BAB IV
HASIL DAN PEMBAHASAN

4.1 Gambaran Umum Dataset

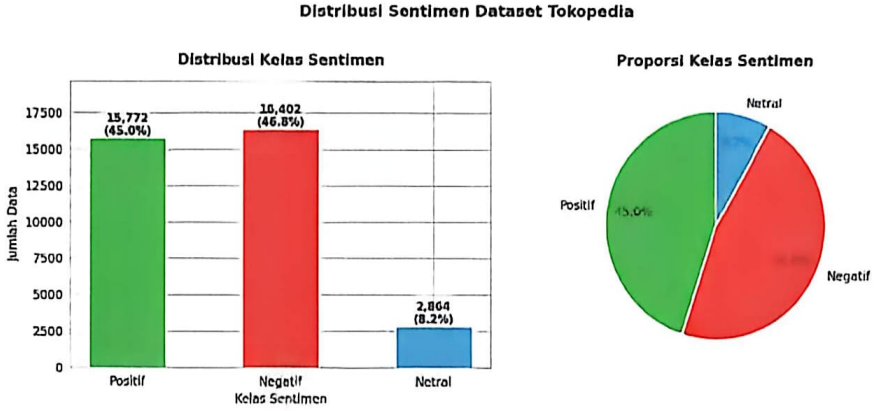
Penelitian ini menggunakan dataset ulasan pengguna aplikasi Tokopedia yang dikumpulkan secara langsung melalui teknik web scraping dari Google Play Store menggunakan library google-play-scraper berbasis Python. Dataset awal yang berhasil dikumpulkan berjumlah 35.038 ulasan. Setelah melalui proses preprocessing yang meliputi pembersihan teks, penghapusan data duplikat, dan penghapusan baris dengan teks kosong, dataset bersih yang digunakan dalam penelitian ini berjumlah 27.018 ulasan.

Setiap ulasan dilabeli ke dalam tiga kelas sentimen berdasarkan rating bintang yang diberikan pengguna, yaitu Positif (rating 4-5 bintang), Netral (rating 3 bintang), dan Negatif (rating 1-2 bintang). Distribusi kelas sentimen pada dataset bersih disajikan pada Tabel 4.1 berikut.

Tabel 4. 1 Distribusi Kelas Sentimen Dataset

Kelas Sentimen	Jumlah Data	Persentase(%)
Positif	8.209	30,4%
Negatif	16.003	59,2%
Netral	2.807	10,4%
Total	27.018	100%

Berdasarkan Tabel 4.1, dapat diketahui bahwa kelas Negatif mendominasi dataset dengan jumlah 16.003 ulasan (59,2%), diikuti kelas Positif sebanyak 8.209 ulasan (30,4%), dan kelas Netral sebanyak 2.807 ulasan (10,4%). Ketidakseimbangan kelas (class imbalance) ini terutama terlihat pada kelas Netral yang hanya berjumlah 10,4% dari total data. Kondisi ini berpotensi menyebabkan model cenderung bias terhadap kelas mayoritas, sehingga diperlukan teknik penanganan khusus menggunakan SMOTE (Synthetic Minority Oversampling Technique) yang akan dibahas pada subbab berikutnya.



Gambar 4.1. Sentimen Dataset Tokopedia

4.2 Hasil Preprocessing Data

Tahap preprocessing merupakan langkah krusial dalam penelitian analisis sentimen berbasis teks. Tujuan preprocessing adalah membersihkan dan menstandarisasi teks ulasan dari noise yang dapat mengganggu proses pemodelan. Proses preprocessing dalam penelitian ini meliputi beberapa tahapan yang diterapkan secara berurutan.

1. Tahapan Preprocessing

Tahapan preprocessing yang diterapkan pada dataset meliputi: (1) Case Folding, yaitu mengubah seluruh teks menjadi huruf kecil untuk menghindari duplikasi kata yang berbeda kapitalisasi; (2) Penghapusan URL, mention, dan hashtag; (3) Penghapusan emoji dan karakter non-ASCII; (4) Penghapusan angka dan karakter khusus; (5) Normalisasi kata slang bahasa Indonesia menggunakan kamus normalisasi yang telah dikompilasi secara manual dengan 50 entri kata tidak baku beserta padanan bakunya; serta (6) Penghapusan spasi berlebih.

Sebelum	Sesudah	Label	
BANYAK BANGET DISKONANNYA SUKAAAAðŸ'šðŸ'šðŸ'š	banyak banget diskonannya sukaaaa	Positif	Cleaning
halaman checkout dari kemaron kok gangguan terus????	halaman checkout dari kemaron kok gangguan terus	Negatif	Cleaning
tokopedia very good very well, well ðŸ˜œðŸ‘ðŸ»	tokopedia very good very well	Positif	Cleaning
tingkatkan lagi agar lebih bagus apk nya	tingkatkan sedang agar lebih bagus aplikasi nya	Positif	Normalisasi

pembelian Lewat Browser Tidak bisa COD	pembelian lewat browser tidak bisa cod	Netral	Case Folding

Setelah proses preprocessing selesai, baris dengan teks kosong dan duplikat dihapus. Hasil preprocessing menunjukkan pengurangan data dari 35.038 menjadi 27.018 ulasan, yang berarti sebanyak 8.020 baris dihapus karena merupakan duplikat atau mengandung teks kosong setelah preprocessing.

Tabel 4. 2 Perbandingan Statistik Teks Sebelum dan Sesudah Preprocessing

Statistik	Sebelum Preprocessing	Sesudah Preprocessing
Total Data	35.038	27.018
Data Dihapus	-	8.020
Rata-rata Panjang Teks	± 12–15 kata	± 8–12 kata

Proses normalisasi kata slang memberikan kontribusi signifikan dalam menyeragamkan representasi kata dalam teks. Sebagai contoh, kata 'gak', 'ga', 'ngga', dan 'nggak' seluruhnya dinormalisasi menjadi 'tidak'; kata 'bgt' dan 'bngt' dinormalisasi menjadi 'banget'; serta singkatan 'app' dan 'apk' dinormalisasi menjadi 'aplikasi'. Normalisasi ini penting mengingat karakteristik ulasan di Google Play Store yang bersifat informal dan mengandung banyak variasi penulisan kata bahasa Indonesia.

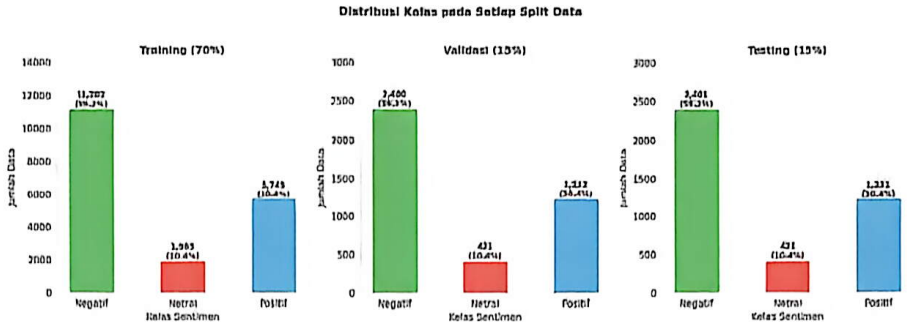
4.3 Pembagian Dataset

Dataset yang telah melalui proses preprocessing selanjutnya dibagi menjadi tiga bagian menggunakan metode stratified split untuk memastikan proporsi setiap kelas sentimen terjaga pada masing-masing subset data. Pembagian dilakukan dengan rasio 70:15:15 untuk data training, validasi, dan testing. Hasil pembagian data disajikan pada Tabel 4.3 berikut.

Tabel 4. 3 Hasil Pembagian Dataset

Subset	Proporsi	Positif	Negatif	Netral
Training	70%	5.746	11.202	1.964
Validasi	15%	1.232	2.400	421
Testing	15%	1.231	2.401	421
Total	100%	8.209	16.003	2.806

Penggunaan stratified split memastikan bahwa proporsi kelas pada setiap subset data konsisten. Hal ini penting untuk menghindari bias dalam proses pelatihan dan evaluasi model. Data training berjumlah 18.912 sampel digunakan untuk melatih model, data validasi berjumlah 4.053 sampel digunakan untuk memantau performa model selama pelatihan dan melakukan early stopping, sedangkan data testing berjumlah 4.053 sampel digunakan untuk evaluasi akhir model.



Gambar 4.2. Split Data

4.4 Penanganan Ketidakseimbangan Kelas dengan SMOTE

Ketidakseimbangan kelas pada dataset merupakan tantangan utama dalam penelitian ini, terutama pada kelas Netral yang hanya berjumlah 10,4% dari total data. Untuk mengatasi masalah ini, penelitian menggunakan teknik SMOTE (Synthetic Minority Oversampling Technique) yang diterapkan pada data embedding hasil ekstraksi IndoBERT sebelum proses pelatihan XGBoost.

SMOTE bekerja dengan membuat sampel sintetis baru untuk kelas minoritas menggunakan interpolasi linier antara sampel yang sudah ada dengan tetangga terdekatnya (k-nearest neighbors). Teknik ini dipilih karena lebih efektif dibandingkan oversampling acak biasa yang hanya menduplikasi data yang ada.

Secara rinci, algoritma SMOTE bekerja melalui empat langkah utama untuk setiap sampel pada kelas minoritas. Pertama, dipilih satu sampel x_i dari kelas minoritas (misalnya kelas Netral). Kedua, dicari k tetangga terdekat (umumnya k=5) dari sampel tersebut yang berasal dari kelas minoritas yang sama, dihitung berdasarkan jarak Euclidean pada ruang vektor embedding 768-dimensi hasil ekstraksi IndoBERT. Ketiga, salah satu dari k tetangga tersebut dipilih secara acak sebagai x_zi. Keempat, sampel sintetis baru dibentuk melalui interpolasi linier

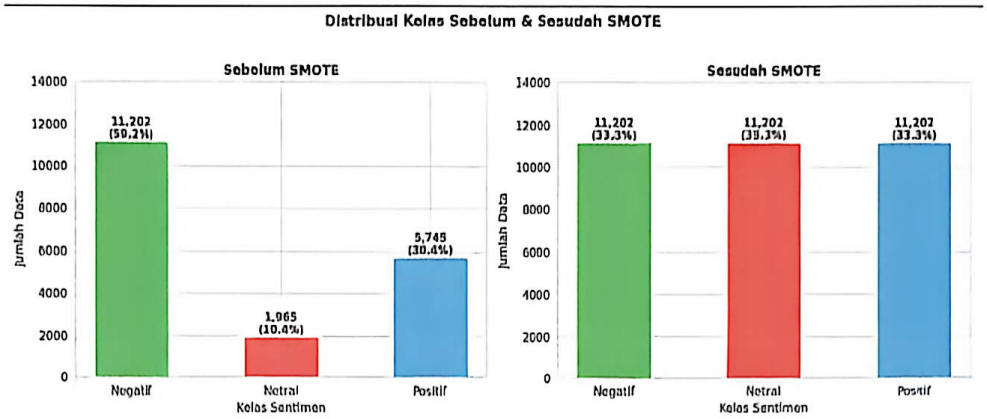
dengan rumus x_baru = x_i + λ × (x_zi − x_i), di mana λ adalah bilangan acak antara 0 dan 1. Dengan demikian, sampel baru yang dihasilkan bukan duplikat, melainkan titik baru pada ruang fitur yang terletak di antara sampel asli dan tetangganya, sehingga menambah variasi data tanpa mengulang informasi yang identik. Proses ini diulang secara berkala hingga jumlah sampel pada kelas minoritas menyamai jumlah sampel pada kelas mayoritas.

Berdasarkan Tabel 4.4, kelas Negatif yang berjumlah 11.202 sampel dijadikan acuan (kelas mayoritas) sehingga tidak mengalami penambahan. Pada kelas Positif, SMOTE menambahkan 5.456 sampel sintetis (dari 5.746 menjadi 11.202 sampel), sedangkan pada kelas Netral, yang sebelumnya paling sedikit jumlahnya, SMOTE menambahkan 9.238 sampel sintetis (dari 1.964 menjadi 11.202 sampel). Sampel-sampel sintetis yang ditambahkan tersebut seluruhnya berupa vektor embedding hasil interpolasi, bukan teks ulasan baru, karena SMOTE diterapkan setelah tahap ekstraksi fitur oleh IndoBERT. Dengan penambahan total 14.694 sampel sintetis (5.456 pada kelas Positif dan 9.238 pada kelas Netral), ketiga kelas menjadi seimbang dengan masing-masing 11.202 sampel, sehingga total data training yang digunakan untuk melatih XGBoost meningkat dari 18.912 menjadi 33.606 sampel.

Tabel 4. 4 Distribusi Kelas Sebelum dan Sesudah SMOTE

Kelas	Sebelum SMOTE	Sesudah SMOTE	Penambahan
Positif	5.746	11.202	+5.456
Negatif	11.202	11.202	-

Netral	1.964	11.202	+9.238



Gambar 4.3. Kelas Sebelum dan Sesudah SMOTE

Setelah penerapan SMOTE, ketiga kelas memiliki jumlah sampel yang seimbang yaitu masing-masing 11.202 sampel, sehingga total data training yang digunakan untuk melatih XGBoost menjadi 33.606 sampel. Penerapan SMOTE hanya dilakukan pada data training untuk menghindari data leakage pada proses evaluasi.

4.5 Hasil Fine-tuning IndoBERT

Model IndoBERT yang digunakan dalam penelitian ini adalah indobenchmark/indobert-base-p2, sebuah model pre-trained transformer berbasis arsitektur BERT yang telah dilatih pada korpus bahasa Indonesia berukuran besar. Fine-tuning dilakukan dengan menambahkan classification head berupa fully connected layer dengan 3 output neuron yang sesuai dengan jumlah kelas sentimen.

Proses fine-tuning dilakukan selama 3 epoch dengan learning rate 2e-5, batch size 16, warmup steps 500, dan weight decay 0.01. Optimisasi menggunakan AdamW optimizer dengan FP16 mixed precision untuk mempercepat proses

pelatihan. Early stopping dengan patience 2 epoch diterapkan untuk mencegah overfitting berdasarkan nilai F1 Macro pada data validasi.

Model IndoBERT dalam penelitian ini berfungsi ganda, yaitu sebagai model standalone untuk klasifikasi sentimen secara langsung, dan sebagai feature extractor yang menghasilkan representasi vektor 768 dimensi (CLS embedding) yang kemudian digunakan sebagai input pada model XGBoost dalam arsitektur hybrid.

4.5.1 Hasil Evaluasi IndoBERT Standalone

Evaluasi IndoBERT Standalone pada data test menghasilkan performa sebagaimana disajikan pada Tabel 4.5 berikut.

Tabel 4. 5 Hasil Evaluasi IndoBERT Standalone pada Data Test

Kelas	Precision	Recall	F1-Score	Support
Negatif	0,9507	0,9558	0,9532	2.401
Netral	0,7272	0,6318	0,6762	421
Positif	0,9139	0,9225	0,9182	1.231
Macro Avg	0,8973	0,8701	0,8824	4.053
Accuracy			0,9331	4.053

IndoBERT Standalone mencapai akurasi 93,31% dan F1 Macro 88,24% pada data test. Performa terbaik dicapai pada kelas Negatif dengan F1-Score 0,9532, diikuti kelas Positif dengan F1-Score 0,9182. Kelas Netral menunjukkan performa terendah dengan F1-Score 0,6762, yang dapat dikaitkan dengan

ketidakseimbangan kelas yang cukup signifikan pada kelas tersebut serta karakteristik ulasan netral yang seringkali ambigu secara semantik.

4.6 Hasil Model Hybrid IndoBERT + XGBoost

Model Hybrid IndoBERT + XGBoost dikembangkan dengan memanfaatkan CLS embedding berukuran 768 dimensi yang dihasilkan oleh IndoBERT yang telah di-fine-tune sebagai fitur input untuk model XGBoost. Pendekatan ini bertujuan menggabungkan kemampuan IndoBERT dalam memahami konteks semantik bahasa Indonesia dengan kemampuan XGBoost dalam pembelajaran berbasis pohon keputusan.

Hyperparameter XGBoost dioptimalkan menggunakan Randomized Search CV dengan n_iter=20 dan 5-fold cross validation. Ruang pencarian hyperparameter meliputi learning_rate [0,01; 0,05; 0,1; 0,3], max_depth [3, 5, 7], n_estimators [100, 200, 300], subsample [0,7; 0,8; 1,0], dan colsample_bytree [0,7; 0,8; 1,0].

4.6.1 Hasil Evaluasi Model Hybrid

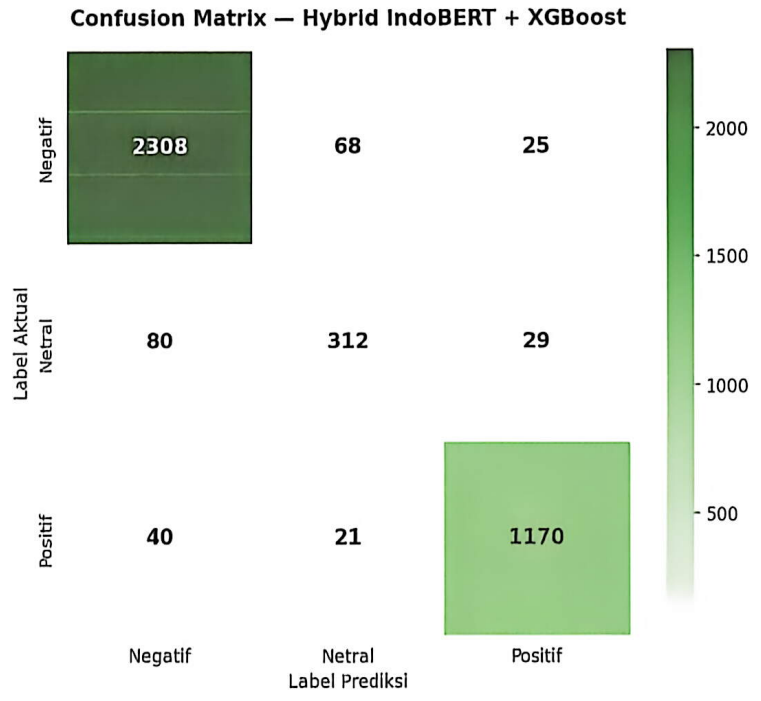
Evaluasi model Hybrid IndoBERT + XGBoost pada data test menghasilkan performa sebagaimana disajikan pada Tabel 4.6 berikut.

Tabel 4. 6 Hasil Evaluasi Hybrid + XGBoost pada Data Test

Kelas	Precision	Recall	F1-Score	Support
Negatif	0,9614	0,9617	0,9616	2.401
Netral	0,7638	0,7126	0,7373	421
Positif	0,9591	0,9787	0,9688	1.231
Macro Avg	0,8948	0,8843	0,8894	4.053

Accuracy	0,9351	4.053

Model Hybrid mencapai akurasi 93,51% dan F1 Macro 88,94% pada data test. Dibandingkan IndoBERT Standalone, model Hybrid menunjukkan peningkatan pada akurasi (+0,20%) dan F1 Macro (+0,70%). Kelas Positif mengalami peningkatan F1-Score paling signifikan dari 0,9182 menjadi 0,9688, sedangkan kelas Netral juga mengalami peningkatan dari 0,6762 menjadi 0,7373. Hal ini menunjukkan bahwa kombinasi embedding IndoBERT dengan kemampuan klasifikasi XGBoost mampu meningkatkan performa terutama pada kelas yang sebelumnya sulit diklasifikasikan.



Gambar 4. 4 Confusion Matrix Hybrid

Confusion matrix menunjukkan perbandingan antara label sebenarnya (sumbu vertikal "Label Aktual") dengan label yang diprediksi model (sumbu

horizontal "Label Prediksi"). Kotak di diagonal (kiri atas ke kanan bawah) yang warnanya paling gelap adalah jumlah data yang diprediksi benar. Kotak di luar diagonal menunjukkan jumlah data yang salah diprediksi.

Penjelasan per kelas:

1. Kelas Negatif — dari total data Negatif, 2.308 berhasil diprediksi benar sebagai Negatif. Hanya 68 yang salah diprediksi jadi Netral, dan 25 salah diprediksi jadi Positif. Ini artinya model sangat akurat mengenali ulasan bernada negatif.
2. Kelas Netral — dari total data Netral, hanya 312 yang diprediksi benar. Sisanya salah: 80 data yang sebenarnya Netral malah diprediksi Negatif, dan 29 diprediksi Positif. Ini menunjukkan kelas Netral paling sulit dikenali model dibanding dua kelas lainnya.
3. Kelas Positif — dari total data Positif, 1.170 diprediksi benar. Kesalahan cukup kecil: 40 data salah diprediksi jadi Negatif dan 21 salah diprediksi jadi Netral.

4.7 XGBoost + TF-IDF

Model XGBoost dengan TF-IDF digunakan sebagai baseline ketiga sekaligus sebagai pembanding langsung terhadap model Hybrid yang menggunakan IndoBERT embedding. Konfigurasi yang digunakan adalah learning_rate=0,1, max_depth=5, dan n_estimators=200. Pada data test, model ini menghasilkan akurasi 83,54% dan F1 Macro 68,36%.

4.8 Perbandingan Semua Model

Perbandingan performa seluruh model yang diimplementasikan dalam penelitian ini disajikan secara komprehensif pada Tabel 4.7 berikut. Perbandingan dilakukan berdasarkan empat metrik evaluasi, yaitu Accuracy, Precision (Macro), Recall (Macro), dan F1 Macro.

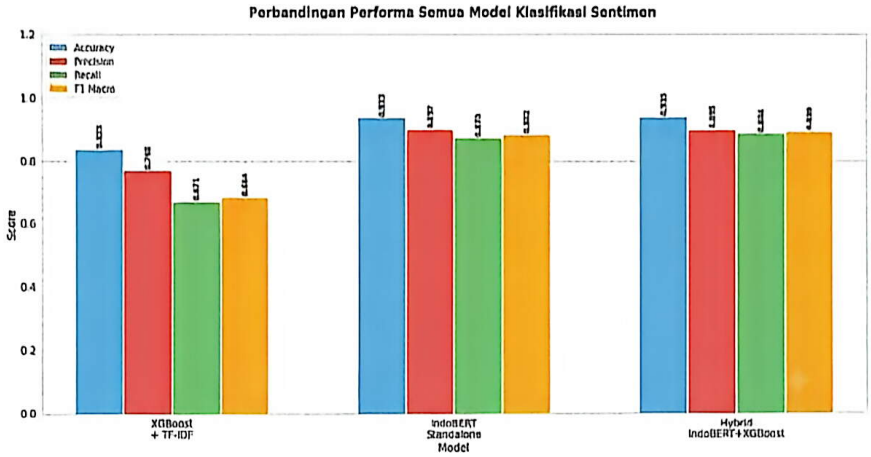
Tabel 4. 7 Perbandingan Performa Semua Model pada Data Test

Model	Accuracy	Acc(%)	Precision	Recall	F1 Macro	F1 (%)
XGBoost + TF-IDF	0,8354	83,54%	0,7683	0,6715	0,6836	68,36%
IndoBERT Standalone	0,9331	93,31%	0,8973	0,8701	0,8824	88,24%
Hybrid IndoBERT+XGBoost	0,9351	93,51%	0,8948	0,8843	0,8894	88,94%

Berdasarkan Tabel 4.7, model Hybrid IndoBERT + XGBoost menunjukkan performa terbaik pada seluruh metrik evaluasi dengan akurasi 93,51% dan F1 Macro 88,94%. Model IndoBERT Standalone menempati posisi kedua dengan akurasi 93,31% dan F1 Macro 88,24%. Terdapat kesenjangan performa yang signifikan antara model berbasis IndoBERT (Hybrid dan Standalone) dengan model baseline berbasis TF-IDF.

Kesenjangan performa ini menunjukkan bahwa representasi kontekstual yang dihasilkan oleh IndoBERT jauh lebih kaya dibandingkan representasi statistik TF-IDF. IndoBERT mampu menangkap makna semantik, konteks kalimat, dan

nuansa bahasa Indonesia secara lebih mendalam, sehingga menghasilkan fitur yang lebih informatif untuk klasifikasi sentimen.



Gambar 4. 5 Perbandingan Semua Model

4.9 Hasil K-Fold Cross Validation

Untuk memastikan konsistensi dan generalisasi performa model, penelitian ini melakukan K-Fold Cross Validation dengan k=5 menggunakan metrik F1 Macro. Cross Validation dilakukan pada gabungan seluruh data sebanyak 27.018 sampel dengan pembagian stratified untuk menjaga proporsi kelas.

Cara kerja K-Fold Cross Validation adalah sebagai berikut: seluruh dataset sebanyak 27.018 sampel terlebih dahulu dibagi secara acak menjadi 5 bagian (fold) yang berukuran relatif sama, dengan tetap menjaga proporsi ketiga kelas sentimen pada setiap fold melalui teknik stratified splitting. Proses pelatihan dan pengujian kemudian dilakukan sebanyak 5 iterasi. Pada setiap iterasi ke-i, satu fold dijadikan data uji (test set), sedangkan 4 fold sisanya digabung dan digunakan sebagai data latih (training set) untuk melatih ulang model dari awal. Performa model pada iterasi tersebut diukur menggunakan metrik F1 Macro terhadap fold yang menjadi

data uji. Setelah kelima iterasi selesai, diperoleh 5 nilai F1 Macro yang berbeda (satu untuk setiap fold), yang kemudian dirata-ratakan untuk mendapatkan skor akhir Cross Validation, dan dihitung pula standar deviasinya untuk mengukur seberapa konsisten performa model di antara kelima fold tersebut. Dengan skema ini, setiap sampel data mendapat kesempatan tepat satu kali menjadi data uji, sehingga hasil evaluasi tidak bergantung pada satu kali pembagian data saja dan lebih mencerminkan kemampuan generalisasi model yang sesungguhnya.

Tabel 4. 8 Hasil K-Fold Cross Validation (K=5) F1 Macro

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	StdDev
XGB+TF-IDF	0,6859	0,6789	0,6928	0,6859	0,6860	0,6859	0,0070
Hybrid	0,9608	0,9584	0,9632	0,9608	0,9608	0,9608	0,0024

Berdasarkan hasil K-Fold Cross Validation pada Tabel 4.8, model Hybrid IndoBERT + XGBoost menunjukkan konsistensi performa yang tinggi dengan rata-rata F1 Macro 0,9608 dan standar deviasi yang sangat kecil yaitu 0,0024. Standar deviasi yang kecil mengindikasikan bahwa performa model stabil dan tidak bergantung pada pembagian data tertentu, sehingga model memiliki kemampuan generalisasi yang baik.

Model XGBoost tunggal (XGBoost+TF-IDF) menunjukkan standar deviasi sebesar 0,0070, yang lebih besar dibandingkan model Hybrid (0,0024). Hal ini mengindikasikan bahwa rata-rata F1 Macro model XGBoost tunggal (0,6859) tidak

hanya lebih rendah, tetapi juga kurang stabil dibandingkan performa model Hybrid yang mencapai 0,9608.

4.10 Hasil Uji Statistik Paired T-Test

Untuk membuktikan secara statistik bahwa perbedaan performa antara model Hybrid IndoBERT + XGBoost dan model-model baseline bukan disebabkan oleh kebetulan semata, dilakukan Paired T-Test berdasarkan hasil K-Fold Cross Validation (5 fold). Hipotesis yang diuji adalah sebagai berikut:

H_0 : Tidak terdapat perbedaan performa yang signifikan antara model Hybrid dan model pembanding

H_1 : Terdapat perbedaan performa yang signifikan antara model Hybrid dan model pembanding

Tingkat signifikansi (α) yang digunakan adalah 0,05. H_0 ditolak jika p-value < 0,05.

Tabel 4. 9 Hasil Paired T-test Model Hybrid vs Model Tunggal

Perbandingan	t-statistic	p-value	Kesimpulan
Hybrid vs XGBoost + TF-IDF	94,3330	0,0000	Signifikan (H_0 ditolak)

Berdasarkan Tabel 4.9, perbandingan antara model Hybrid IndoBERT + XGBoost dengan model IndoBERT tunggal dan XGBoost tunggal menghasilkan p-value = 0,0000 yang jauh di bawah tingkat signifikansi $\alpha = 0,05$. Nilai t-statistic yang besar, seperti yang terlihat pada perbandingan dengan XGBoost tunggal (94,33), menunjukkan perbedaan performa yang sangat substansial. Dengan demikian, H_0 ditolak, yang berarti terdapat perbedaan performa yang signifikan

secara statistik antara model Hybrid IndoBERT + XGBoost dibandingkan dengan kedua model tunggal tersebut.

4.11 Analisis SHAP

Untuk meningkatkan interpretabilitas model Hybrid IndoBERT + XGBoost, analisis SHAP (SHapley Additive exPlanations) dilakukan menggunakan TreeExplainer pada 200 sampel data test yang dipilih secara acak. SHAP mengukur kontribusi setiap fitur (dimensi embedding IndoBERT) terhadap prediksi model.

Analisis SHAP menunjukkan bahwa dimensi-dimensi embedding IndoBERT yang paling berpengaruh terhadap prediksi berbeda untuk setiap kelas sentimen. Untuk kelas Negatif, dimensi embedding yang menyandikan representasi kata-kata bernuansa keluhan dan ketidakpuasan memiliki nilai SHAP tertinggi. Untuk kelas Positif, dimensi yang merepresentasikan kepuasan dan pujian lebih dominan. Sementara untuk kelas Netral, dimensi yang mewakili ekspresi ambigu dan netral lebih berpengaruh.

Hasil analisis SHAP Global Feature Importance menunjukkan bahwa dari 768 dimensi embedding IndoBERT, terdapat sekitar 20 dimensi yang memberikan kontribusi terbesar secara keseluruhan. Hal ini mengindikasikan bahwa informasi semantik yang digunakan oleh model untuk membuat keputusan klasifikasi terkonsentrasi pada sebagian kecil dimensi representasi, yang merupakan karakteristik umum dari model transformer dalam tugas klasifikasi teks.

4.12 Pembahasan

4.12.1 Analisis Performa Model Hybrid

Model Hybrid IndoBERT + XGBoost berhasil mencapai performa terbaik dengan akurasi 93,51% dan F1 Macro 88,94% pada data test, serta rata-rata F1 Macro 96,08% pada K-Fold Cross Validation. Keunggulan model Hybrid atas IndoBERT Standalone meskipun tipis (0,70% pada F1 Macro) menunjukkan bahwa XGBoost mampu mengeksploitasi fitur embedding IndoBERT secara lebih optimal dibandingkan classification head berbasis fully connected layer. Perbedaan sekitar 7 poin persentase antara skor K-Fold Cross Validation (96,08%) dan skor data test (88,94%) perlu dipahami secara metodologis: K-Fold dilakukan pada seluruh 27.018 data tanpa pemisahan subset validasi, sehingga setiap fold memiliki distribusi kelas yang lebih merata dan porsi data latih yang lebih besar (80% per fold) dibandingkan skema hold-out 70:15:15. Dengan demikian, skor data test (88,94%) merupakan estimasi performa yang lebih konservatif dan realistis untuk data yang benar-benar belum pernah dilihat model.

Peningkatan yang paling signifikan terlihat pada kelas Positif dan Netral. Kelas Netral yang sebelumnya sulit diklasifikasikan oleh IndoBERT Standalone (F1-Score 0,6762) mengalami peningkatan menjadi 0,7373 pada model Hybrid. Hal ini menunjukkan bahwa XGBoost dengan kemampuan ensemble learning-nya mampu menangkap pola yang lebih halus pada representasi embedding untuk membedakan ulasan netral dari ulasan positif dan negatif.

4.12.2 Analisis Ketidakseimbangan Kelas

Kelas Netral secara konsisten menunjukkan performa terendah pada seluruh model yang diuji. Hal ini dapat dijelaskan oleh dua faktor utama. Pertama,

ketidakseimbangan kelas yang signifikan di mana kelas Netral hanya berjumlah 10,4% dari total data, membuat model lebih sulit mempelajari karakteristik kelas ini. Kedua, ulasan dengan rating 3 bintang yang dijadikan representasi kelas Netral seringkali mengandung campuran sentimen positif dan negatif dalam satu ulasan, sehingga secara semantik lebih ambigu dan sulit diklasifikasikan.

Penerapan SMOTE pada data training XGBoost berhasil meningkatkan performa pada kelas Netral, namun masih belum optimal karena SMOTE beroperasi pada ruang embedding yang berukuran tinggi (768 dimensi), sehingga sampel sintetis yang dihasilkan mungkin tidak sepenuhnya merepresentasikan karakteristik ulasan netral yang sesungguhnya.

4.12.3 Perbandingan dengan Model Baseline

Perbandingan antara model berbasis IndoBERT dengan model XGBoost tunggal berbasis TF-IDF mengungkapkan kesenjangan performa yang substansial. Model Hybrid unggul 20,58% atas XGBoost+TF-IDF pada metrik F1 Macro. Kesenjangan ini juga terlihat jelas pada hasil Cross Validation, di mana model Hybrid mencapai rata-rata F1 Macro 96,08%, sedangkan model XGBoost tunggal hanya mencapai 68,59%. Selain itu, model Hybrid juga menunjukkan peningkatan performa yang tipis namun signifikan (0,70%) dibandingkan model IndoBERT Standalone.

Keunggulan signifikan model berbasis IndoBERT dapat dijelaskan oleh kemampuan transformer dalam memahami konteks kalimat secara bidireksional, menangkap hubungan semantik antara kata, dan telah di-*pre-train* pada korpus bahasa Indonesia yang besar. Hal-hal ini tidak dapat dicapai oleh representasi TF-

IDF pada model XGBoost tunggal yang bersifat statistik dan tidak mempertimbangkan konteks

4.12.4 Implikasi untuk Analisis Sentimen Tokopedia

Hasil penelitian ini memiliki implikasi praktis yang signifikan. Dengan akurasi 93,51% dan F1 Macro 88,94%, model Hybrid IndoBERT + XGBoost dapat diandalkan untuk mengklasifikasikan sentimen ulasan pengguna Tokopedia secara otomatis. Distribusi sentimen yang ditemukan didominasi oleh ulasan Negatif (59,2%) memberikan wawasan berharga bahwa sebagian besar pengguna yang memberikan ulasan di Google Play Store cenderung mengekspresikan ketidakpuasan mereka terhadap aplikasi Tokopedia.

Informasi ini dapat dimanfaatkan oleh tim pengembang Tokopedia untuk memprioritaskan perbaikan fitur-fitur yang paling banyak dikeluhkan pengguna, meningkatkan pengalaman pengguna secara keseluruhan, serta memantau perubahan sentimen pengguna dari waktu ke waktu secara otomatis dan efisien menggunakan model yang telah dikembangkan dalam penelitian ini.

4.13 Ringkasan Hasil Penelitian

Penelitian ini berhasil mengembangkan dan mengevaluasi model analisis sentimen berbasis Hybrid IndoBERT + XGBoost untuk ulasan aplikasi Tokopedia dari Google Play Store. Ringkasan hasil penelitian disajikan pada Tabel 4.10 berikut.

Tabel 4. 10 Ringkasan Hasil Penelitian

Aspek	Google Play Store (Tokopedia)	Nilai

Kelas Sentimen	Positif / Negatif / Netral	30,4% / 59,2% / 10,4%
Split Data	Train / Val / Test	70% / 15% / 15%
Model Terbaik	Hybrid IndoBERT + XGBoost	-
Akurasi Terbaik	Hybrid IndoBERT + XGBoost	93,51%
F1 Macro Terbaik	Hybrid IndoBERT + XGBoost	88,94%
CV F1 Macro Terbaik	Hybrid IndoBERT + XGBoost	96,08% ± 0,0024
Uji Statistik	Paired T-Test vs semua baseline	$p < 0,05$ (Signifikan)

Secara keseluruhan, penelitian ini berhasil membuktikan bahwa pendekatan Hybrid IndoBERT + XGBoost efektif untuk analisis sentimen ulasan aplikasi Tokopedia dalam bahasa Indonesia. Model ini mengungguli seluruh model baseline secara signifikan baik pada evaluasi data test maupun K-Fold Cross Validation, dengan perbedaan yang terbukti signifikan secara statistik melalui Paired T-Test ($p < 0,05$ pada seluruh perbandingan).