Populer: Jurnal Penelitian Mahasiswa
Volume. 5 Nomor. 3 September 2026
E-ISSN: 2963-5306, P-ISSN: 2962-116X, Hal. 153-160
DOI: https://doi.org/10.58192/populer.v5i3.5517
Tersedia: https://journal.unimar-amni.ac.id/index.php/Populer

Deteksi *Cyberbullying* pada Twitter menggunakan Distilbert dan *(Explainable Ai)* Shap

Hadana Maulana[1*], Mustafa[2], Ghufron[3]
[1-3]Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Islam Sultan Agung Semarang, Indonesia
**Penulis Korespondensi: maulanahadana333@gmail.com*

Abstract*. The prevalence of cyberbullying on Twitter demands a fast and transparent automated moderation system. This study designs an efficient and explainable binary cyberbullying detection system using DistilBERT and the SHAP-based Explainable AI (XAI) method. The dataset, sourced from Kaggle, consists of 13,169 raw data filtered into 12,548 clean tweets. Preprocessing was conducted without stemming and stopword removal to preserve the language's semantic context, setting a max_length of 48 tokens. The pre-trained distilbert-base-multilingual-cased model was evaluated on an 80% training and 20% testing data split, while model transparency was validated using SHAP's PartitionExplainer algorithm. Evaluation results show DistilBERT performed robustly, achieving 84% accuracy, 84% precision, 86% F1-score, and 88% recall in detecting bullying, with an Area Under Curve (AUC) score reaching 0.9207. Locally, SHAP evaluation proved that the model concentrated heavy penalty weights on abusive words like "brengsek". Globally, SHAP revealed a unique phenomenon where the most important features were dominated by neutral vocabularies such as "times. roman" and "pendidikan", which the model cleverly utilized as strong discriminative indicators to validate normal texts. In conclusion, the integration of DistilBERT and SHAP yields a classification system that is not only accurate and efficient but also transparent and logically reasoned.*

Keywords*: Cyberbullying; DistilBERT; Explainable AI; NLP; SHAP.*

Abstrak. Maraknya perundungan siber (*cyberbullying*) di Twitter menuntut sistem moderasi otomatis yang cepat dan transparan. Penelitian ini merancang sistem deteksi *cyberbullying biner* yang efisien dan dapat dijelaskan menggunakan DistilBERT serta metode *Explainable AI* (XAI) berbasis SHAP. Dataset bersumber dari Kaggle sebanyak 13.169 data mentah yang disaring menjadi 12.548 tweet bersih. Proses prapemrosesan dilakukan secara minimal tanpa *stemming* dan *stopword removal* untuk mempertahankan konteks semantik bahasa, dengan penetapan *max_length* sebesar 48 token. Model *pre-trained distilbert-base-multilingual-cased* dievaluasi pada proporsi data latih 80% dan uji 20%, sedangkan transparansi model divalidasi menggunakan algoritma *PartitionExplainer* pada SHAP. Hasil pengujian menunjukkan DistilBERT berkinerja tangguh dengan akurasi 84%, presisi 84%, *F1-score* 86%, dan *recall* 88% dalam mendeteksi perundungan, serta skor *Area Under Curve* (AUC) mencapai 0.9207. Evaluasi SHAP membuktikan bahwa secara lokal, model memusatkan bobot pinalti besar pada kata makian seperti "brengsek". Secara global, SHAP mengungkap fenomena unik di mana fitur terpenting didominasi oleh kosakata netral seperti "times, roman" dan "pendidikan", yang dimanfaatkan model sebagai indikator diskriminatif kuat untuk memvalidasi teks normal. Kesimpulannya, integrasi DistilBERT dan SHAP menghasilkan sistem klasifikasi yang tidak hanya akurat dan efisien, tetapi juga transparan dan beralasan.

Kata Kunci: *Cyberbullying*; DistilBERT; *Explainable AI*; NLP; SHAP.

1. LATAR BELAKANG

Perkembangan teknologi informasi telah mengubah *platform* media sosial seperti Twitter (X) menjadi ruang publik utama untuk pertukaran informasi dengan jangkauan dan kecepatan yang sangat masif (Khaira et al., 2020). Namun, kemudahan ini memicu dampak negatif yang mengkhawatirkan, salah satunya adalah fenomena perundungan siber atau *cyberbullying* (Asalnaije et al., 2024). Laporan *Digital Civility Index* (DCI) tahun 2020 oleh *Microsoft* menempatkan warganet Indonesia pada posisi terburuk di kawasan Asia Tenggara terkait tingkat keadaban digital (Susanto et al., 2024). Hal ini diperkuat oleh data *ChildFund International* di Indonesia yang mencatat lebih dari 35% remaja pernah mengalami pelecehan

Naskah Masuk: 25 Juni 2026; Revisi: 16 Agustus 2026; Diterima: 18 September 2026; Terbit: 24 September 2026

daring, sebuah tindakan yang secara jelas melanggar hukum berdasarkan Undang-Undang ITE (Yudhistira et al., 2025). Dampak *cyberbullying* terhadap kesehatan mental korban sangat serius, meliputi kecemasan, depresi, hingga risiko bunuh diri (Khaira et al., 2020).

Banyak penelitian telah dilakukan untuk mendeteksi *cyberbullying* secara otomatis (Muftie et al., 2024). Arsitektur *Deep Learning* berbasis Transformer seperti BERT *(Bidirectional Encoder Representations from Transformers)* telah terbukti mengungguli metode *Machine Learning* klasik seperti SVM (Lukman & Novianto, 2025) karena kemampuannya dalam memahami konteks semantik secara utuh (Muftie et al., 2024). Akan tetapi, penggunaan model BERT memunculkan dua masalah fundamental. Pertama, model ini menuntut sumber daya komputasi yang sangat besar dan waktu pelatihan yang lama, mencapai 24 menit hingga 7200 detik dalam beberapa eksperimen (Basbeth, 2024) (Saputra et al., 2024). Kedua, BERT beroperasi sebagai model "kotak hitam" *(black box)* yang hasil prediksinya sulit diinterpretasikan logikanya (Tolla, 2025).

Penelitian ini diajukan untuk merancang sistem yang efisien dan transparan. Efisiensi dicapai menggunakan DistilBERT, yang berukuran 40% lebih kecil dan 60% lebih cepat dibandingkan BERT, namun tetap sebanding secara akurasi . Transparansi sistem dicapai melalui implementasi *Explainable AI* (XAI) berbasis SHAP *(SHapley Additive exPlanations)* yang memungkinkan visualisasi alasan di balik pengambilan keputusan model.(Basbeth, 2024; Ade Chandra Sapu (Tolla, 2025).

2. KAJIAN TEORITIS

Perundungan siber *(cyberbullying)* merupakan perilaku agresif yang dilakukan melalui teknologi digital seperti media sosial, yang mencakup tindakan penghinaan, pelecehan, ancaman, dan penyebaran informasi palsu sehingga membahayakan kesehatan mental korban (Asalnaije et al., 2024). Untuk mendeteksi perilaku tersebut secara otomatis, digunakan *Natural Language Processing* (NLP) yang merupakan cabang Kecerdasan Buatan (AI) untuk memampukan komputer memahami bahasa manusia, yang sekaligus menjadi fondasi utama dalam analisis sentimen serta klasifikasi teks (Azmi Verdikha et al., 2023).

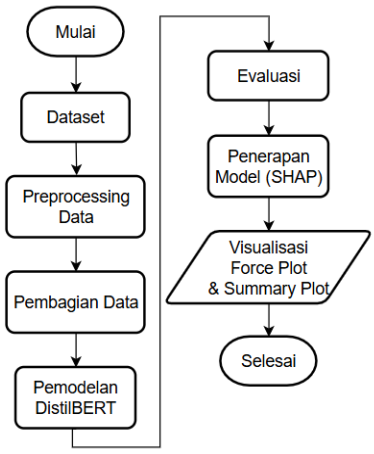
Dalam pemrosesan teks, tahap *preprocessing* umumnya digunakan untuk mengeliminasi ketidakteraturan data asli agar siap dianalisis (Muftie et al., 2024). Namun, arsitektur Transformer seperti DistilBERT memiliki metodologi *preprocessing* yang berbeda, di mana penerapan *stemming* justru berisiko menghilangkan makna kontekstual kalimat (Saraswati et al., 2025). DistilBERT sendiri adalah model Transformer yang dirancang lebih ringan melalui proses *Knowledge Distillation*, di mana model ini bertindak sebagai "siswa"

yang meniru probabilitas model "guru" atau BERT. Model ini terbukti mampu mempertahankan 97% pemahaman bahasanya dengan waktu komputasi yang jauh lebih singkat dibandingkan BERT(Saputra et al., 2024) (Fajri et al., 2022)(Basbeth, 2024)

Guna mengatasi sifat "kotak hitam" *(black box)* pada model *Deep Learning* seperti DistilBERT, diterapkan metode *Explainable AI* (XAI) agar logika pengambilan keputusan model dapat dipahami secara transparan (Tolla, 2025). Metode XAI yang digunakan adalah SHAP *(SHapley Additive exPlanations)*, yang bekerja dengan menghitung bobot pengaruh atau nilai *Shapley* dari setiap kata (Tolla, 2025). Analisis SHAP menyediakan visualisasi *Force Plot* untuk penjelasan lokal pada tingkat prediksi kalimat individu, serta *Summary Plot* untuk merangkum penjelasan global terhadap keseluruhan fitur yang paling penting bagi model (Anjani et al., 2023).

3. METODE PENELITIAN

Penelitian ini menggunakan dataset sekunder publik *"Indonesian Abusive and Hate Speech Twitter Text"* bersumber dari *platform Kaggle* berisi 13.169 *tweet* berbahasa Indonesia . Teks dikategorikan secara biner menjadi Label 1 *(Bullying)* jika kolom HS atau *Abusive* bernilai 1, dan Label 0 (Bukan *Bullying*) jika keduanya bernilai 0. Secara keseluruhan, tahapan dalam penelitian ini divisualisasikan pada bagan berikut:



Gambar 1. *Flowchart* Alur Penelitian.

Pada tahap *preprocessing* data, penelitian ini tidak melakukan *stemming* dan *stopword removal* karena model bahasa (Transformer) dilatih untuk memahami konteks secara utuh, dan pemotongan kata berisiko menyebabkan kehilangan makna *(loss of sense)* dan nuansa *(loss of nuance)*(Saraswati et al., 2025). Pembersihan dilakukan secara minimal meliputi *Decoding* HTML, *Hex/Encoding Cleaning*, *Removal of Platform Specifics* (URL, RT, USER), *Character Filtering*, *Whitespace Normalization*, dan *Case Folding* .

Data dibagi menggunakan teknik *random shuffling* (seed=42), menghasilkan 80% data latih (10.038 baris) dan 20% data uji (2.510 baris). Model pemrosesan menggunakan tokenisasi *sub-word* dari *distilbert-base-multilingual-cased* dengan *max_length* 48 token. Proses dilanjutkan dengan pemodelan *fine-tuning* selama 3 *epoch*, di mana lapisan klasifikasi menggunakan fungsi aktivasi *Softmax* dengan persamaan berikut:

$$softmax(z_i) = \frac{exp(z_i)}{\sum_{j=1}^{V} exp(z_j)} \quad \text{(i)}$$

Kinerja model kemudian diukur menggunakan perhitungan metrik *Confusion Matrix* (Akurasi, Presisi, *Recall*, *F1-Score*) dan kurva ROC-AUC . Secara matematis, rumus perhitungan metrik performa tersebut adalah sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(ii)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{(iii)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{(iv)}$$

$$F1 - Score = 2\text{x} \frac{presisi\ x\ recall}{(presisi+recall)} \quad \text{(v)}$$

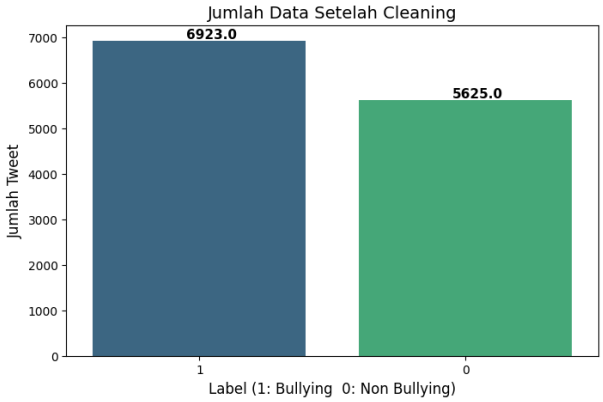
Tahap akhir adalah interpretasi menggunakan *PartitionExplainer* dari modul SHAP. Perhitungan bobot pengaruh kata didasarkan pada nilai *Shapley* (ϕi) yang dirumuskan sebagai berikut:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} [f(S \cup \{i\}) - f(S)] \quad \text{(vi)}$$

4. HASIL DAN PEMBAHASAN

Hasil *Preprocessing* dan Karakteristik Data

Berdasarkan proses data *filtering*, program berhasil membersihkan 621 baris data sampah, menyisakan 12.548 data teks bersih.

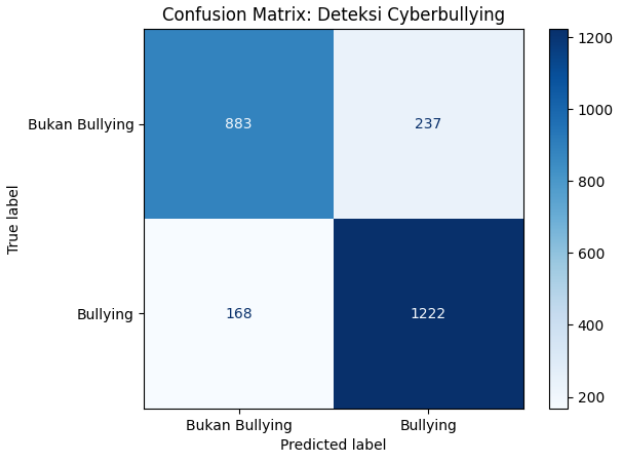


Gambar 2. Grafik Distribusi Kelas setelah *Cleaning.*

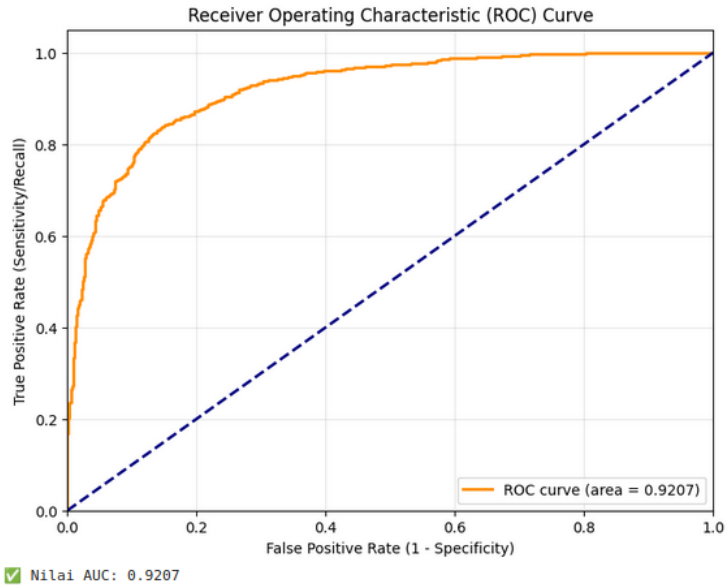
Distribusi kelas pada data bersih menunjukkan 6.923 teks *Bullying* dan 5.625 teks Bukan *Bullying*. Proporsi ini tergolong seimbang dan layak diproses langsung ke tahap pelatihan tanpa perlu teknik penyeimbangan data sintetis (Thabtah et al., 2019).

Evaluasi Kinerja Deteksi DistilBERT

Evaluasi kinerja model pada 2.510 data uji menghasilkan 1.222 *True Positive*, 883 *True Negative*, 237 *False Positive*, dan 168 *False Negative*.



Gambar 3. *Confusion Matrix.*

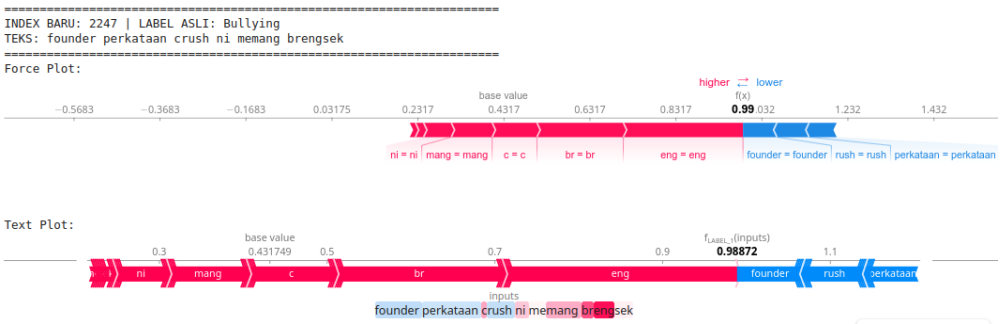


Gambar 4. Kurva ROC.

Secara performa metrik, model DistilBERT mencatatkan hasil tingkat akurasi sebesar 84%, presisi 84%, *recall* 88%, dan *F1-score* 86%. Kurva ROC menunjukkan skor *Area Under Curve* (AUC) mencapai 0.9207, yang menegaskan ketangguhan model dalam memisahkan konteks kalimat perundungan dan *non*-perundungan secara sangat istimewa (Wisdayani et al., 2019).

Evaluasi Penjelasan Model (*Explainable AI* SHAP)

Melalui hasil validasi lokal *(Local Explanation)* SHAP, visualisasi *Force Plot* menunjukkan bahwa probabilitas prediksi *Bullying* didorong secara dominan oleh blok warna merah pada token yang membentuk kata makian seperti "brengsek".



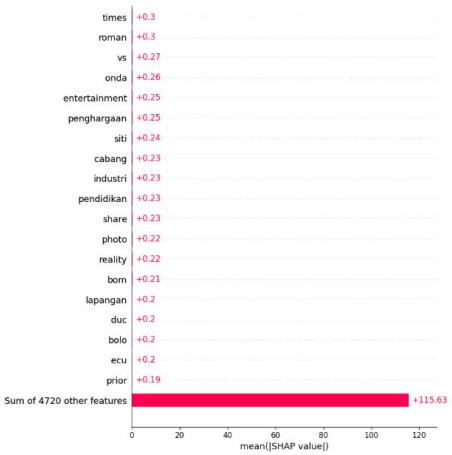
Gambar 5. *Force Plot Bullying*.



Gambar 6. *Force Plot Non-Bullying.*

Sedangkan pada teks normal, gaya tarik positif warna biru dari kata-kata informatif seperti "presiden" atau "kabinet" secara aktif menahan laju probabilitas agar tidak masuk ke ambang batas perundungan.

Dalam penjelasan global *(Global Explanation)*, grafik SHAP *Summary Bar Plot* menunjukkan pola anomali.



Gambar 7. SHAP *Summary Bar Plot* .

Fitur dengan bobot pengaruh tertinggi justru didominasi oleh kata netral, formal, dan informatif seperti "times. roman", "pendidikan", dan "industri". Fenomena ini terjadi akibat dispersi kosakata makian yang bervariasi luas di media sosial, sementara model secara cerdas mampu mengidentifikasi kosakata baku sebagai filter pembatas utama untuk memvalidasi kelas Bukan *Bullying*.

5. KESIMPULAN DAN SARAN

Berdasarkan pengujian, model klasifikasi *cyberbullying* menggunakan arsitektur DistilBERT berhasil dibangun melalui teknik *fine-tuning* dengan prapemrosesan teks minimal tanpa *stemming* dan *stopword removal*. Model berkinerja tangguh dengan mencatatkan akurasi 84%, presisi 84%, *recall* 88%, *F1-score* 86%, dan AUC 0,9207. Implementasi *Explainable AI* berbasis SHAP berhasil memetakan bahwa model dipicu oleh fitur kata makian secara spesifik pada interpretasi lokal, serta menggunakan kosakata berkonteks formal sebagai filter utama kelas Bukan *Bullying* pada interpretasi global.

Penelitian selanjutnya disarankan untuk memanfaatkan atribut dataset secara menyeluruh guna merancang klasifikasi multi-kelas *(multi-class classification)*, mengkomparasikan performa dengan *pre-trained* model spesifik bahasa Indonesia (seperti IndoBERT), serta menambah dataset kalimat sarkasme untuk mengenali bentuk perundungan secara implisit.

DAFTAR REFERENSI

Anjani, A. F., Anggraeni, D., & Tirta, I. M. (2023). Implementasi Random Forest Menggunakan SMOTE untuk Analisis Sentimen Ulasan Aplikasi Sister for Students UNEJ. *Jurnal Nasional Teknologi dan Sistem Informasi*, *9*(2), 163–172. https://doi.org/10.25077/TEKNOSI.v9i2.2023.163-172

Asalnaije, E., Bete, Y., Tira, A. D., & Lian, Y. P. (2024). *Bentuk-Bentuk CyberbullyingDi Indonesia*.

Azmi Verdikha, N., Habid, R., & Johar Latipah, A. (2023). Analisis DistilBERT dengan Support Vector Machine (SVM) untuk Klasifikasi Ujaran Kebencian pada Sosial Media Twitter. *METIK JURNAL*, *7*(2), 101–110. https://doi.org/10.47002/metik.v7i2.583

Basbeth, F. (2024). *Klasifikasi Emosi Data Text Bahasa Indonesia Menggunakan Algoritma BERT, RoBERTa, dan DistilBERT*.

Fajri, F., Tutuko, B., & Sukemi, S. (2022). Membandingkan Nilai Akurasi BERT dan DistilBERT pada Dataset Twitter. *JUSIFO (Jurnal Sistem Informasi)*, *8*(2), 71–80. https://doi.org/10.19109/jusifo.v8i2.13885

Khaira, U., Johanda, R., Utomo, P. E. P., & Suratno, T. (2020). Sentiment Analysis Of Cyberbullying On Twitter Using SentiStrength. *Indonesian Journal of Artificial Intelligence and Data Mining*, *3*(1), 21. https://doi.org/10.24014/ijaidm.v3i1.9145

Lukman, K., & Novianto, S. (2025). Komparasi Algoritma Naïve Bayes dan SVM untuk Identifikasi Cyberbullying Selebriti di Media Sosial Twitter. *Jurnal Algoritma*, *22*(1), 970–981. https://doi.org/10.33364/algoritma/v.22-1.2196

Muftie, F., Yafi, K. M., & Addina, Q. M. (2024). Perbandingan performa deteksi cyberbullying dengan transformer, deep learning, dan machine learning. *Jurnal Pendidikan Informatika dan Sains*, *13*(1), 75–87. https://doi.org/10.31571/saintek.v13i1.4002

Saputra, A. C., Saragih, A. S., & Ronaldo, D. (2024). PERBANDINGAN NILAI AKURASI DISTILBERT DAN BERT PADA DATASET ANALISIS SENTIMEN LEMBAGA KURSUS. *Jurnal Teknologi Informasi: Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika*, *18*(2), 160–171. https://doi.org/10.47111/jti.v18i2.13278

Saraswati, N. W. S., Yanti, C. P., Muku, I. D. M. K., & Dewi, D. A. P. R. (2025). Evaluation Analysis of the Necessity of Stemming and Lemmatization in Text Classification. *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, *24*(2), 321–332. https://doi.org/10.30812/matrik.v24i2.4833

Susanto, E., Budimansyah, D., Komalasari, K., & Masyitoh, I. S. (2024). Pendidikan Kewarganegaraan Digital: Upaya Meningkatkan Keadaban Digital pada Warga Negara Muda Indonesia. *Jurnal Moral Kemasyarakatan*, *9*(2). https://doi.org/10.21067/jmk.v9i2.10337

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2019). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Tolla, Y. (2025). *Deteksi Stres dan Depresi Unggahan Media Sosial dengan Machine Learning*. *15*(1).

Wisdayani, D. S., Nur, I. M., & Wasono, R. (2019). *PERBANDINGAN ALGORITMA K-NEAREST NEIGHBOR DAN NAIVE BAYES UNTUK KLASIFIKASI TINGKAT KEPARAHAN KORBAN KECELAKAAN LALU LINTAS DI KABUPATEN PATI JAWA TENGAH*.

Yudhistira, D., Saraswati, G. W., & Lutfina, E. (2025). *ANALISIS SENTIMEN OPINI MASYARAKAT INDONESIA TERHADAP KASUS CYBERBULLYING DI MEDIA SOSIAL X (TWITTER) MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE.*