QOMARUNA
Journal of
Multidisciplinary Studies

E-ISSN: 3026-5266
https://jurnal.uqgresik.ac.id/index.php/qjms

Perbandingan Kinerja IndoBERT dan IndoRoBERTa dengan Penerapan SMOTE dalam Deteksi Ujaran Kebencian Berbahasa Indonesia

Muhammad Mutawakkil Alallah[1*], Indra Rosyidah[2]

[1] Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia, 65144
[2] Universitas Nurul Jadid, Probolinggo, Indonesia, 67291
* Korespondensi: 240605220005@student.uin-malang.ac.id

Received: 21 April 2026
Revised: 17 May 2026
Accepted: 21 May 2026

Citation:

Alallah, M. M. ., & Rosyidah, I. . Perbandingan Kinerja IndoBERT dan IndoRoBERTa dengan Penerapan SMOTE dalam Deteksi Ujaran Kebencian Berbahasa Indonesia. *Qomaruna: Journal of Multidisciplinary Studies*, *3*(2), 149–163. https://doi.org/10.62048/qjms.v3i2.167

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by-nc-sa/4.0/).

ABSTRACT

The rapid growth of social media in Indonesia has increased digital interaction while also giving rise to hate speech issues that affect communication quality and social stability. This study aims to compare the performance of two Transformer-based models, IndoBERT and IndoRoBERTa, in Indonesian-language hate speech classification and to evaluate the effect of the SMOTE data balancing technique. The dataset consisted of Indonesian-language Twitter data that underwent preprocessing and was divided using an 80:20 stratified train-test split. Model training was conducted through fine-tuning, while evaluation employed accuracy, precision, recall, and F1-score metrics. The results show that IndoRoBERTa outperformed IndoBERT across all evaluation metrics and was more effective in reducing classification errors. The application of SMOTE also improved the models' ability to detect minority classes, particularly in terms of recall. These findings indicate that the combination of Transformer-based models and data balancing techniques is effective in improving both classification accuracy and class balance in hate speech detection. Furthermore, the results suggest that the combination of IndoRoBERTa and SMOTE has strong potential to support the development of more accurate and adaptive automated content moderation systems for Indonesian-language social media platforms.

Keywords: *hate speech, NLP, Transformer, IndoBERT, IndoRoBERTa*

ABSTRAK

Perkembangan media sosial di Indonesia meningkatkan interaksi digital sekaligus memunculkan masalah ujaran kebencian yang berdampak pada kualitas komunikasi dan stabilitas sosial. Penelitian ini bertujuan membandingkan performa dua model Transformer, IndoBERT dan IndoRoBERTa, dalam klasifikasi ujaran kebencian berbahasa Indonesia serta mengevaluasi pengaruh teknik data balancing SMOTE. Dataset berupa data Twitter berbahasa Indonesia yang telah melalui tahap pre-processing dan dibagi menggunakan *stratified train-test split* 80:20. Pelatihan model dilakukan melalui *fine-tuning*, sedangkan evaluasi menggunakan accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa IndoRoBERTa lebih unggul dibandingkan IndoBERT pada seluruh metrik serta mampu mengurangi kesalahan klasifikasi. Penerapan SMOTE juga meningkatkan kemampuan model dalam mendeteksi kelas minoritas, terutama pada recall. Temuan ini menunjukkan bahwa kombinasi model Transformer dan teknik data balancing efektif meningkatkan akurasi serta keseimbangan klasifikasi ujaran kebencian. Temuan ini mengindikasikan bahwa kombinasi IndoRoBERTa dan SMOTE berpotensi mendukung pengembangan sistem

moderasi konten otomatis yang lebih akurat dan adaptif pada media sosial berbahasa Indonesia.

Kata kunci: ujaran kebencian, pemrosesan bahasa alami, transformer, IndoBERT, IndoRoBERTa

Pendahuluan

Perkembangan teknologi informasi dan komunikasi telah mendorong peningkatan signifikan penggunaan media sosial sebagai sarana interaksi digital di masyarakat (Suciati, 2024; Hakimi et al., 2025). Platform seperti Twitter, YouTube, dan Facebook menjadi ruang utama bagi pengguna untuk menyampaikan opini, berdiskusi, serta berbagi informasi secara real-time (Pananookooln et al., 2023). Namun demikian, perkembangan ini juga diikuti oleh meningkatnya fenomena ujaran kebencian (*hate speech*) dan bahasa kasar (*abusive language*) yang berpotensi menimbulkan konflik sosial, diskriminasi, serta gangguan terhadap stabilitas sosial (Yoon et al., 2021; Purnomo & Sutopo, 2024). Studi terbaru menunjukkan bahwa toksisitas daring telah menjadi permasalahan serius dalam ekosistem digital Indonesia dan berdampak langsung terhadap kesejahteraan psikologis pengguna serta kualitas interaksi sosial di ruang digital (Selvaraj et al., 2022; Tsugawa & Watabe, 2023; Alamsyah & Sagama, 2024).

Dalam konteks bahasa Indonesia, deteksi ujaran kebencian menjadi tantangan yang kompleks karena karakteristik linguistik yang beragam (Tsugawa & Watabe, 2023). Penggunaan slang, campuran bahasa daerah, serta fenomena *code-switching* menyebabkan teks sulit dipahami secara literal maupun semantik (Bao & Gu, 2022). Selain itu, ujaran kebencian seringkali bersifat implisit dan kontekstual sehingga tidak dapat diidentifikasi secara efektif menggunakan pendekatan berbasis aturan (*rule-based*) (Shoeb & Melo, 2021). Oleh karena itu, diperlukan pendekatan berbasis *Natural Language Processing* (NLP) yang mampu memahami konteks bahasa secara lebih mendalam. Penelitian menunjukkan bahwa pendekatan berbasis Transformer memiliki kemampuan superior dalam menangkap representasi semantik teks dibandingkan metode tradisional berbasis *machine learning* (Ramos et al., 2024).

Perkembangan terbaru dalam NLP menunjukkan bahwa model berbasis Transformer, seperti BERT dan variannya, telah menjadi pendekatan dominan dalam tugas klasifikasi teks, termasuk deteksi ujaran kebencian (Sarkar et al., 2021; Amalia & Suyanto, 2024; Ghosh, 2025). Studi komprehensif terbaru menunjukkan bahwa model Transformer secara konsisten memberikan performa terbaik dalam berbagai tugas NLP karena kemampuannya dalam memahami konteks bidirectional dan representasi bahasa yang kompleks (Ramos et al., 2024). Dalam konteks bahasa Indonesia, penggunaan model seperti IndoBERT terbukti efektif dalam menangani kompleksitas linguistik lokal dan meningkatkan akurasi pada berbagai tugas NLP, termasuk analisis sentimen dan klasifikasi teks (Purnomo & Sutopo, 2024).

Lebih lanjut, penelitian terkini menunjukkan bahwa model berbasis Transformer juga efektif dalam deteksi ujaran kebencian pada media sosial, termasuk dalam lingkungan bahasa dengan sumber daya terbatas (*low-resource languages*). Model-model ini mampu mengatasi tantangan linguistik seperti variasi bahasa dan konteks implisit dengan lebih baik dibandingkan pendekatan sebelumnya (Fetahi et al., 2025). Dalam konteks Indonesia, beberapa penelitian telah menunjukkan efektivitas model berbasis Transformer dalam deteksi ujaran kebencian. Penelitian oleh (Amalia & Suyanto (2024) menunjukkan bahwa model BERT mampu meningkatkan performa deteksi offensive language dan hate speech dibandingkan pendekatan konvensional. Purnomo & Sutopo (2024) juga menemukan bahwa model berbasis BERT memberikan hasil yang lebih baik dibandingkan model deep learning tradisional dalam klasifikasi teks berbahasa Indonesia. Selain itu, Pamungkas & Purworini (2025) menunjukkan bahwa pendekatan Transformer efektif dalam mendeteksi hate speech pada tweet berbahasa Indonesia yang mengandung *code-mixing* dan variasi linguistik informal. Penelitian Fetahi et al. (2025) turut menegaskan bahwa Transformer memiliki kemampuan yang baik dalam menangani *low-resource language* melalui representasi konteks yang lebih mendalam. Temuan-temuan tersebut menunjukkan bahwa pendekatan Transformer memiliki potensi besar dalam meningkatkan akurasi sistem deteksi ujaran kebencian berbahasa Indonesia. Selain itu, tren penelitian global menunjukkan peningkatan

signifikan dalam kajian terkait hate speech sejak tahun 2021, dengan fokus utama pada deteksi otomatis menggunakan pendekatan berbasis deep learning dan Transformer (Nurindah et al., 2025). Penelitian lain juga menegaskan bahwa kombinasi teknik NLP dan model Transformer berperan penting dalam meningkatkan efektivitas sistem deteksi ujaran kebencian di media sosial (Alkomah & Ma, 2022).

Meskipun demikian, terdapat beberapa keterbatasan dalam penelitian terdahulu. Pertama, sebagian besar penelitian masih berfokus pada penggunaan satu model tanpa melakukan analisis komparatif antar model Transformer yang berbeda. Padahal, perbandingan performa antar model seperti IndoBERT dan IndoRoBERTa penting untuk menentukan pendekatan terbaik dalam konteks bahasa Indonesia. Kedua, permasalahan ketidakseimbangan data (*imbalanced data*) masih menjadi tantangan utama dalam klasifikasi ujaran kebencian. Distribusi data yang tidak seimbang menyebabkan model cenderung bias terhadap kelas mayoritas dan menurunkan kemampuan deteksi pada kelas minoritas. Beberapa penelitian menunjukkan bahwa teknik augmentasi data dan balancing dapat meningkatkan performa model, khususnya dalam meningkatkan nilai *recall* dan *F1-score* pada kelas minoritas (Pamungkas & Purworini, 2025).

Berdasarkan keterbatasan tersebut, penelitian ini menawarkan kebaruan (*novelty*) berupa integrasi analisis komparatif antara model IndoBERT dan IndoRoBERTa dengan penerapan teknik *data balancing* dalam satu *pipeline* eksperimental yang terstruktur. Berbeda dengan penelitian sebelumnya yang cenderung mengkaji aspek tersebut secara terpisah, penelitian ini tidak hanya membandingkan kinerja model, tetapi juga mengevaluasi secara sistematis pengaruh distribusi data terhadap performa klasifikasi, khususnya dalam meningkatkan kemampuan deteksi pada kelas minoritas. Selain itu, penelitian ini mengaitkan performa model dengan karakteristik linguistik khas bahasa Indonesia, seperti penggunaan *slang*, variasi informal, dan fenomena *code-switching*, sehingga memberikan analisis yang lebih kontekstual terhadap kinerja model Transformer dalam mendeteksi ujaran kebencian.

Dengan demikian, *research gap* dalam penelitian ini terletak pada kurangnya studi yang secara simultan melakukan (1) perbandingan kinerja model IndoBERT dan IndoRoBERTa, serta (2) integrasi teknik *data balancing* dalam pipeline model Transformer untuk meningkatkan performa klasifikasi ujaran kebencian berbahasa Indonesia. Sejalan dengan latar belakang dan *research gap* yang telah diuraikan, rumusan masalah dalam penelitian ini adalah: (1) bagaimana kinerja model IndoBERT dalam mendeteksi ujaran kebencian berbahasa Indonesia, (2) bagaimana kinerja model IndoRoBERTa dalam tugas yang sama, serta (3) bagaimana pengaruh penerapan teknik *data balancing* terhadap peningkatan performa kedua model tersebut.

Penelitian ini bertujuan untuk menganalisis dan membandingkan performa model IndoBERT dan IndoRoBERTa dalam klasifikasi ujaran kebencian pada media sosial Twitter, serta mengevaluasi efektivitas teknik *data balancing* dalam meningkatkan performa klasifikasi, khususnya pada kelas minoritas. Kontribusi penelitian ini meliputi dua aspek utama. Secara keilmuan, penelitian ini memberikan evaluasi komparatif terhadap dua model Transformer pada dataset bahasa Indonesia serta memperkaya kajian NLP terkait deteksi ujaran kebencian berbasis bahasa lokal. Secara praktis, hasil penelitian ini diharapkan dapat menjadi dasar pengembangan sistem moderasi konten otomatis yang lebih akurat dan adaptif dalam mengidentifikasi ujaran kebencian di media sosial.

Tinjauan Pustaka

Deteksi Ujaran Kebencian dan Natural Language Processing (NLP)

Deteksi ujaran kebencian (*hate speech detection*) merupakan salah satu cabang utama dalam Natural Language Processing (NLP) yang berfokus pada identifikasi dan klasifikasi teks berdasarkan kandungan semantik serta konteks linguistik (Tita, 2021; Rivadeneira, 2025). Dalam media sosial, ujaran kebencian umumnya muncul dalam bentuk ekspresi implisit yang mengandung unsur penghinaan, diskriminasi, atau permusuhan terhadap individu maupun kelompok tertentu (Kovács et al., 2021; Idris et al., 2026).

Pada konteks bahasa Indonesia, tantangan dalam deteksi ujaran kebencian menjadi semakin kompleks akibat karakteristik linguistik yang beragam, seperti penggunaan bahasa tidak baku, slang, serta fenomena *code-switching*. Kondisi tersebut menyebabkan pendekatan berbasis aturan (*rule-based*)

sulit menghasilkan performa yang optimal dalam memahami makna teks secara kontekstual (Okky & Budi, 2023).

Pendekatan awal dalam klasifikasi teks menggunakan metode *machine learning* seperti Naïve Bayes, Support Vector Machine (SVM), dan Logistic Regression dengan representasi fitur *bag-of-words* dan TF-IDF. Namun, pendekatan tersebut memiliki keterbatasan dalam menangkap hubungan semantik yang kompleks dalam teks (Ramos et al., 2024). Oleh karena itu, pendekatan berbasis *deep learning* mulai dikembangkan untuk meningkatkan kemampuan representasi konteks dalam teks.

Transformer dalam NLP

Perkembangan signifikan dalam NLP ditandai dengan munculnya arsitektur Transformer yang berbasis mekanisme *attention*. Model Transformer mampu memproses hubungan antar kata secara paralel dan memahami konteks secara *bidirectional*, sehingga menghasilkan representasi teks yang lebih kaya dan akurat (Przybyła & Soto, 2021; Przybyła & Soto, 2021; Vaswani et al., 2023).

Salah satu model yang paling berpengaruh adalah Bidirectional Encoder Representations from Transformers (BERT), yang dirancang untuk memahami konteks kata berdasarkan keseluruhan kalimat. Keunggulan utama BERT terletak pada kemampuan *contextual embedding* yang lebih baik dibandingkan pendekatan sebelumnya seperti CNN dan LSTM (Koto & Baldwin, 2020).

Dalam perkembangan selanjutnya, berbagai varian BERT dikembangkan untuk bahasa dan domain tertentu. Dalam konteks bahasa Indonesia, IndoBERT dan IndoRoBERTa merupakan model yang diadaptasi untuk menangani karakteristik linguistik lokal. IndoRoBERTa sebagai pengembangan dari RoBERTa menggunakan strategi pelatihan yang lebih optimal dengan penghapusan *next sentence prediction* dan peningkatan skala data pre-training, sehingga menghasilkan performa yang lebih stabil dalam berbagai tugas NLP.

Penelitian Terdahulu dan Kesenjangan Penelitian

Sejumlah penelitian menunjukkan bahwa model berbasis Transformer memiliki performa yang lebih unggul dibandingkan metode konvensional dalam berbagai tugas klasifikasi teks. Purnomo et al. (2024) menunjukkan bahwa IndoBERT mampu meningkatkan akurasi klasifikasi teks berbahasa Indonesia secara signifikan dibandingkan model berbasis LSTM dan CNN. Selain itu, IndoRoBERTa juga menunjukkan performa kompetitif melalui optimalisasi proses pre-training dan representasi bahasa yang lebih kuat.

Dalam konteks deteksi ujaran kebencian, penelitian oleh Fetahi et al. (2025) menunjukkan bahwa model Transformer mampu meningkatkan nilai F1-score secara signifikan, terutama pada bahasa dengan sumber daya terbatas (*low-resource language*). Hal ini menunjukkan bahwa Transformer lebih efektif dalam menangani variasi linguistik dan konteks implisit dibandingkan pendekatan sebelumnya.

Penelitian lain oleh Pamungkas & Purworini (2025) juga menunjukkan bahwa model berbasis Transformer seperti IndoBERT dan RoBERTa mampu melampaui performa model berbasis LSTM dan CNN dalam deteksi ujaran kebencian pada media sosial Indonesia. Selain itu, tren penelitian global menunjukkan peningkatan signifikan penggunaan Transformer dalam tugas hate speech detection sejak tahun 2021 (Nurindah et al., 2025).

Namun demikian, terdapat beberapa kesenjangan penelitian yang masih ditemukan. Pertama, sebagian besar penelitian masih berfokus pada satu model tanpa melakukan analisis komparatif yang mendalam antara IndoBERT dan IndoRoBERTa. Kedua, integrasi teknik penanganan ketidakseimbangan data seperti SMOTE masih belum banyak diterapkan secara sistematis dalam pipeline berbasis Transformer. Ketiga, penelitian yang menggabungkan analisis komparatif model dan teknik data balancing dalam satu kerangka eksperimen masih terbatas.

Oleh karena itu, penelitian ini diarahkan untuk mengisi kesenjangan tersebut melalui analisis komparatif IndoBERT dan IndoRoBERTa serta penerapan SMOTE dalam meningkatkan performa klasifikasi ujaran kebencian berbahasa Indonesia.

Metode

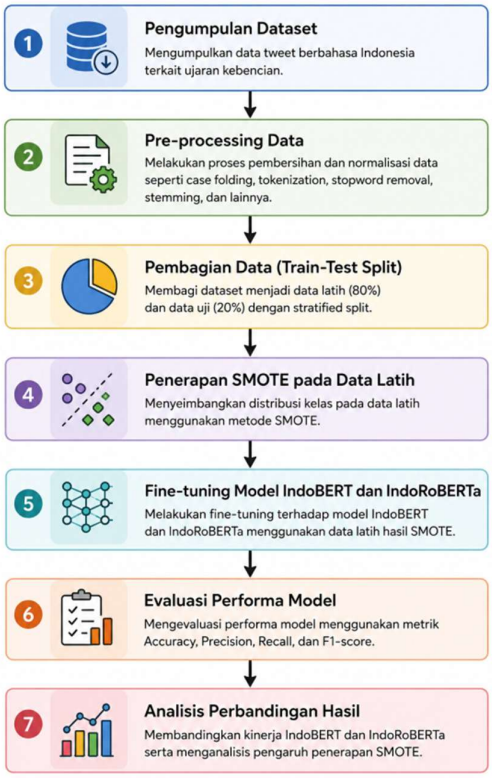
Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen untuk mengevaluasi performa model *Natural Language Processing* dalam tugas klasifikasi ujaran kebencian. Desain penelitian dilakukan secara komparatif dengan membandingkan dua model berbasis Transformer, yaitu IndoBERT dan IndoRoBERTa, serta menganalisis pengaruh penerapan teknik *data balancing* terhadap kinerja kedua model tersebut.

Eksperimen dilakukan melalui beberapa tahapan utama, yaitu pengumpulan dataset, praproses data, pembagian data, penerapan teknik *data balancing*, pelatihan model (*training*), pengujian model (*testing*), serta evaluasi performa menggunakan metrik klasifikasi.

Alur Penelitian

Secara keseluruhan, penelitian ini dilakukan melalui tujuh tahapan utama. Tahap pertama adalah pengumpulan dataset ujaran kebencian berbahasa Indonesia dari sumber yang telah tersedia. Selanjutnya dilakukan pre-processing untuk membersihkan dan menormalkan teks agar siap digunakan dalam proses pemodelan. Dataset kemudian dibagi menjadi data latih dan data uji menggunakan metode stratified train-test split dengan proporsi 80:20. Untuk mengatasi ketidakseimbangan kelas, teknik Synthetic Minority Oversampling Technique (SMOTE) diterapkan pada data latih. Setelah itu, model IndoBERT dan IndoRoBERTa dilatih melalui proses fine-tuning menggunakan dataset yang telah dipersiapkan. Performa kedua model dievaluasi menggunakan metrik accuracy, precision, recall, dan F1-score. Tahap terakhir adalah analisis perbandingan hasil untuk mengidentifikasi model yang memiliki performa terbaik serta mengevaluasi pengaruh penerapan SMOTE terhadap kemampuan klasifikasi ujaran kebencian. Alur penelitian secara sistematis ditunjukkan pada Gambar 1.



Gambar 1. Alur penelitian

Dataset dan Tahapan Eksperimen

Dataset yang digunakan dalam penelitian ini berasal dari platform Kaggle, yaitu *Indonesian Abusive and Hate Speech Twitter Text* yang dapat diakses melalui tautan https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text. Dataset ini berisi kumpulan teks berupa komentar dari media sosial Twitter yang ditulis dalam bahasa Indonesia dan telah melalui proses anotasi. Data diklasifikasikan ke dalam beberapa kategori, yaitu *hate speech*, *abusive language*, dan *non-hate speech*. Dalam implementasi penelitian ini, label digunakan untuk membedakan antara teks yang mengandung ujaran kebencian dan yang tidak, sehingga membentuk skema klasifikasi biner.

Secara karakteristik, dataset ini terdiri dari sekitar 13.169 data tweet berbahasa Indonesia dengan tipe anotasi *multi-label classification*. Label utama yang digunakan dalam penelitian ini adalah: (1) nilai 1 untuk menunjukkan keberadaan ujaran kebencian (*hate speech*), dan (2) nilai 0 untuk menunjukkan tidak adanya ujaran kebencian (*non-hate speech*).

Dataset ini dipilih karena beberapa alasan. Pertama, dataset menggunakan bahasa Indonesia sehingga relevan dengan konteks penelitian. Kedua, dataset memiliki anotasi label yang jelas dan terstruktur. Ketiga, dataset ini telah digunakan secara luas dalam berbagai penelitian NLP, khususnya dalam deteksi ujaran kebencian, sehingga mendukung aspek reusabilitas dan komparabilitas penelitian.

Selain itu, penggunaan dataset ujaran kebencian berbahasa Indonesia juga telah banyak dimanfaatkan dalam penelitian sebelumnya, salah satunya oleh Imaduddin et al. (2023). Penelitian tersebut menunjukkan bahwa dataset Twitter berbahasa Indonesia memiliki kompleksitas tinggi dan relevan untuk pengujian model klasifikasi teks.

Sebelum digunakan dalam proses pelatihan model, dataset dianalisis untuk mengetahui distribusi kelas guna mengidentifikasi adanya permasalahan *imbalanced data*, yang selanjutnya ditangani menggunakan teknik *data balancing* pada tahap eksperimen. Contoh beberapa sampel teks beserta labelnya disajikan pada Tabel 1.

Tabel 1. Contoh Dataset

text	label
- disaat semua cowok berusaha melacak perhatia...	1
RT USER: USER siapa yang telat ngasih tau elu?...	0
41. Kadang aku berfikir, kenapa aku tetap perc...	0
USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...	0
USER USER Kaum cebong kapir udah keliatan dong...	1

Pre-processing Data

Tahap *pre-processing* dilakukan untuk membersihkan dan menormalkan data teks agar siap digunakan dalam model NLP. Tahapan praproses meliputi:

1. **Cleaning**: Menghapus URL, mention (@user), hashtag, angka, dan simbol non-alfabet
2. **Case Folding**: Mengubah seluruh teks menjadi huruf kecil
3. **Tokenization**: Memecah kalimat menjadi token kata
4. **Stopword Removal**: Menghapus kata umum yang tidak memiliki makna signifikan
5. **Normalization**: Mengubah kata tidak baku menjadi kata baku

Setelah praproses, data dikonversi ke dalam format yang sesuai dengan tokenizer dari masing-masing model (IndoBERT dan IndoRoBERTa).

Pembagian Data

Dataset dibagi menjadi dua bagian:

- Data latih (training set)**:** 80%
- Data uji (testing set)**:** 20%

Pembagian dilakukan secara *stratified* untuk menjaga proporsi distribusi kelas pada masing-masing subset data.

Penanganan Imbalanced Data

Untuk mengatasi ketidakseimbangan data, penelitian ini menggunakan teknik *data balancing* berupa *Synthetic Minority Oversampling Technique* (SMOTE)**.** Teknik ini bekerja dengan menghasilkan data sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang.

Penerapan SMOTE dilakukan hanya pada data latih untuk menghindari *data leakage*. Dengan demikian, model dapat belajar secara lebih optimal terhadap kelas minoritas tanpa mempengaruhi data uji.

Model Transformer

Penelitian ini menggunakan dua model berbasis Transformer. Pertama adalah IndoBERT yang merupakan model *pre-trained language model* berbasis BERT yang dikembangkan khusus untuk bahasa Indonesia. Model ini mampu memahami konteks bahasa secara *bidirectional* dan efektif dalam berbagai tugas NLP. Kedua adalah IndoRoBERTa yang merupakan pengembangan dari RoBERTa yang diadaptasi untuk bahasa Indonesia. Model ini memiliki keunggulan pada strategi pelatihan yang lebih optimal, seperti penghapusan *next sentence prediction* dan penggunaan data pelatihan yang lebih besar.

Pelatihan Model

Model dilatih menggunakan pendekatan *fine-tuning* dengan parameter sebagai berikut:

- Epoch: 3–5
- Batch size: 16 atau 32
- Learning rate: 2e-5
- Optimizer: AdamW
- Loss function: Cross-Entropy Loss

Proses *fine-tuning* dilakukan dengan menyesuaikan bobot model pre-trained terhadap dataset yang digunakan dalam penelitian ini.

Evaluasi Model

Evaluasi performa model dilakukan menggunakan metrik klasifikasi sebagai berikut:

- Accuracy
- Precision
- Recall
- F1-score

Untuk klasifikasi, digunakan confusion matrix sebagai dasar perhitungan metrik evaluasi:

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

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

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

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

Evaluasi dilakukan pada data uji untuk memastikan kemampuan generalisasi model.

Analisis Perbandingan

Hasil dari kedua model akan dibandingkan berdasarkan nilai evaluasi yang diperoleh. Analisis difokuskan pada:

- Perbandingan performa IndoBERT vs IndoRoBERTa
- Dampak penerapan *data balancing*
- Kemampuan model dalam mendeteksi kelas minoritas

Hasil dan Pembahasan

Distribusi dan Pembagian Data

Dataset yang digunakan dalam penelitian ini berjumlah 13.169 data teks yang telah melalui tahap *pre-processing*. Selanjutnya, dataset dibagi menjadi data latih dan data uji menggunakan metode *train-test split* dengan perbandingan 80:20.

Hasil pembagian menunjukkan bahwa data latih berjumlah 10.535 data, sedangkan data uji sebanyak 2.634 data. Pembagian dilakukan dengan pendekatan *stratified sampling*, sehingga proporsi distribusi kelas pada data latih dan data uji tetap terjaga dan merepresentasikan distribusi data secara keseluruhan.

Distribusi label pada data latih menunjukkan bahwa kelas 0 (non-ujaran kebencian) berjumlah 6.086 data, sedangkan kelas 1 (ujaran kebencian) sebanyak 4.449 data. Sementara itu, pada data uji terdapat 1.522 data untuk kelas 0 dan 1.112 data untuk kelas 1.

Hasil tersebut mengindikasikan bahwa dataset memiliki ketidakseimbangan kelas, di mana jumlah data pada kelas mayoritas lebih tinggi dibandingkan kelas minoritas. Meskipun perbedaan ini tidak terlalu ekstrem, kondisi tersebut tetap berpotensi memengaruhi performa model dalam mendeteksi kelas minoritas.

Penggunaan teknik *stratified sampling* memastikan bahwa distribusi kelas tetap konsisten antara data latih dan data uji, sehingga evaluasi model dapat dilakukan secara lebih representatif dan tidak bias terhadap salah satu kelas. Distribusi dan Pembagian Data disajikan pada Tabel 2.

Tabel 2. Distribusi dan Pembagian Data

Dataset	Kelas 0	Kelas 1	Total
Train	6086	4449	10535
Test	1522	1112	2634

Hasil Pelatihan Model

Proses pelatihan model dilakukan melalui pendekatan *fine-tuning* pada dua arsitektur Transformer, yaitu IndoBERT dan IndoRoBERTa. Evaluasi selama pelatihan dilakukan dengan memantau nilai *training loss* dan *validation loss* pada setiap epoch untuk mengidentifikasi performa pembelajaran dan kemampuan generalisasi model.

a. Hasil Pelatihan IndoBERT

Berdasarkan hasil pelatihan, nilai *training loss* pada model IndoBERT menunjukkan tren penurunan yang konsisten dari 0,357952 pada epoch pertama menjadi 0,159274 pada epoch ketiga. Hal ini mengindikasikan bahwa model semakin mampu menyesuaikan diri dengan data latih.

Namun demikian, nilai *validation loss* mengalami peningkatan dari 0,428260 pada epoch pertama menjadi 0,589752 pada epoch ketiga. Kenaikan ini menunjukkan bahwa performa model terhadap data validasi cenderung menurun seiring bertambahnya epoch.

Fenomena tersebut mengindikasikan terjadinya overfitting, di mana model mulai menghafal pola pada data latih dan kehilangan kemampuan generalisasi. Dengan demikian, epoch pertama dapat dianggap sebagai titik optimal karena memiliki nilai *validation loss* terendah. Hasil Pelatihan Model IndoBERT disajikan pada Tabel 3.

Tabel 3. Hasil Pelatihan Model IndoBERT

Epoch	Training Loss	Validation Loss
1	0.357952	0.428260
2	0.226831	0.437860
3	0.159274	0.589752

b. Hasil Pelatihan IndoRoBERTa

Pada model IndoRoBERTa, nilai *training loss* juga menunjukkan penurunan dari 0,306601 pada epoch pertama menjadi 0,164882 pada epoch ketiga. Hal ini menandakan bahwa model berhasil mempelajari pola dari data latih secara efektif.

Sementara itu, nilai *validation loss* mengalami peningkatan dari 0,335549 pada epoch pertama menjadi 0,471252 pada epoch ketiga. Pola ini serupa dengan yang terjadi pada IndoBERT, di mana performa pada data validasi menurun setelah epoch awal.

Kondisi ini juga menunjukkan indikasi overfitting, meskipun nilai *validation loss* pada IndoRoBERTa relatif lebih rendah dibandingkan IndoBERT pada setiap epoch. Hasil Pelatihan Model IndoRoBERTa disajikan pada Tabel 4.

Tabel 4. Hasil Pelatihan Model IndoRoBERTa

Epoch	Training Loss	Validation Loss
1	0.306601	0.335549
2	0.179965	0.355909
3	0.164882	0.471252

c. Analisis Perbandingan Pelatihan

Secara umum, kedua model menunjukkan pola pembelajaran yang serupa, yaitu penurunan *training loss* yang diikuti oleh peningkatan *validation loss*. Hal ini mengindikasikan bahwa kedua model cenderung mengalami overfitting setelah epoch pertama.

Namun demikian, IndoRoBERTa menunjukkan performa yang lebih stabil dengan nilai *validation loss* yang lebih rendah pada seluruh epoch dibandingkan IndoBERT. Hal ini mengindikasikan bahwa IndoRoBERTa memiliki kemampuan generalisasi yang lebih baik dalam menangani data validasi.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa epoch pertama merupakan titik optimal untuk kedua model, karena memberikan keseimbangan terbaik antara kemampuan pembelajaran dan generalisasi.

Evaluasi Performa Model

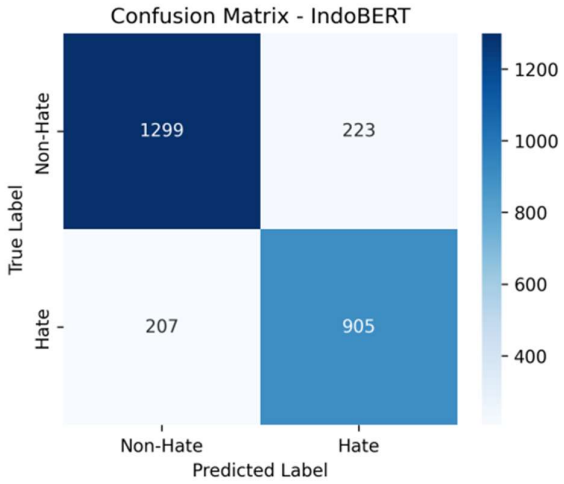
Evaluasi performa model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score, serta analisis *confusion matrix* untuk memberikan gambaran distribusi hasil prediksi terhadap label aktual. Evaluasi ini bertujuan untuk mengukur kemampuan model dalam mengklasifikasikan data ujaran kebencian secara akurat dan seimbang pada setiap kelas.

a. Hasil Evaluasi IndoBERT

Berdasarkan hasil pengujian, model IndoBERT memperoleh nilai akurasi sebesar **0,8368**. Pada kelas 0 (non-ujaran kebencian), model mencapai nilai presisi sebesar 0,8625, recall 0,8535, dan F1-score

0,8580. Sementara itu, pada kelas 1 (ujaran kebencian), diperoleh presisi sebesar 0,8023, recall 0,8138, dan F1-score sebesar 0,8080.

Hasil ini menunjukkan bahwa model IndoBERT memiliki performa yang cukup baik dalam mengklasifikasikan kedua kelas, meskipun terdapat sedikit penurunan performa pada kelas ujaran kebencian dibandingkan kelas non-ujaran kebencian. Distribusi hasil klasifikasi model IndoBERT disajikan pada Gambar 1 dan Tabel 5.



Gambar 2. Confusion Matrix IndoBERT

Tabel 5. Confusion Matrix IndoBERT

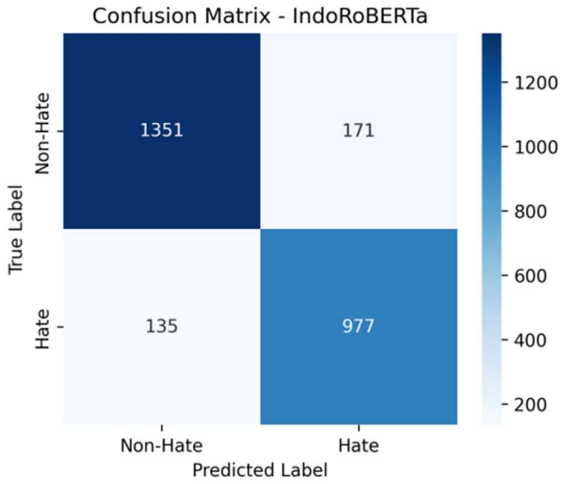
	Prediksi 0	**Prediksi 1**
Aktual 0	1299	223
Aktual 1	207	905

Berdasarkan *confusion matrix*, terlihat bahwa masih terdapat kesalahan klasifikasi, yaitu sebanyak 223 data kelas 0 yang diprediksi sebagai kelas 1, serta 207 data kelas 1 yang diprediksi sebagai kelas 0. Hal ini mengindikasikan bahwa model masih mengalami kesulitan dalam membedakan beberapa data dengan karakteristik yang ambigu.

b. Hasil Evaluasi IndoRoBERTa

Model IndoRoBERTa menunjukkan performa yang lebih baik dibandingkan IndoBERT dengan nilai akurasi sebesar **0,8838**. Pada kelas 0, model mencapai presisi sebesar 0,9092, recall 0,8876, dan F1-score sebesar 0,8983. Sementara itu, pada kelas 1 diperoleh presisi sebesar 0,8510, recall 0,8786, dan F1-score sebesar 0,8646.

Hasil ini menunjukkan bahwa IndoRoBERTa memiliki kemampuan yang lebih baik dalam mengklasifikasikan kedua kelas secara seimbang, terutama dalam meningkatkan kemampuan deteksi pada kelas ujaran kebencian. Distribusi hasil klasifikasi model IndoRoBERTa ditampilkan pada gambar 2 Tabel 6. Analisis *confusion matrix* menunjukkan penurunan jumlah kesalahan klasifikasi dibandingkan IndoBERT, dengan 171 data kelas 0 yang salah diprediksi sebagai kelas 1 dan 135 data kelas 1 yang salah diprediksi sebagai kelas 0. Hal ini menunjukkan peningkatan kemampuan model dalam memahami konteks data.



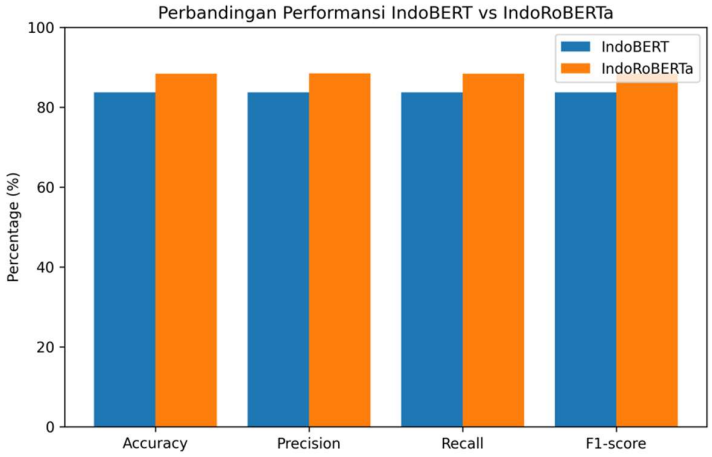
Gambar 2. Confusion Matrix IndoRoBERTa

Tabel 6. *Confusion Matrix* IndoRoBERTa

	Prediksi 0	Prediksi 1
Aktual 0	1351	171
Aktual 1	135	977

c. Perbandingan Performa Model

Perbandingan hasil evaluasi antara kedua model disajikan pada gambar 3 dan Tabel 7.



Gambar 3. Perbandingan Performa IndoBERT dan IndoRoBERTa

Tabel 7. Perbandingan Performa IndoBERT dan IndoRoBERTa

Model	Accuracy	Precision (avg)	Recall (avg)	F1-score (avg)
IndoBERT	0.8368	0.8324	0.8337	0.8330
IndoRoBERTa	0.8838	0.8801	0.8831	0.8814

Berdasarkan Gambar 3 dan Tabel 6, dapat dilihat bahwa model IndoRoBERTa secara konsisten menunjukkan performa yang lebih unggul dibandingkan IndoBERT pada seluruh metrik evaluasi. Selisih akurasi sebesar sekitar 4,7% menunjukkan peningkatan yang signifikan dalam kemampuan klasifikasi.

Peningkatan performa ini juga terlihat pada nilai F1-score, yang menunjukkan bahwa IndoRoBERTa memiliki keseimbangan yang lebih baik antara presisi dan recall. Dengan demikian,

model ini lebih efektif dalam mengurangi kesalahan klasifikasi, baik pada kelas mayoritas maupun minoritas.

Perbandingan IndoBERT dan IndoRoBERTa

Perbandingan kinerja antara model IndoBERT dan IndoRoBERTa dilakukan untuk mengidentifikasi model yang paling efektif dalam tugas klasifikasi ujaran kebencian berbahasa Indonesia. Analisis ini didasarkan pada metrik evaluasi yang telah diperoleh, yaitu akurasi, presisi, recall, dan F1-score, serta distribusi kesalahan yang terlihat pada *confusion matrix*.

a. Perbandingan Kinerja Umum

Berdasarkan hasil evaluasi, IndoRoBERTa menunjukkan performa yang lebih unggul dibandingkan IndoBERT pada seluruh metrik. Model IndoRoBERTa memperoleh akurasi sebesar 0,8838, lebih tinggi dibandingkan IndoBERT yang mencapai 0,8368. Selisih ini menunjukkan peningkatan performa yang signifikan dalam kemampuan klasifikasi.

Selain itu, nilai F1-score pada IndoRoBERTa (0,8814) juga lebih tinggi dibandingkan IndoBERT (0,8330), yang menunjukkan bahwa IndoRoBERTa memiliki keseimbangan yang lebih baik antara presisi dan recall dalam mengklasifikasikan kedua kelas.

b. Analisis Berdasarkan Kelas

Jika ditinjau berdasarkan masing-masing kelas, IndoRoBERTa menunjukkan peningkatan performa yang konsisten. Pada kelas ujaran kebencian (kelas 1), IndoRoBERTa memiliki nilai recall sebesar 0,8786, lebih tinggi dibandingkan IndoBERT yang hanya mencapai 0,8138. Hal ini menunjukkan bahwa IndoRoBERTa lebih efektif dalam mendeteksi data ujaran kebencian, sehingga mampu mengurangi jumlah *false negative*.

Selain itu, nilai presisi pada kelas 1 juga meningkat dari 0,8023 (IndoBERT) menjadi 0,8510 (IndoRoBERTa), yang mengindikasikan bahwa prediksi positif yang dihasilkan IndoRoBERTa lebih akurat. Peningkatan ini berkontribusi langsung terhadap nilai F1-score yang lebih tinggi.

c. Analisis Kesalahan Klasifikasi

Berdasarkan *confusion matrix*, IndoRoBERTa menghasilkan jumlah kesalahan klasifikasi yang lebih rendah dibandingkan IndoBERT. Pada IndoBERT, terdapat 223 kesalahan *false positive* dan 207 kesalahan *false negative*. Sementara itu, IndoRoBERTa menurunkan kesalahan tersebut menjadi 171 *false positive* dan 135 *false negative*.

Penurunan jumlah kesalahan ini menunjukkan bahwa IndoRoBERTa memiliki kemampuan yang lebih baik dalam memahami konteks teks, sehingga dapat mengurangi ambiguitas dalam proses klasifikasi.

d. Analisis Penyebab Perbedaan Performa

Keunggulan IndoRoBERTa dapat dijelaskan dari sisi arsitektur dan strategi pelatihan. IndoRoBERTa merupakan pengembangan dari model BERT dengan pendekatan pelatihan yang lebih optimal, termasuk penggunaan data yang lebih besar dan penghapusan komponen *next sentence prediction*. Hal ini memungkinkan model untuk menangkap representasi konteks bahasa secara lebih mendalam.

Selain itu, hasil pelatihan menunjukkan bahwa IndoRoBERTa memiliki nilai *validation loss* yang lebih rendah dibandingkan IndoBERT pada setiap epoch, yang mengindikasikan kemampuan generalisasi yang lebih baik serta ketahanan yang lebih tinggi terhadap overfitting.

e. Implikasi Hasil Penelitian

Hasil penelitian ini menunjukkan bahwa pemilihan arsitektur model memiliki pengaruh signifikan terhadap performa sistem klasifikasi teks. IndoRoBERTa terbukti lebih efektif dalam mendeteksi ujaran kebencian, terutama dalam meningkatkan kemampuan deteksi pada kelas minoritas.

Dengan demikian, IndoRoBERTa dapat direkomendasikan sebagai model yang lebih optimal untuk digunakan dalam pengembangan sistem deteksi ujaran kebencian berbahasa Indonesia.

Pembahasan

Hasil penelitian ini menunjukkan bahwa model berbasis Transformer, khususnya IndoBERT dan IndoRoBERTa, memiliki kemampuan yang baik dalam menangani tugas klasifikasi ujaran kebencian berbahasa Indonesia. Temuan ini sejalan dengan studi sebelumnya yang menyatakan bahwa arsitektur Transformer mampu menangkap konteks semantik secara bidirectional sehingga lebih unggul dibandingkan metode berbasis machine learning tradisional maupun deep learning konvensional seperti LSTM dan CNN (Ramos et al., 2024; Pamungkas & Purworini, 2025).

Perbedaan performa antara IndoBERT dan IndoRoBERTa mengindikasikan bahwa variasi arsitektur dan strategi pre-training memiliki pengaruh signifikan terhadap kemampuan generalisasi model. IndoRoBERTa yang tidak menggunakan mekanisme *next sentence prediction* serta dilatih dengan optimasi data yang lebih besar menunjukkan performa yang lebih stabil. Hal ini konsisten dengan temuan Liu et al. dalam pengembangan RoBERTa yang menyatakan bahwa penghilangan objective tertentu dalam pre-training dapat meningkatkan representasi bahasa secara lebih efektif.

Selain itu, hasil penelitian ini memperkuat temuan bahwa deteksi ujaran kebencian dalam bahasa Indonesia memiliki tantangan utama berupa variasi linguistik dan konteks implisit. Model Transformer terbukti mampu mengatasi kompleksitas tersebut melalui mekanisme attention yang memungkinkan model memahami hubungan antar kata dalam konteks kalimat secara lebih mendalam. Hal ini sejalan dengan penelitian Fetahi et al. (2025) yang menunjukkan bahwa Transformer lebih efektif dalam menangani low-resource language dibandingkan pendekatan sebelumnya.

Dari sisi ketidakseimbangan data, penerapan teknik SMOTE pada data latih memberikan kontribusi dalam meningkatkan kemampuan model pada kelas minoritas. Hal ini terlihat dari peningkatan nilai recall pada kelas ujaran kebencian, yang menunjukkan bahwa model lebih mampu mengidentifikasi kasus positif secara lebih sensitif. Temuan ini mendukung penelitian Zhang & Chen (2021) mengenai SMOTE yang menyatakan bahwa oversampling berbasis sintetik dapat membantu meningkatkan representasi kelas minoritas tanpa menghilangkan karakteristik distribusi data asli.

Secara keilmuan, penelitian ini memberikan kontribusi dalam pengembangan studi Natural Language Processing (NLP) berbahasa Indonesia, khususnya dalam konteks klasifikasi ujaran kebencian berbasis Transformer. Studi ini memperkuat bukti empiris bahwa model IndoRoBERTa memiliki potensi lebih baik dibandingkan IndoBERT dalam menangani teks media sosial yang bersifat tidak formal dan kontekstual.

Dari sisi praktis, hasil penelitian dapat digunakan sebagai dasar pengembangan sistem moderasi konten otomatis pada platform media sosial, khususnya untuk bahasa Indonesia. Penerapan model berbasis Transformer yang dikombinasikan dengan teknik penyeimbangan data seperti SMOTE dapat meningkatkan akurasi deteksi konten berbahaya sehingga membantu menciptakan ruang digital yang lebih aman dan sehat.

Implikasi hasil penelitian ini menunjukkan bahwa pemilihan arsitektur Transformer berpengaruh signifikan terhadap efektivitas sistem klasifikasi ujaran kebencian berbahasa Indonesia. IndoRoBERTa terbukti lebih adaptif dalam memahami konteks bahasa informal pada media sosial dan lebih efektif dalam mendeteksi kelas minoritas. Oleh karena itu, kombinasi IndoRoBERTa dan teknik data balancing seperti SMOTE dapat dipertimbangkan sebagai pendekatan yang potensial dalam pengembangan sistem moderasi konten otomatis yang lebih akurat, adaptif, dan kontekstual pada platform media sosial Indonesia.

Kesimpulan

Penelitian ini menunjukkan bahwa model berbasis Transformer efektif untuk klasifikasi ujaran kebencian berbahasa Indonesia. IndoRoBERTa secara konsisten memberikan performa yang lebih baik dibandingkan IndoBERT pada aspek akurasi, kemampuan generalisasi, keseimbangan antara precision dan recall, serta pengurangan kesalahan klasifikasi. Hasil ini mengindikasikan bahwa perbedaan

arsitektur dan strategi pre-training berpengaruh terhadap kualitas representasi konteks teks, terutama pada data media sosial yang bersifat informal dan kompleks. Selain itu, penerapan SMOTE terbukti membantu meningkatkan kemampuan model dalam mendeteksi kelas minoritas. Namun, penelitian ini masih memiliki keterbatasan berupa cakupan dataset yang terbatas dan indikasi overfitting selama pelatihan. Penelitian selanjutnya disarankan menggunakan dataset yang lebih beragam, menerapkan validasi yang lebih komprehensif, serta mengeksplorasi model dan strategi pelatihan yang lebih mutakhir untuk meningkatkan robustitas dan akurasi sistem.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan, baik berupa bantuan teknis, masukan ilmiah, maupun dukungan lainnya selama proses penelitian dan penyusunan artikel ini.

Pernyataan Konflik Kepentingan (*Declaration of Conflict of Interest*)

Para penulis menyatakan tidak ada potensi konflik kepentingan terkait dengan penelitian, penulisan, dan/atau publikasi dari artikel ini.

Daftar Pustaka

Alamsyah, A., & Sagama, Y. (2024). Empowering Indonesian internet users : An approach to counter online toxicity and enhance digital well-being. Intelligent Systems with Applications, 22(August 2023), 200394. https://doi.org/10.1016/j.iswa.2024.200394

Alkomah, F., & Ma, X. (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. Information, 1–22. https://doi.org/https://doi.org/10.3390/info13060273

Amalia, F. S., & Suyanto, Y. (2024). Offensive Language and Hate Speech Detection Using Bert Model. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 18(4). https://doi.org/https://doi.org/10.22146/ijccs.99841

Bao, R., & Gu, B. (2022). An Accelerated Doubly Stochastic Gradient Method with Faster Explicit Model Identification. In Proceedings of the 31st ACM International Conference on Information and Knowledge Management (Vol. 1, Nomor 1). Association for Computing Machinery. https://doi.org/10.1145/3511808.3557234

Fetahi, E., Susuri, A., Hamiti, M., Kastrati, Z., Canhasi, E., & Misini, A. (2025). Enhancing social media hate speech detection in low - resource languages using transformers and explainable AI. Social Network Analysis and Mining, 15(1), 1–30. https://doi.org/10.1007/s13278-025-01497-w

Ghosh, K. (2025). Hate speech detection in low-resourced Indian languages : An analysis of transformer-based monolingual and multilingual models with cross-lingual experiments. Natural Language Processing, 393–414. https://doi.org/10.1017/nlp.2024.28

Hakimi, M., Kohistani, A. J., Azimy, A. S., & Ardi, I. M. (2025). The Influence of Emerging Technologies on Communication Practices in the Digital Age. Jurnal Ilmiah Dinamika Sosial, 9(1), 136–153. https://doi.org/https://doi.org/10.38043/jids.v9i1.6500

Idris, U., Salihu, S., Abdulalim, N., Ali, S., Shawulu, J. C., & Adam, A. (2026). Machine Learning for Hate Text Speech Detection : A Comprehensive Review of Techniques , Dataset and Challenges. Asian Journal of Research in Computer Science Volume, 19(2), 204–218. https://doi.org/10.9734/ajrcos/2026/v19i2832

Imaduddin, H., Kusumaningtias, L. A., & A, F. Y. (2023). Application of LSTM and GloVe Word Embedding for Hate Speech Detection in Indonesian Twitter Data. Ingénierie des Systèmes d ' Information, 28(4), 1107–1112. https://doi.org/https://doi.org/10.18280/isi.280430

Koto, F., & Baldwin, T. (2020). IndoLEM and IndoBERT : A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. Proceedings ofthe 28th International Conference on Computational Linguistics, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of Hate Speech Detection in Social Media. SN

Computer Science, 2(2), 1–15. https://doi.org/10.1007/s42979-021-00457-3
Nurindah, A. A., Hasanati, N., & Aini, Q. (2025). Bibliometrik Hate Speech : Tren Metode Penelitian dan Domain Implementasi. JUSIFOR: Jurnal Sistem Informasi dan Informatika, 4(2), 270–278. https://doi.org/https://doi.org/10.70609/jusifor.v4i2.8652
Okky, M., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media : Progress and challenges. Heliyon, 9(8), e18647. https://doi.org/10.1016/j.heliyon.2023.e18647
Pamungkas, E. W., & Purworini, D. (2025). Enhancing Hate Speech Detection in Low- Resource Code-Mixed Indonesian Tweets via GPT-Based Data Augmentation. Engineering, Technology & Applied Science Research, 15(6), 30649–30656. https://doi.org/https://etasr.com/index.php/ETASR/article/view/14342/6045
Pananookooln, C., Akaranee, J., & Silpasuwancha, C. (2023). Comparing Selective Masking Methods for Depression Detection in Social Media. Computational Linguistics, February. https://doi.org/10.1162/coli a 00479
Przybyła, P., & Soto, A. J. (2021). When classification accuracy is not enough : Explaining news credibility assessment. Information Processing and Management, 58(5), 102653. https://doi.org/10.1016/j.ipm.2021.102653
Purnomo, T. D., & Sutopo, J. (2024). Comparison of Pre-Trained BERT-based Transformer Models fo Regional. Internasional Journal Science and Technology, 3(3), 11–21. https://doi.org/https://doi.org/10.56127/ijst.v3i3.1739
Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. Social Network Analysis and Mining, 14(1), 1–25. https://doi.org/10.1007/s13278-024-01361-3
Rivadeneira, R. (2025). applied sciences Emotional Tone Detection in Hate Speech Using Machine Learning and NLP : Methods , Challenges , and Future Directions — A Systematic Review. Applied Sciences. https://doi.org/https://doi.org/10.3390/app152312686
Sarkar, D., Zampieri, M., Ranasinghe, T., & Ororbia, A. (2021). fBERT : A Neural Transformer for Identifying Offensive Content. Antologi ACL, 1792–1798. https://doi.org/10.18653/v1/2021.findings-emnlp.154
Selvaraj, P., Nc, G., Kumar, P., & Khapra, M. (2022). OpenHands : Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages. Antologi ACL, 1, 2114–2133. https://doi.org/0.18653/v1/2022.acl-long.150
Shoeb, A. A., & Melo, G. De. (2021). Assessing Emoji Use in Modern Text Processing Tools. ACL Antology. https://doi.org/10.18653/v1/2021.acl-long.110
Suciati. (2024). A bibliometrics analysis of interpersonal communication in social media. Cogent Social Sciences, 1886. https://doi.org/10.1080/23311886.2024.2424472
Tita, T. (2021). Cross-lingual Hate Speech Detection using Transformer Models. arXiv. https://doi.org/https://doi.org/10.48550/arXiv.2111.00981
Tsugawa, S., & Watabe, K. (2023). Identifying Influential Brokers on Social Media from Social Network Structure. Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023), Icwsm. https://doi.org/https://doi.org/10.1609/icwsm.v17i1.22193
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Kaiser, Ł. (2023). Attention Is All You Need. arxiv, Nips. https://doi.org/https://doi.org/10.48550/arXiv.1706.03762
Yoon, M., Gervet, T., Shi, B., Niu, S., He, Q., & Yang, J. (2021). Performance-Adaptive Sampling Strategy Towards Fast and Accurate Graph Neural Networks. Research Track Paper, 2046–2056. https://doi.org/https://doi.org/10.1145/3447548.34672
Zhang, Y., & Chen, L. (2021). A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method. Theoretical Economics Letters, 258–267. https://doi.org/10.4236/tel.2021.112019