Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November

OPEN ACCESS

https://doi.org/ 10.63447/jmt.v3i2.2197

Perbandingan Ketahanan Lintas-Domain Indobert dan SVM Pada Ulasan Game dan Media Sosial Berbahasa Indonesia

Dyo Arya Purnama [1], Alsella Meiriza [2*], Hardini Novianti [3], Rizka Dhini Kurnia [4], Ken Ditha Tania [5]

[1,2*,3,4,5] Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Sriwijaya, Kota Palembang, Provinsi Sumatera Selatan, Indonesia.

Corresponding Email: dyoaryap12@gmail.com [1]

Histori Artikel:
Dikirim 16 Oktober 2026; *Diterima dalam bentuk revisi* 20 Oktober 2026; *Diterima* 30 Oktober 2026; *Diterbitkan* 25 November 2026. Semua hak dilindungi oleh Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) STMIK Indonesia Banda Aceh.

Abstrak

Pengujian pengklasifikasi sentimen untuk ulasan aplikasi berbahasa Indonesia pada umumnya masih dibatasi pada aplikasi yang sama dengan data latihnya. Studi ini membandingkan support vector machine (SVM) linear berfitur TF-IDF dengan IndoBERT yang di-fine-tune. Kedua model dilatih pada ulasan Roblox, lalu diuji pada ulasan Roblox yang disisihkan serta ulasan Minecraft dan TikTok dari Google Play (17 Juli–10 Oktober 2026). Setiap aplikasi menyumbang 2.500 ulasan dengan kelas negatif dan positif yang seimbang, dan label diturunkan dari rating. Kinerja diukur dengan macro-F1 beserta interval kepercayaan bootstrap 95%, sedangkan jarak domain dinilai melalui lima ukuran. Pada Roblox, IndoBERT dan SVM tidak berbeda secara jelas (0,867 vs 0,850). Di lintas domain, IndoBERT unggul sekitar 0,065 dan mengalami penurunan yang sekitar 0,048 lebih kecil. TikTok berjarak lebih jauh dari Roblox pada setiap ukuran dan mengalami penurunan lebih besar; temuan ini sejalan dengan pengaruh jarak, tetapi belum membuktikannya. Pra-pelatihan adaptif domain (DAPT) memperkecil penurunan pada TikTok sebesar 0,020 (macro-F1 naik dari 0,794 menjadi 0,810), tetapi tidak berpengaruh pada Minecraft. Secara praktis, IndoBERT lebih aman dipakai oleh pengembang yang menerapkan pengklasifikasi pada aplikasi lain, sedangkan SVM memadai sebagai baseline murah dalam satu domain.

Kata Kunci: Analisis Sentimen; Pergeseran Domain; IndoBERT; Support Vector Machine; Ulasan Google Play; Pra-Pelatihan Adaptif Domain.

Abstract

Sentiment classifiers for Indonesian app reviews are rarely tested on any app other than the one that provided their training data. This study contrasts a linear support vector machine (SVM) with TF-IDF features and a fine-tuned IndoBERT. Both models were trained on Roblox reviews and then tested on held-out Roblox reviews as well as on Minecraft and TikTok reviews from Google Play. Each app contributed 2,500 class-balanced reviews, with labels derived from ratings. Performance was measured by macro-F1 with 95% bootstrap confidence intervals, and domain distance was assessed using five measures. On Roblox itself, IndoBERT and SVM showed no clear difference (0.867 vs 0.850). Across domains, IndoBERT scored roughly 0.065 higher and lost about 0.048 less. TikTok lay farther from Roblox on every distance measure and suffered larger drops, which is consistent with a distance effect. Domain-adaptive pretraining reduced the TikTok drop by 0.020 (macro-F1 rose from 0.794 to 0.810) but had no effect on Minecraft. In practical terms, IndoBERT is the safer option for developers who apply a classifier to other apps, while SVM is enough as a cheap in-domain baseline.

Keyword: Sentiment Analysis; Domain Shift; IndoBERT; Support Vector Machine; Google Play Reviews; Domain-Adaptive Pretraining.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November

OPEN ACCESS

https://doi.org/ 10.63447/jmt.v3i2.2197

1. Pendahuluan

Ulasan Google Play merupakan catatan yang kaya dan selalu diperbarui tentang bagaimana pengguna menilai suatu aplikasi, dan analisis sentimen banyak digunakan untuk membacanya dalam skala besar (Setiawan, 2024). Penelitian terbaru atas ulasan berbahasa Indonesia kian bertumpu pada IndoBERT, model bahasa Transformer yang dilatih dengan teks Indonesia. Sebagai contoh, IndoBERT mencapai akurasi 96% pada 9.310 ulasan dari tiga aplikasi kesehatan (Imaduddin *et al.*, 2023), dan akurasi 96% pada 9.000 ulasan MyTelkomsel setelah laju belajar dan jumlah epoch disetel (Zaidan *et al.*, 2024). Pada 2.672 ulasan DeepSeek, IndoBERT dengan augmentasi back-translation memperoleh akurasi 93,95%, sedangkan baseline Random Forest berfitur TF-IDF hanya 85,30% (Sunan *et al.*, 2026). Penelitian sejenis juga menyasar ulasan TikTok (Athyabunnaja *et al.*, 2026). Meski begitu, model klasik belum ditinggalkan. SVM berfitur TF-IDF mencapai akurasi 89,81% pada 535 ulasan Free Fire dan mengungguli Naive Bayes (Wahid *et al.*, 2025). Sebuah studi yang terbit di jurnal ini memadukan LinearSVC dengan pseudo-label dari IndoBERT pada tweet Coretax dan memperoleh macro-F1 0,7423 (Shollis & Sanjaya, 2026). Pada ulasan produk Tokopedia, LinearSVC juga unggul atas IndoBERT (macro-F1 0,5510 vs 0,5088), walaupun penulisnya menilai selisih itu terutama berasal dari perbedaan strategi pengambilan sampel (Zahra *et al.*, 2026). Pada perbandingan klasifikasi teks umum, LinearSVC (98%) dan BERT (97%) nyaris seimbang (Sabiri *et al.*, 2023). Hasil yang beragam ini menandakan bahwa model terbaik bergantung pada data dan kondisi pengujiannya.

Kebanyakan studi tersebut memecah satu sumber data secara acak, sehingga ulasan untuk pelatihan dan pengujian sama-sama berasal dari satu aplikasi. Padahal dalam penerapan sebenarnya, pengklasifikasi kerap dipakai pada aplikasi lain dengan pengguna, kosakata, dan keluhan yang berbeda. Pergeseran domain semacam ini dikenal sebagai penyebab menurunnya kinerja, dan cara untuk mengukur ketahanan model terhadapnya sudah pernah diusulkan (Calderon *et al.*, 2024). Untuk menekan penurunan tersebut, metode adaptasi memanfaatkan representasi internal yang kokoh (Rostami *et al.*, 2023), adapter ringan (Malik *et al.*, 2023), atau pra-pelatihan lanjutan pada teks target tak berlabel (Xie *et al.*, 2024). Tiga hal masih terbuka: apakah IndoBERT dan SVM berbeda ketahanannya pada ulasan aplikasi berbahasa Indonesia, bagaimana ketahanan itu berkaitan dengan jarak domain yang terukur, dan apakah adaptasi dengan teks tak berlabel memberi manfaat. Sepengetahuan penulis, studi-studi ulasan berbahasa Indonesia di atas menguji satu sumber data, sehingga ketahanan lintas aplikasi, kaitannya dengan jarak domain yang terukur, dan manfaat adaptasi tanpa label belum dibandingkan secara terkontrol untuk kedua model. Kesulitan lain ialah label ulasan kerap diturunkan dari rating bintang. Rating dan teks ulasan tidak selalu seiring, sebab bintang bisa mengabaikan detail yang tertulis dalam teks (Zuhir *et al.*, 2025); selain itu, derau label yang realistis menjadi masalah tersendiri bagi pengklasifikasi berbasis BERT (Agro & Aldarmaki, 2023). Karena itu, penelitian ini membandingkan SVM linear berfitur TF-IDF dengan IndoBERT yang di-fine-tune pada tiga aplikasi Google Play, yaitu Roblox, Minecraft, dan TikTok. Penelitian dipandu tiga pertanyaan. Pertama, bagaimana perbandingan kedua model pada data uji dalam domain dan lintas domain. Kedua, seberapa besar kinerja menurun ketika model yang dilatih pada Roblox diterapkan pada Minecraft dan TikTok, serta apakah penurunan itu konsisten dengan jarak domain yang terukur. Ketiga, apakah pra-pelatihan adaptif domain pada ulasan target tak berlabel dapat mengurangi penurunan tersebut. Kebaruan penelitian ini ada pada perbandingan terkontrol, dengan jendela waktu, ukuran sampel, dan keseimbangan kelas yang sama untuk setiap aplikasi, pelaporan ketidakpastian untuk setiap selisih, dan penjelasan eksplisit mengenai batas kesimpulan akibat label berbasis rating. Berbeda dengan studi terdahulu yang menguji satu sumber data, penelitian ini mengukur jarak domain dengan lima ukuran, menguji pra-pelatihan adaptif domain (DAPT) pada teks target tak berlabel, dan melaporkan interval kepercayaan untuk setiap selisih sehingga kesimpulan tidak bergantung pada skor tunggal.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2197

2. Metode Penelitian

2.1 Pengumpulan dan Pembersihan Data

Ulasan dikumpulkan dari halaman Google Play Roblox (com.roblox.client), Minecraft (com.mojang.minecraftpe), dan TikTok (com.ss.android.ugc.trill) dengan bahasa dan negara diatur ke Indonesia. Ketiga aplikasi memakai rentang waktu yang sama, yakni 17 Juli hingga 10 Oktober 2026. Tahap pembersihan data dirinci pada Tabel 1. Ulasan berating 3 disisihkan karena sentimennya ambigu. Teks yang tersisa dibersihkan, ulasan yang kurang dari empat kata dibuang, duplikat persis dihapus, dan hanya ulasan yang dikenali pustaka langdetect sebagai bahasa Indonesia yang dipertahankan. Mengenali bahasa pada teks media sosial Indonesia memang sulit karena penulis sering mencampur bahasa Indonesia, Jawa, dan Inggris (Hidayatullah *et al*., 2023); akibatnya, sebagian ulasan berbahasa Indonesia mungkin ikut tersaring, sementara sebagian ulasan berbahasa asing mungkin masih lolos. Ulasan berating 1 dan 2 dilabeli negatif, sedangkan yang berating 4 dan 5 dilabeli positif. Dari masing-masing kumpulan data yang telah dibersihkan, diambil secara acak 1.250 ulasan negatif dan 1.250 ulasan positif dengan seed tetap (42), sehingga tiap aplikasi memiliki 2.500 ulasan. Data akhir diringkas pada Tabel 2. Urutan pengolahannya adalah pembersihan, pengambilan sampel acak seimbang, pengurutan ulasan terpilih menurut waktu, lalu pembagian temporal; sisa ulasan bersih yang tidak terpilih disimpan sebagai teks tak berlabel untuk adaptasi domain.

Tabel 1. Jumlah ulasan setelah setiap langkah pembersihan

Langkah	Roblox	Minecraft	TikTok
Mentah (hasil scraping)	82.800	10.200	55.600
Dalam jendela waktu	82.797	10.072	55.529
Rating 3 dibuang	78.148	9.545	51.919
Tidak kosong setelah dibersihkan	77.551	9.459	51.375
Minimal 4 kata	54.746	6.923	31.845
Duplikat dihapus	52.495	6.848	31.380
Hanya bahasa Indonesia	44.956	5.743	27.165
Negatif dalam kumpulan	13.403	1.625	12.585
Positif dalam kumpulan	31.553	4.118	14.580

Tabel 2. Deskripsi data akhir

Butir	Roblox	Minecraft	TikTok
Ulasan (neg./pos.)	1.250/1.250	1.250/1.250	1.250/1.250
Median kata	13	13	11
Rerata kata	19,78	18,14	16,31
Latih/valid./uji	1.750/375/375	1.750/375/375	1.750/375/375
Teks tak berlabel tersisa	42.456	3.243	24.665

2.2 Model dan Pelatihan

Roblox berperan sebagai domain sumber. Dari 2.500 ulasannya, ulasan hasil sampel diurutkan menurut waktu terbit, lalu dibagi secara temporal tanpa pengacakan ulang: 1.750 ulasan untuk pelatihan, 375 untuk validasi, dan 375 ulasan paling baru untuk pengujian. Minecraft dan TikTok berperan sebagai domain target, dan keseluruhan 2.500 ulasan dari masing-masing aplikasi digunakan sebagai data uji lintas domain. Model yang dilatih pada Minecraft atau TikTok dengan ukuran pembagian yang sama hanya dibuat untuk menetapkan batas atas dalam domain bagi setiap target. Dengan demikian, ulasan latih selalu lebih lama daripada ulasan validasi dan uji; pengambilan sampel acak dilakukan sebelum pembagian dan hanya menentukan ulasan mana yang terpilih. Model pertama berupa SVM linear dengan fitur TF-IDF; konstanta regularisasi C dipilih dari 0,1; 0,3; 1; 3; dan 10 menurut macro-F1 pada data validasi. Model kedua berupa IndoBERT (indobenchmark/indobert-base-p1) yang di-fine-tune dengan panjang maksimum 128 token,

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2197

sampai 4 epoch, laju belajar 2×10^{-5}, dan ukuran batch 32. Epoch yang menghasilkan macro-F1 validasi terbaik dipilih, dan pelatihan diulang dengan seed 42, 43, dan 44. Karena SVM bersifat deterministik, model ini cukup dijalankan satu kali. SVM linear berfitur TF-IDF dipilih sebagai baseline klasik yang murah, dan IndoBERT sebagai model Transformer berbahasa Indonesia yang umum dipakai pada ulasan berbahasa Indonesia, sehingga keduanya mewakili dua pendekatan yang paling sering dibandingkan. SVM memakai LinearSVC dengan pembobotan kelas seimbang. Hiperparameter IndoBERT berada dalam rentang yang lazim dianjurkan untuk fine-tuning BERT (Devlin *et al.*, 2019), yaitu laju belajar 2×10^{-5} sampai 5×10^{-5}, batch 16 atau 32, dan 2 sampai 4 epoch, serta dipertahankan sama pada semua skenario agar perbandingan terkontrol; pengoptimal yang dipakai adalah AdamW dengan peluruhan bobot 0,01 dan pemanasan linear 10% dari langkah pelatihan. Penyetelan hiperparameter yang menyeluruh tidak dilakukan; hanya konstanta C pada SVM dan epoch terbaik pada IndoBERT yang dipilih menurut data validasi. Pra-pelatihan adaptif domain (DAPT) (Xie *et al.*, 2024) meneruskan pra-pelatihan IndoBERT pada ulasan tak berlabel dari domain target yang tidak mengandung data uji, yakni 3.243 teks untuk Minecraft dan 20.000 teks untuk TikTok. Model hasil adaptasi kemudian di-fine-tune pada Roblox dengan cara yang sama persis seperti sebelumnya. Pada tahap ini digunakan 2 epoch, batch 64, panjang maksimum 64 token, laju belajar 5×10^{-5}, dan pemanasan 6%; jumlah teks dibatasi paling banyak 20.000 per domain target.

2.3 Evaluasi dan Jarak Domain

Metrik utama yang dipakai adalah macro-F1, yakni rata-rata skor F1 kelas negatif dan kelas positif (Opitz, 2024). Penurunan kinerja akibat pemindahan ke domain target dirumuskan pada (1): suku pertama adalah macro-F1 pada data uji Roblox, sedangkan suku kedua adalah macro-F1 pada aplikasi target. Karena selisih antarmodel kerap kecil, setiap skor, selisih, dan penurunan disajikan bersama interval kepercayaan 95% yang diperoleh dari 1.000 kali resampling bootstrap atas ulasan uji (Du, 2025). Dalam setiap resampling, ulasan uji ditarik dengan pengembalian sebanyak data uji yang tersedia; macro-F1 IndoBERT dihitung untuk tiap seed dari ketiga seed, kemudian dirata-ratakan, sehingga intervalnya menggambarkan rata-rata ketiga seed. Selisih antarmodel dihitung pada resampel yang sama (berpasangan), sedangkan pada selisih penurunan, ulasan uji Roblox dan ulasan target diresampel secara terpisah. Batas bawah dan batas atas interval masing-masing adalah persentil ke-2,5 dan ke-97,5. Keputusan statistik didasarkan pada interval dua sisi ini; interval yang memuat nol dianggap tidak meyakinkan.

$$Penurunan = F1_{Roblox} - F1_{target} \qquad (1)$$

Jarak domain diukur melalui lima ukuran. Tingkat kata di luar kosakata (OOV) adalah proporsi token target yang tidak ada dalam kosakata latih Roblox, sementara indeks Jaccard menghitung tumpang tindih kosakata kedua domain. Tidak seperti ukuran lain, Jaccard adalah ukuran kemiripan, sehingga nilai yang lebih kecil menandakan domain yang lebih berbeda. Divergensi Jensen–Shannon (JSD) membandingkan sebaran kata antara kedua domain. Proxy A-distance (PAD) dihitung dari akurasi A sebuah pengklasifikasi yang dilatih untuk membedakan kedua domain, seperti pada (2), dan bernilai 0 bila kedua domain tidak dapat dibedakan. Ukuran kelima adalah jarak kosinus antara rata-rata embedding IndoBERT kedua domain sebelum fine-tuning. Garis dasar tiap ukuran diperoleh dari pembagian acak ulasan Roblox.

$$PAD = 2(2A - 1) \qquad (2)$$

Dilakukan tiga pemeriksaan tambahan. Pertama, model dilatih dan diuji hanya pada ulasan berating 1 dan 5 untuk melihat pengaruh rating yang ambigu. Kedua, skor dipisahkan berdasarkan panjang ulasan, yaitu 4–6, 7–12, dan 13 kata atau lebih. Ketiga, waktu pelatihan dan inferensi turut dicatat.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS

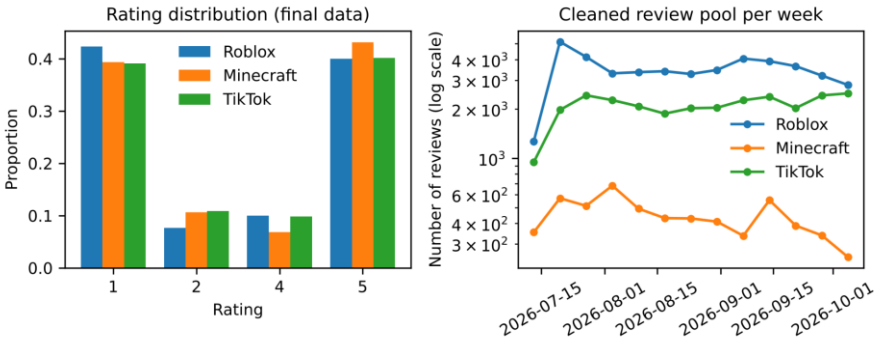
https://doi.org/ 10.63447/jmt.v3i2.2197

3. Hasil dan Pembahasan

3.1 Hasil

3.1.1 Gambaran Data

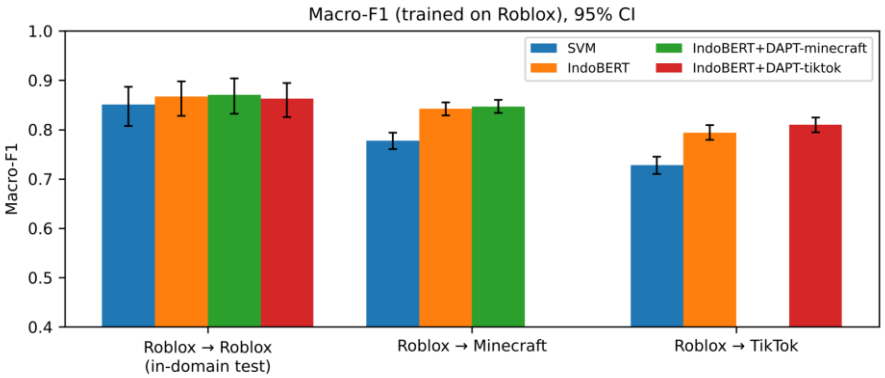
Tiap aplikasi menyumbang 2.500 ulasan yang seimbang, tetapi tahap pembersihan membentuk kumpulan datanya dengan cara berbeda. Filter empat kata membuang banyak ulasan positif, terutama pada TikTok, yang kumpulan positifnya menyusut dari 34.551 menjadi 17.567 ulasan (Tabel 1). Setelah penyeimbangan, rating 2 dan 4 sangat jarang: gabungan keduanya hanya 17,7% dari ulasan Roblox, 17,4% dari Minecraft, dan 20,8% dari TikTok, sehingga sampel didominasi rating 1 dan 5 (Gambar 1). Ulasan TikTok sedikit lebih pendek (median 11 kata) dibandingkan ulasan Roblox dan Minecraft (13 kata; Tabel 2). Setelah pengambilan sampel, kumpulan Minecraft hanya menyisakan 3.243 teks tak berlabel, jauh di bawah Roblox (42.456) maupun TikTok (24.665); hal ini penting bagi hasil adaptasi yang dibahas di bawah.



Gambar 1. Distribusi rating data akhir (kiri) dan jumlah ulasan bersih per minggu pada kumpulan tiap aplikasi (kanan)

3.1.2 IndoBERT versus SVM

Pada 375 ulasan uji Roblox, IndoBERT mencatat macro-F1 0,867 dan SVM 0,850 (Gambar 2). Menurut Tabel 3, selisih 0,016 ini tidak meyakinkan karena intervalnya melewati nol. Pada Minecraft skornya 0,842 dan 0,777, sedangkan pada TikTok 0,794 dan 0,728. Kedua selisih sekitar 0,065 dan intervalnya tidak mencakup nol. Akurasi menunjukkan urutan yang sama (IndoBERT 0,873; 0,844; dan 0,799; SVM 0,859; 0,784; dan 0,740). Di setiap skenario, kedua model memperoleh skor lebih tinggi pada ulasan negatif daripada ulasan positif. Simpangan baku IndoBERT di antara tiga seed berada pada kisaran 0,009 sampai 0,016.



Gambar 2. Macro-F1 setiap model yang dilatih pada Roblox, dengan interval kepercayaan 95%

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2197

Tabel 3. Selisih macro-F1 antara IndoBERT dan SVM

Skenario uji	Selisih	CI 95%
Roblox → Roblox	+0,016	[−0,015; +0,048]
Roblox → Minecraft	+0,065	[+0,052; +0,077]
Roblox → TikTok	+0,066	[+0,053; +0,080]

Studi terdahulu yang mendapati SVM linear setara atau lebih baik daripada model Transformer bertumpu pada satu sumber data (Zahra *et al*., 2026), (Sabiri *et al*., 2023). Hasil dalam domain pada penelitian ini sejalan dengan studi tersebut, sebab kedua model tidak berbeda jelas pada Roblox. Keunggulan IndoBERT baru tampak ketika domain berganti. Kata yang paling khas pada ulasan Roblox antara lain *harga*, *mahal*, *chat*, dan *item*, pada ulasan Minecraft antara lain *login*, *bayar*, *akun*, dan *bug*, dan pada ulasan TikTok antara lain *aplikasi*, *fitur*, *iklan*, dan *video*. Model TF-IDF hanya mempelajari bobot untuk kata-kata yang muncul saat pelatihan, dan ini menjadi alasan yang masuk akal bagi penurunannya yang lebih besar; model bahasa terlatih, sebaliknya, mampu merepresentasikan kata yang tidak muncul saat fine-tuning. Penjelasan ini tidak diuji secara langsung di sini. Perlu dibedakan antara temuan dan dugaan. Temuan eksperimen adalah selisih macro-F1 dan penurunan kinerja beserta intervalnya pada Tabel 3 dan Tabel 4. Alasan keunggulan IndoBERT, yaitu kemampuannya merepresentasikan kata yang tidak muncul saat pelatihan, belum diuji (misalnya dengan memisahkan kinerja pada ulasan yang banyak memuat kata di luar kosakata latih), sehingga statusnya masih dugaan

3.1.3 Penurunan Kinerja dan Jarak Domain

Kinerja kedua model menurun saat dipindahkan ke aplikasi lain (Tabel 4). Penurunan SVM mencapai 0,073 pada Minecraft dan 0,123 pada TikTok. Pada IndoBERT, penurunannya 0,025 di Minecraft, dengan interval yang mencakup nol, dan 0,073 di TikTok. Selisih penurunan SVM terhadap IndoBERT adalah 0,048 pada Minecraft dan 0,049 pada TikTok; karena kedua interval tidak mencakup nol, IndoBERT dapat dikatakan kehilangan lebih sedikit. Bila dibandingkan dengan model yang dilatih langsung pada domain target, SVM berbasis Roblox tertinggal 0,083 dari batas atas pada Minecraft dan 0,061 pada TikTok, sedangkan IndoBERT berbasis Roblox tertinggal 0,038 pada Minecraft dan setara pada TikTok (−0,003).

Tabel 4. Penurunan macro-F1 dari data uji Roblox ke setiap target Penurunan dihitung terhadap macro-F1 Roblox model yang sama

Target	Model	Penurunan	CI 95%
Minecraft	SVM	0,073	[0,030; 0,111]
Minecraft	IndoBERT	0,025	[−0,012; 0,058]
Minecraft	IndoBERT + DAPT	0,024	[−0,011; 0,058]
Minecraft	SVM − IndoBERT	0,048	[0,013; 0,083]
TikTok	SVM	0,123	[0,081; 0,161]
TikTok	IndoBERT	0,073	[0,037; 0,108]
TikTok	IndoBERT + DAPT	0,053	[0,018; 0,087]
TikTok	SVM − IndoBERT	0,049	[0,013; 0,085]
Minecraft	IndoBERT − IndoBERT + DAPT	0,000	[−0,015; 0,014]
TikTok	IndoBERT − IndoBERT + DAPT	0,020	[0,004; 0,037]

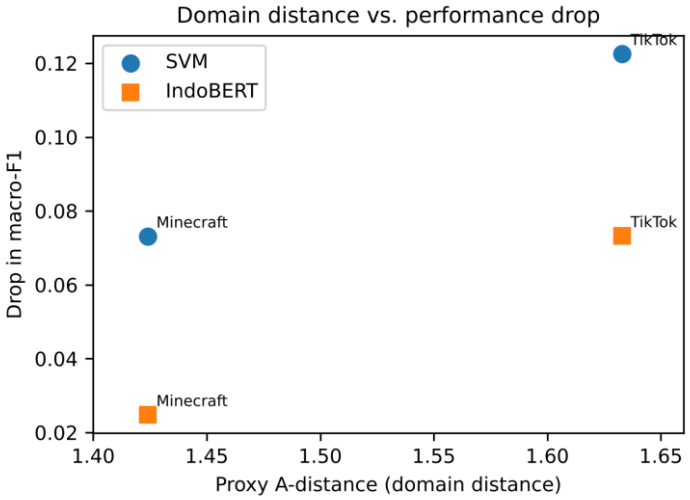
Pada setiap ukuran jarak (Tabel 5), TikTok berada lebih jauh dari Roblox dibandingkan Minecraft, meskipun dua nilai Jaccard-nya hampir sama (0,449 dan 0,444; Jaccard menunjukkan kemiripan, sehingga nilai lebih kecil berarti lebih jauh). Pengklasifikasi domain memisahkan aplikasi dengan akurasi 0,856 untuk Minecraft dan 0,908 untuk TikTok, sementara akurasi garis dasar untuk dua bagian acak Roblox adalah 0,5. Dengan demikian, penurunan yang lebih besar pada TikTok konsisten dengan pengaruh jarak (Gambar 3). Tambahan penurunan dari Minecraft ke TikTok hampir sama pada kedua model (SVM 0,050; IndoBERT 0,048), sehingga keunggulan IndoBERT

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2197

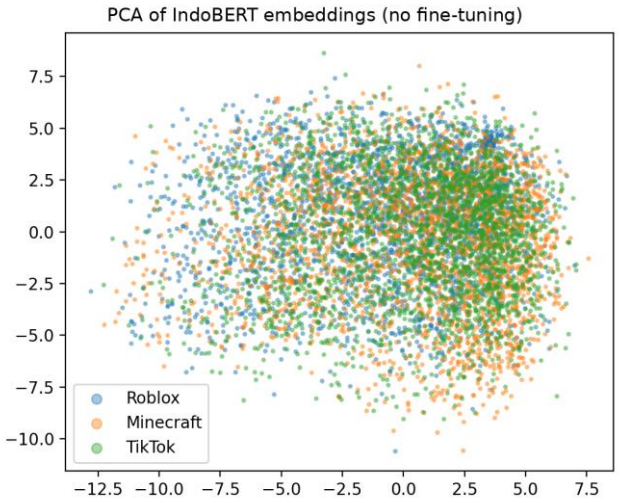
terlihat relatif tetap dan tidak membesar seiring bertambahnya jarak. Akan tetapi, karena hanya ada dua domain target, korelasi tidak dapat dihitung dan tidak ada klaim sebab-akibat yang diajukan. Embedding IndoBERT sebelum fine-tuning saling tumpang tindih kuat pada ketiga aplikasi (Gambar 4), sehingga domain-domain tersebut berbeda dalam derajat, bukan terpisah secara tegas.

Tabel 5. Jarak setiap domain target dari Roblox

Ukuran	Split Roblox	Minecraft	TikTok
Tingkat OOV	0,114	0,173	0,202
Indeks Jaccard (kemiripan)	–	0,449	0,444
JSD	0,104	0,216	0,253
PAD	0	1,42	1,63
Jarak embedding	0,0006	0,0195	0,0325



Gambar 3. Proxy A-distance setiap domain target terhadap penurunan macro-F1



Gambar 4. PCA embedding IndoBERT sebelum fine-tuning untuk ketiga aplikasi

Penurunan pada TikTok tidak semestinya dibaca sebagai pergeseran domain murni. IndoBERT yang dilatih dan diuji pada TikTok mencetak 0,778, nilai dalam domain terendah, sementara IndoBERT yang dilatih pada Roblox memperoleh 0,781 pada partisi uji TikTok yang sama. Karena

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2197

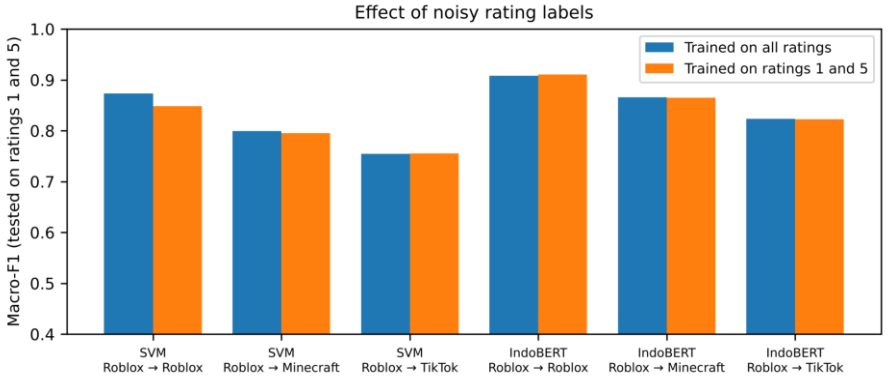
itu, sebagian penurunan pada TikTok mencerminkan rating yang lebih berderau atau ulasan uji yang lebih sulit. Sebaliknya, model yang dilatih pada TikTok memperoleh 0,862 pada Minecraft. Data uji Roblox dalam domain pun hanya terdiri atas 375 ulasan, sehingga interval penurunannya menjadi lebar.

3.1.4 Pra-pelatihan Adaptif Domain

DAPT hanya memperbaiki IndoBERT pada TikTok. Macro-F1 naik dari 0,794 menjadi 0,810, dan penurunan berkurang dari 0,073 menjadi 0,053, yakni pengurangan 0,020 (CI 95% 0,004–0,037); angka ini kecil, tetapi intervalnya tidak mencakup nol. Pengurangan ini sedikit lebih besar daripada kenaikan macro-F1 di TikTok (0,016) karena penurunan dihitung terhadap skor Roblox model yang sama, dan skor Roblox model DAPT-TikTok (0,863) lebih rendah daripada IndoBERT tanpa adaptasi (0,867). Pada Minecraft skor bergeser dari 0,842 ke 0,846 dengan pengurangan 0,000 (CI 95% −0,015–0,014). DAPT tidak merugikan kinerja Roblox dalam domain (0,871 dan 0,863 dibandingkan 0,867). Penyebab yang mungkin bagi tidak adanya efek pada Minecraft terletak pada jumlah dan komposisi teks adaptasinya. Teks Minecraft tak berlabel yang tersisa hanya 3.243, dan karena 1.250 ulasan negatif telah terpakai pada data akhir, sisanya sangat condong ke ulasan positif; TikTok, sebaliknya, menyediakan 20.000 teks. Hal ini sejalan dengan studi pra-pelatihan lanjutan yang menekankan pentingnya jumlah dan pemilihan teks domain (Xie *et al.*, 2024), tetapi statusnya penjelasan, bukan hasil yang diuji. Alternatif yang lebih ringan, seperti adapter (Malik *et al.*, 2023) atau adaptasi berbasis representasi (Rostami *et al.*, 2023), tidak diperiksa dan diserahkan kepada penelitian berikutnya.

3.1.5 Derau Label, Panjang Ulasan, dan Biaya

Melatih model hanya dengan ulasan berating 1 dan 5 hampir tidak mengubah hasil (Gambar 5). Perubahan yang jelas hanya satu, yaitu penurunan 0,025 pada SVM Roblox dalam domain (CI 95% −0,044 sampai −0,005). Sebaliknya, menguji hanya pada ulasan berating 1 dan 5 menaikkan skor; contohnya IndoBERT pada Roblox naik dari 0,867 menjadi 0,908. Dengan demikian, sebagian galat bersumber dari label uji yang ambigu, bukan dari model. Hal ini sejalan dengan laporan bahwa rating dan teks ulasan bisa tidak selaras (Zuhir *et al.*, 2025) dan bahwa derau label yang realistis berdampak pada pengklasifikasi berbasis BERT (Agro & Aldarmaki, 2023). Oleh sebab itu, seluruh skor dalam naskah ini mengukur kesesuaian dengan rating, bukan sentimen yang sebenarnya.



Gambar 5. Macro-F1 saat dilatih dengan semua rating atau hanya rating 1 dan 5, diuji pada rating 1 dan 5

3.1.6 Keterbatasan dan Implikasi

Ada beberapa keterbatasan. Label diturunkan dari rating, bukan dari anotasi manual, dan arti sebuah rating bisa berbeda antaraplikasi. Penelitian ini hanya memakai tiga seed dan satu sampel acak, sehingga klaimnya terbatas pada konfigurasi ini. Filter empat kata dan filter bahasa turut membentuk populasi data, dan data uji Roblox berukuran kecil. Dua domain target tidak cukup

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2197

untuk mendukung uji statistik terhadap pengaruh jarak, dan kumpulan adaptasi Minecraft juga kecil serta condong. Secara praktis, hasil ini menyarankan IndoBERT sebagai pilihan yang lebih aman bila pengklasifikasi sentimen akan menghadapi ulasan dari aplikasi lain, sementara SVM tetap berguna sebagai baseline murah dalam satu domain. Adaptasi dengan teks tak berlabel layak dicoba jika tersedia banyak teks target yang seimbang, tetapi hasilnya sebaiknya hanya diharapkan berupa perbaikan kecil. Dengan demikian, yang merupakan temuan empiris penelitian ini adalah selisih macro-F1 dan penurunan kinerja beserta interval kepercayaannya, sedangkan penjelasan mengenai penyebabnya masih berupa dugaan yang belum diuji. Karena alasan keunggulan IndoBERT terhadap derau label lintas-domain dalam pengujian ini masih berupa dugaan awal, penelitian mendatang disarankan untuk merancang eksperimen faktorial yang secara spesifik mampu mengisolasi pengaruh masing-masing metrik jarak domain (seperti JSD dan Jaccard) terhadap degradasi performa klasifikasi.

3.2 Pembahasan

Hasil dalam domain menunjukkan bahwa IndoBERT dan SVM linear berfitur TF-IDF tidak berbeda jelas pada ulasan Roblox, dengan macro-F1 0,867 dan 0,850. Selisih sebesar 0,016 ini masih berada dalam rentang ketidakpastian karena interval kepercayaannya melewati nol. Temuan ini sejalan dengan Zahra dkk. (2026), yang mendapati LinearSVC unggul atas IndoBERT pada ulasan produk Tokopedia, dan dengan Sabiri dkk. (2023), yang melaporkan LinearSVC (98%) dan BERT (97%) yang nyaris seimbang pada klasifikasi teks umum. Kedua studi tersebut, seperti penelitian ini, menunjukkan bahwa keunggulan model Transformer tidak selalu muncul ketika data latih dan data uji berasal dari sumber yang sama. Hasil ini juga mendukung Shollis dan Sanjaya (2026), yang menilai bahwa performa model klasik sangat bergantung pada data dan kondisi pengujiannya. Keunggulan IndoBERT baru tampak ketika model diterapkan pada aplikasi lain. Pada skenario lintas domain, IndoBERT unggul sekitar 0,065 macro-F1 atas SVM di Minecraft dan TikTok, dan penurunan kinerjanya lebih kecil sekitar 0,048. Pola ini sejalan dengan Imaduddin dkk. (2023) dan Sunan dkk. (2026), yang sama-sama memperlihatkan keunggulan IndoBERT atas baseline klasik pada ulasan berbahasa Indonesia, meskipun kedua studi itu menguji dalam satu sumber data. Dengan demikian, hasil penelitian ini menambah bukti bahwa keunggulan IndoBERT mungkin lebih terlihat pada kondisi pergeseran domain dibandingkan pada pengujian dalam domain. Pergeseran domain sendiri telah diakui sebagai penyebab utama menurunnya kinerja model, dan Calderon dkk. (2024) mengusulkan cara mengukur ketahanan model terhadapnya.

Kemungkinan penjelasan atas perbedaan ketahanan ini berkaitan dengan cara kedua model merepresentasikan kosakata. Model TF-IDF hanya mempelajari bobot untuk kata yang muncul saat pelatihan, sedangkan model bahasa terlatih dapat merepresentasikan kata baru melalui konteksnya. Dugaan ini masuk akal mengingat ulasan TikTok memiliki tingkat kata di luar kosakata yang lebih tinggi (0,202) dibandingkan Minecraft (0,173). Namun, penelitian ini belum menguji dugaan tersebut secara langsung, misalnya dengan memisahkan kinerja pada ulasan yang banyak memuat kata di luar kosakata latih. Oleh karena itu, penjelasan ini perlu diperlakukan sebagai hipotesis untuk penelitian berikutnya. Pola penurunan pada TikTok juga perlu dibaca dengan hati-hati. TikTok berada lebih jauh dari Roblox pada kelima ukuran jarak domain, dan penurunan kinerja pada kedua model memang lebih besar di sana. Hal ini konsisten dengan pengaruh jarak domain, tetapi hanya dua domain target yang diuji sehingga hubungan tersebut belum dapat diuji secara statistik. Selain itu, IndoBERT yang dilatih dan diuji pada TikTok hanya mencapai 0,778, hampir sama dengan model Roblox yang memperoleh 0,781 pada partisi uji TikTok yang sama. Artinya, sebagian penurunan mencerminkan ulasan TikTok yang lebih sulit atau rating yang lebih berderau, bukan semata pergeseran domain. Temuan ini sejalan dengan Zuhir dkk. (2025), yang melaporkan bahwa rating bintang tidak selalu selaras dengan sentimen dalam teks ulasan, serta Athyabunnaja dkk. (2026), yang meneliti emosi pada ulasan TikTok dan menunjukkan bahwa karakteristik teks media sosial memang berbeda dari ulasan aplikasi lain. Hasil pada pra-pelatihan adaptif domain (DAPT) memberikan gambaran yang lebih terbatas. DAPT mengurangi penurunan pada TikTok sebesar 0,020 dan menaikkan macro-F1 dari 0,794 menjadi 0,810, tetapi tidak berpengaruh pada Minecraft.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2197

Perbedaan ini kemungkinan besar berkaitan dengan jumlah dan komposisi teks adaptasi. Teks Minecraft tak berlabel yang tersisa hanya 3.243, dan sisanya condong ke ulasan positif, sedangkan teks TikTok berjumlah 20.000. Xie dkk. (2024) menekankan pentingnya jumlah dan pemilihan teks domain dalam pra-pelatihan lanjutan, sehingga temuan ini sejalan dengan argumen tersebut. Meski demikian, penjelasan ini belum diuji. Alternatif yang lebih ringan, seperti adapter yang diusulkan Malik dkk. (2023) atau adaptasi berbasis representasi internal yang kokoh dari Rostami dkk. (2023), belum diperiksa dalam penelitian ini dan layak menjadi arah penelitian selanjutnya. Pemeriksaan tambahan terhadap label juga memberikan implikasi penting. Ketika model hanya dilatih dan diuji pada ulasan berating 1 dan 5, skor naik cukup banyak, misalnya IndoBERT pada Roblox naik dari 0,867 menjadi 0,908. Hal ini menunjukkan bahwa sebagian galat berasal dari label uji yang ambigu, bukan semata dari model. Temuan ini konsisten dengan Agro dan Aldarmaki (2023), yang menunjukkan bahwa derau label yang realistis berdampak pada pengklasifikasi berbasis BERT. Oleh karena itu, seluruh skor dalam penelitian ini sebaiknya dimaknai sebagai tingkat kesesuaian dengan rating, bukan sentimen yang sebenarnya. Kehati-hatian ini juga sejalan dengan Zaidan dkk. (2024), yang menunjukkan bahwa akurasi IndoBERT pada ulasan MyTelkomsel sangat dipengaruhi oleh penyetelan laju belajar dan jumlah epoch, sehingga perbandingan antarstudi perlu mempertimbangkan kondisi pelatihan masing-masing. Secara praktis, hasil ini mendukung penggunaan IndoBERT sebagai pilihan yang lebih aman bagi pengembang yang akan menerapkan pengklasifikasi pada aplikasi lain, sedangkan SVM tetap memadai sebagai baseline yang murah dalam satu domain. Wahid dkk. (2025) juga menunjukkan bahwa SVM dengan TF-IDF dapat memberikan hasil yang cukup baik pada ulasan berbahasa Indonesia dengan data terbatas, sehingga pilihan model dapat menyesuaikan kebutuhan dan sumber daya yang tersedia. Meskipun demikian, penelitian ini memiliki keterbatasan, seperti penggunaan tiga seed, satu sampel acak, dan hanya dua domain target. Karena itu, kesimpulan yang dihasilkan sebaiknya dipahami sebagai temuan awal yang perlu diperluas dengan lebih banyak aplikasi, data uji beranotasi manual, dan eksperimen yang secara spesifik mengisolasi pengaruh tiap ukuran jarak domain.

4. Kesimpulan

Pada data dalam domain Roblox, IndoBERT tidak jelas lebih baik daripada SVM (0,867 vs 0,850). Namun, saat dilatih pada Roblox dan diuji pada Minecraft atau TikTok, IndoBERT unggul sekitar 0,065 macro-F1 dan mengalami penurunan sekitar 0,048 lebih kecil, sehingga lebih tahan terhadap pergeseran domain pada kondisi ini. TikTok, yang lebih jauh dari Roblox pada kelima ukuran jarak, mengalami penurunan lebih besar pada kedua model; hal ini konsisten dengan pengaruh jarak, tetapi dua domain target belum cukup untuk memastikannya. Pra-pelatihan adaptif domain mengurangi penurunan di TikTok sebesar 0,020 dan tidak berefek di Minecraft, yang teks tak berlabelnya sedikit dan condong ke ulasan positif. Skor mengukur kesesuaian dengan rating bintang, bukan sentimen yang sebenarnya; hasilnya berasal dari tiga seed dan satu sampel; dan sebagian penurunan pada TikTok mencerminkan label yang lebih berderau. Penelitian lanjutan disarankan untuk (1) menambah domain target agar hubungan antara jarak dan kinerja dapat diuji secara statistik, (2) memakai data uji beranotasi manual, (3) memperbanyak dan menyeimbangkan teks tak berlabel untuk adaptasi, serta (4) menguji bentuk adaptasi yang lebih ringan seperti adapter.

5. Ucapan Terima Kasih

Penulis berterima kasih kepada Google Colaboratory atas lingkungan komputasi awan yang dipakai untuk menjalankan seluruh eksperimen. Ucapan terima kasih juga ditujukan kepada Google Play Store yang menjadi sumber ulasan pengguna sebagai data penelitian, serta kepada komunitas sumber terbuka, terutama Hugging Face dan scikit-learn, atas model dan pustaka yang digunakan. Penelitian ini dikerjakan tanpa pendanaan dari lembaga maupun donatur.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November

OPEN ACCESS

https://doi.org/ 10.63447/jmt.v3i2.2197

6. Daftar Pustaka

Agro, M. T., & Aldarmaki, H. (2023). Handling realistic label noise in BERT text classification. *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP)*, 11–20. https://doi.org/10.48550/arXiv.2305.16337.

Athyabunnaja, S., Tamrin, T., & Mulyo, H. (2026). Implementation of IndoBERT for text-based emotion classification on TikTok application reviews in Google Play Store. *Journal Dinda: Data Science, Information Technology, and Data Analytics, 6*(1), 18–24. https://doi.org/10.20895/dinda.v6i1.10174.

Calderon, N., Porat, N., Ben-David, E., Chapanin, A., Gekhman, Z., Oved, N., Shalumov, V., & Reichart, R. (2024). Measuring the robustness of NLP models to domain shifts. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 126–154. https://doi.org/10.18653/v1/2024.findings-emnlp.7.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Du, W. (2025). *When +1% is not enough: A paired bootstrap protocol for evaluating small improvements* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2511.19794.

Hidayatullah, A. F., Apong, R. A., Lai, D. T. C., & Qazi, A. (2023). Corpus creation and language identification for code-mixed Indonesian-Javanese-English tweets. *PeerJ Computer Science, 9*, Article e1312. https://doi.org/10.7717/peerj-cs.1312.

Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment analysis in Indonesian healthcare applications using IndoBERT approach. *International Journal of Advanced Computer Science and Applications, 14*(8). https://doi.org/10.14569/IJACSA.2023.0140813.

Malik, B., Kashyap, A. R., Kan, M.-Y., & Poria, S. (2023). UDAPTER – Efficient domain adaptation using adapters. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2249–2263. https://doi.org/10.18653/v1/2023.eacl-main.165.

Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics, 12*, 820–836. https://doi.org/10.1162/tacl_a_00675.

Rostami, M., Bose, D., Narayanan, S., & Galstyan, A. (2023). Domain adaptation for sentiment analysis using robust internal representations. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11484–11498. https://doi.org/10.18653/v1/2023.findings-emnlp.769.

Sabiri, B., Khtira, A., El Asri, B., & Rhanoui, M. (2023). Analyzing BERT's performance compared to traditional text classification models. *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS), 1*, 572–582. https://doi.org/10.5220/0011983100003467.

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2197

Setiawan, B. (2024). A review of sentiment analysis applications in Indonesia between 2023–2024. *JIEET (Journal of Information Engineering and Educational Technology), 8*(2), 71–83. https://doi.org/10.26740/jieet.v8n2.p71-83.

Shollis, R. F., & Sanjaya, U. P. (2026). Comparison of LinearSVC and complement naive Bayes on Coretax sentiment analysis using IndoBERT pseudo-labeling and ADASYN. *JIKO (Jurnal Informatika dan Komputer), 9*(2). https://doi.org/10.33387/jiko.v9i2.11615.

Sunan, M., Resiloy, U. D. A., Endriani, D., Suhaeni, C., Sartono, B., & Dito, G. A. (2026). IndoBERT optimization for sentiment analysis on DeepSeek app reviews. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 20*(1), 95–106. https://doi.org/10.22146/ijccs.107507.

Wahid, Y. A., Sanatang, & Andayani, D. D. (2025). Performance comparison of SVM and Naïve Bayes for Indonesian-language sentiment analysis on Free Fire reviews using TF-IDF and SMOTE. *JESSI, 6*(4). https://doi.org/10.59562/jessi.v6i4.10818.

Xie, Y., Aggarwal, K., & Ahmad, A. (2024). Efficient continual pre-training for building domain specific large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, 10184–10201. https://doi.org/10.18653/v1/2024.findings-acl.606.

Zahra, N. Z., Farhanatussaidah, S., Afifah, N. N., Muthoharoh, L., Satria, A., & Manullang, M. C. T. (2026). Benchmarking logistic regression, SVM, naive Bayes, and IndoBERT fine-tuning for sentiment analysis on Indonesian product reviews [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.03439.

Zaidan, M. N., Sibaroni, Y., & Prasetiyowati, S. S. (2024). Learning rate and epoch optimization in the fine-tuning process for IndoBERT's performance on sentiment analysis of MyTelkomsel app reviews. *Jurnal Teknik Informatika (JUTIF), 5*(5), 1443–1450. https://doi.org/10.52436/1.jutif.2024.5.5.2396.

Zuhir, N., Salim, A. M., Premkumar, P., & Farazi, M. (2025). Beyond stars: Bridging the gap between ratings and review sentiment with LLM [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2509.20953.