Journal Software, Hardware and Information Technology
p-ISSN: 2808-3385 | e-ISSN: 2776-8961
Volume 6 Issue 1, January 2026

Penerapan *Explainable AI* LIME pada Klasifikasi Sentimen Isu IKN, PPN, dan MBG Menggunakan IndoBERTweet

Application of Explainable AI (LIME) to Sentiment Classification of IKN, VAT, and MBG Issues Using IndoBERTweet

Luthfi Delvian, Esa Firmansyah, Beben Sutara
Jurusan Informatika, Universitas Sebelas April, Sumedang, Indonesia
220660121029@student.unsap.ac.id, esa@unsap.ac.id, beben@unsap.ac.id

Notes

Submitted: 21-11-2025
Revised: 17-12-2025
Accepted: 18-12-2025
Published: 30-01-2026

How to cite: L. Delvian, E. Firmansyah, and B. Sutara, "Penerapan Explainable AI LIME pada Klasifikasi Sentimen Isu IKN, PPN, dan MBG Menggunakan IndoBERTweet," *Journal Software, Hardware and Information Technology (SHIFT)*, vol. 6, no. 1, pp. 40–52, Jan. 2026, doi: 10.24252/shift.v6i1.236.

DOI: 10.24252/shift.v6i1.236

Copyright 2026© the Author(s)

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.




ABSTRACT

Social media, particularly the X platform (formerly Twitter), has become a primary space for the public to express their opinions on public policy. This study aimed to classify public sentiment towards three strategic issues in Indonesia: the relocation of the national capital (IKN), the 11% VAT increase policy, and the Free Nutritious Meals (MBG) program by applying the IndoBERTweet model combined with a LIME-based Explainable Artificial Intelligence (XAI) approach. A dataset of 2,845 tweets was obtained through scraping using relevant keywords. The sentiment distribution in the data reveals a dominance of negative sentiment, with 1,575 tweets, followed by positive sentiment, with 1,005 tweets, and neutral sentiment, with 265 tweets. The IndoBERTweet model, trained for three epochs, produced an accuracy of 89% and an average macro F1-score of 0.78. The highest performance was achieved in both the positive and negative classes, with an F1-score of 0.92, while neutral sentiment remained challenging, with an F1-score of 0.49. To improve prediction transparency, the LIME method was employed and successfully displayed consistent visualizations that identified the dominant words influencing the classification. This study demonstrates the potential application of explainable Indonesian language transformer models for sentiment analysis of national policies in the digital realm.

Keywords: explainable artificial intelligence, IndoBERTweet, LIME, sentiment, social media.

ABSTRAK

Media sosial, khususnya platform X (sebelumnya Twitter), menjadi ruang utama bagi masyarakat untuk menyuarakan opini terhadap kebijakan publik. Penelitian ini telah bertujuan untuk mengklasifikasikan sentimen publik terhadap tiga isu strategis di Indonesia pemindahan Ibu Kota Negara

(IKN), kebijakan kenaikan PPN 11%, dan program Makan Bergizi Gratis (MBG) dengan menerapkan model IndoBERTweet yang dipadukan dengan pendekatan Explainable Artificial Intelligence (XAI) berbasis LIME. Dataset sebanyak 2.845 tweet diperoleh melalui proses scraping dengan kata kunci yang relevan. Distribusi sentimen dalam data menunjukkan dominasi sentimen negatif sebanyak 1.575 tweet, disusul sentimen positif sebanyak 1.005 tweet, dan netral sebanyak 265 tweet. Model IndoBERTweet yang dilatih selama tiga epoch menghasilkan akurasi sebesar 89% dan nilai F1-score rata-rata makro sebesar 0,78. Performa tertinggi dicapai pada kelas positif dan negatif dengan F1-score sebesar 0,92, sementara sentimen netral masih menantang dengan nilai F1 sebesar 0,49. Untuk meningkatkan transparansi prediksi, metode LIME digunakan dan berhasil menampilkan visualisasi yang konsisten dalam mengidentifikasi kata-kata dominan yang memengaruhi klasifikasi. Penelitian ini menunjukkan potensi penggunaan model transformer bahasa Indonesia yang dapat dijelaskan untuk analisis sentimen terhadap kebijakan nasional di ruang digital.

Kata kunci: explainable artificial intelligence, IndoBERTweet, LIME, media sosial, sentimen.

1. PENDAHULUAN

Media sosial telah menjadi bagian tak terpisahkan dari kehidupan masyarakat modern, khususnya dalam menyampaikan opini terhadap isu-isu kebijakan publik. Platform *X* merupakan salah satu media sosial yang populer digunakan untuk menyuarakan pendapat secara terbuka, cepat, dan masif [1]. Fenomena ini menciptakan ruang digital yang dinamis, di mana berbagai topik strategis nasional dibahas dan dikomentari oleh masyarakat dalam skala besar.

Dalam konteks Indonesia, beberapa kebijakan pemerintah seperti pemindahan Ibu Kota Negara (IKN), kenaikan Pajak Pertambahan Nilai (PPN) menjadi 11%, serta program Makan Bergizi Gratis (MBG) telah menjadi bahan perbincangan luas di media sosial. Ketiga isu ini menimbulkan reaksi publik yang beragam dan polaristik, mulai dari dukungan hingga penolakan [2], [3], [4]. Program Makan Bergizi Gratis (MBG) memiliki urgensi sosial yang tinggi karena berperan langsung dalam peningkatan status gizi dan prestasi belajar anak sekolah, sebagaimana ditunjukkan oleh studi global dan nasional mengenai efektivitas program makan gratis dalam memperbaiki ketahanan pangan dan pemerataan nutrisi [5]. Oleh karena itu, analisis terhadap opini publik menjadi penting sebagai bahan evaluasi terhadap penerimaan kebijakan di masyarakat.

Salah satu pendekatan yang dapat digunakan untuk memahami opini publik secara masif adalah analisis sentimen berbasis *machine learning*. Beberapa studi sebelumnya telah menerapkan algoritma seperti Naive Bayes, Support Vector Machine, hingga decision tree seperti C4.5 untuk klasifikasi sentimen pada isu-isu kebijakan [2], [6], [7], [8]. Namun, sebagian besar dari pendekatan tersebut masih menggunakan metode konvensional dan belum mengadopsi pendekatan berbasis transformer maupun *Explainable AI (XAI)* secara optimal. Studi-studi terkini

menunjukkan bahwa algoritma modern berbasis transformer seperti BERT [9], IndoBERT [10], dan IndoBERTweet [11], memiliki kemampuan yang jauh lebih unggul dalam memahami konteks semantik dan sintaksis bahasa alami dibandingkan metode klasik. Pendekatan ini terbukti fmenghasilkan peningkatan signifikan dalam tugas klasifikasi teks berbahasa Indonesia, terutama pada domain media sosial yang bersifat informal dan kontekstual [12], [13].

Tabel 1. Ringkasan Referensi Penelitian Terdahulu

No	Penulis	Topik	Metode	Dataset	Kelemahan	Ref
1	Hanin A, Maryam M	MBG	C4.5	1.000 tweet	Tidak menggunakan XAI	[2]
2	Hermansyah I, Hasibuan M	MBG	Linear Regression, Random Forest	1.200 tweet	Belum gunakan deep learning	[6]
3	Farhan Saleh M, Imanda R	MBG	Naive Bayes, SVM	1.000 tweet	Tidak menampilkan interpretasi model	[7]
4	Fauziah N	PPN 11%	VADER	1.500 tweet	Tanpa validasi berbasis transformer	[14]
5	Koto F, Lau J, Baldwin T	IndoBERTweet	IndoBERTweet (Transformer)	Multi-label tweet	Fokus ke akurasi, tanpa interpretasi	[11]
6	Sawant t	XAI (LIME)	LIME	—	Belum dikaitkan dengan isu sosial	[15]
7	Putra D, Kurniawan A	Media Sosial (Twitter)	IndoBERT (Transformer)	3.000 tweet		[10]
8	Herlambang S, Lestari F, Wibawa T	NLP (Low-Resource Bahasa Indonesia)	mBERT, IndoBERTweet, NusaBERT	2 korpus public	Fokus hanya pada perbandingan akurasi antar-model	[12]

Melalui tabel tersebut terlihat bahwa meskipun telah banyak penelitian mengenai analisis sentimen terhadap isu-isu kebijakan, sebagian besar masih menggunakan pendekatan konvensional atau hanya berfokus pada akurasi prediksi tanpa mempertimbangkan transparansi hasil model. Penelitian-penelitian terbaru yang mengadopsi algoritma transformer seperti IndoBERT dan NusaBERT [10], [12] menunjukkan peningkatan signifikan dalam performa klasifikasi teks berbahasa Indonesia, namun sebagian besar belum mengeksplorasi integrasi Explainable AI seperti LIME untuk menjelaskan keputusan model. Padahal, dalam konteks kebijakan publik, interpretabilitas model menjadi aspek yang krusial untuk memastikan akuntabilitas hasil analisis [16].

Seiring berkembangnya teknologi NLP, model berbasis transformer seperti IndoBERTweet menjadi solusi yang efektif dalam memahami struktur dan konteks bahasa Indonesia di media sosial [4], [11]. Penelitian sebelumnya juga membuktikan bahwa varian IndoBERTweet yang dilatih khusus pada korpus Twitter berbahasa Indonesia, seperti IndoBERTweet dan IndoBERTweet-Lite, menunjukkan performa unggul pada tugas klasifikasi sentimen dibandingkan model multilingual seperti mBERT atau XLM-R [12], [13]. Namun, model ini bersifat “black-box”, sehingga sulit menjelaskan alasan di balik keputusan prediktifnya. Oleh karena itu, penggunaan metode XAI seperti LIME menjadi sangat penting dalam memberikan pemahaman terhadap cara kerja model [15], [17]. Survei terbaru juga menunjukkan bahwa penerapan XAI pada tugas-tugas NLP, termasuk klasifikasi teks, menjadi salah satu fokus utama penelitian beberapa tahun terakhir [18].

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi sentimen berbasis IndoBERTweet terhadap opini publik mengenai isu IKN, PPN 11%, dan MBG, serta menerapkan metode LIME untuk menjelaskan hasil prediksi model. Dengan pendekatan ini, penelitian diharapkan dapat memberikan kontribusi terhadap pengembangan sistem analisis sentimen yang tidak hanya akurat tetapi juga dapat dijelaskan secara transparan dan dapat dipercaya.

2. METODE PENELITIAN

Penelitian ini dirancang melalui serangkaian tahapan sistematis yang dimulai dari pengumpulan data, pemrosesan awal teks, pelabelan sentimen, pelatihan model berbasis IndoBERTweet, hingga interpretasi hasil klasifikasi menggunakan pendekatan Explainable AI (LIME). Setiap tahapan disusun secara terstruktur agar hasil klasifikasi tidak hanya akurat tetapi juga dapat dijelaskan secara transparan.

2.1. Arsitektur dan Alur Penelitian

Secara umum, alur penelitian ini dapat dilihat pada Gambar 1, yang menggambarkan keseluruhan proses mulai dari akuisisi data hingga interpretasi hasil model.

Tahap pertama adalah proses pengambilan data (scraping) dari platform media sosial X menggunakan kata kunci “IKN”, “PPN”, dan “MBG”. Tweet yang terkumpul kemudian masuk ke tahap pra-pemrosesan teks, meliputi penghapusan elemen non-teks seperti URL, mention, hashtag, emotikon, serta normalisasi kata tidak baku agar seragam.

Selanjutnya dilakukan proses anotasi manual menggunakan Doccano, di mana setiap tweet diberi label sentimen positif, negatif, atau netral berdasarkan makna opini yang terkandung. Data berlabel ini kemudian dibagi menjadi data latih, validasi, dan uji (80:10:10) untuk proses pelatihan model.

Tahap berikutnya adalah pelatihan (fine-tuning) model IndoBERTweet menggunakan konfigurasi hyperparameter tertentu, dengan evaluasi berbasis metrik akurasi, precision, recall, dan F1-score. Setelah model terbaik diperoleh, dilakukan interpretasi hasil klasifikasi menggunakan metode Explainable AI berbasis LIME untuk menjelaskan keputusan model secara lokal terhadap setiap prediksi.



Gambar 1. Arsitektur Alur Penelitian

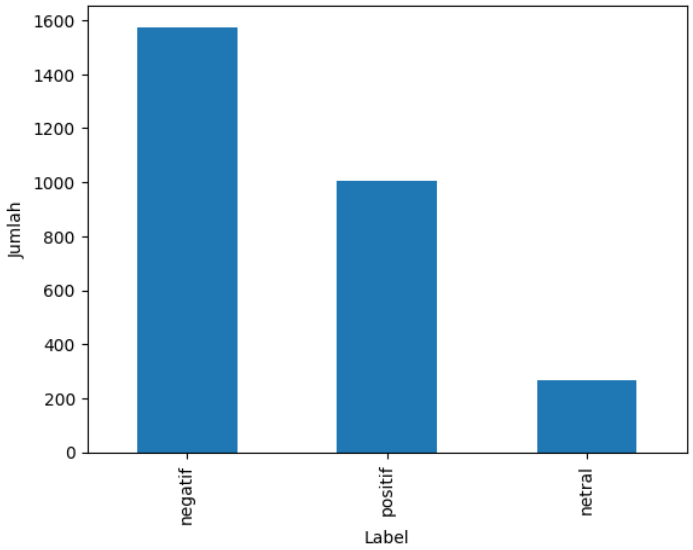
2.2. Dataset dan Pra-pemrosesan

Dataset dalam penelitian ini diperoleh melalui proses *web scraping* menggunakan tools *tweet harvest* terhadap media sosial X menggunakan kata kunci “IKN”, “PPN”, dan “MBG”. Dari proses ini berhasil dikumpulkan sebanyak 2.845 tweet berbahasa Indonesia yang mencerminkan opini publik terhadap tiga isu kebijakan nasional tersebut.

Setelah proses pengumpulan data, seluruh tweet diberi label sentimen secara manual menggunakan alat anotasi Doccano, sebuah platform pelabelan teks berbasis web yang mendukung klasifikasi multi-kategori secara sistematis. Proses pelabelan dilakukan oleh penulis sendiri dan sebagian dibantu oleh satu penilai tambahan dengan latar belakang yang serupa. Tiga kelas sentimen yang digunakan adalah positif, negatif, dan netral, berdasarkan ekspresi atau opini yang terkandung dalam teks. Kriteria pelabelan ditetapkan sebagai berikut: positif

untuk tweet yang menyuarakan dukungan, optimisme, atau sikap setuju terhadap isu; negatif untuk tweet yang memuat kritik, penolakan, atau sentimen sinis; dan netral untuk tweet yang bersifat informatif tanpa opini eksplisit.

Meskipun sebagian besar tweet dilabeli langsung oleh penulis, diskusi dan klarifikasi terhadap tweet yang ambigu dilakukan bersama penilai tambahan guna menjaga konsistensi interpretasi. Proses ini menghasilkan dataset yang cukup representatif dan siap digunakan dalam pelatihan model klasifikasi sentimen. Diskusi dan klarifikasi terhadap tweet ambigu dilakukan untuk menjaga konsistensi label. Menurut studi oleh Kenyon-Dean et al. (2018) [19] dalam sentiment analisis pelabelan manual tetap menjadi pendekatan yang valid dan andal untuk tugas klasifikasi teks, asalkan dilakukan dengan panduan konsistensi dan pemeriksa antar-anotator. Pendekatan ini memastikan bahwa model memperoleh data pelatihan yang merepresentasikan persepsi manusia secara akurat. Oleh karena itu, metode anotasi manual yang diterapkan dalam penelitian ini dianggap sahih secara metodologis dan sesuai dengan praktik penelitian NLP terkini. Distribusi label sentimen pada dataset dapat dilihat pada Gambar 2.



Gambar 2. Distribusi Label

Sebelum data digunakan dalam pelatihan model IndoBERTweet, dilakukan tahap pra-pemrosesan untuk memastikan bahwa teks yang diinputkan bersih, seragam, dan sesuai dengan format yang dapat ditangani oleh tokenizer model. Tahapan ini penting mengingat data berasal dari media sosial yang sangat bervariasi dari segi struktur, gaya bahasa, dan keberadaan simbol-simbol informal.

Langkah pertama adalah penghapusan elemen non-teks seperti tautan (URL), *mention* (@username), *hashtag*, dan emotikon yang tidak relevan. Selanjutnya, seluruh teks diubah menjadi huruf kecil (lowercase) untuk menghindari perbedaan pemrosesan antar kata yang sama tetapi dengan kapitalisasi berbeda. Kata-kata yang dianggap tidak membawa makna penting dalam konteks sentiment seperti *stopwords* dalam Bahasa Indonesia juga dihapus untuk mengurangi noise dalam data.

Langkah berikutnya adalah normalisasi kata tidak baku dan singkatan khas Twitter (seperti "gk" menjadi "nggak", atau "yg" menjadi "yang") agar lebih mudah dipahami oleh model. Terakhir, dilakukan tokenisasi dan pembatasan panjang teks maksimal hingga 128 token sesuai

konfigurasi model. Proses ini menghasilkan teks bersih (clean text) yang siap digunakan dalam tahap pelatihan dan evaluasi model klasifikasi.

Tabel 2. Contoh Data Pre Processing

Raw Text	Clean Text
MBG bukan cuma soal makanan tapi perhatian. Anak-anak ngerasain kasih sayang negara lewat setiap suapan. Di hari ulang tahun Pak Prabowo mereka bilang: Terima kasih Pak semoga panjang umur & makin kuat memimpin bangsa. ðŸ‡®ðŸ‡© https://t.co/MvdMGw7rxw	mbg bukan cuma soal makanan tapi perhatian anakanak ngerasain kasih sayang negara lewat setiap suapan di hari ulang tahun pak prabowo mereka bilang terima kasih pak semoga panjang umur amp makin kuat memimpin bangsa
@hoteliercrypto Tapi kok ga kena ppn 12% ky temen2 ya bli ms sama ky bln lalu https://t.co/cR1FZ8kJUY	tapi kok ga kena ppn 12 ky temen2 ya bli ms sama ky bln lalu

2.3. Model IndoBERTweet

Model dasar yang digunakan dalam penelitian ini adalah IndoBERTweet, secara spesifik menggunakan varian 'indolem/indobertweet-base-uncased' yang tersedia pada pustaka Hugging Face. Model ini merupakan arsitektur berbasis Transformer yang telah melalui proses *pre-training* menggunakan korpus Twitter berbahasa Indonesia. Proses tokenisasi dilakukan memanfaatkan modul AutoTokenizer yang disesuaikan dengan model tersebut, dengan batasan panjang sekuens (*max length*) ditetapkan sebesar 128 token.

2.4. Proses Pelatihan Model

Dataset dibagi menjadi tiga subset dengan proporsi 90% data latih, dan 10% data uji. Proses *fine-tuning* model dilakukan menggunakan optimizer AdamW sesuai standar implementasi BERT, dengan konfigurasi parameter sebagai berikut:

Tabel 3. Parameter Model Training

Parameter	Nilai
Epoch	3
Batch Size	16
Learning Rate	2e-5
Max sequence length	128

Konfigurasi hyperparameter pada Tabel 3 merupakan pengaturan utama yang digunakan dalam proses fine-tuning model IndoBERTweet. Pemilihan tiga epoch bertujuan untuk memberikan kesempatan bagi model untuk belajar secara berulang tanpa menyebabkan overfitting pada data latih. Nilai batch size sebesar 16 dipilih untuk menyeimbangkan stabilitas proses pembelajaran dan keterbatasan sumber daya komputasi. Learning rate sebesar 2e-5 mengikuti praktik umum pada pelatihan model berbasis Transformer, yakni menggunakan laju pembelajaran yang relatif kecil agar pembaruan bobot berjalan stabil. Sementara itu, panjang sekuens maksimum ditetapkan 128 token untuk menyesuaikan karakteristik tweet yang umumnya berisi teks pendek sekaligus menjaga efisiensi komputasi. Kombinasi pengaturan ini menghasilkan proses pelatihan yang stabil dan mendukung tercapainya performa model yang dilaporkan pada bagian hasil. Konfigurasi ini sejalan dengan praktik fine-tuning model Transformer seperti BERT dan IndoBERTweet pada tugas klasifikasi teks [9], [11].

2.5. Explainable AI: Metode LIME

Untuk menjelaskan keputusan model yang bersifat "black-box", digunakan pendekatan XAI dengan metode LIME (Local Interpretable Model-Agnostic Explanations).

LIME memberikan penjelasan lokal terhadap hasil klasifikasi dengan cara memodelkan ulang prediksi model menggunakan model interpretable di sekitar instance yang diamati. Dalam konteks ini, LIME digunakan untuk menampilkan kontribusi kata (fitur) terhadap prediksi sentimen IndoBERTweet.

LIME memberikan kontribusi dalam menjelaskan hasil klasifikasi dengan menunjukkan kata-kata yang dominan dalam menentukan sentimen. Berdasarkan evaluasi manual terhadap sejumlah visualisasi LIME, ditemukan bahwa kontribusi fitur yang ditampilkan konsisten dengan sentimen yang diklasifikasikan pada lebih dari 90% sampel. Hal ini menunjukkan bahwa interpretasi lokal dari model sesuai dengan konteks semantik tweet yang dianalisis.

2.6. Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa sistem klasifikasi sentimen yang dibangun, khususnya pada data uji (test set). Metrik evaluasi yang digunakan meliputi akurasi, precision, recall, dan F1-score dengan pendekatan rata-rata macro average. Penghitungan metrik dilakukan menggunakan pustaka sklearn.metrics dan fungsi classification_report.

Selain evaluasi secara agregat, sistem juga menghasilkan laporan per kelas untuk masing-masing label sentimen yaitu *positif*, *negatif*, dan *netral*. Laporan ini memuat nilai precision, recall, dan F1-score untuk setiap kelas sehingga memudahkan analisis performa model secara menyeluruh.

Penilaian dilakukan pada hasil prediksi model terhadap data uji sebanyak 285 tweet, setelah pelatihan diselesaikan pada model terbaik yang disimpan berdasarkan checkpoint dengan akurasi validasi tertinggi.

3. HASIL PENELITIAN DAN PEMBAHASAN

3.1. Hasil Evaluasi Model

Dataset dibagi dengan rasio 90:10, menghasilkan 2.560 data latih dan 285 data uji. Model klasifikasi sentimen yang dibangun menggunakan IndoBERTweet kemudian dievaluasi pada data uji tersebut. Hasil evaluasi menunjukkan bahwa model mampu mengklasifikasikan tweet dengan akurasi sebesar 89% dan F1-score makro sebesar 0,78. Tabel 4 menyajikan metrik precision, recall, dan F1-score untuk masing-masing kelas sentimen.

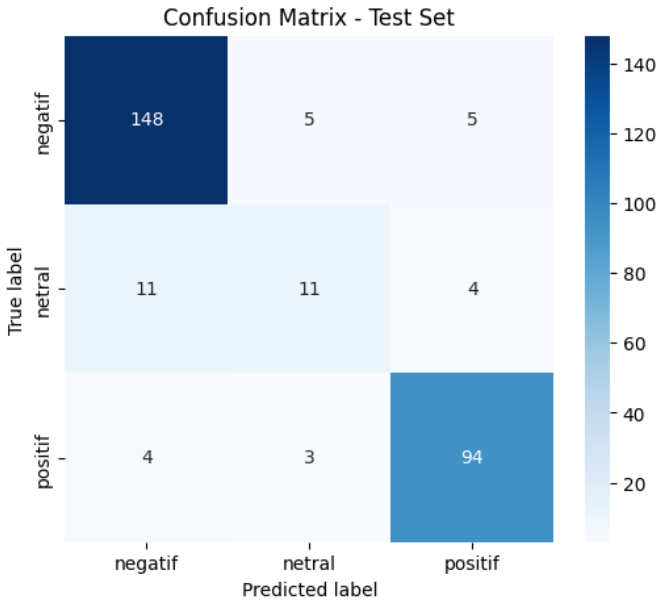
Tabel 4. Hasil Evaluasi Model pada Data Uji

	Precision	Recall	F1-Score	Support
Negatif	0.91	0.94	0.92	158
Netral	0.58	0.42	0.49	26
Positif	0.91	0.93	0.92	101
Accuracy			0.89	285
Macro avg	0.80	0.76	0.78	285
Weighted avg	0.88	0.89	0.88	285

Model menunjukkan performa terbaik dalam mengidentifikasi tweet dengan sentimen negatif dan positif, terlihat dari nilai F1-score yang tinggi (>0,90) untuk kedua kelas tersebut.

Namun, performa terhadap kelas netral masih rendah, hanya menghasilkan F1-score sebesar 0,49. Hal ini dapat disebabkan oleh konteks tweet netral yang cenderung ambigu atau minim kata kunci eksplisit yang kuat secara semantik.

Sebagai pelengkap, confusion matrix digunakan untuk menampilkan jumlah prediksi benar dan salah antar kelas. Dari visualisasi confusion matrix, diketahui bahwa sebagian besar kesalahan klasifikasi terjadi pada tweet netral yang diklasifikasikan sebagai negatif, serta sebagian kecil tweet positif yang juga diklasifikasikan negatif.



Gambar 3. Confusion Matrix Hasil Klasifikasi IndoBERTweet

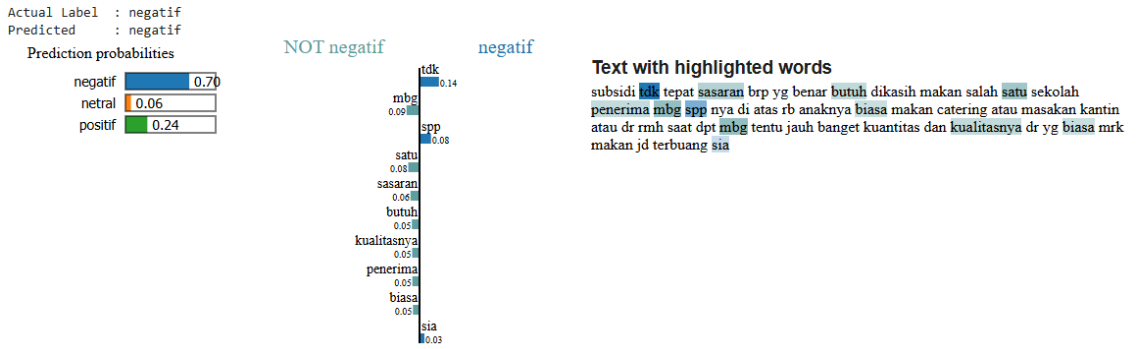
Berdasarkan Gambar 3 terlihat bahwa model mampu mengklasifikasikan tweet negatif dan positif dengan cukup konsisten. Dari total 158 tweet dengan label negatif, sebanyak 148 tweet berhasil diprediksi benar sebagai negatif, sedangkan masing-masing 5 tweet salah diklasifikasikan sebagai netral dan positif. Untuk kelas positif, 94 dari 101 tweet diklasifikasikan dengan tepat sebagai positif, sementara 4 tweet positif diprediksi sebagai negatif dan 3 tweet lainnya sebagai netral. Pola ini sejalan dengan nilai F1-score yang tinggi (>0,90) pada kedua kelas tersebut.

Sebaliknya, performa model pada kelas netral terlihat jauh lebih menantang. Dari 26 tweet netral, hanya 11 yang berhasil dikenali dengan benar sebagai netral. Sebanyak 11 tweet netral justru diklasifikasikan sebagai negatif dan 4 lainnya sebagai positif. Hal ini menunjukkan bahwa model cenderung "menarik" tweet yang ambigu atau informatif tanpa ekspresi emosi eksplisit ke kelas sentimen yang lebih dominan, khususnya negatif. Kondisi ini juga dipengaruhi oleh distribusi data yang tidak seimbang, di mana jumlah tweet netral jauh lebih sedikit dibandingkan tweet negatif dan positif, sehingga representasi kelas netral menjadi kurang kuat selama proses pelatihan.

3.2. Analisis Visualisasi LIME

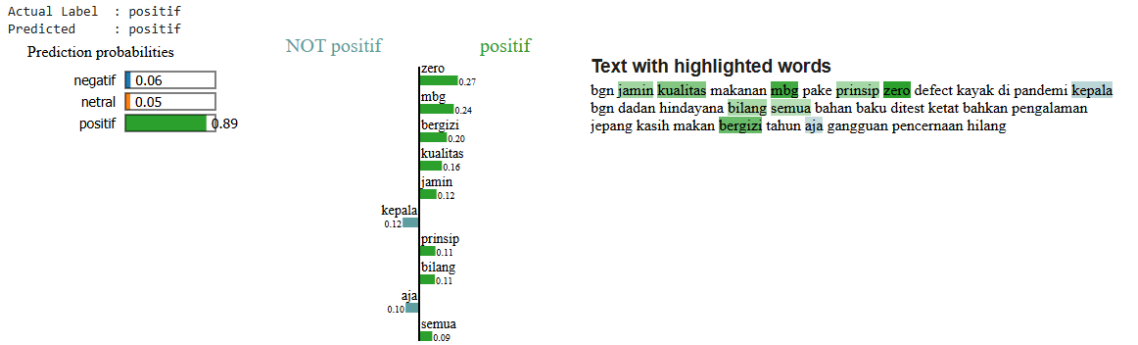
Untuk memahami keputusan yang diambil oleh model klasifikasi sentimen, digunakan pendekatan XAI berupa LIME (Local Interpretable Model-Agnostic Explanations). Metode ini membantu menjelaskan prediksi model terhadap tweet tertentu dengan menyoroti kata-kata yang memberikan pengaruh terbesar dalam klasifikasi.

Berdasarkan beberapa visualisasi hasil LIME yang diterapkan pada berbagai contoh tweet, ditemukan bahwa model IndoBERTweet secara konsisten menggunakan fitur-fitur yang secara semantik relevan dengan sentimen yang ditargetkan. Gambar 4, 5, dan 6 memperlihatkan tiga contoh representatif untuk masing-masing kategori sentimen.



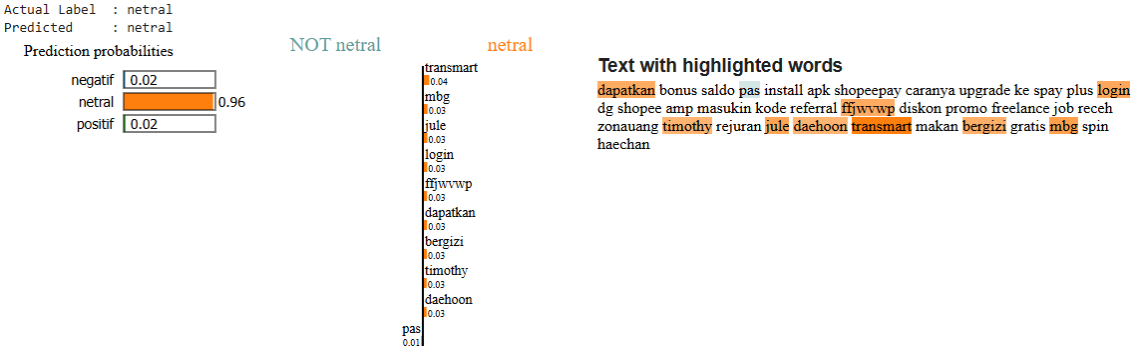
Gambar 4. Visualisasi LIME untuk Tweet Bernada Negatif

Pada Gambar 4, terlihat bahwa tweet bernada negatif terhadap program Makan Bergizi Gratis (MBG) diklasifikasikan secara akurat oleh model. Kata-kata seperti *"tdk"*, *"spp"*, *"butuh"*, dan *"sia"* memiliki bobot negatif yang signifikan, menandakan keluhan terkait ketidaktepatan sasaran dan pemborosan sumber daya.



Gambar 5. Visualisasi LIME untuk Tweet Bernada Positif

Gambar 5 menampilkan tweet dengan sentimen positif yang juga diklasifikasikan dengan benar oleh model. Kata-kata seperti *"jamin"*, *"kualitas"*, *"bergizi"*, dan *"prinsip"* menjadi penentu utama prediksi positif. Hal ini menunjukkan bahwa model mampu mengenali kata-kata yang mengandung apresiasi terhadap kualitas dan manfaat program.



Gambar 6. Visualisasi LIME untuk Tweet Netral

Sementara itu, Gambar 6 menunjukkan tweet dengan sentimen netral, berisi promosi yang tidak mengandung opini eksplisit terhadap isu kebijakan. Model mengenali kata-kata seperti *"login"*, *"transmart"*, *"ffwwwp"*, dan *"berlaku"* yang tidak memiliki arah sentimen kuat, dan dengan demikian berhasil mengklasifikasikan tweet tersebut sebagai netral.

Analisis ini membuktikan bahwa visualisasi LIME dapat membantu peneliti memahami dasar pengambilan keputusan model, serta memvalidasi bahwa prediksi dilakukan berdasarkan fitur-fitur linguistik yang memang relevan. Di sisi lain, analisis ini juga membuka ruang untuk identifikasi kesalahan prediksi yang mungkin berasal dari sarkasme, ambiguitas, atau konteks yang sulit ditangkap oleh model berbasis teks.

3.3. Pembahasan Sentimen terhadap Isu IKN, PPN, dan MBG

Berdasarkan klasifikasi sentimen yang dilakukan terhadap 2.845 tweet terkait tiga isu kebijakan nasional yakni Ibu Kota Negara (IKN), kenaikan Pajak Pertambahan Nilai (PPN 11%), dan program Makan Bergizi Gratis (MBG) terlihat dinamika persepsi publik yang cukup beragam. Model IndoBERTweet berhasil mengidentifikasi kecenderungan opini dengan bantuan metode interpretasi LIME, sehingga memungkinkan analisis yang lebih bermakna terhadap isu-isu tersebut.

IKN (Ibu Kota Negara Baru)

Sebagian besar tweet terkait IKN diklasifikasikan ke dalam kategori sentimen negatif. Visualisasi LIME menunjukkan dominasi kata-kata seperti *"boros"*, *"tidak penting"*, *"prioritas"*, dan *"rakyat"*, yang sering digunakan dalam konteks kritik terhadap pembangunan Ibu Kota baru. Hasil ini konsisten dengan temuan dari [20] yang menyebutkan bahwa persepsi publik terhadap IKN cenderung skeptis akibat isu prioritas anggaran dan ketimpangan pembangunan antar daerah.

PPN 11% (Kenaikan Pajak)

Topik PPN 11% juga memicu sentimen negatif yang kuat di media sosial. Tweet-tweet yang diklasifikasikan sebagai negatif banyak memuat kata-kata seperti *"berat"*, *"rakyat kecil"*, *"kenaikan"*, dan *"mencekik"*. Hal ini mengindikasikan bahwa masyarakat menganggap kebijakan ini memberatkan, terutama di tengah kondisi ekonomi pascapandemi. Penelitian sebelumnya oleh [14] yang menggunakan VADER dan oleh [21] menggunakan model indoBERT juga menunjukkan pola serupa, yakni dominasi sentimen negatif terhadap isu kenaikan pajak.

MBG (Makan Bergizi Gratis)

Berbeda dengan dua isu sebelumnya, MBG menunjukkan distribusi sentimen yang lebih beragam dan seimbang. Banyak tweet dengan sentimen positif mengandung kata seperti *"bergizi"*, *"gratis"*, *"jamin"*, dan *"bermanfaat"*, yang mengapresiasi program ini sebagai upaya meningkatkan kualitas gizi anak sekolah. Namun, terdapat pula sentimen negatif yang mengkritisi aspek implementasi, seperti efisiensi anggaran, akurasi sasaran, atau dugaan pemborosan.

Visualisasi LIME memperkuat interpretasi ini: kata-kata seperti *"tidak tepat sasaran"*, *"sia"*, dan *"spp"* muncul dalam tweet negatif, sementara *"jamin"*, *"kualitas"*, dan *"bergizi"* muncul dalam tweet positif. Studi terkait oleh [2], [6], [7] menyebutkan bahwa persepsi masyarakat terhadap MBG sangat bergantung pada konteks lokal, faktor politis, dan pengalaman penerima.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa model IndoBERTweet yang digunakan untuk klasifikasi sentimen masyarakat terhadap isu-isu kebijakan publik menunjukkan kinerja yang sangat baik. Model berhasil mencapai akurasi sebesar 89% dan nilai macro average F1-score sebesar 0,78 pada data uji, dengan performa optimal pada kelas sentimen positif dan negatif. Sementara itu, kelas netral masih memiliki akurasi dan konsistensi prediksi yang lebih rendah, mengindikasikan perlunya perhatian lebih dalam penyeimbangan label saat pelatihan.

Penelitian ini juga membuktikan bahwa metode Explainable AI, khususnya LIME, efektif digunakan untuk meningkatkan transparansi dalam sistem klasifikasi. LIME memberikan visualisasi kontribusi kata terhadap hasil prediksi yang dilakukan oleh model, sehingga pengguna dapat memahami alasan di balik keputusan yang diambil oleh model. Interpretabilitas ini penting dalam konteks pengambilan keputusan berbasis data dan sangat berguna dalam studi opini publik yang bersifat sensitif.

Analisis terhadap tiga isu yang menjadi fokus penelitian yakni IKN, PPN 11%, dan MBG menunjukkan dinamika opini publik yang beragam. Sentimen terhadap IKN dan PPN didominasi oleh opini negatif, sementara MBG mendapat tanggapan yang lebih berimbang antara positif dan negatif. Hasil ini memperkuat peran media sosial sebagai ruang kritik publik terhadap kebijakan pemerintah.

Temuan ini menegaskan bahwa model transformer berbahasa Indonesia yang dipadukan dengan pendekatan XAI dapat menjadi alat yang efektif untuk memetakan opini publik terhadap kebijakan nasional di ruang digital. Dengan perbaikan pada distribusi label netral, penambahan platform media sosial lain, serta eksplorasi metode interpretasi alternatif seperti SHAP atau Integrated Gradients, penelitian lanjutan berpotensi memberikan gambaran yang lebih komprehensif mengenai dinamika sentimen masyarakat Indonesia.

Sebagai saran, penelitian selanjutnya disarankan untuk meningkatkan distribusi label data, khususnya pada kelas sentimen netral agar model tidak bias pada label dominan. Selain itu, pelibatan lebih banyak anotator saat proses pelabelan manual dapat meningkatkan validitas data latih. Pengambilan data dari platform lain seperti TikTok atau YouTube juga direkomendasikan untuk memperluas cakupan opini masyarakat. Terakhir, perbandingan metode interpretasi lain seperti SHAP atau Integrated Gradients patut dijelajahi lebih lanjut agar pemahaman terhadap perilaku model menjadi lebih komprehensif. Dengan penguatan metodologis tersebut, penelitian sejenis dapat memberikan kontribusi yang lebih luas dalam mendukung pengambilan keputusan berbasis data

PERNYATAAN PENGGUNAAN ARTIFICIAL INTELLIGENCE

Penulis memanfaatkan ChatGPT dan Claude untuk membantu memperbaiki tata bahasa, menyempurnakan terminologi akademik. Hasil AI diperiksa secara untuk menjaga ketepatan makna dan konsistensi ilmiah. Penulis bertanggung tanggung jawab penuh atas hasil akhir artikel.

KONTRIBUSI PENULIS

Luthfi Delvian: Konsep, Metode, Analisis Data, Penulisan Artikel.

Esa Firmansyah: Kurasi Data, Analisis Data, Supervisi, Review.

Beben Sutara: Konsep, Metode, Supervisi, Review.

PENDANAAN

Penelitian ini tidak menerima hibah khusus dari lembaga pendanaan mana pun, baik sektor publik, komersial, maupun nirlaba.

KONFLIK KEPENTINGAN

Penulis menyatakan tidak terdapat konflik kepentingan terkait publikasi artikel ini.

DAFTAR PUSTAKA

[1] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif Intell Rev*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[2] A. S. Hanin and M. Maryam, "Sentiment Analysis of Twitter Towards the Free Lunch Program Using the C4.5 Algorithm," *International Journal of Advances in Data and Information Systems*, vol. 6, no. 1, pp. 31–45, Apr. 2025, doi: 10.59395/ijadis.v6i1.1357.

[3] C. Juditha, "The Phenomenon of 'Indonesia Gelap' on Social Media: Sentiment Analysis and Public Opinion Polarization," *Jurnal Komunikasi*, vol. 17, no. 1, pp. 157–170, Jul. 2025, doi: 10.24912/jk.v17i1.33846.

[4] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 1071–1078, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[5] M. N. P. Abadi *et al.*, "Unraveling Future Trends in Free School Lunch and Nutrition: Global Insights for Indonesia from Bibliometric Approach and Critical Review," Sep. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/nu17172777.

[6] I. Hermansyah and M. S. Hasibuan, "Sentiment Analysis of Free Nutritious Meal Programme on Social Media X using Linear Regression and Random Forest Algorithms," *PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic*, vol. 13, no. 1, pp. 55–68, Mar. 2025, doi: 10.33558/piksel.v13i1.10633.

[7] M. Farhan Saleh and R. Imanda, "Public Sentiment Analysis of the Free Meal Program: A Comparison of Naive Bayes and Support Vector Machine Methods on the Twitter (X) Social Media Platform," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[8] A. Elbasiony, I. M. El-Hasnony, and S. Abdelrazek, "XAI-Based Sentiment Analysis Using Machine Learning Approaches," *Mansoura Journal for Computer and Information Sciences*, vol. 19, no. 1, pp. 23–42, Dec. 2024, doi: 10.21608/mjcis.2024.275844.1002.

[9] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171-4186. 2019.

[10] T. Dwi Purnomo and J. Sutopo, "Comparison Of Pre-Trained Bert-Based Transformer Models For Regional Language Text Sentiment Analysis In Indonesia" no. 3, pp. 11–21, 2024, doi: 10.56127/ijst.v3i3.1.

[11] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.04607

[12] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[13] Dhendra and V. Gayuh Utomo, "Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews," *Jurnal Transformatika*, vol. 23, no. 1, pp. 86–95, Jul. 2025, doi: 10.26623/transformatika.v23i1.12095.
[14] N. Fauziah, "Analisis Sentimen Publik Terhadap Kenaikan Tarif PPN di Indonesia dengan Pendekatan VADER," *Jurnal Akuntansi dan Keuangan*, vol. 12, no. 2, p. 228, Sep. 2024, doi: 10.29103/jak.v12i2.16796.
[15] A. Sawant, "A Comparative Study of SHAP and LIME: Interpreting Black-Box Machine Learning Models," *International Journal Of Scientific Research In Engineering And Management*, vol. 09, no. 07, pp. 1–9, Jul. 2025, doi: 10.55041/IJSREM51586.
[16] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, "XAI-Explainable artificial intelligence," *Sci Robot*, vol. 4, no. 37, Dec. 2019, doi: 10.1126/scirobotics.aay7120.
[17] J. Dieber and S. Kirrane, "Why model why? Assessing the strengths and limitations of LIME," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2012.00093
[18] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A Survey of the State of Explainable AI for Natural Language Processing," Oct. 2020, doi: 10.18653/v1/2020.aacl-main.46.
[19] K. Kenyon-Dean *et al.*, "Sentiment Analysis: It's Complicated! 1." [Online]. Available: https://github.com/networkdynamics/
[20] I. Sa'diyah, A. T. Aviolla Terza, C. C. B. Lima, M. R. M. Ariefean, and I. Athallah, "Sentiment Analysis of Netizen's Comments on YouTube about IKN (Capital City) Development in Indonesia," *GHANCARAN: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, vol. 6, no. 2, Jan. 2025, doi: 10.19105/ghancaran.v6i2.15432.
[21] M. R. Manoppo *et al.*, "Analisis Sentimen Publik Di Media Sosial Terhadap Kenaikan PPN 12% Di Indonesia Menggunakan Indobert" *Jurnal Kecerdasan Buatan dan Teknologi Informasi*, vol. 4, no. 2, pp. 152–163, May 2025, doi: 10.69916/jkbti.v4i2.322.