JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika

[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

ANALISIS SENTIMEN PUBLIK INDONESIA TERHADAP KONFLIK ISRAEL-IRAN DI MEDIA SOSIAL MENGGUNAKAN INDOBERT DAN EXPLAINABLE AI (LIME)

Agus Sehatman Saragih [a,1,*], Deddy Ronaldo [b,2], Ade Chandra Saputra [c,3]
[a,b,c] Universitas Palangka Raya, Kampus UPR Tanjung Nyaho, Jl. Yos Sudarso Palangka Raya, Kalimantan Tengah
[1] assaragih@it.upr.ac.id*; [2] deddy.ronaldo@it.upr.ac.id; [3] adechandra@it.upr.ac.id
* corresponding author

ARTICLE INFO

Keywords
Israel-Iran, **IndoBERT, explainable AI, LIME**

ABSTRACT (10PT)

This study analyzes Indonesian public sentiment toward the Israel–Iran conflict on social media using a fine-tuned IndoBERT model and Explainable AI (XAI) techniques based on LIME. Data were collected from YouTube and X (Twitter) between April 2024 and June 2025, yielding 7,922 unique entries that were preprocessed and automatically annotated. The IndoBERT model achieved an accuracy of 74.90% and an F1-score of 0.7521 on the test set (n = 1,513). LIME-based interpretations reveal trigger words such as "zionist" (negative) and "allahu akbar" (positive), indicating opinion polarization driven by anti-Zionist narratives and religious solidarity. This approach enhances the transparency of black-box models and provides insights for public opinion monitoring. The findings contribute to Indonesian NLP and geopolitical analysis, with limitations related to the quality of automatic annotation. The results are consistent with studies on the Israel–Palestine conflict showing dominant negative sentiment on social media. Similar analyses using BERT for Reddit discussions highlight the importance of hybrid approaches for stance. The integration of XAI techniques such as LIME has been shown to be effective in explaining sentiment predictions. Therefore, this methodology enriches the understanding of opinion dynamics in sensitive geopolitical contexts.

1. Pendahuluan

Konflik Israel–Iran telah berlangsung sejak Revolusi Iran tahun 1979 dan melibatkan ketegangan geopolitik yang kompleks, khususnya terkait isu nuklir, militer, dan ideologis di kawasan Timur Tengah [1]. Eskalasi terkini, termasuk serangan Iran pada April 2024 dan respons Israel pada Oktober 2024, memicu diskusi global di media sosial yang mencerminkan dinamika opini publik internasional [2]. Media sosial berperan penting sebagai ruang diskursif, di mana masyarakat mengekspresikan pandangan politik, ideologis, dan emosional terhadap konflik bersenjata modern.

Platform seperti YouTube dan X (sebelumnya Twitter) menjadi kanal utama bagi masyarakat untuk menyampaikan reaksi spontan terhadap konflik Israel–Iran. Diskursus tersebut kerap dipengaruhi oleh faktor religius dan politik domestik, yang memperkuat narasi solidaritas terhadap Palestina dan Iran [3]. Konflik ini telah menyebabkan ratusan korban sipil di kedua belah pihak serta kerusakan infrastruktur strategis, termasuk fasilitas nuklir Iran. Hingga Desember 2025, ketegangan dilaporkan tetap tinggi, dengan indikasi bahwa Iran mempersiapkan kemungkinan perluasan konflik regional apabila Israel kembali melakukan serangan terhadap Hezbollah. Dalam konteks Indonesia, sentimen publik yang berkembang di media sosial menunjukkan kecenderungan kuat anti-Israel, sering kali dikaitkan dengan isu kemanusiaan dan pengalaman konflik lokal.

Analisis sentimen berbasis pemrosesan bahasa alami (Natural Language Processing/NLP) merupakan pendekatan efektif untuk mengukur opini publik dalam skala besar. Namun, model deep learning seperti Bidirectional Encoder Representations from Transformers (BERT) bersifat black-box, sehingga sulit untuk diinterpretasikan secara transparan [4]. Penelitian sebelumnya menunjukkan bahwa BERT unggul dalam menangkap konteks dua arah dan nuansa semantik dalam teks, tetapi

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

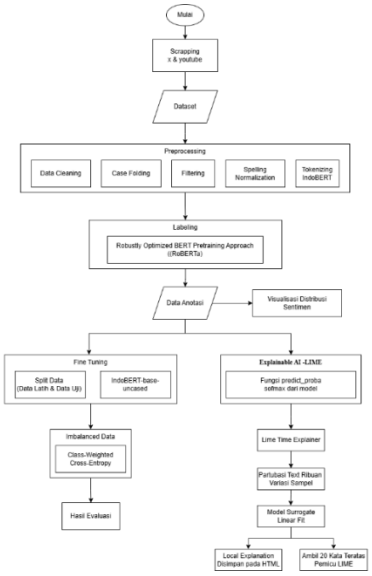
kekurangan aspek keterjelasan dalam proses pengambilan keputusan model [4].
Pendekatan Explainable Artificial Intelligence (XAI), seperti Local Interpretable Model-agnostic Explanations (LIME) [5] dan SHapley Additive exPlanations (SHAP) [6], dikembangkan untuk mengatasi keterbatasan tersebut dengan memberikan penjelasan lokal terhadap prediksi model. Meski demikian, integrasi antara IndoBERT—adaptasi BERT untuk bahasa Indonesia—dan metode XAI, khususnya LIME, masih jarang diterapkan dalam konteks konflik geopolitik internasional. Hal ini menjadi celah penelitian, terutama pada analisis opini publik Indonesia yang bersifat polar antara kelompok pro-Palestina/Iran dan pro-Israel [7].

Penelitian ini bertujuan untuk: (1) mengklasifikasikan sentimen masyarakat Indonesia terhadap konflik Israel–Iran berdasarkan data media sosial, (2) mengevaluasi kinerja model IndoBERT dalam tugas klasifikasi sentimen, dan (3) menginterpretasikan hasil prediksi menggunakan LIME guna memperoleh wawasan yang transparan dan dapat dipertanggungjawabkan. Hipotesis penelitian ini adalah bahwa sentimen publik Indonesia didominasi oleh sentimen negatif terhadap Israel, dengan faktor religius berperan signifikan dalam membentuk polarisasi opini.

Penelitian ini berkontribusi pada pengembangan metodologi NLP yang transparan untuk analisis opini geopolitik serta memberikan landasan empiris bagi pembuat kebijakan dalam merespons dinamika opini publik berbasis media sosial [8]. Studi sebelumnya terkait konflik Israel–Palestina menunjukkan dominasi sentimen negatif menggunakan pendekatan CNN pada data Twitter [9][10]. Selain itu, pendekatan hybrid BERT–VADER terbukti efektif dalam mendeteksi polarisasi opini pada platform Reddit [11]. Dalam konteks Indonesia, IndoBERT telah berhasil digunakan untuk analisis sentimen Pemilu 2024 dengan kemampuan menangkap nuansa bahasa lokal [12][13]. Integrasi XAI seperti LIME menjadi aspek krusial untuk meningkatkan transparansi dan kepercayaan terhadap hasil analisis sentimen [14]. Temuan penelitian ini juga selaras dengan survei yang menunjukkan dukungan kuat masyarakat Indonesia terhadap Palestina.

2. Metodologi Penelitian

Penelitian ini mengadopsi pendekatan kuantitatif dengan menggunakan analisis sentimen berbasis kecerdasan buatan untuk mengukur dan mengklasifikasikan sentimen publik terhadap konflik *Israel-Iran* yang tercermin di media sosial. Pendekatan kuantitatif dipilih karena memungkinkan pengolahan data besar dan memberikan hasil yang dapat diukur dan dianalisis secara statistik. Dalam penelitian ini, data yang dikumpulkan berupa teks dari platform media sosial seperti Twitter yang akan dianalisis untuk mengetahui perasaan dan opini masyarakat terhadap berbagai peristiwa besar dalam konflik Israel-Iran. Tahapan metode penelitian ditunjukkan pada Gambar 1.



JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

Gambar 1. Metode penelitian

Penelitian ini dimulai dengan proses pengumpulan data melalui teknik scraping pada dua platform media sosial, yaitu X (*Twitter*) dan *YouTube*. Data yang dikumpulkan berupa komentar dan teks relevan yang mencerminkan opini atau respon pengguna terhadap topik tertentu. Seluruh data yang berhasil diperoleh kemudian dihimpun dalam satu dataset sebagai fondasi utama untuk proses analisis selanjutnya. Setelah dataset terkumpul, dilakukan tahap preprocessing yang bertujuan meningkatkan kualitas dan konsistensi data. Proses ini mencakup pembersihan teks dari elemen yang tidak relevan seperti *URL*, *emoji*, *tag HTML*, serta karakter khusus lain yang dapat mengganggu proses pemodelan. Selain itu, dilakukan *case folding* untuk menyeragamkan seluruh huruf menjadi *lowercase*, *filtering* untuk menghilangkan kata-kata umum yang tidak memberikan kontribusi informasi, normalisasi ejaan untuk menyempurnakan penulisan kata tidak baku, dan tokenisasi agar teks dapat dipecah menjadi unit-unit yang siap diproses oleh model bahasa.

Setelah data siap, tahap berikutnya adalah pelabelan sentimen menggunakan model *RoBERTa* (*Robustly Optimized BERT Pretraining Approach*). *RoBERTa* dipilih karena memiliki sejumlah keunggulan signifikan dibandingkan model sejenis. Model ini dilatih dengan jumlah data yang lebih besar, tanpa penggunaan masking statis seperti pada *BERT*, serta menggunakan dinamika hyperparameter yang telah dioptimalkan secara lebih agresif. Keunggulan tersebut membuat *RoBERTa* lebih stabil, lebih akurat, dan lebih sensitif terhadap konteks pada berbagai jenis teks. Pada penelitian ini, *RoBERTa* berperan melakukan pelabelan awal secara otomatis terhadap data mentah. Jika diperlukan, anotasi manual juga dilakukan untuk menjamin konsistensi dan akurasi label, terutama pada kasus-kasus ambigu. Hasil pelabelan kemudian dianalisis melalui visualisasi distribusi sentimen guna melihat proporsi data positif, negatif, dan netral serta mendeteksi adanya ketidakseimbangan kelas.
Tahap berikutnya adalah fine-tuning terhadap model I*ndoBERT-base-uncased*, yaitu model *BERT* yang dirancang khusus untuk bahasa Indonesia. Pemilihan IndoBERT didasari oleh keunggulannya dalam memahami struktur morfologi, gaya bahasa, dan konteks khas bahasa Indonesia yang sulit ditangani oleh model multibahasa. Selain itu, IndoBERT dilatih dengan dataset yang sangat luas untuk bahasa Indonesia sehingga mampu menghasilkan representasi semantik yang lebih tajam dan akurat. Proses fine-tuning dilakukan setelah dataset dibagi menjadi data latih dan data uji. Mengingat distribusi data sering kali tidak seimbang, penelitian ini menerapkan metode *class-weighted cross-entropy* untuk memberi bobot tambahan pada kelas minoritas. Langkah ini penting untuk menghindari bias model terhadap kelas yang jumlah datanya dominan.

Setelah *fine-tuning* selesai, model dievaluasi menggunakan metrik seperti akurasi, *precision*, *recall*, dan *F1-score* untuk menilai performa prediksi secara menyeluruh. Tahap selanjutnya adalah penerapan *Explainable AI (XAI)* menggunakan metode *Local Interpretable Model-agnostic Explanations* (*LIME*). *LIME* digunakan untuk menjelaskan alasan di balik prediksi model dengan mengidentifikasi kata atau fitur yang memiliki kontribusi terbesar. Penggunaan *LIME* memberikan nilai tambah berupa transparansi model, sehingga prediksi tidak hanya akurat tetapi juga dapat dipertanggungjawabkan dan mudah dipahami. Hasil interpretasi ini kemudian dianalisis kembali untuk memastikan bahwa keputusan model selaras dengan konteks yang sebenarnya.

Rangkaian proses ini diakhiri dengan penarikan kesimpulan serta rekomendasi untuk penelitian selanjutnya. Dengan penggunaan kombinasi *RoBERTa* untuk pelabelan awal dan *IndoBERT* untuk *fine-tuning*, penelitian ini memanfaatkan dua model berbasis transformer yang terbukti unggul dalam pemahaman konteks, generalisasi, serta performa analisis sentimen pada teks berbahasa Indonesia.

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

3. Hasil dan Pembahasan

3.1 Scrapping Data

Pengumpulan data primer dalam penelitian ini dilakukan melalui teknik scraping pada dua platform media sosial utama yang banyak digunakan oleh publik Indonesia, yaitu YouTube dan X (dulunya Twitter). Pemilihan kedua platform ini didasarkan pada karakteristik interaksi penggunanya yang berbeda namun saling melengkapi: YouTube menyediakan komentar panjang, naratif, dan bersifat reflektif dengan konteks video berita, sedangkan X menawarkan opini singkat, spontan, dan sering kali viral melalui *hashtag* dan *retweet*. Kombinasi keduanya memungkinkan representasi yang lebih holistik terhadap dinamika sentimen publik Indonesia terhadap isu konflik Israel-Iran.

Semua data dari kedua platform digabungkan ke dalam satu *dataframe* dengan kolom terstandar: Penulis (dianonimkan), Komentar (teks asli), Jumlah_Suka, Tanggal, dan File_Sumber (untuk membedakan YouTube vs X). Total 7.922 entri unik berhasil dikumpulkan setelah proses *deduplikasi* berdasarkan kemiripan teks menggunakan *Levenshtein distance* (< 90%) dan penghapusan komentar bot (panjang < 10 karakter atau berisi URL promosi).

Tabel 1. Ringkasan Hasil Pengumpulan Data

Platform	Sumber Akun/Media	Video/Tweet Awal	Data Unik
YouTube	CNN Indonesia, Kompas TV, dll	12 video (7.512 komentar)	4.512 komentar
X (*Twitter*)	Pencarian kata kunci	5.000 tweet	3.410 tweet
Total		12.512 entri	7.922 entri

3.2 Preprocessing Data

Tahapan utama preprocessing yang diterapkan dalam penelitian ini meliputi:

1. *Case Folding*
   Seluruh teks diubah menjadi huruf kecil (*lowercase*) untuk menghilangkan sensitivitas terhadap kapitalisasi. Contoh: *"ISRAEL ZIONIS" → "israel zionis"*.
2. Pembersihan Elemen Non-Teks
   Tahap ini bertujuan untuk menghilangkan elemen yang tidak berkontribusi terhadap makna sentimen, dengan rincian sebagai berikut:
   a) URL dihapus menggunakan pola *regular expression* r'https?://\S+|www\.\S+'. Contoh: *"Lihat di sini: https://t.co/abc123" "Lihat di sini"*.
   b) Mention (@username) dihapus menggunakan pola r'@\w+'.
   c) Hashtag diproses dengan menghapus simbol # namun mempertahankan kata kunci di belakangnya untuk menjaga konteks analisis. Contoh: *#IsraelIran → israeliran*.
   d) Emoji dikonversi menjadi representasi teks menggunakan pustaka emoji.demojize() untuk mempertahankan muatan emosional dalam bentuk token. Contoh: *"Kasian Palestina 🤲" → "kasian palestina wajah berdoa"*.
3. Normalisasi *Slang* dan Singkatan
   Normalisasi dilakukan menggunakan kamus slang bahasa Indonesia khusus yang berisi sekitar 1.250 entri, dikompilasi dari dataset publik dan sumber daring. Contoh normalisasi meliputi:
   a) *"gak" → "tidak"*
   b) *"bgt" → "banget"*
   c) *"zionis"* → dipertahankan sebagai kata kunci domain
   Proses normalisasi diterapkan melalui fungsi *replacement* berbasis token untuk meminimalkan kesalahan substitusi (*false positive*).
4. Pembersihan Simbol Berlebih dan Spasi
   a) Tanda baca berulang (misalnya *"!!!!"* atau *"..."*) direduksi menjadi satu karakter.
   b) Spasi berlebih dihapus dan dinormalisasi menjadi satu spasi.
   c) Karakter non-alfanumerik dihapus, kecuali tanda hubung yang merupakan bagian dari istilah majemuk (misalnya *Israel-Iran*).
5. Penghapusan Teks Tidak Informatif

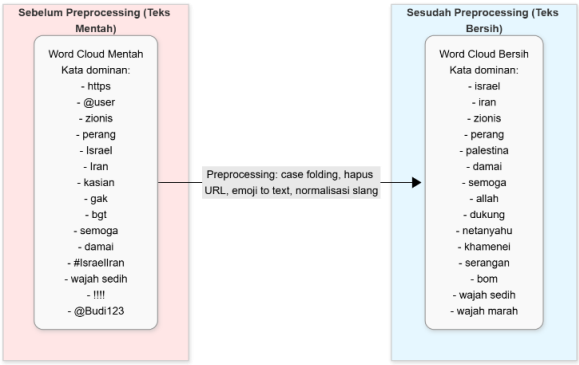
JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

Komentar dengan panjang kurang dari 10 karakter atau yang hanya berisi angka dan simbol dihapus karena dianggap tidak mengandung informasi sentimen yang bermakna.

Hasil akhir *preprocessing* disimpan dalam berkas komentar_clean_final.csv dengan penambahan kolom Komentar_Clean, yang berisi teks bersih siap digunakan pada tahap anotasi sentimen. Setelah seluruh tahapan preprocessing dan penyaringan diterapkan, diperoleh 7.562 komentar bersih. Sebagai bentuk kontrol kualitas, dilakukan evaluasi sebelum dan sesudah preprocessing melalui perbandingan sampel acak serta visualisasi word cloud untuk memastikan bahwa kata-kata bermakna tetap dipertahankan sementara noise berhasil diminimalkan.

Tabel 2 Hasil transformasi preprocesing

Teks Mentah	Teks Bersih
@Budi123 Israel ZIONIS banget!!!! Lihat: https://t.co/xyz #PerangDunia3	israel zionis banget lihat perangdunia
Ya Allah semoga damai 🤲	ya allah semoga damai wajah berdoa



Gambar 2 Perbandingan Word Cloud Sebelum dan Sesudah Preprocessing

Gambar 2 menunjukkan perbandingan visualisasi word cloud sebelum dan sesudah tahap preprocessing. Pada teks mentah, kata-kata yang tidak bermakna secara semantik seperti URL, mention, hashtag mentah, serta variasi slang masih mendominasi. Setelah preprocessing, kata-kata dominan menjadi lebih relevan secara kontekstual terhadap isu konflik Israel–Iran, yang menunjukkan bahwa proses pembersihan teks berhasil menghilangkan noise tanpa menghilangkan informasi penting.

3.3 Anotasi Data Sentimen

Proses anotasi data dalam penelitian ini dilakukan secara otomatis menggunakan model pre-trained berbasis transformer yang telah di fine-tune khusus untuk tugas klasifikasi sentimen bahasa Indonesia, yaitu w11wo/*indonesian-roberta-base-sentiment-classifier*. Model berbasis *RoBERTa* ini digunakan sebagai anotator otomatis awal (*weak supervision*) untuk menghasilkan label sentimen pada dataset yang belum memiliki anotasi manual.

Seluruh probabilitas keluaran untuk ketiga kelas sentimen (P_NEG, P_NET, dan P_POS) disimpan sebagai fitur tambahan guna mendukung analisis tingkat kepastian prediksi. Proses anotasi dilakukan secara batch dengan ukuran batch sebesar 64 dan memanfaatkan akselerasi *GPU* untuk efisiensi komputasi. Penerapan *preprocessing* lanjutan, khususnya normalisasi slang menggunakan kamus slang, turut membantu meningkatkan konsistensi anotasi dengan mengurangi variasi ejaan non-baku pada teks media sosial.

Tabel 3. Distribusi data sentimen publik konflik israel dan iran

Sentimen		Jumlah	Persentase (%)
NEGATIF		3.045	40,3
NETRAL		2.568	34,0
POSITIF		1.949	25,7
Total		7.562	100,0

3.4 Pemodelan Sentimen dengan Finetuning IndoBERT

Data yang digunakan pada tahap pemodelan berasal dari 7.562 komentar bersih hasil preprocessing dan anotasi otomatis. Setelah dilakukan pembersihan lanjutan (penghapusan nilai kosong, label tidak valid, dan duplikasi teks), seluruh data dinyatakan valid dan digunakan dalam proses

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

pemodelan.Pembagian data dilakukan menggunakan strategi *stratified split* dengan rasio 80:20 untuk memastikan proporsi kelas sentimen tetap terjaga pada data latih dan data uji. Pendekatan ini penting untuk meminimalkan bias akibat ketidakseimbangan kelas, khususnya pada kelas POSITIF yang memiliki frekuensi relatif lebih rendah.

Pembagian data dilakukan dengan strategi stratified split (rasio 80:20) untuk memastikan proporsi kelas tetap terjaga pada kedua subset. Hal ini penting untuk mencegah *bias* akibat distribusi kelas yang tidak seimbang, terutama pada kelas POSITIF yang memiliki frekuensi terendah**.**Distribusi data latih ditunjukkan pada tabel 4.

Tabel 4. Distribusi pembagian data latih dan data ujirsitekturr dan konfigurasi model

Dataset	Negatif	Netral	Positif	Total
Latih (80%)	2.426	2.062	1.561	6.049
Uji (20%)	619	506	388	1.513
Total	3.045	2.568	1.949	7.562

Model dasar yang digunakan adalah *IndoBERT-base-uncased*, sebuah model transformer dengan 12 lapisan *encoder* yang dilatih pada korpus *IndoLEM* (Wikipedia, berita, dan teks web berbahasa Indonesia). Spesifikasi model ditunjukkan pada Tabel 5.

Tabel 5. Arsitektur Model IndoBERT

Komponen	Spesifikasi
Arsitektur	12-layer Transformer
Dimensi Tersembunyi	768
Jumlah Parameter	110 juta
Korpus Pelatihan	IndoLEM (Wikipedia, Berita, Web)

Pada tahap *finetuning*, lapis klasifikasi teratas (*classification head*) diganti dengan lapisan baru yang terdiri dari 3 neuron output sesuai jumlah kelas sentimen (NEGATIF, NETRAL, POSITIF). Proses pelatihan dilakukan dengan konfigurasi hiperparameter pada table 6.

Tabel 6. Distribusi data sentimen publik konflik israel dan iran

Parameter	Nilai
Max sequence length	256 token
Batch size	16
Learning rate	5×10^{-5}
Optimizer	AdamW
Epoch maksimum	10
Weight decay	0,01
Warmup steps	150
Learning rate scheduler	Cosine with restarts
Mixed precision (FP16)	Aktif
Early stopping	Aktif
Class-weighted loss	Aktif

3.4.1 Penanganan Ketidakseimbangan Kelas

Untuk mengatasi ketidakseimbangan kelas, diterapkan *class-weighted cross-entropy loss* dengan bobot yang dihitung menggunakan metode *balanced*:

Tabel 7. Bobot kelas

Sentimen	Bobot
NEGATIF	0.83
NETRAL	0.98
POSITIF	1.30

Bobot ini memberikan penalti lebih besar pada kesalahan prediksi kelas minoritas (POSITIF), sehingga mendorong model untuk lebih memperhatikan pola-pola langka.

3.4.2 Pelatihan

Proses pelatihan model dilakukan menggunakan *framework Hugging Face Transformers* (Trainer API) pada lingkungan Google Colab Pro dengan dukungan akselerasi GPU NVIDIA T4. Model IndoBERT dilatih menggunakan data latih hasil anotasi otomatis.

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

Tabel 8. Performa model pada validatio set epoch

Epoch	Training Loss	Validation Loss	Akurasi	F1-Score Tertimbang
1	0,7131	0,7076	0,7074	0,7083
2	0,6457	0,5778	0,7537	0,7522
3	0,4552	0,6829	0,7603	0,7578
4	0,3277	0,8076	0,7471	0,7495

Hasil pelatihan menunjukkan bahwa *validation loss* mencapai nilai minimum pada *epoch* ke-2, kemudian meningkat pada *epoch* ke-3 dan ke-4. Pola ini mengindikasikan bahwa model mulai mengalami kecenderungan overfitting setelah *epoch* ke-2, meskipun *training loss* terus menurun. Oleh karena itu, mekanisme *early stopping* diaktifkan, dan model terbaik dipilih berdasarkan nilai *validation loss* terendah, yaitu pada *checkpoint epoch* ke-2.

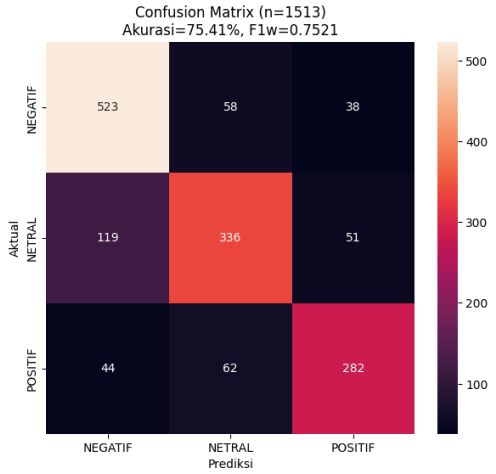
3.4.3 Validasi

Model dievaluasi menggunakan metrik akurasi, *F1-score* (*weighted F1*), serta nilai *loss* pada data uji. Penggunaan F1-score tertimbang dipilih untuk memberikan gambaran performa yang lebih representatif mengingat distribusi kelas sentimen yang tidak sepenuhnya seimbang.

Tabel 9. Performa model indobert pada data uji

Model	Akurasi	F1-Score
IndoBERT Finetuned (GPU + Class-Weighted Loss)	**75.41%**	0.7521

Evaluasi lebih lanjut terhadap performa model *IndoBERT* hasil *finetuning* dilakukan menggunakan confusion matrix pada data uji independen yang tidak pernah digunakan selama proses pelatihan maupun validasi (n = 1.513).



Gambar 3 Heatmap Confusion Matrix Model IndoBERT Hasil Finetuning pada Data Uji (n = 1.513)

Untuk melengkapi visualisasi heatmap pada gambar 3, Gambar 4 menyajikan confusion matrix dalam bentuk numerik beserta ringkasan metrik klasifikasi berupa *precision*, *recall*, dan *F1-score* pada masing-masing kelas sentimen.

```
Accuracy: 0.7541308658294779 F1 weighted: 0.7521007812300369
              precision    recall  f1-score   support

     NEGATIF       0.76      0.84      0.80       619
      NETRAL       0.74      0.66      0.70       506
     POSITIF       0.76      0.73      0.74       388

    accuracy                           0.75      1513
   macro avg       0.75      0.75      0.75      1513
weighted avg       0.75      0.75      0.75      1513

array([[523,  58,  38],
       [119, 336,  51],
       [ 44,  62, 282]])
```

Gambar 4 Confusion Matrix Numerik dan Classification Report Model IndoBERT pada Data Uji

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

3.5 Interpretasi Model Sentimen dengan LIME

Interpretasi dilakukan pada beberapa komentar representatif dari data uji, dengan parameter num_features antara 10 hingga 20 untuk menampilkan token yang paling berpengaruh.

Berdasarkan Tabel X, dapat diamati bahwa pada kelas NEGATIF, token-token yang dominan berkaitan dengan isu politik, konflik, dan ideologi, seperti *freedom*, *democracy*, *usa*, *kills*, *israhell*, dan *iran*. Meskipun secara leksikal kata seperti *freedom* dan *democracy* memiliki konotasi positif atau netral, dalam konteks diskursus konflik Israel–Iran kata-kata tersebut sering digunakan secara sarkastik atau ironis untuk menyampaikan kritik. Hal ini menyebabkan token-token tersebut berkontribusi kuat terhadap prediksi sentimen negatif.

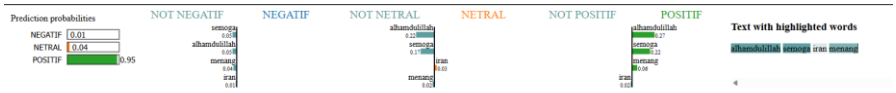
Pada kelas NETRAL, token seperti *israel*, *law*, *rights*, dan *civilians* mendominasi. Token-token ini umumnya muncul dalam komentar yang bersifat informatif, faktual, atau deskriptif, misalnya penyampaian berita, rujukan hukum internasional, atau penjelasan kondisi sipil, tanpa disertai evaluasi emosional yang eksplisit. Temuan ini menunjukkan bahwa model cenderung mengasosiasikan terminologi institusional dan faktual dengan sentimen netral.

Sementara itu, pada kelas POSITIF, token seperti *thank*, *support*, *rights*, dan *freedom* menjadi pemicu utama prediksi positif. Token-token tersebut umumnya muncul dalam konteks ekspresi dukungan politik, solidaritas, atau harapan terhadap pihak tertentu. Meskipun frekuensi token positif relatif lebih rendah dibandingkan kelas negatif, pola kontribusinya tetap konsisten dan menunjukkan kemampuan model dalam mengenali ekspresi afirmatif.

Secara keseluruhan, agregasi hasil LIME menunjukkan bahwa model IndoBERT tidak hanya mempelajari polaritas kata secara leksikal, tetapi juga mampu menangkap konteks penggunaan kata dalam diskursus konflik geopolitik. Token yang sama dapat berkontribusi pada kelas sentimen yang berbeda bergantung pada konteks kalimatnya, yang menegaskan kompleksitas analisis sentimen pada topik sensitif berbasis opini publik.

Setiap hasil penjelasan LIME disimpan dalam bentuk file HTML interaktif, sehingga memungkinkan eksplorasi visual terhadap kontribusi kata secara langsung.

1. Komentar bernuansa religious
   Pada komentar yang memuat ekspresi religius, model cenderung mengasosiasikannya sebagai sinyal afirmasi/dukungan emosional sehingga meningkatkan probabilitas prediksi ke kelas POSITIF



Gambar 5 Hasil Penjelasan LIME untuk Komentar Bernuansa Religius (“alhamdulillah semoga iran menang”)

2. Komentar sarkastik atau ejekan
   Komentar yang mengandung istilah seperti *israhell* atau frasa ironi terkait *freedom* dan *democracy* cenderung didorong kuat ke kelas NEGATIF. LIME memperlihatkan bahwa token-token tersebut memiliki bobot besar dalam mempengaruhi prediksi

Gambar 6 Hasil Penjelasan LIME untuk Komentar Sarkastik (“zionis penjajah kok masih dibela”)

3. Koment Pada komentar yang hanya berisi Beberapa token umum, LIME menunjukkan bahwa prediksi sering kali didorong oleh kata-kata yang relatif netral seperti *possible* atau *according*.ar sangat pendek atau generik

   Pada komentar yang hanya berisi beberapa token umum, LIME menunjukkan bahwa prediksi sering kali didorong oleh kata-kata yang relatif netral seperti *possible* atau *according*.

4. Kesimpulan

Penelitian ini berhasil mengimplementasi dan mengevaluasi sistem analisis sentimen untuk memetakan opini publik Indonesia terhadap konflik Israel–Iran dengan memanfaatkan model IndoBERT yang difinetune serta pendekatan *Explainable AI* menggunakan LIME. Rangkaian proses penelitian meliputi pengumpulan data dari dua platform (YouTube dan X), *preprocessing* lanjutan (termasuk normalisasi slang), anotasi otomatis berbasis *weak supervision* menggunakan RoBERTa, *fine-tuning IndoBERT*, evaluasi kuantitatif, serta interpretasi keputusan model. Temuan utama penelitian dapat dirangkum sebagai berikut.

1. Data dikumpulkan melalui web scraping dari YouTube dan X sehingga diperoleh 7.922 entri unik. Setelah pipeline preprocessing diterapkan secara konsisten (case folding, pembersihan URL/mention/hashtag, konversi emoji, normalisasi slang dengan kamus ±1.250 entri, reduksi simbol, serta penyaringan teks tidak informatif), dataset menghasilkan 7.562 komentar bersih yang siap dianalisis. Distribusi hasil anotasi otomatis menunjukkan dominasi sentimen NEGATIF (40,3%), diikuti NETRAL (34,0%) dan POSITIF (25,7%), yang menggambarkan kuatnya respons kritik/emosi negatif publik sekaligus tingginya komentar faktual/ambigu yang ditangani melalui kebijakan anotasi konservatif.
2. Pendekatan *weak supervision* dapat digunakan untuk membangun data latih sentimen pada domain spesifik. Anotasi sentimen dilakukan menggunakan w11wo/indonesian-roberta-base-sentiment-classifier sebagai anotator otomatis, dengan kebijakan konservatif berbasis ambang probabilitas ($P_{max} > 0{,}60$ dan $|P\_POS - P\_NEG| > 0{,}15$) sehingga komentar yang ambigu dipetakan ke kelas NETRAL. Dalam rancangan penelitian ini, RoBERTa berperan sebagai anotator otomatis, sedangkan IndoBERT adalah model utama yang dilatih menggunakan label hasil anotasi tersebut.
3. Ketiga, fine-tuning IndoBERT menghasilkan performa yang stabil dan dapat dipertanggungjawabkan pada data uji independen. IndoBERT-base-uncased difinetune menggunakan 7.562 data dengan pembagian stratified split 80:20, menghasilkan data latih 6.049 dan data uji 1.513. Proses pelatihan dilakukan dengan konfigurasi utama: max length 256 token, batch size 16, learning rate 5×10^{-5}, optimizer AdamW, weight decay 0,01, warmup 150, cosine with restarts, mixed precision (FP16), serta early stopping. Untuk mengatasi ketidakseimbangan kelas, diterapkan class-weighted cross-entropy loss dengan bobot [NEGATIF=0,83; NETRAL=0,98; POSITIF=1,30]. Model terbaik dipilih berdasarkan validation loss minimum pada epoch ke-2. Evaluasi pada data uji menghasilkan akurasi 75,41% dan F1-score tertimbang 0,7521, yang menunjukkan kemampuan generalisasi yang memadai pada data yang tidak pernah dilihat saat pelatihan.
4. Analisis kesalahan memperlihatkan tantangan utama berada pada pembedaan kelas NETRAL dengan kelas lain. Confusion matrix menunjukkan prediksi benar pada diagonal utama (NEGATIF=523; NETRAL=336; POSITIF=282), namun terdapat pola kesalahan dominan NETRAL→NEGATIF (119 kasus) yang menandakan kecenderungan model mengasosiasikan kosakata konflik deskriptif sebagai sinyal negatif. Selain itu, kesalahan POSITIF→NETRAL (62 kasus) menunjukkan bahwa dukungan yang implisit (misalnya doa/harapan tanpa kata evaluatif kuat) sering kali belum cukup kuat dikenali sebagai positif. Performa kelas menunjukkan recall tertinggi pada NEGATIF (0,84) dan terendah pada NETRAL (0,66), sehingga aspek ambiguitas dan konteks ganda tetap menjadi batas utama performa model pada domain konflik geopolitik.
5. Kelima, penerapan LIME berhasil meningkatkan transparansi dan akuntabilitas model. LIME menunjukkan bahwa prediksi IndoBERT dipengaruhi token yang relevan secara linguistik dan kontekstual. Secara global, agregasi kata pemicu mengindikasikan bahwa kelas NEGATIF banyak dipengaruhi token bertema konflik/ideologi (misalnya kills, israhell, iran serta kata bernuansa sarkastik seperti freedom/democracy), kelas NETRAL didominasi istilah faktual (israel, law, rights, civilians), sedangkan kelas POSITIF dipicu token dukungan/afirmasi (thank, support, alhamdulillah). Secara lokal, studi kasus menunjukkan: (1) komentar bernuansa religius dapat mendorong prediksi POSITIF secara kuat, (2) komentar bernada kritik/ejekan dengan istilah

ideologis mendorong NEGATIF, dan (3) komentar sangat pendek seperti "menurut saya" cenderung menghasilkan prediksi dengan keyakinan lebih rendah karena minim konteks. Dengan demikian, LIME tidak hanya menjadi alat interpretasi teknis, tetapi juga mendukung analisis sosial-linguistik atas dinamika opini publik.

Daftar Pustaka

[1] S. Bazai, A. Rahman, and M. Khalid, "Geopolitical conflicts in the Middle East and public opinion dynamics: The Israel–Iran rivalry," Journal of International Relations and Security Studies, vol. 8, no. 2, pp. 115–132, 2023.
[2] A. Jamison, D. Karpf, and D. Kreiss, "Social media, global crises, and public discourse in times of war," New Media & Society, vol. 26, no. 3, pp. 987–1004, 2024.
[3] K. Joseph, "Religious narratives and political polarization in social media discussions of Middle East conflicts," Journal of Media and Religion, vol. 23, no. 1, pp. 22–39, 2024.
[4] A. Liaqat, K. Dashtipour, and A. Hussain, "Explainability in deep learning-based sentiment analysis: A survey," ACM Computing Surveys, vol. 55, no. 4, pp. 1–38, 2022.
[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.
[7] P. Ranjan, R. Singh, and S. Verma, "Polarized opinion mining in geopolitical conflicts using transformer models," Information Processing & Management, vol. 60, no. 4, 2023.
[8] E. Nabankema, "Artificial intelligence for policy analysis: Opportunities and challenges in public opinion mining," Policy & Internet, vol. 16, no. 1, pp. 89–104, 2024.
[9] A. Fuadi, R. Pratama, and A. Hidayat, "Twitter sentiment analysis on the Israel–Palestine conflict using convolutional neural networks," Journal of Information Systems Engineering and Business Intelligence, vol. 9, no. 1, pp. 45–56, 2023.
[10] A. Nurlatifah, M. Sari, and Y. Kurniawan, "Public sentiment analysis on the Israel–Palestine conflict using deep learning approaches," International Journal of Advanced Computer Science and Applications, vol. 14, no. 5, pp. 312–320, 2023.
[11] M. A. Keerio, H. Xu, and S. Ali, "Hybrid BERT–VADER approach for stance detection on Reddit discussions," Expert Systems with Applications, vol. 213, 2023.
[12] R. Geni, A. Nugroho, and D. P. Putra, "Sentiment analysis of Indonesia's 2024 general election using IndoBERT," Indonesian Journal of Artificial Intelligence and Data Science, vol. 6, no. 1, pp. 1–12, 2024.
[13] S. Yulfa, F. Ramadhan, and D. Lestari, "Evaluating IndoBERT for Indonesian political sentiment classification," Journal of Big Data Analytics in Politics, vol. 3, no. 2, pp. 55–68, 2024.
[14] V. Ranjbar, "Explainable artificial intelligence for sentiment analysis: A systematic review," Artificial Intelligence Review, vol. 56, pp. 2145–2173, 2023