DATAMATIKA

Tersedia secara daring di:
https://conference.ipb.ac.id/datamatika

PENGARUH TAHAPAN PREPROCESSING TERHADAP PERFORMA MODEL INDOBERT DAN INDOBERTWEET PADA SENTIMEN BANJIR ACEH 2025

Naiya Dzil Izzati[1‡], Farit Mochamad Afendi[2], and Mohammad Masjkur[3]

[1]Statistics and Data Science Study Program, IPB University, Indonesia
[‡]corresponding author: naiyadzilizzati@apps.ipb.ac.id

Abstract

Indonesia is highly vulnerable to hydrometeorological disasters such as floods, which often generate public responses on social media. These responses provide valuable insights for sentiment analysis but are characterized by unstructured and noisy text. This study aims to analyze the effect of preprocessing stages on the performance of IndoBERT and IndoBERTweet models for sentiment classification of tweets related to the Aceh flood in 2025. The dataset was collected from platform X and manually labeled into three sentiment classes: positive, negative, and neutral. Three preprocessing scenarios were applied, starting from case folding as minimal preprocessing and extending to more complex stages such as cleaning, normalization, stopword removal, and stemming. The models were evaluated using accuracy, precision, recall, and F1-score with a macro approach. The results show that preprocessing affects model performance, although the differences are relatively small. IndoBERT achieved its best performance under minimal preprocessing, while IndoBERTweet showed improved performance with more complex preprocessing and achieved the highest F1-score of 0.8435. Overall, IndoBERTweet demonstrated slightly better and more stable performance compared to IndoBERT. These findings indicate that the effectiveness of preprocessing depends on the compatibility between the model and data characteristics, particularly for social media text.

Keywords: IndoBERT, IndoBERTweet, Preprocessing, Sentiment Analysis, Social Media.

1. Pendahuluan

Indonesia merupakan negara dengan tingkat kerawanan bencana alam yang tinggi, khususnya bencana hidrometeorologi seperti banjir. Data Badan Nasional Penanggulangan Bencana (2024) menunjukkan bahwa dari 3.472 kejadian bencana di Indonesia, sebanyak 1.420 di antaranya merupakan banjir. Tingginya frekuensi tersebut menjadikan banjir sebagai isu publik yang berdampak luas terhadap kondisi sosial dan ekonomi masyarakat. Pada akhir tahun 2025, beberapa wilayah di Aceh mengalami

curah hujan ekstrem yang menyebabkan banjir serta memunculkan berbagai respons masyarakat. Perkembangan teknologi informasi mendorong masyarakat menyampaikan opini secara cepat melalui media sosial. *Platform* X menjadi salah satu ruang publik digital yang aktif digunakan untuk membahas peristiwa kebencanaan. Data yang dihasilkan bersifat real-time dan mencerminkan persepsi masyarakat secara langsung. Karakteristik teks media sosial yang tidak terstruktur, mengandung bahasa informal, singkatan, dan variasi ejaan menimbulkan tantangan dalam proses analisis, sehingga diperlukan pendekatan komputasional yang mampu menangani kompleksitas tersebut.

Analisis sentimen sebagai bagian dari *Natural Language Processing* (NLP) digunakan untuk mengklasifikasikan opini ke dalam kategori positif, negatif, atau netral (Liu 2012). Salah satu model bahasa yang berpengaruh dalam perkembangan NLP adalah *Bidirectional Encoder Representations from Transformers* (BERT). BERT menggunakan arsitektur Transformer dengan mekanisme *self-attention* yang memungkinkan model memahami hubungan antar kata dalam dua arah secara simultan (Devlin *et al.* 2019). Berbeda dengan model sekuensial tradisional yang membaca teks secara satu arah, pendekatan dua arah pada BERT memungkinkan representasi konteks yang lebih kaya sehingga meningkatkan performa pada berbagai tugas klasifikasi teks, termasuk analisis sentimen. Pengembangan BERT untuk bahasa Indonesia menghasilkan IndoBERT yang dilatih menggunakan korpus besar berbahasa Indonesia dengan cakupan topik umum, termasuk teks formal dan semi formal (Wilie *et al.* 2020). IndoBERTweet dikembangkan dengan korpus khusus yang berasal dari data Twitter berbahasa Indonesia (Koto *et al*. 2021). Perbedaan sumber korpus pelatihan tersebut menyebabkan IndoBERT dan IndoBERTweet memiliki karakteristik representasi bahasa yang berbeda.

Penelitian yang dilakukan oleh Qolbu *et al.* (2025) membandingkan IndoBERT dengan metode konvensional seperti *Support Vector Machine (*SVM) dan Naïve Bayes pada data berbahasa Indonesia, dan memperoleh hasil bahwa IndoBERT mencapai performa terbaik dengan akurasi 0,96 dan F1-*score makro* 0,95, IndoBERT juga relatif stabil bahkan tanpa penanganan ketidakseimbamgan data. Dewi (2025) membandingkan IndoBERT dan IndoBERTweet pada data yang berasal dari platform X, dan menunjukkan bahwa IndoBERTweet memiliki performa lebih tinggi dengan akurasi 82,40% dan F1-*score* 82,38%. Hasil tersebut menunjukkan bahwa kesesuaian antara domain data dan korpus pelatihan berpengaruh terhadap kinerja model. Selain faktor korpus pelatihan, tahapan *preprocessing* juga berperan dalam membentuk representasi teks sebelum diproses oleh model. Khairani *et al.* (2024) membuktikan bahwa penghapusan *stopwords* dan penerapan *stemming* tidak selalu meningkatkan akurasi pada IndoBERT dan IndoBERTweet.

Penelitian terdahulu umumnya berfokus pada perbandingan performa model atau evaluasi satu konfigurasi *preprocessing* tertentu. Analisis yang secara sistematis mengkaji pengaruh beberapa skenario *preprocessing* terhadap IndoBERT dan IndoBERTweet dalam konteks kebencanaan masih terbatas. Karakteristik linguistik pada data banjir Aceh 2025

memungkinkan munculnya pola yang berbeda dari studi sebelumnya. Kesenjangan tersebut menunjukkan perlunya penelitian yang mengevaluasi interaksi antara model dan strategi *preprocessing* secara lebih komprehensif. Penelitian ini bertujuan untuk membandingkan performa IndoBERT dan IndoBERTweet serta menganalisis pengaruh tiga skenario *preprocessing* terhadap klasifikasi sentimen pada data platform X terkait banjir Aceh menggunakan metrik *accuracy, precision, recall,* dan F1-*score*. Hasil penelitian diharapkan memberikan kontribusi teoretis dalam memahami sensitivitas model *Transformer* terhadap variasi *preprocessing* serta kontribusi praktis dalam pengembangan sistem pemantauan opini publik berbasis media sosial pada situasi kebencanaan.

2. Metodologi

2.1 Data

Data yang digunakan adalah data komentar dari platform X yang mengandung kata “Banjir Aceh” pada tanggal 29 November 2025 hingga 6 Desember 2025. Metode pengumpulan data dilakukan dengan mengekstraksi data teks secara otomatis menggunakan metode *scraping* pada Google Colab dengan *Python* dan memanfaatkan library “Tweet Harvest”. Contoh data dapat dilihat pada Tabel 1.

2.2 Class Weight

Ketidakseimbangan data merupakan kondisi ketika jumlah data pada setiap kelas dalam dataset tidak seimbang. Kondisi tersebut dapat menyebabkan model klasifikasi cenderung memprediksi kelas mayoritas dan kurang mampu mengenali kelas minoritas (He dan Garcia 2009). Salah satu pendekatan yang umum digunakan untuk mengatasi masalah tersebut adalah penerapan *class weight.* Metode ini memberikan bobot yang lebih besar pada kelas minoritas dalam fungsi *loss* selama proses pelatihan, sehingga kesalahan klasifikasi pada kelas tersebut akan menghasilkan penalti yang lebih besar, sehingga model terdorong untuk mempelajari pola pada setiap kelas secara lebih seimbang (King dan Zeng 2001). Perhitungan bobot kelas dapat dilihat pada Persamaan (1)

$$w_i = \frac{N}{n_i \times C} \tag{1}$$

dengan w_i adalah bobot kelas ke-i, N adalah total jumlah data latih, n_i adalah jumlah data pada kelas ke-i, dan C adalah total jumlah kelas.

Tabel 1: Contoh data hasil *crawling*

Conversation_ID	*Date*	*Tweet*	*Link Post*
1,99497E+18	30/11/25	@ranggavega Berbagai satwa kehilangan habitat dan memasuki ladang dan perkampungan. Kini habitat mereka semakin ciut untuk berbagai kepentingan. Mereka menjadi korban banjir bandang. Tak mungkin hanya gajah satwa² kecil lain tentu menjadi korban bencana karena ulah manusia (dok FFI Aceh) https://t.co/HCNhPfBRSG	https://x.com/undefined/status/1995281122511687749
1,99659E+18	04/12/25	sederhana tapi pemerintah aceh sumut sumbar gabisa. yap! PERAHU KARET. dari kemaren kemaren gw baca bantuan di lempar lempar padahal banjir di area yang mereka lempari bantuan itu udah ga deras arusnya cuman ya emang masi sepinggang	https://x.com/undefined/status/1996593542152741159

2.3 IndoBERT

IndoBERT merupakan model bahasa pra-latih berbasis arsitektur BERT yang dikembangkan untuk bahasa Indonesia. Arsitektur BERT diperkenalkan oleh Devlin *et al.* (2019) sebagai model representasi bahasa dua arah (*bidirectional*) berbasis *Transformer Encoder* yang memahami setiap token berdasarkan konteks sebelum dan sesudahnya dalam satu urutan teks. Proses pra-pelatihan dilakukan menggunakan *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP) untuk menghasilkan *contextual embedding*, yaitu representasi kata yang bergantung pada konteks penggunaannya (Devlin *et al.* 2019). Representasi kontekstual ini memungkinkan model menangkap makna kalimat secara lebih komprehensif pada berbagai tugas NLP, termasuk klasifikasi teks. IndoBERT menggunakan tokenisasi berbasis *WordPiece* yang memecah kata menjadi unit *subword*, sehingga mampu menangani

variasi morfologi bahasa Indonesia tanpa memerlukan proses *stemming* agresif. Pada analisis sentimen, model diadaptasi melalui proses *fine-tuning* dengan memanfaatkan representasi keseluruhan teks untuk memprediksi kelas sentiment (Devlin *et al.* 2019).

2.4 IndoBERTweet

IndoBERTweet dikembangkan untuk memodelkan bahasa Indonesia pada domain media sosial, khususnya Twitter. Model ini diperkenalkan Koto *et al.* (2021) melalui pelatihan menggunakan korpus besar yang berasal dari cuitan berbahasa Indonesia. Pengembangan tersebut dilatarbelakangi oleh perbedaan karakteristik antara bahasa formal dan bahasa yang digunakan dalam media sosial. Teks Twitter umumnya bersifat informal dan mengandung berbagai bentuk non-standar, seperti slang, singkatan, pengulangan huruf, emotikon, serta variasi ejaan. Kondisi ini menyebabkan distribusi kosakata dan pola bahasa pada media sosial berbeda dari teks berita atau dokumen formal. Koto *et al.* (2021) menekankan pentingnya inisialisasi kosakata berbasis domain agar kata-kata khas Twitter dapat direpresentasikan dengan lebih efektif. IndoBERTweet mengacu pada BERT dengan skema pra-pelatihan berbasis *masked language modeling*. Model kemudian diadaptasi melalui proses *fine-tuning* untuk tugas klasifikasi, termasuk analisis sentimen. Kesesuaian antara korpus pelatihan dan karakteristik data media sosial menjadikan IndoBERTweet secara teoritis lebih adaptif terhadap teks yang mengandung *noise*.

2.5 Prosedur Analisis Data

Prosedur analisis data pada penelitian ini adalah sebagai berikut:

1. *Scraping* data dilakukan dengan *scraping* komentar dari aplikasi X menggunakan syntax Tweet Harvest pada Google Colab.
2. Pelabelan data dilakukan secara manual oleh penulis. label yang digunakan adalah positif, negatif, dan netral, kriteria yang digunakan dapat dilihat pada Tabel 2.

Tabel 2: Kriteria pelabelan data.

Label	Kriteria
Positif	Teks yang mengandung opini, dukungan, apresiasi, doa, atau harapan terhadap pemerintah, relawan, atau pihak terkait
Negatif	Teks yang mengandung kritik, kemarahan, kekecewaan, kesedihan, atau opini buruk terhadap peristiwa banjir, dampaknya, maupun penanganannya
Netral	Teks yang bersifat informatif, berupa laporan kejadian, pertanyaan, atau pernyataan faktual tanpa menunjukkan opini atau emosi yang jelas

3. *Preprocessing* data bertujuan untuk mengubah data teks mentah menjadi format yang lebih terstruktur sehingga dapat dipahami dan diproses lebih lanjut sesuai kebutuhan. *Preprocessing* data mencakup beberapa tahapan, yaitu:
   a. *Case folding*, proses menyeragamkan jenis huruf, untuk mencegah perbedaan dalam pengenalan huruf besar dan huruf kecil dalam data.
   b. *Cleaning*, menghilangkan unsur yang tidak terlalu penting seperti URL, *mention*, emoji, tanda baca, angka, hashtag, dan pengulangan karakter yang sama.
   c. *Normalization*, mengganti kata tidak formal atau slang menjadi kata yang lebih umum, bertujuan untuk memastikan konsistensi dalam gaya penulisan.
   d. *Stopword removal,* menghilangkan kata umum yang tidak memberikan banyak informasi penting dalam analisis teks. Proses ini menggunakan *library Sastrawi* (Khairani *et al.* 2024).
   e. *Stemming*, proses menyederhanakan kata menjadi bentuk dasarnya, Pada penelitian ini, proses stemming menggunakan *library Sastrawi* (Khairani *et al.* 2024).

   Pada penelitian ini terdapat tiga skenario *preprocessing*, yaitu skenario 1 hanya melalui *case folding*, skenario 2 melalui *case folding, cleaning,* dan *normalization*, sedangkan skenario 3, melalui seluruh tahapan *preprocessing.*
4. Melakukan eksplorasi data menggunakan *word cloud* untuk melihat kata yang muncul di setiap sentimen dan seberapa sering kata tersebut muncul.
5. Pemodelan IndoBERT dan IndoBERTweet
   a. Data dibagi menjadi data latih, data validasi, dan data uji dengan proporsi 80%, 10%, dan 10%.
   b. Memeriksa ketidakseimbangan distribusi kelas pada data latih, jika ketidakseimbangan terdeteksi, dilakukan penanganan dengan *class weight.*
   c. Melakukan pelatihan model IndoBERT dan IndoBERTweet menggunakan data latih dan data validasi. *Hyperparameter* yang digunakan pada pemodelan adalah *epoch* 4, *learning rate* 2e-5 dan 3e-5, *batch size* 16 dan 32, *optimizer* adamW, *dropout rate* 0.1 dan *max length* 128.
   d. Melakukan *tuning hyperparameter* dengan *grid search* serta menerapkan *early stopping* untuk mencegah *overfitting* (Devlin *et al.* 2019).
6. Melakukan evaluasi pemodelan untuk melihat efektivitas model pada setiap klasifikasi. Evaluasi dilakukan dengan menggunakan *confusion matrix* untuk melihat perbandingan antara prediksi dengan kejadian sebenarnya.

Tabel 3: *Confusion Matrix*

Actual Class	*Predicted Class*		
	Positive	*Neutral*	*Negative*
Positive	*True Positive* (TP)	*False Neutral* (FNt)	*False Negative* (FN)
Neutral	*False Positive* (FP)	*True Neutral* (TNt)	*False Negative* (FN)
Negative	*False Positive* (FP)	*False Neutral* (FNt)	*True Negative* (TN)

Selain itu dilakukan evaluasi juga menggunakan *Balanced Accuracy* untuk melihat ukuran kebaikan model secara keseluruhan dan F1-*score* yang menggabungkan *precision* dan *recall* dalam satu metrik untuk melihat keseimbangan performa model. Persamaan *Balanced Accuracy*, F1-*score* dan F1-*score macro* dapat dilihat pada persamaan (2) dan (3). Hasil evaluasi dibandingkan antar skenario *preprocessing* dan antar model untuk menganalisis pengaruh tahapan *preprocessing* terhadap performa klasifikasi.

$$Balanced\ Accuracy = \frac{\sum_{i=1}^{C} Recall_i}{C} \quad (2)$$

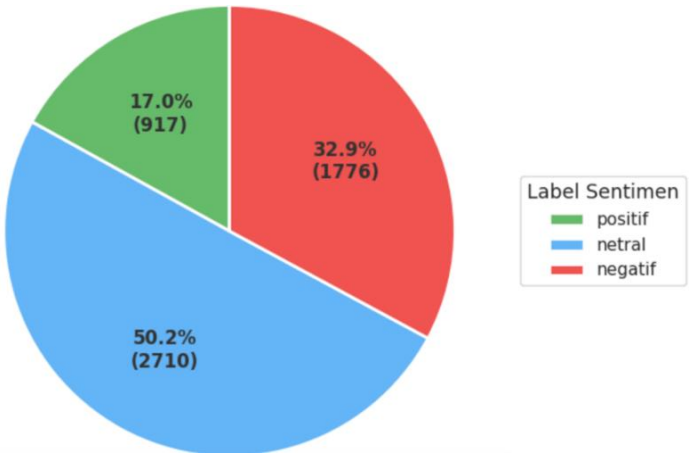
$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

$$Macro\ F1 - score = \frac{\sum_{i=1}^{C} F1_i}{C} \quad (4)$$

3. Hasil dan Pembahasan

3.1 Pelabelan Data

Pelabelan sentimen pada data dibagi dalam tiga kategori, yaitu positif, negatif, dan netral. Proses pelabelan dilakukan secara manual oleh penulis dengan total data 5403.



Gambar 1: *Pie chart* distribusi data tiap kelas

Distribusi data setiap kelas dapat dilihat pada Gambar 1, grafik *pie chart* menunjukkan adanya ketidakseimbangan data. Kelas netral mendominasi dengan persentase sebesar 50,2 %, diikuti kelas negatif dengan persentase sebesar 32,9%, dan kelas positif memiliki persentase terkecil sebesar 17%. Ketidakseimbangan data dapat membuat model cenderung memprediksi kelas mayoritas secara tidak proporsional dan menyulitkan model dalam mengenali kelas minoritas. Kondisi tersebut perlu ditangani sebelum proses pemodelan dilakukan agar model mampu mempelajari karakteristik dari setiap kelas dengan seimbang, sehingga hasil akurasi klasifikasi lebih adil dan optimal.

3.2 Preprocessing Output

Tahapan *preprocessing* dilakukan untuk meningkatkan kualitas data teks sebelum digunakan dalam proses klasifikasi sentimen. Tabel 4 menunjukkan perubahan teks setelah melalui setiap tahapan *preprocessing*. Pada data awal, teks masih mengandung huruf kapital, singkatan, *hashtag*, dan kata tidak baku yang umum ditemukan pada media sosial. Tahapan *case folding* menyeragamkan seluruh huruf menjadi huruf kecil sehingga penulisan kata menjadi lebih konsisten*, data cleaning* menghilangkan simbol dan karakter yang tidak diperlukan, *normalization* mengubah kata tidak baku seperti “blm” menjadi “belum” dan “ga” menjadi “tidak”, *stopword removal* menghapus kata yang kurang informatif sehingga teks menjadi lebih ringkas, dan *stemming* mengubah kata menjadi bentuk dasar.

Tabel 4: Contoh data hasil *preprocessing*.

Preprocessing	Hasil
Isi Tweet	Katanya banjir didaerah Aceh blm BENCANA NASIONAL coba rasain dan liat dulu parahnya gimana bantuan ga ada orang orang pada kelaparan harga harga yg ga masuk logika BBM yg sulit miris bgt liat PEMERINTAH. #PrayForAceh
Case folding	katanya banjir didaerah aceh blm bencana nasional coba rasain dan liat dulu parahnya gimana bantuan ga ada orang orang pada kelaparan harga harga yg ga masuk logika bbm yg sulit miris bgt liat pemerintah. #prayforaceh
Data cleaning	katanya banjir didaerah aceh blm bencana nasional coba rasain dan liat dulu parahnya gimana bantuan ga ada orang orang pada kelaparan harga harga yg ga masuk logika bbm yg sulit miris bgt liat pemerintah
Normalization	katanya banjir didaerah aceh belum bencana nasional coba rasain dan lihat dulu parahnya gimana bantuan tidak ada orang orang pada kelaparan harga harga yang tidak masuk logika bbm yang sulit miris banget lihat pemerintah

Preprocessing	Hasil
Stopword removal	katanya banjir didaerah aceh bencana nasional coba rasain lihat dulu parahnya gimana bantuan ada orang orang kelaparan harga harga tidak masuk logika bbm sulit miris banget lihat pemerintah
Stemming	kata banjir daerah aceh bencana nasional coba rasain lihat dulu parah gimana bantu ada orang orang lapar harga harga tidak masuk logika bbm sulit miris banget lihat perintah

3.3 Visualisasi Data

Visualisasi data yang digunakan adalah *word cloud. Word cloud* adalah teknik analisis teks yang menampilkan representasi grafis dari frekuensi kata dalam teks asli. Semakin sering kata muncul dalam dokumen, semakin besar ukuran kata dalam visualisasi.

(a) (b) (c)

Gambar 2: Visualisasi *Word Cloud* (a) skenario 1, (b) skenario 2, (c) skenario 3

Pada Gambar 2 (a) ditampilkan hasil *word cloud* pada data yang hanya melalui tahapan *case folding*. Terlihat bahwa kata-kata yang dominan tidak hanya berasal dari konteks utama penelitian, tetapi juga mencakup kata-kata yang kurang informatif seperti “di”, “dan”, “yang”, serta elemen non-teks seperti “https”. Dominasi elemen tersebut menunjukkan bahwa data masih mengandung *noise* yang cukup tinggi, terutama dalam bentuk *stopwords* dan URL yang belum dihilangkan. Kondisi ini mengindikasikan bahwa penerapan *case folding* saja belum cukup untuk menghasilkan representasi teks yang bersih dan informatif. Pada Gambar 2 (b), yang merupakan hasil dari penerapan *case folding*, *cleaning*, dan *normalization*, terlihat adanya penurunan elemen *noise* seperti URL yang sebelumnya dominan pada skenario pertama, meskipun beberapa kata umum seperti “di” dan “dan” masih muncul dalam visualisasi, Di sisi lain, kata-kata yang relevan dengan konteks kebencanaan seperti “banjir”, “aceh”, “dampak”, dan “korban” mulai lebih menonjol. Hal ini menunjukkan bahwa tahapan *cleaning* dan *normalization* mampu meningkatkan kualitas data, meskipun belum sepenuhnya menghilangkan kata-kata yang kurang informatif. Pada Gambar 2 (c) yang merepresentasikan skenario dengan seluruh tahapan *preprocessing*, Visualisasi menjadi lebih bersih dan terarah pada kata-kata kunci seperti “bantu”, “longsor”, “dampak”, dan “korban”. Meskipun

demikian, penerapan *preprocessing* yang lebih intensif tetap berpotensi mengurangi variasi kata dan sebagian informasi kontekstual.

3.4 Pemodelan IndoBERT dan IndoBERTweet

Pemodelan dilakukan dengan menerapkan *class weight* sebagai teknik penanganan ketidakseimbangan data pada data latih. Bobot kelas dihitung berdasarkan proporsi jumlah data pada tiap kelas, dimana kelas dengan jumlah data lebih sedikit memperoleh bobot yang lebih besar.

Tabel 5: Hasil perhitungan *class weight*.

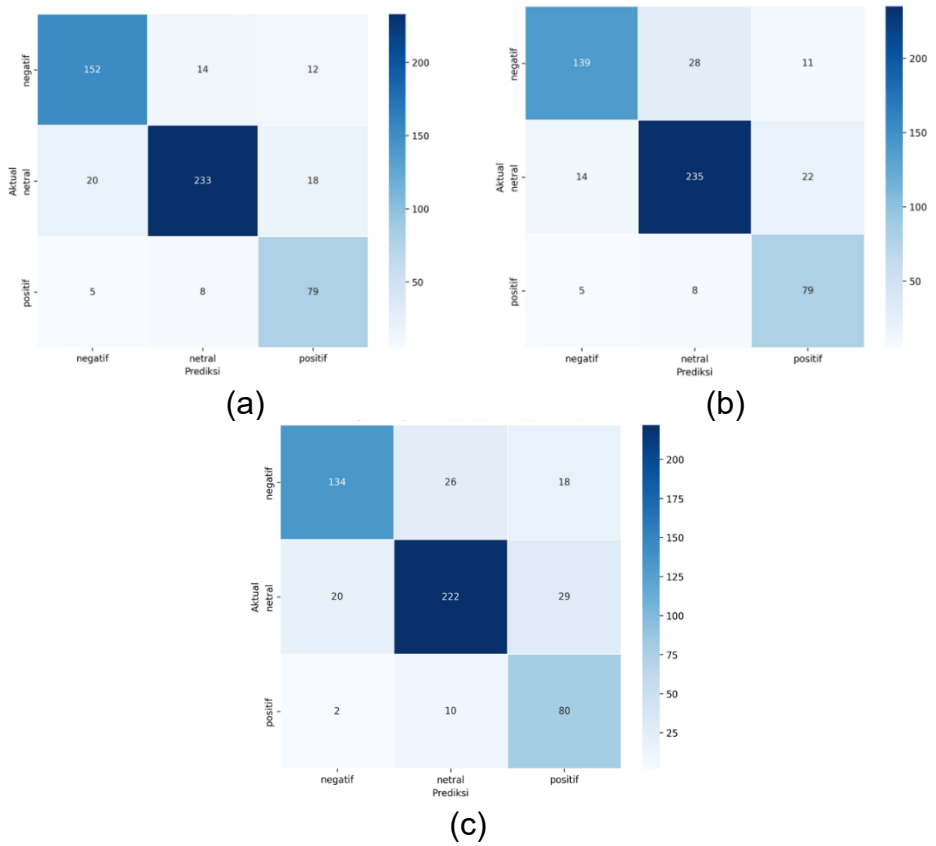
Kelas	*Class Weight*
Negatif	1,0138
Netral	0,6645
Positif	1,9654

Berdasarkan Tabel 5, kelas positif memiliki bobot terbesar karena jumlah datanya paling sedikit dibandingkan kelas lainnya. Sebaliknya, kelas netral memiliki bobot paling kecil karena merupakan kelas mayoritas. Pemberian bobot tersebut bertujuan agar model dapat mempelajari setiap kelas secara lebih seimbang selama proses pelatihan. Selanjutnya, pemodelan dilakukan menggunakan beberapa kombinasi *hyperparameter* untuk memperoleh konfigurasi terbaik pada setiap skenario *preprocessing*. Proses pemilihan *hyperparameter* dilakukan menggunakan *grid search* dengan evaluasi pada data validasi. Kombinasi *hyperparameter* terbaik dipilih berdasarkan nilai F1-*score macro* tertinggi pada data validasi. Hasil tuning *hyperparameter* pada masing-masing model dan skenario ditampilkan pada Tabel 6.

Tabel 6: Hasil *tuning hyperparameter* pada data validasi.

Skenario	Model	*Epoch*	*Lrate*	*Batch*	*Val Acc*	*Val Prec Macro*	*Val Rec Macro*	*Val F1-score Macro*
Skenario 1	IndoBERT	3	3e-05	32	0.864	0.852	0.855	0.853
	IndoBERTweet	2	3e-05	32	0.874	0.856	0.883	0.867
Skenario 2	IndoBERT	2	2e-05	32	0.855	0.844	0.845	0.844
	IndoBERTweet	3	3e-05	32	0.872	0.858	0.870	0.864
Skenario 3	IndoBERT	4	2e-05	16	0.837	0.827	0.822	0.824
	IndoBERTweet	4	2e-05	32	0.848	0.829	0.852	0.839

Hasil deteksi kelas sentimen ditampilkan menggunakan *confusion matrix* untuk menentukan kinerja model dalam setiap skenario.

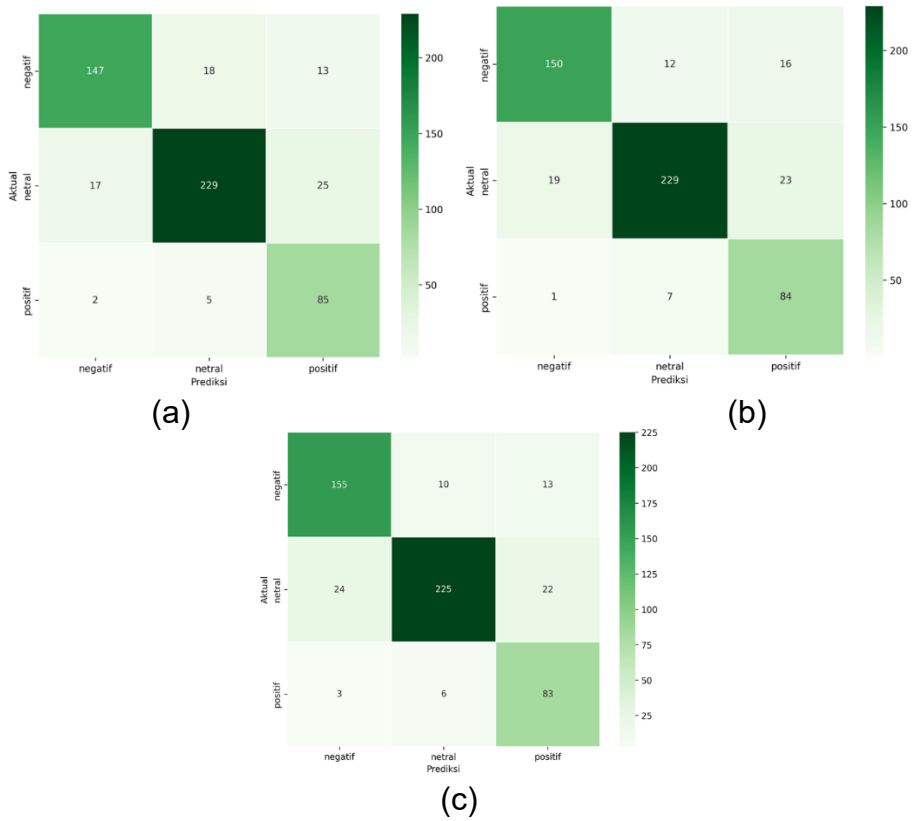


Gambar 3: *Confusion matrix* model IndoBERT (a) skenario 1 (b) skenario 2 (c) Skenario 3

Berdasarkan *confusion matrix* pada Gambar 3, performa model IndoBERT menunjukkan bahwa tidak semua kelas mencapai nilai tertinggi pada skenario yang sama. Pada skenario 1, kelas negatif menghasilkan jumlah prediksi benar tertinggi sebesar 152 data, sedangkan kelas netral mencapai 233 data dan kelas positif 79 data. Selain itu, distribusi kesalahan klasifikasi pada skenario ini relatif lebih seimbang dibandingkan skenario lainnya. Pada skenario 2, kelas netral mengalami peningkatan dengan jumlah prediksi benar sebesar 235 data, yang merupakan nilai tertinggi untuk kelas tersebut di antara seluruh skenario. Namun, pada saat yang sama terjadi penurunan performa pada kelas negatif menjadi 139 data, serta peningkatan kesalahan klasifikasi dari negatif ke netral dari 14 menjadi 28 data. Hal ini menunjukkan bahwa peningkatan pada satu kelas tidak selalu diikuti oleh peningkatan performa secara keseluruhan. Pada skenario 3 kelas positif menunjukkan peningkatan jumlah prediksi benar menjadi 80 data. Namun, performa pada kelas negatif mengalami penurunan menjadi 134 data, dan kesalahan klasifikasi pada kelas netral meningkat. Kondisi ini menunjukkan bahwa model IndoBERT yang melalui *preprocessing*

kompleks tidak selalu menghasilkan peningkatan performa yang konsisten pada seluruh kelas.

Secara keseluruhan skenario 1 memberikan performa terbaik diantara 3 skenario. Hal tersebut disebabkan oleh keseimbangan prediksi yang lebih baik antar kelas serta jumlah kesalahan klasifikasi yang lebih rendah dibandingkan skenario lainnya. Pada skenario 2 dan 3, meskipun menunjukkan peningkatan pada satu kelas tertentu, penurunan performa pada kelas lain serta bertambahnya kesalahan prediksi menyebabkan performa total menjadi lebih rendah.



Gambar 4: *Confusion matrix* model IndoBERTweet (a) skenario 1 (b) skenario 2 (c) skenario 3

Berdasarkan *confusion matrix* pada Gambar 4, performa model IndoBERTweet skenario 1 menunjukkan bahwa jumlah prediksi benar tertinggi terdapat pada kelas netral sebesar 229 data, diikuti kelas negatif sebesar 147 data dan kelas positif sebesar 85 data. Kesalahan klasifikasi masih cukup tinggi, terutama pada data negatif yang diprediksi sebagai netral dengan 18 data, serta data netral yang diprediksi sebagai positif dengan 25 data, yang menunjukkan adanya kesulitan model dalam membedakan batas antar kelas tersebut. Pada skenario 2, jumlah prediksi benar pada kelas negatif meningkat menjadi 150 data, sementara kelas netral tetap 229 data dan kelas positif sedikit menurun menjadi 84 data. Kesalahan pada data negatif yang diprediksi sebagai netral menurun menjadi 12 data, namun kesalahan pada kelas lainnya tidak mengalami

perubahan yang signifikan. Pada skenario 3, jumlah prediksi benar pada kelas negatif kembali meningkat menjadi 155 data, sedangkan kelas netral dan positif menurun menjadi 225 data dan 83 data. Kesalahan klasifikasi pada kelas netral masih cukup tinggi, yaitu diprediksi sebagai negatif 24 data dan sebagai positif 22 data, hal tersebut menunjukkan bahwa kelas netral tetap menjadi kelas yang paling sulit diprediksi. Secara keseluruhan, perubahan performa antar skenario tidak menunjukkan peningkatan yang konsisten pada semua kelas. Peningkatan pada satu kelas cenderung diikuti oleh penurunan pada kelas lain. Sehingga diperlukan evaluasi lebih lanjut menggunakan matriks evaluasi untuk menentukan skenario *preprocessing* yang memberikan performa terbaik secara keseluruhan.

3.5 Perbandingan Hasil Evaluasi

Evaluasi dilakukan menggunakan metrik *balanced accuracy* dan F1-*score macro*. Dimana F1-*score macro* digunakan sebagai acuan utama dalam membandingkan performa model.

Tabel 7: Perbandingan performa model IndoBERT dan IndoBERTweet pada data uji.

Skenario	Model	*Balanced Accuracy*	F1-*Score Macro*
Skenario 1	IndoBERT	0,8575	0,8428
Skenario 2	IndoBERT	0,8356	0,8230
Skenario 3	IndoBERT	0,8139	0,7908
Skenario 1	IndoBERTweet	0,8649	0,8404
Skenario 2	IndoBERTweet	0,8669	0,8420
Skenario 3	IndoBERTweet	0,8677	0,8435

Secara umum, model IndoBERT menunjukkan performa terbaik pada skenario 1 dengan nilai *balanced accuracy* sebesar 0,8575 dan F1-*score macro* sebesar 0,8428. Hal tersebut mengindikasikan bahwa IndoBERT cenderung optimal ketika melalui *preprocessing* minimal, yaitu hanya *case folding*. Kondisi tersebut diduga karena struktur dan konteks kalimat masih terjaga dengan baik, sehingga model dapat memanfaatkan informasi kontekstual secara lebih optimal. Pada skenario 2 dan skenario 3, performa IndoBERT mengalami penurunan dengan nilai *balanced accuracy* 0,8356 dan 0,8139 serta F1-*score macro* sebesar 0,8230 dan 0,7908. Hal tersebut mengindikasikan bahwa tahapan *preprocessing* yang lebih kompleks, seperti *cleaning, normalization, stopword removal,* dan *stemming* berpotensi menghilangkan informasi kontekstual yang penting, sehingga memengaruhi kemampuan model dalam memahami makna kalimat secara utuh. Sedangkan model IndoBERTweet menunjukkan tren performa yang meningkat seiring bertambahnya kompleksitas *preprocessing*. Nilai *balanced accuracy* meningkat dari 0,8649 pada skenario 1 menjadi 0,8669 pada skenario 2 dan mencapai nilai tertinggi sebesar 0,8677 pada skenario 3. Peningkatan serupa juga terlihat pada nilai F1-*score macro*, yaitu dari

0,8404 menjadi 0,8420 dan mencapai 0,8435 pada skenario 3. Nilai tersebut merupakan performa terbaik secara keseluruhan dibandingkan seluruh model dan skenario yang diuji. Nilai *balanced accuracy* yang relatif tinggi menunjukkan bahwa IndoBERTweet mampu mengenali setiap kelas secara lebih seimbang, termasuk pada kelas minoritas. Perbedaan performa antara kedua model diduga dipengaruhi oleh karakteristik data pelatihan masing-masing model. IndoBERT dilatih menggunakan korpus teks berbahasa Indonesia yang cenderung formal dan terstruktur, sehingga lebih sensitif terhadap perubahan struktur kalimat akibat *preprocessing* yang berlebihan. Sebaliknya, IndoBERTweet dilatih menggunakan data media sosial yang bersifat tidak baku dan mengandung banyak variasi bahasa. Oleh karena itu, *preprocessing* yang lebih kompleks justru membantu mengurangi *noise* dan meningkatkan kualitas representasi teks pada IndoBERTweet.

4. Simpulan

Perbedaan skenario *preprocessing* menghasilkan variasi nilai *balanced accuracy* dan F1-*score macro* pada model IndoBERT dan IndoBERTweet, meskipun perbedaannya relatif tidak terlalu besar. Hasil tersebut menunjukkan bahwa tahapan *preprocessing* memengaruhi performa dan stabilitas model dalam melakukan klasifikasi sentimen. Model IndoBERT menunjukkan performa terbaik ketika *preprocessing* minimal, yaitu skenario 1 dengan nilai F1-*score macro* 0,8428, kemudian performanya menurun pada *preprocessing* yang lebih kompleks. Sebaliknya, model IndoBERTweet menunjukkan peningkatan performa seiring dengan bertambahnya kompleksitas *preprocessing*, dengan hasil terbaik pada skenario 3 dan nilai F1-*score macro* sebesar 0,8435. Secara keseluruhan, IndoBERTweet memiliki performa yang lebih baik dan lebih stabil dibandingkan IndoBERT pada data media sosial yang digunakan. Hal tersebut menunjukkan bahwa kesesuaian antara karakteristik data dan model memiliki peran penting dalam menentukan hasil klasifikasi sentimen.

Daftar Pustaka

Badan Nasional Penanggulangan Bencana. (2024). *Data bencana Indonesia 2024*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Dewi, B. E. S. (2025). Model analisis sentimen pada kendaraan listrik menggunakan algoritman IndoBERTweet dan IndoBERT. *Jurnal Ilmiah Teknologi Informasi, 19*(1), 169–179. https://doi.org/10.35457/antivirus.v19i1.4416

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

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

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*, 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10660–10668). Association for Computational Linguistics. https://arxiv.org/abs/2109.04607

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Muzakir, A., Adi, K., & Kusumaningrum, R. (2023). Short text classification based on hybrid semantic expansion and bidirectional GRU (BiGRU) based method to improve hate speech detection. *Revue d'Intelligence Artificielle, 37*(6), 1471–1481. https://doi.org/10.18280/ria.370611

Qolbu, A. A., Fitriyati, N., & Inayah, N. (2025). Performa Naïve Bayes, SVM, dan IndoBERT pada analisis sentimen Twitter IndiHome dengan strategi penanganan data tidak seimbang. *Jurnal Fourier, 14*(1), 29–44. https://doi.org/10.14421/fourier.2025.141.29-44

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems, 121*, 102342. https://doi.org/10.1016/j.is.2023.102342

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & others. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 2020 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 843–857). Association for Computational Linguistics.