

Preprocessing Effects on Word2Vec and IndoRoBERTa for Shopee Sentiment Classification

Lilik Avitadia Prichanti ^{1*}, Kinanti Rizky Pangestutik ^{2*}, Cici Suhaeni ^{3*}, Bagus Sartono ^{4*}

* Statistika dan Sains Data, Institut Pertanian Bogor

lilikavitadia@apps.ipb.ac.id ¹, kinantirizkypangestutik@apps.ipb.ac.id ², cici_suhaeni@apps.ipb.ac.id ³, bagusco@apps.ipb.ac.id ⁴

Article Info

Article history:

Received 2026-05-15

Revised 2026-09-10

Accepted 2026-10-14

Keyword:

IndoRoBERTa,
Indonesian Sentiment Analysis,
Shopee Reviews,
Text Preprocessing,
Word2Vec.

ABSTRACT

This study evaluates the effects of heavy and minimal preprocessing on static and contextual text representations for sentiment classification of Indonesian Shopee reviews. The dataset comprised 29,618 Google Play Store reviews collected in January 2026 and weakly labeled from star ratings into negative, neutral, and positive classes. The data were divided into training, validation, and test sets using an 80:10:10 stratified split. Heavy preprocessing included case folding, text cleaning, slang normalization, stopword removal, and stemming, whereas minimal preprocessing applied case folding only. Four experimental configurations combined the two preprocessing strategies with Word2Vec–Logistic Regression and fine-tuned IndoRoBERTa. Model performance was evaluated using accuracy, weighted precision, weighted recall, weighted F1-score, macro F1-score, and confusion matrices. Heavy preprocessing produced a marginal weighted F1-score improvement for Word2Vec–Logistic Regression, from 86.0% to 86.2%. Minimal preprocessing improved IndoRoBERTa’s weighted F1-score from 88.4% to 89.7% and achieved the best accuracy of 90.7% and macro F1-score of 65.8%. The difference between weighted and macro F1-scores indicates that the neutral minority class remained difficult to identify. The findings show that preprocessing effectiveness depends on model architecture. Minimal preprocessing better preserves contextual information for IndoRoBERTa, whereas aggressive preprocessing provides only a marginal benefit for the static baseline.



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

I. PENDAHULUAN

Ulasan pengguna aplikasi *e-commerce* merupakan sumber data tekstual yang penting untuk memahami pengalaman dan persepsi konsumen terhadap kualitas layanan digital. Pada aplikasi Shopee, ulasan Google Play Store memuat opini pengguna dalam bentuk bahasa Indonesia informal, termasuk singkatan, *slang*, kesalahan ejaan, campuran bahasa, dan ekspresi emosional. Kondisi tersebut menjadikan analisis sentimen sebagai pendekatan yang relevan untuk mengelompokkan opini pengguna ke dalam kelas positif, netral, dan negatif secara otomatis.

Tahap preprocessing dapat mengubah struktur teks, jumlah token, dan variasi kosakata yang digunakan sebagai masukan model klasifikasi. Pengolahan teks bahasa Indonesia perlu mempertimbangkan *slang*, singkatan, kesalahan ejaan, serta bentuk bahasa tidak baku yang banyak ditemukan pada data

buatan pengguna [1]. Dampak setiap teknik preprocessing juga dapat berbeda berdasarkan jenis representasi teks dan karakteristik tugas klasifikasi [2]. Penghapusan stopwords dan stemming bahkan tidak selalu meningkatkan performa model transformer berbahasa Indonesia [3].

Model klasik umumnya menggunakan representasi kata statis untuk mengubah teks menjadi vektor numerik. Word2Vec mempelajari representasi suatu kata berdasarkan konteks kemunculannya dalam korpus [4]. Pendekatan FastText kemudian menggunakan informasi subword untuk menangani variasi morfologi dan kata yang tidak terdapat dalam kosakata [5]. Representasi Word2Vec juga telah diterapkan bersama Long Short-Term Memory untuk melakukan klasifikasi sentimen pada ulasan hotel berbahasa Indonesia [6].

Penerapan model transformer pada ulasan *e-commerce* menunjukkan kemampuan yang baik dalam memahami teks

pengguna berbahasa Indonesia. IndoBERT telah digunakan untuk mengklasifikasikan sentimen ulasan Shopee dan Bukalapak yang diperoleh dari Google Play Store [7]. Model yang sama juga telah diterapkan pada ulasan aplikasi Lazada, Shopee, dan Tokopedia dengan penanganan ketidakseimbangan kelas melalui oversampling [8]. Penelitian lain menunjukkan bahwa IndoBERT mampu mengungguli metode berbasis LSTM dan word embedding pada ulasan produk Tokopedia [9]. Perbandingan IndoBERT dan Bi-LSTM pada ulasan aplikasi Shopee juga menunjukkan keunggulan model transformer dalam menangkap informasi kontekstual [10].

Model transformer menghasilkan representasi kontekstual karena setiap token diproses berdasarkan hubungannya dengan token lain dalam kalimat. BERT menggunakan arsitektur Transformer encoder dua arah untuk membentuk representasi token berdasarkan konteks kiri dan kanan [11]. RoBERTa mengembangkan pendekatan BERT melalui pengoptimalan prosedur prapelatihan dan penghapusan tugas next sentence prediction [12]. Evaluasi pada analisis sentimen berbasis BERT menunjukkan bahwa dampak preprocessing bergantung pada jenis operasi yang diterapkan dan karakteristik data penelitian [13].

Kajian terbaru pada ulasan aplikasi Indonesia memperlihatkan bahwa representasi teks berpengaruh terhadap performa klasifikasi sentimen. Perbandingan TF-IDF, FastText, dan IndoBERT menunjukkan bahwa representasi kontekstual memberikan hasil yang lebih tinggi dibandingkan representasi sparse dan subword pada ulasan aplikasi perjalanan [14]. Fine-tuning IndoBERT pada ulasan pariwisata juga masih menghadapi kesulitan dalam membedakan sentimen netral dan sentimen campuran [15]. Penggabungan IndoBERT dengan Recurrent Convolutional Neural Network menghasilkan peningkatan performa klasifikasi tiga kelas pada data ulasan bahasa Indonesia [16]. Penerapan IndoBERT pada ulasan aplikasi BRImo menunjukkan bahwa penyeimbangan data dapat membantu model memperoleh performa yang lebih merata pada kelas negatif, netral, dan positif [17].

Ketidakseimbangan kelas menjadi tantangan penting karena model dapat lebih banyak mempelajari pola kelas mayoritas daripada kelas minoritas. Generasi kalimat sintesis berbasis GPT-3 telah digunakan untuk meningkatkan performa klasifikasi pada data sentimen yang tidak seimbang [18], [19]. Random oversampling juga dapat meningkatkan performa IndoBERT pada kelas minoritas, meskipun penerapannya berpotensi meningkatkan risiko overfitting [20]. Kondisi tersebut menunjukkan bahwa strategi penanganan ketidakseimbangan kelas perlu dipisahkan dari evaluasi pengaruh preprocessing agar sumber perubahan performa dapat ditafsirkan secara lebih jelas.

Kajian terdahulu telah mengevaluasi model transformer, teknik preprocessing, dan penanganan ketidakseimbangan kelas pada berbagai data ulasan. Kajian yang secara terkontrol membandingkan heavy preprocessing dan minimal preprocessing pada Word2Vec-Logistic Regression dan IndoRoBERTa dengan partisi data serta parameter yang sama

masih terbatas. Penelitian ini mengisi keterbatasan tersebut melalui empat konfigurasi eksperimen yang membandingkan dua strategi preprocessing pada representasi kata statis dan representasi kontekstual.

Penelitian ini bertujuan membandingkan pengaruh heavy preprocessing dan minimal preprocessing terhadap kinerja Word2Vec-Logistic Regression dan fine-tuned IndoRoBERTa pada klasifikasi sentimen ulasan Shopee. Penelitian ini mengukur perubahan performa pada setiap model, membandingkan representasi statis dan kontekstual, serta mengidentifikasi strategi preprocessing yang lebih sesuai untuk setiap arsitektur. Kontribusi penelitian terletak pada penggunaan partisi data dan parameter yang sama dalam seluruh konfigurasi, pelaporan weighted F1-score dan macro F1-score pada data tidak seimbang, serta analisis kesalahan pada kelas netral.

II. METODE

A. Sumber Data dan Pelabelan

Data penelitian berupa ulasan pengguna aplikasi Shopee yang diperoleh dari Google Play Store pada periode 1-31 Januari 2026. *Dataset* memuat teks ulasan, *rating* bintang, nama pengguna, dan waktu ulasan. Pelabelan sentimen dilakukan secara otomatis berbasis *rating* bintang, yaitu *rating* 1-2 sebagai kelas negatif, *rating* 3 sebagai kelas netral, dan *rating* 4-5 sebagai kelas positif. Skema pelabelan berbasis *rating* dipilih karena efisien untuk data berskala besar, meskipun tetap berpotensi mengandung *noise* ketika isi teks tidak sepenuhnya konsisten dengan *rating* pengguna.

Pelabelan berbasis *rating* memiliki keterbatasan karena *rating* tidak selalu sepenuhnya konsisten dengan isi teks ulasan. Misalnya, pengguna dapat memberikan *rating* rendah tetapi menulis komentar singkat yang netral, atau memberikan *rating* tinggi tetapi tetap menyampaikan keluhan tertentu. Oleh karena itu, label dalam penelitian ini diperlakukan sebagai *weak* label yang merepresentasikan kecenderungan sentimen berdasarkan penilaian pengguna, bukan sebagai anotasi manual yang sepenuhnya bebas dari *noise*. Keterbatasan ini dipertimbangkan dalam pembahasan hasil, terutama ketika menafsirkan kesalahan klasifikasi pada kelas netral.

Data yang disertakan dalam analisis adalah ulasan yang memiliki teks ulasan tidak kosong, *rating* bintang valid pada skala 1 sampai 5, dan berada pada periode 1–31 Januari 2026. Data yang dikeluarkan adalah baris dengan teks kosong, *rating* tidak valid, tanggal di luar periode penelitian, serta duplikasi ulasan apabila ditemukan pada tahap pemeriksaan data. Setelah proses pemeriksaan data, jumlah observasi akhir yang digunakan dalam eksperimen adalah 29.618 ulasan.

B. Skenario Preprocessing

Penelitian ini membandingkan dua skenario preprocessing. Skenario A atau *heavy preprocessing* mencakup *case folding*, *text cleaning*, normalisasi kata informal atau slang, *stopword removal*, dan *stemming* menggunakan Sastrawi. Tahap *text cleaning* pada Skenario A meliputi penghapusan URL, angka,

tanda baca, mention, dan karakter khusus. Normalisasi slang dilakukan menggunakan kamus normalisasi berbasis pasangan kata informal dan kata baku. Contoh normalisasi yang digunakan antara lain “gk”, “ga”, dan “gak” menjadi “tidak”; “yg” menjadi “yang”; “dgn” menjadi “dengan”; “apk” dan “app” menjadi “aplikasi”; serta “lemot” menjadi “lambat”. Tahap ini bertujuan mengurangi variasi bentuk kata yang memiliki makna serupa agar representasi teks pada model Word2Vec menjadi lebih konsisten.

Skenario B atau *minimal preprocessing* hanya menerapkan *case folding*. Pada skenario ini, teks ulasan dipertahankan sedekat mungkin dengan bentuk aslinya. Penelitian ini tidak menghapus URL, angka, tanda baca, emoji, karakter khusus, maupun *stopword* pada Skenario B. Strategi ini digunakan agar struktur kalimat, urutan token, dan konteks linguistik tetap tersedia bagi model transformer. Perbedaan kedua skenario dirancang untuk menguji apakah pembersihan teks agresif yang umum digunakan pada model klasik tetap relevan untuk model transformer kontekstual.

C. Analisis Deskriptif

Analisis deskriptif dilakukan untuk memahami perubahan karakteristik data teks pada setiap skenario *preprocessing*. Analisis ini mencakup distribusi label sentimen, distribusi *rating*, panjang ulasan, jumlah *token*, ukuran kosakata, dan kata dominan. Visualisasi deskriptif yang digunakan meliputi diagram batang, histogram, grafik perbandingan jumlah *token*, dan *wordcloud*. Analisis ini penting karena perubahan performa model dapat dipengaruhi oleh perubahan struktur teks akibat *preprocessing*, bukan hanya oleh arsitektur model.

D. Pemodelan

Dua pendekatan pemodelan digunakan dalam penelitian ini. Model *baseline* menggunakan Word2Vec dengan arsitektur *Skip-gram* untuk menghasilkan representasi vektor kata. Representasi dokumen diperoleh melalui *mean pooling* dari vektor kata penyusun ulasan, kemudian digunakan sebagai fitur masukan bagi Logistic Regression untuk klasifikasi tiga kelas sentimen. Model modern menggunakan *fine-tuning* IndoRoBERTa, yaitu model *transformer* berbahasa Indonesia yang menghasilkan representasi kontekstual berdasarkan urutan *token* dan konteks kalimat. IndoRoBERTa diadaptasi untuk klasifikasi sentimen dengan menambahkan lapisan klasifikasi pada keluaran model.

Seluruh model diuji pada dua skenario *preprocessing* sehingga terdapat empat kombinasi eksperimen, yaitu Word2Vec + Logistic Regression pada Skenario A, Word2Vec + Logistic Regression pada Skenario B, IndoRoBERTa pada Skenario A, dan IndoRoBERTa pada Skenario B. Data dibagi menjadi data latih, validasi, dan uji dengan rasio 80:10:10 menggunakan *stratified sampling* agar distribusi label pada setiap partisi tetap proporsional. Pembagian data dilakukan secara bertingkat menggunakan *stratified sampling* agar proporsi kelas negatif, netral, dan positif tetap terjaga pada setiap subset. Tahap pertama memisahkan 80% data sebagai data latih dan 20% data sebagai data sementara. Tahap kedua membagi data

sementara tersebut menjadi data validasi dan data uji dengan proporsi yang sama, sehingga komposisi akhir data menjadi 80% data latih, 10% data validasi, dan 10% data uji. Proses pembagian data menggunakan *random seed* yang sama pada seluruh eksperimen agar perbandingan antarskenario dan antarmodel dilakukan pada partisi data yang identik.

Penelitian ini tidak menerapkan *class weighting*, *oversampling*, *undersampling*, atau *resampling* pada tahap pelatihan. Keputusan ini diambil agar pengaruh strategi *preprocessing* terhadap performa model dapat diamati tanpa intervensi tambahan terhadap distribusi kelas. Dengan demikian, seluruh model dibandingkan pada kondisi distribusi data yang sama. Dampak ketidakseimbangan kelas kemudian dianalisis melalui *weighted F1-score*, *macro F1-score*, dan *confusion matrix*.

E. Konfigurasi Model

Word2Vec dilatih menggunakan data latih pada setiap skenario *preprocessing* dengan arsitektur Skip-gram. Model menggunakan dimensi vektor 100, window size 5, minimum count 2, dan 10 epoch. Representasi setiap ulasan dibentuk melalui *mean pooling* terhadap vektor kata yang tersedia. Logistic Regression menggunakan solver *lbfgs*, skema multinomial, maksimum 1.000 iterasi, dan *random seed* 42.

Model transformer menggunakan checkpoint *flax-community/indonesian-roberta-base* dengan lapisan klasifikasi tiga kelas. Proses *fine-tuning* menggunakan maksimum panjang input 128 token, tiga epoch, batch size pelatihan 16, batch size evaluasi 32, learning rate 2×10^{-5} , weight decay 0,01, dan *random seed* 42. Evaluasi dan penyimpanan model dilakukan pada setiap epoch. Model terbaik dipilih berdasarkan *weighted F1-score* pada data validasi, sedangkan *early stopping* menggunakan *patience* sebanyak dua evaluasi.

Penelitian ini tidak menerapkan hyperparameter tuning melalui grid search atau random search. Seluruh parameter ditetapkan secara sama pada Skenario A dan Skenario B agar perbedaan hasil lebih dapat dikaitkan dengan strategi *preprocessing*, bukan dengan konfigurasi pelatihan yang berbeda.

F. Evaluasi

Evaluasi dilakukan menggunakan akurasi (*accuracy*), *precision*, *recall*, *F1-score weighted*, *macro F1-score*, dan *confusion matrix*. *Weighted F1-score* digunakan sebagai metrik utama karena distribusi kelas pada data ulasan tidak seimbang. *Macro F1-score* dilaporkan sebagai metrik tambahan untuk melihat kemampuan model pada setiap kelas secara setara tanpa dominasi kelas mayoritas. *Confusion matrix* digunakan untuk mengidentifikasi pola kesalahan prediksi, terutama pada kelas netral yang cenderung menjadi kelas minoritas. Hasil evaluasi divisualisasikan melalui *confusion matrix* pada data validasi dan data uji, serta grafik perbandingan *weighted F1-score* pada data uji.

III. HASIL DAN PEMBAHASAN

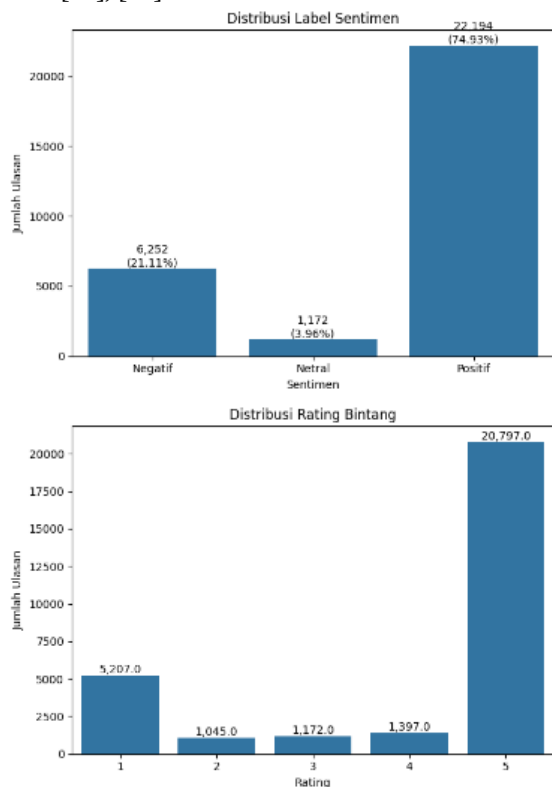
A. Hasil Analisis Deskriptif

Dataset akhir terdiri atas 29.618 ulasan Shopee. Distribusi label menunjukkan ketidakseimbangan kelas yang cukup kuat. Kelas positif mendominasi dengan 22.194 ulasan atau 74,93%, diikuti kelas negatif sebanyak 6.252 ulasan atau 21,11%, dan kelas netral sebanyak 1.172 ulasan atau 3,96%. Pola ini mengindikasikan bahwa sebagian besar pengguna memberikan *rating* tinggi terhadap aplikasi Shopee, tetapi proporsi ulasan negatif tetap cukup besar untuk merepresentasikan keluhan pengguna.

TABEL I
DISTRIBUSI LABEL SENTIMEN

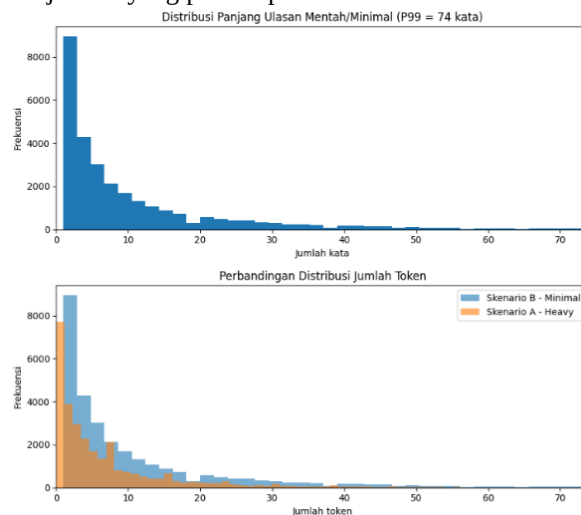
Label (<i>Rating</i>)	Jumlah	Persentase
Negatif (1-2)	6252	21,11%
Netral (3)	1172	3,96%
Positif (4-5)	22194	74,93%
Total	29618	100,00%

Ketidakseimbangan label tersebut berdampak langsung pada strategi evaluasi. Akurasi berpotensi memberikan kesan performa yang tinggi ketika model cenderung memprediksi kelas mayoritas. Oleh karena itu, pembahasan hasil perlu menekankan *weighted F1-score* sebagai metrik utama dan *macro F1-score* sebagai indikator kemampuan model mengenali kelas minoritas. Temuan pada penelitian *class imbalance* berbasis generasi kalimat sintetis juga memperkuat pentingnya memperhatikan distribusi kelas dalam analisis sentimen [18], [19].



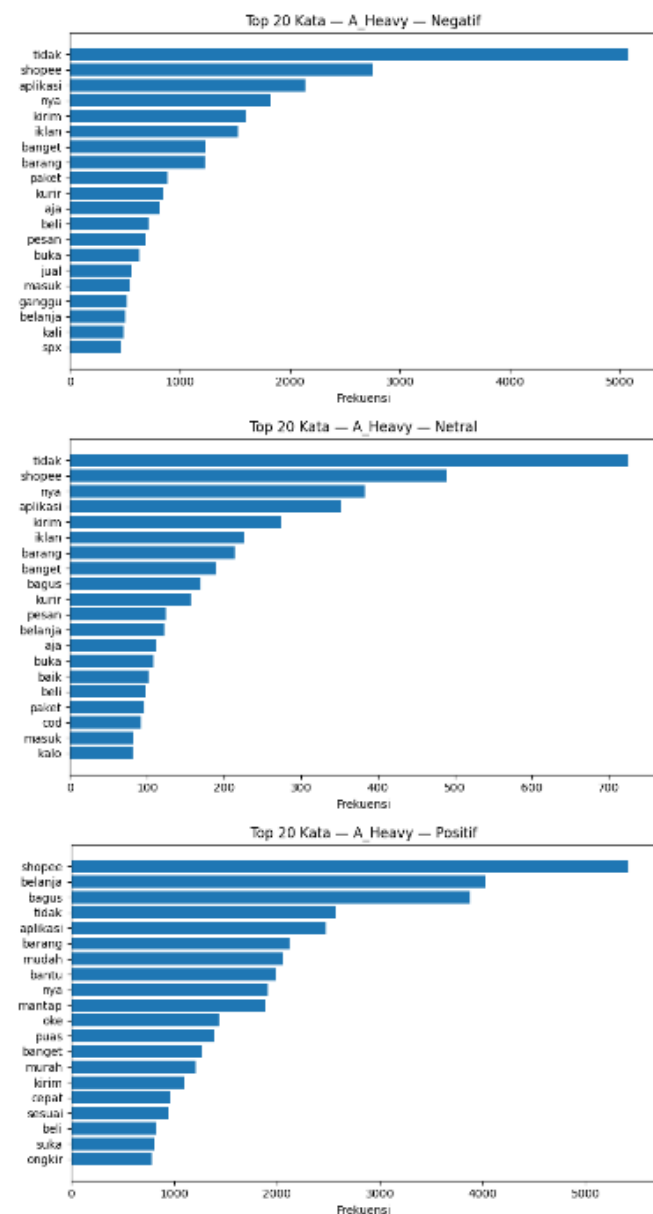
Gambar 1. Distribusi label sentimen dan *rating* bintang ulasan Shopee.

Gambar 1 memperlihatkan dominasi kelas positif pada data ulasan Shopee. Distribusi *rating* juga menunjukkan bahwa *rating* 5 menjadi *rating* yang paling banyak diberikan oleh pengguna, yaitu sebanyak 20.797 ulasan atau 70,22% dari total data. Pola ini menunjukkan bahwa sebagian besar pengguna memberikan penilaian positif terhadap aplikasi Shopee. Namun, jumlah *rating* 1 yang mencapai 5.207 ulasan tetap menunjukkan bahwa keluhan pengguna masih muncul dalam jumlah yang perlu diperhatikan.

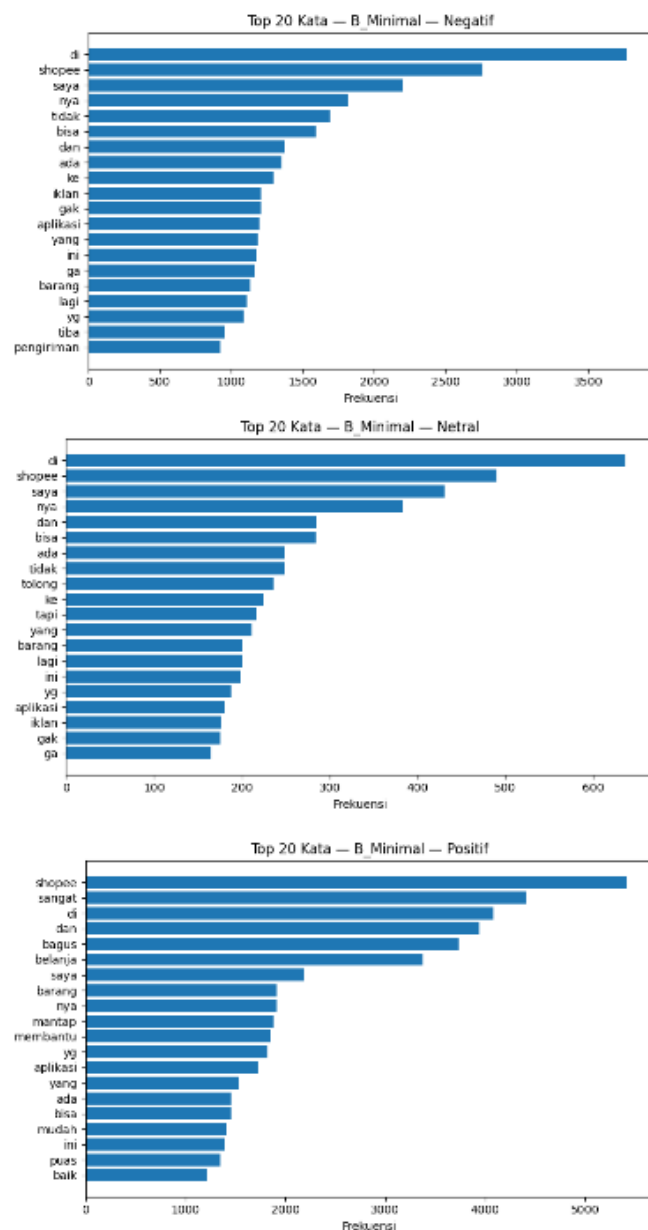


Gambar 2. Distribusi panjang ulasan sebelum *preprocessing* dan perbandingan jumlah *token* pada Skenario A dan Skenario B.

Gambar 2 menunjukkan bahwa sebagian besar ulasan memiliki panjang teks yang relatif pendek. Median panjang ulasan sebelum *preprocessing* adalah 5 kata, sedangkan rata-rata panjang ulasan adalah 11,34 kata. Setelah *preprocessing*, Skenario A menghasilkan rata-rata 7,01 *token*, sedangkan Skenario B menghasilkan rata-rata 11,34 *token*. Perbedaan ini menunjukkan bahwa *heavy preprocessing* mengurangi panjang teks secara cukup jelas karena proses *cleaning*, *stopword removal*, dan *stemming* menghapus atau menyederhanakan sebagian *token*.



Gambar 3. Dua puluh kata dominan pada Skenario A berdasarkan kelas sentimen.



Gambar 4. Dua puluh kata dominan pada Skenario B berdasarkan kelas sentimen.

Gambar 3 dan Gambar 4 menunjukkan perbedaan kata dominan pada Skenario A dan Skenario B. Pada Skenario A, kata-kata yang muncul lebih banyak berupa kata bermakna inti karena *stopword removal* dan *stemming* menghapus sebagian kata fungsi. Pada Skenario B, kata dominan masih mencakup kata fungsi seperti “di”, “dan”, “nya”, dan “yang” karena skenario ini hanya menerapkan *case folding*. Kondisi ini menunjukkan bahwa Skenario B mempertahankan struktur bahasa yang lebih utuh, sedangkan Skenario A menghasilkan representasi teks yang lebih ringkas.



Gambar 5. Wordcloud ulasan Shopee pada Skenario A (a) dan Skenario B (b).

Wordcloud pada Gambar 5 memperkuat pola yang terlihat pada grafik kata dominan. Skenario A menghasilkan visualisasi kata yang lebih terfokus pada kata inti seperti “shopee”, “aplikasi”, “belanja”, “barang”, dan “bayar”. Sebaliknya, Skenario B masih mempertahankan banyak kata fungsi yang muncul secara alami dalam ulasan pengguna. Perbedaan tersebut menjadi bukti awal bahwa strategi *preprocessing* dapat mengubah struktur representasi teks sebelum tahap pemodelan dilakukan.

B. Hasil Evaluasi Model

Rekapitulasi hasil evaluasi empat kombinasi eksperimen disajikan pada Tabel II. Hasil evaluasi menunjukkan bahwa model IndoRoBERTa dengan Skenario B menghasilkan performa terbaik dibandingkan kombinasi eksperimen lainnya. Model tersebut memperoleh akurasi sebesar 90,7%, *weighted precision* sebesar 89,3%, *weighted recall* sebesar 90,7%, *weighted F1-score* sebesar 89,7%, dan *macro F1-score* sebesar 65,8%. Hasil ini menunjukkan bahwa *minimal preprocessing* memberikan keuntungan bagi IndoRoBERTa karena teks ulasan tetap mempertahankan struktur kalimat, urutan *token*, dan konteks semantik yang dibutuhkan oleh model *transformer*.

TABEL II
REKAPITULASI KINERJA MODEL PADA DATA UJI

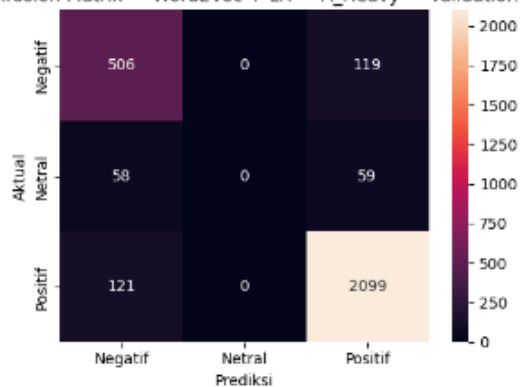
Eksperimen	Acc. (%)	P. (%)	R. (%)	F1 W. (%)	Macro F1 (%)
W2V+LR A	87,8	84,7	87,8	86,2	56,8
W2V+LR B	87,6	84,4	87,6	86,0	56,5
IndoRoBERTa A	89,6	87,8	89,6	88,4	61,8
IndoRoBERTa B	90,7	89,3	90,7	89,7	65,8

Keterangan: Acc. = akurasi; P. = *precision*; R. = *recall*; F1 W. = *weighted F1-score*; dan Macro F1 = *macro F1-score*.

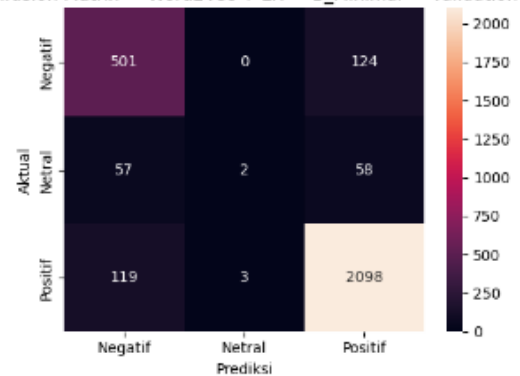
Tabel II menunjukkan bahwa IndoRoBERTa dengan Skenario B menghasilkan kinerja agregat tertinggi, yaitu akurasi 90,7%, *weighted precision* 89,3%, *weighted recall* 90,7%, *weighted F1-score* 89,7%, dan *macro F1-score* 65,8%. Hasil tersebut menunjukkan bahwa minimal preprocessing menghasilkan performa terbaik pada IndoRoBERTa dalam eksperimen ini. Nilai *macro F1-score* yang lebih rendah daripada *weighted F1-score* memperlihatkan bahwa performa model belum merata pada seluruh kelas.

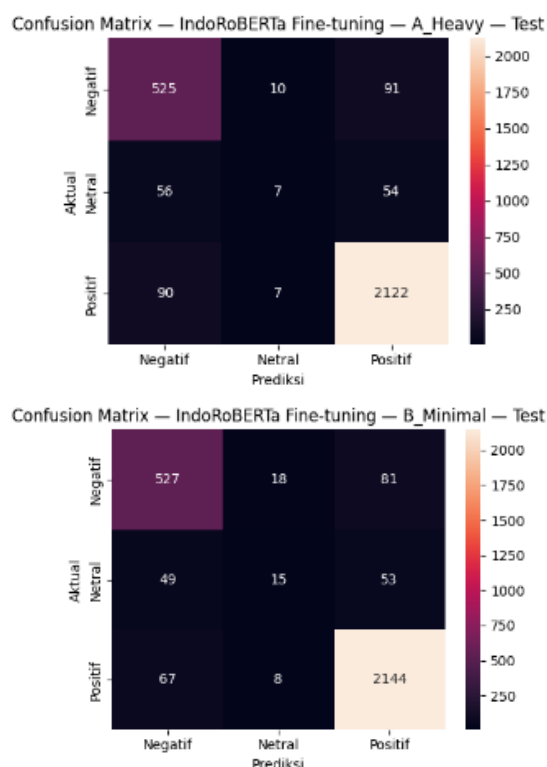
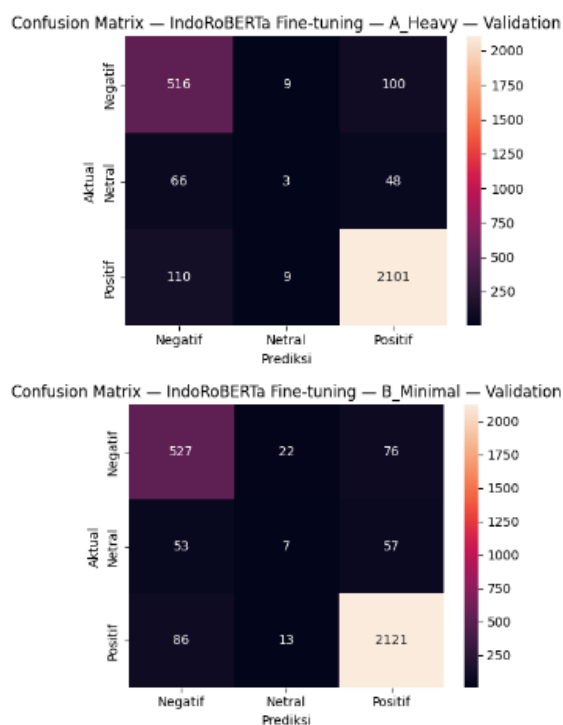
Word2Vec-Logistic Regression dengan Skenario A menghasilkan *weighted F1-score* 86,2%, sedangkan Skenario B menghasilkan 86,0%. Selisih sebesar 0,2 poin persentase menunjukkan bahwa keuntungan heavy preprocessing pada model baseline bersifat marginal. IndoRoBERTa memperoleh peningkatan *weighted F1-score* sebesar 1,3 poin persentase ketika menggunakan minimal preprocessing, yaitu dari 88,4% menjadi 89,7%. Perbedaan tersebut bersifat deskriptif karena penelitian ini tidak menerapkan repeated cross-validation atau uji signifikansi antarkonfigurasi.

Confusion Matrix — Word2Vec + LR — A_Heavy — Validation

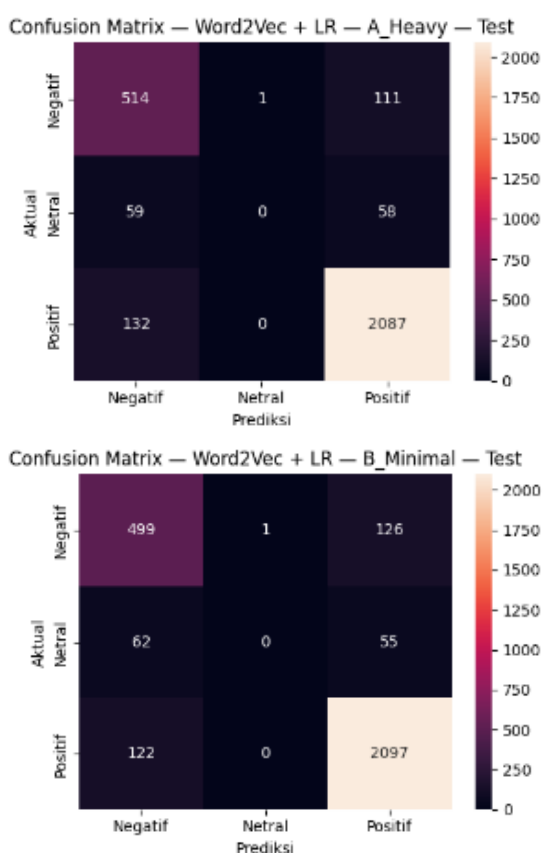


Confusion Matrix — Word2Vec + LR — B_Minimal — Validation





Gambar 6. Confusion matrix pada data validasi untuk empat kombinasi eksperimen.



Gambar 7. Confusion matrix pada data uji untuk empat kombinasi eksperimen.

Gambar 6 dan Gambar 7 menunjukkan pola kesalahan prediksi pada data validasi dan data uji. Pada model Word2Vec + Logistic Regression, kelas netral cenderung sulit dikenali karena sebagian besar data netral diprediksi sebagai kelas negatif atau positif. Pola tersebut menunjukkan bahwa model *baseline* masih belum mampu membedakan ulasan netral yang memiliki karakteristik campuran. Pada model IndoRoBERTa, jumlah prediksi benar pada kelas netral meningkat, terutama pada Skenario B, meskipun performa kelas netral masih lebih rendah dibandingkan kelas positif dan negatif. Kondisi ini konsisten dengan distribusi data yang sangat tidak seimbang, karena kelas netral hanya mencakup 3,96% dari total data.

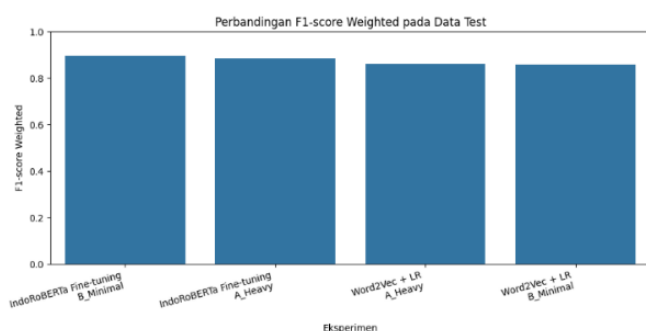
TABEL III
KINERJA PER KELAS INDOBERTA SKENARIO B PADA DATA UJI

Kelas	Precision	Recall	F1-score	Support
Negatif	0,820	0,842	0,831	626
Netral	0,366	0,128	0,190	117
Positif	0,941	0,966	0,954	2.219

Tabel III menunjukkan bahwa IndoRoBERTa Skenario B menghasilkan F1-score 0,831 pada kelas negatif dan 0,954 pada kelas positif. Kinerja tersebut berbeda cukup jauh dari kelas netral yang hanya memperoleh precision 0,366, recall 0,128, dan F1-score 0,190. Recall kelas netral menunjukkan bahwa model hanya mengidentifikasi 15 dari 117 ulasan netral secara benar.

Rendahnya kinerja kelas netral berkaitan dengan proporsi kelas yang hanya mencapai 3,96%, karakteristik ulasan rating

3 yang sering memuat pujian dan keluhan secara bersamaan, serta potensi ketidaksesuaian antara rating dan isi teks. Kondisi ini menjelaskan perbedaan yang cukup besar antara weighted F1-score dan macro F1-score.



Gambar 8. Perbandingan *weighted F1-score* pada data uji.

Gambar 8 memperlihatkan bahwa IndoRoBERTa dengan Skenario B memperoleh *weighted F1-score* tertinggi pada data uji. Model tersebut menghasilkan *weighted F1-score* sebesar 89,7%, diikuti oleh IndoRoBERTa Skenario A sebesar 88,4%, Word2Vec + Logistic Regression Skenario A sebesar 86,2%, dan Word2Vec + Logistic Regression Skenario B sebesar 86,0%. Hasil ini memperkuat temuan bahwa model *transformer* lebih diuntungkan oleh teks yang mempertahankan struktur dan konteks aslinya, sedangkan model berbasis Word2Vec hanya memperoleh sedikit keuntungan dari *heavy preprocessing*.

C. Analisis Kesalahan Klasifikasi

Analisis kesalahan dilakukan dengan meninjau pola prediksi pada *confusion matrix* data uji. Hasil *confusion matrix* menunjukkan bahwa kelas netral menjadi kelas yang paling sulit dikenali oleh seluruh model. Kesalahan ini dapat terjadi karena ulasan dengan rating 3 sering memiliki karakteristik semantik yang campuran, misalnya mengandung pujian dan keluhan dalam satu komentar. Selain itu, sebagian ulasan netral memiliki panjang teks yang pendek sehingga informasi kontekstual yang tersedia bagi model menjadi terbatas.

Pada model Word2Vec + Logistic Regression, kesalahan pada kelas netral cenderung lebih tinggi karena representasi kata yang digunakan bersifat statis dan tidak mempertimbangkan konteks penggunaan kata dalam kalimat. Pada model IndoRoBERTa, jumlah prediksi benar pada kelas netral meningkat, tetapi kelas ini tetap memiliki performa lebih rendah dibandingkan kelas positif. Temuan ini menunjukkan bahwa representasi kontekstual membantu model memahami variasi sentimen, tetapi tidak sepenuhnya mengatasi masalah ketidakseimbangan kelas dan ambiguitas label berbasis *rating*.

Kesalahan klasifikasi juga dapat dipengaruhi oleh karakteristik ulasan Shopee yang cenderung pendek, informal, dan sering memuat ekspresi tidak baku. Ulasan pendek seperti “lumayan”, “biasa aja”, atau “oke tapi lemot” dapat memiliki interpretasi sentimen yang ambigu. Kondisi

tersebut membuat model lebih mudah memprediksi ulasan ke kelas mayoritas atau ke kelas yang memiliki kata sentimen lebih dominan. Oleh karena itu, hasil ini menunjukkan bahwa peningkatan performa kelas minoritas tidak cukup hanya dilakukan melalui pemilihan model, tetapi juga memerlukan strategi penanganan ketidakseimbangan kelas dan peningkatan kualitas label.

D. Pembahasan Perbandingan Skenario Preprocessing

Pada model Word2Vec + Logistic Regression, Skenario A menghasilkan *weighted F1-score* sebesar 86,2%, sedangkan Skenario B menghasilkan *weighted F1-score* sebesar 86,0%. Perbedaan sebesar 0,2 poin persentase menunjukkan bahwa *heavy preprocessing* memberikan peningkatan yang sangat kecil terhadap model *baseline*. Hasil ini dapat dijelaskan oleh karakteristik Word2Vec yang menghasilkan representasi kata statis. Model tersebut lebih sensitif terhadap variasi bentuk kata, imbuhan, *slang*, dan ejaan tidak baku. Oleh karena itu, normalisasi *slang*, *stopword removal*, dan *stemming* masih dapat membantu mengurangi variasi kosakata pada data ulasan.

Pada model IndoRoBERTa, Skenario B menghasilkan performa terbaik dengan *weighted F1-score* sebesar 89,7%, sedangkan Skenario A menghasilkan *weighted F1-score* sebesar 88,4%. Hasil ini menunjukkan bahwa *minimal preprocessing* lebih sesuai untuk model *transformer*. IndoRoBERTa memanfaatkan struktur kalimat, urutan *token*, dan konteks semantik secara lebih utuh. *Preprocessing* agresif berpotensi menghilangkan kata fungsi, negasi, dan ekspresi informal yang penting dalam pembentukan makna sentimen. Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa *stopword removal* dan *stemming* tidak selalu meningkatkan performa model *transformer* berbahasa Indonesia [3], serta dengan kajian *preprocessing* pada *BERT-based sentiment analysis* [13].

E. Pembahasan Perbandingan Word2Vec dan IndoRoBERTa

Perbandingan antara Word2Vec + Logistic Regression dan IndoRoBERTa menunjukkan bahwa model *transformer* menghasilkan performa yang lebih tinggi pada kedua skenario *preprocessing*. Pada Skenario A, IndoRoBERTa memperoleh *weighted F1-score* sebesar 88,4%, sedangkan Word2Vec + Logistic Regression memperoleh *weighted F1-score* sebesar 86,2%. Pada Skenario B, IndoRoBERTa memperoleh *weighted F1-score* sebesar 89,7%, sedangkan Word2Vec + Logistic Regression memperoleh *weighted F1-score* sebesar 86,0%. Hasil ini menunjukkan bahwa representasi kontekstual pada IndoRoBERTa lebih mampu menangkap makna sentimen yang implisit, informal, dan bergantung pada konteks kalimat.

Keunggulan IndoRoBERTa dapat dijelaskan melalui arsitektur *transformer* yang memproses *token* berdasarkan konteks kalimat secara menyeluruh. Pendekatan ini berbeda dari Word2Vec yang memberikan satu representasi tetap untuk setiap kata tanpa mempertimbangkan konteks penggunaannya. Secara konseptual, hasil ini sejalan dengan

pengembangan BERT dan RoBERTa sebagai model bahasa kontekstual [11], [12]. Hasil ini juga konsisten dengan penelitian sebelumnya yang menunjukkan bahwa representasi transformer memberikan performa lebih baik dibandingkan representasi teks klasik pada klasifikasi sentimen bahasa Indonesia [10], [14].

F. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan. Pertama, pelabelan sentimen dilakukan secara otomatis berdasarkan *rating* bintang sehingga label yang digunakan berpotensi mengandung *noise*. Pelabelan berbasis *rating* efisien untuk data berskala besar, tetapi belum sepenuhnya menggantikan anotasi manual karena isi teks ulasan tidak selalu konsisten dengan *rating* yang diberikan pengguna.

Kedua, penelitian ini belum menerapkan teknik penanganan ketidakseimbangan kelas seperti *class weighting*, *resampling*, *focal loss*, atau *synthetic text generation*. Keputusan tersebut dilakukan agar pengaruh *preprocessing* dapat diamati tanpa intervensi tambahan terhadap distribusi kelas. Namun, rendahnya *macro F1-score* dibandingkan *weighted F1-score* menunjukkan bahwa strategi penanganan kelas minoritas perlu diuji pada penelitian selanjutnya.

Ketiga, perbandingan model dalam penelitian ini dibatasi pada Word2Vec + Logistic Regression sebagai *baseline* klasik dan IndoRoBERTa sebagai model transformer. Model lain seperti FastText, IndoBERT, dan XLM-RoBERTa belum diuji. Oleh karena itu, hasil penelitian ini lebih tepat ditafsirkan sebagai bukti empiris mengenai pengaruh *preprocessing* pada dua paradigma representasi teks, yaitu representasi kata statis dan representasi kontekstual transformer, bukan sebagai perbandingan menyeluruh terhadap seluruh model NLP modern.

Keempat, penelitian ini belum melakukan *ablation study* terhadap setiap komponen *heavy preprocessing* secara individual. Dengan demikian, penelitian ini belum dapat memastikan apakah peningkatan kecil pada Word2Vec + Logistic Regression berasal dari normalisasi *slang*, *stopword removal*, *stemming*, atau kombinasi seluruh tahapan tersebut. Penelitian lanjutan dapat menguji tahapan *preprocessing* secara bertahap untuk mengetahui kontribusi masing-masing tahap.

Kelima, penelitian ini belum membandingkan kebutuhan komputasi dan waktu pelatihan secara sistematis. Secara konseptual, IndoRoBERTa membutuhkan sumber daya komputasi lebih besar dibandingkan Word2Vec + Logistic Regression karena proses *fine-tuning* melibatkan pembaruan parameter model transformer. Analisis waktu pelatihan dan efisiensi komputasi dapat menjadi fokus penelitian lanjutan agar rekomendasi model tidak hanya mempertimbangkan performa klasifikasi, tetapi juga biaya komputasi.

Keenam, generalisasi hasil penelitian ini masih terbatas pada ulasan aplikasi Shopee di Google Play Store pada periode Januari 2026. Karakteristik bahasa, distribusi *rating*, dan pola keluhan pengguna dapat berbeda pada platform *e-commerce* lain atau pada media sosial. Oleh karena itu, penelitian selanjutnya perlu menguji strategi *preprocessing*

yang sama pada domain lain, seperti Tokopedia, Lazada, atau komentar media sosial, untuk menilai konsistensi temuan.

IV. KESIMPULAN

Penelitian ini mengevaluasi pengaruh dua strategi *preprocessing* terhadap performa analisis sentimen ulasan Shopee menggunakan model Word2Vec + Logistic Regression dan IndoRoBERTa. Hasil penelitian menunjukkan bahwa efektivitas *preprocessing* bergantung pada arsitektur model yang digunakan. Pada model Word2Vec + Logistic Regression, Skenario A atau *heavy preprocessing* menghasilkan *weighted F1-score* sebesar 86,2%, sedangkan Skenario B atau *minimal preprocessing* menghasilkan *weighted F1-score* sebesar 86,0%. Hasil tersebut menunjukkan bahwa *heavy preprocessing* memberikan peningkatan kecil pada model berbasis representasi kata statis.

Pada model IndoRoBERTa, Skenario B menghasilkan performa terbaik dengan akurasi sebesar 90,7%, *weighted F1-score* sebesar 89,7%, dan *macro F1-score* sebesar 65,8%. Hasil tersebut menunjukkan bahwa *minimal preprocessing* lebih sesuai untuk IndoRoBERTa karena strategi ini mempertahankan struktur kalimat dan konteks *token* yang dibutuhkan oleh model *transformer*. Dengan demikian, penelitian ini menyimpulkan bahwa *preprocessing* tidak dapat diperlakukan sebagai prosedur universal. Strategi *preprocessing* perlu disesuaikan dengan karakteristik model yang digunakan.

Meskipun IndoRoBERTa dengan *minimal preprocessing* menghasilkan performa terbaik, selisih antara *weighted F1-score* dan *macro F1-score* menunjukkan bahwa model masih menghadapi kesulitan dalam mengenali kelas minoritas, terutama kelas netral. Hal ini berkaitan dengan distribusi data yang tidak seimbang dan potensi *noise* pada pelabelan berbasis *rating*. Oleh karena itu, hasil penelitian ini tidak hanya menunjukkan keunggulan *minimal preprocessing* pada model transformer, tetapi juga menegaskan perlunya evaluasi yang mempertimbangkan performa setiap kelas secara lebih seimbang.

Penelitian selanjutnya disarankan untuk memperluas kamus normalisasi *slang*, mengevaluasi pengaruh emoji dan negasi secara khusus, serta membandingkan pelabelan berbasis *rating* dengan anotasi manual. Selain itu, strategi penanganan ketidakseimbangan kelas seperti *class weighting*, *focal loss*, *oversampling*, atau *synthetic text generation* dapat dipertimbangkan untuk meningkatkan performa pada kelas minoritas, khususnya kelas netral.

DAFTAR PUSTAKA

- [1] C. Suhaeni, S. A. Kamila, F. Fahira, M. Yusran, and G. A. Dito, "Exploring a Large Language Model on the ChatGPT Platform for Indonesian Text Preprocessing Tasks," *Indonesian Journal of Statistics and Its Applications*, vol. 9, no. 1, pp. 100–116, 2025, doi: 10.29244/ijsa.v9i1p100-116.
- [2] N. Babanejad, A. Agrawal, A. An, and M. Papagelis, "A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks," in *Proceedings of the 58th Annual*

- Meeting of the Association for Computational Linguistics, 2020, pp. 5799–5810, doi: 10.18653/v1/2020.acl-main.514.
- [3] U. Khairani, V. Mutiawani, and H. Ahmadian, “Pengaruh Tahapan Preprocessing terhadap Model IndoBERT dan IndoBERTweet untuk Mendeteksi Emosi pada Komentar Akun Berita Instagram,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 4, pp. 887–894, 2024, doi: 10.25126/jtiik.1148315.
- [4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv:1301.3781, 2013.
- [5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017, doi: 10.1162/tac1_a_00051.
- [6] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, “Sentiment Analysis Using Word2Vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews,” *Procedia Computer Science*, vol. 179, pp. 728–735, 2021, doi: 10.1016/j.procs.2021.01.061.
- [7] K. C. Pradhisa and R. Fajriyah, “Analisis Sentimen Ulasan Pengguna E-commerce di Google Play Store Menggunakan Metode IndoBERT,” *Building of Informatics, Technology and Science*, vol. 6, no. 1, pp. 92–104, 2024, doi: 10.47065/bits.v6i1.5247.
- [8] A. R. Aisy and G. Karyono, “Analisis Sentimen terhadap Ulasan Google Play Store Aplikasi Lazada, Shopee, dan Tokopedia Menggunakan Algoritma IndoBERT,” *Building of Informatics, Technology and Science*, vol. 7, no. 3, pp. 2127–2135, 2025, doi: 10.47065/bits.v7i3.8745.
- [9] B. P. Aji and C. S. K. Aditya, “Klasifikasi Sentimen Ulasan Produk pada Platform E-Commerce di Indonesia dengan Menggunakan Model Pre-Trained IndoBERT,” *Building of Informatics, Technology and Science*, vol. 6, no. 4, pp. 2491–2500, 2025, doi: 10.47065/bits.v6i4.6968.
- [10] T. Ramadhan and K. Kusnawi, “Performance Comparison of IndoBERT and Bi-LSTM Models for Sentiment Analysis of Shopee App Users in Indonesia,” *Journal of Applied Informatics and Computing*, vol. 10, no. 2, pp. 2076–2085, 2026, doi: 10.30871/jaic.v10i2.10996.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [12] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, 2019.
- [13] M. Pota, M. Ventura, H. Fujita, and M. Esposito, “Multilingual Evaluation of Pre-processing for BERT-Based Sentiment Analysis of Tweets,” *Expert Systems with Applications*, vol. 181, Art. no. 115119, 2021, doi: 10.1016/j.eswa.2021.115119.
- [14] C. T. Tangdilomban, S. Y. Ramdhan, M. Rizal, C. Suhaeni, and B. Sartono, “From Sparse Features to Transformers: A Statistical Evaluation of TF-IDF, FastText, and IndoBERT for Sentiment Classification of Indonesian Travel App Reviews,” *Journal of Applied Informatics and Computing*, vol. 10, no. 3, pp. 3071–3083, 2026, doi: 10.30871/jaic.v10i3.12610.
- [15] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, “Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 30–40, 2025, doi: 10.20473/jisebi.11.1.30-40.
- [16] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, “Sentiment Analysis of Indonesian Reviews Using Fine-Tuning IndoBERT and R-CNN,” *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.
- [17] A. A. P. Simarmata and T. B. Sasongko, “Sentiment Analysis on BRImo Application Reviews Using IndoBERT,” *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 851–862, 2025, doi: 10.30871/jaic.v9i3.8162.
- [18] C. Suhaeni and H.-S. Yong, “Enhancing Imbalanced Sentiment Analysis: A GPT-3-Based Sentence-by-Sentence Generation Approach,” *Applied Sciences*, vol. 14, no. 2, Art. no. 622, 2024, doi: 10.3390/app14020622.
- [19] C. Suhaeni and H.-S. Yong, “Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences,” *Applied Sciences*, vol. 13, no. 17, Art. no. 9766, 2023, doi: 10.3390/app13179766.
- [20] D. R. Alfinsyah and B. P. Hartato, “Evaluating the Impact of Random Over Sampling on IndoBERT Performance for Indonesian Sentiment Analysis,” *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 3270–3282, 2025, doi: 10.30871/jaic.v9i6.11488.