

## Analisis Sentimen Teks *Code-Mixed* Bahasa Indonesia-Jawa Menggunakan Metode *Fine-Tuning* Model Nusbart

<sup>1</sup>Sopian Syauri, <sup>2</sup>Sofi Defiyanti, <sup>3</sup>Dadang Yusup  
<sup>1,2,3</sup>Informatika, Universitas Singaperbangsa Karawang, Karawang

E-mail: <sup>1</sup>2210631170104@student.unsika.ac.id, <sup>2</sup>sofi.defiyanti@unsika.ac.id,  
<sup>3</sup>dadang.dyf@staff.unsika.ac.id

### ABSTRAK

Pesatnya pertumbuhan pengguna platform X di Indonesia memicu lahirnya fenomena *code-mixing* antara Bahasa Indonesia dan Bahasa Jawa dalam komunikasi digital. Teks campuran ini mengandung nuansa linguistik lokal yang sulit diproses oleh model NLP berbasis bahasa umum. Penelitian ini mengimplementasikan model NusaBERT dengan teknik *fine-tuning* untuk melakukan analisis sentimen pada teks *code-mixed* Indonesia-Jawa yang dikumpulkan dari platform X. Data sebanyak 1.685 teks diperoleh melalui teknik *crawling* menggunakan kata kunci berbahasa Jawa, kemudian dilabeli secara otomatis menggunakan AI (Gemini dan Claude) dengan validasi manual oleh penutur asli. Eksperimen dilakukan dalam sembilan skenario yang mengombinasikan tiga variasi pembagian data (70:30, 80:20, 90:10) dan tiga nilai *learning rate* ( $2 \times 10^{-5}$ ,  $3 \times 10^{-5}$ ,  $5 \times 10^{-5}$ ). Hasil terbaik diperoleh pada Skenario 3 dengan pembagian data 70:30 dan *learning rate*  $5 \times 10^{-5}$ , menghasilkan nilai *Accuracy* 0,8538, *Precision* 0,8539, *Recall* 0,8538, dan *F1-Score* 0,8504. Penelitian ini membuktikan bahwa NusaBERT yang telah dilatih pada korpus bahasa daerah Indonesia mampu menangani kompleksitas linguistik teks *code-mixed* Indonesia-Jawa secara efektif.

**Kata kunci :** Analisis Sentimen, *Code-Mixed*, NusaBERT, *Fine-Tuning*

### ABSTRACT

The rapid growth of X platform users in Indonesia has triggered a widespread phenomenon of *code-mixing* between Indonesian and Javanese in digital communication. Such mixed-language text contains local linguistic nuances that are difficult to process by general-purpose NLP models. This study implements the NusaBERT model with a *fine-tuning* technique to perform sentiment analysis on Indonesian-Javanese *code-mixed* text collected from the X platform. A total of 1,685 texts were obtained through *crawling* using Javanese keywords, then automatically labeled using AI tools (Gemini and Claude) with manual validation by native speakers. Experiments were conducted across nine scenarios combining three data split variations (70:30, 80:20, 90:10) and three *learning rate* values ( $2 \times 10^{-5}$ ,  $3 \times 10^{-5}$ ,  $5 \times 10^{-5}$ ). The best result was achieved in Scenario 3 with a 70:30 data split and a *learning rate* of  $5 \times 10^{-5}$ , yielding *Accuracy* of 0.8538, *Precision* of 0.8539, *Recall* of 0.8538, and *F1-Score* of 0.8504. This research demonstrates that NusaBERT, pre-trained on Indonesian regional language corpora, is effective in handling the linguistic complexity of Indonesian-Javanese *code-mixed* text.

**Keyword :** Sentiment Analysis, *Code-Mixed*, NusaBERT, *Fine-Tuning*

## 1. PENDAHULUAN

Indonesia menempati posisi ke-3 sebagai negara dengan jumlah pengguna platform X terbanyak di dunia, dengan pengguna aktif mencapai 25,2 juta jiwa (Kemp, 2025). Karakteristik platform X yang mengutamakan kecepatan dan batasan karakter mendorong terciptanya aliran data tekstual yang sangat masif, spontan, dan kaya akan konteks emosional. Data ini menjadi aset berharga yang dapat diekstraksi nilainya melalui Analisis Sentimen untuk memahami polaritas emosi masyarakat secara otomatis (Gozali et al., 2026).

Tantangan utama dalam pemrosesan teks media sosial Indonesia adalah fenomena *code-mixing*, yaitu percampuran Bahasa Indonesia dengan bahasa daerah dalam satu tuturan (Sirait et al., 2026). Bahasa Jawa merupakan bahasa daerah dengan penutur terbanyak di Indonesia, mencakup sekitar 42% dari total populasi (Badan Pusat Statistik, 2024), sehingga campuran Indonesia-Jawa sangat lazim ditemukan di media sosial. Penggunaan ekspresi seperti “Lagi mumet mikir tugas, padahal deadline wis cedek” menunjukkan integrasi sintaksis dan leksikal yang kompleks dan sulit dianalisis oleh model NLP standar akibat ambiguitas makna serta fenomena *Out-of-Vocabulary* (OOV) (Purnomo & Sutopo, 2024).

Pendekatan pembelajaran mesin konvensional seperti K-Nearest Neighbor dan Naive Bayes memiliki keterbatasan karena sangat bergantung pada feature engineering manual yang tidak fleksibel terhadap variasi dialek (Prasetyo et al., 2023). Kehadiran *Pre-trained Language Models* (PLMs) berbasis arsitektur Transformer, khususnya BERT, membawa revolusi signifikan melalui mekanisme *self-attention* yang mampu memahami konteks dua arah (Devlin et al., 2019). Namun, model multibahasa

global seperti bert-base-multilingual-cased sering kali kurang optimal untuk menangani nuansa lokal, metafora, dan slang bahasa daerah di Indonesia karena minimnya representasi data lokal (Wongso et al., 2025).

NusaBERT hadir sebagai solusi yang dikembangkan dari IndoBERT dengan melakukan ekspansi kosakata sebanyak 10.000 target ukuran baru, termasuk 1.511 token unik yang tidak tumpang tindih, guna menangkap nuansa linguistik bahasa daerah yang selama ini tergolong low-resource (Cahyawijaya et al., 2023). Penelitian terdahulu menunjukkan bahwa model IndoBERT yang dikombinasikan dengan SVM mampu mencapai akurasi 95,91% pada teks media sosial (Gozali et al., 2026), sedangkan perbandingan model *Transformer* pada analisis sentimen bahasa daerah menunjukkan NusaBERT bersaing kompetitif pada domain tertentu (Ramdani et al., 2025).

Kesenjangan penelitian yang ada adalah belum adanya penelitian yang secara khusus mengeksplorasi kemampuan NusaBERT melalui *fine-tuning* pada teks *code-mixed* bilingual Indonesia-Jawa yang dikumpulkan langsung dari platform X menggunakan pendekatan pelabelan berbantuan AI.

Penelitian ini bertujuan untuk: (1) mengimplementasikan *fine-tuning* model NusaBERT pada teks *code-mixed* Indonesia-Jawa, dan (2) mengetahui nilai performa model berdasarkan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

## 2. LANDASAN TEORI

### 2.1 *Code-Mixed* dan Tantangan NLP

*Code-mixing* merupakan fenomena sociolinguistik di mana seorang penutur menyisipkan serpihan bahasa lain dalam satu tuturan yang sama (Sirait et al., 2026). Dalam konteks NLP, teks *code-*

*mixed* memiliki tingkat kesulitan yang jauh lebih tinggi dibandingkan teks monolingual karena tiga tantangan utama: (1) *Syntactic Irregularity*, yaitu pencampuran dua tata bahasa yang menciptakan pola kalimat yang sulit diprediksi; (2) *Out-of-Vocabulary* (OOV), di mana model gagal mengenali kosakata bahasa Jawa terutama yang mengalami perubahan bentuk dialek; dan (3) ambiguitas kata, yaitu kata-kata yang memiliki penulisan sama namun makna berbeda antara Bahasa Indonesia dan Jawa (Hidayatullah et al., 2024).

## 2.2 Arsitektur Transformer & BERT

Arsitektur Transformer diperkenalkan oleh Vaswani et al. melalui penelitian “*Attention is All You Need*”. Transformer menggunakan mekanisme *Self-Attention* yang memungkinkan pemrosesan paralel melalui perhitungan *Scaled Dot-Product Attention* (Vaswani et al., 2017):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

BERT (*Bidirectional Encoder Representations from Transformers*) adalah model yang menonjol karena penerapan pelatihan dua arah (*bidirectional*) yang mendalam, sehingga mampu mempertimbangkan konteks dari sisi kiri maupun kanan secara bersamaan (Devlin et al., 2019). Fase *pre-training* BERT menggunakan dua strategi: *Masked Language Model* (MLM) dan *Next Sentence Prediction* (NSP). Proses *fine-tuning* memungkinkan adaptasi model untuk tugas spesifik secara efisien dengan rekomendasi *hyperparameter batch size* 16 atau 32, *learning rate* Adam pada rentang  $2 \times 10^{-5}$  hingga  $5 \times 10^{-5}$ , dan *epoch* sebanyak 2 hingga 4 (Devlin et al., 2019).

## 2.3 NusaBERT

NusaBERT adalah model bahasa berbasis Transformer yang dikembangkan

dari IndoBERT untuk mengatasi tantangan linguistik unik Indonesia (Cahyawijaya et al., 2023). Pengembangannya mencakup: (1) ekspansi kosakata (*vocabulary expansion*) dari 30.521 menjadi 32.032 token dengan menambahkan 1.511 token unik bahasa daerah; (2) penggunaan objektif *Masked Language Modeling* (MLM) dari RoBERTa menggantikan NSP; dan (3) pelatihan pada korpus 13 bahasa termasuk Bahasa Jawa dari CulturaX dan Wikipedia. Pengujian pada benchmark IndoRobusta-Blend menunjukkan NusaBERT memiliki ketahanan kompetitif dan peningkatan signifikan pada tugas analisis sentimen dibandingkan IndoBERT dasar (Cahyawijaya et al., 2023).

## 2.4 Metrik Evaluasi

Evaluasi performa menggunakan empat metrik berbasis *Confusion Matrix* (Koto et al., 2021):

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

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

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

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

Untuk klasifikasi multikelas, digunakan *Macro-Averaged F1-Score* yang memberikan bobot sama pada setiap kelas sentimen tanpa mempedulikan frekuensi kemunculannya (Cahyawijaya et al., 2023).

## 3. METODOLOGI

Penelitian ini mengikuti prosedur eksperimental NLP yang tersusun secara sistematis melalui lima tahapan: (1) Akuisisi Data, (2) Pra-pemrosesan Data, (3) Pelabelan Data, (4) Implementasi Model, dan (5) Evaluasi Performa.

### 3.1 Akuisisi Data

Data diperoleh melalui teknik *crawling* pada platform X menggunakan alat bantu *tweet-harvest* yang dijalankan pada lingkungan Google Colab. Pengumpulan data dibatasi pada rentang 1 Januari 2024 hingga 31 Desember 2025 dan difokuskan pada koordinat geografis -7,6298, 111,5239 (Madiun) dengan radius 350 km yang mencakup wilayah Jawa Tengah, DIY, dan Jawa Timur. Data dikumpulkan dalam 10 batch dengan kata kunci yang terbagi dalam lima kategori: (1) kombinasi ekspresi sentimen dan kata kasar (asu, bajingan, koplak); (2) kosakata sapaan dan slang emosional (jancok, lur); (3) ekspresi harapan dan syukur (maturnuwun, mugo); (4) partikel penegas (ojo, tenan, wis); dan (5) kata keterangan aktivitas sehari-hari (mumet, turu, ambyar). Total data yang terkumpul adalah 5.064 baris.

### 3.2 Pra-pemrosesan Data

Tahapan pra-pemrosesan dilakukan secara berurutan sebagai berikut:

1. Penyaringan bahasa: Memfilter data berdasarkan kolom lang == 'in' sehingga dataset berkurang dari 5.064 menjadi 2.352 baris.
2. Pembersihan komponen platform: Menghapus URL, mention (@username), dan hashtag (#) menggunakan teknik Regular Expression (Regex).
3. Pembersihan karakter: Mengganti karakter non-alfanumerik (tanda baca, emotikon) dengan spasi menggunakan Regex.
4. *Case folding*: Mengubah seluruh huruf menjadi huruf kecil (lowercase).
5. Reduksi karakter duplikat: Membatasi pengulangan karakter beruntun dan menghapus spasi berlebih.
6. Penghapusan duplikat: Mengeliminasi data identik sehingga tersisa 2.114 baris.

7. Validasi manual: Membaca ulang seluruh data untuk mengidentifikasi dan menghapus teks berbahasa Sunda yang ikut terjaring, sehingga tersisa 1.685 baris.
8. Normalisasi kata: Menstandarkan kata-kata non-formal dan singkatan (seperti bgt → banget, yg → yang) menggunakan kamus normalisasi (*slang dictionary*) yang disusun secara manual. Tahap stemming dan *stopword removal* secara sadar tidak dilakukan untuk mempertahankan partikel bahasa Jawa (seperti wis, ora, wae) yang berfungsi sebagai penegas sentimen.

### 3.3 Pelabelan Data

Data dilabeli ke dalam tiga kelas: Positif, Negatif, dan Netral, menggunakan pendekatan *AI-assisted labeling* dengan dua model LLM, yaitu Gemini dan Claude. Sebuah prompt terstruktur dikirimkan ke kedua model yang memuat instruksi peran sebagai pakar linguistik komputasional, kriteria pelabelan dengan contoh, serta instruksi khusus untuk menangani konteks *code-mixed*, slang, dan sarkasme. Hasil dari kedua model dibandingkan dan ditemukan 485 baris data dengan label yang tidak konsisten (*discrepancy*). Baris-baris tersebut kemudian dilabeli ulang secara manual (*human-in-the-loop*) dengan bantuan penutur asli sebagai validator akhir. Distribusi label akhir yang diperoleh adalah: Netral 909 (53,95%), Negatif 541 (32,11%), dan Positif 235 (13,95%).

### 3.4 Implementasi Model dan *Fine-Tuning*

Model dasar yang digunakan adalah LazarusNLP/NusaBERT-base yang menggunakan pustaka Hugging Face *Transformers* berbasis PyTorch. Konfigurasi *fine-tuning* yang tetap pada semua skenario meliputi *batch size* 16, *weight decay* 0,01, *max sequence length* 512 token, dan maksimum 20 *epoch* dengan mekanisme *Early Stopping*

(*patience* = 3). Model terbaik pada setiap skenario dipilih berdasarkan skor F1 tertinggi pada data validasi. Eksperimen dilakukan pada sembilan kombinasi skenario dari tiga variasi *data split* (70:30, 80:20, 90:10) dan tiga nilai *learning rate* ( $2 \times 10^{-5}$ ,  $3 \times 10^{-5}$ ,  $5 \times 10^{-5}$ ).

### 3.5 Evaluasi Performa

Evaluasi dilakukan menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score* yang dihitung terhadap data validasi pada masing-masing skenario. *F1-Score* menjadi fokus utama evaluasi karena distribusi kelas dalam dataset tidak seimbang, di mana kelas Netral mendominasi sebesar 53,95%.

## 4. HASIL DAN PEMBAHASAN

### 4.1 Hasil Evaluasi Performa

Rekapitulasi hasil evaluasi dari sembilan skenario eksperimen disajikan pada Tabel 1 dan Tabel 2 berikut.

Tabel 1. Hasil Evaluasi Performa

Sken	Data Split	Learning Rate	Training Loss	Evaluation Loss
1	70:30	$2 \times 10^{-5}$	0.057573	0.751166
2	70:30	$3 \times 10^{-5}$	0.281864	0.461836
3	70:30	$5 \times 10^{-5}$	0.147395	0.619406
4	80:20	$2 \times 10^{-5}$	0.420535	0.426427
5	80:20	$3 \times 10^{-5}$	0.340533	0.421775
6	80:20	$5 \times 10^{-5}$	0.994330	0.983333
7	90:10	$2 \times 10^{-5}$	0.413158	0.516284
8	90:10	$3 \times 10^{-5}$	0.345099	0.638326
9	90:10	$5 \times 10^{-5}$	0.606913	0.625966

Tabel 2. Hasil Evaluasi Metrik Performa

Sken	Accuracy	Precision	Recall	F1 Score
1	0.8281	0.8324	0.8281	0.8291
2	0.8202	0.8193	0.8202	0.8185
3	0.8538	0.8539	0.8538	0.8504
4	0.8427	0.8424	0.8427	0.8425
5	0.8368	0.8352	0.8368	0.8356

Sken	Accuracy	Precision	Recall	F1 Score
6	0.5638	0.7281	0.5638	0.4348
7	0.8284	0.8366	0.8284	0.8221
8	0.8047	0.8044	0.8047	0.8027
9	0.7692	0.6752	0.7692	0.7047

### 4.2 Analisis Model Terbaik

Skenario 3 (*data split* 70:30, *learning rate*  $5 \times 10^{-5}$ ) ditetapkan sebagai model terbaik dengan *F1-Score* tertinggi sebesar 0,8504. Nilai *Precision* 0,8539 mengindikasikan kemampuan model dalam meminimalisir kesalahan klasifikasi *false positive*, sementara *Recall* 0,8538 menunjukkan bahwa model mampu menangkap sebagian besar sentimen yang ada dalam dataset secara benar. Capaian ini membuktikan bahwa NusaBERT berhasil mempelajari fitur semantik dan kontekstual secara optimal melalui *fine-tuning*.

Secara keseluruhan, tujuh dari sembilan skenario menghasilkan nilai metrik yang stabil di atas 80%, yang membuktikan efektivitas tinggi NusaBERT dalam menangani kompleksitas linguistik teks campuran bahasa Indonesia-Jawa. Keberhasilan ini didukung oleh fase *pre-training* NusaBERT yang telah melibatkan bahasa-bahasa daerah Indonesia, sehingga model memiliki pemahaman semantik yang mendalam terhadap leksikal bahasa Jawa (Cahyawijaya et al., 2023). Keputusan untuk tidak melakukan *stemming* dan *stopword removal* juga terbukti krusial; partikel bahasa Jawa seperti “*wis*”, “*ora*”, dan “*wae*” dipertahankan karena berfungsi sebagai penegas sentimen yang esensial (Khairani et al., 2024).

### 4.3 Analisis Skenario Bermasalah

Penurunan performa drastis terjadi pada Skenario 6 (*Accuracy* 0,5638, *F1-Score* 0,4348). Penggunaan *learning rate* tertinggi ( $5 \times 10^{-5}$ ) pada pembagian data 80:20 menyebabkan model gagal



## 5. KESIMPULAN

- a. Model NusaBERT-base berhasil diimplementasikan dan dioptimasi melalui *fine-tuning* untuk menangani karakteristik unik teks *code-mixed* Indonesia-Jawa. Fase *pre-training* NusaBERT yang melibatkan bahasa-bahasa daerah Indonesia memberikan pemahaman semantik yang mendalam terhadap leksikal bahasa Jawa, termasuk partikel penegas seperti “ora”, “wis”, dan “tenan” yang esensial bagi penentuan polaritas sentimen.
- b. Skenario 3 (*data split* 70:30, *learning rate*  $5 \times 10^{-5}$ ) ditetapkan sebagai model terbaik dengan hasil: *Accuracy* 0,8538, *Precision* 0,8539, *Recall* 0,8538, dan *F1-Score* 0,8504. Tujuh dari sembilan skenario menghasilkan nilai *F1-Score* di atas 80%, membuktikan efektivitas dan ketahanan NusaBERT dalam menangani kompleksitas linguistik teks campuran. Penggunaan *learning rate* yang terlalu besar ( $5 \times 10^{-5}$ ) pada pembagian data tertentu (skenario 6 dan 9) terbukti memicu *Majority Class Bias* dan penurunan performa yang signifikan, sehingga pemilihan *learning rate* harus dilakukan dengan hati-hati.

## 6. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Dosen Pembimbing yang telah memberikan bimbingan, arahan, dan masukan selama proses penelitian dan penulisan artikel ini berlangsung. Penulis juga mengucapkan terima kasih kepada penutur asli Bahasa Jawa yang telah bersedia meluangkan waktu untuk membantu proses validasi pelabelan data,

serta kepada seluruh pihak di lingkungan institusi yang telah memberikan dukungan fasilitas dan motivasi selama penelitian ini dilaksanakan.

## DAFTAR PUSTAKA

- Badan Pusat Statistik. (2024). *Profil Suku dan Keragaman Bahasa Daerah Hasil Long Form Sensus Penduduk 2020*.
- Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G. I., Wilie, B., Koto, F., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., Santoso, J., Moeljadi, D., Wirawan, C., Hudi, F., Wicaksono, M. S., Parmonangan, I. H., Alfina, I., Putra, I. F., Rahmadani, S., ... Purwarianti, A. (2023). NusaCrowd: Open Source Initiative for Indonesian NLP Resources. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 13745–13818. <https://doi.org/10.18653/v1/2023.findings-acl.868>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1*, 4171–4186.
- Gozali, M. R. F., Tibyani, & Brata, D. W. (2026). Analisis Sentimen Pengguna Twitter / X Terhadap Fenomena #KaburAjaDulu Menggunakan Metode Support Vector Machine dan IndoBERT Embedding. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 10(2), 2548–2964. <http://j-ptiik.ub.ac.id>
- Hidayatullah, A. F., Apong, R. A., Lai, D. T. C., & Qazi, A. (2024). Word

- Level Language Identification in Indonesian-Javanese-English Code-Mixed Text. *Procedia Computer Science*, 244, 105–112. <https://doi.org/10.1016/j.procs.2024.10.183>
- Hossin, M., & Sulaiman, M. N. (2015). A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS. *International Journal of Data Mining & Knowledge Management Process (IJDMP)*, 5(2), 1–11.
- Kemp, S. (2025). *X Users, Stats, Data & Trends for 2025*. Data Reportal. <https://datareportal.com/essential-x-stats>
- Khairani, U., Mutiawani, V., & Ahmadian, H. (2024). Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 11(4), 887–894. <https://doi.org/10.25126/jtiik.1148315>
- Koto, F., Lau, J. H., & Baldwin, T. (2021). INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 10660–10668. <https://doi.org/10.18653/v1/2021.emnlp-main.833>
- Mosbach, M., Andriushchenko, M., & Klakow, D. (2021). ON THE STABILITY OF FINE-TUNING BERT: MISCONCEPTIONS, EXPLANATIONS, AND STRONG BASELINES. *ICLR*.
- Prasetyo, S. D., Hilabi, S. S., & Nurapriani, F. (2023). Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN. *Jurnal KomtekInfo*, 1–7. <https://doi.org/10.35134/komtekinfo.v10i1.330>
- Purnomo, T. D., & Sutopo, J. (2024). COMPARISON OF PRE-TRAINED BERT-BASED TRANSFORMER MODELS FOR REGIONAL LANGUAGE TEXT SENTIMENT ANALYSIS IN INDONESIA. *International Journal Science and Technology*, 3(3), 11–21. <https://doi.org/10.56127/ijst.v3i3.1739>
- Ramdani, D., Irfan, M., & Lukman, N. (2025). Perbandingan Kinerja IndoBERT, IndoRoBERTa dan NusaBERT dalam Analisis Sentimen Isu LGBT di Media Sosial X. *Information Sistem Journal*, 8(2), 91–107.
- Sirait, G. A., Br Lingga, Y. E., Hutasoit, T. M., Tarigan, B., & Sari, Y. (2026). Analisis Alih Kode dan Campur Kode dalam Komunikasi Mahasiswa di Media Sosial serta Perannya terhadap Pembentukan Identitas Bahasa Generasi Muda. *DEIKTIS: Jurnal Pendidikan Bahasa Dan Sastra*, 6(1), 50–54. <https://doi.org/10.53769/deiktis.v6i1.2941>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). The Transformer. *IEEE Industry Applications Magazine*, 8(1), 8–15. <https://doi.org/10.1109/2943.974352>
- Wongso, W., Setiawan, D. S., Limcorn, S., & Joyoadikusumo, A. (2025). NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural. *Proceedings - International Conference on Computational Linguistics, COLING*, 10–26. <https://github.com/LazarusNLP/NusaBERT>
- Zhang, C., Bengio, S., Hardt, M., Recht,

B., & Vinyals, O. (2017).  
Understanding Deep Learning (Still)  
Requires Rethinking Generalization.  
*COMMUNICATIONS OF THE  
ACM*, 64, 107–115.  
<https://doi.org/10.1145/3446776>

