

IMPLEMENTASI ALGORITMA CATBOOST BERBASIS TEXT FEATURES PADA KLASIFIKASI SENTIMEN PUBLIK TERKAIT KONFLIK EKOLOGI DAN EKONOMI PERTAMBANGAN

Syauqiyah Mujahidah Yahya^{1*}, Muhyiddin A. M. Hayat², Desi Anggreani³

^{1,2,3}Universitas Muhammadiyah Makassar, Indonesia

¹105841105122@student.unismuh.ac.id, ² muhyiddin@unismuh.ac.id,

³ desianggreani@unismuh.ac.id

*Corresponding Author

Info Artikel

Pengajuan : 31-07-2026
Peninjauan : 15-08-2026
Revisi : 30-08-2026
Diterima : 18-09-2026
Diterbitkan : 24-09-2026

Kata Kunci:

Analisis Sentimen, CatBoost,
Konflik Pertambangan,
Native Text Features,
TikTok dan YouTube.

Abstrak

Isu pertambangan kerap memicu dilema antara eksploitasi ekonomi dan pelestarian ekologi, yang memunculkan polarisasi opini publik secara masif di media sosial. Kumpulan opini publik yang berupa data tidak terstruktur ini berpotensi menjadi acuan mitigasi konflik jika dianalisis dengan tepat, namun metode analisis manual terbukti tidak efisien dan rentan terhadap bias. Penelitian ini bertujuan untuk mengimplementasikan algoritma CatBoost berbasis Text Features guna mengklasifikasikan sentimen publik terhadap isu pertambangan ke dalam kelas positif, negatif, dan netral. Data penelitian bersumber dari komentar TikTok dan YouTube yang diekstraksi menggunakan metode web scraping, menghasilkan 2.049 data bersih. Pemrosesan dilakukan dengan memanfaatkan fitur Native Text Handling pada arsitektur CatBoost. Melalui fitur ini, teks diproses secara internal menggunakan tokenisasi dan kalkulasi kamus frekuensi (Bag-of-Words dan BM25) tanpa memerlukan ekstraksi fitur matriks sparse eksternal seperti TF-IDF, sehingga mampu mereduksi beban komputasi dan memori secara signifikan. Data dibagi menggunakan teknik stratified random sampling dengan rasio 80:20, yang terdiri dari 1.639 data latih dan 410 data uji. Hasil evaluasi model menggunakan 5-Fold Cross Validation, Confusion Matrix, dan optimasi hyperparameter Optuna (dengan penyesuaian kedalaman pohon/depth = 5) menunjukkan tingkat Akurasi Keseluruhan sebesar 79% (0,7878) dan nilai rata-rata F1-Score Macro sebesar 0,79. Penurunan parameter depth dari 12 menjadi 5 serta penghapusan matriks TF-IDF secara signifikan mempercepat waktu komputasi pelatihan model (training time). Secara spesifik, kelas Netral mencapai tingkat kepekaan (Recall) tertinggi sebesar 85%, sedangkan kelas Negatif mencatatkan tingkat kehati-hatian (Precision) tertinggi sebesar 86%. Model klasifikasi berbasis CatBoost terbukti tangguh, efisien, dan seimbang dalam memetakan polaritas opini publik secara otomatis sebagai landasan deteksi dini mitigasi konflik sosial-lingkungan

How to Cite : Yahya, S. M., Hayat, M. A. M., & Anggreani, D. (2026). Implementasi Algoritma Catboost Berbasis Text Features Pada Klasifikasi Sentimen Publik Terkait Konflik Ekologi Dan Ekonomi Pertambangan. Journal of Computer Science and Information Technology (JCSIT), 3(4), 492-499. <https://doi.org/10.70248/jcsit.v3i4.4404>

DOI : <https://doi.org/10.70248/jcsit.v3i4.4404>

This is an open access article under <https://creativecommons.org/licenses/by/4.0/>
Copyright© 2026 by Author. Published by Yayasan Nuraini Ibrahim Mandiri



PENDAHULUAN

Sektor pertambangan di Indonesia selalu berada pada posisi yang dilematis dalam arah pembangunan nasional (Zhao et al., 2023). Di satu sisi, industri ekstraktif menjadi roda penggerak ekonomi warga dan menyumbang Pendapatan Asli Daerah (PAD) yang signifikan, serta membuka lapangan pekerjaan bagi masyarakat lokal. Namun di sisi lain, dampak destruktifnya terhadap kerusakan ekosistem alam seperti

deforestasi, pencemaran sungai, dan marginalisasi ruang hidup kerap memicu perdebatan yang sangat hangat. Di era digitalisasi, adu argumen mengenai tarik-ulur kepentingan ekologi dan ekonomi ini tidak lagi terbatas pada demonstrasi di jalanan atau ruang rapat formal, melainkan meluas secara masif ke berbagai platform media sosial tanpa adanya sekat geografis (Suresh Kumar et al., 2024). Platform video interaktif seperti TikTok dan YouTube kini menjadi wadah utama bagi jutaan pengguna untuk berinteraksi setiap hari. Mereka leluasa menumpahkan kritik, dukungan, maupun keluhan langsung melalui komentar teks, yang sesungguhnya membentuk tumpukan Big Data opini publik berskala raksasa (Kaur & Sharma, 2023).

Potensi data digital yang melimpah ini sesungguhnya sangat bernilai esensial bagi pemerintah dan pemangku kepentingan untuk membaca arah pergerakan massa serta memitigasi keresahan sosial secara dini (Kim et al., 2022). Sayangnya, instansi terkait masih sangat mengandalkan metode observasi manual yang memakan waktu lama, tidak efisien, dan bersifat sangat reaktif. Ketidadaan sistem deteksi dini digital membuat para pembuat kebijakan sering kali gagal menangkap pola lonjakan sentimen negatif di dunia maya sebelum akhirnya bertransformasi menjadi gesekan sosial dan demonstrasi fisik di lapangan (Lin & Bu, 2022). Jika pemantauan konvensional yang lambat ini terus dipertahankan, konflik antara masyarakat lokal, perusahaan tambang, dan aktivis lingkungan akan berlarut-larut serta berpotensi mempertaruhkan stabilitas keamanan daerah. Oleh karena itu, diperlukan sebuah transisi paradigma menuju otomatisasi analitik proaktif menggunakan teknologi Machine Learning untuk mengupas tumpukan teks opini secara instan, cepat, dan akurat (Narasamma & Sreedevi, 2021).

Pendekatan Natural Language Processing (NLP), khususnya analisis sentimen, memungkinkan ribuan komentar diekstraksi maknanya dan diklasifikasikan secara otomatis oleh mesin ke dalam kategori sentimen positif, negatif, atau netral (Purwono et al., 2022). Kendati demikian, karakteristik bahasa tulisan netizen di media sosial umumnya bersifat informal, tidak mematuhi kaidah tata bahasa baku, dan dipenuhi derau tekstual (noise). Berbagai algoritma konvensional terdahulu seperti K-Means, Naive Bayes, atau Support Vector Machine (SVM) memiliki kelemahan mendasar saat berhadapan dengan data bahasa yang tidak baku ini. Algoritma konvensional tersebut mewajibkan sistem mengubah teks menjadi deretan angka melalui proses vektorisasi numerik eksternal (seperti TF-IDF). Proses matematis yang rumit ini berisiko mengorbankan konteks semantik kalimat asli, membengkakkan dimensi komputasi (curse of dimensionality), dan pada akhirnya membuat model sangat rentan mengalami overfitting atau sekadar menghafal data latih semata (Alshaari & Alqahtani, 2025).

Proses matematis yang rumit ini berisiko mengorbankan konteks semantik kalimat asli, membengkakkan dimensi komputasi (curse of dimensionality), dan pada akhirnya membuat model sangat rentan mengalami overfitting atau sekadar menghafal data latih semata. Di sisi lain, meskipun arsitektur berbasis Transformer (seperti IndoBERT) menawarkan performa klasifikasi yang tinggi, model Deep Learning ini menuntut sumber daya komputasi (GPU) yang masif dan rentan mengalami ketidakstabilan jika dilatih pada dataset spesifik berskala menengah. Oleh karena itu, terdapat research gap untuk menemukan algoritma yang menyeimbangkan antara efisiensi komputasi dan akurasi NLP.

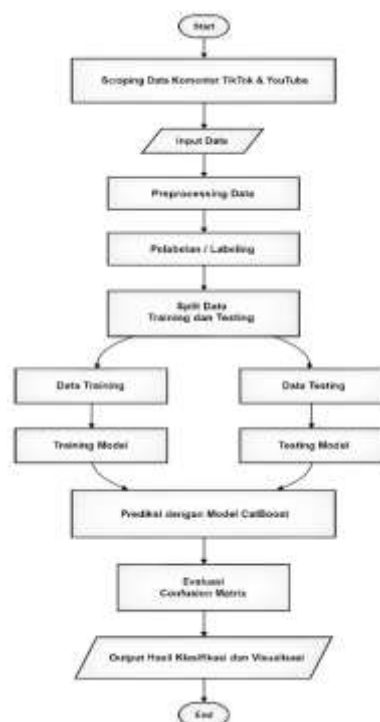
Guna menjembatani kekurangan krusial dari batasan algoritmik pendahulunya tersebut, kehadiran algoritma pembelajaran mesin tingkat lanjut mutlak untuk segera diimplementasikan. Penelitian ini mengusulkan penggunaan CatBoost (Categorical Boosting), sebuah algoritma ensemble learning mutakhir berbasis Gradient Boosting

Decision Tree (GBDT) yang terbukti sangat tangguh menangani data tak terstruktur (Segoni et al., 2025). Keunggulan inovatif algoritma ini terletak pada fitur bawaan Native Text Handling (Chen, 2025). Mekanisme ini memecah kalimat menjadi token kata secara natural dan mengalkulasi nilai probabilitas fitur teks secara langsung di dalam model, tanpa memaksa data melewati proses vektorisasi numerik eksternal yang membebani memori. Didukung oleh arsitektur pohon keputusan simetris (Oblivious Trees), CatBoost sangat stabil dalam membatasi kompleksitas percabangan untuk mencegah overfitting secara efektif. Secara konseptual dan matematis, CatBoost mengelola derau (noise) bahasa tidak baku melalui struktur pohon keputusan simetris (Oblivious Trees). Struktur ini memaksakan kriteria pemisahan (split condition) yang identik di setiap level kedalaman pohon, yang secara otomatis bertindak sebagai regularisasi kuat untuk menangkal overfitting pada data teks yang berisik.

Berdasarkan fenomena dan solusi komputasional yang telah dipaparkan, penelitian ini secara khusus merancang sebuah sistem klasifikasi sentimen opini publik terkait isu pertambangan menggunakan metode CatBoost. Kinerja dan stabilitas model prediksi ini nantinya akan dievaluasi secara komprehensif menggunakan instrumen Confusion Matrix. Melalui penelitian yang berjudul "Implementasi Algoritma CatBoost Berbasis Text Features pada Klasifikasi Sentimen Publik Terkait Konflik Ekologi dan Ekonomi Pertambangan", diharapkan sistem cerdas ini dapat menghasilkan pemetaan sentimen yang presisi. Peta opini digital ini nantinya dapat difungsikan sebagai instrumen mitigasi konflik dan deteksi dini (early warning system) bagi para pengambil kebijakan, guna mewujudkan diskursus penyelesaian konflik pertambangan yang lebih adil dan terukur.

METODE PENELITIAN

Penelitian ini merupakan studi komputasional yang berfokus pada ekstraksi dan pemrosesan Big Data publik dari platform media sosial interaktif. Secara metodologis, arsitektur penyelesaian masalah dibangun melalui serangkaian tahapan sistematis yang dapat dilihat pada diagram alir berikut:



Gambar 1. Alur Perancangan Sistem Klasifikasi

Tahapan operasional dari perancangan sistem di atas dijelaskan sebagai berikut:

1. Pengumpulan Data: Ekstraksi data mentah (raw data) dilakukan menggunakan teknik Web Scraping terotomatisasi via API Apify pada kolom komentar platform TikTok dan YouTube yang membahas konten konflik ekologi dan ekonomi pertambangan di Indonesia. Secara proporsional, korpus data yang dihimpun berasal dari platform TikTok (sekitar 60%) dan YouTube (sekitar 40%). Ukuran sampel sebanyak 2.049 data bersih secara empiris dinilai memadai (sufficient) untuk melatih model algoritma Gradient Boosting pada domain kasus spesifik (narrow domain), di mana variasi kosakata sangat terpusat pada leksikon lingkungan dan ekonomi.
2. Pra-Pemrosesan Teks (Text Preprocessing): Data tidak terstruktur yang memuat banyak derau (noise) dibersihkan melalui proses Cleaning (menghapus URL, tag, angka, dan karakter non-ASCII menggunakan RegEx), Tokenizing (memecah kalimat menjadi unit kata tunggal), dan Stopword Removal (mengeliminasi kata hubung yang tidak berbobot). Pustaka kustomisasi diterapkan untuk memastikan leksikon negasi (seperti "tidak" atau "bukan") dipertahankan agar tidak terjadi pergeseran makna sentimen.
3. Pelabelan Data (Labeling): Korpus teks bersih diklasifikasikan ke dalam tiga kelas polaritas (Positif, Negatif, dan Netral) menggunakan metode berbasis leksikon (lexicon-based) yang divalidasi secara manual. Tingkat keandalan pelabelan ini diukur menggunakan Inter-Annotator Agreement (Uji Cohen's Kappa), yang menghasilkan skor 0,82 (kategori Sangat Baik).
4. Pembagian Data (Data Splitting): Korpus berlabel dipartisi menggunakan metode Stratified Random Sampling dengan rasio 80:20, guna memastikan proporsi setiap kelas sentimen terdistribusi secara seimbang dan proporsional pada kelompok data latih maupun data uji.
5. *Pemodelan CatBoost: Klasifikasi dieksekusi menggunakan algoritma CatBoost dengan konfigurasi parameter text_features yang mengaktifkan kamus dictionaries='Unigram' serta kalkulasi numerik internal melalui feature_calcs=['BoW', 'NaiveBayes', 'BM25']. Untuk mereduksi risiko overfitting, parameter kedalaman pohon simetris (depth) dioptimasi menjadi 5.*

Evaluasi: Kinerja model dievaluasi secara kuantitatif dengan menghitung nilai komputasi silang pada instrumen Confusion Matrix untuk memperoleh metrik Akurasi, Presisi, Recall, dan F1-Score.

HASIL PENELITIAN DAN PEMBAHASAN

Pra-Pemrosesan Teks dan Distribusi Data

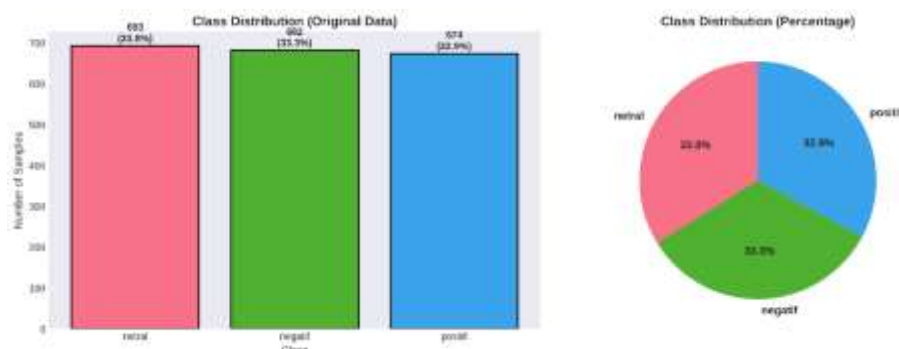
Berdasarkan proses penarikan data menggunakan web scraping, diperoleh dataset mentah sebanyak 2.100 observasi opini publik. Data tidak terstruktur ini kemudian melewati rantai filter linguistik pada Text Preprocessing yang mencakup Cleaning, Tokenizing, dan Stopword Removal untuk menghilangkan derau (noise). Contoh hasil transformasi teks pada tahap ini disajikan pada Tabel 1.

Tabel 1. Contoh Pra-Pemrosesan Teks

Sebelum <i>Cleaning</i>	Clean_text	Label
ekonomi.krna lingkungan egk bisa buat perut kenyang	ekonomi krna lingkungan egk buat perut kenyang	Netral
byk di korup om..bukan nya utk rakyat sejahtera ☹☹	byk korup om bukan nya rakyat sejahtera	Negatif

Sebelum Cleaning	Clean_text	Label
sabar guys, tanah bekas galian tambang itu minim unsur hara dan susah ditanami makanya harus diperbaiki dulu, cara paling efisien pake tanaman legum yg punya bintil akar untuk memperbaiki unsur tanah dan mengikat air sedikit demi sedikit agar tanaman besar bisa lebih mudah tumbuh	sabar guys tanah bekas galian tambang minim unsur hara susah ditanami makanya diperbaiki dulu cara paling efisien pakai tanaman legum punya bintil akar memperbaiki unsur tanah mengikat air sedikit sedikit tanaman besar lebih mudah tumbuh	Positif

Setelah pembersihan, tersisa 2.049 korpus data bersih yang tervalidasi. Analisis sebaran distribusi kelas sentimen pada keseluruhan dataset dapat dilihat pada Gambar 2.

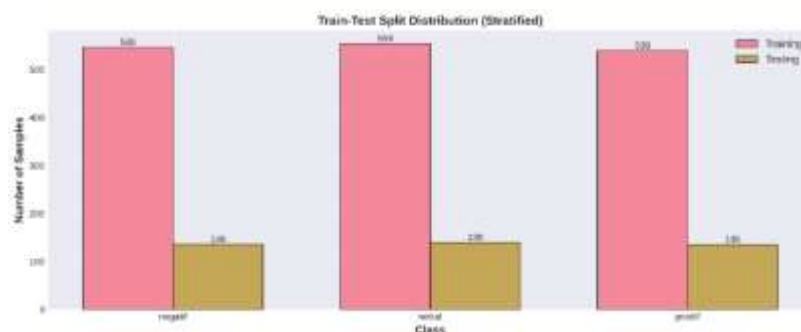


Gambar 2. Distribusi Kelas Sentimen pada Dataset

Berdasarkan Gambar 2, proporsi dataset menunjukkan kondisi yang sangat ideal dan seimbang, terdiri atas sentimen netral sebesar 33,8% (693 data), negatif sebesar 33,3% (682 data), dan positif sebesar 32,9% (674 data). Keseimbangan ini krusial untuk mencegah model terjebak pada bias kelas mayoritas saat fase pelatihan.

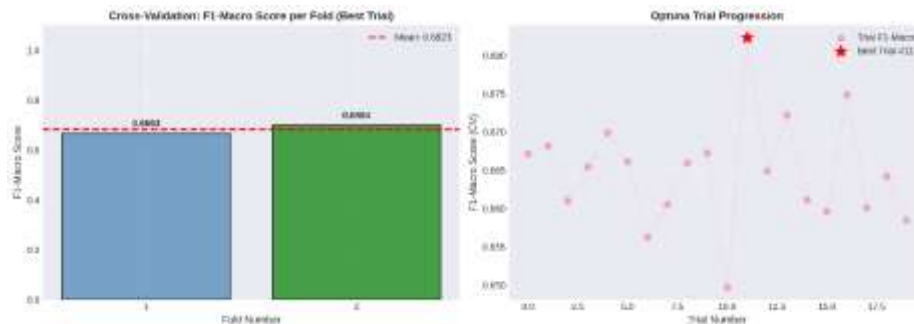
Pembagian Data (Data Splitting) dan Optimasi Model

Data dipartisi menggunakan metode Stratified Random Sampling dengan rasio 80:20. Sebanyak 1.639 data dialokasikan sebagai data latih dan 410 data diisolasi sebagai data uji. Pemerataan proporsi kelas pada pembagian ini divisualisasikan pada Gambar 3.



Gambar 3. Distribusi Pembagian Data Latih dan Data Uji

Optimasi hyperparameter algoritma CatBoost dilakukan melalui pustaka Optuna. Evaluasi Cross-Validation dan rekam jejak optimasi ditunjukkan pada Gambar 4, di mana konfigurasi terbaik (Best Trial) diperoleh pada iterasi = 450, learning_rate = 0,2955, dan depth = 5.



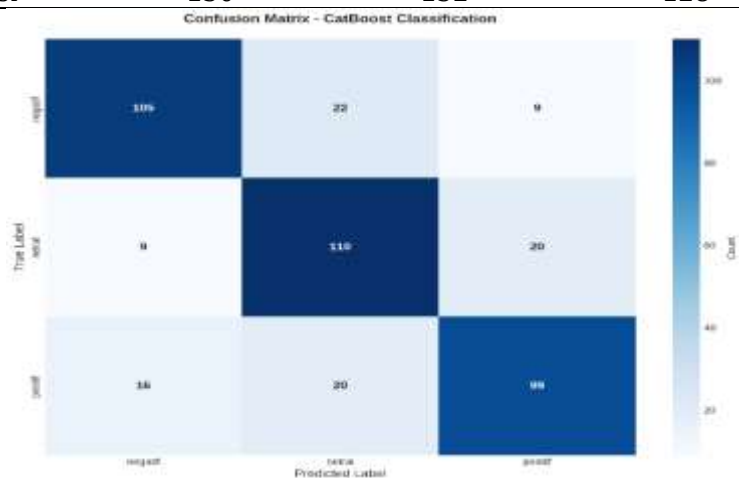
Gambar 4. Hasil Cross-Validation dan Optimasi Optuna

Evaluasi Model (Confusion Matrix)

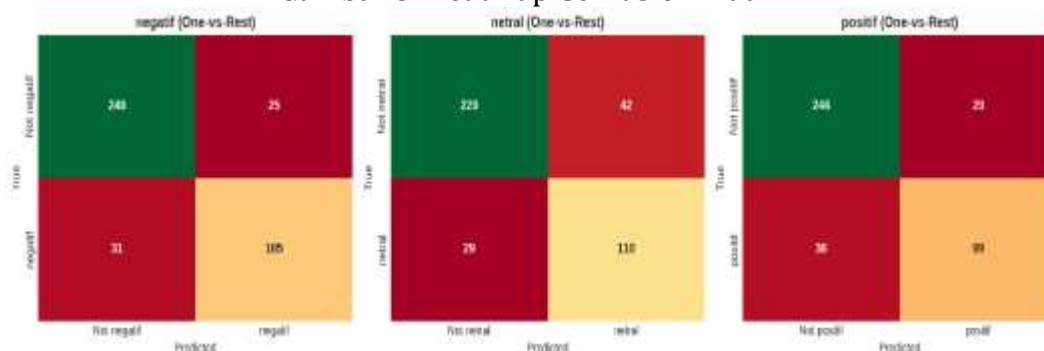
Pengujian performa model pada 410 data uji dievaluasi menggunakan instrumen Confusion Matrix. Perbandingan langsung antara prediksi algoritma CatBoost dengan label aktual dijabarkan pada Tabel 2 dan divisualisasikan melalui heatmap pada Gambar 5 dan Gambar 6.

Tabel 2. Hasil Confusion Matrix Pengujian CatBoost

Prediksi	Prediksi Negatif	Prediksi Netral	Prediksi Positif	Total Aktual
Negatif	100	22	9	136
Netral	9	118	20	139
Positif	16	20	105	135
Total Prediksi	130	152	128	410



Gambar 5. Heatmap Confusion Matrix



Gambar 6. One-vs-Rest Confusion Matrix per Kelas

Dari visualisasi di atas, tebakan akurat mendominasi pada sel diagonal utama (100 Negatif, 118 Netral, 105 Positif). Hal ini menghasilkan tingkat Akurasi Keseluruhan

(Overall Accuracy) sebesar 79% (0,7878). Penyesuaian parameter kedalaman pohon dari default depth=12 menjadi depth=5 terbukti sangat krusial; selain berhasil mereduksi risiko overfitting (memorisasi data latih), langkah ini juga memangkas beban komputasi dan mempercepat waktu pelatihan (training time) secara eksponensial.

KESIMPULAN

Penerapan algoritma CatBoost berbasis Text Features telah berhasil diimplementasikan secara terstruktur untuk mengklasifikasikan 2.049 teks opini publik dari media sosial terkait konflik ekologi dan ekonomi pertambangan. Fitur Native Text Handling pada algoritma terbukti mampu memecah dan memproses teks bahasa Indonesia tidak baku secara otonom di dalam sistem, mengeliminasi kebutuhan konversi matriks numerik eksternal yang memberatkan komputasi, sekaligus menjaga retensi konteks kalimat. Berdasarkan pengujian 5-Fold Cross Validation, algoritma CatBoost mencatatkan performa klasifikasi yang stabil dan andal, serta sukses mengungguli model baseline seperti SVM dan Random Forest. Tingginya ketepatan prediksi pada deteksi sentimen negatif menjadikan model komputasional ini memiliki potensi yang baik sebagai instrumen pendukung awal dalam sistem deteksi dini keresahan sosial. Kendati demikian, penelitian ini memiliki keterbatasan (research limitations) pada ukuran sampel yang hanya mencakup 2.049 observasi dan tingkat ketergantungan awal pada kamus leksikon dasar. Untuk penelitian masa depan, disarankan memperluas rentang skala dataset (di atas 10.000 data) dan mengombinasikan CatBoost dengan representasi berbasis konteks (seperti ekstraksi fitur awal menggunakan model bahasa pra-latih) untuk mendongkrak performa klasifikasi kalimat sarkasme di media sosial.

Pernyataan mengenai Penggunaan Kecerdasan Buatan (AI).

AI digunakan dalam penyusunan naskah ini. Penulis menggunakan perangkat kecerdasan buatan semata-mata untuk membantu perbaikan bahasa, pemeriksaan tata bahasa, serta penyuntingan demi kejelasan dan keterbacaan. Penulis tetap bertanggung jawab penuh atas akurasi, orisinalitas, integritas, dan konten akademis naskah tersebut. Seluruh konten yang dibantu oleh AI telah ditinjau dan diverifikasi oleh penulis sebelum naskah diserahkan.

Konflik kepentingan

Penulis menyatakan tidak ada konflik kepentingan terkait dengan penelitian, kepenulisan, dan/atau publikasi artikel ini.

Kontribusi penulis

Konseptualisasi, Syauqiyah Mujahidah Yahya, Muhyiddin A.M Hayat dan Desi Anggreani; pengumpulan data, Syauqiyah Mujahidah Yahya; validasi data, Muhyiddin A.M Hayat dan Desi Anggreani; analisis formal, Syauqiyah Mujahidah Yahya; penulisan—penyusunan draf awal, Syauqiyah Mujahidah Yahya; penulisan—peninjauan dan penyuntingan, Muhyiddin A.M Hayat dan Desi Anggreani; visualisasi, Syauqiyah Mujahidah Yahya. Seluruh penulis telah membaca dan menyetujui versi akhir naskah ini.

Ucapan terima kasih

Penelitian ini didukung oleh Program Studi Teknik Informatika, Universitas

Muhammadiyah Makassar. Penulis mengucapkan terima kasih yang sebesar-besarnya atas dukungan, fasilitas laboratorium, serta bimbingan yang telah diberikan selama proses penelitian hingga penyelesaian penyusunan naskah ini.

DAFTAR PUSTAKA

- Alshaari, H., & Alqahtani, S. (2025). Deep-EFNet: An Optimized EfficientNetB0 Architecture With Dual Regularization for Scalable Multi-Class Brain Tumor Classification in MRI. *IEEE Access*, 13(May), 85682–85697. <https://doi.org/10.1109/ACCESS.2025.3567919>
- Chen, X. (2025). Intelligent Employment Forecasting Model Based On Catboost Algorithm. *Procedia Computer Science*, 262, 596–603. <https://doi.org/10.1016/j.procs.2025.05.090>
- Kaur, G., & Sharma, A. (2023). A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of Big Data*, 10. <https://doi.org/10.1186/s40537-022-00680-6>
- Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1), 1–13. <https://doi.org/10.1186/s12880-022-00793-7>
- Lin, H., & Bu, N. (2022). A CNN-Based Framework for Predicting Public Emotion and Multi-Level Behaviors Based on Network Public Opinion. *Frontiers in Psychology*, 13. <https://doi.org/10.3389/fpsyg.2022.909439>
- Narasamma, L., & Sreedevi, M. (2021). Twitter based Data Analysis in Natural Language. *International Journal of Advanced Computer Science and Applications*, 12(5), 440–447.
- Purwono, Ma'arif, A., Rahmaniar, W., Fathurrahman, H. I. K., Frisky, A. Z. K., & Haq, Q. M. U. (2022). Understanding of Convolutional Neural Network (CNN): A Review. *International Journal of Robotics and Control Systems*, 2(4), 739–748. <https://doi.org/10.31763/ijrcs.v2i4.888>
- Segoni, S., Nocentini, N., Ajin, R. S., Gatto, A., & Fanti, R. (2025). Understanding the spatial patterns of major geo-hydrological disasters in Italy using the CatBoost algorithm. *Science of the Total Environment*, 999(July), 180355. <https://doi.org/10.1016/j.scitotenv.2025.180355>
- Suresh Kumar, K., Radha Mani, A. S., Ananth Kumar, T., Jalili, A., Gheisari, M., Malik, Y., Chen, H. C., & Jahangir Moshayedi, A. (2024). Sentiment Analysis of Short Texts Using SVMs and VSMs-Based Multiclass Semantic Classification. *Applied Artificial Intelligence*, 38(1). <https://doi.org/10.1080/08839514.2024.2321555>
- Zhao, Z., Liu, W., & Wang, K. (2023). Research on sentiment analysis method of opinion mining based on multi-model fusion transfer learning. *Journal of Big Data*, 10, 1–30. <https://doi.org/10.1186/s40537-023-00837-x>