JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 2 Tahun 2026
ISSN 3025-0919 (media online)
Hal 459-464

Klasifikasi Pelanggaran Etika Siber pada Media Sosial X Menggunakan Algoritma Extreme Gradient Boosting Berbasis TF-IDF

Faisal Aditya Pratama[1], Fijriani Silviana[2], Muhammad Karifki[3], Muhammad Wahyu Muges[4], Sherly Septiani[5], Rahmawati[6]

[123456] Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1] faisalpratama831@gmail.com, [2] fijrianis@gmail.com, [3] karifkimuhammad@gmail.com, [4]wahyumuges@gmail.com, [5] sherlyseptiani1809@gmail.com, [6] dosen02394@unpam.ac.id
(* : coressponding author)

Abstrak−Perkembangan platform sosial memiliki dampak signifikan terhadap cara orang berkomunikasi dalam dunia digital. Salah satu situs yang populer adalah media sosial X, yang memberikan kesempatan bagi penggunanya untuk menyampaikan pendapat dengan cepat dan tanpa batas. Namun, kebebasan ini sering kali memicu berbagai persoalan terkait pelanggaran norma etika di dunia maya, seperti ujaran kebencian, perundungan siber, penghinaan, dan komentar negatif lainnya. Tujuan dari penelitian ini adalah untuk mengklasifikasikan pelanggaran etika di media sosial X melalui algoritma Extreme Gradient Boosting (XGBoost) dengan menggunakan data dari Kaggle. Metode yang digunakan dalam studi ini mencakup pengumpulan data, pemrosesan awal teks, ekstraksi fitur dengan TF-IDF, pelatihan model dengan algoritma XGBoost, dan penilaian kinerja model melalui pengukuran akurasi, presisi, recall, dan skore F1. Dataset yang digunakan berasal dari Kaggle, berisi komentar-komentar yang dikelompokkan sebagai pelanggaran etika di dunia maya serta komentar yang normal. Temuan dari penelitian ini menunjukkan bahwa algoritma XGBoost dapat mengklasifikasikan kemungkinan terjadinya pelanggaran etika di komentar media sosial dengan tingkat akurasi yang sangat baik.

Kata Kunci: Klasifikasi Pelanggaran Etika Siber; Komentar Media Sosial X; Machine Learning; Extreme Gradient Boosting (XGBoost); Text Mining

Abstract−*The development of social platforms has had a significant impact on how people communicate in the digital world. One popular site is social media X, which provides its users with the opportunity to express their opinions quickly and without limits. However, this freedom often triggers various issues related to violations of ethical norms online, such as hate speech, cyberbullying, insults, and other negative comments. The aim of this research is to classify ethical violations on social media X using the Extreme Gradient Boosting (XGBoost) algorithm with data from Kaggle. The methods used in this study include data collection, text preprocessing, feature extraction using TF-IDF, model training with the XGBoost algorithm, and model performance evaluation through accuracy, precision, recall, and F1-score measurements. The dataset used is sourced from Kaggle and contains comments categorized as online ethical violations as well as normal comments. The findings of this study show that the XGBoost algorithm can classify the occurrence of ethical violations in social media comments with excellent accuracy.*

Keywords: *Cyber Ethics Violation Classification; X Social Media Comments; Machine Learning; Extreme Gradient Boosting (XGBoost); Kaggle Dataset; Text Mining*

1. PENDAHULUAN

Kemajuan teknologi informasi pada era digital telah mengubah cara masyarakat dalam berkomunikasi dan bertukar informasi. Internet serta media sosial menjadi sarana utama yang digunakan untuk menyampaikan pendapat, berbagi informasi, hingga membangun interaksi sosial secara daring. Salah satu platform yang banyak dimanfaatkan masyarakat adalah X yang memungkinkan pengguna memberikan komentar dan opini secara cepat serta terbuka kepada publik. Namun, tingginya kebebasan dalam penggunaan media sosial juga memunculkan berbagai permasalahan, terutama terkait pelanggaran etika siber seperti ujaran kebencian, penghinaan, cyberbullying, sarkasme, dan penyebaran komentar negatif (Mahmudah & Yudhistira, 2025).

Pelanggaran etika siber pada media sosial menjadi perhatian penting karena dapat memberikan dampak negatif bagi pengguna. Cyberbullying atau perundungan siber merupakan tindakan menyerang, merendahkan, maupun menghina seseorang melalui media digital. Fenomena

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 2 Tahun 2026
ISSN 3025-0919 (media online)
Hal 459-464

tersebut semakin meningkat seiring berkembangnya media sosial karena pengguna dapat bertindak secara anonim sehingga lebih mudah melakukan tindakan yang merugikan orang lain. Dampak dari cyberbullying tidak hanya memengaruhi hubungan sosial, tetapi juga berdampak pada kondisi psikologis korban seperti stres, kecemasan, depresi, serta menurunnya rasa percaya diri. Selain itu, penyebaran komentar negatif pada media sosial berlangsung sangat cepat sehingga sulit dilakukan pengawasan secara manual.

Beberapa penelitian sebelumnya menunjukkan bahwa media sosial sering dimanfaatkan sebagai sarana penyebaran komentar yang mengandung unsur pelanggaran etika siber. Penelitian mengenai analisis sentimen cyberbullying pada media sosial X menjelaskan bahwa rendahnya pemahaman pengguna terhadap etika bermedia sosial menjadi salah satu faktor meningkatnya perilaku negatif di dunia maya (Mahmudah & Yudhistira, 2025). Penelitian tersebut juga membuktikan bahwa penerapan machine learning mampu membantu proses deteksi cyberbullying secara otomatis melalui analisis data teks.

Dalam proses analisis komentar media sosial, tahapan text preprocessing menjadi bagian penting untuk meningkatkan kualitas data sebelum dilakukan klasifikasi. Tahapan tersebut meliputi cleaning, case folding, normalization, tokenization, stopword removal, dan stemming. Setelah data dibersihkan, teks diubah menjadi bentuk numerik menggunakan metode TF-IDF agar dapat diproses oleh algoritma machine learning (Felix Fernando, 2025).

Salah satu algoritma machine learning yang memiliki performa baik dalam analisis data teks adalah Extreme Gradient Boosting (XGBoost). Algoritma XGBoost merupakan metode boosting berbasis decision tree yang mampu menghasilkan performa klasifikasi dengan tingkat akurasi tinggi dan efisiensi komputasi yang baik (Felix Fernando, 2025). Pada penelitian sebelumnya mengenai klasifikasi cyberbullying menggunakan algoritma SVM dan XGBoost, metode XGBoost memperoleh hasil akurasi yang lebih baik dibandingkan SVM dengan nilai akurasi mencapai 83%. Hasil tersebut menunjukkan bahwa algoritma XGBoost memiliki kemampuan yang cukup efektif dalam mendeteksi komentar yang mengandung unsur cyberbullying maupun pelanggaran etika siber.

Berdasarkan permasalahan tersebut, penelitian ini dilakukan untuk mengklasifikasikan pelanggaran etika siber pada komentar media sosial X menggunakan algoritma Extreme Gradient Boosting (XGBoost) berbasis metode pembobotan TF-IDF. Dataset yang digunakan berupa kumpulan komentar media sosial yang telah dikategorikan berdasarkan komentar normal dan komentar yang mengandung pelanggaran etika siber. Penelitian ini diharapkan dapat membantu proses deteksi otomatis terhadap komentar negatif sehingga dapat mendukung terciptanya lingkungan digital yang lebih aman, nyaman, dan sehat bagi pengguna media sosial.

2. METODE

2.1 Tahapan Penelitian

Tahapan penelitian yang dilakukan pada penelitian ini terdiri dari beberapa proses yaitu pengumpulan dataset, preprocessing data, ekstraksi fitur menggunakan TF-IDF, pembagian data training dan testing, implementasi algoritma XGBoost, serta evaluasi model.
Tahapan Penelitian:

1. Pengumpulan dataset komentar media sosial dari Kaggle.
2. Melakukan preprocessing data teks.
3. Mengubah data teks menjadi numerik menggunakan TF-IDF.
4. Membagi dataset menjadi data training dan testing.
5. Melakukan pelatihan model menggunakan algoritma XGBoost.
6. Mengevaluasi performa model menggunakan confusion matrix.

2.2 Dataset penelitian

Data tambahan yang digunakan dalam studi ini berasal dari platform penyimpanan data Kaggle yang dikenal sebagai Cyberbullying Classification Dataset. Dataset ini terdiri dari 47.692 entri cuitan (tweet) yang ditulis dalam bahasa Inggris dari media sosial X. Terdapat dua elemen utama dalam dataset ini yang digunakan untuk tahap pemodelan, yaitu kolom tweet_text yang

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 2 Tahun 2026
ISSN 3025-0919 (media online)
Hal 459-464

berfungsi sebagai atribut meta (data teks yang akan dipelajari) dan kolom cyberbullying_type yang berfungsi sebagai target atau label kelas yang mengidentifikasi jenis perundungan.

2.3 Preprocessing Data

Untuk mempersiapkan teks mentah agar dapat dipahami oleh algoritma klasifikasi, dilakukan serangkaian proses pembersihan teks (preprocessing) melalui widget Preprocess Text pada Orange Data Mining. Tahapan ini meliputi:

1. Case Folding merupakan tahap awal untuk mengubah seluruh huruf kapital dalam teks menjadi huruf kecil (lowercase) agar format data menjadi seragam.
2. Tokenization dilakukan guna memecah kalimat utuh menjadi potongan kata per kata (word) sehingga mesin lebih mudah dalam melakukan analisis frekuensi.
3. Cleaning diterapkan melalui fitur filtering lexicon dan regexp untuk menghapus karakter khusus, tanda baca, tautan URL, angka, serta simbol yang tidak memiliki makna secara semantik.
4. Stopword Removal berfungsi menghilangkan kata-kata hubung umum dalam bahasa Inggris (seperti is, am, are, the) yang tidak memiliki bobot sentimen.
5. Feature Selection diterapkan untuk membatasi jumlah kata yang diekstrak menjadi hanya 1000 term utama yang paling sering muncul, sehingga kinerja komputasi mesin lebih optimal tanpa membuang informasi penting.

2.4 Metode TF-IDF

Metode Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk mengubah data teks menjadi representasi numerik. TF-IDF merupakan metode pembobotan yang sangat umum digunakan dalam text mining dan analisis sentimen untuk mengevaluasi seberapa penting sebuah kata di dalam suatu dokumen (Munna & Zuliarso, 2024). Metode ini menggabungkan frekuensi kemunculan sebuah kata dalam dokumen tertentu (TF) dan mengukur seberapa jarang kata tersebut muncul di seluruh dokumen dalam dataset (IDF), sehingga memberikan bobot lebih rendah pada kata yang terlalu umum (Kirana et al., 2025).

$$\text{TF-IDF(t,d)} = \text{TF(t,d)} \times \text{IDF(t)}$$

2.5 Algoritma XGBoost

Extreme Gradient Boosting (XGBoost) adalah algoritma pembelajaran mesin yang mengandalkan teknik peningkatan untuk memperbaiki kemampuan prediksi dengan menggabungkan beberapa pohon keputusan. Algoritma ini sering dipakai dalam pengklasifikasian data karena menawarkan kecepatan pemrosesan yang tinggi serta akurasi yang memadai. Di samping itu, XGBoost mampu mengelola data dalam jumlah besar, memiliki kemampuan adaptasi dalam berbagai jenis prediksi, dan dapat mengurangi risiko terjadinya overfitting pada model-model pembelajaran mesin. (Felix Fernando, 2025).

$$x_i = W_{t,d} = TF(t,d)x\ log_{10}\left(\frac{N}{DF(t)}\right) \quad (1)$$

Nilai x_i ini selanjutnya akan dimasukkan ke dalam fungsi prediksi kumulatif XGBoost guna menetapkan kelas akhir (sentimen/kategori bullying):

$$\hat{y}i = \sum_{k=1}^{K} f_k(x_i) \quad (2)$$

Keterangan:

1. x_i : Fitur input untuk sampel ke-i (berupa nilai bobot TF-IDF kata).
2. $W_{t,d}$: Nilai bobot TF-IDF kata t pada dokumen d.
3. $\hat{y}i$: Hasil prediksi akhir dari model XGBoost.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 2 Tahun 2026
ISSN 3025-0919 (media online)
Hal 459-464

4. $f_k(x_i)$: Hasil prediksi dari pohon keputusan (decision tree) ke-k.
5. K : Total seluruh pohon (tree) yang dibangun dalam model XGBoost.

2.6 Evaluasi Model

Pengujian performa algoritma dilakukan dengan menggunakan metode Data Sampler yang menerapkan fixed proportion of data, di mana dataset dibagi menjadi 80% sebagai data latih (training data) dan 20% sebagai data uji (testing data). Hasil prediksi kemudian diukur menggunakan Test on test data berdasarkan parameter Classification Accuracy (Akurasi), Precision (Presisi), Recall, dan F1-Score. Berdasarkan nilai TP, TN, FP, dan FN yang terdapat dalam tabel di atas, efektivitas model ditentukan dengan memanfaatkan empat parameter utama melalui rumus yang adadi bawah ini:

1. Akurasi Klasifikasi
   Menilai persentase total dari prediksi yang tepat (baik positif maupun negatif) yang dihasilkan oleh model.

$$Accuracy = \frac{TP \ + \ TN}{TP + TN + FP + FN}$$

2. Presisi
   Menilai seberapa akurat data yang diminta dibandingkan dengan hasil prediksi yang diberikan oleh model.

$$Precision = \frac{TP}{TP + FP}$$

3. Sensitivitas
   Menilai seberapa efisien model dalam mengidentifikasi kembali informasi atau kelas yang ditargetkan.

$$Recall \ = \frac{TP}{TP + FN}$$

4. F1-Score
   Nilai rata-rata harmonis yang menghimpun presisi dan sensitivitas untuk menilai keseimbangan kinerja model secara keseluruhan.

$$F1 - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision\ +\ Recall}$$

Keterangan variabel:

1. TP (*True Positive*) : Jumlah data komentar melanggar etika siber yang dengan tepat diidentifikasi sebagai pelanggaran etika siber oleh model.
2. TN (True Negative) : Jumlah data komentar normal yang diprediksi secara benar sebagai komentar normal oleh model.
3. FP (*False Positive*) : Jumlah data komentar yang biasa yang secara keliru diidentifikasi sebagai komentar pelanggaran etika siber oleh model (salah anggap).
4. FN (*False Negative*):Jumlah data komentar pelanggaran etika siber yang secara keliru diidentifikasi sebagai komentar biasa oleh model (terlewat deteksi).

3. ANALISA DAN PEMBAHASAN

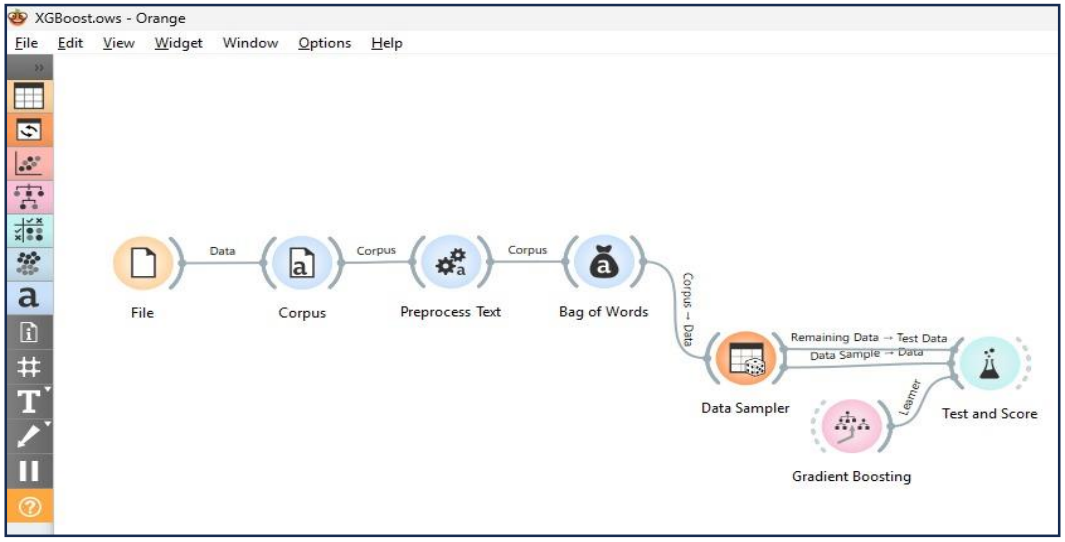
3.1 Hasil Preprocessing

Proses pra-pemrosesan dilaksanakan pada 47.692 cuitan yang diambil dari platform media sosial X. Teks asli yang menunjukkan variasi dalam format, penggunaan huruf kapital, simbol, serta gangguan lainnya telah berhasil dibersihkan. Dengan menerapkan metode pemilihan fitur pada tahap penyaringan, jumlah ribuan kata yang unik dalam kumpulan data teks berhasil dikurangi dan difokuskan menjadi 1000 istilah utama. Tahapan pengurangan ini terbukti sangat penting untuk

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 2 Tahun 2026
ISSN 3025-0919 (media online)
Hal 459-464

menghindari masalah kelebihan beban komputasi ketika algoritma mengubah teks menjadi matriks angka.

Dataset yang telah dikonversi menjadi bobot angka melalui metode TF-IDF kemudian didistribusikan ke dalam proses pemodelan. Berdasarkan pembagian rasio 80:20, sebanyak sekitar 38.153 baris data digunakan oleh algoritma XGBoost (Gradient Boosting) untuk mempelajari pola perundungan siber (training phase). Model membangun struktur decision trees secara berulang untuk membedakan antara komentar normal dan komentar yang mengandung pelanggaran etika siber berdasarkan 1000 fitur kata yang telah disaring sebelumnya.



3.2 Hasil Evaluasi Model

Setelah model XGBoost dilatih, algoritma kemudian diuji menggunakan sisa 20% data uji (sekitar 9.538 baris data) yang belum pernah dilihat oleh sistem sebelumnya. Berdasarkan hasil pengujian pada widget Test and Score, diperoleh metrik evaluasi kinerja model sebagai berikut:

1. Classification Accuracy (CA): 0.824 (82.4%)
2. Precision (Prec): 0.843 (84.3%)
3. Recall: 0.824 (82.4%)
4. F1-Score (F1): 0.824 (82.4%)

Tabel 1. Hasil Evaluasi Model

Metode	Accuracy	Precision	Recall	F1-Score
XGBoost	82,4%	84,3%	82,4%	82,4%

3.3 Pembahasan

Berdasarkan metrik evaluasi di atas, penerapan algoritma XGBoost menunjukkan kinerja yang sangat baik dalam menganalisis pelanggaran etika siber. Nilai Akurasi (CA) sebesar 82.4% mengindikasikan bahwa model mampu menebak kategori sebuah cuitan secara tepat pada 82.4% kasus dari total data uji.

Nilai Presisi yang menyentuh angka 84.3% menunjukkan tingkat kehati-hatian model. Artinya, ketika model memprediksi bahwa suatu komentar adalah pelanggaran etika, prediksi tersebut 84.3% dapat dipercaya dan bukan tuduhan keliru (false positive). Di sisi lain, nilai Recall dan F1-Score yang sama-sama berada di angka 82.4% membuktikan bahwa model XGBoost yang dikembangkan sangat stabil dan seimbang. Kemampuan komputasi XGBoost berhasil membedah kerumitan semantik bahasa gaul, singkatan, dan konteks pada dataset media sosial menjadi prediksi klasifikasi yang valid dan robust.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 2 Tahun 2026
ISSN 3025-0919 (media online)
Hal 459-464

4. KESIMPULAN

Berdasarkan hasil analisis prediksi pelanggaran etika siber pada komentar media sosial X menggunakan algoritma Extreme Gradient Boosting (XGBoost) berbasis dataset Kaggle, dapat disimpulkan bahwa:

1. Implementasi algoritma XGBoost dengan penggabungan metode ekstraksi teks TF-IDF telah terbukti berhasil dalam menangani data bahasa alami. Pembatasan jumlah token hingga maksimum 1000 kata dapat dengan jelas meningkatkan efisiensi waktu pemrosesan tanpa mengorbankan kualitas representasi informasi yang penting.
2. Model klasifikasi XGBoost menunjukkan kemampuan prediksi yang sangat baik dengan tingkat Akurasi (CA) mencapai 82.4%, Presisi 84.3%, Recall 82.4%, dan F1-Score 82.4%. Temuan ini mengindikasikan bahwa algoritma pembelajaran mesin yang didasarkan pada pohon keputusan ini sangat sesuai dan dapat diandalkan untuk pengembangan lebih lanjut sebagai sistem penyaringan atau deteksi otomatis terhadap ujaran kebencian, penindasan daring, dan pelanggaran etika lainnya demi mewujudkan ekosistem digital yang lebih baik.

REFERENCES

Applications, E. (2025). *Implementation of TF-IDF and XGBoost Algorithms in Scientific Paper Classification*. *5*(1), 1–5.

Felix Fernando. (2025). Klasifikasi Tweet Cyberbullying Dengan Menggunakan Algoritma Svm Dan Xgboost. *Jurnal Ilmu Komputer Dan Sistem Informasi*, *13*(1). https://doi.org/10.24912/jiksi.v13i1.32857

Kairupan, I. Y., Angdresey, A., & Arif, H. (2023). An Extreme Gradient Boosting Approach for Classification and Sentiment Analysis. *The Asian Journal of Technology Management (AJTM)*, *16*(3), 211–225. https://doi.org/10.12695/ajtm.2023.16.3.5

Kipkosgei, D., & Mackenzie, S. (2026). *Performance Evaluation of Hybrid SVM- RF and XGBoost-RF Architectures for Classifying Gender-Based Violence Tweets on X*. *28*(5), 61–72.

Kirana, A. S., Roeswidiah, R., & Pudoli, A. (2025). *Analisis Sentimen Pada Media Sosial Terhadap Layanan Samsat Digital Nasional*. *8*, 53–63.

Mahmudah, S. A., & Yudhistira, A. (2025). Analisis Sentimen Terhadap Cyberbullying pada Platform Media Sosial X Menggunakan Algoritma Naive Bayes. *Jurnal Pendidikan Dan Teknologi Indonesia*, *5*(1), 189–200. https://doi.org/10.52436/1.jpti.628

Munna, A., & Zuliarso, E. (2024). Interpretation of Stacking Ensemble model for sentiment analysis of online loan application reviews using LIME. *Aiti*, *21*(2), 183–196.

Putu, I., Purnama Widiarta, A., Dwiyansaputra, R., & Aranta, A. (2023). ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN PENERAPAN PPKM DI MEDIA SOSIAL TWITTER DENGAN MENGGUNAKAN METODE XGBOOST (Analysis Of Community Sentiment On The Policy Of Implementation Of PPKM On Twitter Social Media Using Xgboost Method). *Jurnal Teknologi Informasi, Komputer Dan Aplikasinya (JTIKA)*, *5*(2), 154–163. http://jtika.if.unram.ac.id/index.php/JTIKA/