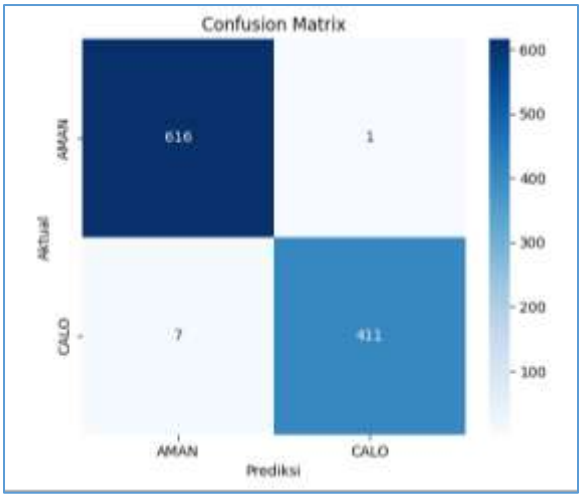
Pengujian performa dilakukan menggunakan 1,035 data uji (*test set*). Evaluasi metrik standar menunjukkan kemampuan klasifikasi yang sangat tinggi, dengan akurasi keseluruhan mencapai 99%, sebagaimana disajikan pada Tabel 5.

Tabel 5 Laporan klasifikasi (classification report) model v1

Kelas / Metrik	Precision	Recall	F1-Score	Support
Aman (Organik)	0.99	1.00	0.99	617
Calo (Spam)	1.00	0.98	0.99	418
Accuracy			0.99	1035
Macro Avg	0.99	0.99	0.99	1035
Weighted Avg	0.99	0.99	0.99	1035

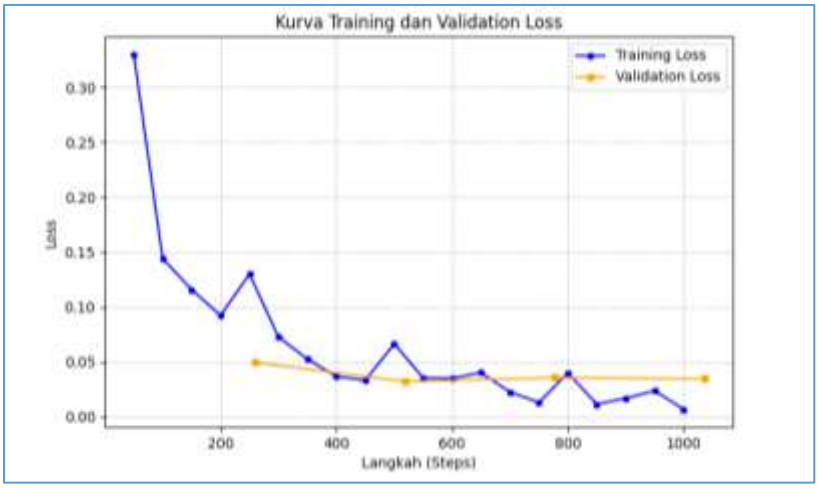
Tingkat presisi yang tinggi tersebut didukung oleh hasil matriks konfusi (*Confusion Matrix*) pada Gambar 4. Model IndoBERT-Lite terbukti hanya melakukan 7 kesalahan prediksi (gabungan *False Positive* dan *False Negative*) dari total 1,035 data pengujian.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 4 Confusion matrix prediksi model v1

Pemantauan terhadap riwayat pelatihan (training history) justru memunculkan pola yang saling bertolak belakang, meskipun secara kasat mata metrik akurasi dan matriks konfusi mengindikasikan model yang sangat handal. Pola tersebut terekam dalam kurva pergerakan *Training Loss* dan *Validation Loss* selama proses pelatihan berlangsung (Gambar 5).



Gambar 5 Kurva training loss dan validation loss model v1

Gambar 5 memperlihatkan anomali di mana Training Loss terus mengalami penurunan secara eksponensial mendekati nol, namun metrik Validation Loss justru mengalami stagnasi mendatar sejak tahap (step) ke-500. Stagnasi pada kurva validasi di tengah akurasi yang nyaris sempurna (99%) ini merupakan indikator akademis yang kuat terjadinya permasalahan fundamental pada distribusi data [25], yang memerlukan analisis komprehensif mengenai potensi overfitting pada subbab selanjutnya.

4.1.3 Analisis Overfitting & Kebocoran Data (Data Leakage)

Anomali yang terdeteksi pada kurva evaluasi Model v1 (Gambar 4) mengindikasikan adanya kendala fundamental pada proses pembelajaran model. Metrik *Training Loss* yang menurun secara konstan di bawah 0.01 menunjukkan bahwa model mampu menekan kesalahan prediksi pada data latih. Namun, tertahannya *Validation Loss* pada ambang ~0.03 sejak langkah iterasi ke-500 memperlihatkan bahwa model kehilangan kemampuan generalisasinya terhadap data baru. Tingginya akurasi yang kontras dengan stagnasi metrik validasi ini merupakan indikator kuat terjadinya *overfitting* dalam konteks pemrosesan bahasa alami (NLP) [25].

Analisis komputasional terhadap distribusi teks dalam korpus data latih dan uji dilakukan untuk mencari akar penyebab *overfitting* tersebut. Pengecekan kemiripan (*similarity check*) dieksekusi menggunakan skrip Python untuk mengevaluasi tumpang tindih rekaman (duplikasi) dalam dataset awal yang berjumlah 5,171 baris. Hasil pemindaian sistem mengungkap terjadinya *data leakage*

 ISSN:2302-8149
e-ISSN:2540-9719

berskala signifikan, di mana 545 baris data (10.5% dari populasi) teridentifikasi sebagai duplikat eksak.

Analisis lebih lanjut pada tingkat kelas menunjukkan karakteristik yang mendasari kebocoran data tersebut. Pada kelas Ham (Organik), tingkat duplikasi tercatat sebesar 14.3% (442 dari 3,083 data), yang mayoritas didominasi oleh teks pendek repetitif seperti *"Cek dm min"* dan *"Balas Inbox saya min"*. Sementara itu, pada kelas Spam (Calo), tingkat duplikasi absolut berada di angka 4,9% (103 dari 2.088 data).

Temuan paling kritis pada kelas Spam tidak terletak pada duplikasi eksak, melainkan pada tingginya rasio kemiripan antar teks (near-duplicate).. Karakteristik komentar calo di lapangan ternyata didominasi oleh penggunaan *template* promosi yang disalin-tempel (*copy-paste*). Beberapa pasangan teks spam terdeteksi memiliki tingkat kemiripan hingga 98%, sebagaimana direkapitulasi pada Tabel 6.

Tabel 6 Temuan pasangan teks spam dengan tingkat kemiripan tertinggi (top similarity)

Pasangan ID	Tingkat Kemiripan	Contoh Variasi Teks (Template)
1	98.3%	• "JASA PENCAIRAN BPJS ,ADM SETELAH CAIR !!! Wa : 089516150381 DATA LEN.." • "Saeful Anwar JASA PENCAIRAN BPJS ,ADM SETELAH CAIR !!! Wa : 0895161503..."
2	97.9%	• "JASA PENCAIRAN BPJS ,ADM SETELAH CAIR !!! Wa : 089516150381 DATA LEN..." • "Elang Trisnanda JASA PENCAIRAN BPJS ,ADM SETELAH CAIR !!! Wa : 0895161.."
3	94.9%	• "BPJS Ketenagakerjaan tolong cek inbox..." • "BPJS Ketenagakerjaan min tolong cek inbox..."

Tabel 6 menunjukkan bahwa perbedaan antar teks spam sering kali hanya terletak pada elemen minor, seperti penambahan nama akun (contoh: **[NAMA AKUN 1]**, **[NAMA AKUN 2]**) di awal kalimat promosi yang sama persis. Keberadaan ratusan *template* identik yang tersebar melintasi partisi *Training Set* dan *Testing Set* ini menyebabkan kebocoran informasi. Model IndoBERT-Lite beralih melakukan hafalan (*memorization*) terhadap *template* teks, alih-alih mengekstraksi fitur linguistik secara semantik.

Prapemrosesan korektif berupa deduplikasi ekstrem diterapkan berdasarkan luaran program pemindai tersebut sebagai langkah mitigasi. Sebanyak 545 data duplikat dihapus, menyisakan 4,626 sampel teks yang sepenuhnya unik (*unseen*). Dataset bersih (Dataset v2) inilah yang selanjutnya digunakan untuk melatih ulang arsitektur IndoBERT-Lite pada tahap evaluasi akhir.

4.1.4 Karakteristik Dataset Bersih (Model v2)

Arsitektur IndoBERT-Lite dilatih ulang menggunakan dataset v2 yang telah dibersihkan secara ekstrem setelah melalui tahap mitigasi kebocoran data. Evaluasi pada tahap ini bertujuan untuk membuktikan bahwa model mampu melakukan klasifikasi berdasarkan pemahaman semantik teks, bukan sekadar hafalan terhadap *template* repetitif.

Dataset v2 terdiri dari 4,626 rekaman unik, yang terbagi menjadi 2,641 data organik (57.1%) dan 1.985 data spam (42.9%). Pembagian dataset dilakukan secara konsisten dengan rasio 80:20, menghasilkan 3,700 data latih dan 926 data uji yang sepenuhnya terasing (*unseen*) satu sama lain. Kondisi ini menjamin bahwa setiap teks yang muncul pada tahap evaluasi belum pernah dilihat oleh model selama proses pelatihan.

4.1.5 Metrik Performa Model v2

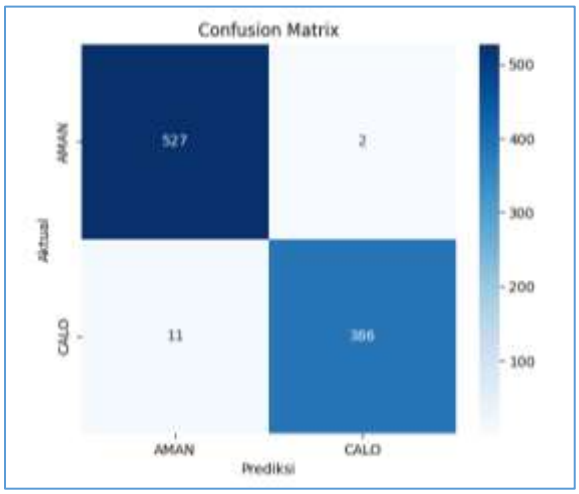
Pengujian pada Model v2 menghasilkan metrik evaluasi yang lebih kredibel secara metodologis. Hasil ini merepresentasikan performa model yang sebenarnya dalam menangani variasi teks yang unik, meskipun nilai akurasi secara nominal mengalami penurunan kecil dibandingkan Model v1. Laporan klasifikasi Model v2 disajikan pada Tabel 7.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 7 Laporan klasifikasi (classification report) model v2

Kelas / Metrik	Precision	Recall	F1-Score	Support
Aman (0)	0.98	0.99	0.98	530
Calo (1)	0.99	0.97	0.98	396
Accuracy			0.98	926
Macro Avg	0.98	0.98	0.98	926
Weighted Avg	0.98	0.98	0.98	926

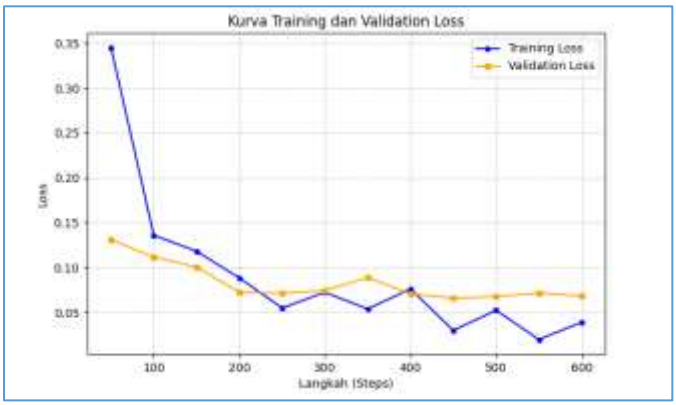
Model tetap mempertahankan performa tinggi dengan F1-Score sebesar 0.98 untuk kedua kelas berdasarkan Tabel 7. Hasil perhitungan ini diperoleh dengan menerapkan metrik evaluasi pada persamaan (4), yang membuktikan bahwa model memiliki keseimbangan yang sangat baik antara presisi dan daya ingat (*recall*). Analisis kesalahan prediksi secara lebih detail ditampilkan melalui *Confusion Matrix* pada Gambar 6.



Gambar 6 Confusion matrix prediksi model v2

Gambar 6 mencatat total 13 kesalahan prediksi dari 926 data uji, yang terdiri dari 4 kasus *False Positive* (organik terdeteksi calo) dan 9 kasus *False Negative* (calo terdeteksi organik). Peningkatan jumlah kesalahan dari 7 kasus (pada Model v1) menjadi 13 kasus (pada Model v2) merupakan konsekuensi logis dari penghapusan data duplikat, namun hal ini memberikan jaminan bahwa model bekerja melalui generalisasi fitur linguistik yang sehat.

Perbaikan paling signifikan terlihat pada dinamika riwayat pelatihan. Kurva loss pada Model v2 menunjukkan tren konvergensi yang sinkron antara data latih dan data validasi, berbeda dengan Model v1 yang mengalami stagnasi validasi sebagaimana diilustrasikan pada Gambar 7.



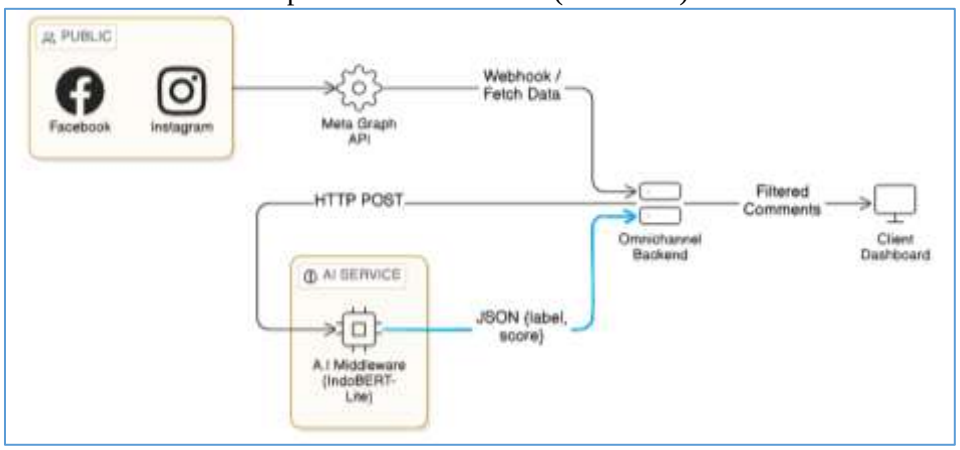
Gambar 7 Kurva training dan validation loss model v2

 ISSN:2302-8149
e-ISSN:2540-9719

Gambar 7 memperlihatkan bahwa *Validation Loss* terus menurun mengikuti pergerakan *Training Loss* hingga mencapai titik optimal. Pola kurva yang melandai secara bersamaan ini membuktikan bahwa risiko *overfitting* telah berhasil dimitigasi secara efektif. Model v2 menunjukkan kemampuan generalisasi yang stabil, sehingga bobot (weights) dari model ini siap untuk diekspor dan diimplementasikan pada sistem moderasi operasional instansi.

4.1.6 Implementasi Sistem Klasifikasi

Tahap implementasi merupakan pembuktian operasional untuk menunjukkan bahwa model IndoBERT-Lite yang telah dilatih dapat diintegrasikan ke dalam sistem moderasi komentar secara *real-time*. Fokus utama pada tahap ini adalah efisiensi penggunaan sumber daya dan kecepatan inferensi model. Model v2 diekspor dan diintegrasikan ke dalam middleware API. Frontend instansi mengirimkan data mentah melalui protokol HTTP POST (Gambar 8).



Gambar 8 Diagram arsitektur integrasi model IndoBERT-Lite

Pendekatan arsitektur *microservice* ini memisahkan inferensi AI dari antarmuka utama, memastikan stabilitas sistem tanpa mengganggu aliran data transaksional pada basis data utama [26].

4.1.7 Hasil Uji Inferensi dan Integrasi Sistem

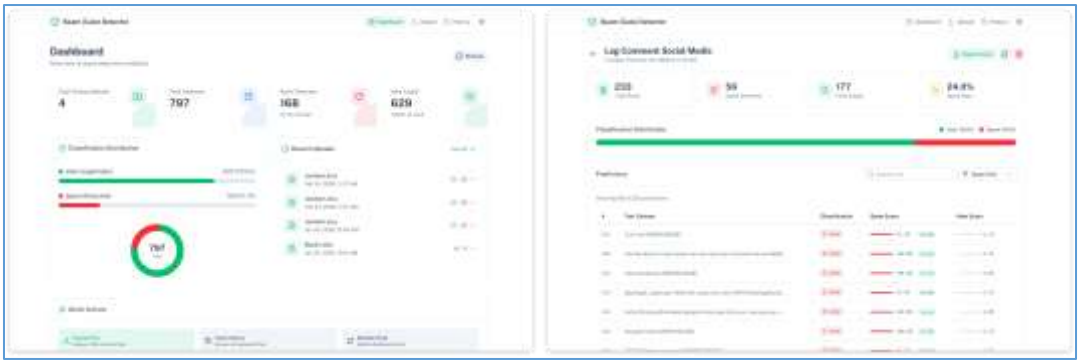
Pengujian dilakukan pada modul API *microservice* sesuai dengan tujuan penelitian untuk menghasilkan sistem yang responsif. Pengujian ini menggantikan analisis teknis perangkat keras dengan fokus pada waktu respons sistem. Hasil uji coba pengiriman data melalui protokol HTTP POST disajikan pada Tabel 8.

Tabel 8 Hasil uji waktu respons inferensi API

Uji Coba Ke-	Jumlah Karakter Teks	Waktu Pemrosesan (ms)	Status
1	45	54.2	Sukses
2	120	56.8	Sukses
3	88	55.5	Sukses
4	210	58.1	Sukses
5	150	57.3	Sukses
Rata-rata	122.6	56.3	-

Sistem juga divisualisasikan melalui antarmuka dasbor web (Gambar 9) yang memisahkan komentar organik dan spam secara dinamis, mengonfirmasi keandalan operasional model di lingkungan produksi.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 9 Dasbor antarmuka moderasi komentar berbasis IndoBERT-Lite

4.2 Pembahasan

Penelitian ini menunjukkan bahwa penggunaan arsitektur IndoBERT-Lite yang dikombinasikan dengan strategi deduplikasi ekstrem mampu menghasilkan model yang tangguh terhadap manipulasi teks spam calo di sektor asuransi ketenagakerjaan. Strategi pembersihan data identik terbukti efektif dalam memitigasi *overfitting* yang sering terjadi pada penelitian klasifikasi teks media sosial sebelumnya [7].

Sistem yang diusulkan dalam riset ini berhasil mencapai rata-rata waktu respons sebesar 56.3 ms, sebuah pencapaian yang berbeda dengan penelitian terdahulu yang seringkali mengabaikan aspek kecepatan implementasi. Kecepatan ini berada jauh di bawah ambang batas latensi *real-time* industri yaitu 100 ms, yang menunjukkan keunggulan efisiensi model IndoBERT-Lite dibandingkan model monolingual BERT standar yang digunakan pada penelitian lain [15], [27]. Temuan ini menegaskan bahwa kualitas data pelatihan (bebas duplikasi) memegang peranan yang lebih krusial daripada sekadar kuantitas data mentah untuk menjaga integritas hasil evaluasi model.

5 Kesimpulan

Optimasi IndoBERT-Lite melalui *fine-tuning* yang dibarengi deduplikasi ekstrem terbukti menjadi solusi efektif untuk moderasi teks spam informal. Penghapusan data identik berhasil mengembalikan kemampuan generalisasi model dan mencegah metrik evaluasi yang menyesatkan akibat kebocoran data. Sistem yang dibangun mencapai tingkat akurasi dan F1-Score 98%, serta membuktikan kesiapan implementasi via API *microservice* dengan latensi inferensi di bawah 60 milidetik dan konsumsi VRAM di bawah 1.5 GB. Alur kerja ini memberikan keseimbangan optimal antara akurasi prediksi dan efisiensi operasional, menjadikannya standar baru bagi pengembangan sistem pengamanan saluran komunikasi digital pada instansi pelayanan publik.

Referensi

[1] A. Elmahdy, H. A. Inan, and R. B. Sim, "*Privacy Leakage in Text Classification A Data Extraction Approach*," pp. 13–20, Jan. 2022, DOI: 10.18653/v1/2022.privatenlp-1.3.

[2] J.-S. Kim, H.-J. Lee, H. W. Lee, and S.-H. Choi, "*Advanced Analysis of Learning-based Spam Email Filtering Methods based on Feature Distribution Differences of Dataset*," *IEEE Access*, Vol. 12, pp. 167313–167323, Jan. 2024, DOI: 10.1109/access.2024.3495830.

[3] K. Kamdan, M. P. Anugrah, M. J. Almutaali, R. Ramdani, and I. L. Kharisma, "*Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments*," in *The 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society*, MDPI, Sep. 2025, p. 66. DOI: 10.3390/engproc2025107066.

[4] M. Isnaini, Y. Triana, and I. Afrita, "*Law Enforcement in Online Fraud Cases in the Jurisdiction of the Pekanbaru City Resort Police*," *JILPR J. Indones. Law Policy Rev.*, Vol. 7, No. 2, pp. 391–405, Feb. 2026, DOI: 10.56371/jirpl.v7i2.594.

[5] R. L. Mustofa and C. E. Widodo, "Analisis Sentimen berbasis Aspek pada Aplikasi Elektronik Survei Kepuasan Masyarakat (E-SKM) Jawa Tengah menggunakan IndoBERT".

[6] R. A. Supono and M. I. Imani, "Implementasi *Machine Learning* untuk Klasifikasi Email Spam menggunakan I*ndobert, Hugging Face Transfomers* dan *Streamlit*," *J. Sos. Teknol.*, Vol. 6, No. 1, pp. 420–440, Feb. 2026, DOI: 10.59188/jurnalsostech.v6i1.32659.
[7] F. Destryanto, P. Rizqiyah, and P. Sokibi, "*Sentiment Analysis of Public Response to the Free Nutritious Meal Program on Instagram using IndoBERT*," *J. Artif. Intell. Eng. Appl. JAIEA*, Vol. 5, No. 1, pp. 1924–1928, Oct. 2025, DOI: 10.59934/jaiea.v5i1.1755.
[8] V. E. Sidauruk and W. Herowati, "*Indobert-based Sentiment Analysis of Political Discourse on Platform X: The Case of Prabowo-Gibran Administration*," *J. Appl. Inform. Comput.*, Vol. 10, No. 1, pp. 673–683, Feb. 2026, DOI: 10.30871/jaic.v10i1.11586.
[9] S. Agustian, M. I. Syah, N. Fatiara, and R. Abdillah, "*New Directions in Text Classification Research: Maximizing The Performance of Sentiment Classification from Limited Data*," *ArXiv Cornell Univ.*, Jul. 2024, DOI: 10.48550/arxiv.2407.05627.
[10] D. Purwitasari, A. S. S. Ansyah, A. P. Kurniawan, and A. N. Kholifah, "*A Hybrid Method on Emotion Detection for Indonesian Tweets of COVID-19*," *J. RESTI Rekayasa Sist. dan Teknol. Inf.*, Vol. 7, No. 2, pp. 254–262, Mar. 2023, DOI: 10.29207/resti.v7i2.4816.
[11] R. Jeffmarvin, H. Dzaky, Y. Ardiyanto, A. D. Saputra, D. Irawan, and J. B. Ardianto, "Analisis Perbandingan: *SMOTE* dan *Undersampling* pada Klasifikasi Spam *Naïve Bayes*," *J. Inform. Interact. Technol.*, Vol. 2, No. 2, pp. 377–383, Aug. 2025, DOI: 10.63547/jiite.v2i2.92.
[12] S. Ni *et al.*, "*Training on the Benchmark is not All You Need*," *ArXiv Cornell Univ.*, Sep. 2024, DOI: 10.48550/arxiv.2409.01790.
[13] W. Elouataoui, "*AI-Driven Frameworks for Enhancing Data Quality in Big Data Ecosystems: Error_Detection, Correction, and Metadata Integration*," *ArXiv Cornell Univ.*, May 2024, DOI: 10.48550/arxiv.2405.03870.
[14] A. G. M. Meque, N. Hussain, G. Sidorov, and A. Gelbukh, "*Machine Learning-based Guilt Detection in Text*," *SCI. Rep.*, Vol. 13, No. 1, Jul. 2023, DOI: 10.1038/s41598-023-38171-0.
[15] K. F. H. Holle, D. N. Munna, and E. W. Ekaputri, "*Performance Evaluation of Transformer Models: Scratch, Bart, and Bert for News Document Summarization*," *J. Tek. Inform. Jutif*, Vol. 6, No. 2, pp. 787–802, Apr. 2025, DOI: 10.52436/1.jutif.2025.6.2.2534.
[16] M. A. Hadi and F. H. Fard, "*Evaluating Pre-Trained Models for User Feedback Analysis in Software Engineering: A Study on Classification of App-Reviews*," *Empir. Softw. Eng.*, Vol. 28, No. 4, May 2023, DOI: 10.1007/s10664-023-10314-x.
[17] A. T. Riadi, F. Indriani, M. I. Mazdadi, M. R. Faisal, and R. Herteno, "*Cross-Temporal Generalization of IndoBERT for Indonesian Hoax News Classification*," *J. Tek. Inform. Jutif*, Vol. 6, No. 5, pp. 5291–5304, Oct. 2025, DOI: 10.52436/1.jutif.2025.6.5.4757.
[18] N. K. Nissa and E. Yulianti, "*Multi-Label Text Classification of Indonesian Customer Reviews using Bidirectional Encoder Representations from Transformers Language Model*," *Int. J. Power Electron. Drive Syst. J. Electr. Comput. Eng.*, Vol. 13, No. 5, pp. 5641–5641, Jun. 2023, DOI: 10.11591/ijece.v13i5.pp5641-5652.
[19] D. Z. Abidin, L. Afuan, A. N. Toscany, and N. Nurhadi, "*A Comprehensive Benchmarking Pipeline for Transformer-based Sentiment Analysis using Cross-Validated Metrics*," *J. Tek. Inform. Jutif*, Vol. 6, No. 4, pp. 1797–1810, Aug. 2025, DOI: 10.52436/1.jutif.2025.6.4.4894.
[20] Wildan Amru Hidayat and V. R. S. Nastiti, "Perbandingan Kinerja *Pre-trained* IndoBERT-base dan IndoBERT-Lite pada Klasifikasi Sentimen Ulasan TikTok Tokopedia *Seller Center* dengan Model IndoBERT" *JSiI J. Sist. Inf.*, Vol. 11, No. 2, pp. 13–20, Sep. 2024, DOI: 10.30656/jsii.v11i2.9168.
[21] H. D. Jayanti and A. Rohman, "*Cyberbullying Detection in Indonesian TikTok Comments using IndoBERT with Fairness Evaluation*," *J. Inf. Syst. Inform.*, Vol. 8, No. 1, pp. 907–927, Mar. 2026, DOI: 10.63158/journalisi.v8i1.1448.
[22] M. A. Nur, N. Umar, Z. Feng, and H. Gani, "*Evaluation of IndoBERT and RoBERTa: Performance of Indonesian Language Transformer Models in Sentiment Classification*," *JIKO J. Inform. Dan Komput.*, Vol. 8, No. 2, pp. 121–127, Jul. 2025, DOI: 10.33387/jiko.v8i2.9988.
[23] M. H. Humaidi, S. Sutrisno, and P. W. Laksono, "*Implementation of Machine Learning for Text Classification using the Naive Bayes Algorithm in Academic Information Systems at Sebelas Maret University Indonesia*," *E3S Web Conf.*, Vol. 465, pp. 2048–2048, Jan. 2023, DOI: 10.1051/e3sconf/202346502048.

 ISSN:2302-8149
e-ISSN:2540-9719

[24] S. Beddar-Wiesing, A. Moallemy-Oureh, M. Kempkes, and J. M. Thomas, "*Absolute Evaluation Measures for Machine Learning: A Survey*," *ArXiv Cornell Univ.*, Jul. 2025, [Online]. Available: http://arxiv.org/abs/2507.03392

[25] S. Sunardi, A. Yudhana, and M. Fahmi, "*Improving Waste Classification using Convolutional Neural Networks: An Application of Machine Learning for Effective Environmental Management*," *Rev. Intell. Artif.*, Vol. 37, No. 4, pp. 845–855, Aug. 2023, DOI: 10.18280/ria.370404.

[26] A. Nuril Wahyuni, T. Listyorini, and E. Supriyati, "Implementasi Model IndoBERT dan mBERT untuk Deteksi Berita *Hoaks* berbasis *Web*," *JATI J. Mhs. Tek. Inform.*, Vol. 10, No. 2, pp. 2736–2743, Mar. 2026, DOI: 10.36040/jati.v10i2.17752.

[27] M. B. M. Amin *et al.*, "Deteksi Spam Berbahasa Indonesia berbasis Teks menggunakan Model Bert," *J. Teknol. Inf. Dan Ilmu Komput.*, Vol. 11, No. 6, pp. 1291–1302, Dec. 2024, DOI: 10.25126/jtiik.1168121.