

Analisis Sentimen Komentar TikTok pada Konten Sound Horeg Menggunakan XLM-RoBERTa, Kamus Bahasa Jawa dan CatBoost

Sentiment Analysis of TikTok Comments on Sound Horeg Content Using XLM-RoBERTa, a Javanese Lexicon, and CatBoost

Fiona Evelyne Luther^{1,*}, Radius Tanone²

¹ Informatics Engineering, Faculty of Information Technology, Universitas Kristen Satya Wacana, Salatiga, Indonesia

² Artificial Intelligence Research Center, Universitas Kristen Satya Wacana, Salatiga, Indonesia

Email: ^{1,*}672023195@student.uksw.edu, ²radius.tanone@uksw.edu

Email Penulis Korespondensi: 672023195@student.uksw.edu

Received: 31/08/2026, Revised: 19/09/2026, Accepted: 21/09/2026, Published: 22/09/2026

Abstrak—Komentar pada media sosial TikTok dapat menjadi sumber data untuk menganalisis sentimen masyarakat terhadap fenomena sosial dan budaya, tetapi karakteristik bahasa yang informal, multibahasa, *code-mixing*, serta keterbatasan sumber daya bahasa daerah menjadi tantangan dalam proses klasifikasi sentimen. Sound Horeg dipilih sebagai objek penelitian karena fenomena sosial dan budaya ini memunculkan beragam respons publik di media sosial, baik dalam bentuk dukungan maupun penolakan, sehingga menyediakan konteks yang relevan untuk mengkaji persepsi masyarakat melalui analisis sentimen. Penelitian ini bertujuan mengembangkan pipeline analisis sentimen hibrida untuk mengklasifikasikan komentar TikTok pada konten Sound Horeg dengan mengintegrasikan XLM-RoBERTa, kamus sentimen bahasa Jawa, dan CatBoost. Kebaruan penelitian ini terletak pada pengintegrasian representasi semantik multibahasa XLM-RoBERTa dengan sumber leksikal bahasa Jawa dalam konteks analisis sentimen Sound Horeg, serta evaluasi kontribusi setiap komponen melalui *ablation study*. Data penelitian diperoleh dari lima video Sound Horeg dan menghasilkan 7.330 komentar yang kemudian melalui tahap pembersihan, *preprocessing*, pelabelan berbasis *Lexicon-Based*, ekstraksi *embedding* XLM-RoBERTa, serta penyeimbangan kelas menggunakan *Random Undersampling*. Hasil pelabelan menghasilkan 960 komentar yang terdiri atas 559 positif dan 401 negatif, kemudian diperoleh 802 data seimbang untuk pelatihan model utama. Representasi *embedding* berdimensi 768 digunakan sebagai fitur CatBoost. Model utama menghasilkan *accuracy* sebesar 89,44%, F1-Macro sebesar 89,44%, dan ROC-AUC sebesar 90,25%. *Ablation study* menunjukkan bahwa konfigurasi *Lexicon-only* menghasilkan *accuracy* 94,79%, XLM-R only sebesar 93,23%, dan XLM-R + Lexicon sebesar 91,15%. Hasil ini menunjukkan bahwa konfigurasi dengan komponen lebih sederhana, khususnya Lexicon-only dan XLM-R only, menghasilkan performa yang lebih tinggi dibandingkan pipeline utama, sehingga pipeline hibrida belum dapat dianggap sebagai konfigurasi paling optimal untuk implementasi operasional. Hasil tersebut menunjukkan bahwa XLM-RoBERTa mampu memberikan representasi semantik yang baik tanpa memasukkan skor *lexicon* sebagai fitur utama, sedangkan kontribusi kosakata bahasa Jawa masih terbatas karena hanya ditemukan pada 1,6% komentar Kelompok A dan 0,8% komentar Kelompok B. Temuan ini memberikan implikasi bahwa pemilihan komponen dalam sistem analisis sentimen perlu mempertimbangkan keseimbangan antara performa, kompleksitas pemrosesan, dan karakteristik bahasa pada data yang dianalisis.

Kata Kunci: Analisis Sentimen; TikTok; XLM-RoBERTa; CatBoost; Bahasa Jawa

Abstract—Comments on TikTok can provide valuable data for analyzing public sentiment toward social and cultural phenomena. However, informal language, multilingual expressions, code-mixing, and limited linguistic resources for regional languages present challenges for sentiment classification. Sound Horeg was selected as the research subject because this social and cultural phenomenon generates diverse public responses on social media, including support and opposition, making it a relevant context for examining public perceptions through sentiment analysis. This study aims to develop a hybrid sentiment analysis pipeline for classifying TikTok comments related to Sound Horeg by integrating XLM-RoBERTa, a Javanese sentiment lexicon, and CatBoost. The methodological novelty of this study lies in integrating the multilingual semantic representations of XLM-RoBERTa with a Javanese lexical resource in the context of Sound Horeg sentiment analysis, while evaluating the contribution of each component through an ablation study. The dataset was collected from five Sound Horeg videos and consisted of 7,330 comments, which were processed through data cleaning, preprocessing, Lexicon-Based sentiment labeling, XLM-RoBERTa embedding extraction, and Random Undersampling. Lexicon-based labeling produced 960 labeled comments, consisting of 559 positive and 401 negative comments, while Random Undersampling produced 802 balanced samples for the main model. The 768-dimensional XLM-RoBERTa embeddings were used as input features for CatBoost. The main model achieved an accuracy of 89.44%, an F1-Macro score of 89.44%, and an ROC-AUC of 90.25%. The ablation study showed that the Lexicon-only configuration achieved an accuracy of 94.79%, followed by XLM-R only at 93.23% and XLM-R + Lexicon at 91.15%. These results indicate that the simpler configurations, particularly Lexicon-only and XLM-R only, achieved higher performance than the main pipeline, suggesting that the hybrid pipeline cannot yet be considered the most optimal configuration for operational implementation. These findings indicate that XLM-RoBERTa can provide effective semantic representations without using skor leksikals as the main features, while the contribution of Javanese vocabulary remains limited because it was detected in only 1.6% of Group A comments and 0.8% of Group B comments. This finding implies that the selection of components in a sentiment analysis system should consider the balance between predictive performance, processing complexity, and the linguistic characteristics of the analyzed data.

Keywords: Sentiment Analysis; TikTok; XLM-RoBERTa; CatBoost; Javanese Language

1. PENDAHULUAN

TikTok merupakan salah satu platform media sosial yang banyak digunakan di Indonesia dan menjadi ruang bagi masyarakat untuk mengekspresikan opini terhadap berbagai fenomena yang sedang berkembang [1][2][3][4]. Salah

satu fenomena yang menarik perhatian publik adalah *Sound Horeg*, yaitu hiburan berbasis sistem tata suara berdaya tinggi yang umum digunakan dalam berbagai kegiatan masyarakat di Pulau Jawa. Fenomena tersebut memunculkan beragam tanggapan dari pengguna TikTok yang tercermin melalui kolom komentar sehingga komentar pada platform tersebut dapat dimanfaatkan sebagai sumber data untuk menganalisis sentimen masyarakat terhadap fenomena sosial dan budaya lokal [1][3][4]. Analisis terhadap komentar tersebut penting dilakukan karena opini yang muncul pada media sosial dapat memberikan gambaran mengenai kecenderungan respons masyarakat terhadap suatu fenomena. Namun, karakteristik komentar TikTok yang bersifat spontan dan tidak terstruktur menimbulkan tantangan dalam proses analisis sentimen. Pengguna dapat menggunakan bahasa informal, slang, singkatan, serta mencampurkan beberapa bahasa dalam satu komentar sehingga struktur teks menjadi lebih kompleks dibandingkan dengan teks formal. Kondisi tersebut semakin relevan ketika data mengandung bahasa Indonesia yang bercampur dengan bahasa Jawa dan bahasa Inggris sehingga metode analisis perlu mempertimbangkan karakteristik multibahasa sekaligus konteks lokal yang terdapat dalam data.

Karakteristik bahasa tersebut menempatkan analisis komentar TikTok pada konteks *low-resource language*, yaitu kondisi ketika ketersediaan sumber daya linguistik, korpus, dan data berlabel untuk suatu bahasa atau dialek masih terbatas [5][6][7]. Penelitian mengenai analisis sentimen pada bahasa dengan sumber daya rendah menunjukkan bahwa keterbatasan dataset berlabel dan sumber daya linguistik menjadi salah satu hambatan dalam menghasilkan model yang memiliki kemampuan generalisasi yang baik [5][6]. Permasalahan tersebut juga ditemukan pada bahasa daerah di Indonesia yang memiliki keragaman kosakata dan penggunaan dalam komunikasi sehari-hari, tetapi belum memperoleh dukungan sumber daya NLP sebesar bahasa dengan sumber daya tinggi [8][9][10]. Ketersediaan sumber daya berupa dataset multibahasa juga menjadi faktor penting dalam pengembangan model analisis sentimen bahasa daerah [9][10]. Kondisi ini menjadi semakin kompleks pada data media sosial karena penggunaan bahasa daerah dapat muncul dalam bentuk kata tidak baku, variasi penulisan, maupun *code-mixing* dengan bahasa lain. Oleh karena itu, pendekatan analisis sentimen untuk data TikTok yang mengandung bahasa Indonesia dan bahasa Jawa perlu mempertimbangkan tidak hanya representasi semantik secara umum, tetapi juga pengetahuan linguistik lokal yang dapat membantu menangkap ekspresi yang khas dari bahasa daerah. Dalam penelitian ini, istilah *low-resource language* digunakan untuk merujuk pada kondisi keterbatasan sumber daya linguistik dan data berlabel yang berkaitan dengan bahasa daerah dalam komentar. Dengan karakteristik tersebut, metode yang digunakan perlu mampu menangani data yang didominasi bahasa Indonesia, tetapi tetap mengandung kosakata lokal, variasi ejaan, dan ekspresi bahasa Jawa.

Salah satu pendekatan yang banyak digunakan untuk menangani analisis sentimen pada data media sosial adalah *Lexicon-Based*. Pendekatan ini menentukan polaritas sentimen berdasarkan keberadaan kata atau ekspresi yang telah diberikan bobot sentimen dalam suatu kamus sehingga tidak membutuhkan data latih dalam jumlah besar [1][2][11][12][13]. Pendekatan berbasis *lexicon* juga telah diterapkan pada berbagai jenis data media sosial untuk mengidentifikasi kecenderungan polaritas berdasarkan karakteristik kosakata yang muncul dalam teks [11][12]. Penggunaan pendekatan tersebut juga relevan pada kondisi *low-resource* karena kamus sentimen dapat digunakan sebagai sumber pengetahuan ketika ketersediaan data berlabel terbatas. Penelitian sebelumnya telah menerapkan pendekatan berbasis *lexicon* pada komentar TikTok dan platform media sosial lainnya untuk memperoleh distribusi sentimen berdasarkan kosakata yang muncul dalam teks [1][2][11][12][13]. Dalam penelitian ini, pendekatan berbasis *lexicon* digunakan untuk menghasilkan label awal atau *pseudo-label* yang diperlakukan sebagai *silver standard*. Pendekatan ini dipilih karena dataset berlabel melalui anotasi manusia belum tersedia, sedangkan proses anotasi manual memerlukan pedoman pelabelan yang konsisten, keterlibatan anotator, serta pengujian kesepakatan antarannotator. Pada skenario penelitian ini, pendekatan berbasis leksikon digunakan sebagai alternatif awal yang transparan dan konsisten untuk membentuk *dataset* berlabel dari komentar dalam jumlah besar. Penggunaan label tersebut tidak dimaksudkan untuk menggantikan *ground truth* yang diperoleh melalui anotasi manusia, tetapi sebagai acuan supervisi awal untuk mengevaluasi kemampuan model dalam mempelajari pola sentimen berdasarkan aturan leksikal. Namun, label tersebut tidak dapat dianggap sebagai *gold standard* karena belum melalui validasi anotator manusia dan masih berpotensi dipengaruhi oleh keterbatasan kosakata, slang, negasi, sarkasme, serta konteks komentar. Dengan demikian, model yang dilatih menggunakan label tersebut dipahami sebagai model yang mempelajari pola pelabelan berbasis leksikon, sehingga hasilnya perlu diinterpretasikan dengan mempertimbangkan keterbatasan label yang digunakan.

Perkembangan model *transformer* memberikan pendekatan alternatif untuk menghasilkan representasi semantik yang lebih kontekstual pada data multibahasa. Salah satu model yang relevan adalah XLM-RoBERTa yang dikembangkan melalui pelatihan pada lebih dari 100 bahasa dan dirancang untuk menghasilkan representasi lintas bahasa [14]. Kemampuan tersebut menjadikan XLM-RoBERTa relevan untuk data yang memiliki karakteristik multibahasa, termasuk data yang mengandung bahasa Indonesia dan bahasa daerah. Penelitian sebelumnya menunjukkan bahwa XLM-RoBERTa dapat digunakan dalam analisis sentimen pada data media sosial multibahasa dan menghasilkan representasi yang mampu menangkap informasi semantik teks [15]. Penerapan XLM-RoBERTa pada data berbahasa Indonesia juga menunjukkan potensinya dalam memahami karakteristik teks yang digunakan dalam analisis sentimen [16][17]. Meskipun demikian, representasi berbasis *transformer* tidak secara khusus memasukkan pengetahuan mengenai kosakata lokal tertentu sehingga ekspresi bahasa daerah yang memiliki variasi penulisan atau makna kontekstual tertentu masih dapat menjadi tantangan.

Oleh karena itu, kamus sentimen bahasa Jawa dalam penelitian ini tidak dimaksudkan untuk menggantikan kemampuan multilingual XLM-RoBERTa, tetapi untuk menguji kontribusi pengetahuan linguistik lokal terhadap klasifikasi sentimen. Keterbatasan cakupan kosakata bahasa Jawa dalam *dataset* menjadi bagian dari aspek yang perlu dievaluasi, bukan dasar untuk mengasumsikan bahwa fitur bahasa daerah akan selalu meningkatkan performa model. Dengan demikian, penggunaan kamus bahasa Jawa diposisikan sebagai eksperimen untuk mengukur apakah sumber daya linguistik lokal dapat memberikan manfaat yang terukur pada data yang secara dominan menggunakan bahasa Indonesia, meskipun cakupannya relatif terbatas.

Sejumlah penelitian telah membahas analisis sentimen pada bahasa campuran maupun bahasa daerah menggunakan pendekatan *transformer*. Hidayatullah [18] membahas analisis sentimen pada teks campuran bahasa Indonesia, Jawa, dan Inggris menggunakan model *transformer*, sedangkan penelitian Purnomo dan Sutopo [19] membandingkan beberapa model *pre-trained transformer* untuk analisis sentimen pada teks bahasa daerah di Indonesia. Di sisi lain, Winata et al. [10] mengembangkan NusaX sebagai dataset sentimen multibahasa yang mencakup sepuluh bahasa lokal Indonesia dan menunjukkan pentingnya ketersediaan sumber daya untuk pengembangan analisis sentimen bahasa daerah. Penelitian Koto et al. [13] juga menunjukkan bahwa *multilingual sentiment lexicon* dapat dimanfaatkan untuk analisis sentimen pada bahasa dengan sumber daya rendah. Meskipun penelitian-penelitian tersebut menunjukkan perkembangan pendekatan *transformer* dan sumber daya leksikal untuk bahasa daerah, integrasi representasi semantik XLM-RoBERTa dengan kamus sentimen bahasa Jawa pada komentar TikTok yang memiliki karakteristik *code-mixing* masih relatif terbatas.

Selain itu, penggunaan sumber informasi yang sama sebagai dasar pembentukan label dan fitur dapat menimbulkan risiko *data leakage* apabila skor *lexicon* secara langsung dimasukkan sebagai fitur model. Dalam kondisi tersebut, model berpotensi memperoleh informasi yang sangat dekat dengan aturan pembentukan label sehingga hasil evaluasi tidak sepenuhnya mencerminkan kemampuan model dalam mempelajari pola dari representasi teks. Oleh karena itu, skor *lexicon* tidak digunakan sebagai fitur pada model utama, meskipun aturan leksikon tetap digunakan untuk membentuk *pseudo-label*. Karena label penelitian diperoleh melalui aturan leksikon, performa model dalam penelitian ini perlu dipahami sebagai kemampuan mempelajari pola dari *silver standard*, bukan sebagai pengukuran langsung terhadap kesesuaian model dengan anotasi manusia. Pemisahan skor leksikon dari fitur model dilakukan untuk mengurangi ketergantungan langsung antara sumber label dan fitur, sementara keterbatasan yang berasal dari penggunaan label berbasis leksikon tetap dipertimbangkan dalam interpretasi hasil.

Berdasarkan kesenjangan tersebut, penelitian ini mengembangkan *pipeline* analisis sentimen hibrida yang mengintegrasikan XLM-RoBERTa Base sebagai pengekstrak representasi semantik multibahasa, kamus sentimen bahasa Jawa sebagai sumber pengetahuan linguistik lokal, dan CatBoost sebagai algoritma klasifikasi. XLM-RoBERTa digunakan untuk menghasilkan representasi embedding dari komentar, sedangkan kamus sentimen digunakan pada tahap pelabelan *pseudo-label* (*silver standard*) dan sebagai komponen pembanding dalam *ablation study*. Rancangan penelitian ini memungkinkan kontribusi masing-masing komponen dievaluasi secara empiris tanpa mengasumsikan bahwa integrasi representasi kontekstual dan informasi leksikal selalu menghasilkan peningkatan performa. CatBoost dipilih karena menggunakan mekanisme *ordered boosting* yang dirancang untuk mengurangi *prediction shift* pada proses pembelajaran [20]. Dalam model utama, skor leksikal tidak digunakan sebagai fitur pelatihan karena informasi tersebut telah digunakan dalam pembentukan label. Pemisahan antara sumber label dan fitur dilakukan untuk mengurangi risiko *data leakage* serta menghasilkan evaluasi yang lebih terkontrol. Namun, karena label yang digunakan tetap berasal dari aturan leksikon, hasil evaluasi dipahami sebagai pengukuran terhadap kemampuan model mempelajari *silver standard*, bukan sebagai validasi langsung terhadap sentimen yang telah diverifikasi oleh anotator manusia. Sementara itu, kontribusi informasi leksikal diuji secara terpisah melalui konfigurasi *Lexicon-only*, *XLM-R-only*, dan *XLM-R + Lexicon (Fused)* sehingga pengaruh masing-masing komponen terhadap performa klasifikasi dapat diamati.

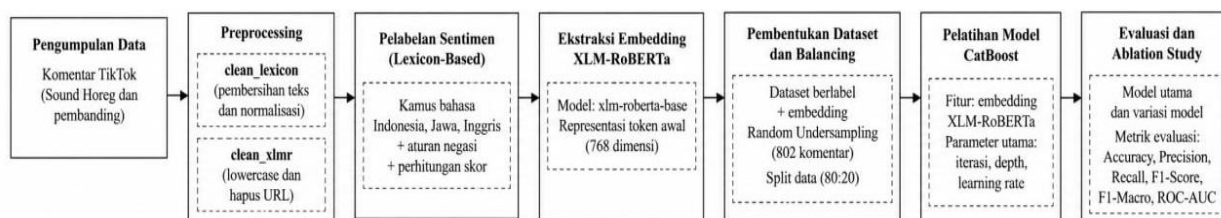
Penelitian ini bertujuan mengembangkan model analisis sentimen hibrida berbasis XLM-RoBERTa dan CatBoost dengan dukungan kamus sentimen bahasa Jawa untuk mengklasifikasikan komentar TikTok pada konten *Sound Horeg*. Selain mengevaluasi performa model utama, penelitian ini juga menguji kontribusi kamus sentimen bahasa Jawa melalui *ablation study* dan membandingkan distribusi sentimen pada konten *Sound Horeg* dan *sholawatan*. Penelitian ini memberikan tiga kontribusi utama. Pertama, penelitian menghasilkan sumber daya leksikal sentimen bahasa Jawa yang terdokumentasi untuk mendukung eksperimen analisis sentimen pada bahasa daerah. Kedua, penelitian mengusulkan *pipeline* hibrida yang memisahkan proses pembentukan label berbasis *lexicon* dari ekstraksi fitur XLM-RoBERTa untuk mengurangi risiko *data leakage*. *Pipeline* tersebut dievaluasi sebagai pendekatan eksperimental, sehingga kontribusinya tidak hanya ditinjau berdasarkan performa model utama, tetapi juga melalui perbandingan dengan konfigurasi yang lebih sederhana.

Ketiga, penelitian melakukan evaluasi melalui *ablation study* dan analisis distribusi sentimen lintas kategori konten untuk mengidentifikasi kontribusi representasi semantik dan informasi leksikal terhadap hasil klasifikasi. Evaluasi tersebut juga digunakan untuk mengetahui apakah penambahan komponen leksikal memberikan manfaat yang terukur dibandingkan dengan konfigurasi berbasis representasi XLM-RoBERTa saja, sehingga kompleksitas *pipeline* dapat dipertimbangkan bersama dengan hasil performa. Dengan pendekatan tersebut, penelitian diharapkan dapat memberikan alternatif metodologis untuk analisis sentimen pada data media sosial yang memiliki karakteristik multibahasa dan keterbatasan sumber daya linguistik bahasa daerah.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan eksperimental untuk mengembangkan *pipeline* analisis sentimen pada komentar TikTok bertema *Sound Horeg* dengan mengintegrasikan *Lexicon-Based*, XLM-RoBERTa, dan CatBoost. Tahapan penelitian terdiri atas pengumpulan data, *preprocessing*, pelabelan sentimen berbasis *lexicon*, ekstraksi representasi semantik menggunakan XLM-RoBERTa, penyeimbangan kelas, pelatihan model CatBoost, serta evaluasi melalui model utama dan *ablation study*. Rangkaian tahapan dan hubungan antarproses dalam penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Pipeline Analisis Sentimen

Gambar 1 memperlihatkan alur penelitian yang dimulai dari tahap pengumpulan komentar TikTok hingga evaluasi performa model. Tahap pengumpulan data menghasilkan komentar dari lima video TikTok bertema *Sound Horeg* sebagai kelompok utama dan satu video bertema sholawatan sebagai data pembanding. Data yang diperoleh selanjutnya melalui tahap pembersihan untuk menghilangkan data kosong dan teks yang tidak dapat digunakan. Setelah *preprocessing*, data diproses menggunakan dua jalur yang disesuaikan dengan kebutuhan pelabelan dan ekstraksi representasi. Jalur pertama digunakan untuk menghasilkan skor sentimen berbasis *lexicon*, sedangkan jalur kedua mempertahankan informasi tekstual yang diperlukan oleh XLM-RoBERTa. Label hasil *lexicon* kemudian dipasang dengan *embedding* XLM-RoBERTa pada data yang sama. *Dataset* berlabel tersebut selanjutnya digunakan dalam dua skenario pengujian. Skenario pertama merupakan model utama yang menggunakan data seimbang melalui *Random Undersampling* untuk mengurangi pengaruh ketidakseimbangan kelas. Skenario kedua merupakan *ablation study* yang menggunakan data sebelum penyeimbangan untuk mempertahankan kondisi label hasil *lexicon scoring* dan mengevaluasi kontribusi masing-masing komponen. Karena kedua skenario menggunakan jumlah dan distribusi data yang berbeda, hasil performanya digunakan untuk mengevaluasi karakteristik masing-masing konfigurasi, bukan sebagai perbandingan langsung dalam kondisi eksperimental yang sepenuhnya identik.

XLM-RoBERTa menggunakan arsitektur *transformer* berbasis mekanisme *multi-head self-attention* dengan 12 lapisan *encoder* dan dimensi *hidden state* sebesar 768, yang dilatih menggunakan teknik *Masked Language Modeling* (MLM) pada 2,5 TB data dari lebih dari 100 bahasa [14]. Kemampuan lintas bahasa diperoleh melalui representasi bersama pada ruang laten yang memungkinkan model memproses teks multibahasa, termasuk bahasa dengan sumber daya terbatas, tanpa memerlukan pelatihan ulang per bahasa. CatBoost menggunakan mekanisme *ordered boosting* yang membangun pohon keputusan secara berurutan dengan permutasi acak pada setiap iterasi untuk menghasilkan estimasi gradien yang tidak bias [20]. Mekanisme ini mengurangi *prediction shift* yang umum terjadi pada implementasi *gradient boosting* standar, sehingga lebih stabil pada dataset dengan distribusi kelas yang tidak seimbang dan kualitas label yang tidak sempurna.

2.2 Pengumpulan dan Preprocessing Data

Data penelitian berupa komentar publik pada platform TikTok yang dikumpulkan menggunakan layanan Apify TikTok Scraper. *Dataset* utama terdiri atas lima video bertema *Sound Horeg*, sedangkan satu video bertema sholawatan digunakan sebagai data pembanding. Proses pengumpulan menghasilkan 7.330 komentar dengan jumlah aktual 1.466 komentar per video pada kelompok utama. Data yang dikumpulkan berupa komentar utama dari setiap video, tanpa mengambil balasan komentar. Parameter pengumpulan data yang digunakan dalam penelitian disajikan pada Tabel 1.

Tabel 1. Parameter Pengumpulan Data

Parameter	Nilai	Keterangan
<i>commentsPerPost</i>	500	Batas komentar per video
<i>maxRepliesPerComment</i>	0	Balasan komentar tidak diambil
<i>scrapeComments</i>	<i>True</i>	Komentar diambil
Jumlah video kelompok A	5	Video <i>Sound Horeg</i>
Jumlah video kelompok B	1	Video sholawatan @inaarbha
Komentar per video (aktual)	1.466	Jumlah komentar utama yang berhasil dikumpulkan
Total komentar kotor	7.330	Sebelum <i>preprocessing</i>

Berdasarkan Tabel 1, proses pengumpulan data dilakukan dengan mengaktifkan parameter pengambilan komentar melalui Apify TikTok Scraper. Parameter *commentsPerPost* ditetapkan sebesar 500, sedangkan parameter *scrapeComments* diaktifkan untuk memungkinkan proses pengambilan komentar. Dataset utama terdiri atas lima video bertema *Sound Horeg*, sementara satu video bertema sholawatan digunakan sebagai kelompok pembanding. Hasil pengumpulan memperoleh 1.466 komentar aktual dari setiap video pada kelompok utama sehingga total komentar kotor yang terkumpul mencapai 7.330 komentar.

Setelah proses pengumpulan, data kosong dihapus menggunakan metode *drop NA* sehingga diperoleh 7.260 komentar. Tahap selanjutnya dilakukan melalui dua jalur *preprocessing* sesuai dengan kebutuhan metode. Jalur pertama menerapkan pembersihan teks secara lebih agresif dengan menghapus URL, *mention*, *hashtag*, emoji, angka, dan karakter nonalfabet. Proses tersebut dilanjutkan dengan normalisasi kata tidak baku menggunakan kamus yang terdiri atas 66 entri, seperti perubahan “gk”, “ga”, dan “nggak” menjadi “tidak”, “bgt” menjadi “banget”, serta “wes” dan “udh” menjadi “sudah”. Jalur kedua hanya melakukan perubahan huruf menjadi huruf kecil dan penghapusan URL untuk mempertahankan informasi tekstual yang masih relevan bagi model XLM-RoBERTa. Penggunaan *preprocessing* yang berbeda bertujuan mempertahankan karakteristik teks pada tahap ekstraksi representasi, sedangkan pembersihan yang lebih agresif diterapkan pada tahap pencocokan kosakata.

Komentar yang menghasilkan teks kosong setelah proses pembersihan pada jalur pertama dihapus sehingga jumlah data menjadi 6.060 komentar. Selanjutnya, data tersebut digunakan secara bersamaan sebagai dasar pelabelan sentimen dan ekstraksi representasi semantik. Rincian perubahan jumlah data pada setiap tahap pemrosesan disajikan pada Tabel 2.

Tabel 2. Rincian *Dataset* pada Setiap Tahap Pemrosesan

Tahap	Jumlah Data	Keterangan
Hasil <i>scraping</i>	7.330	5 video × 1.466 komentar
Setelah <i>drop NA</i>	7.260	Data kosong dihapus
Setelah <i>cleaning</i>	6.060	Baris kosong setelah <i>jalur pertama</i> dihapus
Setelah <i>lexicon scoring</i>	960	559 positif dan 401 negatif
Setelah <i>Random Undersampling</i>	802	401 positif dan 401 negatif
Data latih	641	80% dari data seimbang
Data uji	161	20% dari data seimbang

Berdasarkan Tabel 2, jumlah data berkurang dari 7.330 komentar hasil *scraping* menjadi 7.260 komentar setelah penghapusan data kosong. Proses *cleaning* pada jalur pertama kemudian menghasilkan 6.060 komentar. Tahap *lexicon scoring* menghasilkan 960 komentar berlabel yang terdiri atas 559 komentar positif dan 401 komentar negatif. Setelah dilakukan *random undersampling*, diperoleh 802 data dengan distribusi kelas seimbang, yaitu 401 data positif dan 401 data negatif. Dataset tersebut selanjutnya dibagi menjadi 641 data latih dan 161 data uji. Penurunan jumlah data dari 6.060 menjadi 960 komentar pada tahap *lexicon scoring* terjadi karena hanya komentar dengan skor sentimen positif atau negatif yang dipertahankan, sedangkan komentar dengan skor nol dikategorikan sebagai netral dan tidak digunakan dalam pelatihan model.

2.3 Pelabelan Sentimen Berbasis *Lexicon*

Pelabelan sentimen dilakukan menggunakan pendekatan *Lexicon-Based* dengan memanfaatkan kamus yang mencakup kosakata bahasa Indonesia, bahasa Jawa, dan bahasa Inggris. Kamus tersebut juga dilengkapi dengan daftar kata negasi dan kamus normalisasi *slang*. Komposisi kamus sentimen terdiri atas 70 entri positif dan 66 entri negatif, sedangkan kosakata bahasa Jawa digunakan secara khusus untuk menangkap ekspresi lokal yang muncul pada komentar. Skor sentimen setiap komentar dihitung berdasarkan selisih antara bobot kata positif dan bobot kata negatif yang dinormalisasi terhadap jumlah kata dalam komentar. Perhitungan skor sentimen dinyatakan pada Persamaan (1).

$$\text{Skor Sentimen} = \frac{\sum \text{Bobot Kata Positif} - \sum \text{Bobot Kata Negatif}}{\text{Jumlah Total Kata}} \quad (1)$$

Komentar dengan skor lebih besar dari nol diberi label positif, sedangkan komentar dengan skor kurang dari nol diberi label negatif. Komentar dengan skor sama dengan nol dikategorikan sebagai netral dan tidak digunakan dalam proses pelatihan model. Karena label diperoleh secara otomatis berdasarkan aturan *lexicon* tanpa anotasi manual, label yang dihasilkan diperlakukan sebagai *pseudo-label* atau *silver standard*. Setelah proses tersebut, diperoleh 960 komentar berlabel yang terdiri atas 559 komentar positif dan 401 komentar negatif. Untuk mengevaluasi ketahanan model terhadap ketidaksempurnaan label, sebanyak 5% label dibalik secara acak menggunakan *seed* tetap. Proses ini dilakukan untuk mensimulasikan kemungkinan kesalahan pelabelan pada *silver standard*. Label yang telah diberi *noise* digunakan secara konsisten dalam skenario pengujian yang ditetapkan, sehingga hasil evaluasi merepresentasikan kemampuan model dalam mempelajari label yang mengandung tingkat ketidaksempurnaan tertentu.

Penggunaan *pseudo-label* berbasis *lexicon* dalam penelitian ini merupakan konsekuensi dari tidak tersedianya anotasi manual oleh anotator manusia untuk dataset komentar berbahasa campuran yang digunakan. Pendekatan *silver standard* melalui aturan linguistik merupakan strategi yang lazim diterapkan pada skenario *low-resource* ketika

pengumpulan *gold standard* tidak memungkinkan secara praktis karena keterbatasan sumber daya dan waktu [5][6]. Label yang dihasilkan merepresentasikan pola aturan kamus, bukan persepsi sentimen manusia secara langsung, sehingga seluruh evaluasi dalam penelitian ini mengukur kemampuan model dalam mereplikasi pola *pseudo-label* tersebut, bukan kesesuaian terhadap anotasi manusia.

Konsekuensi metodologis dari penggunaan *pseudo-label* perlu dipertimbangkan secara khusus dalam interpretasi hasil *ablation study*. Karena label diperoleh melalui aturan *lexicon*, konfigurasi *Lexicon-only* yang menggunakan skor leksikal sebagai fitur memiliki hubungan langsung dengan sumber pembentukan label. Kondisi tersebut dapat menghasilkan estimasi performa yang tinggi karena model berpotensi mempelajari kembali pola yang digunakan dalam proses pelabelan. Oleh karena itu, hasil konfigurasi *Lexicon-only* tidak diperlakukan sebagai bukti kemampuan generalisasi terhadap persepsi sentimen manusia, melainkan sebagai ukuran kesesuaian model terhadap *silver standard* yang dibentuk oleh aturan leksikon. Konfigurasi tersebut tetap dipertahankan dalam *ablation study* untuk menunjukkan tingkat keterkaitan antara fitur leksikal dan label yang dihasilkan, bukan untuk dijadikan pembandingan yang bebas dari bias pelabelan. Sebaliknya, konfigurasi XLM-R-only digunakan untuk mengevaluasi kemampuan representasi semantik dalam mempelajari pola label tanpa memasukkan skor leksikal secara langsung sebagai fitur. Meskipun konfigurasi ini tetap dilatih dan dievaluasi menggunakan *silver standard*, hasilnya memberikan pembandingan yang relatif lebih terpisah dari aturan fitur leksikal dibandingkan konfigurasi *Lexicon-only*.

Komposisi kamus sentimen yang digunakan dalam penelitian disajikan pada Tabel 3. Tabel 3 merangkum seluruh lapisan kamus yang digunakan, mencakup contoh entri dari ketiga bahasa beserta bobot polaritas yang ditetapkan untuk masing-masing kategori kosakata.

Tabel 3. Komposisi Kamus Sentimen

Lapisan	Jenis	Contoh Entri	Bobot/Keterangan
Indonesia umum	Positif	bagus, baik, mantap, hebat, keren, senang, suka, cinta, indah	1 atau 2
Jawa	Positif	apik, becik, gayeng, pener, sae, nggumunke, mampesona, nyenengke, endah	1
Inggris (<i>code-mix</i>)	Positif	good, great, awesome, amazing, nice, love, perfect	1 atau 2
Indonesia umum	Negatif	buruk, jelek, parah, bodoh, bising, kesal, marah, benci	-1 atau -2
Jawa	Negatif	elek, ala, ora apik, nyebahi, nggilani, mboseni	-1 atau -2
Inggris (<i>code-mix</i>)	Negatif	bad, terrible, awful, horrible, hate, ugly, worst	-1 atau -2
Negasi	—	tidak, bukan, jangan, belum, tak, tanpa, kurang	Membalik polaritas
Normalisasi <i>slang</i>	—	gk/ga/nggak → tidak; bgt → banget; wes/udh → sudah	66 entri

Berdasarkan Tabel 3, kamus sentimen yang digunakan mencakup kosakata bahasa Indonesia, bahasa Jawa, dan bahasa Inggris untuk menangani variasi bahasa serta penggunaan *code-mixing* pada komentar TikTok. Kosakata positif diberikan bobot antara 1 dan 2, sedangkan kosakata negatif diberikan bobot antara -1 dan -2 sesuai dengan tingkat polaritas yang ditetapkan. Kamus bahasa Jawa digunakan untuk menangkap ekspresi lokal seperti “apik”, “becik”, “elek”, dan “nggilani”. Selain itu, daftar negasi digunakan untuk membalik polaritas kata yang dipengaruhi oleh bentuk negasi, sedangkan kamus normalisasi *slang* digunakan untuk menyamakan bentuk tidak baku dengan bentuk kata yang lebih standar.

2.4 Ekstraksi *Embedding* dan Model Klasifikasi

Representasi semantik setiap komentar diekstraksi menggunakan model XLM-RoBERTa Base. Model dijalankan dalam mode evaluasi tanpa pembaruan bobot sehingga proses penelitian memanfaatkan representasi yang telah diperoleh dari model *pre-trained*. Vektor token CLS pada lapisan terakhir digunakan sebagai representasi setiap komentar dengan dimensi 768. XLM-RoBERTa digunakan karena dirancang untuk menghasilkan representasi lintas bahasa pada berbagai bahasa, termasuk bahasa dengan sumber daya terbatas [14].

Embedding XLM-RoBERTa digunakan sebagai fitur masukan bagi CatBoost. Skor *lexicon* tidak dimasukkan sebagai fitur pada model utama karena skor tersebut telah digunakan dalam pembentukan label sehingga penggunaannya kembali sebagai fitur dapat menyebabkan *data leakage*. Dengan demikian, model utama mempelajari hubungan antara representasi semantik XLM-RoBERTa dan *pseudo-label* yang telah dihasilkan tanpa menggunakan skor yang sama sebagai prediktor.

Sebelum pelatihan model utama, data diseimbangkan menggunakan Random Undersampling dengan mempertahankan seluruh 401 komentar negatif dan memilih 401 komentar positif dari total 559 komentar positif. Proses tersebut menghasilkan 802 sampel dengan distribusi kelas yang seimbang. Dataset kemudian dibagi menggunakan rasio 80:20 sehingga diperoleh 641 data latih dan 161 data uji.

Model CatBoost digunakan sebagai pengklasifikasi dengan konfigurasi maksimum 500 iterasi, kedalaman pohon sebesar 6, dan *learning rate* sebesar 0,0001. Pelatihan menggunakan mekanisme *early stopping* untuk menghentikan proses ketika performa validasi tidak lagi mengalami peningkatan. Konfigurasi model utama tersebut

digunakan untuk memperoleh hasil klasifikasi yang selanjutnya dibandingkan dengan hasil *ablation study*. Penggunaan CatBoost pada penelitian ini mengacu pada mekanisme *ordered boosting* yang menjadi karakteristik algoritma tersebut [20].

2.5 Skenario Pengujian dan Evaluasi

Evaluasi dilakukan melalui dua skenario. Skenario pertama merupakan model utama, yaitu CatBoost yang dilatih menggunakan *embedding* XLM-RoBERTa pada dataset yang telah diseimbangkan melalui Random Undersampling. Skenario kedua merupakan *ablation study* yang menggunakan 960 data sebelum penyeimbangan untuk mempertahankan distribusi label hasil *lexicon scoring*. *Ablation study* terdiri atas tiga konfigurasi, yaitu *Lexicon-only*, XLM-R only, dan XLM-R + *Lexicon* (*Fused*).

Sebelum membahas masing-masing konfigurasi, perlu diperjelas mengapa performa model utama (89,44%) lebih rendah secara numerik dibandingkan konfigurasi *XLM-R only* pada *ablation study* (93,23%). Perbedaan nilai performa antara model utama dan konfigurasi *XLM-R-only* perlu dipahami dengan mempertimbangkan perbedaan kondisi eksperimental. Model utama menggunakan 802 data yang telah diseimbangkan, sedangkan konfigurasi *XLM-R-only* pada *ablation study* menggunakan 960 data dengan distribusi label yang tidak seimbang. Selain itu, kedua skenario memiliki konfigurasi pelatihan yang berbeda. Oleh karena itu, nilai performa dari kedua skenario tidak dapat dibandingkan sebagai perbandingan langsung dalam kondisi data dan parameter yang identik, melainkan digunakan untuk memberikan gambaran mengenai kinerja masing-masing konfigurasi pada skenario pengujian yang ditetapkan.

Konfigurasi *Lexicon-only* menggunakan *skor leksikal* sebagai satu-satunya fitur masukan CatBoost. Konfigurasi XLM-R only menggunakan 768 dimensi *embedding* XLM-RoBERTa tanpa skor leksikal. Sementara itu, konfigurasi XLM-R + *Lexicon* menggunakan mekanisme *late fusion* dengan bobot $\alpha = 0,5$ untuk menggabungkan probabilitas model XLM-RoBERTa dan probabilitas berbasis *lexicon*. Model *Lexicon-only*, XLM-R only, dan *Fused* menggunakan 960 data yang belum diseimbangkan sehingga kontribusi masing-masing komponen dapat diamati dalam kondisi dataset yang sama. Perbandingan tersebut dipahami sebagai evaluasi internal antar-konfigurasi yang tetap menggunakan *silver standard*, bukan sebagai evaluasi terhadap anotasi manusia. Rincian konfigurasi pengujian disajikan pada Tabel 4.

Tabel 4. Konfigurasi *Ablation Study*

Konfigurasi	Fitur/Mekanisme	Iterasi CatBoost	Depth	Data Label
<i>Lexicon-only</i>	<i>skor leksikal</i> skalar (1 dimensi)	200	4	960 data tidak diseimbangkan
XLM-R only	<i>Embedding</i> XLM-RoBERTa (768 dimensi)	300	6	960 data tidak diseimbangkan
XLM-R + <i>Lexicon</i> (<i>Fused</i>)	<i>Late fusion</i> , $\alpha = 0,5$	300	6	960 data tidak diseimbangkan
Model Utama	<i>Embedding</i> XLM-RoBERTa (768 dimensi), tanpa <i>skor leksikal</i>	500	6	Data seimbang (802)

Berdasarkan Tabel 4, konfigurasi *Lexicon-only* menggunakan skor leksikal berdimensi satu sebagai fitur masukan, sedangkan konfigurasi XLM-R only menggunakan *embedding* XLM-RoBERTa dengan 768 dimensi. Konfigurasi XLM-R + *Lexicon* (*Fused*) menggabungkan probabilitas dari model berbasis XLM-RoBERTa dan pendekatan leksikal melalui mekanisme *late fusion* dengan nilai $\alpha = 0,5$. Ketiga konfigurasi dalam *ablation study* menggunakan 960 data yang belum diseimbangkan, tetapi memiliki perbedaan pada fitur dan mekanisme yang digunakan. Sebagai pembandingan, model utama menggunakan *embedding* XLM-RoBERTa tanpa skor leksikal, 500 iterasi CatBoost, kedalaman pohon sebesar 6, serta dataset seimbang yang terdiri atas 802 data.

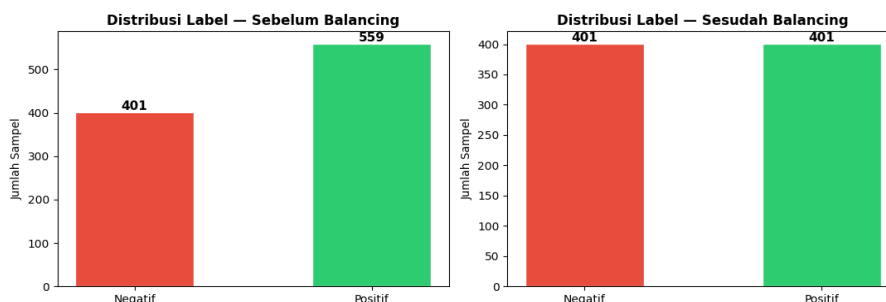
Performa seluruh konfigurasi dievaluasi menggunakan accuracy, precision, recall, F1-score, dan ROC-AUC. Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap keseluruhan data, sedangkan precision dan recall digunakan untuk mengevaluasi ketepatan serta kemampuan model dalam mengenali masing-masing kelas. F1-score digunakan untuk memberikan ukuran keseimbangan antara precision dan recall, sementara ROC-AUC digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas positif dan negatif berdasarkan probabilitas prediksi.

3. HASIL DAN PEMBAHASAN

3.1 Distribusi Data dan Hasil *Preprocessing*

Setelah proses *scraping*, diperoleh 7.330 komentar dari lima video *Sound Horeg*. Penghapusan data kosong menghasilkan 7.260 komentar, sedangkan proses *cleaning* melalui *jalur pertama* menghasilkan 6.060 komentar yang dapat digunakan pada tahap pelabelan dan ekstraksi representasi. Pelabelan berbasis *lexicon* selanjutnya menghasilkan 960 komentar berlabel yang terdiri atas 559 komentar positif dan 401 komentar negatif. Distribusi tersebut menunjukkan bahwa kelas positif memiliki jumlah data lebih besar daripada kelas negatif dengan selisih 158

komentar. Ketidakseimbangan tersebut kemudian ditangani menggunakan Random Undersampling dengan mempertahankan seluruh 401 komentar negatif dan mengambil 401 komentar positif sehingga diperoleh 802 data dengan distribusi kelas yang seimbang. Dataset tersebut kemudian dibagi menjadi 641 data latih dan 161 data uji menggunakan rasio 80:20. Rangkaian perubahan jumlah data menunjukkan bahwa sebagian besar pengurangan data terjadi pada tahap *cleaning* dan penyaringan komentar netral, sedangkan Random Undersampling digunakan secara khusus untuk mengendalikan ketidakseimbangan kelas sebelum proses pelatihan model. Tahapan tersebut menjadi dasar bagi evaluasi model utama yang menggunakan data seimbang. Perbandingan distribusi kelas sebelum dan sesudah proses *Random Undersampling* disajikan pada Gambar 2.



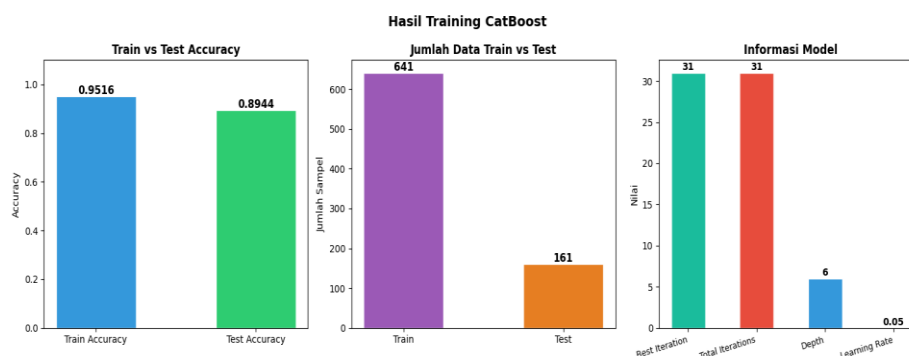
Gambar 2. Distribusi Label Sebelum dan Sesudah Random *Undersampling*

Gambar 2 memperlihatkan perubahan distribusi kelas sebelum dan sesudah *Random Undersampling*. Sebelum penyeimbangan, jumlah komentar positif mencapai 559, sedangkan komentar negatif berjumlah 401. Setelah penyeimbangan, kedua kelas memiliki jumlah yang sama, yaitu 401 komentar. Penyeimbangan tersebut dilakukan untuk mengurangi dominasi kelas positif pada proses pembelajaran model. Dengan distribusi yang sama pada kedua kelas, performa model dapat diamati tanpa memberikan proporsi data yang lebih besar kepada salah satu kelas selama proses pelatihan. Namun, karena Random Undersampling menghilangkan sebagian data positif, jumlah data yang tersedia untuk pembelajaran menjadi lebih kecil daripada data berlabel awal. Konsekuensi tersebut perlu dipertimbangkan ketika menginterpretasikan performa model karena hasil evaluasi tidak hanya dipengaruhi oleh representasi fitur, tetapi juga oleh jumlah data yang digunakan dalam pelatihan.

3.2 Hasil Pelatihan Model Utama

Model utama menggunakan representasi *embedding* XLM-RoBERTa sebagai fitur masukan bagi CatBoost. Model dikonfigurasi dengan maksimum 500 iterasi, kedalaman pohon 6, *learning rate* 0,0001, dan mekanisme *early stopping*. Hasil pelatihan menunjukkan bahwa akurasi data latih mencapai 95,16%, sedangkan akurasi pada data uji mencapai 89,44%. Selisih sekitar 5,72 poin persentase antara akurasi data latih dan validasi menunjukkan adanya perbedaan performa antara data yang digunakan untuk pembelajaran dan data validasi. Perbedaan tersebut mengindikasikan adanya kecenderungan *overfitting*, tetapi model masih mempertahankan performa validasi yang relatif tinggi. Perkembangan akurasi selama proses pelatihan dan validasi model utama CatBoost disajikan pada Gambar 3.

Berdasarkan Gambar 3, performa terbaik dicapai pada iterasi ke-31. Setelah titik tersebut, akurasi validasi tidak menunjukkan peningkatan selama 50 iterasi sehingga mekanisme *early stopping* menghentikan proses pelatihan. Model kemudian mempertahankan bobot terbaik yang diperoleh pada tahap tersebut. Konvergensi yang relatif cepat menunjukkan bahwa representasi 768 dimensi yang dihasilkan XLM-RoBERTa telah menyediakan informasi yang cukup bagi CatBoost untuk membentuk pola klasifikasi pada dataset penelitian. Dengan demikian, penambahan jumlah iterasi hingga batas maksimum 500 tidak secara otomatis menghasilkan peningkatan performa. Penggunaan *early stopping* menjadi penting karena dapat mencegah model terus melakukan pembelajaran ketika performa validasi tidak lagi mengalami peningkatan.



Gambar 3. Kurva *Training* dan *Validation Accuracy* Model Utama CatBoost

3.3 Evaluasi Performa Model Utama

Evaluasi model utama dilakukan menggunakan 161 data uji yang terdiri atas 81 komentar negatif dan 80 komentar positif. Evaluasi mencakup precision, recall, F1-Score, accuracy, F1-Macro, dan ROC-AUC untuk mengukur kemampuan model dalam membedakan kedua kelas sentimen. Hasil evaluasi tersebut disajikan pada Tabel 5.

Tabel 5. *Classification Report Model Utama*

Kelas	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>	Support
Negatif	0,91	0,88	0,89	81
Positif	0,88	0,91	0,90	80
<i>Accuracy</i>			0,8944	161
<i>Macro Average</i>	0,89	0,89	0,89	161
ROC-AUC			0,9025	

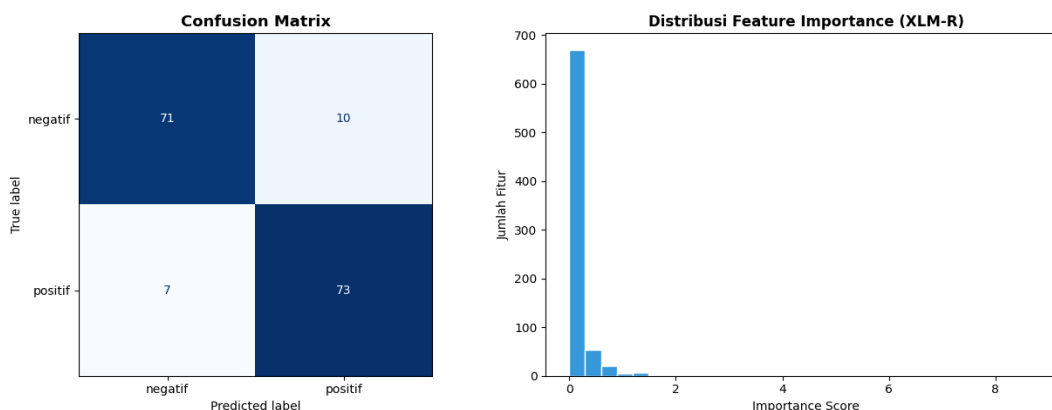
Berdasarkan Tabel 5, model memperoleh *precision* sebesar 0,91, *recall* sebesar 0,88, dan F1-Score sebesar 0,89 untuk kelas negatif. Pada kelas positif, model memperoleh *precision* sebesar 0,88, *recall* sebesar 0,91, dan F1-Score sebesar 0,90. Secara keseluruhan, model menghasilkan *accuracy* sebesar 89,44%, F1-Macro sebesar 89,44%, dan ROC-AUC sebesar 90,25%. Hasil tersebut menunjukkan bahwa model memiliki kemampuan klasifikasi yang relatif seimbang antara kedua kelas. *Recall* kelas positif sebesar 0,91 menunjukkan bahwa model mampu mengenali sebagian besar komentar yang diberi label positif. Sementara itu, *recall* kelas negatif sebesar 0,88 menunjukkan bahwa masih terdapat sebagian komentar negatif yang diklasifikasikan sebagai positif. Perbedaan tersebut relatif kecil sehingga model tidak menunjukkan kecenderungan yang sangat kuat terhadap salah satu kelas. Nilai ROC-AUC sebesar 90,25% juga menunjukkan bahwa model memiliki kemampuan diskriminasi yang baik dalam membedakan dua kelas berdasarkan probabilitas prediksi.

Meskipun demikian, hasil tersebut perlu dipahami dalam konteks sumber label yang digunakan. Label sentimen diperoleh melalui pendekatan *pseudo-label* berbasis *lexicon* dan bukan melalui anotasi manual oleh penilai manusia. Oleh karena itu, nilai *accuracy* dan F1-Score yang diperoleh menggambarkan kemampuan model dalam mereplikasi pola label yang dihasilkan oleh proses *lexicon*. Model yang memiliki performa tinggi terhadap *pseudo-label* belum secara otomatis menunjukkan kesesuaian penuh dengan persepsi sentimen manusia. Kondisi tersebut menjadi salah satu alasan dilakukannya *ablation study* untuk melihat pengaruh sumber informasi leksikal dan representasi semantik secara terpisah.

3.4 Analisis Confusion Matrix dan Feature Importance

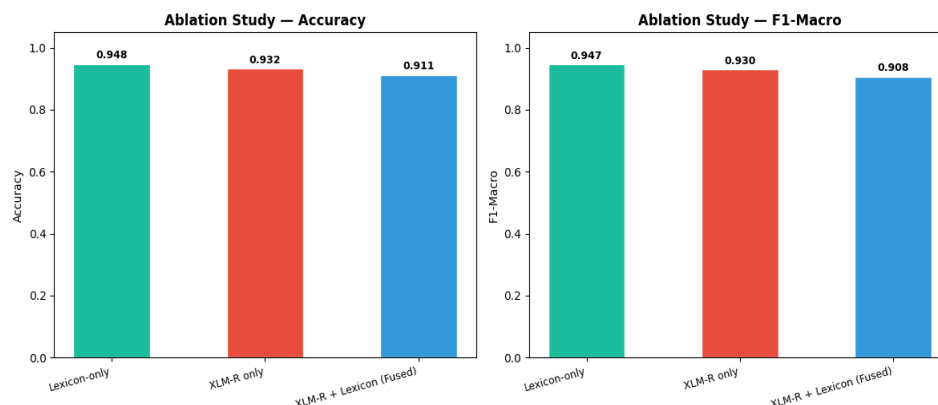
Pola kesalahan klasifikasi model pada data uji dianalisis menggunakan *confusion matrix*. Hasil *confusion matrix* model utama disajikan pada Gambar 4.

Berdasarkan Gambar 4, pola kesalahan klasifikasi dapat diamati melalui jumlah prediksi yang benar dan prediksi yang keliru pada masing-masing kelas. Hasil evaluasi menunjukkan bahwa model memiliki *recall* positif sebesar 0,91 dan *recall* negatif sebesar 0,88. Perbedaan tersebut menunjukkan bahwa model sedikit lebih baik dalam mengenali komentar positif dibandingkan komentar negatif. Kesalahan klasifikasi pada kelas negatif menunjukkan bahwa sebagian komentar negatif masih memiliki representasi yang dapat dianggap menyerupai karakteristik komentar positif oleh model. Kondisi tersebut dapat berkaitan dengan karakteristik komentar TikTok yang menggunakan bahasa informal, ekspresi pendek, serta variasi kosakata yang dapat memiliki makna berbeda bergantung pada konteks.



Gambar 4. *Confusion Matrix Model Utama*

Analisis kontribusi fitur dilakukan untuk mengetahui dimensi representasi XLM-RoBERTa yang paling banyak digunakan dalam proses pengambilan keputusan CatBoost. Hasil analisis *feature importance* model utama disajikan pada Gambar 5.



Gambar 5. Feature Importance Model Utama

Berdasarkan Gambar 5, *feature importance* digunakan untuk melihat kontribusi relatif masing-masing dimensi representasi XLM-RoBERTa terhadap keputusan klasifikasi CatBoost. Karena fitur yang digunakan berupa vektor *embedding* berdimensi 768, nilai *importance* yang tinggi tidak dapat secara langsung diterjemahkan sebagai kata atau istilah tertentu. Nilai tersebut merepresentasikan kontribusi dimensi numerik pada ruang representasi XLM-RoBERTa terhadap keputusan klasifikasi. Oleh karena itu, *feature importance* pada penelitian ini lebih tepat digunakan untuk mengevaluasi kontribusi dimensi representasi terhadap model daripada memberikan interpretasi linguistik secara langsung.

Temuan tersebut juga memperlihatkan keterbatasan interpretabilitas model berbasis *embedding*. Walaupun XLM-RoBERTa mampu menghasilkan representasi semantik yang kaya, hubungan antara dimensi numerik dan makna linguistik tidak bersifat langsung. Meskipun demikian, analisis *feature importance* tetap berguna untuk mengidentifikasi apakah keputusan model cenderung bertumpu pada sejumlah kecil dimensi representasi atau tersebar pada banyak dimensi. Informasi tersebut dapat menjadi dasar evaluasi lanjutan terhadap stabilitas fitur dan interpretabilitas model, meskipun belum cukup untuk menjelaskan hubungan antara dimensi tertentu dan ekspresi linguistik lokal.

Dalam konteks penelitian ini, penggunaan kamus bahasa Jawa tidak hanya diposisikan sebagai sumber informasi dalam proses pelabelan, tetapi juga dievaluasi secara terpisah untuk melihat sejauh mana ekspresi lokal muncul dalam *dataset*. Evaluasi tersebut diperlukan karena kontribusi informasi bahasa daerah tidak dapat disimpulkan hanya berdasarkan nilai *feature importance* dari *embedding* XLM-RoBERTa.

3.5 Hasil Ablation Study

Ablation study dilakukan untuk mengidentifikasi kontribusi masing-masing komponen dalam *pipeline*. Berbeda dengan model utama yang menggunakan 802 data hasil Random Undersampling, tiga konfigurasi *ablation study* menggunakan 960 data sebelum penyeimbangan. Pengaturan tersebut mempertahankan distribusi label hasil *lexicon scoring* sehingga setiap konfigurasi dapat dibandingkan dalam kondisi data yang sama. Tiga konfigurasi yang diuji terdiri atas *Lexicon-only*, *XLM-R only*, dan *XLM-R + Lexicon (Fused)*.

Untuk mengetahui kontribusi masing-masing pendekatan terhadap hasil klasifikasi, penelitian ini melakukan *ablation study* dengan membandingkan tiga konfigurasi model, yaitu *Lexicon-only*, *XLM-R only*, dan *XLM-R + Lexicon (Fused)*. Seluruh konfigurasi menggunakan 960 data berlabel yang sama sebelum proses penyeimbangan kelas sehingga perbandingan difokuskan pada pengaruh fitur dan metode yang digunakan. Hasil pengujian setiap konfigurasi ditampilkan pada Tabel 6.

Tabel 6. Hasil Ablation Study pada Data Uji

Konfigurasi	Accuracy	Macro F1-Score
<i>Lexicon-only</i>	94,79%	94,66%
<i>XLM-R only</i>	93,23%	93,02%
<i>XLM-R + Lexicon (Fused)</i>	91,15%	90,76%
Δ Fused vs XLM-R only	-2,08%	-2,26%

Berdasarkan Tabel 6, konfigurasi *Lexicon-only* menghasilkan nilai accuracy tertinggi sebesar 94,79% dan F1-Macro sebesar 94,66%. Meskipun demikian, hasil tersebut tidak dapat langsung dianggap sebagai bukti bahwa pendekatan berbasis leksikon merupakan metode terbaik karena label sentimen pada penelitian ini juga dibentuk menggunakan aturan leksikal yang sama. Kondisi tersebut menimbulkan *circular reasoning*, yaitu fitur atau aturan yang digunakan untuk menghasilkan label kemudian diuji kembali untuk memprediksi label tersebut. Oleh karena itu, performa *Lexicon-only* lebih tepat dipahami sebagai hasil diagnostik terhadap konsistensi aturan pelabelan, bukan sebagai pembanding yang sepenuhnya independen.

Konfigurasi *Lexicon-only* menghasilkan accuracy tertinggi sebesar 94,79% dan F1-Macro sebesar 94,66%. Namun, performa tersebut perlu ditafsirkan secara hati-hati karena fitur yang digunakan untuk klasifikasi berupa *skor leksikal* berasal dari sumber informasi yang sama dengan label *pseudo-label*. Dengan kata lain, model pada konfigurasi ini mempelajari hubungan antara fitur yang secara langsung digunakan dalam proses pembentukan label dengan label tersebut. Kondisi tersebut dapat menghasilkan estimasi performa yang lebih tinggi dan tidak dapat digunakan sebagai bukti bahwa *Lexicon-only* merupakan pendekatan yang paling efektif untuk klasifikasi sentimen secara independen.

Konfigurasi XLM-R only menghasilkan accuracy sebesar 93,23% dan F1-Macro sebesar 93,02%. Hasil ini lebih representatif untuk melihat kemampuan representasi semantik XLM-RoBERTa karena fitur berupa *embedding* tidak digunakan secara langsung dalam pembentukan label. Perbedaan sebesar 1,56 poin persentase accuracy dibandingkan *Lexicon-only* menunjukkan bahwa representasi semantik XLM-RoBERTa mampu menghasilkan performa klasifikasi yang mendekati fitur leksikal meskipun tidak memperoleh informasi label secara langsung dari fitur tersebut. Hasil ini memperkuat alasan penggunaan XLM-RoBERTa sebagai fitur utama pada model yang dirancang dalam penelitian.

Sebaliknya, konfigurasi XLM-R + *Lexicon (Fused)* menghasilkan accuracy sebesar 91,15% dan F1-Macro sebesar 90,76%. Nilai tersebut lebih rendah dibandingkan XLM-R only dengan penurunan accuracy sebesar 2,08 poin persentase dan F1-Macro sebesar 2,26 poin persentase. Penurunan tersebut menunjukkan bahwa penggabungan dua sumber informasi melalui *late fusion* dengan bobot yang sama belum menghasilkan peningkatan performa. Pemberian bobot 0,5 pada masing-masing sumber mengasumsikan bahwa informasi dari XLM-RoBERTa dan *lexicon* memiliki kontribusi yang setara, padahal karakteristik dan kualitas informasi keduanya berbeda. Selain itu, informasi leksikal memiliki hubungan langsung dengan label *pseudo-label*, sehingga penambahannya pada proses *fusion* dapat membawa kembali informasi yang telah digunakan pada tahap pelabelan.

Secara keseluruhan, hasil *ablation study* menunjukkan bahwa performa numerik tertinggi tidak selalu menunjukkan konfigurasi yang paling valid secara metodologis. *Lexicon-only* memperoleh hasil terbaik karena memiliki hubungan langsung dengan mekanisme pelabelan, sedangkan XLM-R only memberikan alternatif yang lebih terpisah dari aturan leksikal. Hasil konfigurasi *Fused* yang lebih rendah juga memperlihatkan bahwa integrasi informasi leksikal tidak selalu meningkatkan performa, khususnya ketika informasi tersebut memiliki keterkaitan dengan sumber *pseudo-label*. Temuan ini memperkuat pentingnya memisahkan sumber label dan fitur serta mengevaluasi setiap komponen secara terpisah.

3.6 Analisis Kesalahan (Error Analysis)

Analisis kesalahan dilakukan untuk memahami pola komentar yang salah diklasifikasikan oleh model. Berdasarkan *confusion matrix* pada Gambar 4, terdapat sejumlah komentar positif yang diklasifikasikan sebagai negatif (*false negative*) dan sejumlah komentar negatif yang diklasifikasikan sebagai positif (*false positive*). Beberapa pola kesalahan yang teridentifikasi dapat dijelaskan melalui karakteristik linguistik komentar TikTok.

Pertama, kesalahan pada komentar yang menggunakan sarkasme atau ironi. Komentar seperti “mantap banget bikin pusing” secara leksikal mengandung kata positif “mantap” namun konteks keseluruhannya bersifat negatif. Model yang dilatih pada *pseudo-label* berbasis kamus tidak dapat menangkap konteks sarkasme karena label ditetapkan berdasarkan keberadaan kata berbobot, bukan interpretasi kontekstual secara menyeluruh.

Kedua, kesalahan pada komentar yang menggunakan singkatan atau variasi penulisan yang tidak terdaftar dalam kamus normalisasi. Ekspresi seperti “wkwkwk” atau “haha” yang bersifat ambigu secara sentimen sering kali tidak memperoleh skor yang akurat dari pendekatan leksikal sehingga representasi XLM-RoBERTa menjadi satu-satunya sumber informasi yang tersedia.

Ketiga, kesalahan pada komentar pendek yang hanya terdiri dari satu atau dua kata. Komentar seperti “gas” atau “oke” memiliki representasi semantik yang kurang informatif karena konteks kalimat tidak tersedia sehingga model kesulitan menentukan polaritas yang tepat. Temuan tersebut menunjukkan bahwa keterbatasan model bukan semata-mata berasal dari arsitektur, melainkan juga dari karakteristik data komentar TikTok yang sangat bervariasi secara linguistik.

3.7 Distribusi Sentimen pada Kelompok Data

Analisis distribusi sentimen dilakukan untuk menggambarkan kecenderungan respons pengguna terhadap konten Sound Horeg dan video pembandingan. Analisis ini tidak digunakan untuk mengukur hubungan sebab-akibat, tetapi untuk menunjukkan proporsi komentar positif dan negatif berdasarkan hasil pelabelan berbasis leksikon. Distribusi sentimen pada masing-masing kelompok data ditampilkan pada Tabel 7.

Tabel 7. Perbandingan Distribusi Sentimen Kelompok A dan Kelompok B

Dataset	Positif	Negatif	Total	Rasio Positif
Kelompok A (5 video <i>Sound Horeg</i>)	559	401	960	57,81%
Kelompok B (dataset komparatif)	2.105	1.183	3.288	64,02%
Sholawatan @inaarbha	831	476	1.307	63,58%

Berdasarkan Tabel 7, kelompok Sound Horeg A memiliki 559 komentar positif dan 401 komentar negatif dari total 960 komentar yang berhasil diberi label. Dengan demikian, komentar positif memiliki proporsi sebesar 57,81%,

sedangkan komentar negatif menjadi bagian sebesar 42,19%. Hasil tersebut menunjukkan bahwa respons positif lebih dominan, meskipun selisih antara kedua kategori tidak terlalu besar.

Pada kelompok Sound Horeg B, terdapat 2.105 komentar positif dan 1.183 komentar negatif dari total 3.288 komentar. Proporsi komentar positif mencapai 64,02%, sehingga kelompok ini menunjukkan kecenderungan sentimen positif yang lebih tinggi dibandingkan kelompok Sound Horeg A. Sementara itu, video sholawatan memiliki 831 komentar positif dan 476 komentar negatif dari total 1.307 komentar, dengan proporsi positif sebesar 63,58%.

Perbedaan distribusi tersebut menunjukkan adanya variasi proporsi sentimen pada kelompok data yang dianalisis. Namun, hasil ini tidak digunakan untuk menyimpulkan bahwa kategori konten tertentu secara langsung menyebabkan munculnya sentimen positif atau negatif. Analisis yang dilakukan bersifat deskriptif sehingga perbedaan rasio hanya menunjukkan karakteristik distribusi pada dataset yang tersedia. Selain itu, perbedaan ukuran dataset dan prosedur pembentukan masing-masing kelompok perlu dipertimbangkan ketika membandingkan hasil antarkelompok.

3.8 Kontribusi Kosakata Bahasa Jawa

Analisis kosakata bahasa Jawa dilakukan untuk mengetahui sejauh mana ekspresi lokal muncul dalam komentar yang dianalisis. Analisis ini penting karena komentar TikTok dapat memuat bahasa daerah, bahasa Indonesia, bahasa Inggris, maupun bentuk *code-mixing*. Pengukuran dilakukan dengan menghitung jumlah komentar yang mengandung kosakata bahasa Jawa berdasarkan kamus yang digunakan dalam proses pelabelan. Hasil identifikasi kosakata bahasa Jawa ditampilkan pada Tabel 8.

Tabel 8. Statistik Komentar Berkosa Kata Bahasa Jawa

Metrik	Kelompok A	Kelompok B
Total komentar dianalisis	960	3.288
Komentar berkosakata Jawa	15 (1,6%)	25 (0,8%)
Rasio positif komentar Jawa	66,7%	60,0%

Berdasarkan Tabel 8, kosakata bahasa Jawa ditemukan pada 15 komentar dari 960 komentar pada Sound Horeg Kelompok A atau sebesar 1,6%. Dari jumlah tersebut, 66,7% komentar memiliki label positif dan 33,3% memiliki label negatif. Pada Sound Horeg Kelompok B, kosakata bahasa Jawa ditemukan pada 25 komentar dari 3.288 komentar atau sebesar 0,8%. Komentar yang mengandung kosakata Jawa pada kelompok ini terdiri atas 60,0% komentar positif dan 40,0% komentar negatif.

Persentase tersebut menunjukkan bahwa penggunaan kosakata bahasa Jawa dalam data yang dianalisis relatif terbatas. Meskipun demikian, keberadaan kosakata lokal tetap relevan karena beberapa ekspresi sentimen dapat disampaikan melalui istilah yang tidak selalu tercakup dalam kamus bahasa Indonesia atau bahasa Inggris. Oleh sebab itu, kamus bahasa Jawa dapat berfungsi sebagai sumber daya tambahan untuk menangani variasi linguistik lokal, tetapi kontribusinya dalam penelitian ini belum dapat dianggap sebagai faktor dominan karena jumlah kemunculannya relatif kecil.

Rasio sentimen positif pada komentar yang mengandung kosakata bahasa Jawa mencapai 66,7% pada Kelompok A dan 60,0% pada Kelompok B. Hasil tersebut menunjukkan bahwa komentar yang mengandung kosakata bahasa Jawa tidak selalu memiliki polaritas yang sama sehingga keberadaan kosakata lokal tidak dapat dijadikan satu-satunya dasar untuk menentukan sentimen. Makna sentimen tetap bergantung pada kombinasi kata dan konteks yang terdapat dalam komentar.

Proporsi kosakata bahasa Jawa yang relatif kecil juga menunjukkan bahwa kontribusi kamus bahasa Jawa pada dataset masih terbatas dari sisi cakupan data. Kondisi ini tidak berarti bahwa informasi bahasa Jawa tidak relevan, tetapi menunjukkan bahwa jumlah ekspresi lokal yang terdeteksi dalam dataset belum cukup besar untuk mendominasi proses klasifikasi. Temuan tersebut mendukung penggunaan pendekatan yang mengombinasikan representasi semantik multibahasa dengan pengetahuan linguistik lokal. XLM-RoBERTa dapat digunakan untuk menangkap informasi semantik yang lebih luas, sedangkan kamus bahasa Jawa memberikan sumber informasi tambahan ketika komentar mengandung ekspresi yang khas secara lokal.

Dengan demikian, analisis kosakata bahasa Jawa memperlihatkan bahwa sumber daya leksikal lokal telah digunakan dalam proses pelabelan, tetapi cakupannya terhadap keseluruhan data masih terbatas. Temuan ini menunjukkan perlunya pengembangan kamus yang lebih luas serta evaluasi menggunakan data yang memiliki proporsi bahasa daerah lebih besar agar kontribusi pengetahuan linguistik lokal dapat diuji secara lebih komprehensif.

3.9 Pembahasan

Secara keseluruhan, hasil penelitian menunjukkan bahwa *pipeline* yang dikembangkan mampu menghasilkan model klasifikasi dengan performa yang cukup baik pada data komentar TikTok yang memiliki karakteristik informal dan multibahasa. Model utama menghasilkan accuracy 89,44%, F1-Macro 89,44%, dan ROC-AUC 90,25% dengan menggunakan *embedding* XLM-RoBERTa sebagai fitur dan CatBoost sebagai pengklasifikasi. Performa tersebut menunjukkan bahwa representasi semantik XLM-RoBERTa dapat digunakan sebagai dasar klasifikasi sentimen ketika dipadukan dengan algoritma *gradient boosting*. Namun, hasil tersebut tidak dapat dilepaskan dari karakteristik label yang digunakan karena seluruh model belajar berdasarkan *pseudo-label* yang diperoleh melalui pendekatan *lexicon*.

Perbandingan melalui *ablation study* memberikan temuan yang lebih penting daripada sekadar nilai accuracy tertinggi. *Lexicon-only* menghasilkan accuracy 94,79%, tetapi performanya berpotensi bias karena fitur berasal dari sumber yang sama dengan label. XLM-R only menghasilkan accuracy 93,23% dan memberikan gambaran yang lebih representatif mengenai kemampuan fitur semantik dalam mereproduksi label. Sementara itu, konfigurasi *Fused* menghasilkan accuracy 91,15%, sehingga kombinasi langsung antara kedua sumber informasi dengan bobot yang sama belum memberikan keuntungan. Hasil ini menunjukkan bahwa integrasi dua pendekatan tidak selalu meningkatkan performa apabila mekanisme penggabungannya belum mempertimbangkan karakteristik informasi yang digunakan.

Dari sisi bahasa daerah, hanya sebagian kecil komentar yang teridentifikasi mengandung kosakata bahasa Jawa, yaitu 1,6% pada Kelompok A dan 0,8% pada Kelompok B. Temuan tersebut memperlihatkan bahwa kontribusi kamus bahasa Jawa pada dataset lebih tepat dipandang sebagai dukungan terhadap ekspresi lokal daripada sebagai sumber fitur dominan. Keterbatasan jumlah kosakata yang terdeteksi juga menunjukkan bahwa pengembangan kamus dengan cakupan kosakata dan variasi bentuk bahasa Jawa yang lebih luas masih diperlukan. Variasi penulisan, penggunaan slang, *code-mixing*, dan ekspresi kontekstual dapat menyebabkan kata yang sebenarnya membawa informasi sentimen tidak selalu cocok dengan entri kamus.

Jika dibandingkan dengan penelitian sejenis, performa model utama dalam penelitian ini berada pada rentang yang kompetitif. Winata et al. [10] melalui dataset NusaX untuk 10 bahasa daerah Indonesia melaporkan bahwa model XLM-RoBERTa tanpa *fine-tuning* menghasilkan F1-Score rata-rata di bawah 70% untuk bahasa-bahasa dengan sumber daya sangat terbatas, sedangkan penelitian ini mencapai F1-Macro sebesar 89,44% pada data komentar TikTok berbahasa campuran. Darnoto dan Firmawan [9] dalam eksperimen *transfer learning* lintas 12 bahasa daerah Indonesia menunjukkan bahwa performa pendekatan berbasis *transformer* dapat bervariasi bergantung pada karakteristik bahasa sumber dan bahasa target. Sementara itu, Koto et al. [13] menunjukkan bahwa pemanfaatan *lexicon* multibahasa dapat membantu klasifikasi sentimen pada kondisi tertentu. Namun, hasil penelitian ini memperlihatkan bahwa manfaat informasi leksikal tidak bersifat otomatis, karena konfigurasi XLM-R + *Lexicon* (*Fused*) justru menghasilkan performa lebih rendah daripada XLM-R only. Perbedaan tersebut dapat dipengaruhi oleh mekanisme penggabungan fitur, bobot *fusion*, serta keterkaitan informasi leksikal dengan proses pembentukan *pseudo-label*.

Temuan lain yang perlu diperhatikan adalah perbedaan antara performa model utama dan konfigurasi *ablation study*. Model utama memperoleh accuracy 89,44%, sedangkan XLM-R only pada *ablation study* mencapai 93,23%. Perbedaan tersebut tidak dapat langsung diartikan bahwa XLM-R only lebih baik daripada model utama karena kedua konfigurasi menggunakan kondisi data yang berbeda. Model utama menggunakan 802 data setelah Random Undersampling, sedangkan XLM-R only menggunakan 960 data dengan distribusi kelas asli. Oleh karena itu, perbedaan performa merupakan hasil gabungan dari perbedaan jumlah data, distribusi kelas, dan konfigurasi pelatihan. Perbandingan tersebut sekaligus menunjukkan pentingnya menjaga konsistensi skenario ketika membandingkan performa beberapa model.

Dengan demikian, penelitian ini belum memberikan bukti bahwa pipeline hibrida yang dikembangkan lebih unggul daripada penggunaan XLM-RoBERTa secara mandiri. Hasil eksperimen justru menunjukkan bahwa konfigurasi XLM-R only memberikan performa lebih tinggi pada skenario *ablation study* yang digunakan. Temuan tersebut menjadi hasil penting karena menunjukkan bahwa penambahan informasi leksikal tidak selalu menghasilkan peningkatan performa, terutama ketika sumber label juga dibentuk menggunakan aturan leksikal. Oleh karena itu, kontribusi penelitian ini tidak ditekankan pada klaim keunggulan absolut pipeline hibrida, tetapi pada evaluasi empiris mengenai pengaruh representasi semantik, informasi leksikal, dan risiko bias akibat penggunaan *pseudo-label* dalam klasifikasi sentimen komentar TikTok.

Secara metodologis, penelitian ini juga menunjukkan pentingnya memisahkan sumber label dan fitur dalam penggunaan *pseudo-label*. Penggunaan skor *lexicon* sebagai fitur utama dapat menghasilkan performa yang tinggi karena fitur tersebut secara langsung merepresentasikan aturan yang digunakan untuk menghasilkan label. Model utama menghindari kondisi tersebut dengan menggunakan *embedding* XLM-RoBERTa sebagai fitur, sedangkan skor *lexicon* diuji secara terpisah dalam *ablation study*. Strategi tersebut memberikan gambaran yang lebih jelas mengenai kontribusi representasi semantik dan informasi leksikal terhadap proses klasifikasi. Hasilnya menunjukkan bahwa informasi semantik XLM-RoBERTa dapat mempertahankan performa yang tinggi tanpa memasukkan skor leksikal sebagai fitur utama.

4. KESIMPULAN

Penelitian ini mengembangkan pendekatan analisis sentimen terhadap komentar YouTube berbahasa Indonesia menggunakan *pipeline* hibrida yang menggabungkan pelabelan berbasis *lexicon*, *embedding* XLM-RoBERTa, dan algoritma CatBoost dengan penanganan ketidakseimbangan kelas melalui *Random Undersampling*. Hasil pengujian menunjukkan bahwa *pipeline* utama memperoleh akurasi sebesar 89,44%, sedangkan hasil *ablation study* menggunakan XLM-RoBERTa tanpa komponen tambahan (*XLM-R only*) menghasilkan akurasi yang lebih tinggi, yaitu 93,23%. Hasil tersebut menunjukkan bahwa *pipeline* hibrida belum memberikan peningkatan performa dibandingkan penggunaan XLM-RoBERTa secara langsung, sehingga model *XLM-R only* lebih menjanjikan untuk

implementasi awal karena memiliki performa lebih tinggi dengan tahapan pemrosesan yang lebih sederhana. Dari sisi bahasa daerah, komentar yang teridentifikasi mengandung kosakata bahasa Jawa hanya mencapai 1,6% pada Kelompok A dan 0,8% pada Kelompok B, sehingga kontribusi kamus bahasa Jawa dalam *dataset* lebih tepat dipandang sebagai dukungan terhadap ekspresi lokal daripada sebagai sumber fitur dominan. Berdasarkan keseluruhan hasil tersebut, *pipeline* hibrida masih layak digunakan sebagai prototipe eksperimental untuk mengeksplorasi kombinasi fitur leksikal dan representasi kontekstual, tetapi belum menjadi pilihan utama untuk implementasi operasional karena model *XLM-R only* menunjukkan hasil yang lebih baik dan lebih sederhana. Penelitian selanjutnya disarankan untuk menggunakan data berlabel manual, menyamakan kondisi eksperimen dalam perbandingan model, memperluas cakupan kosakata bahasa daerah, serta melakukan *fine-tuning*, analisis kesalahan, dan pengujian pada data dari sumber atau periode yang berbeda guna mengevaluasi kemampuan generalisasi model.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: F. E. L., dan R. T.;

Metodologi: F. E. L., dan R. T.;

Perangkat Lunak: F. E. L.;

Validasi: F. E. L., dan R. T.;

Analisis Formal: F. E. L.;

Investigasi: F. E. L.;

Sumber Daya: F. E. L.;

Kurasi Data: F. E. L.;

Penulisan Draf Awal: F. E. L.;

Penulisan, Peninjauan, dan Penyuntingan: F. E. L., dan R. T.;

Visualisasi: F. E. L.;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] N. Apriliani, N. Suarna, and W. Prihartono, "Analisis sentimen review penggunaan TikTok melalui pendekatan algoritma Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, 2023, doi: <https://doi.org/10.36040/jati.v7i6.8299>.
- [2] A. M. Saprizal and N. Anisa, "Analisis sentimen TikTok: Wajib militer dengan metode Lexicon Based dan Naive Bayes Classifier," *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi*, vol. 4, no. 2, pp. 242–246, 2024, doi: <https://doi.org/10.46880/tamika.Vol4No2.pp242-246>.
- [3] N. Apriliani, N. Suarna, and W. Prihartono, "Analisis sentimen review penggunaan TikTok melalui pendekatan algoritma Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, 2023, doi: <https://doi.org/10.36040/jati.v7i6.8299>.
- [4] M. Isnani, G. N. Elwirehardja, and B. Pardamean, "Sentiment analysis for TikTok review using VADER sentiment and SVM model," *Procedia Computer Science*, vol. 227, pp. 168–175, 2023, doi: <https://doi.org/10.1016/j.procs.2023.10.514>.

- [5] Y. Aliyu, A. Sarlan, K. Usman Danyaro, A. S. B. Abd Rahman, and M. Abdullahi, “Sentiment analysis in low-resource settings: A comprehensive review of approaches, languages, and data sources,” *IEEE Access*, vol. 12, pp. 66883–66909, 2024, doi: <https://doi.org/10.1109/ACCESS.2024.3398635>.
- [6] N. Ahmed, R. Amin, H. Aldabbas, M. Saeed, M. Bilal, and H. Song, “A novel approach for sentiment analysis of a low resource language using deep learning models,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2024, doi: <https://doi.org/10.1145/3696789>.
- [7] L. A. Pha, L. T. Thuy, V. N. Hong, T. T. Phan, and T. K. Tran, “A systematic survey of sentiment analysis in low-resource Southeast Asian languages: Resources, methods, and evaluation challenges,” *ICT Express*, 2026, doi: <https://doi.org/10.1016/j.ict.2026.06.013>.
- [8] A. Ali, M. Khan, K. Khan, R. U. Khan, and A. Aloraini, “Sentiment analysis of low-resource language literature using data processing and deep learning,” *Computers, Materials & Continua*, vol. 79, no. 1, pp. 713–733, 2024, doi: <https://doi.org/10.32604/cmc.2024.048712>.
- [9] B. R. P. Darnoto and D. B. Firmawan, “Language-similarity-guided transfer fine-tuning of pre-trained transformer models for sentiment analysis across 12 Indonesian regional languages,” *Journal of Computing Theories and Applications*, vol. 3, no. 4, pp. 547–563, 2026, doi: <https://doi.org/10.62411/jcta.15975>.
- [10] G. I. Winata *et al.*, “NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 815–834, doi: <https://doi.org/10.18653/v1/2023.eacl-main.57>.
- [11] Z. Fatah and R. A. Ningsih, “Analisis sentimen komentar YouTube terhadap Tragedi Demo 25 Agustus menggunakan pendekatan Lexicon-Based,” *JAMASTIKA*, vol. 4, no. 2, 2025, doi: <https://doi.org/10.35473/jamastika.v4i2.4525>.
- [12] A. D. Pramesti, K. Umam, and M. R. Handayani, “Identification of buzzers in skincare reviews using a Lexicon-Based sentiment analysis method,” *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2598–2606, 2025, doi: <https://doi.org/10.30871/jaic.v9i5.11005>.
- [13] F. Koto, T. Beck, Z. Talat, I. Gurevych, and T. Baldwin, “Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1: Long Papers*, 2024, pp. 298–320, doi: <https://doi.org/10.18653/v1/2024.eacl-long.18>.
- [14] A. Conneau *et al.*, “Unsupervised cross-lingual representation learning at scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451, doi: <https://doi.org/10.18653/v1/2020.acl-main.747>.
- [15] A. Gaurav, B. B. Gupta, S. Sharma, R. Bansal, and K. T. Chui, “XLM-RoBERTa based sentiment analysis of tweets on Metaverse and 6G,” *Procedia Computer Science*, vol. 238, pp. 902–907, 2024. <https://doi.org/10.1016/j.procs.2024.06.110>
- [16] I. D. G. M. Parawangsa, G. S. Mahendra, and R. N. P. Pinatih, “Implementasi metode XLM-RoBERTa untuk analisis sentimen ulasan pengunjung objek wisata Kerta Gosa Klungkung,” *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 4, pp. 9893–9900, 2026, doi: <https://doi.org/10.31004/riggs.v4i4.5368>.
- [17] E. Yulianti and N. K. Nissa, “ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches,” *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024, doi: <https://doi.org/10.11591/eei.v13i5.8032>.
- [18] A. F. Hidayatullah, “Code-mixed sentiment analysis on Indonesian-Javanese-English text using transformer models,” in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2024, pp. 340–345, doi: <https://doi.org/10.1109/ICITISEE63424.2024.10730138>.
- [19] T. D. Purnomo and J. Sutopo, “Comparison of pre-trained BERT-based transformer models for regional language text sentiment analysis in Indonesia,” *International Journal Science and Technology*, vol. 3, no. 3, pp. 11–21, 2024, doi: <https://doi.org/10.56127/ijst.v3i3.1739>.
- [20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: Unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems*, vol. 31, pp. 6638–6648, 2018. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.