



Komparasi FastText dan TF-IDF Berbasis Random Forest pada Analisis Sentimen IKN di Youtube

Fadhil Irsyad Ramadhani, Taghfirul Azhima Yoga*, Naufal Azmi Verdhika

Fakultas Sains dan Teknologi, Program Studi Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia

Email: ¹2111102441044@umkt.ac.id, ^{2,*}tay758@umkt.ac.id, ³nav651@umkt.ac.id

Email Penulis Korespondensi: tay758@umkt.ac.id

Abstrak—Pengembangan Ibu Kota Nusantara (IKN) Indonesia merepresentasikan kebijakan nasional signifikan yang telah memicu beragam respons publik, terutama di berbagai platform media sosial seperti YouTube. Penelitian ini bertujuan untuk menganalisis sentimen publik mengenai proyek IKN dan membandingkan performa dua metode ekstraksi fitur teks yaitu FastText dan Term Frequency Inverse Document Frequency (TF-IDF), dengan menggunakan algoritma Random Forest. Tujuan utamanya adalah untuk mengidentifikasi metode mana yang lebih baik dalam menangkap nuansa opini publik berbahasa Indonesia. Dataset penelitian ini mencakup 4.093 komentar YouTube terkait IKN yang diperoleh menggunakan YouTube Data API v3 pada Agustus 2025. Data dibagi menjadi dua kelas yaitu positif dan negatif, sementara data netral dihapus untuk meminimalkan bias pada model. Pelabelan data dilakukan secara manual yang divalidasi oleh pakar bahasa, tahap pre-processing dilakukan seperti data cleaning, case folding, normalisasi, tokenizing, stopword removal, dan stemming. Penetapan dimensi vektor 200 pada FastText dan batasan 5.000 fitur pada TF-IDF didasarkan pada temuan penelitian analisis sentimen terdahulu yang membuktikan bahwa konfigurasi tersebut memberikan performa klasifikasi yang stabil dibandingkan pengujian pada parameter lain, karena secara statistik lebih efektif dalam menyaring fitur tidak relevan tanpa menghilangkan kedalaman informasi semantik. Performa model dievaluasi menggunakan metode 10-Fold Cross Validation dan Confusion Matrix berdasarkan metrik accuracy, precision, recall, dan F1-score. Hasil menunjukkan bahwa metode FastText mencapai akurasi sebesar 83,67%, precision 84,01%, recall 83,72%, dan F1-score 80,83%, sementara TF-IDF menghasilkan akurasi 80,53%. Temuan ini menyimpulkan bahwa FastText lebih efektif dalam merepresentasikan konteks dan makna semantik dari komentar YouTube berbahasa Indonesia terkait IKN. Selain itu, metode ini memberikan keseimbangan dalam pengenalan pola dan ketepatan hasil klasifikasi sentimen. Penelitian ini berkontribusi dalam membantu para pemangku kepentingan dan peneliti untuk memahami opini publik terhadap IKN dengan lebih tepat.

Kata Kunci: Analisis Sentimen; IKN; Youtube; FastText; TF-IDF; Random Forest

Abstract—The development of Indonesia's New Capital City (IKN) represents a significant national policy that has triggered diverse public responses, particularly across social media platforms like YouTube. This study aims to analyze public sentiment regarding the IKN project and compare the performance of two text feature extraction methods, FastText and Term Frequency-Inverse Document Frequency (TF-IDF), using the Random Forest algorithm. The primary objective is to identify which method is more effective in capturing the nuances of Indonesian-language public opinion. The dataset for this research includes 4,093 YouTube comments related to IKN, obtained using the YouTube Data API v3 in August 2025. The data were categorized into two classes, positive and negative, while neutral data were removed to minimize model bias. Data labeling was conducted manually and validated by a linguistic expert, followed by pre-processing stages such as data cleaning, case folding, normalization, tokenizing, stopword removal, and stemming. The setting of a 200-vector dimension for FastText and a 5,000-feature limit for TF-IDF was based on findings from previous sentiment analysis research, proving that such configurations provide stable classification performance compared to other parameters, as they are statistically more effective in filtering irrelevant features without losing deep semantic information. Model performance was evaluated using the 10-Fold Cross-Validation method and Confusion Matrix based on accuracy, precision, recall, and F1-score metrics. Results indicate that the FastText method achieved an accuracy of 83.67%, precision of 84.01%, recall of 83.72%, and an F1-score of 80.83%, while TF-IDF yielded an accuracy of 80.53%. These findings conclude that FastText is more effective in representing the context and semantic meaning of Indonesian YouTube comments related to IKN. Furthermore, this method provides a balance in pattern recognition and the precision of sentiment classification results. This research contributes to assisting stakeholders and researchers in more accurately understanding public opinion toward IKN.

Keywords: Sentiment Analysis; IKN; Youtube; FastText; TF-IDF; Random Forest

1. PENDAHULUAN

Rencana pembangunan Ibu Kota Nusantara (IKN) telah memicu kekhawatiran dan kritik luas dari masyarakat, terutama terkait dampak lingkungan, sosial, serta keberlanjutan budaya adat (Fitriyah Syarifatul, 2025; Rihastuti & Rosyidi, 2025). Fenomena tersebut memicu lonjakan opini digital di platform YouTube, di mana ekspresi publik sering kali memuat nuansa emosional kompleks seperti ironi dan sarkasme (Prastyo et al., 2024). Komentar tersebut seringkali menggunakan gaya bahasa kasual yang mencampurkan berbagai istilah. Namun, dalam penerapan analisis sentimen pada teks code-mixed (campur kode) menghadirkan tantangan tersendiri. Permasalahan tersebut disebabkan oleh kompleksitas dalam memproses dan mengekstrak sentimen secara tepat dari teks yang mengandung kosakata dari lebih dari satu bahasa (Nurannisa et al., 2026).

Analisis sentimen merupakan teknik komputasi untuk mengidentifikasi serta mengklasifikasikan polaritas opini publik yang diekstraksi dari data tekstual. Pendekatan ini memanfaatkan metode Natural Language Processing (NLP) untuk mentransformasi teks mentah menjadi representasi data terstruktur yang dapat diolah secara otomatis oleh algoritma klasifikasi (Arif & Kustiyo, 2025). Dalam implementasinya berbagai algoritma machine learning seperti Naive Bayes, Support Vector Machine (SVM), dan Random Forest banyak diterapkan untuk memetakan kecenderungan opini dalam berbagai domain (Purnomo et al., 2025).



Sejumlah penelitian terdahulu telah mengeksplorasi analisis sentimen IKN pada platform Twitter dengan akurasi yang cukup stabil berkisar antara 75,69% hingga 79% menggunakan algoritma Naive Bayes, KNN, dan K-Means (Iman & Ujianto, 2024; Melati & Reza, 2024; Priyanto et al., 2025). Namun, karakteristik data pada platform YouTube menghadirkan tantangan yang berbeda. Penelitian oleh Huwaida et al., (2024) pada komentar YouTube mengenai IKN hanya mencapai akurasi tertinggi sebesar 76%, di mana pengabaian terhadap istilah asing dan bahasa daerah menjadi kendala utama dalam akurasi. Sejalan dengan pernyataan tersebut, Laia & Barus, (2025) mengonfirmasi opini publik di YouTube terkait IKN sangat bervariasi dan kontras, mulai dari setuju hingga penolakan keras, yang menuntut ekstraksi fitur lebih mendalam. Kesenjangan utama dalam penelitian terdahulu terlihat pada rendahnya performa model konvensional seperti Naive Bayes yang hanya menghasilkan akurasi sebesar 60% saat diterapkan pada data YouTube IKN (Mola et al., 2024). Rendahnya angka tersebut menunjukkan bahwa model klasifikasi konvensional sering kali gagal menangani struktur bahasa media sosial yang dinamis. Oleh karena itu, diperlukan eksplorasi integrasi algoritma dengan metode ekstraksi fitur yang lebih stabil untuk menghasilkan kinerja klasifikasi yang lebih baik.

Random Forest merupakan sebuah algoritma ensemble berbasis decision trees dikenal karena stabilitasnya yang tinggi kapasitasnya dalam menangani data berdimensi tinggi, serta kemampuannya dalam mengurangi risiko overfitting (Mursyidah et al., 2024). Keunggulan Random Forest dapat dijelaskan oleh kemampuannya dalam menangkap hubungan non-linear dan interaksi kompleks antar fitur linguistik melalui mekanisme ensemble berbasis banyak pohon keputusan (Raihan et al., 2026). Pada penelitian sebelumnya Ayunda et al., (2022) menunjukkan bahwa Random Forest unggul dengan akurasi 78% melalui pemanfaatan Multiple Decision Tree. Keunggulan tersebut didukung oleh kemampuan model dalam mengekstraksi atribut emosi dari opini masyarakat yang kerap menggunakan bahasa slang, kata tidak baku, hingga ekspresi kompleks seperti ironi dan sarkasme, sehingga memerlukan ekstraksi fitur yang mendalam untuk memperkuat representasi makna teksnya.

Metode ekstraksi fitur Term Frequency–Inverse Document Frequency (TF-IDF) banyak digunakan karena kesederhanaan dan efisiensinya dalam merepresentasikan bobot kata (Setiana et al., 2023). Sementara itu, FastText mampu merepresentasikan kata sebagai vektor berbasis subword, sehingga efektif dalam menangkap konteks dan makna semantik kata, terutama pada teks bahasa Indonesia yang bersifat informal (Ermawan & Cahyono, 2025). Kedua metode ekstraksi fitur tersebut memiliki kekuatan masing-masing, TF-IDF unggul dalam efisiensi pengolahan dengan memberikan bobot lebih tinggi pada kata-kata spesifik yang relevan bagi dokumen tertentu. Di sisi lain, FastText memberikan pemahaman konteks yang lebih mendalam melalui representasi vector subword yang akurat, sehingga sangat efektif dalam menangani variasi bahasa dan kata-kata yang tidak ditemui sebelumnya (out-of-vocabulary) pada teks media sosial (Atika et al., 2025)

Meskipun kedua metode ekstraksi fitur tersebut telah banyak digunakan secara terpisah, studi komparatif antara FastText dan TF-IDF menggunakan algoritma Random Forest khususnya dalam konteks analisis sentimen pembangunan IKN di YouTube masih sangat terbatas. Penelitian sebelumnya umumnya hanya berfokus pada satu metode ekstraksi fitur atau menggunakan algoritma klasifikasi konvensional dengan performa yang relatif rendah pada data media sosial yang bersifat informal dan dinamis. Oleh karena itu, penelitian ini dilakukan untuk membandingkan performa ekstraksi fitur TF-IDF dan FastText menggunakan algoritma Random Forest pada analisis sentimen komentar YouTube terkait pembangunan IKN.

Penelitian ini memberikan kontribusi dalam tiga aspek utama. Pertama, menyediakan evaluasi komparatif antara TF-IDF dan FastText dalam menangani karakteristik teks media sosial yang mengandung bahasa informal, slang, dan campur kode, serta mengevaluasi kemampuan FastText dalam merepresentasikan istilah out-of-vocabulary melalui pendekatan subword embedding, serta istilah out-of-vocabulary. Kedua, mengimplementasikan algoritma Random Forest sebagai model klasifikasi untuk meningkatkan stabilitas dan kemampuan generalisasi pada data komentar YouTube yang kompleks. Ketiga, penelitian ini memberikan rekomendasi metode ekstraksi fitur yang paling efektif dalam meningkatkan performa analisis sentimen pembangunan IKN di media sosial, sehingga dapat menjadi referensi bagi penelitian serupa pada isu sosial berbasis opini publik digital.

2. METODOLOGI PENELITIAN

2.1 Kerangka Dasar Penelitian

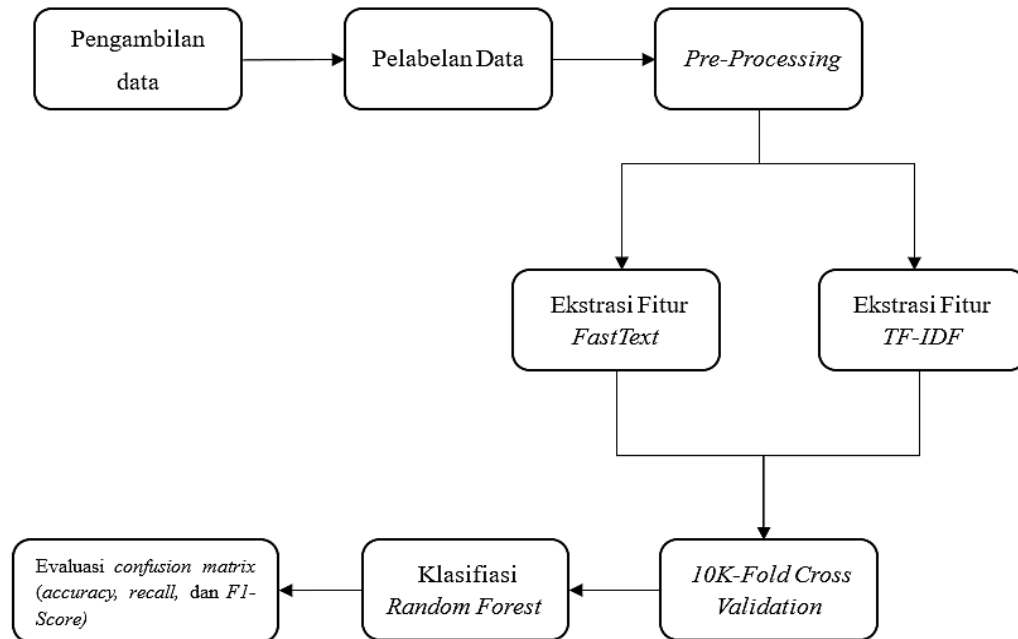
Penelitian ini merupakan studi kuantitatif eksperimen komparatif yang bertujuan mengklasifikasikan sentimen komentar YouTube terkait pembangunan IKN menggunakan teknik Natural Language Processing (NLP). Data dikumpulkan secara daring melalui YouTube Data API v3 mencakup pada bulan Agustus 2025 dan diproses dalam lingkungan digital menggunakan bahasa Python. Pendekatan ini dipilih untuk menjawab tantangan linguistik pada komentar YouTube yang cenderung informal dan variatif. Berdasarkan landasan penelitian terdahulu yang menggunakan TF-IDF (Mola et al., 2024; Saputri & Alita, 2024) serta penggunaan FastText (Ermawan & Cahyono, 2025). Penelitian ini melakukan evaluasi komparatif antara fitur berbasis statistik (TF-IDF) dan berbasis semantik (FastText). Algoritma Random Forest diterapkan sebagai model klasifikasi utama karena stabilitasnya dalam menangani data tidak terstruktur (Ayunda et al., 2022).

Dalam struktur metodologinya, variabel bebas yang diuji meliputi metode ekstraksi fitur dan tahapan pre-processing, sementara variabel terikatnya adalah kinerja model yang diukur melalui metrik akurasi, precision, recall, dan F1-score melalui skema 10-Fold Cross Validation. Analisis data dilakukan secara komparatif untuk

mengidentifikasi teknik ekstraksi fitur yang paling efektif dalam meningkatkan akurasi. Kerangka pemikiran ini didasarkan pada asumsi bahwa efektivitas klasifikasi sentimen sangat bergantung pada kemampuan representasi fitur dalam menangkap konteks semantik, dengan tetap mengutamakan validasi label oleh pakar bahasa guna menjaga objektivitas hasil penelitian.

2.2 Tahapan Penelitian

Tahapan penelitian terdiri dari beberapa tahap untuk memastikan alur penyelesaian masalah penelitian yang ditampilkan pada Gambar 1 sebagai berikut



Gambar 1. Tahapan Penelitian

Gambar 1 mengilustrasikan alur penelitian yang dimulai dengan pengumpulan komentar YouTube terkait IKN, diikuti validasi label oleh pakar bahasa untuk kategori positif dan negatif. Data kemudian melalui tahap pre-processing agar siap dianalisis menggunakan ekstraksi fitur FastText dan TF-IDF. Selanjutnya, diterapkan skema 10-Fold Cross Validation untuk pembagian data yang diproses oleh algoritma Random Forest. Tahap akhir melibatkan evaluasi menggunakan confusion matrix untuk membandingkan efektivitas kedua metode ekstraksi fitur dalam mengklasifikasikan sentimen opini publik secara akurat.

2.2.1 Pengambilan Data

Penelitian ini menggunakan dataset yang diperoleh dari platform YouTube, dengan fokus khusus pada topik terkait pembangunan Ibu Kota Nusantara (IKN) Indonesia (Laia & Barus, 2025). Proses pengambilan data dilakukan menggunakan YouTube Data API v3 untuk mengambil komentar (Rahmadhani et al., 2025). Pada penelitian sebelumnya lima kanal YouTube yang berbeda dipilih untuk mengumpulkan opini yang objektif, mendukung akurasi model sebesar 86% dengan menggunakan total 3.078 data (Muhayat et al., 2023) Pengumpulan data dilakukan selama periode satu bulan, merujuk pada penelitian sebelumnya yang memperoleh total 4.774 data dalam satu bulan dan mencapai akurasi model sebesar 92% (Larasakti et al., 2023).

2.2.2 Pelabelan Data

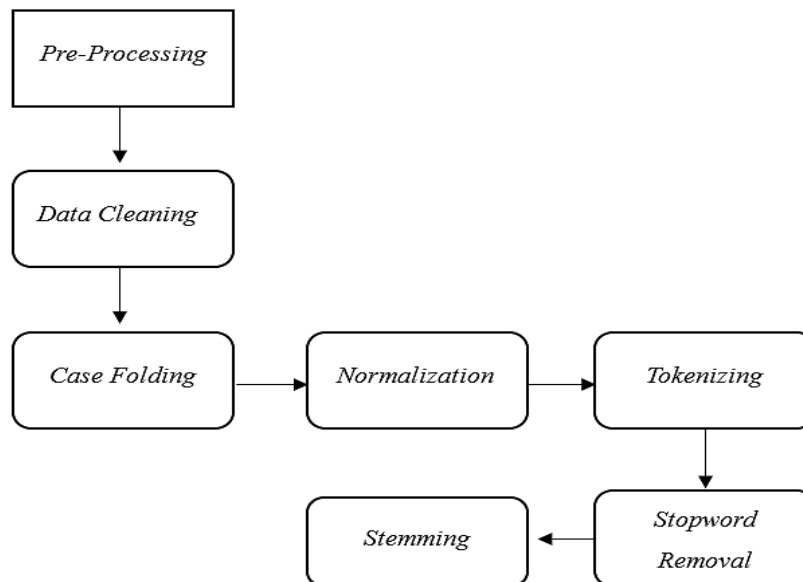
Pelabelan data yang dilakukan secara manual dan divalidasi oleh seorang pakar bahasa dapat memberikan label sentimen berdasarkan makna kontekstual yang terkandung dalam teks (Romadhoni & Holle Hayati, 2022). Validitas tersebut diuji menggunakan analisis berbasis konsensus (consensus-based assessment), yaitu setiap label sentimen disepakati secara penuh oleh peneliti dan pakar untuk meminimalkan bias subjektif. Prosedur ini dilakukan melalui metode systematic spot-checking pada sampel data yang representatif, di mana pakar bahasa melakukan audit terhadap akurasi pelabelan yang dilakukan peneliti. Dalam mengukur reliabilitas antar penilai penelitian ini menerapkan metode konsensus absolut, yakni setiap perbedaan interpretasi dianalisis secara mendalam hingga mencapai skor kesepakatan 1,00 (full agreement).

Sentimen diklasifikasikan ke dalam kelas positif (dukungan dan optimisme) dan negatif (kritik dan keluhan). Pendekatan binary classification ini dipilih karena efektivitasnya yang terbukti pada berbagai penelitian analisis sentimen terdahulu (Iman & Ujianto, 2024; Jasmine et al., 2025). Kelas netral ditiadakan karena ambiguitas bawaannya karena sering kali mengandung sentimen campuran atau pernyataan deskriptif murni yang memberikan nilai informatif

terbatas bagi analisis (Kadir & Fairuzabadi, 2025). Melalui prosedur tersebut hasil pelabelan akhir berfungsi sebagai gold standard, yang merepresentasikan hasil validasi pakar terhadap data latih model.

2.2.3 Pre-Processing

Tujuan dari pre-processing untuk menyaring data yang tidak relevan dan mengubahnya menjadi format yang lebih terstruktur untuk memastikan teks yang dapat diproses secara efisien oleh sistem dan layak untuk dianalisis (Calvin Jonathan et al., 2023). Berikut adalah alur dari preprocessing yang di tampilkan pada Gambar 2.



Gambar 2. Tahapan Pre-Processing

Gambar 2 menampilkan tahapan pre-processing teks komentar yang dilakukan secara berurutan, dimulai dari data cleaning, case folding, normalization, tokenizing, stopword removal, hingga stemming. Tahapan preprocessing data merupakan langkah krusial untuk mempersiapkan data mentah agar dapat diproses oleh algoritma pembelajaran mesin. Tahapan ini dilakukan sebelum proses ekstraksi fitur menggunakan FastText dan TF-IDF (Salman Al Markas et al., 2025; Wijaya & Hayati, 2025). Penempatan preprocessing di awal tersebut selaras dengan prinsip metodologis dalam penelitian komparatif untuk menjamin bahwa seluruh metode ekstraksi fitur menerima input yang seragam. komputer hanya mampu memproses input numerik, representasi teks ke dalam nilai numerik (ekstraksi fitur) sangat bergantung pada kualitas data bersih yang dihasilkan di tahap awal. Oleh karena itu, standarisasi preprocessing dilakukan untuk menjaga konsistensi pipeline serta memastikan bahwa perbedaan performa yang dihasilkan murni mencerminkan karakteristik masing-masing metode, bukan akibat variasi perlakuan awal pada data. Adapun penjelasan mendalam untuk setiap tahapan tersebut adalah sebagai berikut:

- Data Cleaning bertujuan untuk menghapus elemen yang tidak relevan seperti simbol, emotikon, tanda baca, angka, URL, dan karakter non-alfabet, serta untuk menghilangkan data duplikat (Prastyo et al., 2024).
- Case Folding merupakan proses mengubah semua karakter menjadi huruf kecil untuk mencegah inkonsistensi yang disebabkan oleh variasi penggunaan huruf kapital (Huwaida et al., 2024).
- Normalization adalah tahapan yang digunakan untuk mengubah variasi teks tidak baku, seperti bahasa gaul (slang), singkatan, dan salah ketik (typos), menjadi bentuk bakunya, sehingga dapat meningkatkan kualitas data dan akurasi model (Muhammad Rizki Syafapri et al., 2024).
- Tokenizing adalah proses memecah teks menjadi unit kata individual (tokens), yang berfungsi sebagai unit dasar analisis dalam pemrosesan bahasa (Wajhillah & Wibowo, 2021).
- Stopword Removal adalah proses menghapus kata-kata dari teks yang dianggap kurang relevan atau tidak signifikan (Rafi et al., 2024).
- Terakhir, Stemming bertujuan untuk mengembalikan kata ke bentuk dasar atau kata fungsinya dengan menghapus imbuhan, sehingga kata-kata yang memiliki akar kata yang sama dianggap sebagai satu fitur tunggal (Dina et al., 2025).

2.2.4 Ekstraksi Fitur FastText

FastText adalah metode word embedding yang menghasilkan representasi kata berdasarkan informasi sub-word menggunakan karakter n-grams. FastText menangani variasi bahasa dan imbuhan yang sering ditemukan dalam bahasa Indonesia, dengan setiap kata dipecah menjadi serangkaian unit kecil bernama n-grams. Vektor akhir dari sebuah kata didapat dengan merata-ratakan vektor dari unit-unit n-grams penyusunnya (Pangestu et al., 2024). Meskipun FastText memiliki keunggulan utama dalam menangkap informasi sub-word yang memungkinkan pembentukan representasi kata di luar kosakata pelatihan (out-of-vocabulary) (Atika et al., 2025). Namun, dalam penelitian ini metode tersebut tetap



diterapkan pada data yang telah melalui serangkaian text preprocessing yang terstandarisasi. Hal tersebut dilakukan karena metode ini merepresentasikan setiap kata sebagai kumpulan n-gram karakter, sehingga sangat bergantung pada kualitas input untuk menangani variasi bentuk kata secara efektif (Mukti & Agustian, 2022). Melalui pendekatan ini, performa FastText dapat dievaluasi secara murni berdasarkan kemampuannya menghasilkan vektor fitur yang kemudian diklasifikasikan menggunakan algoritma Random Forest untuk mengukur tingkat akurasi dan F1-score dalam analisis sentimen. Secara matematis, representasi kata dalam FastText dinyatakan sebagai berikut (Ermawan & Cahyono, 2025):

$$\vec{v}_w = \frac{1}{|G(w)|} \sum_{g \in G(w)} z_g \quad (1)$$

2.2.5 Ekstraksi Fitur TF-IDF

Term Frequency–Inverse Document Frequency (TF-IDF) adalah metode pembobotan kata yang digunakan untuk menentukan tingkat kepentingan suatu kata dalam sebuah dokumen berdasarkan frekuensi kemunculannya dan distribusinya di seluruh kumpulan dokumen atau korpus (Sari et al., 2023). TF-IDF menentukan bobot kata dengan mengombinasikan frekuensi kemunculan kata dalam dokumen (TF) dan tingkat kelangkaannya secara global (IDF) untuk menonjolkan istilah yang paling representatif (Bayu et al., 2021). Secara matematis, rumus TF-IDF dinyatakan sebagai berikut (Sari et al., 2023) :

$$IDF(t) = \log \left(\frac{N+1}{df_t} \right) \quad (2)$$

$$W_{d,t} = TF_{t,d} \times IDF_t \quad (3)$$

2.2.6 10 Fold Cross Validation

Pembagian data dilakukan menggunakan metode 10-fold cross-validation, kumpulan data dibagi menjadi sepuluh bagian dengan ukuran yang sama. Setiap bagian digunakan secara bergantian sebagai data uji, sementara sembilan bagian lainnya berfungsi sebagai data latih. Pendekatan K-fold cross-validation ini membantu mengurangi risiko overfitting, karena model dievaluasi melalui berbagai kombinasi data latih dan data uji (Wajhillah & Wibowo, 2021; Wijiyanto et al., 2024).

2.2.7 Klasifikasi Random Forest

Random Forest adalah algoritma ensemble learning yang menggabungkan prediksi dari banyak pohon keputusan (decision trees). Penentuan kelas akhir dilakukan berdasarkan pemungutan suara terbanyak (majority voting), di mana kelas yang memperoleh suara tertinggi dari seluruh pohon keputusan akan dipilih sebagai hasil akhir (Tamba, 2022). Hasil dari Random Forest ditentukan dengan menggabungkan prediksi dari setiap pohon individu, menggunakan mekanisme majority voting untuk tugas klasifikasi dan perhitungan rata-rata untuk tugas klasifikasi (Siswa Yoga & Pranoto, 2023). Proses ini dapat dirumuskan melalui persamaan berikut (Mahmuda, 2024):

$$l(y) = \arg_c \max \left(\sum_{n=1}^N I_{h_n(y)=c} \right) \quad (4)$$

2.2.8 Evaluasi Confusion Matrix

Confusion matrix adalah teknik yang digunakan untuk mengevaluasi kinerja algoritma klasifikasi. Dalam penggunaannya confusion matrix memberikan gambaran yang jelas mengenai seberapa baik model dalam mengelompokkan data (Cucun et al., 2024). Evaluasi model dilakukan menggunakan confusion matrix melalui empat metrik utama, akurasi untuk mengukur proporsi klasifikasi yang benar, recall untuk kemampuan menemukan informasi relevan, presisi untuk tingkat ketepatan prediksi, dan F1-score sebagai keseimbangan antara keduanya. Formulasi metrik tersebut dirumuskan sebagai berikut (Pasaribu & Ida Ayu, 2025):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 \text{ Score} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (8)$$

Penerapan keempat metrik evaluasi tersebut bertujuan untuk memberikan gambaran yang lebih menyeluruh mengenai performa model klasifikasi dari berbagai sudut pandang. Akurasi digunakan untuk melihat efektivitas model secara umum, sedangkan presisi dan recall berfungsi untuk memastikan bahwa setiap label sentimen diprediksi dengan tingkat kesalahan yang minimal. Penggunaan F1-score menjadi sangat penting karena metrik ini menggabungkan nilai presisi dan recall untuk menjamin stabilitas performa sistem pada data tekstual yang kompleks. Seluruh hasil perhitungan ini nantinya akan menjadi dasar utama dalam membandingkan keunggulan antara metode FastText dan TF-IDF yang diterapkan dalam penelitian ini.

2.2.9 Ringkasan Konfigurasi Parameter

Seluruh parameter penting dalam teknis yang digunakan dalam penelitian ini dirangkum dalam tabel 1 berikut.

Tabel 1. Parameter Teknis dan Pengaturan Model

Komponen	Parameter	Nilai/Spesifikasi
Ekstraksi Fitur (FastText)	Vector Dimension	200
Ekstraksi Fitur (TF-IDF)	Max Features	5000
Validasi Data	K-Fold Splits	10 (Stratified)
	Shuffle & Random State	Shuffle & Random State
Model Random Forest	Number of Trees	400
	Splitting Criterion	Gini Index
	Max Features	5000
Metrik Evaluasi	Class Weight	Balanced
	Average Type	Weighted

Berdasarkan Tabel 1, penetapan parameter penelitian ini didasarkan pada rujukan ilmiah untuk menjaga validitas eksperimen. Dimensi vektor FastText merujuk pada (Rahman et al., 2021), sementara batas fitur (max features) TF-IDF dan Random Forest merujuk pada (Magnolia et al., 2022). Prosedur K-Fold didasarkan pada (Wijiyanto et al., 2024). Pada model Random Forest, jumlah pohon mengikuti (Putra & Rumini, 2025), penggunaan Gini index berdasarkan (Suci Amaliah et al., 2022), dan pengaturan class weight merujuk pada (Akbar & Sanjaya, 2023). Terakhir, pendekatan weighted average pada evaluasi merujuk pada (Bifadhlillah et al., 2025).

3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

Bab ini menyajikan hasil penelitian dari implementasi dari komparasi FastText dan TF-IDF terhadap kinerja model Random Forest dalam analisis sentimen pembangunan IKN di YouTube. Berikut pembahasan yang didasarkan pada hasil pelaksanaan tahapan penelitian yang telah dijelaskan pada bab sebelumnya sebagai berikut.

3.1.1 Pengambilan Data Komentar Youtube

Dalam penelitian ini, data diambil melalui YouTube Data API V3 menggunakan Google Colab dari lima saluran YouTube, yaitu @TitikNol_Nusantara, @iknindonesia4397, @DIAN-RANA, @topkontruksi, dan @tempovideochannel dalam periode Agustus 2025. Periode tersebut dipilih untuk menangkap lonjakan interaksi publik di media sosial terkait progres dan capaian pembangunan proyek IKN. Sebanyak 80 video berhasil dikumpulkan, yang menghasilkan 4.093 komentar berdasarkan kata kunci "IKN" dan "Nusantara". Data tersebut awalnya disimpan ke dalam format XLSX (data.xlsx). Hasil dari proses pembagian data ditunjukkan pada gambar 3.

1	video_id	komentar
2	dywMpVoqE-w	Gpp yg penting IKN 100% pindah , asal jangan dijakarta lagi karna disana pusatnya kadrun2 ba'alawi 🤔🤔
3	dywMpVoqE-w	IKN bakar duit negara
4	dywMpVoqE-w	Kalao sampe sekelas presiden ngomongin ginian... keadaanya sebaliknya apa yg d ucapkan
5	dywMpVoqE-w	Jelas aja Ekosistem di sekitar nya blm terbentuk, Dari Srb undangan akhir nya turun menjadi +- 1rb undangan, akibat terburu buru & serba di paksakan, membangun kota, apalagi ibukota, Ekosistem nya sekitar nya lah yg harus di tumbuh kembangkan, Kota Balikpapan sbg kota penyangga saja masih blm siap sepenuhnya,Salam Dari IKN..
6	dywMpVoqE-w	Bila Presiden/wakil dilantik bukan di istana negara yang ada di ibukota negara maka presiden/wakil tersebut tidak sah dan tidak mewakili rakyat dan negara Indonesia.Sampai detik ini RI tidak/belum punya ibukota negara,status DKI Jakarta sudah dicabut,seandainya status IKN sebagai Ibukota negara belum resmi dan belum ada undang2nya yang sah.

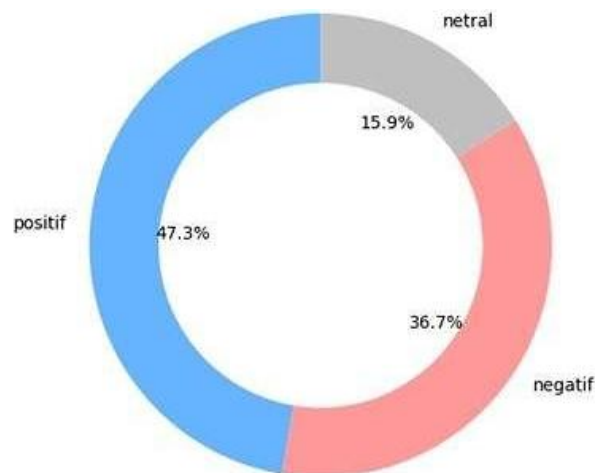
Gambar 3. Data Komentar IKN

Gambar 3 menampilkan kolom Video_id sebagai identitas sumber video dan kolom Comment sebagai data utama penelitian. Mayoritas data berupa komentar panjang berisi tanggapan, kritik, dan opini, di samping komentar pendek dengan konteks terbatas. Tingginya penggunaan bahasa informal, singkatan (slang), serta gangguan (noise) seperti emoji meningkatkan kompleksitas pemaknaan, terutama pada kalimat dengan struktur tidak beraturan dan minim konteks.

3.1.2 Hasil Pelabelan Data

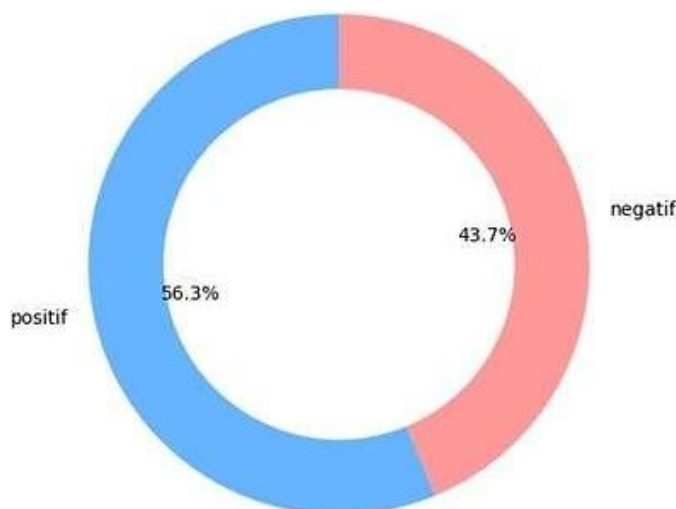
Proses anotasi data dilakukan secara manual melalui kolaborasi antara pakar linguistik dan peneliti untuk mengategorikan dataset ke dalam tiga label awal, yaitu positif (apresiasi), negatif (kritik), dan netral (informatif). Dari jumlah awal sebanyak 4.093 komentar, sebaran frekuensi menunjukkan dominasi sentimen positif sebanyak 1.937 data,

diikuti sentimen negatif sebanyak 1.504 data, dan netral sebanyak 652 data. Untuk meningkatkan efektivitas klasifikasi serta mereduksi ambiguitas fitur pada ruang vektor, dilakukan eliminasi terhadap kelas netral. Langkah ini merupakan strategi untuk mempertegas batas keputusan (decision boundary) model dalam melakukan pemisahan sentimen secara biner. Distribusi tiga kelas sentimen sebelum proses eliminasi disajikan pada gambar 4.



Gambar 4. Distribusi Tiga Kelas Sentimen

Visualisasi pada Gambar 4 menunjukkan bahwa distribusi sentimen awal memiliki proporsi sebesar 47.3% positif, 36.7% negatif, dan 15.9% netral. Pada tahap pengembangan selanjutnya, fokus studi dialihkan pada skema klasifikasi biner (binary classification) dengan mengeksklusi kategori netral. Transformasi ini bertujuan untuk memaksimalkan polaritas data sehingga model dapat mengidentifikasi sentimen publik terhadap IKN dengan tingkat presisi yang lebih tinggi. Hasil akhir dari penyesuaian distribusi data tersebut menghasilkan dataset baru sebanyak 3.441 komentar, sebagaimana diilustrasikan pada Gambar 5



Gambar 5. Distribusi Dua Kelas Sentimen

Sebagaimana diilustrasikan pada Gambar 5, penghapusan kelas netral menghasilkan distribusi akhir sebesar 56.3% untuk kelas positif dan 43.7% untuk kelas negatif. Meskipun terdapat selisih proporsi antar kelas, menunjukkan tingkat ketidakseimbangan yang proporsional, sehingga masa data pada kelas minoritas tetap berada di atas ambang batas signifikan untuk menjaga stabilitas batas keputusan (decision boundary) dan mencegah pergeseran fungsi objektif model pada arah kelas mayoritas dalam proses minimalisasi error rate. distribusi ini cukup representatif untuk melatih model Random Forest dalam mengenali karakteristik fitur semantik dari kedua ekstraksi fitur.

3.1.3 Hasil Pre-Processing

Konsistensi data di seluruh rangkaian preprocessing telah dipastikan, dengan volume dataset yang tetap terjaga sebanyak 3.441 komentar, mulai dari data cleaning sampai tahap akhir stemming. Pendekatan sistematis ini menjamin bahwa proses hanya berfokus pada perbaikan struktur dan bahasa seperti membakukan kata-kata gaul (slang) dan mengembalikan kata ke bentuk dasarnya tanpa mengurangi jumlah sampel data. Tahap stemming sebagai akhir dari preprocessing divisualisasikan menggunakan word cloud untuk sentimen positif dan negatif, ditampilkan pada Gambar 6 dan Gambar 7.



Gambar 6. Word Cloud Positif

Gambar 6 menyajikan kata-kata berorientasi positif hasil dari proses stemming, yang didominasi oleh istilah seperti "bangun", "jadi", "moga", dan "nusantara". Dominasi kata-kata tersebut mencerminkan dukungan publik, optimisme, serta harapan terkait pembangunan Ibu Kota Negara (IKN) yang baru.



Gambar 7. Word Cloud Negatif

Gambar 7 menunjukkan kata-kata berorientasi negatif hasil dari proses stemming, yang didominasi oleh istilah "tidak", "bangun", "bukan", dan "kalau". Munculnya kata-kata tersebut mengindikasikan adanya ungkapan keraguan, kritik, serta penolakan terhadap pembangunan Ibu Kota Negara (IKN). Perlu dicatat bahwa kata "bangun" sering kali muncul dalam konteks penyangkalan, sehingga memperkuat sentimen negatif yang ada.

3.1.4 Hasil Ekstraksi Fitur FastText dan TF-IDF

Fasttext melatih model word embedding (penanaman kata) tanpa pengawasan (unsupervised) dengan pendekatan skip-gram. Setiap komentar direpresentasikan sebagai vektor numerik dengan dimensi sebesar 200. Proses tersebut menghasilkan matriks fitur berukuran 3.441×200 , yang selanjutnya menjadi input bagi model klasifikasi. Contoh hasil ekstraksi fitur FastText disajikan pada Tabel 2.

Tabel 2. Vektor FastText

No	Komentar	Sentimen	Vektor FastText
1	tidak apa apa penting ikn pindah asal jangan ...	Negatif	-0.0148, -0.0423, -0.0186, ...
2	ikn bakar duit negara	Negatif	-0.0131, -0.0080, -0.0581, ...
3	kalau kelas presiden ngomongim ginian keadaany...	Negatif	-0.0061, -0.0694, -0.0595, ...
4	jelas ekosistem nya belum bentuk undang akh...	Negatif	-0.0194, -0.0120, 0.0060, ...
5	bila presiden wakil lantik bukan istana negara...	Negatif	0.0016, -0.0305, 0.0072, ...

Tabel 2 mengilustrasikan transformasi teks ke dalam ruang vektor kontinu, di mana setiap komentar dikonversi menjadi urutan 200 nilai desimal. Berbeda dengan metode berbasis frekuensi, nilai-nilai seperti -0.0148, -0.0423, -0.0186, merupakan pemetaan semantik yang menangkap konteks dan hubungan antar kata di seluruh korpus. Dengan nilai tersebut memungkinkan algoritma Random Forest mengenali pola sentimen melalui representasi makna.

Ekstraksi fitur TF-IDF menggunakan batasan 5.000 fitur dengan penggabungan unigram dan bigram untuk mempertahankan informasi kontekstual dan menjaga efisiensi komputasi model. Perhitungan frekuensi istilah secara sublinear (sublinear term frequency scaling) diterapkan untuk meminimalkan pengaruh kata-kata berfrekuensi tinggi yang berisiko mendominasi ruang fitur. Distribusi statistik dan bobot dari 10 kata atau istilah teratas dirangkum dalam tabel 3.

**Tabel 3.** Vektor TF-IDF

Term	TF	DF	IDF	Average TF-IDF
indonesia	987	734	2.543937	0.024564
tidak	1140	803	2.454209	0.024374
bangun	837	617	2.717320	0.019980
jadi	798	593	2.756929	0.019782
jokowi	706	518	2.891904	0.019228
nusantara	424	379	3.203637	0.018213
maju	489	416	3.110722	0.017624
mantap	204	197	3.855541	0.016140
rakyat	544	402	3.144871	0.015610
kota	494	358	3.260486	0.014777

Tabel 3 menyajikan kata dengan nilai TF, DF, IDF, dan rata-rata TF-IDF tertinggi dalam korpus komentar IKN. Sebagai contoh, kata "indonesia" memiliki nilai DF sebesar 734, menunjukkan kemunculannya di sebagian besar dokumen. Berdasarkan nilai DF tersebut, diperoleh nilai IDF sebesar 2,543937. Nilai rata-rata TF-IDF mencerminkan kontribusi keseluruhan dari setiap istilah terhadap matriks fitur. Istilah dengan bobot tinggi seperti "indonesia", "tidak", "bangun", "jadi", dan "jokowi" berfungsi sebagai indikator kunci bagi pengklasifikasi Random Forest. Penggunaan istilah "tidak" (dengan bobot rata-rata 0,024374) sangat menonjol, karena kata ini sering kali menjadi titik penentu dalam ekspresi sentimen negatif terkait pembangunan ibu kota baru tersebut.

3.1.5 Hasil Pembagian Data (10-Fold Cross-Validation)

Dataset penelitian ini terdiri dari 3.441 komentar (1.937 positif dan 1.504 negatif) yang dibagi menggunakan metode Stratified 10-Fold Cross-Validation. Penggunaan random state yang konsisten menjamin distribusi data identik pada pengujian TF-IDF maupun FastText, sehingga menghasilkan perbandingan performa yang objektif. Rincian distribusi data pelatihan dan pengujian untuk setiap fold disajikan pada Tabel 4.

Tabel 4. Distribusi 10-Fold Cross-Validation

Fold ke	Training Set			Test Set		
	Positif	Negatif	Total	Positif	Negatif	Total
1	1743	1353	3096	194	151	345
2	1744	1353	3097	193	151	344
3	1744	1353	3097	193	151	344
4	1744	1353	3097	193	151	344
5	1743	1354	3097	194	150	344
6	1743	1354	3097	194	150	344
7	1743	1354	3097	194	150	344
8	1743	1354	3097	194	150	344
9	1743	1354	3097	194	150	344
10	1743	1354	3097	194	150	344

Tabel 4 menunjukkan konsistensi distribusi data di seluruh fold, dengan jumlah sampel pelatihan berkisar 3.096–3.097 dan data pengujian 344–345 komentar, memastikan bahwa metrik performa yang diperoleh pada tahap evaluasi model bersifat representatif terhadap keseluruhan dataset.

3.1.6 Hasil Pengujian dan Evaluasi Model

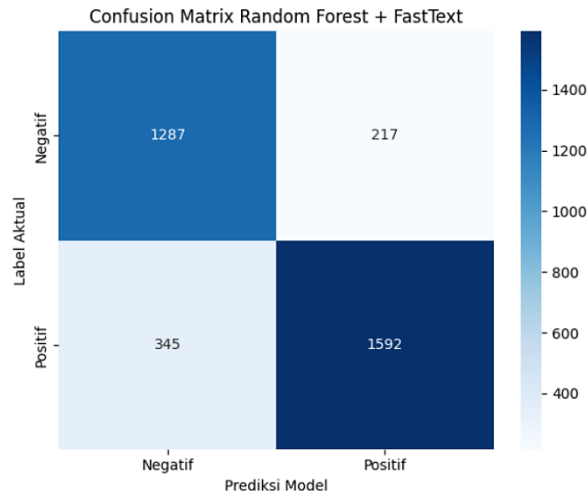
Proses evaluasi melalui skema 10-Fold Cross Validation memastikan bahwa model memiliki tingkat generalisasi yang baik dan tidak hanya unggul pada sebagian data saja. Dengan membagi dataset menjadi sepuluh bagian yang berbeda, setiap iterasi memberikan gambaran mengenai konsistensi performa algoritma saat menghadapi variasi data teks yang beragam. Penggunaan metrik akurasi pada setiap fold berfungsi sebagai parameter statistik untuk mengukur sejauh mana metode ekstraksi fitur mampu merepresentasikan pola informasi dari komentar YouTube secara tepat. Melalui perbandingan skor dari tiap iterasi, dapat diidentifikasi stabilitas antara model yang menggunakan FastText dengan model yang menggunakan TF-IDF. Pendekatan ini sangat penting untuk meminimalkan risiko bias pada hasil akhir penelitian dan memastikan bahwa kesimpulan yang diambil didasarkan pada data pengujian yang komprehensif. Hasil performanya ditampilkan dalam tabel 5.

Tabel 5. Perbandingan Akurasi Model pada Setiap Fold

Fold Ke	RF + FastText	RF + TF-IDF
1	81.45%	79.71%
2	81.69%	79.07%
3	81.69%	81.69%

Fold Ke	RF + FastText	RF + TF-IDF
4	84.01%	79.07%
5	87.21%	81.40%
6	85.76%	83.72%
7	86.34%	81.40%
8	84.59%	78.78%
9	82.56%	83.43%
10	81.40%	77.03%
Rata-Rata	83.67%	80.53%

Berdasarkan Tabel 5, model Random Forest dengan fitur FastText menunjukkan performa yang lebih unggul dengan mencapai rata-rata akurasi sebesar 83.67%, sementara TF-IDF mencapai 80.53%. Model berbasis FastText menunjukkan akurasi yang stabil di seluruh fold dan mencapai akurasi maksimum sebesar 87,21%, sedangkan model berbasis TF-IDF menunjukkan variasi yang lebih besar. Dalam membuktikan validitas hasil penelitian, dilakukan pengujian untuk melihat apakah selisih rata-rata akurasi antara FastText sebesar 83.67% dan TF-IDF sebesar 80.53% memiliki signifikansi secara statistik, dilakukan uji statistik Paired Sample T-Test terhadap hasil 10-fold cross-validation. Berdasarkan perhitungan tersebut diperoleh nilai T-Statistic sebesar 4.1457 dan P-Value sebesar 0.0025. Karena nilai P-value sebesar 0.0025, oleh karena itu hipotesis nol ditolak. Hal tersebut membuktikan bahwa penggunaan fitur FastText memberikan peningkatan akurasi yang signifikan secara statistik dibandingkan TF-IDF, sehingga perbedaan performa tersebut bukan merupakan hasil kebetulan (noise). Analisis lebih lanjut menggunakan confusion matrix diperlukan untuk menilai keseimbangan prediksi model, sebagaimana disajikan pada Gambar 8.



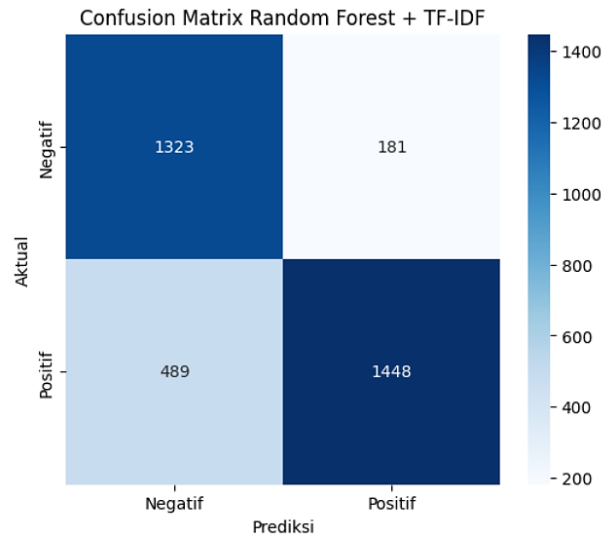
Gambar 8. Confusion Matrix Random Forest dan FasText

Gambar 8 menyajikan hasil confusion matrix untuk model berbasis FastText, di mana nilai True Negative (TN) adalah 1.287, False Positive (FP) sebesar 217, False Negative (FN) sebesar 345, dan True Positive (TP) mencapai 1.592. Keempat nilai ini menjadi dasar dalam penghitungan metrik evaluasi, yaitu Akurasi, Presisi, Recall, dan F1-score, sebagaimana ditunjukkan pada Tabel 6.

Tabel 6. Metrik Evaluasi pada Random Forest dan FastText

Kelas	Precision	Recall	F1-Score	Accuracy
Negatif	0,79	0,86	0,82	0,8367
Positif	0,87	0,82	0,85	

Hasil evaluasi pada Tabel 6 menunjukkan bahwa model mencapai akurasi 83.67%. Namun, analisis mendalam pada confusion matrix mengungkapkan adanya 345 kasus False Negative (FN), terdapat sejumlah komentar positif yang salah diklasifikasikan sebagai negatif. Tingginya angka FN tersebut berdampak pada nilai recall kelas positif yang hanya mencapai 0.82, lebih rendah dibandingkan recall kelas negatif sebesar 0.86. Fenomena tersebut menunjukkan adanya bias model dalam menangani distribusi data yang tidak seimbang. Nilai precision kelas positif memiliki skor lebih tinggi sebesar 0.87 dibandingkan kelas negatif yang hanya mencapai 0.79. Hasil tersebut mengindikasikan bahwa ketika model memprediksi sebuah komentar sebagai positif, tingkat kepastiannya sangat kuat, namun model menjadi terlalu selektif sehingga banyak data positif lainnya yang terlewat menjadi FN. Secara teknis, perbedaan recall tersebut mengindikasikan bahwa batas keputusan model lebih condong ke arah kelas mayoritas, dipicu oleh adanya tumpang tindih (overlap) fitur semantik pada FastText. Pola fitur kelas negatif yang lebih dominan memberikan pengaruh lebih besar pada hasil prediksi dibandingkan kelas positif. Selanjutnya analisis TF-IDF dari confusion matrix di visualisasikan pada gambar 9.



Gambar 9. Confusion Matrix Random Forest dan TF-IDF

Gambar 9 menyajikan hasil confusion matrix untuk model Random Forest dengan fitur TF-IDF, di mana nilai True Negative (TN) adalah 1.323, False Positive (FP) sebesar 181, False Negative (FN) sebesar 489, dan True Positive (TP) mencapai 1.448. Keempat nilai ini menjadi dasar dalam penghitungan metrik evaluasi yang disajikan pada Tabel 7.

Tabel 7. Metrik Evaluasi pada Random Forest dan TF-IDF

Kelas	Precision	Recall	F1-Score	Accuracy
Negatif	0,73	0,88	0,80	0,8053
Positif	0,89	0,75	0,81	

Berdasarkan Tabel 7, model TF-IDF kurang optimal dalam mengenali kelas positif dengan angka False Negative mencapai 489 data, sehingga recall positif turun menjadi 0.75. Hal ini membuktikan kerentanan TF-IDF terhadap ketidakseimbangan data atau data imbalance. Model lebih cenderung memprediksi kelas negatif yang memiliki pola kata lebih konsisten. Perbedaan antara presisi positif sebesar 0.89 dan recall positif sebesar 0.75 yang memiliki selisih cukup lebar yaitu 0.14 mempertegas bahwa TF-IDF mengalami kesulitan dalam melakukan generalisasi pada sentimen positif. Skor F1-score yang lebih rendah pada kelas negatif sebesar 0.80 dibandingkan kelas positif sebesar 0.81 menunjukkan bahwa meskipun recall negatif tinggi dan model banyak melakukan kesalahan salah tangkap atau False Positive pada kelas tersebut. Pengamatan pada distribusi fitur menunjukkan bahwa pendekatan berbasis frekuensi belum mampu memberikan pemisahan yang cukup kuat antar kelas. Dengan nilai FN mencapai 14.2% dari total dataset, variasi kata pada kelas positif nampaknya tertutup oleh bobot informasi dari kelas negatif. Akibatnya model mengalami kesulitan melakukan generalisasi pada sentimen positif yang tidak memiliki kata kunci unik dengan bobot IDF yang tinggi. Kondisi tersebut menunjukkan bahwa pembobotan fitur yang didasarkan pada statistik kemunculan kata memerlukan pola distribusi yang lebih spesifik untuk dapat membedakan setiap label sentimen secara maksimal. Karakteristik TF-IDF yang menitikberatkan pada keunikan kata kunci cenderung memberikan hasil yang berbeda ketika dihadapkan dengan variasi bahasa informal yang sangat beragam pada data teks.

3.2 Pembahasan

Berdasarkan evaluasi menggunakan confusion matrix, penelitian ini mengukur metrik akurasi, presisi, recall, dan F1-score dari model Random Forest dengan membandingkan representasi teks FastText dan TF-IDF. Hasil pengukuran menunjukkan bahwa FastText secara konsisten memberikan performa yang lebih unggul dibandingkan TF-IDF pada mayoritas metrik evaluasi. Hal tersebut diperkuat oleh hasil uji signifikansi menggunakan Paired Sample T-Test yang menghasilkan nilai p-value sebesar 0.0025. Karena nilai $p < 0.05$ menegaskan perbedaan akurasi antara kedua model adalah signifikan secara statistik dan bukan merupakan hasil dari kebetulan pada pembagian data. FastText menghasilkan akurasi sebesar 83.67%, mengungguli TF-IDF yang mencapai 80.53%. Meskipun TF-IDF mencatat nilai presisi kelas positif sedikit lebih tinggi yaitu 0.89 dibandingkan FastText sebesar 0.87, hal tersebut diikuti dengan penurunan drastis pada nilai recall positif sebesar 0.75. Fenomena tersebut menunjukkan adanya masalah False Negative (FN) yang kritis pada TF-IDF dengan total 489 data terklasifikasi salah. Tingginya angka FN pada TF-IDF membuktikan bahwa pendekatan berbasis frekuensi kaku gagal melakukan generalisasi pada sentimen positif yang tidak memiliki bobot informasi (IDF) yang kuat, sehingga model cenderung bersifat konservatif dalam memberikan prediksi positif.

Penggunaan FastText terbukti mampu mereduksi angka FN sebesar 29.4% jika dibandingkan dengan TF-IDF. Secara teknis, representasi vektor semantik pada FastText memberikan daya tahan lebih baik terhadap distribusi data yang tidak seimbang (56.3% positif dan 43.7% negatif) dengan menangkap hubungan antar kata secara kontekstual.



Namun, munculnya 345 kasus FN pada FastText mengonfirmasi bahwa batas keputusan (decision boundary) model tetap mengalami sedikit pergeseran ke arah kelas mayoritas (negatif), di mana pola-pola kritis pada dataset ini memiliki densitas fitur yang lebih dominan sehingga menekan probabilitas prediksi pada kelas positif yang memiliki struktur bahasa lebih informal. Meskipun pengaturan `class_weight="balanced"` telah diimplementasikan, efektivitas model tetap ditentukan oleh kualitas ekstraksi fitur. Hal ini terefleksi pada F1-score, di mana FastText mencapai 0.84 sedangkan TF-IDF tertahan pada 0.81. Perbedaan performa tersebut menunjukkan bahwa FastText memiliki keunggulan dalam menjaga keseimbangan antara ketepatan (*precision*) dan kelengkapan (*recall*) pada dataset. Secara keseluruhan, integrasi FastText dengan Random Forest memberikan stabilitas yang lebih konsisten dalam menangani ambiguitas semantik, sementara TF-IDF tetap menunjukkan efektivitasnya dalam memetakan frekuensi kata yang spesifik. Meskipun demikian, kendala bias akibat ketidakseimbangan data tetap menjadi batasan teknis bagi kedua model yang memerlukan optimasi lebih lanjut pada tahap pemrosesan data.

4. KESIMPULAN

Berdasarkan hasil analisis sentimen terhadap komentar YouTube terkait pembangunan Ibu Kota Nusantara (IKN), kombinasi metode ekstraksi fitur FastText dan algoritma Random Forest menunjukkan performa terbaik dengan rata-rata akurasi sebesar 83,67%, lebih tinggi dibandingkan TF-IDF yang memperoleh akurasi sebesar 80,53%. Hasil tersebut menunjukkan bahwa pendekatan berbasis word embedding lebih efektif dalam merepresentasikan konteks semantik pada komentar media sosial dibandingkan pendekatan berbasis frekuensi kata. Keunggulan FastText terletak pada kemampuannya dalam merepresentasikan hubungan semantik melalui informasi subword, sehingga lebih mampu menangani variasi bahasa informal, singkatan, dan kata tidak baku yang umum ditemukan pada komentar YouTube. Temuan tersebut mengindikasikan bahwa keberhasilan klasifikasi sentimen pada media sosial tidak hanya dipengaruhi oleh algoritma klasifikasi, tetapi juga oleh kemampuan metode ekstraksi fitur dalam memahami karakteristik linguistik data nonformal. Namun, tingginya nilai False Negative menunjukkan bahwa model masih mengalami keterbatasan dalam mengenali komentar yang mengandung istilah gaul, dialek daerah, serta bentuk bahasa informal yang ambigu. Kondisi ini menunjukkan bahwa representasi semantik FastText belum sepenuhnya mampu memahami variasi linguistik nonformal yang sangat dinamis pada media sosial. Oleh karena itu, penelitian selanjutnya disarankan untuk mengintegrasikan normalisasi slang berbasis kamus bahasa informal, lexicon media sosial, atau fine-tuning embedding menggunakan korpus bahasa Indonesia nonformal agar kemampuan model dalam memahami konteks sentimen dapat ditingkatkan secara lebih optimal. Secara praktis, hasil penelitian ini menunjukkan bahwa pendekatan FastText dan Random Forest berpotensi diterapkan pada sistem monitoring opini publik berbasis media sosial untuk membantu proses analisis respons masyarakat terhadap isu strategis nasional secara lebih cepat dan terukur. Namun, penelitian ini masih memiliki keterbatasan pada penggunaan klasifikasi biner dan sumber data yang hanya berasal dari platform YouTube, sehingga pengembangan pada klasifikasi multi-kelas dan integrasi data lintas platform media sosial menjadi peluang penelitian selanjutnya.

REFERENCES

- Akbar, H., & Sanjaya, W. K. (2023). Kajian Performa Metode Class Weight Random Forest pada Klasifikasi Imbalance Data Kelas Curah Hujan. *Jurnal Sains, Nalar, Dan Aplikasi Teknologi Informasi (SNATI)*, 3(1), 42–49. <https://doi.org/10.20885/snati.v3i1.3094>
- Arif, M. W., & Kustiyono, K. (2025). Analisis Sentimen Kebijakan Makan Bergizi Gratis di Media Sosial Menggunakan Natural Language Processing Berbasis Python TextBlob di Indonesia. *Jurnal Pendidikan Dan Teknologi Indonesia*, 5(9), 2463–2471. <https://doi.org/10.52436/1.jpti.931>
- Atika, P., Surya, A., Jasril, & Iis, A. (2025). Eksplorasi Fitur FastText, TF-IDF Dan IndoBERT Pada Metode K-Nearest Neighbor Untuk Klasifikasi Sentimen. *Jurnal Sistem Informasi (ZONASI)*, 7(1), 49–60. <https://doi.org/10.31849/zn.v7i1.24779>
- Ayunda, C., Soemedhy, A., Trivetisia, N., Winanti, N. A., Martiyaningsih, D. P., Utami, T. W., & Sudianto, S. (2022). Analisis Komparasi Algoritma Machine Learning untuk Sentiment Analysis (Studi Kasus: Komentar YouTube ‘Kekerasan Seksual’). *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, 7(2), 80–84. <https://doi.org/10.30591/jpit.v7i2.3547>
- Bayu, B. B., Susanto, I., & Khomsah, S. (2021). Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs Tripadvisor). *INISTA*, 3(2), 21–29. <https://doi.org/10.20895/inista.v3i2.203>
- Bifadhilillah, M. I., Sucipto, & Arie, N. (2025). Analisis Algoritma KNN dan Penerapan SMOTE Dalam Deteksi Dini Kanker Paru-Paru. *Jurnal Qua Teknika*, 15(2), 38–50. <https://doi.org/10.35457/quateknika.v15i02.4603>
- Calvin Jonathan, Theresia Herlina Rochadiani, & Thamrin Sofian. (2023). Analisis Sentimen Komentar Video Youtube Flat Earth Theory Dengan Menggunakan Metode Unsupervised Dan Supervised Learning. *Decode: Jurnal Pendidikan Teknologi Informasi*, 3(2), 378–387. <https://doi.org/10.51454/decode.v3i2.210>
- Cucun, V. A., Mochamad Adrian, N. T., Budi, D. S., & Ari, K. (2024). Optimasi Klasifikasi Sentimen Menggunakan Random Forest dengan Preprocessing K-Means Clustering dan SMOTE. *JEPIN : Jurnal Edukasi Dan Penelitian Informatika*, 10(3), 389–400. <https://doi.org/10.26418/jp.v10i3.84514>



- Dina, S. N., Nining, R., Raditya, D. D., & Cep Lukman, R. (2025). Penerapan Algoritma Naive Bayes Dalam Analisis Sentimen Ulasan Aplikasi Kita Lulus Di Google Play Store. *Jurnal Informatika Terpadu*, 11(1), 213–223. <https://doi.org/10.54914/jit.v11i1.1544>
- Ermawan, B. R., & Cahyono, N. (2025). Optimasi Metode Klasifikasi Menggunakan FastText dan GRID Search Pada Analisis Sentimen Ulasan Aplikasi Seabank. *JIKO (Jurnal Informatika Dan Komputer)*, 9(1), 226–238. <https://doi.org/10.26798/jiko.v9i1.1523>
- Fitriyah Syarifatul. (2025). Analisis Penerapan Teori Pembangunan Rostow Terhadap Pembangunan Ibu Kota Negara (IKN) di Kalimantan Timur. *JEPP: Jurnal Ekonomi Pembangunan Dan Pariwisata*, 5(1), 1–11. <https://doi.org/10.52300/jep.v5i1.19822>
- Huwaida, S. F., Kusumawati, R., & Isnaini, B. (2024). Analisis Sentimen Komentar YouTube terhadap Pemindahan Ibu Kota Negara Menggunakan Metode Naive Bayes. *Jambura Journal of Informatics*, 6(1), 26–39. <https://doi.org/10.37905/jji.v6i1.24718>
- Iman, A. K., & Ujianto, E. I. H. (2024). Analisis Sentimen Pemindahan Ibu Kota Indonesia Menggunakan K-Nearest Neighbor. *Jurnal Pendidikan Dan Teknologi Indonesia*, 4(12), 759–768. <https://doi.org/10.52436/1.jpti.546>
- Jasmine, A. M., Kinaya, K. K., Wildan, H., Muhammad Galuh, G., Reza, P., & Humannisa, R. L. (2025). Analisis Sentimen Ulasan Aplikasi HeyJapan di Google Play Store Menggunakan Algoritma NLP. *Pragmatik: Jurnal Rumpun Ilmu Bahasa Dan Pendidikan*, 3(3), 157–167. <https://doi.org/10.61132/pragmatik.v3i3.1801>
- Kadir, S. F., & Fairuzabadi, A. (2025). Analisis Sentimen Ulasan Shopee di Google Play dengan TF-IDF dan Logistic Regression. *RIGGS: Journal of Artificial Intelligence and Digital Business*, 4(2), 7940–7944. <https://doi.org/10.31004/riggs.v4i2.2850>
- Laia, N. A., & Barus, S. P. (2025). Analisis Sentimen YouTube: ‘Di Balik Ambisi Jokowi dalam IKN’. *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, 5(1), 7–12. <https://doi.org/10.55382/jurnalpustakaai.v5i1.891>
- Larasakti, D. N., Aziz, A., & Aditya, D. (2023). Analisis Sentimen Komentar Video Youtube Dengan Metode K-Nearest Neighbor. *Jurnal Ilmiah Wahana Pendidikan*, 9(5), 132–142. <https://doi.org/10.5281/zenodo.7728573>
- Magnolia, C., Nurhopipah, A., & Kusuma, A. B. (2022). Penanganan Imbalanced Dataset untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter. *Du Komputika: Jurnal Ilmiah Pendidikan Informatika*, 9(2), 105–113. <https://doi.org/10.15294/edukomputika.v9i2.61854>
- Mahmuda, S. (2024). Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube. *Jurnal Jendela Matematika*, 2(01), 21–31. <https://doi.org/10.57008/jjm.v2i01.633>
- Melati, R., & Reza, M. (2024). Analisis Sentimen Data Twitter Menggunakan Metode K-Means Clustering Pada Studi Kasus Pemindahan Ibu Kota Nusantara (IKN). *Jurnal TAM (Technology Acceptance Model)*, 15(1), 66–73. <https://doi.org/10.56327/jurnaltam.v15i1.1670>
- Mola, S. A. S., Iqbal Muhammad Iskandar, Pidu Dimu, J. E., & Seran, W. Y. (2024). Analisis Sentimen Pembangunan Ibu Kota Negara Indonesia Menggunakan Metode Naive Bayes, dan K-Nearest Neighbor. *HOAQ (High Education of Organization Archive Quality): Jurnal Teknologi Informasi*, 15(2), 151–157. <https://doi.org/10.52972/hoaq.vol15no2.p151-157>
- Muhammad Rizki Syafapri, Elin Haerani, Iwan Iskandar, & Liza Afriyanti. (2024). Klasifikasi sentimen terhadap larangan pernikahan beda agama menggunakan metode Naive Bayes Classifier. *Jurnal CoSciTech (Computer Science and Information Technology)*, 5(1), 10–18. <https://doi.org/10.37859/coscitech.v5i1.6889>
- Muhayat, T., Fauzi, A., & Indra, J. (2023). Analisis Sentimen Terhadap Komentar Video Youtube Menggunakan Support Vector Machines. *Progresif: Jurnal Ilmiah Komputer*, 19(1), 231–240. <https://doi.org/10.35889/progresif.v19i1.1060>
- Mukti, M. K., & Agustian, S. (2022). Metode SVM dengan Fitur Representasi FastText untuk Klasifikasi Sentimen Twitter Mengenai Program Vaksinasi Covid-19. *Jurnal Teknologi Informasi & Komunikasi Digital Zone*, 13(1), 140–150. <https://doi.org/10.31849/digitalzone.v13i2.11531>
- Mursyidah, Davi, M., & Novitri, D. S. (2024). Klasifikasi Sentimen Review Pengguna terhadap Aplikasi Instagram menggunakan Algoritma Random Forest. *Jurnal Informatika Dan Mesin (JIM)*, 9(2), 106–115. <https://doi.org/10.30811/jim.v9i2.6069>
- Nurannisa, N., Indriati, & Arif Rahman, M. (2026). Analisis Sentimen Opini Terhadap Bilingualisme dalam Teks Code-Mixed Indonesia-Inggris di Platform X Menggunakan Metode Attention-Based BiLSTM-CNN. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 10(1), 2548–2964. <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/16075>
- Pangestu, A. F., Rahmat, B., & Sihananto, A. N. (2024). Analisis Sentimen Pada Media Sosial X Terhadap Implementasi Kurikulum Merdeka Menggunakan Metode Fasttext Dan Long Short-Term Memory (LSTM). *JIPi (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, 9(4), 2271–2280. <https://doi.org/10.29100/jipi.v9i4.5665>
- Pasaribu, R., & Ida Ayu, G. P. S. (2025). Analisis Sentimen Ulasan Aplikasi dengan Multinomial Naive Bayes, Logistic Regression, dan SVM. *Jurnal Nasional Teknologi Informasi Dan Aplikasinya (JNATIA)*, 4(1), 63–72. <https://doi.org/10.24843/JNATIA.2025.v04.i01.p08>
- Prastyo, D., Irawan, D., & Mursyidin, I. H. (2024). Klasifikasi Sentimen Komentar YouTube dengan NLP pada Debat Pilkada Banten 2024. *Bit-Tech*, 7(2), 413–421. <https://doi.org/10.32877/bt.v7i2.1833>



- Priyanto, C. R. D. M., Azahari, & Sa'ad, I. M. (2025). Analisis Sentimen Terhadap Kontroversi Pembangunan IKN di Media Sosial Twitter Menggunakan Metode Naïve Bayes. *Bulletin of Information Technology (BIT)*, 6(2), 97–108. <https://doi.org/10.47065/bit.v6i2.1993>
- Purnomo, D., Firgiawan, W., & Nur, N. (2025). Komparasi Algoritma Random Forest, Naïve Bayes, dan SVM pada Sentimen Kebijakan PPN 12%. *Jurnal Tekno Kompak*, 19(2), 155–167. <https://doi.org/10.33365/jtk.v19i2.122>
- Putra, H., & Rumi. (2025). Comparative Study of Logistic Regression, Random Forest, and XGBoost for Bank Loan Approval Classification. *Journal of Applied Informatics and Computing (JAIC)*, 9(5), 2822–2835. <https://doi.org/10.30871/jaic.v9i5.10862>
- Rafi, R., Rahim, A., & Rudiman, R. (2024). Analisis Sentimen Ulasan “Ojol The Game” Di Google Play Store Menggunakan Algoritma Naive Bayes Dan Model Ekstraksi Fitur TF-IDF Untuk Meningkatkan Kualitas Game. *Jurnal Informatika Dan Teknik Elektro Terapan*, 12(3), 2928–2936. <https://doi.org/10.23960/jitet.v12i3.4988>
- Rahmadhani, S. A., Rusanti, L. D., & Rosyid, H. Al. (2025). Klasifikasi Sentimen Komentar Youtube Demonstrasi DPR RI Menggunakan Support Vector Machine. *Arcitech: Journal of Computer Science and Artificial Intelligence*, 5(2), 356–375. <https://doi.org/10.29240/arcitech.v5i2.15316>
- Rahman, M. D., Arif Djunaidy, & Faizal Mahananto. (2021). Penerapan Weighted Word Embedding pada Pengklasifikasian Teks Berbasis Recurrent Neural Network untuk Layanan Pengaduan Perusahaan Transportasi. *Jurnal Sains Dan Seni ITS*, 10(1), 1–6. <https://doi.org/10.12962/J23373520.V10I1.56145>
- Raihan, A., Hasan, A., Azhim, M. F., & Fadilah, I. (2026). Klasifikasi Kompleksitas Gameplay Berbasis Struktur Kalimat pada Deskripsi Game. *Jurnal Ilmu Komputer Dan Teknologi Informasi (JIKTI)*, 3(1), 11–21. <https://doi.org/10.63447/jikti.v3i1.1824>
- Rihastuti, S., & Rosyidi, A. (2025). Progres Pembangunan IKN Dengan Metode Random Forest. *Journal of Computer Science and Technology*, 5(1), 19–23. <https://doi.org/10.54840/jcstech.v5i1.345>
- Romadhoni, Y., & Holle Hayati, F. K. (2022). Analisis Sentimen Terhadap PERMENDIKBUD No.30 pada Media Sosial Twitter Menggunakan Metode Naive Bayes dan LSTM. *JPIT (Jurnal Informatika: Jurnal Pengembangan IT)*, 7(2), 118–123. <https://doi.org/10.30591/jpit.v7i2.3191>
- Salman Al Markas, M., Anraeni, S., & Budiman Ilmuwan, L. (2025). Implementasi Fitur Vector Bag Of Word Dan TF IDF untuk Analisis Sentiment. *LINIER: Literatur Informatika Dan Komputer*, 2(2), 136–146. <https://doi.org/10.33096/linier.v2i2.3104>
- Saputri, G. A., & Alita, D. (2024). Analisis Sentimen Twitter Terhadap Pemindahan Ibu Kota Negara Menggunakan Support Vector Machine. *Jurnal Informatika: Jurnal Pengembangan IT*, 9(3), 213–223. <https://doi.org/10.30591/jpit.v9i3.6612>
- Sari, S. N., Faisal, M. R., Kartini, D., Budiman, I., Saragih, T. H., & Muliadi. (2023). Perbandingan Ekstraksi Fitur dengan Pembobotan Supervised dan Unsupervised pada Algoritma Random Forest untuk Pemantauan Laporan Penderita Covid-19 di Twitter. *Jurnal Komputasi*, 11(1), 33–42. <https://doi.org/10.23960/komputasi.v11i1.6650>
- Setiana, E., Marwondo, Venia Retreva Danestiara, & Wiyandudin. (2023). Analisis Sentimen Pelaksanaan Kuliah Online Menggunakan Algoritma Support Vector Machine. *Nuansa Informatika*, 17(2), 66–70. <https://doi.org/10.25134/ilkom.v17i2.11>
- Siswa Yoga, A. T., & Pranoto, J. W. (2023). Implementasi Seleksi Fitur Information Gain Ratio Pada Algoritma Random Forest untuk Model Data Klasifikasi Pembayaran Kuliah. *Dinamika Informatika*, 15(1), 41–49. <https://doi.org/10.35315/informatika.v15i1.9465>
- Suci Amaliah, Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, 4(3), 121–127. <https://doi.org/10.35580/variansiunm31>
- Tamba, P. S. E. (2022). Prediksi Penyakit Gagal Jantung Dengan Menggunakan Random Forest. *Jurnal Sistem Informasi Dan Ilmu Komputer Prima (JOSIKOM PRIMA)*, 5(2), 176–181. <https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2445>
- Wajhillah, R., & Wibowo, A. (2021). Information Retrieval Pemetaan Peta Jalan Penelitian Perguruan Tinggi Berbasis Dokumen Publikasi Ilmiah Dosen. *Larik: Jurnal Lapisan Riset Informatika*, 2(2), 49–56. <https://doi.org/10.31294/larik.v2i2.1816>
- Wijaya, H., & Hayati, N. (2025). Natural Language Processing (NLP) Untuk Analisis Sentimen Ulasan Seblak Bandung Pedas Kudus. *Journal of Business and Audit Information System*, 8(1), 13–22. <https://doi.org/10.30813/jbase.v8i1.8035>
- Wijiyanto, W., Pradana, A. I., Sopongi, S., & Atina, V. (2024). Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa. *Jurnal Algoritma*, 21(1), 239–248. <https://doi.org/10.33364/algoritma/v.21-1.1618>