JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analisis Komparatif IndoBERT, FinBERT, dan IndoLEM dengan Fine-Tuning untuk Klasifikasi Sentimen Berita Finansial Indonesia

Naufal Latif Ramadhan*,Majid Rahardi

Informatika, Teknologi Informasi, Universitas AMIKOM, Yogyakarta, Indonesia
Email: [1,*]naufalatif7114@gmail.com, majid@amikom.ac.id
Email Penulis Korespondensi: naufalatif7114@gmail.com*
Submitted **04-06-2026**; Accepted **02-10-2026**; Published **31-10-2026**

Abstrak

Analisis sentimen berita finansial memiliki peran penting dalam mendukung pengambilan keputusan investasi di pasar modal Indonesia. Penelitian terdahulu umumnya membandingkan satu model tanpa perbandingan lintas model, sehingga rekomendasi pemilihan model masih bersifat parsial. Penelitian ini membandingkan tiga model bahasa *pre-trained*, yaitu IndoBERT, IndoLEM, dan FinBERT, melalui pendekatan *fine-tuning* untuk klasifikasi sentimen tiga kelas (positif, netral, negatif) pada berita finansial berbahasa Indonesia. Dataset terdiri dari 9.015 artikel dari empat portal keuangan, dilabeli dengan strategi hybrid *weak-supervision* (*lexicon-based labeling* dan anotasi manusia). Ketiga model dilatih dengan konfigurasi hyperparameter identik menggunakan *class-weighted cross-entropy loss*. Evaluasi menggunakan accuracy, macro F1-score, weighted F1-score, dan Cohen's Kappa pada validation set held-out (N = 2.149). Hasil eksperimen menunjukkan IndoBERT-Fine-Tuned mencapai Macro F1 tertinggi sebesar (0,863), diikuti IndoLEM-Fine-Tuned (0,843) dengan perbedaan prediksi antara kedua model yang signifikan berdasarkan uji McNemar (p = 0,0002), sementara FinBERT-Fine-Tuned mengalami penurunan tajam ke (0,437), bahkan di bawah *baseline* klasik TF-IDF + LinearSVC (0,809). Uji McNemar dengan koreksi Bonferroni mengonfirmasi bahwa kedua model BERT bahasa Indonesia secara signifikan mengungguli FinBERT-Fine-Tuned maupun *baseline* klasik (p < 0,001). Evaluasi pada gold subset berlabel manusia mengindikasikan adanya kemungkinan pengaruh pola label lexicon terhadap performa model pada label training. Kontribusi utama penelitian ini adalah evaluasi komparatif IndoBERT, IndoLEM, dan FinBERT pada berita finansial Indonesia dengan konfigurasi fine-tuning yang setara, disertai uji signifikansi statistik dan evaluasi pada gold standard berlabel manusia untuk menilai keterbatasan pendekatan weak-supervision.

Kata Kunci: Analisis sentimen; Berita finansial; FinBERT; Fine-tuning; IndoBERT; IndoLEM

Abstract

Financial news sentiment analysis plays a vital role in supporting investment decision-making in the Indonesian capital market. Previous studies typically employ a single model without cross-model comparison, leading to partial model selection recommendations. This study compares three *pre-trained* language models, namely IndoBERT, IndoLEM, and FinBERT, through a *fine-tuning* approach for three-class sentiment classification (positive, neutral, negative) on Indonesian financial news. The dataset consists of 9,015 articles from four financial portals, labeled using a hybrid *weak-supervision* approach combining *lexicon-based labeling* and human annotation. The three models were trained with identical hyperparameter configurations using *class-weighted cross-entropy loss*. Evaluation employed accuracy, macro F1-score, weighted F1-score, and Cohen's Kappa on a *held-out validation set* of 2,149 samples. Experimental results show IndoBERT-Fine-Tuned achieved the highest Macro F1 of (0.863), followed by IndoLEM-Fine-Tuned (0.843) with prediction differences between the two models found significant by McNemar's test (p = 0.0002), while FinBERT-Fine-Tuned achieved a substantially lower Macro F1 of (0.437), even below the classical TF-IDF + LinearSVC *baseline* (0.809). McNemar test with Bonferroni correction confirmed that both Indonesian BERT models significantly outperformed FinBERT-Fine-Tuned and the classical *baseline*s (p < 0.001). Additional evaluation on the human-labeled gold subset showed substantial performance drops, suggesting a possible influence of *lexicon-derived label patterns* on model performance. The main contribution of this study is a comparative evaluation of IndoBERT, IndoLEM, and FinBERT on Indonesian financial news under equivalent fine-tuning configurations, supported by statistical significance testing and gold-standard human-labeled evaluation to assess the limitations of the weak-supervision approach.

Keywords: Financial news; FinBERT; Fine-tuning; IndoBERT; IndoLEM; Sentiment analysis

1. PENDAHULUAN

Pasar modal Indonesia mengalami pertumbuhan yang signifikan dalam dekade terakhir, ditandai dengan peningkatan keterlibatan investor ritel dan volume transaksi harian. Berita finansial yang dipublikasikan oleh portal media keuangan menjadi salah satu sumber informasi utama yang dirujuk investor sebelum mengambil keputusan transaksi. Pemberitaan dengan nada negatif terhadap sebuah emiten dapat memicu aksi jual yang signifikan, sebaliknya berita dengan nada positif berpotensi mendorong aksi beli secara kolektif. Studi yang dilakukan pada dataset berita pasar modal Indonesia menunjukkan bahwa terdapat korelasi yang dapat diukur antara sentimen berita finansial dengan pergerakan harga saham domestik [1][2]. Kebutuhan untuk mengotomatisasi analisis sentimen pada berita finansial Indonesia karenanya memiliki relevansi praktis bagi industri investasi dan kebutuhan akademis bagi pengembangan *natural language processing* (NLP) berbahasa Indonesia.

Analisis sentimen pada teks finansial memiliki karakteristik berbeda dibandingkan domain umum seperti ulasan produk atau komentar media sosial. Terminologi finansial seperti aksi profit-taking, rebound, atau cuan memiliki muatan sentimen spesifik yang tidak selalu tertangkap oleh model bahasa generik. Model FinBERT yang diperkenalkan oleh

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1301
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Huang et al. dirancang untuk menjawab tantangan ini pada teks finansial berbahasa Inggris melalui *pre-training* pada korpus laporan keuangan, analyst report, dan komunikasi bisnis berskala besar [3]. Namun ketika diterapkan pada teks berbahasa Indonesia, FinBERT membutuhkan tahapan translasi terlebih dahulu, dan proses tersebut berpotensi menghilangkan konteks entitas finansial lokal seperti IHSG, BBRI, atau BI rate yang menjadi penanda sentimen penting pada teks berita.

Komunitas riset bahasa Indonesia telah menghasilkan beberapa model BERT yang dilatih spesifik pada korpus berbahasa Indonesia. IndoBERT yang dirilis bersama benchmark IndoLEM oleh Koto et al. [4] dan varian IndoBERT dari benchmark IndoNLU merupakan dua keluarga model yang paling banyak digunakan untuk tugas-tugas NLP berbahasa Indonesia. Beberapa penelitian terbaru telah membuktikan efektivitas IndoBERT pada berbagai tugas analisis sentimen. Yulianti dan Nissa menerapkan IndoBERT untuk aspect-based sentiment analysis pada ulasan pelanggan dan memperoleh peningkatan F1 yang signifikan dibandingkan baseline machine learning klasik [5]. Imaduddin et al. menggunakan IndoBERT untuk analisis sentimen pada ulasan aplikasi kesehatan dan mencapai akurasi 90% [6]. Ahmadian et al. menunjukkan bahwa hybrid IndoBERT-BiLSTM-BiGRU dapat memperbaiki performa pada IndoNLU benchmark untuk sentiment analysis dan emotion classification [7]. Nabiilah et al. menerapkan IndoBERT embedding dengan klasifier MBERT untuk klasifikasi multilabel teks bahasa Indonesia [8]. Pada domain finansial, Kang dan Choi melakukan komparasi FinBERT dengan GPT-4o pada berita keuangan berbahasa Inggris [9], namun belum melibatkan teks berbahasa Indonesia.

Beberapa penelitian terkait juga telah dilakukan pada domain berdekatan. Merayo et al. mengembangkan model *hybrid deep learning* untuk analisis sentimen media sosial [10]. Khan et al. memperkenalkan ElecBERT, varian BERT yang dilatih ulang untuk konteks pemilihan umum [11]. Dave dan Chowanda mengembangkan model ekstraksi entitas finansial Indonesia berbasis IndoBERT-BiGRU-CRF [12]. Pada konteks data tidak seimbang, Kunaefi et al. mengevaluasi IndoBERT dengan *focal loss* untuk klasifikasi berita hoaks [13], sementara Hanum et al. melakukan analisis kinerja BERT untuk tugas serupa [14]. Penelitian-penelitian tersebut menunjukkan bahwa pendekatan *fine-tuning* BERT efektif untuk berbagai tugas klasifikasi teks bahasa Indonesia, namun berdasarkan tinjauan literatur yang dilakukan, belum ditemukan studi yang secara eksplisit membandingkan model bahasa-spesifik (IndoBERT, IndoLEM) dengan model domain-spesifik finansial (FinBERT) pada teks berita finansial Indonesia menggunakan konfigurasi fine-tuning yang setara.

Ringkasan posisi penelitian ini terhadap penelitian terdahulu yang relevan disajikan pada Tabel 1, yang memperlihatkan bahwa berdasarkan tinjauan literatur yang dilakukan, belum ditemukan studi model bahasa yang spesifik dan model domain spesifik finansial pada teks berita finansial Indonesia menggunakan konfigurasi fine-tuning yang setara.

Tabel 1. Ringkasan Perbandingan Penelitian Terdahulu yang Relevan

Referensi	Model/Metode	Domain & Bahasa	Gap terhadap Penelitian Ini
[3]	FinBERT	Teks finansial Bahasa Inggris	Belum diuji pada teks Indonesia; translation noise belum dieksplorasi
[4]	IndoBERT/IndoLEM	NLP umum Bahasa Indonesia	Bukan domain finansial; belum dibandingkan setara
[5]	IndoBERT (ABSA)	Ulasan pelanggan	Domain berbeda; tanpa perbandingan lintas model
[7]	Hybrid IndoBERT-BiLSTM-BiGRU	Benchmark IndoNLU	Tidak dibandingkan dengan FinBERT
[9]	FinBERT vs GPT-4o	Berita keuangan Bahasa Inggris	Tidak melibatkan teks Bahasa Indonesia
[12]	IndoBERT-BiGRU-CRF	Entitas finansial Indonesia	Task ekstraksi entitas, bukan klasifikasi sentimen
[13]	IndoBERT + focal loss	Berita hoaks Indonesia	Domain berbeda; tanpa evaluasi gold-standard terpisah
Penelitian ini	IndoBERT, IndoLEM, FinBERT (fine-tuning setara)	Berita finansial Indonesia	Perbandingan setara tiga model + uji McNemar + evaluasi gold subset

Berdasarkan tinjauan tersebut, dapat diidentifikasi gap penelitian sebagai berikut. Pertama, evaluasi komparatif yang ketat antara IndoBERT, IndoLEM, dan FinBERT pada domain berita finansial Indonesia belum tersedia. Kedua, dampak proses translasi terhadap performa FinBERT ketika diterapkan pada teks Indonesia belum dieksplorasi secara sistematis, khususnya terkait hilangnya entitas finansial lokal. Ketiga, penelitian-penelitian sebelumnya umumnya menggunakan dataset berskala kecil (di bawah 5.000 sampel) atau menggunakan label dari satu sumber tanpa validasi reliabilitas.

Tantangan lain dalam riset analisis sentimen berita finansial Indonesia adalah ketersediaan label berkualitas. Anotasi manual oleh ahli memerlukan biaya tinggi dan tidak skalabel untuk dataset berukuran besar. Pendekatan *weak-*

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1302
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

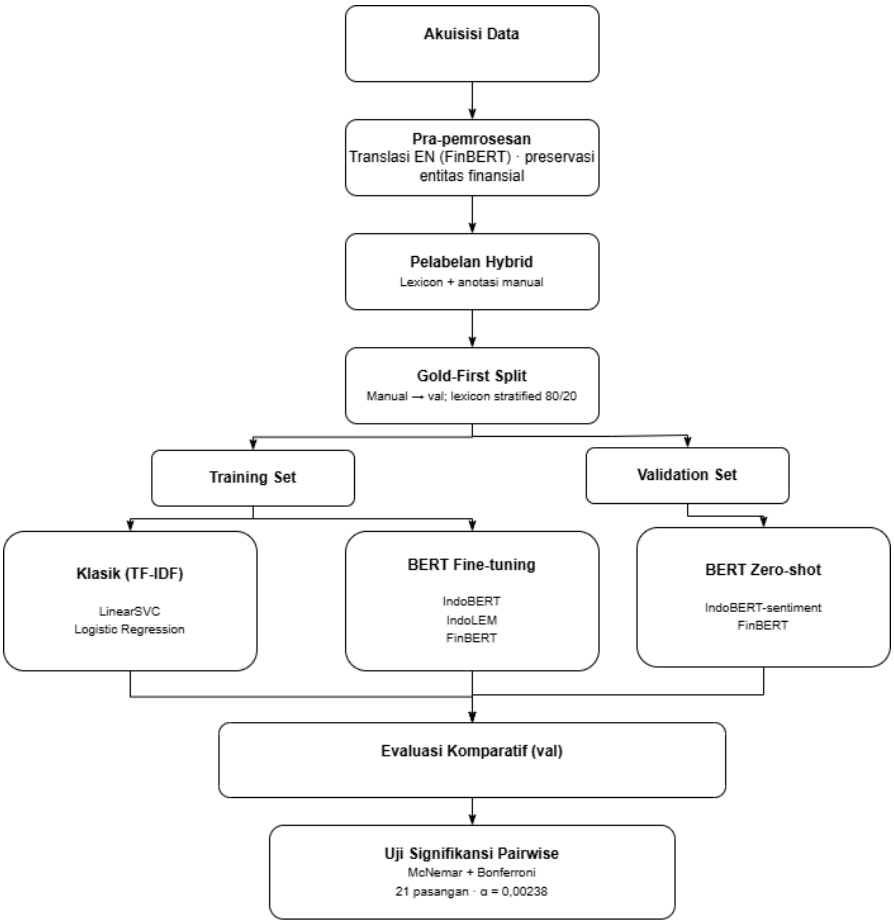
supervision berbasis lexicon menawarkan alternatif yang efisien namun rentan terhadap bias dan kesalahan sistematis [15]. Strategi *hybrid* yang menggabungkan *lexicon-based labeling* untuk skala dengan anotasi manusia pada subset *gold standard* dapat menyeimbangkan efisiensi dan kualitas, sekaligus menyediakan dasar evaluasi yang lebih ketat untuk mengatasi potensi *circularity* antara label training dan label evaluasi.

Berdasarkan motivasi di atas, penelitian ini bertujuan untuk: (1) membandingkan performa IndoBERT, IndoLEM, dan FinBERT pada klasifikasi sentimen berita finansial Indonesia dengan konfigurasi *fine-tuning* yang identik untuk menjamin komparabilitas; (2) menganalisis secara teknis mengapa salah satu model lebih unggul dalam menangani leksikon finansial bahasa Indonesia; dan (3) mengevaluasi reliabilitas hasil komparasi menggunakan uji signifikansi statistik dengan koreksi *multiple comparison* serta evaluasi terpisah pada subset *gold standard* berlabel manusia. Hasil yang diharapkan adalah memberikan bukti empiris mengenai karakteristik performa masing-masing model pada konteks berita finansial Indonesia, beserta diagnosa kelemahan masing-masing pendekatan yang dapat menjadi acuan pengembangan model selanjutnya.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini mengikuti alur pipeline klasifikasi teks *supervised* yang terdiri dari empat tahapan utama: pengumpulan data, pra-pemrosesan dan pelabelan, pelatihan model *fine-tuning*, serta evaluasi komparatif. Arsitektur umum sistem digambarkan pada Gambar 1. Setiap tahapan dirancang dengan prinsip reproduksibilitas: random seed ditetapkan pada nilai 42 di seluruh komponen yang melibatkan stokastisitas, dan pembagian data train-validation dilakukan satu kali kemudian diturunkan secara konsisten ke seluruh eksperimen.



Gambar 1. Diagram alur tahapan penelitian dari pengumpulan data hingga evaluasi komparatif model.

2.2. Pengumpulan dan Pelabelan Data

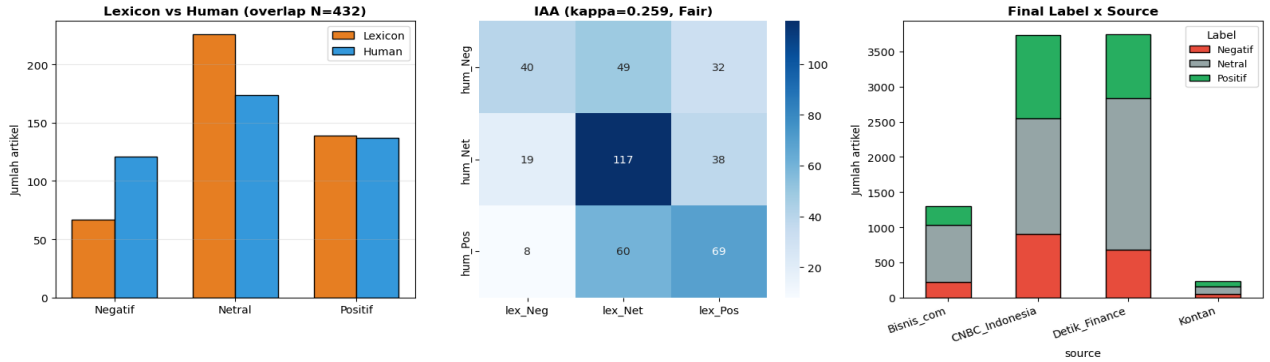
Data berita finansial dikumpulkan dari empat portal media keuangan online di Indonesia, yaitu Detik Finance, CNBC Indonesia, Bisnis.com, dan Kontan. Pemilihan keempat portal didasarkan pada cakupan tematik yang komprehensif (saham, makroekonomi, perbankan, komoditas) dan volume publikasi harian yang tinggi. Proses pengumpulan dilakukan menggunakan teknik *web scraping* dengan kombinasi parsing HTML dan agregasi RSS feed, dilengkapi rotasi *User-Agent* dan jeda permintaan acak 1,0–2,5 detik untuk menghindari pemblokiran. Setelah deduplikasi berbasis hash dari

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1303
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

kombinasi URL dan judul artikel, serta filter kualitas tambahan (panjang judul minimum 10 karakter, ringkasan tidak kosong), diperoleh 9.015 artikel unik dengan distribusi sumber: Detik Finance 3.745 artikel, CNBC Indonesia 3.732 artikel, Bisnis.com 1.302 artikel, dan Kontan 236 artikel.

Pelabelan dilakukan dengan strategi *hybrid weak-supervision*. Komponen lexicon terdiri atas 53 kata bermuatan positif (misal naik, cuan, untung) dan 59 kata bermuatan negatif (misal anjlok, rugi, merosot), masing-masing dengan bobot intensitas 1,0, 1,5, atau 2,0. Aturan negasi dengan window dua token diterapkan untuk membalik polaritas ketika kata penanda negasi (tidak, bukan, belum) muncul di depan kata sentimen. Threshold netral ditetapkan pada selisih skor di bawah 0,5 dengan total skor di bawah 1,5. Subset gold standard dibentuk dari 500 artikel yang disampling secara *stratified* per portal sebagai kandidat awal untuk anotasi manusia, di mana 432 artikel (86,4% dari 500 kandidat) berhasil diselesaikan proses anotasinya secara manual oleh satu annotator sebagai subset berlabel manusia (human-labeled reference subset), sementara 68 artikel sisanya tidak diikutsertakan ke dalam subset ini dan tidak digunakan pada tahap evaluasi mana pun. Subset berlabel manusia berukuran final 432 artikel ini digunakan untuk validasi reliabilitas. Anotasi dilakukan ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif, dengan mempertimbangkan konteks sentimen dalam berita. Seluruh artikel yang telah dianotasi manusia dialokasikan ke validation set dan tidak digunakan dalam proses training; label manusia digunakan sebagai referensi eksternal untuk mengevaluasi kemampuan model terhadap label yang tidak dihasilkan melalui proses lexicon-based labeling. Istilah "gold standard"/"gold subset" pada naskah ini merujuk pada subset berlabel manusia tersebut sebagai referensi eksternal, bukan ground truth yang sempurna. Reliabilitas lexicon dievaluasi dengan menghitung Cohen's Kappa yang mengukur tingkat kesepakatan antara label lexicon dan anotasi manusia pada 432 artikel tersebut (bukan agreement antar-annotator manusia), memperoleh nilai 0,259 yang termasuk kategori fair agreement. Nilai Kappa yang moderat ini mengindikasikan bahwa label lexicon hanya merupakan aproksimasi parsial terhadap persepsi manusia, sebagaimana divisualisasikan pada Gambar 2, sehingga evaluasi terpisah pada *subset gold standard* diperlukan untuk menilai kemampuan model di luar pola lexicon.



Gambar 2. Visualisasi distribusi label lexicon vs human pada subset overlap (N = 432), agreement matrix antara label lexicon dan anotasi manusia (Kappa = 0,259, fair), dan komposisi label final per sumber portal.

2.3 Pra-pemrosesan dan Translasi

Teks dinormalisasi dengan menghapus karakter HTML residu, URL, dan *whitespace* berlebih. Input model dibentuk dari penggabungan judul dan ringkasan dengan separator titik. Untuk FinBERT yang dilatih pada teks bahasa Inggris, teks Indonesia diterjemahkan menggunakan *API Google Translator*. Strategi *entity preservation* diterapkan untuk menjaga nama entitas finansial Indonesia (misalnya IHSG, BBRI, OJK, dan Bank Indonesia) tetap dalam bentuk aslinya tanpa diterjemahkan. Dalam proses translasi ini, entitas finansial tersebut dipertahankan dalam bentuk aslinya untuk mengurangi perubahan terhadap informasi entitas pada teks sumber. Strategi serupa dalam menjaga entitas domain finansial Indonesia digunakan pada pengembangan model ekstraksi entitas oleh Dave dan Chowanda [12].

2.4 Arsitektur *Fine-Tuning* dan Fungsi Kerugian

Tiga model *pre-trained* dibandingkan dalam konfigurasi *fine-tuning*. IndoBERT-base-p1 dari IndoBenchmark yang dilatih pada korpus Indo4B [16], IndoBERT-base-uncased dari IndoLEM yang dilatih pada teks formal Indonesia [4], dan FinBERT yang dilatih pada korpus teks finansial berbahasa Inggris [3]. Setiap model dilengkapi dengan classification head berupa fully-connected layer yang memetakan *hidden state token* [CLS] ke ruang tiga kelas (Negatif, Netral, Positif). Untuk menjamin komparabilitas, ketiga model dilatih dengan konfigurasi hyperparameter identik: learning rate 1×10^{-5}, batch size 16, max sequence length 256 token, 5 *epoch*, warmup ratio 0,1, weight decay 0,01, optimizer AdamW, dan mixed precision FP16. Konfigurasi ini sejalan dengan praktik standar *fine-tuning* BERT pada berbagai tugas klasifikasi teks bahasa Indonesia [13], [17]. Learning rate 1×10^{-5} dipilih sebagai nilai konservatif untuk menjaga stabilitas proses fine-tuning pada representasi hasil pre-training. Batch size 16 dan mixed precision FP16 digunakan untuk menyesuaikan keterbatasan memori GPU tanpa mengorbankan stabilitas gradien. Max sequence length 256 token dipilih berdasarkan karakteristik panjang teks gabungan judul dan ringkasan berita yang umumnya ringkas, sehingga cukup untuk mencakup mayoritas konten tanpa truncation berlebihan. Jumlah 5 *epoch* ditetapkan sebagai batas pelatihan dengan

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1304
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

mempertimbangkan ukuran training set yang terdiri dari 6.866 sampel dan untuk membatasi risiko overfitting, sementara warmup ratio 0,1 dan weight decay 0,01 berfungsi sebagai regularisasi untuk menstabilkan proses optimasi pada tahap awal pelatihan. Optimizer AdamW digunakan untuk proses fine-tuning model Transformer. Seluruh nilai ini diterapkan secara identik pada ketiga model untuk menjamin konsistensi komparasi antar-model, bukan hasil pencarian hyperparameter khusus untuk penelitian ini.

Untuk mengatasi imbalance antar kelas, digunakan *class-weighted cross-entropy loss* dengan bobot per kelas dihitung berdasarkan distribusi label pada training set. Fungsi kerugian yang dioptimasi selama *fine-tuning* ditunjukkan pada Persamaan (1):

$$L = (-\frac{1}{N}) \sum_{i=1}^{N} \sum_{j=1}^{c} w_j \cdot y_{ij} \cdot log(\hat{y}_{ij}) \quad (1)$$

Dengan N adalah jumlah sampel pada batch, C adalah jumlah kelas (C = 3), w_j adalah bobot kelas j, y_{ij} adalah label one-hot ground truth untuk sampel i pada kelas j, dan $\hat{y}_{ij}$ adalah probabilitas prediksi model untuk sampel i pada kelas j. Selain *class weighting*, label smoothing dengan koefisien 0,1 diterapkan untuk meredam pengaruh label noise yang berasal dari *weak-supervision* [14].

2.5 Metrik Evaluasi

Dataset dibagi menggunakan strategi *gold-first split*: seluruh 432 artikel berlabel manusia dialokasikan sepenuhnya ke validation set sebagai gold standard held-out, sementara 8.583 artikel berlabel lexicon dibagi 80/20 secara stratified sampling, menghasilkan lexicon validation sebesar 1.717 sampel dan training set sebesar 6.866 sampel. Hasil akhirnya adalah training set 6.866 sampel (76,2%) dan validation set 2.149 sampel (23,8%), yang terdiri atas 1.717 sampel lexicon validation dan 432 sampel human validation (1.717 + 432 = 2.149), di mana subset gold standard berlabel manusia N = 432 menjadi bagian dari validation set tanpa overlap dengan training. Strategi ini secara eksplisit menghilangkan train-set contamination dari label manusia, sehingga evaluasi pada gold subset benar-benar mencerminkan generalisasi model. Indikator boolean is_val ditambahkan pada dataset agar pembagian konsisten lintas seluruh eksperimen, sehingga mencegah data leakage antar model. Evaluasi performa klasifikasi dilakukan menggunakan empat metrik utama yang dihitung berdasarkan matriks konfusi dengan empat komponen: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Keempat metrik dirumuskan pada Persamaan (2) hingga (5).

a. Akurasi (Accuracy) : menghitung persentase prediksi yang akurat secara keseluruhan, termasuk kelas positif dan negatif. Formula untuk menghitung akurasi adalah:

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \quad (2)$$

b. Ketepatan(*Precision*) : menilai seberapa tepat ramalan positif yang dihasilkan oleh model. Presisi tinggi menandakan bahwa model jarang mengalami kesalahan tipe I (false positive). Rumus ketepatan adalah:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

c. *Recall*: mengukur kemampuan model untuk mengidentifikasi semua contoh positif yang sebenarnya. *Recall* yang tinggi menunjukkan model jarang melakukan kesalahan tipe II (*false negative*). Rumus *recall* yaitu:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

d. *F1-Score*: *F1-Score* adalah rata-rata harmonis dari presisi dan recall, mencerminkan keseimbangan antara kedua elemen tersebut. Metrik ini bermanfaat saat distribusi kelas tidak seimbang. Rumus F1-Score ialah

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Pada konteks klasifikasi multi-kelas tiga kategori, metrik Precision, Recall, dan F1-Score dihitung per kelas kemudian diagregasi dengan dua skema: macro averaging yang memberikan bobot setara untuk setiap kelas tanpa memandang frekuensi, dan weighted averaging yang memberikan bobot proporsional terhadap support kelas. Macro F1 dipilih sebagai metrik utama karena tidak bias terhadap kelas mayoritas, sehingga lebih sesuai untuk evaluasi pada kondisi imbalance. Selain itu, Cohen's Kappa dihitung untuk mengukur kesepakatan prediksi-label dengan koreksi terhadap chance agreement. Untuk pengujian signifikansi statistik perbedaan antar model, digunakan uji McNemar dengan continuity correction pada matriks kontingensi pasangan model, disertai koreksi Bonferroni untuk multiple comparison (21 pasangan, α adjusted = 0,00238). Sebagai pembanding, evaluasi juga dilakukan pada subset gold standard human-labeled berukuran N = 432 yang seluruhnya menjadi bagian validation set untuk mengatasi potensi *circularity* dari penggunaan label lexicon pada training dan evaluasi, serta dibandingkan dengan *baseline* klasik TF-IDF + Logistic Regression dan TF-IDF + Linear SVC.

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Dataset dan *Ablation Threshold Lexicon*

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1305
This Journal is licensed under a Creative Commons Attribution 4.0 International License

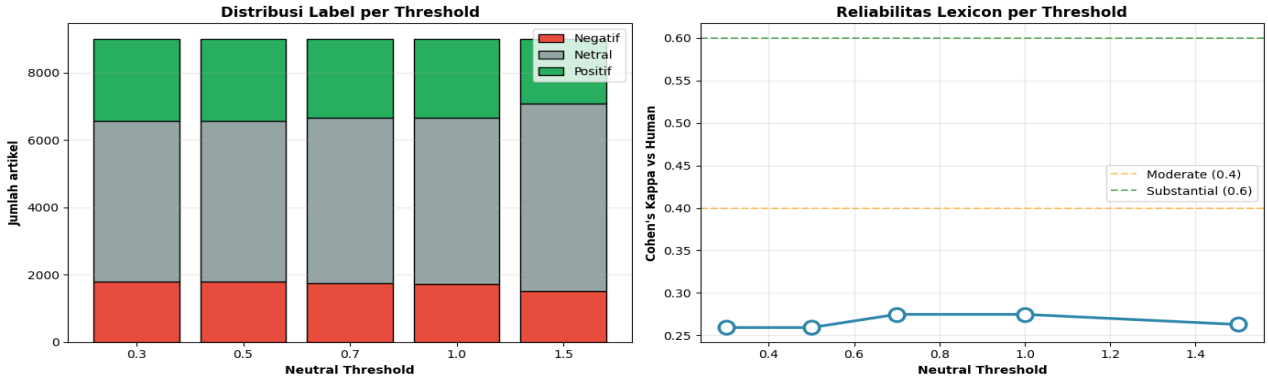
JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Setelah deduplikasi dan filter kualitas, dataset final terdiri dari 9.015 artikel berita finansial dengan distribusi sumber yang didominasi oleh Detik Finance dan CNBC Indonesia. Pemilihan max sequence length sebesar 256 token didasarkan pada karakteristik teks input berupa judul dan ringkasan berita yang umumnya tidak panjang, sehingga konfigurasi ini dinilai cukup untuk mencakup mayoritas teks tanpa truncation berlebihan. Distribusi kelas yang dihasilkan *lexicon-based labeling* pada threshold operasional 0,5 menunjukkan kondisi imbalance moderat dengan komposisi 20,6% negatif (1.801 artikel), 52,3% netral (4.769 artikel), dan 27,1% positif (2.445 artikel), dengan rasio imbalance 2,65. Dominasi kelas netral konsisten dengan karakteristik berita finansial yang banyak memuat pelaporan faktual seperti rilis laporan keuangan atau pernyataan resmi tanpa muatan emosional kuat.

Sebelum digunakan untuk training, dilakukan *ablation study* terhadap parameter threshold pada *lexicon-based labeling* untuk memvalidasi pemilihan threshold operasional. Threshold yang diuji berkisar dari 0,3 hingga 1,5 dengan mengukur distribusi kelas yang dihasilkan dan Cohen's Kappa terhadap anotasi manusia pada subset gold standard berukuran 432 artikel. Hasil lengkap ditampilkan pada Tabel 2, dengan visualisasi distribusi label dan reliabilitas Kappa pada Gambar 3.

Tabel 2. Ablation Study Threshold Lexicon untuk Pelabelan Otomatis

Threshold	Negatif	Netral	Positif	Imbalance Ratio	Kappa vs Human
0,3	1.801	4.769	2.445	2,65	0,259
0,5	1.801	4.769	2.445	2,65	0,259
0,7	1.739	4.920	2.356	2,83	0,275
1,0	1.736	4.933	2.346	2,84	0,275
1,5	1.510	5.584	1.921	3,70	0,263



Gambar 3. Visualisasi ablation study threshold lexicon — distribusi label per threshold (kiri) dan reliabilitas Cohen's Kappa lexicon terhadap anotasi manusia (kanan).

Tabel 2 menunjukkan bahwa threshold 0,3 dan 0,5 menghasilkan distribusi label yang identik dengan Kappa sebesar 0,259, sementara threshold 0,7 dan 1,0 memberikan Kappa marginal lebih tinggi (0,275). Selisih Kappa antar threshold relatif kecil dan seluruh nilai berada pada kategori fair agreement, mengindikasikan bahwa pendekatan lexicon hanya menangkap pola permukaan sentimen dan belum sepenuhnya selaras dengan persepsi manusia. Threshold yang lebih tinggi (1,5) menggeser distribusi ke arah dominasi kelas netral dengan rasio imbalance hingga 3,70 dan Kappa menurun ke 0,263. Pemilihan threshold operasional 0,5 dilakukan untuk mempertahankan distribusi kelas yang relatif lebih seimbang sambil menjaga reliabilitas. Konsekuensi dari nilai Kappa yang moderat adalah perlunya evaluasi terpisah pada subset human-labeled untuk memastikan kemampuan model di luar pola lexicon, sebagaimana dilakukan pada Sub-bab 3.5. Tren serupa pada *lexicon-based labeling* teks bahasa Indonesia juga dilaporkan oleh Daqiqil et al. [18]

3.2 Performa Komparatif Model pada *Validation Set*

Tabel 3 menampilkan hasil komparatif tujuh konfigurasi model pada validation set held-out berukuran 2.149 sampel yang mencakup seluruh 432 artikel berlabel manusia. Selain tiga model *fine-tuned* yang menjadi fokus utama penelitian, ditampilkan pula dua *baseline* zero-shot dan dua *baseline* klasik berbasis TF-IDF untuk memberikan konteks performa *fine-tuning* yang lebih komprehensif.

Tabel 3. Performa Komparatif Model pada Validation Set Penuh (N = 2.149)

Model	Accuracy	Macro F1	Weighted F1	Cohen's Kappa
IndoLEM-Fine-Tuned	0,857	0,843	0,856	0,767
IndoBERT-Fine-Tuned	0,875	0,863	0,874	0,797

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1306
This Journal is licensed under a Creative Commons Attribution 4.0 International License

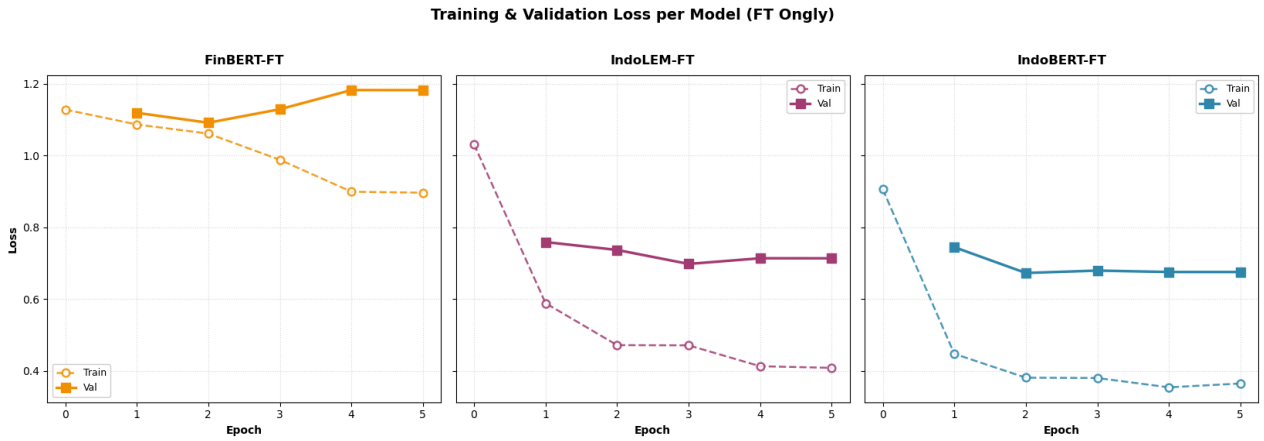
JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Model	Accuracy	Macro F1	Weighted F1	Cohen's Kappa
TF-IDF + LinearSVC	0,827	0,809	0,825	0,715
TF-IDF + LogReg	0,807	0,786	0,805	0,683
FinBERT-Fine-Tuned	0,474	0,437	0,478	0,166
IndoBERT-Zero-Shot	0,497	0,280	0,375	0,019
FinBERT-Zero-Shot	0,453	0,422	0,455	0,128

IndoBERT-Fine-Tuned mencapai Macro F1 sebesar 0,863, sedangkan IndoLEM-Fine-Tuned mencapai 0,843. Perbedaan prediksi kedua model signifikan berdasarkan uji McNemar (p = 0,0002), meskipun selisih Macro F1 secara deskriptif hanya sebesar 2,0 poin persentase. Hasil ini mengindikasikan bahwa IndoBERT yang dilatih pada korpus Indo4B (mencakup ragam genre berita umum) memiliki sedikit keunggulan dibanding IndoLEM yang lebih fokus pada teks formal jurnalistik untuk tugas klasifikasi sentimen berita finansial pada dataset ini. Kedua model BERT bahasa Indonesia mengungguli *baseline* klasik TF-IDF + LinearSVC (0,809) dan TF-IDF + LogReg (0,786) dengan margin sekitar 3,4 hingga 7,7 poin persentase Macro F1. Hasil uji McNemar pada pasangan model yang diuji menunjukkan adanya perbedaan prediksi yang signifikan namun lebih moderat dibanding hasil pada split lexicon-only. Temuan ini sejalan dengan studi transfer learning sentiment classification berbasis BERT yang dilaporkan oleh Areshey dan Mathkour [19]

Temuan paling mencolok pada Tabel 3 adalah performa FinBERT-Fine-Tuned yang justru jatuh ke Macro F1 0,437, di bawah *baseline* klasik TF-IDF + LinearSVC dengan selisih sekitar 37,2 poin persentase. Hasil ini mematahkan asumsi awal bahwa model domain-spesifik finansial akan unggul pada tugas sentimen berita finansial. Dinamika training dan validation loss seluruh model *fine-tuned* divisualisasikan pada Gambar 4 yang memperlihatkan validation loss FinBERT-FT secara divergen meningkat dari *epoch* 2, indikasi *overfitting* yang konsisten dengan rendahnya Macro F1. Akar penyebab penurunan akan dibahas pada Sub-bab 3.7, namun secara ringkas berkaitan dengan dua sumber degradasi: *translation noise* pada teks Indonesia yang diterjemahkan ke bahasa Inggris, dan domain *mismatch* antara korpus *pre-training* FinBERT (laporan keuangan formal berbahasa Inggris) dengan berita finansial konsumen Indonesia.

Performa model zero-shot (IndoBERT-zeroshot yang menggunakan model mdhugol sentiment Twitter Indonesia, dan FinBERT-zeroshot langsung dari Hugging Face) berada jauh di bawah *baseline* klasik, masing-masing pada Macro F1 0,280 dan 0,422. Hal ini mengindikasikan bahwa *fine-tuning* pada domain target tidak dapat digantikan oleh inference langsung dari model *pre-trained* yang dilatih pada domain berbeda. IndoBERT-zeroshot yang dilatih pada sentiment Twitter sulit memetakan istilah formal-jurnalistik yang lazim pada teks berita, sementara FinBERT-zeroshot terkendala oleh *translation noise* sekaligus *domain mismatch*. Pola serupa juga diobservasi pada studi sentiment analysis bahasa Indonesia oleh Setiawan et al. [20].



Gambar 4. Visualisasi training dan validation loss per *epoch* untuk ketiga model *fine-tuned*. FinBERT-Fine-Tuned menunjukkan validation loss yang secara divergen meningkat dari *epoch* 2 (indikasi *overfitting*/*translation noise*), sedangkan IndoBERT-Fine-Tuned dan IndoLEM-Fine-Tuned menunjukkan konvergensi stabil.

3.3 Pengujian Signifikansi Statistik dengan *McNemar Test*

Untuk menguji apakah perbedaan performa antar model bersifat signifikan dan bukan sekadar fluktuasi statistik, dilakukan pairwise McNemar test pada prediksi setiap model di *validation set*. McNemar test menggunakan matriks kontingensi yang menghitung jumlah sampel di mana satu model benar sedangkan model lainnya salah (b dan c), kemudian menghitung statistik chi-kuadrat dengan continuity correction. Tabel 4 menampilkan hasil pengujian untuk enam pasangan model utama. Total perbandingan pairwise yang dilakukan untuk tujuh model adalah 21 pasangan, sehingga koreksi Bonferroni menghasilkan α adjusted sebesar 0,05 dibagi 21 atau 0,00238.

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1307
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 4. Hasil Pairwise McNemar Test († = signifikan setelah koreksi Bonferroni; ns = tidak signifikan)

Perbandingan	b	c	χ^2	p-value	Signifikansi
IndoBERT-Fine-Tuned vs IndoLEM-Fine-Tuned	72	33	13,75	0,0002	*** †
IndoBERT-Fine-Tuned vs FinBERT-Fine-Tuned	951	89	712,81	< 0,0001	*** †
IndoLEM-Fine-Tuned vs FinBERT-Fine-Tuned	925	102	657,92	< 0,0001	*** †
IndoBERT-Fine-Tuned vs TF-IDF + LogReg	199	53	83,43	< 0,0001	*** †
IndoLEM-Fine-Tuned vs TF-IDF + LogReg	184	77	43,05	< 0,0001	*** †
FinBERT-Zero-Shot vs FinBERT-Fine-Tuned	346	390	2,51	0,1130	ns

Hasil uji McNemar menunjukkan dua temuan penting. Pertama, perbandingan IndoBERT-Fine-Tuned versus IndoLEM-Fine-Tuned menghasilkan chi-kuadrat sebesar 13,75 dengan p-value 0,0002, yang berarti perbedaan prediksi antara kedua model signifikan berdasarkan uji McNemar bahkan setelah koreksi Bonferroni, sejalan dengan selisih 2,0 poin persentase pada Macro F1 yang teramati. Implikasinya, IndoBERT-Fine-Tuned menghasilkan Macro F1 tertinggi pada validation set penuh dalam konfigurasi dataset dan pelabelan yang digunakan, dengan IndoLEM-Fine-Tuned menghasilkan performa yang berdekatan sebagai alternatif yang kompetitif. Kedua, Pasangan model fine-tuned terhadap FinBERT maupun baseline klasik yang ditampilkan pada Tabel 4 menghasilkan p-value di bawah 0,0001 dan tetap signifikan setelah koreksi Bonferroni dengan α adjusted 0,00238. Margin terbesar tercatat pada perbandingan IndoBERT-Fine-Tuned versus FinBERT-Fine-Tuned dengan chi-kuadrat sebesar 712,81, yang menunjukkan bahwa perbedaan prediksi FinBERT dengan IndoBERT pada validation set bersifat signifikan secara statistik. Pengecualian muncul pada perbandingan FinBERT-Zero-Shot versus FinBERT-Fine-Tuned (chi-kuadrat = 2,51; p = 0,1130) yang tidak signifikan, mengindikasikan bahwa *fine-tuning* FinBERT pada teks Indonesia hasil translasi tidak memberikan perbaikan yang berarti dibanding inference langsung. Penggunaan koreksi Bonferroni penting karena tanpa koreksi tersebut peluang false positive akan meningkat secara substansial pada skenario multiple comparison.

3.4 Analisis Pola Kesalahan

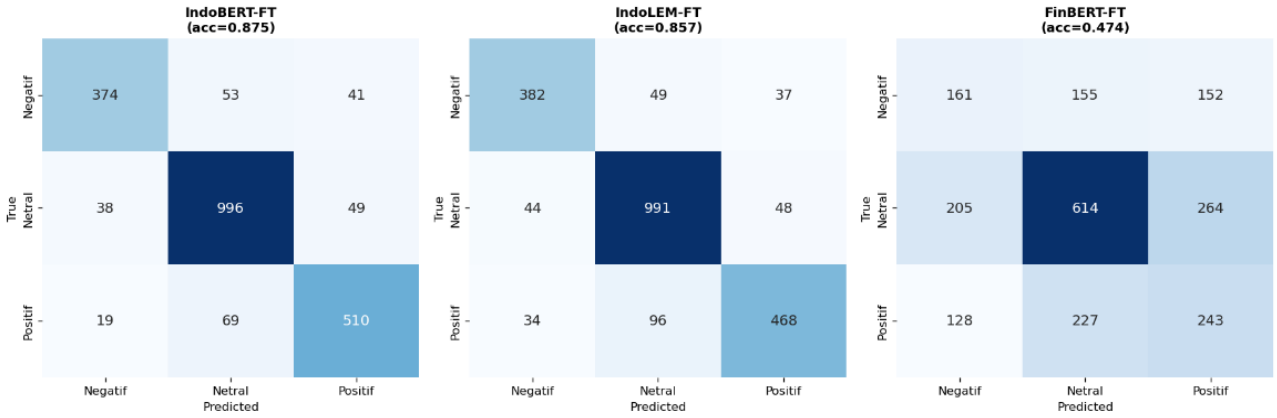
Analisis confusion matrix mengungkap pola kesalahan yang berbeda untuk masing-masing model *fine-tuned*. IndoBERT-Fine-Tuned mencatat distribusi kesalahan yang relatif merata antar kelas dengan F1 per kelas berkisar di sekitar 0,83–0,91, menunjukkan keseimbangan antara presisi dan recall pada ketiga kelas. IndoLEM-*Fine-Tuned* menunjukkan pola serupa pada level yang sedikit lebih rendah (F1 per kelas 0,83–0,86), dengan keunggulan marginal pada kelas Negatif. Sebaliknya, FinBERT-*Fine-Tuned* mengalami penurunan tajam pada seluruh kelas dengan F1 di bawah 0,50. Penurunan paling tajam tercatat pada kelas Netral, di mana recall berada di sekitar 0,33, yang berarti FinBERT-*Fine-Tuned* hanya mampu mengidentifikasi sepertiga dari berita berlabel netral dan cenderung salah mengklasifikasikannya sebagai negatif atau positif. Pola ini konsisten dengan hipotesis *translation noise*: nuansa pelaporan faktual netral dalam bahasa Indonesia (misalnya frasa 'mencatat kenaikan' atau 'tetap stabil') sering kehilangan konteks setelah translasi, sehingga FinBERT cenderung memberikan polaritas yang lebih tegas dari semestinya. Visualisasi confusion matrix ketiga model *fine-tuned* ditampilkan pada Gambar 5, sementara perbandingan F1-score per kelas dirangkum pada Gambar 6.

Tabel 5. Perbandingan Precision, Recall, dan F1-Score Per Kelas untuk Tiga Model Fine-Tuned pada Validation Set (N = 2.149)

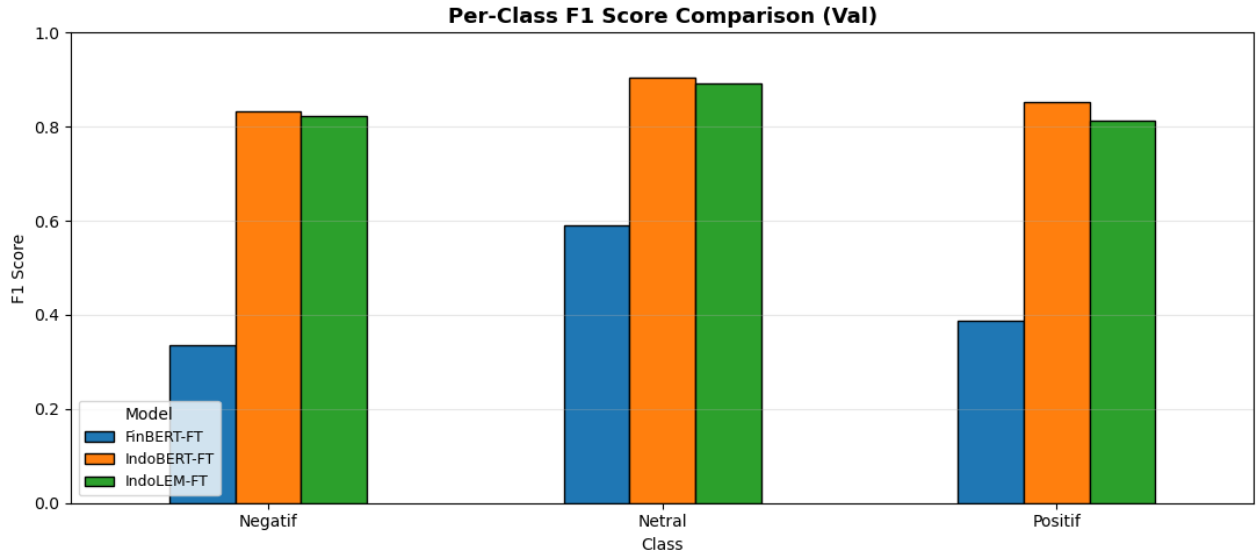
Model	Class	Precision	Recall	F1-Score	Support
IndoBERT-*Fine-Tuned*	Negatif	0,867749	0,799145	0,832036	468
IndoBERT-*Fine-Tuned*	Netral	0,890877	0,919668	0,905043	1083
IndoBERT-*Fine-Tuned*	Positif	0,850000	0,852843	0,851419	598
IndoLEM-*Fine-Tuned*	Negatif	0,830435	0,816239	0,823276	468
IndoLEM-*Fine-Tuned*	Netral	0,872359	0,915051	0,893195	1083
IndoLEM-*Fine-Tuned*	Positif	0,846293	0,782609	0,813206	598
FinBERT-*Fine-Tuned*	Negatif	0,325911	0,344017	0,334719	468
FinBERT-*Fine-Tuned*	Netral	0,616466	0,566944	0,590669	1083
FinBERT-*Fine-Tuned*	Positif	0,368741	0,406355	0,386635	598

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1308
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 5. Confusion matrix ketiga model *fine-tuned* pada validation set.



Gambar 6. Visualisasi perbandingan F1-score per kelas (Negatif, Netral, Positif) pada ketiga model *fine-tuned*. FinBERT-*Fine-Tuned* mengalami degradasi merata di seluruh kelas dengan F1 di bawah 0,60, sedangkan IndoBERT-*Fine-Tuned* dan IndoLEM-*Fine-Tuned* menunjukkan keseimbangan tinggi pada semua kelas.

3.5 Evaluasi pada *Subset Gold Standard*

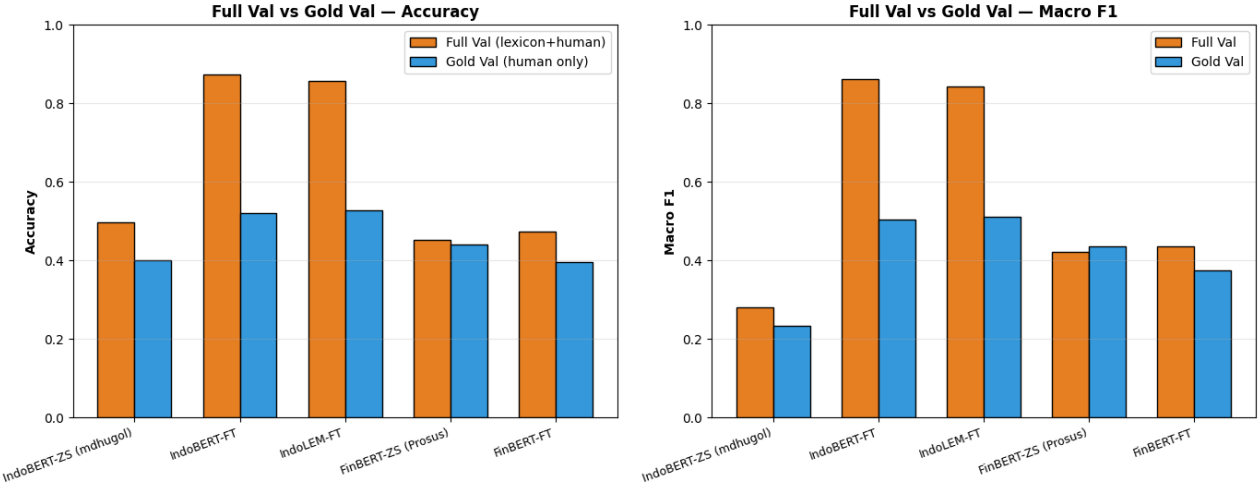
Untuk mengatasi potensi *circularity* dari penggunaan label lexicon pada training dan evaluasi, dilakukan evaluasi terpisah pada subset gold standard berukuran 432 artikel yang seluruhnya berlabel anotasi manusia dan tergabung sepenuhnya pada validation set berkat strategi gold-first split. Subset ini menjadi tolak ukur kemampuan model dalam menangkap sentimen sebagaimana dipersepsikan manusia, terlepas dari pola lexicon. Tabel 6 membandingkan performa pada validation set penuh dengan performa pada subset gold standard, dengan visualisasi grafis ditampilkan pada Gambar 7.

Tabel 6. Perbandingan Performa pada Validation Set Penuh dan Subset Gold Standard (N gold = 432)

Model	*Full Val Macro F1*	*Gold Macro F1*	*Full Val Accuracy*	*Δ Accuracy (poin %)*
IndoLEM- *Fine-Tuned*	0,843	0,512	0,857	−32,89
IndoBERT- *Fine-Tuned*	0,863	0,504	0,875	−35,40
TF-IDF + LogReg	0,786	0,521	0,807	−26,06
TF-IDF + LinearSVC	0,809	0,506	0,827	−29,73
FinBERT- *Fine-Tuned*	0,437	0,376	0,474	−7,79

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1309
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 7. Visualisasi perbandingan Accuracy dan Macro F1 pada Full Validation (lexicon + human) versus Gold Validation (human-only, N = 432). Penurunan substansial performa model BERT fine-tuned pada subset gold mengindikasikan adanya kemungkinan pengaruh pola label lexicon terhadap performa model.

Hasil pada Tabel 6 mengungkap pola yang penting untuk interpretasi. Pertama, pola performa model pada subset gold standard berbeda dari full validation signifikan dibanding ranking pada validation set penuh. Pada subset berlabel manusia, TF-IDF + LogReg memperoleh Macro F1 sebesar 0,521, IndoLEM-*Fine-Tuned* memperoleh Macro F1 sebesar 0,512, TF-IDF + LinearSVC sebesar 0,506, IndoBERT-*Fine-Tuned* sebesar 0,504, dan FinBERT-*Fine-Tuned* sebesar 0,376. Pada subset human-labeled, *baseline* klasik TF-IDF + LogReg justru tampil setara bahkan sedikit unggul dibanding kedua model BERT *fine-tuned*, sebuah temuan yang berlawanan dengan hierarki pada full validation. Kedua, terdapat penurunan performa yang substansial untuk model BERT yang dilatih dengan label lexicon, sebesar 35,40 poin persentase *accuracy* untuk IndoBERT-*Fine-Tuned* dan 32,89 poin persentase untuk IndoLEM-*Fine-Tuned*. Baseline klasik mengalami penurunan yang lebih moderat (26–30 poin persentase), sementara perilaku berbeda muncul pada FinBERT-Fine-Tuned yang penurunannya hanya 7,79 poin persentase, namun bukan karena model tersebut *robust*, melainkan karena performa pada *full validation* pun sudah rendah sehingga jarak ke *gold standard* menjadi kecil. Pergeseran performa pada gold subset mengindikasikan bahwa sebagian performa pada full validation dipengaruhi oleh pola yang terdapat pada label lexicon.

Selisih performa antara full validation dan gold subset menunjukkan adanya pengaruh pola label lexicon terhadap performa model yang dilatih menggunakan label weak-supervision. Temuan ini menunjukkan pentingnya evaluasi pada subset berlabel manusia sebagai pelengkap full validation agar kemampuan generalisasi model tidak ditaksir terlalu tinggi. Pada gold subset, baseline TF-IDF + LogReg memiliki performa yang berdekatan dengan model BERT, sementara pada full validation IndoBERT dan IndoLEM menunjukkan performa lebih tinggi. Perbedaan ini menunjukkan bahwa kualitas dan karakteristik label training perlu menjadi perhatian dalam pengembangan model selanjutnya. Implikasinya, angka pada *full validation* sebaiknya selalu dilaporkan bersama angka pada *gold subset* agar kemampuan generalisasi model tidak ditaksir terlalu tinggi. Meskipun pada *gold subset* TF-IDF + LogReg menyamai performa kedua model BERT, IndoBERT dan IndoLEM tetap unggul signifikan pada *full validation* dan menawarkan fleksibilitas representasi kontekstual yang tidak dimiliki pendekatan klasik, sepanjang kualitas label training terus ditingkatkan.

3.6 Posisi Hasil Penelitian terhadap Studi Sejenis

Untuk memperjelas posisi dan kontribusi penelitian ini terhadap studi sejenis, Tabel 7 membandingkan hasil penelitian ini dengan temuan pada penelitian terdahulu yang relevan. Perbandingan ini menegaskan bahwa performa FinBERT yang tinggi umumnya hanya tercapai pada domain aslinya, yaitu teks finansial berbahasa Inggris, sebagaimana ditunjukkan Kang dan Choi [9], sementara pada teks Indonesia hasil translasi performanya turun tajam pada penelitian ini. Sebaliknya, performa IndoBERT-Fine-Tuned pada penelitian ini menunjukkan bahwa model bahasa Indonesia juga dapat memberikan kinerja tinggi pada domain finansial, sebagaimana efektivitas IndoBERT yang telah dilaporkan pada domain aplikasi kesehatan oleh Imaduddin et al. [6]. Temuan ini memberikan indikasi bahwa IndoBERT dapat mempertahankan kinerja yang baik pada lebih dari satu domain teks bahasa Indonesia. Perbandingan pada tabel ini bersifat deskriptif karena dataset, task, skema pelabelan, dan metrik evaluasi antarpenelitian tidak sepenuhnya identik.

Tabel 7. Perbandingan Hasil Penelitian Ini dengan Studi Sejenis

Studi	Model	Domain & Bahasa	Metrik	Hasil
Imaduddin et al. [6]	IndoBERT	Ulasan aplikasi kesehatan (Indonesia)	Akurasi	0,900

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1310
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9867
Hal 1301-1313
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Studi	Model	Domain & Bahasa	Metrik	Hasil
Kang dan Choi [9]	FinBERT	Berita keuangan (Bahasa Inggris, domain asli)	-	Kompetitif terhadap GPT-4o
Penelitian ini	IndoBERT-Fine-Tuned	Berita finansial (Indonesia)	Macro F1	0,863
Penelitian ini	IndoLEM-Fine-Tuned	Berita finansial (Indonesia)	Macro F1	0,843
Penelitian ini	FinBERT-Fine-Tuned	Berita finansial (Indonesia, hasil translasi)	Macro F1	0,437

3.7 Pembahasan: Analisis Perbedaan Kinerja IndoBERT, IndoLEM, dan FinBERT

Keunggulan tipis namun signifikan IndoBERT-Fine-Tuned atas IndoLEM-Fine-Tuned pada *full validation*, disertai performa yang berdekatan pada gold subset, dapat dijelaskan dari karakteristik korpus *pre-training* masing-masing model pada domain berita finansial Indonesia. IndoBERT-base-p1 dari IndoNLU dan IndoBERT-base-uncased dari IndoLEM sama-sama merupakan model BERT berbasis arsitektur Transformer-base dengan 12 encoder layer dan dimensi hidden state 768. Walaupun korpus *pre-training* kedua model berbeda, IndoBERT dari IndoNLU dilatih pada Indo4B yang mencakup berbagai genre teks termasuk berita umum [16], sementara IndoLEM lebih fokus pada teks formal jurnalistik [4], keduanya telah cukup terpapar struktur kalimat berita ekonomi Indonesia. Pada ukuran dataset 9.015 artikel dalam penelitian ini, kedua model menunjukkan performa yang relatif berdekatan, dengan IndoBERT memperoleh keunggulan tipis atas IndoLEM. Keberagaman genre dalam Indo4B tampaknya memberikan keuntungan tipis namun konsisten bagi IndoBERT, sehingga perbedaan prediksinya dengan IndoLEM signifikan berdasarkan uji McNemar ($p = 0{,}0002$), meskipun secara magnitudo selisih Macro F1 antara kedua model hanya sekitar 2,0 poin persentase. Hasil ini melengkapi studi attribution authorship berbasis IndoBERT oleh Saputra dan Riccosan [21] yang melaporkan IndoBERT dan IndoLEM memberikan performa berdekatan pada teks jurnalistik Indonesia, dengan tambahan temuan bahwa perbedaan kecil tersebut dapat menjadi signifikan secara statistik pada ukuran sampel yang cukup besar.

Sebaliknya, performa rendah FinBERT pada konteks ini diduga berkaitan dengan dua sumber degradasi yang saling memperkuat. Sumber pertama adalah *translation noise*: meskipun entitas finansial Indonesia seperti IHSG, BBRI, dan OJK dipertahankan dalam bentuk aslinya melalui strategi entity preservation, banyak nuansa idiomatik bahasa Indonesia tetap hilang dalam proses translasi. Sebagai contoh, ungkapan ‘saham anjlok 5%’ dalam teks asli memiliki muatan negatif kuat, namun setelah translasi menjadi ‘stocks dropped 5%’, FinBERT cenderung mempersepsikannya sebagai berita netral-faktual. Hal ini terkonfirmasi pada pola kesalahan Sub-bab 3.4, di mana recall FinBERT-Fine-Tuned pada kelas Negatif hanya 0,344 yang menunjukkan model kesulitan mengenali sentimen negatif dan cenderung mengklasifikasikannya ke kelas lain. Sumber kedua adalah *domain mismatch* pada level konten: FinBERT dilatih pada laporan keuangan korporat dan komunikasi bisnis berbahasa Inggris [3], yang memiliki struktur narasi formal-teknikal berbeda dengan berita konsumen yang lazim dijumpai pada portal Indonesia. Berita finansial Indonesia kerap menggunakan gaya naratif yang lebih kasual dan langsung, kerap mencampurkan istilah teknis dengan istilah informal khas pasar seperti ‘cuan’ atau ‘nyangkut’. Gaya naratif ini tidak terwakili dalam korpus *pre-training* FinBERT, sehingga walaupun sudah di-*fine-tune* pada teks terjemahan, model tetap kesulitan menangkap polaritas yang sesungguhnya.

Implikasi praktis dari temuan ini adalah bahwa untuk tugas analisis sentimen berita finansial Indonesia, pendekatan model bahasa-spesifik dengan *fine-tuning* pada data domain lebih efektif dibanding pendekatan domain-spesifik yang memerlukan cross-lingual transfer. Hasil komparasi pada studi ini menunjukkan bahwa IndoBERT-base-p1 menghasilkan Macro F1 tertinggi pada konfigurasi dataset ini, sementara IndoLEM-base-uncased menghasilkan performa berdekatan dan dapat menjadi alternatif kompetitif terutama jika ketersediaan model atau kompatibilitas tooling menjadi pertimbangan praktis. Hal ini selaras dengan rekomendasi dari beberapa studi domain-spesifik Indonesia lainnya [22], [23] yang menekankan pentingnya kesesuaian bahasa pada lapisan *pre-training*. Untuk pengembangan selanjutnya, pendekatan *continued pre-training* pada IndoBERT atau IndoLEM menggunakan korpus berita finansial Indonesia berskala besar berpotensi menggabungkan keunggulan bahasa-spesifik dengan keunggulan domain-spesifik, sebagaimana berhasil ditunjukkan oleh Raza et al. pada konteks fake news detection [24]. Subset gold standard pada studi ini telah diperluas mencakup 432 artikel human-labeled yang sepenuhnya *held-out* dari training, namun perluasan lebih lanjut beserta peningkatan kualitas anotasi melalui pendekatan multi-annotator atau LLM-as-judge berpotensi mengangkat Cohen’s Kappa lexicon-human di atas 0,5 sehingga gap full-vs-gold pada model BERT dapat diperkecil dan klaim generalisasi menjadi lebih kuat.

Keterbatasan penelitian ini adalah anotasi manusia dilakukan oleh satu annotator sehingga reliabilitas antar-annotator belum dapat dievaluasi. Penelitian selanjutnya dapat melibatkan beberapa annotator independen dan proses adjudication untuk meningkatkan reliabilitas subset berlabel manusia.

4. KESIMPULAN

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1311
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Penelitian ini melakukan analisis komparatif tiga model BERT *pre-trained*, yaitu IndoBERT, IndoLEM, dan FinBERT, dalam konfigurasi *fine-tuning* untuk klasifikasi sentimen berita finansial Indonesia pada dataset 9.015 artikel yang dilabeli melalui pendekatan hybrid *weak-supervision*. IndoBERT-Fine-Tuned mencapai performa tertinggi dengan Macro F1 sebesar 0,863, diikuti IndoLEM-Fine-Tuned pada 0,843, dengan perbedaan prediksi antara kedua model yang signifikan berdasarkan uji McNemar ($p = 0,0002$), sementara kedua model secara signifikan mengungguli FinBERT-Fine-Tuned (Macro F1 = 0,437) dan *baseline* klasik TF-IDF ($p < 0,0001$) yang konsisten dengan kemungkinan pengaruh *translation noise* dan *domain mismatch* pada FinBERT. Evaluasi tambahan pada subset gold standard berlabel manusia (N = 432) mengungkap penurunan performa substansial pada model berlabel lexicon (35,40 poin persentase Accuracy untuk IndoBERT-Fine-Tuned dan 32,89 poin persentase untuk IndoLEM-Fine-Tuned), mengindikasikan adanya kemungkinan pengaruh pola label lexicon terhadap performa model pada label training sehingga evaluasi pada gold subset perlu selalu dilaporkan bersama full validation. Berdasarkan temuan tersebut, IndoBERT-Fine-Tuned menghasilkan Macro F1 tertinggi pada validation set penuh dalam konfigurasi dataset dan pelabelan yang digunakan, dengan IndoLEM-Fine-Tuned menghasilkan performa yang berdekatan sebagai alternatif kompetitif, sementara pengembangan selanjutnya diarahkan pada validasi k-fold cross-validation, *continued pre-training* pada korpus berita finansial Indonesia berskala besar, dan peningkatan kualitas label training untuk memperkecil gap antara full validation dan gold subset.

REFERENCES

[1] J. Hartanto, T. Liundi, R. Sutoyo, dan E. W. Andangsari, "ID-SMSA: Indonesian stock market dataset for sentiment analysis," *Data Brief*, vol. 60, hlm. 111571, Jun 2025, doi: 10.1016/j.dib.2025.111571.

[2] A. Alamsyah *dkk.*, "Deciphering news sentiment and stock price relationships in Indonesian companies: an AI-based exploration of industry affiliation and news co-occurrence," *Discover Artificial Intelligence*, vol. 5, no. 1, hlm. 87, Jun 2025, doi: 10.1007/s44163-025-00350-5.

[3] A. H. Huang, H. Wang, dan Y. Yang, "FinBERT" : A Large Language Model for Extracting Information from Financial Text*," *Contemporary Accounting Research*, vol. 40, no. 2, hlm. 806–841, Mei 2023, doi: 10.1111/1911-3846.12832.

[4] F. Koto, A. Rahimi, J. H. Lau, dan T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," dalam *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, hlm. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[5] E. Yulianti dan N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single- sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, hlm. 3579–3589, Okt 2024, doi: 10.11591/eei.v13i5.8032.

[6] H. Imaduddin, F. Y. A'la, dan Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140813.

[7] H. Ahmadian, T. F. Abidin, H. Riza, dan K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, Jan 2024, doi: 10.1155/2024/2826773.

[8] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, dan M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, hlm. 1071, Feb 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[9] J. W. Kang dan S. Y. Choi, "Comparative Investigation of GPT and FinBERT's Sentiment Analysis Performance in News Across Different Sectors," *Electronics (Switzerland)*, vol. 14, no. 6, Mar 2025, doi: 10.3390/electronics14061090.

[10] N. Merayo, J. Vegas, C. Llamas, dan P. Fernández, "Social Network Sentiment Analysis Using Hybrid Deep Learning Models," *Applied Sciences*, vol. 13, no. 20, hlm. 11608, Okt 2023, doi: 10.3390/app132011608.

[11] A. Khan, H. Zhang, N. Boudjellal, A. Ahmad, dan M. Khan, "Improving Sentiment Analysis in Election-Based Conversations on Twitter with ElecBERT Language Model," *Computers, Materials & Continua*, vol. 76, no. 3, hlm. 3345–3361, 2023, doi: 10.32604/cmc.2023.041520.

[12] E. Dave dan A. Chowanda, "IPerFEX-2023: Indonesian personal financial entity extraction using indoBERT-BiGRU-CRF model," *J. Big Data*, vol. 11, no. 1, hlm. 139, Sep 2024, doi: 10.1186/s40537-024-00987-6.

[13] A. Kunaefi, Z. Abidin, dan R. Kusumawati, "KLASIFIKASI BERITA HOAKS BAHASA INDONESIA MENGGUNAKAN INDOBERT FINE-TUNING DENGAN PENDEKA-TAN FOCAL LOSS PADA DATA TIDAK SEIMBANG," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 2, hlm. 1706–1714, Mei 2025, doi: 10.29100/jipi.v10i2.7811.

[14] A. R. Hanum *et al.*, "Analisis Kinerja Algoritma Klasifikasi Teks Bert dalam Mendeteksi Berita Hoaks," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 3, hlm. 537–546, Jul 2024, doi: 10.25126/jtiik.938093.

[15] A. Jazuli, Widowati, dan R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Applied Sciences*, vol. 15, no. 1, hlm. 172, Des 2024, doi: 10.3390/app15010172.

[16] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," dalam *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*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, hlm. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[17] L. Geni, E. Yulianti, dan D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 3, hlm. 746–757, Agu 2023, doi: 10.26555/jiteki.v9i3.26490.

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1312
This Journal is licensed under a Creative Commons Attribution 4.0 International License

[18] I. Daqiqil ID, H. Saputra, S. Syamsudhuha, R. Kurniawan, dan Y. Andriyani, "Sentiment analysis of student evaluation feedback using transformer-based language models," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 2, hlm. 1127, Nov 2024, doi: 10.11591/ijeecs.v36.i2.pp1127-1139.

[19] A. Areshey dan H. Mathkour, "Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model," *Sensors*, vol. 23, no. 11, hlm. 5232, Mei 2023, doi: 10.3390/s23115232.

[20] D. U. Iswavigra, V. D. Setiawan, M. Ulfa, dan B. Ommr, "Sentiment Analysis Using Bidirectional Encoder Representations from Transformers for Indonesian Stock Price Prediction with Long Short-Term Memory and Gated Recurrent Unit Models," *Jurnal Teknik Informatika (Jutif)*, vol. 7, no. 2, hlm. 961–976, Apr 2026, doi: 10.52436/1.jutif.2026.7.2.5383.

[21] K. E. Saputra dan R. Riccosan, "Indonesian news article authorship attribution multilabel multiclass classification using IndoBERT," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 4, hlm. 4688, Des 2024, doi: 10.11591/ijai.v13.i4.pp4688-4694.

[22] A. Riyadi, M. Kovacs, U. Serdült, dan V. Kryssanov, "IndoGovBERT: A Domain-Specific Language Model for Processing Indonesian Government SDG Documents," *Big Data and Cognitive Computing*, vol. 8, no. 11, hlm. 153, Nov 2024, doi: 10.3390/bdcc8110153.

[23] C. Shaw, P. LaCasse, dan L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, hlm. 22, Mar 2025, doi: 10.1007/s13278-025-01439-6.

[24] S. Raza, D. Paulen-Patterson, dan C. Ding, "Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data," *Knowl. Inf. Syst.*, vol. 67, no. 4, hlm. 3267–3292, Apr 2025, doi: 10.1007/s10115-024-02321-1.

Copyright © 2026 The **Naufal Latif Ramadhan**, Page 1313
This Journal is licensed under a Creative Commons Attribution 4.0 International License