

Analisis Perbandingan Metode *Natural Language Processing* (NLP) Klasik (*Naïve Bayes*, *Support Vektor Machine*) Dan Model *Bidirectional Encoder Representations From Transformers* (BERT) Dalam Klasifikasi Teks Berbahasa Indonesia

Ummul Farihah¹, Zainal Azis²

^{1,2} Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received September 23, 2026
Revised September 23, 2026
Accepted September 24, 2026

Kata Kunci:

Analisis Sentimen,
Klasifikasi Teks,
IndoBERT,
Support Vector Machine,
Uji McNemar

Keywords:

*Sentiment Analysis,
Text Classification,
IndoBERT,
Support Vector Machine,
McNemar's Test*

ABSTRAK

Pertumbuhan data teks berbahasa Indonesia di platform digital menuntut sistem klasifikasi sentimen yang akurat dan dapat dipertanggungjawabkan. Penelitian ini membandingkan Multinomial Naïve Bayes, Support Vector Machine (SVM) linear, dan IndoBERT pada dataset SmSA IndoNLU, sekaligus menguji signifikansi perbedaan performa dengan Uji McNemar. Dataset berjumlah 12.760 sampel yang terbagi menjadi 11.000 data latih, 1.260 data validasi, dan 500 data uji dengan distribusi kelas tidak seimbang. Model klasik menggunakan TF-IDF unigram-bigram dan preprocessing eksplisit, sedangkan IndoBERT menggunakan tokenisasi WordPiece. Macro F1-Score digunakan sebagai metrik utama. IndoBERT mencapai akurasi 88,60% dan Macro F1-Score 83,77%, diikuti SVM 76,20% dan 71,68%, serta Naïve Bayes 66,00% dan 60,99%. Pada kelas netral, F1-Score IndoBERT mencapai 67,67%, lebih tinggi daripada SVM 55,56% dan Naïve Bayes 46,48%. Seluruh perbedaan antar model signifikan secara statistik ($p < 0,05$). Hasil eksperimen diintegrasikan ke platform web Ummu NLP Lab yang menghasilkan metrik identik dengan notebook laboratorium serta mendukung reproduksibilitas eksperimen secara konsisten dan transparan.

ABSTRACT

The growth of Indonesian-language text on digital platforms requires accurate and accountable sentiment classification systems. This study compares Multinomial Naïve Bayes, linear Support Vector Machine (SVM), and IndoBERT on the SmSA IndoNLU dataset and tests performance differences using McNemar's Test. The dataset contains 12,760 samples divided into 11,000 training samples, 1,260 validation samples, and 500 test samples with an imbalanced class distribution. Classical models use unigram-bigram TF-IDF with explicit preprocessing, whereas IndoBERT uses WordPiece tokenization. Macro F1-Score is the primary evaluation metric. IndoBERT achieves 88.60% accuracy and 83.77% Macro F1-Score, followed by SVM with 76.20% and 71.68%, and Naïve Bayes with 66.00% and 60.99%. For the neutral class, IndoBERT reaches 67.67% F1-Score, compared with SVM at 55.56% and Naïve Bayes at 46.48%. All pairwise differences are statistically significant ($p < 0.05$). The results are integrated into Ummu NLP Lab, which reproduces the laboratory notebook metrics identically while preserving experimental reproducibility across environments.

This is an open access article under the [CC BY](https://creativecommons.org/licenses/by/4.0/) license



Corresponding Author:

Ummul Fariyah
Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara,
Medan, Indonesia
Email: ummulfariyah20@gmail.com

1. PENDAHULUAN

Perkembangan platform digital di Indonesia menghasilkan volume data teks yang terus meningkat dan mendorong kebutuhan pemrosesan bahasa alami untuk mengubah ulasan, komentar, serta opini daring menjadi informasi yang dapat dianalisis secara sistematis. Penetrasi pengguna internet yang telah melampaui 200 juta jiwa memperbesar skala pertukaran opini digital sehingga analisis manual menjadi semakin tidak memadai [1]. Dalam konteks ini, analisis sentimen menjadi salah satu penerapan Natural Language Processing (NLP) yang relevan untuk mengelompokkan orientasi opini ke dalam kelas sentimen dan mendukung pengambilan keputusan berbasis data.

Pemrosesan teks berbahasa Indonesia memiliki tantangan tersendiri karena teks digital banyak memuat bahasa informal, singkatan tidak baku, pencampuran istilah, negasi, serta struktur kalimat yang tidak selalu mengikuti kaidah gramatikal [2]. Karakteristik tersebut memengaruhi representasi fitur dan kemampuan model dalam menangkap makna. Model Multinomial Naïve Bayes menawarkan proses pelatihan yang ringan, tetapi menggunakan asumsi independensi fitur sehingga urutan kata dan relasi antarterm tidak dimodelkan secara eksplisit [3]. Support Vector Machine (SVM) mampu bekerja efektif pada ruang fitur berdimensi tinggi dengan membentuk hyperplane ber-margin optimal [4], sedangkan BERT menggunakan representasi kontekstual dwiarah melalui arsitektur Transformer sehingga setiap token diproses dengan mempertimbangkan konteks sekitarnya [5].

Benchmark NLP bahasa Indonesia telah menyediakan dataset dan model pralatih yang memungkinkan evaluasi antarmodel secara lebih terstandar. IndoNLU menyediakan benchmark termasuk Sentiment Multi-level Sentence Analysis (SmSA) dan model IndoBERT untuk berbagai tugas pemahaman bahasa Indonesia [8]. Penelitian lain mengenai IndoBERT juga menunjukkan manfaat representasi monolingual untuk pemrosesan bahasa Indonesia [9]. Namun, perbandingan model sering berhenti pada selisih metrik agregat, padahal perbedaan prediksi antarmodel pada data uji yang sama perlu diuji secara inferensial agar klaim keunggulan tidak hanya bergantung pada variasi sampel [6], [7].

Kesenjangan penelitian dalam skripsi ini dirumuskan pada tiga aspek. Pertama, keterbatasan penggunaan validasi statistik formal dalam perbandingan model. Kedua, evaluasi terhadap kelas minoritas netral masih perlu dianalisis secara granular karena distribusi SmSA tidak seimbang dan akurasi keseluruhan dapat menutupi kelemahan pada kelas dengan jumlah sampel lebih sedikit [10]. Ketiga, reproduksibilitas eksperimen memerlukan lingkungan yang mampu menyimpan konfigurasi, hasil evaluasi, dan proses komputasi secara konsisten; aspek ini sejalan dengan pentingnya transparansi dan reproduksibilitas dalam penelitian machine learning [11].

Berdasarkan kesenjangan tersebut, penelitian membandingkan Multinomial Naïve Bayes, SVM linear, dan IndoBERT pada dataset SmSA IndoNLU menggunakan Accuracy, Precision, Recall, Weighted F1-Score, dan Macro F1-Score; menguji signifikansi perbedaan prediksi menggunakan Uji McNemar pada taraf $\alpha = 0,05$; mengevaluasi ketahanan terhadap kelas minoritas netral; serta mengimplementasikan seluruh hasil ke dalam platform web Ummu NLP Lab. Pendekatan ini

menempatkan evaluasi performa, validasi statistik, karakteristik komputasi, dan reproduksibilitas dalam satu kerangka komparatif.

2. METODE

2.1 Dataset dan Desain Eksperimen

Penelitian menggunakan dataset sekunder Sentiment Multi-level Sentence Analysis (SmSA) dari benchmark IndoNLU [8]. Dataset berjumlah 12.760 kalimat teranotasi dengan tiga kelas sentimen, yaitu positif, negatif, dan netral. Pembagian data mengikuti predefined split untuk menjaga konsistensi evaluasi: 11.000 sampel sebagai data latih, 1.260 sampel sebagai data validasi, dan 500 sampel sebagai data uji terisolasi. Data validasi digunakan untuk penalaan hiperparameter, sedangkan data uji hanya digunakan pada tahap evaluasi akhir dan Uji McNemar. Distribusi dataset menunjukkan dominasi kelas positif dan proporsi kelas netral yang jauh lebih kecil sebagaimana dirangkum pada Tabel 1.

Tabel 1. Distribusi Sampel Dataset SmSA IndoNLU

Kelas Sentimen	Data Latih	Data Validasi	Data Uji	Total	Persentase
Positif	6.416	735	208	7.359	57,67%
Negatif	3.436	393	204	4.033	31,61%
Netral	1.148	132	88	1.368	10,72%
Total	11.000	1.260	500	12.760	100,00%

Sumber: Benchmark Dataset IndoNLU SmSA [8].

2.2 Dual Preprocessing Pipeline

Perbedaan arsitektur model klasik dan Transformer ditangani melalui dual preprocessing pipeline. Untuk Multinomial Naïve Bayes dan SVM, teks melalui case folding, pembersihan berbasis regular expression, normalisasi kata tidak baku, penghapusan stopword, serta ekstraksi fitur TF-IDF. Representasi fitur menggunakan kombinasi unigram dan bigram dengan ngram_range=(1,2), min_df=5, dan normalisasi L2. Pembobotan term mengikuti prinsip TF-IDF yang menekankan term yang informatif pada dokumen [12]. Tahapan ini ditujukan untuk mereduksi derau dan dimensionalitas sebelum pemodelan klasik.

Jalur IndoBERT mempertahankan struktur kalimat, kata tugas, serta tanda baca karena representasi kontekstual Transformer bergantung pada urutan token dan konteks dwiarah [5], [13]. Tokenisasi menggunakan tokenizer indobenchmark/indobert-base-p1 berbasis WordPiece dengan penyisipan token [CLS] dan [SEP], panjang maksimum 128 token, padding, truncation, dan attention mask. Perbedaan strategi preprocessing ini digunakan agar setiap model menerima representasi masukan yang sesuai dengan karakteristik arsitekturnya.

2.3 Model dan Penalaan Hiperparameter

Tiga model dibandingkan dalam lingkungan eksperimen yang sama. Multinomial Naïve Bayes menggunakan fitur TF-IDF dengan parameter smoothing α . SVM menggunakan fitur TF-IDF dan mengevaluasi parameter penalti C serta kernel linear dan RBF; implementasi model klasik menggunakan scikit-learn [14]. IndoBERT menggunakan indobenchmark/indobert-base-p1 dan melakukan fine-tuning untuk klasifikasi tiga kelas. Penalaan parameter dilakukan hanya pada 1.260 sampel data validasi untuk mencegah data leakage, dengan ruang pencarian parameter disusun secara sistematis [15]. Ruang parameter dan konfigurasi terbaik disajikan pada Tabel 2.

Tabel 2. Konfigurasi Hiperparameter Terbaik

Model	Ruang Parameter	Konfigurasi Terbaik	Macro F1 Validasi
Multinomial Naïve Bayes	$\alpha \in \{0,1; 0,5; 1,0; 1,5; 2,0\}$	$\alpha = 0,1$	81,20%
SVM	$C \in \{0,1; 1,0; 10,0\};$ kernel $\in \{\text{linear, RBF}\}$	$C = 1,0; \text{kernel} = \text{linear};$ $\text{gamma} = \text{scale}$	82,72%
IndoBERT	$\text{lr} \in \{2 \times 10^{-5}; 5 \times 10^{-5}\};$ batch $\in \{8; 16\}$	$\text{lr} = 2 \times 10^{-5}; \text{batch} = 8;$ epoch = 3	90,66%

2.4 Metrik Evaluasi dan Uji McNemar

Evaluasi kinerja dilakukan pada 500 sampel data uji menggunakan Accuracy, Precision, Recall, Weighted F1-Score, dan Macro F1-Score. Macro F1-Score ditetapkan sebagai metrik primer karena memberikan bobot yang setara pada setiap kelas dan lebih informatif ketika distribusi kelas tidak seimbang [10]. Selain metrik agregat, Precision, Recall, dan F1-Score per kelas dianalisis untuk menilai kemampuan setiap model mendeteksi kelas netral sebagai kelas minoritas.

Perbedaan prediksi antarmodel diuji menggunakan Uji McNemar karena seluruh model dievaluasi pada sampel uji yang sama. Pengujian berfokus pada pasangan prediksi diskordan, yaitu kondisi ketika satu model benar dan model lain salah. Nilai p dihitung dengan metode exact binomial dan hipotesis nol ditolak jika $p < 0,05$ [6], [16]. Dengan desain ini, perbandingan tidak hanya didasarkan pada besar selisih metrik, tetapi juga pada bukti statistik mengenai konsistensi perbedaan prediksi.

2.5 Implementasi dan Reprodusibilitas

Eksperimen menggunakan ekosistem Python dengan PyTorch, Transformers, scikit-learn, Pandas, NumPy, dan SciPy. Hasil pemodelan diintegrasikan ke platform Ummu NLP Lab berbasis Flask dan basis data SQLite dengan mode Write-Ahead Logging. Platform menyimpan konfigurasi dan hasil eksperimen, menyediakan modul pengelolaan dataset, pelatihan model, prediksi, Uji McNemar, serta telemetri komputasi. Random seed dikunci pada nilai 42 dan hasil pada platform divalidasi silang terhadap hasil notebook laboratorium untuk menguji konsistensi dan reprodusibilitas [11].

3. HASIL DAN PEMBAHASAN

3.1 Distribusi Data dan Hasil Penalaan

Analisis deskriptif memperlihatkan ketidakseimbangan kelas pada dataset SmSA. Dari 12.760 sampel, kelas positif berjumlah 7.359 sampel (57,67%), kelas negatif 4.033 sampel (31,61%), dan kelas netral 1.368 sampel (10,72%). Pada data uji, kelas netral hanya berjumlah 88 dari 500 sampel. Kondisi ini menjadikan Macro F1-Score dan metrik per kelas penting agar performa pada kelas minoritas tidak tertutupi oleh dominasi kelas mayoritas.

Penalaan pada data validasi menghasilkan $\alpha=0,1$ untuk Multinomial Naïve Bayes dengan Macro F1-Score validasi 81,20%. SVM memperoleh konfigurasi terbaik $C=1,0$ dengan kernel linear dan Macro F1-Score 82,72%. IndoBERT memperoleh konfigurasi learning rate 2×10^{-5} , batch size 8, dan 3 epoch dengan Macro F1-Score validasi 90,66%. Konfigurasi tersebut selanjutnya digunakan pada pelatihan akhir dan evaluasi data uji.

3.2 Perbandingan Kinerja Model

Hasil evaluasi data uji menunjukkan perbedaan yang konsisten antartetiga paradigma. IndoBERT mencapai akurasi 88,60% dan Macro F1-Score 83,77%, lebih tinggi daripada SVM linear dengan akurasi 76,20% dan Macro F1-Score 71,68%, serta Multinomial Naïve Bayes dengan akurasi 66,00% dan Macro F1-Score 60,99%. Weighted F1-Score juga menunjukkan pola yang sama, yaitu 87,57%

pada IndoBERT, 75,49% pada SVM, dan 64,37% pada Naïve Bayes. Ringkasan metrik utama dan performa kelas netral ditampilkan pada Tabel 3.

Tabel 3. Perbandingan Kinerja Model pada Data Uji

Metrik	Naïve Bayes	SVM Linear	IndoBERT
Akurasi	66,00%	76,20%	88,60%
Macro F1-Score	60,99%	71,68%	83,77%
Weighted F1-Score	64,37%	75,49%	87,57%
Precision kelas netral	61,11%	71,43%	100,00%
Recall kelas netral	37,50%	45,45%	51,14%
F1-Score kelas netral	46,48%	55,56%	67,67%

Perbedaan terbesar terlihat pada kelas minoritas netral. Naïve Bayes memperoleh Precision 61,11%, Recall 37,50%, dan F1-Score 46,48%. SVM meningkatkan nilai tersebut menjadi Precision 71,43%, Recall 45,45%, dan F1-Score 55,56%. IndoBERT menghasilkan Precision 100,00%, Recall 51,14%, dan F1-Score 67,67%. Pada 88 sampel netral di data uji, Naïve Bayes mendeteksi 33 sampel dengan benar, SVM 40 sampel, dan IndoBERT 45 sampel. Hasil ini menunjukkan peningkatan kemampuan representasi kontekstual dalam memisahkan kelas netral, walaupun Recall IndoBERT tetap menunjukkan bahwa 43 sampel netral masih salah diklasifikasikan.

3.3 Validasi Signifikansi Statistik

Uji McNemar pada 500 sampel data uji menghasilkan p-value yang jauh di bawah $\alpha=0,05$ untuk seluruh pasangan model. Perbandingan SVM terhadap Naïve Bayes menghasilkan $p=3,3216 \times 10^{-8}$, IndoBERT terhadap SVM menghasilkan $p=9,9415 \times 10^{-11}$, dan IndoBERT terhadap Naïve Bayes menghasilkan $p=9,6596 \times 10^{-23}$. Hasil lengkap disajikan pada Tabel 4. Seluruh hipotesis nol ditolak, sehingga perbedaan prediksi antarmodel dinyatakan signifikan secara statistik.

Tabel 4. Hasil Uji McNemar ($\alpha = 0,05$)

Pasangan Model	b/c	p-value	Keputusan
SVM vs Naïve Bayes	69 / 18	$3,3216 \times 10^{-8}$	H0 ditolak
IndoBERT vs SVM	79 / 17	$9,9415 \times 10^{-11}$	H0 ditolak
IndoBERT vs Naïve Bayes	130 / 17	$9,6596 \times 10^{-23}$	H0 ditolak

Karena terdapat tiga perbandingan berpasangan, skripsi juga mempertimbangkan ambang Bonferroni sebesar $0,05/3$ atau sekitar 0,0167. Seluruh p-value masih jauh di bawah ambang tersebut sehingga kesimpulan signifikansi tidak berubah. Temuan ini memperkuat interpretasi bahwa selisih performa bukan hanya akibat variasi acak pada pembagian data uji [6], [7].

3.4 Efisiensi Komputasi dan Implikasi Implementasi

Keunggulan akurasi IndoBERT disertai kebutuhan komputasi yang lebih besar. Waktu pelatihan Multinomial Naïve Bayes hanya 0,1527 detik dengan ukuran model 0,4229 MB, sedangkan SVM memerlukan 66,3377 detik dan ukuran 1,2899 MB. IndoBERT memerlukan 917,12 detik atau sekitar 15,3 menit pada GPU dengan ukuran model 474,72 MB. Latensi inferensi per teks tercatat 0,91 ms untuk Naïve Bayes pada CPU, 1,32 ms untuk SVM pada CPU, dan 15,04 ms untuk IndoBERT pada GPU; pada CPU, IndoBERT sekitar 85 ms. Perbandingan karakteristik operasional dirangkum pada Tabel 5.

Tabel 5. Karakteristik Operasional Model

Aspek	Naïve Bayes	SVM Linear	IndoBERT
Waktu pelatihan	0,1527 detik	66,3377 detik	917,12 detik (~15,3 menit)
Latensi inferensi	0,91 ms (CPU)	1,32 ms (CPU)	15,04 ms (GPU); ~85 ms (CPU)
Ukuran model	0,4229 MB	1,2899 MB	474,72 MB
F1 kelas netral	46,48%	55,56%	67,67%

Hasil tersebut menunjukkan adanya trade-off antara performa prediktif dan kebutuhan sumber daya. Naïve Bayes dan SVM memiliki keunggulan pada waktu pelatihan, ukuran model, dan latensi, sedangkan IndoBERT memberikan ketahanan semantik yang lebih tinggi terutama pada kelas netral. SVM menjadi opsi yang berada di antara kedua ekstrem tersebut: performanya lebih tinggi daripada Naïve Bayes dengan kebutuhan komputasi yang tetap jauh lebih ringan daripada IndoBERT. Pola ini konsisten dengan karakteristik SVM sebagai pengklasifikasi efektif untuk ruang fitur teks berdimensi tinggi [4], [17] dan dengan kemampuan Transformer dalam membentuk representasi kontekstual [5], [13], [18].

3.5 Konsistensi Platform Ummu NLP Lab

Seluruh model dan pipeline evaluasi kemudian diintegrasikan ke platform web Ummu NLP Lab. Validasi silang menunjukkan metrik Accuracy, Macro F1-Score, Weighted F1-Score, dan F1-Score per kelas yang dihasilkan platform identik dengan hasil notebook laboratorium, dengan selisih 0,00%. Pengujian fungsional sistem juga menunjukkan sebelas dari dua belas skenario memperoleh status valid sepenuhnya, sedangkan satu skenario Prediction Lab memperoleh status valid dengan catatan terkait latensi IndoBERT pada CPU. Konsistensi ini menunjukkan bahwa penguncian random seed, preprocessing terpisah, dan implementasi algoritma pada platform mampu mempertahankan hasil eksperimen secara reproduksibel [11].

4. KESIMPULAN

Perbandingan pada dataset SmSA IndoNLU menunjukkan bahwa IndoBERT menghasilkan performa tertinggi dengan akurasi 88,60% dan Macro F1-Score 83,77%, diikuti SVM linear dengan 76,20% dan 71,68%, serta Multinomial Naïve Bayes dengan 66,00% dan 60,99%. Pada kelas minoritas netral, F1-Score meningkat dari 46,48% pada Naïve Bayes menjadi 55,56% pada SVM dan 67,67% pada IndoBERT. Uji McNemar mengonfirmasi bahwa seluruh perbedaan antarmodel signifikan secara statistik pada $\alpha=0,05$, bahkan tetap berada di bawah ambang koreksi Bonferroni.

Strategi preprocessing yang berbeda terbukti sesuai dengan karakteristik masing-masing arsitektur: model klasik menggunakan pembersihan eksplisit dan TF-IDF, sedangkan IndoBERT mempertahankan struktur kalimat untuk tokenisasi WordPiece dan pemodelan kontekstual. Dari sisi komputasi, Naïve Bayes dan SVM jauh lebih ringan, sedangkan IndoBERT membutuhkan GPU, waktu pelatihan, dan ruang penyimpanan yang lebih besar untuk memperoleh performa prediktif yang lebih tinggi. Implementasi pada Ummu NLP Lab menghasilkan metrik identik dengan notebook laboratorium sehingga mendukung transparansi dan reproduksibilitas eksperimen. Pengembangan berikutnya dapat diarahkan pada teknik penyeimbangan kelas, optimasi hiperparameter lanjutan, pengujian lintas domain, dan eksplorasi model Transformer yang lebih besar sebagaimana disarankan dalam skripsi.

REFERENSI

- [1] We Are Social and Meltwater, Digital 2026: Global Overview Report. We Are Social Ltd., 2026.

- [2] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in Proc. 2017 Int. Conf. Advanced Computer Science and Information Systems (ICACSYS), 2017, pp. 233–238.
- [3] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in AAAI-98 Workshop on Learning for Text Categorization, 1998, pp. 41–48.
- [4] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT 2019, 2019, pp. 4171–4186.
- [6] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, vol. 10, no. 7, pp. 1895–1923, 1998.
- [7] E. Alpaydin, "Combined 5×2 cv F test for comparing supervised classification learning algorithms," Neural Computation, vol. 11, no. 8, pp. 1885–1892, 1999.
- [8] B. Wilie et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in Proc. AACL-IJCNLP 2020, 2020, pp. 843–857.
- [9] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in Proc. COLING 2020, 2020, pp. 757–770.
- [10] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.
- [11] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program," Journal of Machine Learning Research, vol. 22, no. 164, pp. 1–20, 2021.
- [12] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.
- [13] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems 30, 2017, pp. 5998–6008.
- [14] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [15] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.
- [16] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.
- [17] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in Proc. 10th European Conf. Machine Learning (ECML 98), 1998, pp. 137–142.
- [18] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, F. Koto, A. Rahimi, P. Bahar, and A. Purwarianti, "Indo4B: An open-source Indonesian language foundation model," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3230–3242, 2021.