Journal of Information Technology and Computer Science (INTECOMS)
Volume 9 Nomor 3, Tahun 2026
e-ISSN : 2614-1574
p-ISSN : 2621-3249

KLASIFIKASI SENTIMEN YOUTUBE TERKAIT DEMONSTRASI AGUSTUS 2025 MENGGUNAKAN METODE TF-IDF DAN ALGORITMA XGBOOST

YOUTUBE SENTIMENT CLASSIFICATION REGARDING THE AUGUST 2025 DEMONSTRATIONS USING THE TF-IDF METHOD AND THE XGBOOST ALGORITHM

Muhammad Basyrah Diyanulhaq[1], Saeffurohman[2]
Program Studi Teknik Informatika, Universitas Stikubank Semarang[1,2]
Basyrah05@gmail.com[1]

ABSTRACT

The rapid growth of social media has produced large volumes of unstructured Indonesian-language opinion text that is difficult to analyze manually, particularly on platforms such as YouTube. This study implements a sentiment analysis model for Indonesian YouTube comments using Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction and Extreme Gradient Boosting (XGBoost) as the classifier. A total of 2,000 comments were collected via the YouTube Data API v3, preprocessed through case folding, cleaning, tokenizing, stopword removal, and Sastrawi-based stemming, then labeled into positive, neutral, and negative classes using a crowdsourcing majority-vote scheme. RandomOverSampler was applied to the training data to address class imbalance, and XGBoost hyperparameters were optimized using GridSearchCV. The best configuration, evaluated on an 80:20 train-test split, achieved an overall accuracy of 82.50%, with the negative class reaching an F1-score of 0.90, while the neutral and positive classes obtained F1-scores of 0.40 and 0.51, respectively. These results indicate that the TF-IDF and XGBoost combination performs strongly on the dominant negative class but remains limited in distinguishing minority sentiment classes due to the bag-of-words nature of TF-IDF.

Keywords: *sentiment analysis, TF-IDF, XGBoost, YouTube comments, Indonesian text*

ABSTRAK

Pertumbuhan pesat media sosial menghasilkan volume besar teks opini berbahasa Indonesia yang tidak terstruktur dan sulit dianalisis secara manual, khususnya pada platform YouTube. Penelitian ini mengimplementasikan model analisis sentimen pada komentar YouTube berbahasa Indonesia menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) sebagai teknik ekstraksi fitur dan Extreme Gradient Boosting (XGBoost) sebagai algoritma klasifikasi. Sebanyak 2.000 komentar dikumpulkan melalui YouTube Data API v3, diproses melalui tahap case folding, cleaning, tokenizing, stopword removal, dan stemming menggunakan library Sastrawi, kemudian diberi label positif, netral, dan negatif menggunakan metode crowdsourcing dengan majority vote. RandomOverSampler diterapkan pada data training untuk mengatasi ketidakseimbangan kelas, dan parameter XGBoost dioptimalkan menggunakan GridSearchCV. Konfigurasi terbaik yang diuji pada pembagian data 80:20 menghasilkan akurasi keseluruhan sebesar 82,50%, dengan kelas negatif mencapai F1-score 0,90, sedangkan kelas netral dan positif memperoleh F1-score masing-masing 0,40 dan 0,51. Hasil ini menunjukkan bahwa kombinasi TF-IDF dan XGBoost unggul pada kelas negatif yang dominan, namun masih terbatas dalam membedakan kelas sentimen minoritas akibat karakteristik TF-IDF yang bersifat bag-of-words.

Kata Kunci: analisis sentimen, TF-IDF, XGBoost, komentar YouTube, teks bahasa Indonesia

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi yang semakin pesat telah mengubah cara masyarakat dalam mengakses, berbagi, dan mengekspresikan pendapat di era digital. Platform seperti Twitter, Facebook, Instagram, YouTube, dan Reddit telah menjadi saluran komunikasi digital yang paling umum digunakan untuk menyampaikan opini publik (Setiawan & Fathonah, 2025). Media sosial sebagai platform microblogging memungkinkan pengguna memposting komentar dalam bahasa gaul dengan simbol, idiom, kata yang salah eja, dan kalimat sarkastik, sehingga menciptakan arus data teks yang besar dan beragam (Zhafira et al., 2021).

Salah satu pendekatan untuk menafsirkan data opini publik di media

sosial adalah analisis sentimen, yaitu metode untuk mengidentifikasi dan mengkategorikan opini atau perasaan seseorang terhadap suatu topik berdasarkan data tekstual (Setiawan & Fathonah, 2025). Platform YouTube memiliki potensi besar sebagai sumber data analisis sentimen; pada tahun 2022 YouTube menempati posisi kedua website paling banyak dikunjungi masyarakat Indonesia dengan total 241 miliar pengunjung (Prabowo et al., 2023). Kolom komentar pada video YouTube menjadi representasi nyata opini publik yang dapat diekstraksi dan dianalisis.

Meskipun potensinya besar, analisis sentimen pada teks komentar media sosial menghadapi tantangan signifikan karena sifatnya yang informal, tidak terstruktur, dan mengandung bahasa gaul, singkatan, emoji, code-mixing, ironi, dan sarkasme. Bahasa Indonesia juga memiliki karakteristik linguistik yang berbeda dari bahasa Inggris, sehingga tidak semua metode yang efektif untuk bahasa Inggris dapat langsung diterapkan (Savana et al., 2025).

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan salah satu metode representasi fitur teks yang paling banyak digunakan, yang mengubah teks menjadi representasi numerik berdasarkan frekuensi kemunculan kata dalam dokumen dan dalam keseluruhan koleksi dokumen (Prabowo et al., 2023). Di sisi algoritma klasifikasi, Extreme Gradient Boosting (XGBoost) terbukti menjadi salah satu algoritma machine learning dengan performa terbaik dalam berbagai tugas klasifikasi teks. Berdasarkan Systematic Literature Review terhadap 49 jurnal ilmiah, XGBoost termasuk dalam lima algoritma dengan akurasi tertinggi, mencapai 93% pada berbagai dataset media sosial (Setiawan & Fathonah, 2025).

Penelitian terdahulu menunjukkan efektivitas kombinasi TF-IDF dan XGBoost dalam analisis sentimen berbahasa Indonesia. Prabowo et al. (2023) menerapkan TF-IDF dan XGBoost untuk Aspect-Based Sentiment Analysis pada komentar YouTube review iPhone 14 Pro, menghasilkan akurasi 89%-97%. Sinaga dan Agustian (2022) membuktikan keunggulan XGBoost dibandingkan Decision Tree, Naive Bayes, SVM, dan LSTM dalam klasifikasi sentimen tweet vaksin COVID-19 berbahasa Indonesia, dengan akurasi terbaik 66% dan F1-score 57% pada dataset tidak seimbang.

Berdasarkan pemaparan tersebut, penelitian ini bertujuan mengimplementasikan dan mengevaluasi metode analisis sentimen menggunakan kombinasi TF-IDF sebagai teknik ekstraksi fitur dan algoritma XGBoost sebagai classifier pada data komentar media sosial berbahasa Indonesia, dengan harapan menghasilkan model yang efektif untuk memahami persepsi publik pada berbagai topik di media sosial.

METODE

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimen, karena menggunakan data numerik dan pengujian model secara sistematis (Savana et al., 2025). Data komentar media sosial berbahasa Indonesia yang telah dilabeli sentimen diproses menggunakan metode TF-IDF untuk representasi fitur, kemudian diklasifikasikan menggunakan algoritma XGBoost. Penelitian ini terdiri dari tujuh tahapan utama: (1) pengumpulan data, (2) preprocessing teks, (3) pelabelan data, (4) ekstraksi fitur TF-IDF, (5) pembagian dataset, (6) klasifikasi menggunakan XGBoost, dan (7) evaluasi model (Prabowo et al., 2023; Sinaga & Agustian, 2022).

Pengumpulan Data

Data bersumber dari kolom komentar pada video-video YouTube yang diambil menggunakan YouTube Data API v3 dengan dukungan skrip Python. Data komentar disimpan dalam format CSV dan difilter berdasarkan kriteria: (1) ditulis dalam bahasa Indonesia, (2) merupakan opini yang mengandung sentimen tertentu, (3) tidak bersifat spam atau duplikat, dan (4) memiliki panjang minimal lima kata

(Prabowo et al., 2023; Savana et al., 2025). Total komentar yang terkumpul dan lolos seleksi sebanyak 2.000 komentar yang kemudian dibagi menjadi data training dan testing dengan perbandingan 80%:20% (Sinaga & Agustian, 2022).

Preprocessing Teks

Terdapat lima tahap preprocessing yang dilakukan secara berurutan menggunakan library Sastrawi, NLTK, pandas, dan numpy, yaitu case folding (mengubah seluruh karakter menjadi huruf kecil), cleaning (membersihkan angka, simbol, tanda baca, URL, mention, hashtag, dan emoji), tokenizing (memecah kalimat menjadi token kata), stopword removal (menghapus kata yang tidak relevan), dan stemming menggunakan library Sastrawi untuk mengembalikan kata berimbuhan ke bentuk dasarnya (Prabowo et al., 2023; Sinaga & Agustian, 2022).

Tabel 1. Ringkasan Hasil Preprocessing

Proses	Input	Output
Case Folding	Dynamic Island inovasinya Keren	dynamic island inovasinya keren
Cleaning	produk bagus!!! @toko #review	produk bagus
Tokenizing	ruu segera disahkan	['ruu','segera','disahkan']
Stopword Removal	produk ini sangat bagus	produk bagus
Stemming	disahkan, keadilan	sah, adil

Pelabelan Data

Data yang telah bersih diberi label sentimen ke dalam tiga kategori: positif (pujian, dukungan, kesenangan), negatif (keluhan, kritik, kekecewaan), dan netral (data yang tidak memenuhi kriteria positif maupun negatif) (Sinaga & Agustian, 2022). Pelabelan dilakukan menggunakan metode crowdsourcing dengan majority vote dari tiga anotator; data tanpa mayoritas dihapus dari dataset.

Ekstraksi Fitur TF-IDF

Data yang telah diproses dan dilabeli dikonversi menjadi vektor numerik menggunakan TfidfVectorizer dari scikit-learn. Bobot TF-IDF dihitung dengan mengalikan nilai Term Frequency (TF) dan Inverse Document Frequency (IDF), di mana kata yang sering muncul pada satu dokumen tetapi jarang muncul di seluruh corpus memperoleh bobot yang lebih tinggi (Prabowo et al., 2023; Sinaga & Agustian, 2022).

Pembagian Dataset dan Penanganan Imbalanced Data

Dataset sebanyak 2.000 data dibagi menggunakan metode hold-out dengan proporsi 80% data training (1.600 data) dan 20% data testing (400 data), serta divalidasi menggunakan K-Fold Cross Validation (Savana et al., 2025). Mengingat distribusi label tidak seimbang, penanganan dilakukan menggunakan RandomOverSampler yang hanya diterapkan pada data training agar evaluasi model tetap mencerminkan distribusi kelas yang sebenarnya (Sinaga & Agustian, 2022).

Implementasi XGBoost

Model XGBoost diimplementasikan menggunakan library xgboost dengan antarmuka scikit-learn. Parameter tuning dilakukan menggunakan teknik GridSearchCV pada ruang parameter subsample, colsample_bytree, max_depth, min_child_weight, learning_rate, dan n_estimators untuk memperoleh kombinasi parameter dengan performa terbaik (Sinaga & Agustian, 2022).

Tabel 2. Skenario Pengujian Model

Skenario	Training	Testing	Keterangan
1	90%	10%	Proporsi training besar
2	80%	20%	Baseline
3	70%	30%	Proporsi testing besar

Evaluasi Model

Evaluasi performa model dilakukan menggunakan confusion matrix yang menghasilkan empat elemen, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) (Prabowo et al., 2023). Keempat elemen tersebut diolah untuk memperoleh nilai accuracy, precision, recall, dan F1-score, yang digunakan secara bersamaan untuk menilai performa model klasifikasi sentimen secara komprehensif (Setiawan & Fathonah, 2025).

HASIL DAN PEMBAHASAN

Proses scraping menggunakan YouTube Data API v3 berhasil mengumpulkan 2.000 komentar berbahasa Indonesia yang berkaitan dengan topik legislasi dan kebijakan publik. Setelah melalui proses filtering dan cleaning sesuai kriteria yang ditetapkan, seluruh 2.000 komentar tersebut layak digunakan sebagai dataset penelitian.

Hasil Pelabelan Data

Hasil pelabelan terhadap 2.000 data menghasilkan distribusi tiga kategori sentimen sebagaimana disajikan pada Tabel 3.

Tabel 3. Distribusi Label Sentimen Hasil Pelabelan

Label Sentimen	Jumlah Data	Persentase
Negatif	1.537	76,85%
Netral	254	12,70%
Positif	209	10,45%
Total	2.000	100%

Distribusi sentimen menunjukkan ketidakseimbangan kelas yang signifikan, dengan sentimen Negatif mendominasi dataset (76,85%). Dominasi ini konsisten dengan karakteristik komentar pada video bertopik legislasi dan kebijakan publik, di mana masyarakat cenderung lebih banyak mengekspresikan kritik dibandingkan dukungan (Sinaga & Agustian, 2022). Untuk mengatasi ketidakseimbangan ini, metode RandomOverSampler diterapkan pada data training sebelum pelatihan model dimulai.

Hasil Parameter Tuning XGBoost

Proses GridSearchCV menghasilkan konfigurasi parameter optimal sebagaimana disajikan pada Tabel 4.

Tabel 4. Parameter XGBoost Optimal Hasil GridSearchCV

Parameter	Nilai Optimal
n_estimators	200
max_depth	6
learning_rate	0,1
subsample	0,75
colsample_bytree	0,75
min_child_weight	3

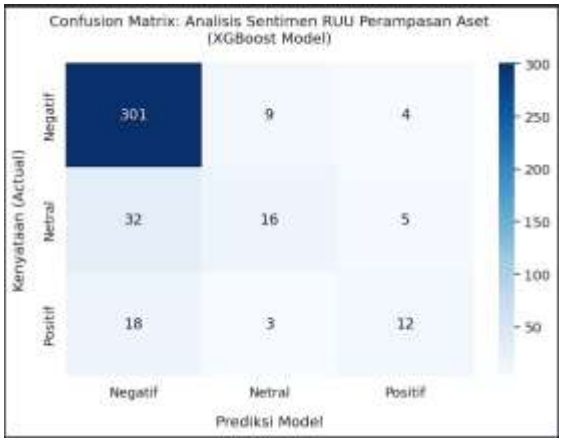
Hasil Evaluasi Model

Pengujian model pada 400 data testing menghasilkan confusion matrix sebagaimana disajikan pada Tabel 5.

Tabel 5. Confusion Matrix Model XGBoost

Aktual \ Prediksi	Negatif	Netral	Positif
Negatif	301	11	2
Netral	37	16	0
Positif	20	1	12

Model berhasil mengklasifikasikan kelas Negatif dengan sangat baik, yaitu 301 dari 314 data uji berhasil diprediksi dengan benar. Untuk kelas Netral, model hanya benar pada 16 dari 53 data, sedangkan 37 data Netral salah diprediksi sebagai Negatif. Kelas Positif menunjukkan 12 prediksi benar dari 33 data. Pola kesalahan ini mengindikasikan adanya bias model terhadap kelas mayoritas, meskipun oversampling telah diterapkan pada data training.



Gambar 1. Visualisasi Confusion Matrix Model XGBoost

Tabel 6. Classification Report Model XGBoost

Kelas	Precision	Recall	F1-Score	Support
Negatif	0,84	0,96	0,90	314
Netral	0,57	0,30	0,40	53
Positif	0,86	0,36	0,51	33
Accuracy	-	-	0,82	400
Macro Avg	0,76	0,54	0,60	400
Weighted Avg	0,81	0,82	0,80	400

Model XGBoost menghasilkan akurasi keseluruhan sebesar 82,50% pada data testing. Kelas Negatif memperoleh performa tertinggi dengan F1-score 0,90, sejalan dengan dominasi kelas tersebut dalam dataset yang memberikan model lebih banyak contoh pelatihan. Performa kelas Netral (F1-score 0,40) dan Positif (F1-score 0,51) relatif lebih rendah, menunjukkan bahwa model masih mengalami kesulitan mengidentifikasi sentimen netral dan positif secara tepat meskipun oversampling telah dilakukan. Fenomena ini umum ditemukan pada analisis sentimen teks bahasa Indonesia yang memiliki ambiguitas semantik tinggi, terutama pada kalimat yang mengandung sarkasme atau ironi (Savana et al., 2025).

Perbandingan Skenario Pengujian

Tabel 7. Perbandingan Akurasi pada Setiap Skenario Pengujian

Skenario	Training	Testing	Akurasi
1	90%	10%	81,00%
2 (Terpilih)	80%	20%	82,50%
3	70%	30%	80,25%

Skenario 2 dengan pembagian 80:20 menghasilkan akurasi tertinggi sebesar 82,50%, lebih tinggi dibandingkan skenario 1 (90:10, akurasi 81,00%) dan skenario 3 (70:30, akurasi 80,25%). Hasil ini konsisten dengan penelitian Sinaga dan Agustian (2022) yang merekomendasikan proporsi 80:20 sebagai konfigurasi optimal untuk dataset analisis sentimen berbahasa Indonesia.

Pembahasan

Penelitian ini berhasil membangun model analisis sentimen berbasis TF-IDF dan XGBoost dengan akurasi 82,50% pada dataset komentar YouTube berbahasa Indonesia. Hasil ini sebanding dengan penelitian sebelumnya; Prabowo et al. (2023) memperoleh akurasi 83,50% pada klasifikasi sentimen iPhone 14 Pro menggunakan XGBoost, sementara Sinaga dan Agustian (2022) melaporkan akurasi 82,70% untuk klasifikasi sentimen vaksin COVID-19.

Kinerja model yang baik pada kelas Negatif (F1-score 0,90) mengonfirmasi bahwa kombinasi TF-IDF dan XGBoost efektif untuk mendeteksi sentimen negatif pada teks berbahasa Indonesia, terutama ketika jumlah data latih pada kelas tersebut mencukupi. Namun keterbatasan model terlihat jelas pada kelas minoritas, yang disebabkan oleh dua faktor utama: (1) ketidakseimbangan distribusi kelas yang signifikan menyulitkan model mempelajari pola yang representatif meskipun oversampling telah diterapkan, dan (2) representasi fitur TF-IDF yang bersifat bag-of-words tidak mampu menangkap informasi kontekstual, urutan kata, atau nuansa semantik yang penting untuk membedakan sentimen pada teks bahasa Indonesia yang kaya akan ekspresi informal (Savana et al., 2025).

Nilai macro average F1-score sebesar 0,60 mencerminkan performa rata-rata yang moderat jika setiap kelas diperlakukan setara, menunjukkan masih terdapat ruang signifikan untuk peningkatan pada kelas Netral dan Positif. Penggunaan model berbasis deep learning seperti IndoBERT atau FastText yang mampu menangkap representasi kontekstual berpotensi meningkatkan performa secara signifikan, khususnya untuk kelas minoritas (Savana et al., 2025; Setiawan & Fathonah, 2025).

SIMPULAN

Penelitian ini berhasil membangun sistem analisis sentimen terhadap komentar berbahasa Indonesia pada platform YouTube menggunakan kombinasi TF-IDF sebagai teknik ekstraksi fitur dan algoritma XGBoost sebagai classifier. Tahapan preprocessing yang diterapkan secara berurutan terbukti efektif mereduksi noise dan menyederhanakan representasi teks. Metode TF-IDF berhasil mengonversi teks komentar menjadi representasi vektor numerik yang informatif bagi proses klasifikasi.

Algoritma XGBoost dengan konfigurasi parameter optimal hasil GridSearchCV (n_estimators=200, max_depth=6, learning_rate=0,1, subsample=0,75, colsample_bytree=0,75, min_child_weight=3) mampu membangun model klasifikasi sentimen tiga kelas dengan akurasi keseluruhan 82,50% pada skenario pembagian data 80:20. Kelas Negatif memperoleh performa terbaik (F1-score 0,90), sedangkan kelas Netral dan Positif memperoleh F1-score masing-masing 0,40 dan 0,51, yang mengindikasikan keterbatasan model dalam mengklasifikasikan kelas minoritas secara akurat.

Penelitian selanjutnya disarankan mengeksplorasi model berbasis konteks seperti IndoBERT atau RoBERTa, teknik penanganan ketidakseimbangan kelas yang lebih canggih seperti SMOTE, peningkatan volume dan keberagaman dataset, serta pendekatan ensemble yang menggabungkan beberapa algoritma klasifikasi untuk meningkatkan performa pada kelas sentimen minoritas.

DAFTAR PUSTAKA

Adriana, N. M. T. O., Suarjaya, I. M. A. D., & Githa, D. P. (2023). Analisis sentimen publik terhadap aksi demonstrasi di Indonesia menggunakan support vector machine dan random forest. Decode: Jurnal Pendidikan Teknologi Informasi, 3(2), 257–267. https://doi.org/10.51454/decode.v3i2.187

Al Rasyid, R., Handayani, D., & Ningsih, U. (2024). Penerapan algoritma TF-IDF dan cosine similarity untuk query pencarian pada dataset destinasi wisata. Jurnal Teknologi Informasi dan Komunikasi, 8(1). https://doi.org/10.35870/jti

Alfaruqy, M. Z., & Sari, I. A. (2025). Peringatan Darurat: Political engagement of Indonesian university students amid political uncertainty. Journal of Social and Industrial Psychology, 14(2).

Fatah, Z., & Ningsih, R. A. (2025). Analisis sentimen komentar YouTube terhadap Tragedi Demo 25 Agustus menggunakan pendekatan lexicon-based. JAMASTIKA, 4(2), 217–224.

Prabowo, T., Rahmat, R., & Wahanani, H. E. (2023). Aspect-based sentiment analysis pada komentar YouTube review iPhone 14 Pro menggunakan TF-IDF dan XGBoost. Jurnal Ilmu Komputer dan Informatika.

Savana, M., Arifianto, A., & Muharom, F. (2025). Analisis sentimen komentar video pidato politik di YouTube menggunakan FastText. Jurnal Teknologi dan Sistem Informasi.

Setiawan, R., & Fathonah, N. (2025). Systematic literature review: Analisis sentimen persepsi publik terhadap kecerdasan buatan di media sosial menggunakan machine learning dan

deep learning. Jurnal Informatika dan Sistem Informasi.

Sinaga, H. H., & Agustian, S. (2022). Perbandingan metode Decision Tree dan XGBoost untuk klasifikasi sentimen vaksin Covid-19 di Twitter. Jurnal Nasional Teknologi dan Sistem Informasi, 8(3), 107–114.

Yutika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis sentimen berbasis aspek pada review Female Daily menggunakan TF-IDF dan Naïve Bayes. Jurnal Media Informatika Budidarma, 5(2), 422. https://doi.org/10.30865/mib.v5i2.2845

Zhafira, D. F., Rahayudi, B., & Indriati. (2021). Analisis sentimen kebijakan Kampus Merdeka menggunakan Naive Bayes dan pembobotan TF-IDF berdasarkan komentar pada YouTube. Jurnal Sistem Informasi, Teknologi Informasi, dan Edukasi Sistem Informasi (JUST-SI), 2(1), 55–63.