BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547

Analisis Sentimen Masyarakat Terhadap Pembatasan BBM Pertalite Menggunakan Random Forest dan K-Nearest Neighbor

Farhan Muhammad Fadillah*, Yana Cahyana, Rahmat, Ahmad Fauzi

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Buana Perjuangan, Karawang, Indonesia
Email: [1,*]if21.farhanfadillah@mhs.ubpkarawang.ac.id, [2] yana.cahyana@ubpkarawang.ac.id, [3] rahmat@ubpkarawang.ac.id, [4]afauzi@ubpkarawang.ac.id
Email Penulis Korespondensi: if21.farhanfadillah@mhs.ubpkarawang.ac.id

Abstrak−Penelitian ini bertujuan untuk menganalisis opini publik terhadap kebijakan pembatasan penggunaan bahan bakar jenis Pertalite, dengan menganalisis komentar warganet dari platform Instagram. Untuk mengelompokkan opini tersebut, digunakan pendekatan klasifikasi berbasis algoritma K-Nearest Neighbor (KNN) dan Random Forest. Komentar dikategorikan ke dalam tiga bentuk ekspresi sentimen, yaitu positif, negatif, maupun bersifat netral. Adapun tahapan penelitian meliputi pengumpulan data (crawling), pembersihan dan normalisasi data teks, proses pelabelan sentimen, pembobotan menggunakan teknik TF-IDF, pembangunan model klasifikasi, hingga evaluasi kinerja model yang dihasilkan. Data yang digunakan berjumlah 3.081 komentar, dengan 1.000 data dilabeli oleh pakar bahasa sebagai data latih dan sisanya digunakan untuk pengujian. Evaluasi model dilakukan dengan menggunakan dua rasio pembagian data, yaitu 80:20 dan 70:30, untuk menilai stabilitas dan akurasi klasifikasi. Hasil menunjukkan bahwa algoritma Random Forest secara konsisten unggul dibandingkan KNN, dengan akurasi tertinggi sebesar 73% pada skenario 80:20. Distribusi klasifikasi mengindikasikan dominasi sentimen negatif dalam opini publik terhadap kebijakan tersebut. Temuan ini mencerminkan adanya ketidakpuasan masyarakat dan menjadi masukan penting bagi pemerintah dalam mengevaluasi kebijakan distribusi BBM bersubsidi. Penelitian ini juga menunjukkan potensi media sosial sebagai sumber data alternatif dalam menilai persepsi publik secara real-time berbasis data.

Kata Kunci: Confusion Matrix; Crawling; K-Nearest Neighbor; Pra-Pemrosesan; Random Forest

Abstract–This study aims to analyze public opinion regarding the policy of limiting the use of Pertalite fuel by examining user comments on the Instagram platform. To classify these opinions, classification approaches using K-Nearest Neighbor (KNN) and Random Forest algorithms were employed. Comments were categorized into three sentiment expressions: positive, negative, and neutral. The research stages included data collection (crawling), text cleaning and normalization, sentiment labeling, weighting using the TF-IDF technique, model development, and performance evaluation. A total of 2,081 comments were used, with 1,000 comments labeled by language experts as training data, and the remaining used for testing. Model evaluation was conducted using two data splitting ratios, 80:20 and 70:30, to assess classification stability and accuracy. The results indicate that the Random Forest algorithm consistently outperforms KNN, achieving the highest accuracy of 73% under the 80:20 scenario. The classification distribution suggests a dominance of negative sentiment in public opinion toward the policy. These findings reflect public dissatisfaction and serve as critical input for the government in reviewing the subsidized fuel distribution policy. This research also highlights the potential of social media as an alternative data source for real-time public perception analysis.

Keywords: Confusion Matrix; Crawling; K-Nearest Neighbor; Preprocessing; Random Forest

1. PENDAHULUAN

Bahan bakar minyak (BBM) bersubsidi di Indonesia merupakan salah satu aspek vital yang mendukung mobilitas dan kesejahteraan masyarakat. Di antara jenis BBM bersubsidi tersebut, Pertalite menjadi pilihan utama sebagian besar masyarakat karena harganya yang relatif terjangkau. Data dari Indikator Politik Indonesia menyebutkan bahwa sekitar 90,4% masyarakat Indonesia menggunakan Pertalite sebagai bahan bakar utama untuk kendaraan mereka. Situasi ini menunjukkan bahwa bahan bakar bersubsidi tetap menjadi komponen esensial yang dibutuhkan publik guna mendukung aktivitas mereka sehari-hari. Namun, tingginya konsumsi Pertalite juga menyebabkan beban subsidi negara yang cukup besar, dan hal ini mendorong pemerintah untuk merancang kebijakan pembatasan distribusi Pertalite agar lebih tepat sasaran.

Pemerintah mengusulkan kebijakan bahwa hanya kendaraan dengan kapasitas mesin tertentu (di bawah 1400 cc untuk mobil dan di bawah 250 cc untuk motor) yang dapat mengakses BBM bersubsidi. Meskipun bertujuan baik, kebijakan ini menimbulkan *respons* beragam di masyarakat, terutama dari kelompok yang merasa dirugikan karena tidak lagi memperoleh akses terhadap BBM bersubsidi. Beberapa masyarakat menyuarakan kecemasan, ketidakpuasan, dan bahkan kemarahan terhadap kebijakan ini, karena dinilai mengabaikan kompleksitas kebutuhan masyarakat sehari-hari yang tidak selalu selaras dengan parameter teknis seperti kapasitas mesin kendaraan [1]. Pembahasan mengenai topik ini juga marak terjadi di dunia maya, salah satunya di Instagram sebuah platform dengan akun aktif yang sangat besar di Indonesia, mencapai angka 99,4 juta pada bulan juli 2024.

Salah satu situs media sosial dengan tingkat pertumbuhan tercepat adalah Instagram [2]. Instagram sebagai platform berbasis visual dan interaksi langsung melalui komentar menyediakan ruang ekspresi publik yang luas, termasuk terhadap kebijakan pemerintah. Banyak pengguna menyampaikan opini mereka melalui komentar pada unggahan berita seputar pembatasan Pertalite. Kondisi ini menjadikan Instagram sebagai sumber data yang kaya untuk menganalisis persepsi dan emosi masyarakat menggunakan pendekatan analisis sentimen. Analisis sentimen sendiri merupakan proses identifikasi opini, emosi, atau sikap dari suatu pernyataan tertulis yang umumnya terbagi ke dalam tiga kategori utama, yaitu sentimen positif, negatif, dan netral [3]. Metode ini semakin relevan di era digital karena mampu mengubah data tidak terstruktur menjadi informasi bermakna bagi pengambil kebijakan.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547

Penelitian terdahulu telah memanfaatkan berbagai media sosial untuk melakukan analisis sentimen terhadap isu-isu publik. Samantri pada tahun 2024 menggunakan algoritma *Support Vector Machine* (SVM) dan *Random Forest* dalam menganalisis sentimen masyarakat terkait kenaikan harga BBM di Twitter dan memperoleh akurasi 76% untuk *Random Forest* dan 77% untuk SVM [4]. Heltroyce dkk pada tahun 2024 juga menggunakan algoritma *K-Nearest Neighbor* (KNN) dalam menganalisis opini masyarakat terhadap kenaikan harga BBM, dan mencapai akurasi 70,31% dengan data partisi 70:30 [5]. Alvionika dkk pada tahun 2024 menggunakan KNN untuk menganalisis komentar Instagram terhadap layanan By.U dan memperoleh akurasi 73%, namun menyatakan perlunya dataset yang lebih besar untuk meningkatkan akurasi [6]. Rahayu dkk pada tahun 2022 meneliti sentimen pengguna aplikasi finansial FLIP dengan KNN dan TF-IDF serta mendapatkan akurasi 76,68% dan presisi ulasan positif hingga 82,67% [7]. Amelia dkk pada tahun 2024 membandingkan tiga algoritma (*Random Forest*, SVM, dan *Naïve Bayes*) dalam analisis sentimen terhadap aplikasi MyPertamina, dengan *Random Forest* mencatatkan akurasi tertinggi sebesar 99,77% [8]. Larasati dkk pada tahun 2022 menguji performa *Random Forest* pada data ulasan aplikasi Dana dan memperoleh nilai akurasi serta metrik evaluasi lainnya sebesar 84% [9]. Sementara itu, Taufiqqurrahman dkk pada tahun 2023 membandingkan *Naïve Bayes* dan KNN pada ulasan MyPertamina dan menunjukkan bahwa meskipun KNN cenderung fluktuatif, ia masih mampu memberikan hasil yang cukup baik tergantung pada setting dan balancing data [10]. Dari beberapa penelitian tersebut, terlihat bahwa *Random Forest* unggul dalam menangani data besar dan kompleks, sedangkan KNN cocok untuk klasifikasi berbasis kedekatan fitur. Masing-masing algoritma memiliki kekuatan tersendiri, dan pemilihan metode yang tepat sangat bergantung pada karakteristik data dan tujuan analisis.

Meskipun begitu, terdapat kesenjangan penelitian yang cukup signifikan. Hingga saat ini, belum ada penelitian yang secara khusus menganalisis sentimen masyarakat terhadap pembatasan BBM Pertalite dengan memanfaatkan data dari Instagram serta membandingkan secara langsung performa dua algoritma populer, yaitu *Random Forest* dan KNN. Kebanyakan studi terdahulu menggunakan Twitter atau platform lain sebagai sumber data, sementara Instagram justru menyimpan potensi yang besar sebagai sumber opini publik. Selain itu, kombinasi atau perbandingan langsung antara dua algoritma ini dalam konteks kebijakan publik belum banyak dijelajahi.

Oleh karena itu, penelitian ini dirancang dengan memanfaatkan data komentar dari platform Instagram yang berkaitan dengan isu pembatasan bahan bakar jenis Pertalite. Fokus penelitian ini adalah membangun serta mengevaluasi model analisis sentimen dengan menerapkan dua algoritma klasifikasi, yaitu *Random Forest* dan *K-Nearest Neighbor*. Evaluasi terhadap performa model dilakukan melalui pengukuran sejumlah metrik seperti *accuracy*, *precision*, dan *recall*. Proses klasifikasi diawali dengan tahapan *text preprocessing* serta pemberian bobot kata menggunakan pendekatan TF-IDF, yang kemudian diikuti oleh pelatihan model dan pengujian terhadap data yang belum dikenali sebelumnya. Tujuan utama dari penelitian ini adalah mengidentifikasi pendekatan algoritma yang paling efektif dalam mengklasifikasikan opini publik terkait Regulasi pembatasan distribusi BBM jenis Pertalite yang dianalisis melalui data dari media sosial. Penelitian ini diharapkan dapat memberikan tidak hanya kontribusi ilmiah yang signifikan dalam pengembangan ilmu komputer, khususnya pada ranah machine learning dan penambangan teks, tetapi juga dapat menjadi landasan yang berguna bagi pengambil kebijakan untuk memahami persepsi masyarakat secara data-driven dan objektif.

2. METODOLOGI PENELITIAN

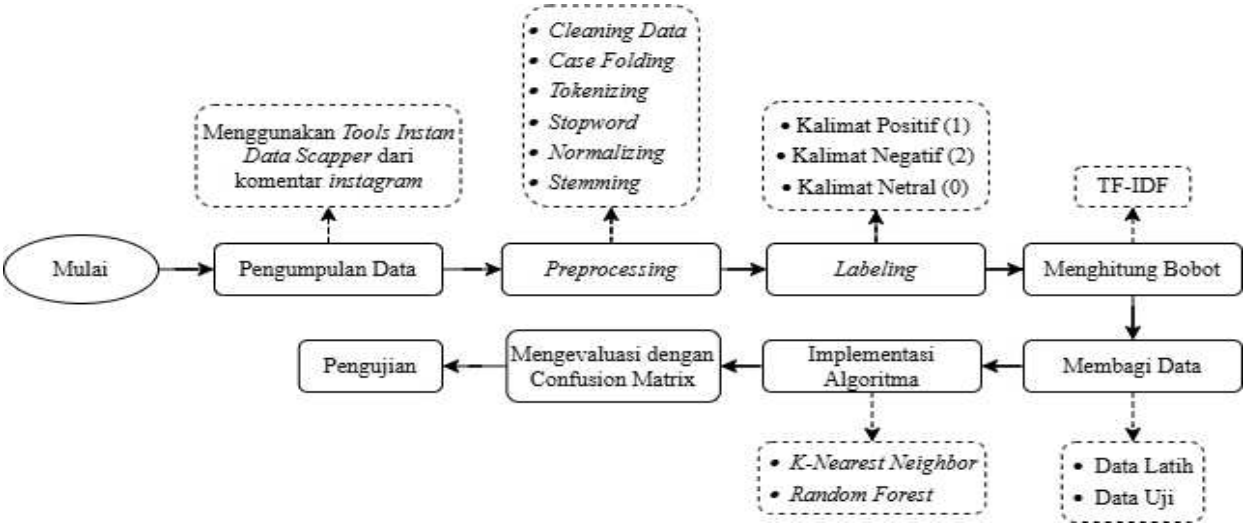
2.1 Tahapan Penelitian

Penelitian ini bertujuan untuk mengevaluasi sentimen publik terhadap kebijakan pembatasan bahan bakar jenis Pertalite dengan memanfaatkan data yang diperoleh dari platform media sosial Instagram. Pendekatan yang digunakan melibatkan algoritma *K-Nearest Neighbor* (KNN) dan *Random Forest* , yang dikenal efisien dalam memproses klasifikasi teks maupun data dengan struktur tidak baku [11]. Subjek yang diteliti difokuskan pada opini masyarakat yang diperoleh dari kolom komentar pengguna Instagram terkait isu pembatasan Pertalite. Rangkaian tahapan penelitian mencakup akuisisi data, proses prapengolahan, pelabelan sentimen, ekstraksi fitur menggunakan metode TF-IDF, pembagian dataset, penerapan algoritma, evaluasi kinerja model, serta pengujian sistem. Setiap langkah dirancang untuk memastikan data telah melalui proses transformasi yang memadai sebelum dianalisis menggunakan metode klasifikasi [12].

Penelitian ini diawali dengan proses pengumpulan data menggunakan *tools* Instan Data Scraper untuk mengambil komentar dari platform Instagram. Setelah data terkumpul, tahap selanjutnya adalah preprocessing yang meliputi beberapa langkah penting, yaitu *cleaning data*, *case folding*, *tokenizing*, *stopword removal*, *normalizing*, dan *stemming*. Tahapan ini bertujuan untuk membersihkan dan menstandarkan data teks agar siap untuk dianalisis. Selanjutnya, data yang telah diproses dilakukan labeling berdasarkan sentimen yaitu kalimat positif (1), kalimat negatif (2), dan kalimat netral (0). Setelah pelabelan dilakukan perhitungan bobot kata menggunakan metode TF-IDF untuk merepresentasikan teks dalam bentuk numerik. Data kemudian dibagi menjadi dua bagian, yaitu data latih dan data uji, untuk keperluan pelatihan dan pengujian model.

Tahap berikutnya adalah implementasi algoritma klasifikasi, yaitu *K-Nearest Neighbor* (KNN) dan *Random Forest*. Setelah model dijalankan, hasil klasifikasi dievaluasi menggunaka *confusion matrix* guna mengukur kinerja model. Proses terakhir adalah pengujian untuk menilai efektivitas dari model klasifikasi yang telah dibangun. Tahapan penelitian ini dapat dilihat pada Gambar 1.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547



Gambar 1. Tahapan Penelitian

2.2 Crawling Data

Proses *crawling* data merujuk pada kegiatan pengambilan atau pengunduhan informasi dari platform basis data, dalam hal ini Instagram [13]. Penelitian ini hanya mengumpulkan komentar publik dari unggahan akun resmi media berita yang bersifat terbuka , tanpa mengakses informasi pribadi pengguna. Untuk menjaga etika dan legalitas, data dikumpulkan dengan batas wajar dan tidak melibatkan scraping masif yang melanggar ketentuan layanan. Pengambilan data dilakukan menggunakan alat bantu *Instant Data Scraper* sebuah ekstensi browser yang memungkinkan ekstraksi data dari halaman web secara langsung. Aktivitas ini bertujuan untuk menghimpun data dari berbagai sumber, termasuk file, database, maupun API, sehingga dapat dimanfaatkan lebih lanjut dalam kegiatan analisis untuk kepentingan riset dan pengembangan[14]. Penelitian ini tidak menggunakan API resmi Instagram karena API tersebut memiliki batasan akses yang tidak mengizinkan ekstraksi komentar publik secara terbuka untuk keperluan riset non-komersial. Seluruh proses dilakukan dengan memperhatikan kode etik penelitian dan kebijakan penggunaan data dari Instagram serta tanpa menyimpan atau menyebarkan identitas pengguna.

2.3 Preprocessing Data

Data yang diperoleh selanjutnya melalui proses preprocessing, yang meliputi tahapan pembersihan (*cleaning*), pengubahan huruf besar-kecil *(case folding*), pemecahan kata (*tokenize*), penghapusan kata umum (*stopword*), normalisasi (*normalize*), dan pengembalian kata ke bentuk dasar (*stemming*) [15]. Tujuan utama preprocessing adalah untuk membersihkan teks dari karakter yang tidak relevan dan menyederhanakan kata menjadi bentuk dasarnya [16]. Proses ini dilakukan menggunakan bahasa pemrograman *python*, dengan dukungan beberapa pustaka (*library*) yaitu:

a. Re untuk membersihkan simbol, tautan, mention, dan karakter non-alfabet
b. Sastrawi untuk melakukan stemming Bahasa Indonesia ke bentuk dasar
c. NLTK (*Natural Language Toolkit*) digunakan untuk *tokenisasi* dan penghapusan *stopword*
d. Pandas dan Numpy untuk manipulasi dan transformasi data tabula
e. *Scikit-learn* digunakan dalam tahapan selanjutnya untuk transformasi TF-IDF dan model klasifikasi

2.4 Labeling Data

Pelabelan dilakukan secara manual oleh ahli bahasa berdasarkan tiga kategori sentimen yaitu negatif, positif, dan netral. Dengan cara masing-masing label pada Instagram yaitu:

a. Apabila kata sentimen kata termasuk kalimat negatif maka bernilai (0)
b. Apabila kata sentimen kata termasuk kalimat positif maka bernilai (1)
c. Apabila kata sentimen kata termasuk kalimat netral maka bernilai (2)

2.5 Pembobotan TF-IDF

Ekstraksi dan pembobotan fitur merupakan proses untuk menentukan fitur yang diperoleh dari kata-kata dalam setiap kalimat dan memberikan bobot pada setiap kata. Salah satu metode ekstraksi dan pembobotan fitur adalah TF-IDF (*Term Frequency – Inverse Document Frequency*)[17]. Melalui metode ini, setiap kata atau fitur dalam dokumen diberi bobot berdasarkan frekuensinya dalam dokumen tertentu dan keberadaannya di seluruh korpus. Nilai bobot diperoleh dari hasil perkalian antara nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Semakin sering suatu kata muncul dalam satu dokumen namun jarang muncul dalam dokumen lain, maka bobotnya akan semakin tinggi [18].
Persamaan (1) dan (2) menunjukkan metode TF-IDF :

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547

$$IDF(w) = log\ (\frac{N}{DF(w)}) \quad (1)$$

$$W_{ij} = TF_{ij} x log(^{D}/_{DF_j}) \quad (2)$$

Langkah pembobotan dimulai dengan menghitung frekuensi kemunculan setiap kata (TF) dalam teks yang telah melalui proses tokenisasi. Setiap kata yang muncul diberikan nilai frekuensi sebesar 1. Tahap berikutnya menghitung jumlah dokumen yang mengandung kata tersebut untuk memperoleh nilai *Document Frequency* (DF). Setelah itu, dilakukan penghitungan nilai *Inverse Document Frequency* (IDF), sesuai formula yang tertera dalam Persamaan (1). Akhirnya, nilai akhir bobot dokumen (BDF) ditentukan berdasarkan kombinasi dari hasil perhitungan sebelumnya [19].

2.6 Split Dataset

Dataset dibagi menjadi dua subset, yaitu data pelatihan dan data pengujian. Data pelatihan digunakan untuk membangun model dengan mempelajari pola-pola dari data yang tersedia, sementara data pengujian dimanfaatkan untuk mengukur performa model terhadap data yang belum pernah dikenali sebelumnya. Penelitian ini menerapkan dua rasio pembagian, yakni 80:20 dan 70:30.

2.7 Implementasi Algoritma

Model machine learning dikembangkan menggunakan tiga algoritma berikut:

2.7.1 K-Nearest Neighbor

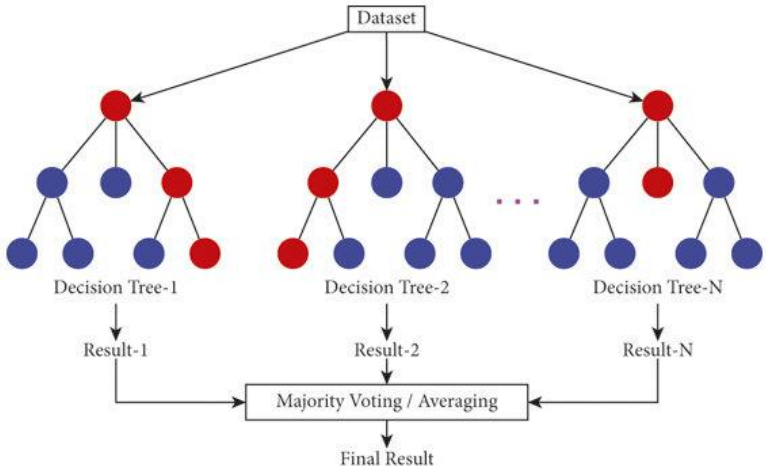
K-Nearest Neighbor (KNN) merupakan metode supervised learning yang digunakan untuk melakukan klasifikasi data berdasarkan kedekatan jarak dengan data yang telah diberi label sebelumnya. Teknik ini bekerja dengan prinsip *Euclidean Distance*, di mana setiap data yang akan diklasifikasikan dibandingkan dengan sejumlah tetangga terdekat sebanyak k [6]. Konsep utama dari KNN adalah bahwa suatu sampel akan dikategorikan berdasarkan kemiripannya dengan sampel lain yang berada di sekitarnya [20]. Dalam penelitian ini digunakan nilai **k = 10**, yaitu sepuluh tetangga terdekat dijadikan acuan dalam proses klasifikasi.
Analisis metode KNN dapat menggunakan rumus *Euclidean Distance* sebagai persamaan (3) :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_1 - y_1)^2} \quad (3)$$

Dalam algoritma *K-Nearest Neighbor* (KNN) notasi d(x, y) digunakan untuk menyatakan jarak antara dua titik data, yaitu antara data latih dan data uji. Simbol xi merepresentasikan sampel data training, sedangkan yi menunjukkan sampel data testing yang akan diklasifikasikan. Adapun n menyatakan jumlah fitur atau atribut yang digunakan dalam perhitungan jarak tersebut. Jarak ini menjadi dasar dalam menentukan kedekatan antar data dan pengambilan keputusan klasifikasi berdasarkan sejumlah tetangga terdekat.

2.7.2 *Random Forest*

Random forest adalah algoritma yang digunakan untuk mengklasifikasikan data dalam jumlah besar. Proses klasifikasinya dilakukan dengan mengkombinasikan beberapa pohon keputusan (tree) dan melatih model menggunakan sampel data yang tersedia [21]. Pada penelitian ini digunakan *random forest classifier* untuk memprediksi teks masuk ke kelas positif, negatif, atau netra [12]. Untuk menghasilkan performa klasifikasi yang optimal parameter yang digunakan pada penelitian ini jumlah pohon keputusan (*n_estimators*) = 100 dan kedalaman maksimum pohon (*max_depth*) = *none*.
Berikut merupakan ilustrasi dari pohon keputusan yang dapat dilihat pada Gambar 2.



Gambar 2. Cara kerja algoritma *random forest*

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547

Pada gambar 2 menggambarkan arsitektur algoritma Random Forest, yang merupakan metode ensemble learning berbasis pohon keputusan (*decision tree*). Proses dimulai dengan satu dataset yang kemudian digunakan untuk membentuk beberapa pohon keputusan yang berbeda (*Decision Tree*-1 hingga *Decision Tree*-N). Setiap pohon keputusan dilatih menggunakan subset acak dari data dan fitur, menghasilkan prediksi individual berupa *Result*-1 hingga *Result*-N. Hasil dari semua pohon tersebut kemudian digabungkan menggunakan metode *majority voting* (untuk klasifikasi) atau *averaging* (untuk regresi). Proses ini menghasilkan prediksi akhir yang lebih akurat dan stabil dibandingkan dengan penggunaan satu pohon keputusan saja. Dengan memanfaatkan banyak pohon dan menggabungkan hasilnya, *Random Forest* mampu mengurangi *overfitting* dan meningkatkan kinerja model secara keseluruhan.

2.8 Evaluasi Model

Evaluasi dilakukan menggunakan *Confusion Matrix* yang mencakup nilai *accuracy*, *precision*, dan *recall*. Nilai-nilai ini dihitung untuk mengukur kinerja model dalam klasifikasi sentimen [4]. Ada beberapa nilai kategori dalam *confusion matrix*:

a. *True* Positif (TP) : Prediksi positif dan nilai aktualnya positif.
b. *True* Negatif (TN) : Prediksi negatif dan nilai aktualnya negatif.
c. *False* Positif (FP) : Prediksi positif dan nilai aktualnya negatif.
d. *False* Negatif (FN) : Prediksi negatif dan nilai aktualnya positif.

Asumsi dalam *Confusion Matriks* diperlihatkan dalam Tabel 1 :

Tabel 1. Asumsi Confusin Matriks

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Pada Tabel 1 jika hasil prediksi dan nilai aktual keduanya benar, maka situasi tersebut dikategorikan sebagai *True Positive* (TP). Sebaliknya, apabila prediksi keliru meskipun nilai aktual benar, atau prediksi tepat namun nilai aktual salah, maka disebut sebagai *False Negative* (FN). Evaluasi performa model dilakukan melalui sejumlah metrik utama seperti *precision*, *recall*, dan *accuracy*. Rumus perhitungan masing-masing metrik ditampilkan pada gambar di bawah, sedangkan formulasi matematisnya disajikan dalam Persamaan (4), (5), dan (6).

Tabel 2. Persamaan menghitung nilai

Matrik Performa	Rumus	
Akurasi	$\frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\%$	(4)
Recall	$\frac{TP}{TP+FN}\ x\ 100\%$	(5)
Presisi	$\frac{TP}{TP+FP}\ x\ 100\%$	(6)

Pada tabel 2 matriks performa merupakan alat evaluasi yang digunakan untuk mengukur seberapa baik suatu model klasifikasi bekerja. Beberapa metrik yang sering digunakan dalam evaluasi model adalah akurasi, recall, dan presisi. Akurasi menggambarkan sejauh mana prediksi model sesuai dengan nilai sebenarnya secara keseluruhan. Recall menunjukkan kemampuan model dalam mendeteksi atau mengenali seluruh data yang tergolong dalam kelas positif. Sementara itu, presisi mengukur tingkat ketepatan model dalam mengklasifikasikan data ke dalam kelas positif, yaitu seberapa banyak dari prediksi positif yang benar-benar relevan. Dengan menggunakan ketiga metrik ini, performa model dapat dinilai secara lebih menyeluruh, terutama saat menghadapi data yang tidak seimbang.

2.9 Pengujian

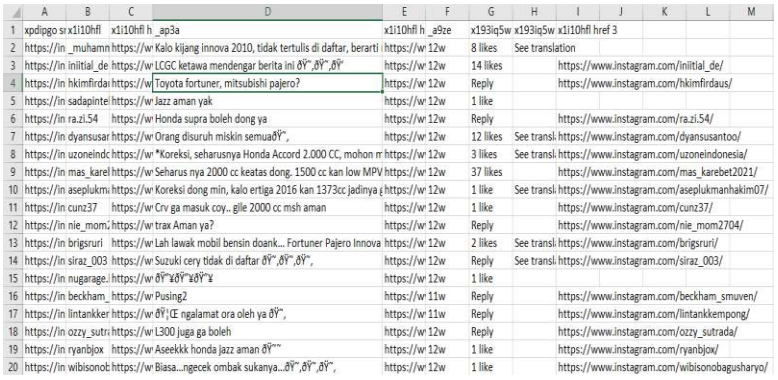
Model yang telah dibangun selanjutnya diterapkan pada data baru yang belum pernah digunakan dalam tahap pelatihan. Tujuan dari pengujian ini adalah untuk menilai sejauh mana tingkat akurasi dan reliabilitas model dalam menghasilkan output yang relevan terhadap teks input yang diberikan. Pengujian dilakukan menggunakan algoritma dengan capaian akurasi tertinggi berdasarkan hasil evaluasi pada tahap pelatihan sebelumnya.

3. HASIL DAN PEMBAHASAN

3.1 Crawling Data

Hasil pengumpulan data komentar pada instagram dengan fokus pada unggahan yang berkaitan dengan isu pembatasan BBM pertalite. Data yang dikumpulkan selama periode Juni 2024 hingga November 2024.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	xpdipgo sr	x1i10hfl	x1i10hfl h	_ap3a	x1i10hfl h	_a9ze	x193iq5w	x193iq5w	x1i10hfl href 3				
2	https://in	_muhamn	https://w	Kalo kijang innova 2010, tidak tertulis di daftar, berarti	https://w	12w	8 likes	See translation					
3	https://in	initial_de	https://w	LCGC ketawa mendengar berita ini ðŸ˜‚ðŸ˜‚ðŸ˜	https://w	12w	14 likes		https://www.instagram.com/initial_de/				
4	https://in	hkimfirdai	https://w	Toyota fortuner, mitsubishi pajero?	https://w	12w	Reply		https://www.instagram.com/hkimfirdaus/				
5	https://in	sadapinte	https://w	Jazz aman yak	https://w	12w	1 like						
6	https://in	ra.zi.54	https://w	Honda supra boleh dong ya	https://w	12w	Reply		https://www.instagram.com/ra.zi.54/				
7	https://in	dyansusar	https://w	Orang disuruh miskin semuaðŸ˜	https://w	12w	12 likes	See transl	https://www.instagram.com/dyansusantoo/				
8	https://in	uzoneindc	https://w	*Koreksi, seharusnya Honda Accord 2.000 CC, mohon m	https://w	12w	3 likes	See transl	https://www.instagram.com/uzoneindonesia/				
9	https://in	mas_karel	https://w	Seharus nya 2000 cc keatas dong. 1500 cc kan low MPV	https://w	12w	37 likes		https://www.instagram.com/mas_karebet2021/				
10	https://in	aseplukm	https://w	Koreksi dong min, kalo ertiga 2016 kan 1373cc jadinya g	https://w	12w	1 like	See transl	https://www.instagram.com/aseplukmanhakim07/				
11	https://in	cunz37	https://w	Crv ga masuk coy.. gile 2000 cc msh aman	https://w	12w	1 like		https://www.instagram.com/cunz37/				
12	https://in	nie_mom2	https://w	trax Aman ya?	https://w	12w	Reply		https://www.instagram.com/nie_mom2704/				
13	https://in	brigsruri	https://w	Lah lawak mobil bensin doank... Fortuner Pajero Innova	https://w	12w	2 likes	See transl	https://www.instagram.com/brigsruri/				
14	https://in	siraz_003	https://w	Suzuki cery tidak di daftar ðŸ˜‚ðŸ˜‚ðŸ˜	https://w	12w	Reply	See transl	https://www.instagram.com/siraz_003/				
15	https://in	nugarage.	https://w	ðŸ˜ŽðŸ˜ŽðŸ˜	https://w	12w	1 like						
16	https://in	beckham_	https://w	Pusing2	https://w	11w	Reply		https://www.instagram.com/beckham_smuven/				
17	https://in	lintankker	https://w	ðŸ¦Œ ngalamat ora oleh ya ðŸ˜	https://w	11w	Reply		https://www.instagram.com/lintankkempong/				
18	https://in	ozzy_sutr	https://w	L300 juga ga boleh	https://w	12w	Reply		https://www.instagram.com/ozzy_sutrada/				
19	https://in	ryanbjox	https://w	Aseekkk honda jazz aman ðŸ˜	https://w	12w	1 like		https://www.instagram.com/ryanbjox/				
20	https://in	wibisonob	https://w	Biasa...ngecek ombak sukanya...ðŸ˜‚ðŸ˜‚ðŸ˜	https://w	12w	1 like		https://www.instagram.com/wibisonobagusharyo/				

Gambar 3. Hasil crawling data

Pada gambar 3 Sebanyak 3.081 entri data dikumpulkan, yang dalam kondisi awalnya masih belum terstruktur dan mengandung berbagai elemen nonteks, seperti simbol, angka, serta kata-kata yang kurang relevan. Oleh karena itu, dilakukan tahap pra-pemrosesan guna menyaring dan menyiapkan data untuk proses analisis lanjutan.

3.1 *Selection* Data

Data yang dikumpulkan disimpan dalam file berformat .csv dan berisi beberapa kolom. Namun, untuk keperluan analisis, hanya kolom berisi komenan.

```
Index(['x1i10hfl href', 'xpdipgo src', 'x1i10hfl', 'x1i10hfl href 2', '_ap3a',
       'x1i10hfl href 3', '_a9ze', 'x193iq5w', 'x193iq5w 3'],
      dtype='object')
```

Gambar 4. Sebelum dilakukan selection

Pada gambar 4 terdapat kolom-kolom yang ditampilkan masih menggunakan nama-nama yang bersifat acak dan tidak deskriptif seperti x1i10hfl href, xpdipgo src, x1i10hfl, x1i10hfl href 2, _ap3a, x1i10hfl href 3, _a9ze, x193iq5w, dan x193iq5w 3 Untuk memudahkan analisis data hanya kolom pada _ap3a yang akan digunakan sebelumnya akan dilakukan proses *rename* agar memudahkan menjadi full_text.

3.2 *Preprocessing* Data

Preprocessing data memiliki peran krusial dalam membersihkan data dari elemen-elemen yang tidak dibutuhkan agar analisis sentimen dapat dilakukan secara lebih tepat dan efisien. Dari total 3.081 komentar yang dikumpulkan melalui proses *crawling* hanya 3.026 data yang lolos hingga tahap akhir *preprocessing*. Penyusutan sebanyak 55 diakibatkan data tersebut mengandung data dengan komentar kosong atau hanya terdapat simbol dan emoji, komentar yang duplikat dan komentar yang sangat pendek atau tidak informatif yang disebut proses *cleaning* seperti tabel 3.

Tabel 3. Data *Cleaning*

Data Sebelum *Cleaning*	Data Sesudah *Cleaning*
ya iyalah disel mana bisa diisi pertalite 😭😭😭😭😭	iya disel mana isi pertalite
Yakali motor 250 cc masih antri pertalite 😂	yakali motor cc antri pertalite
Innova diesel dari dulu dilarang isi pertalite woiii 😩	innova diesel dulu larang isi pertalite woiii
Apa itu pertalite?	apa pertalite

Setelah proses *cleaning* selesai, dilakukan *case folding* untuk menyamakan format huruf menjadi huruf kecil semua. Kemudian diterapkan normalisasi untuk mengonversi kata-kata tidak baku atau singkatan ke bentuk aslinya seperti “bgt” menjadi “banget”. Selanjutnya dilakukan penghapusan *stopword*, seperti “yang”, “dan”, dan kata umum lainnya. Tahapan selanjutnya adalah *tokenisasi*, yaitu proses memisahkan kalimat menjadi unit kata secara individual. Setelah itu, dilakukan proses stemming yang bertujuan untuk mengubah kata ke bentuk dasarnya, seperti kata 'menulis' yang diubah menjadi 'tulis', guna mengurangi variasi kata dan meningkatkan akurasi analisis.

3.3 *Labeling* Data

Proses pelabelan data dilakukan oleh pakar bahasa guna memastikan akurasi pemberian label pada setiap entri, dengan tujuan untuk meningkatkan kinerja dan presisi model klasifikasi. Data diklasifikasikan ke dalam tiga jenis sentimen, yaitu positif, netral, dan negatif. Dari keseluruhan dataset yang tersedia, sebanyak 1.000 entri telah dilabeli oleh ahli, sementara sejumlah 1.000 digunakan dalam tahap pengujian model.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547

Tabel 4. Hasil *Labeling* Data

Sentimen	Jumlah
Positif	352
Netral	74
Negatif	574
Total	1000

Pada Tabel 4 menunjukkan bahwa data pelatihan bersifat tidak seimbang (*imbalanced data*) dengan dominasi signifikan pada kelas negatif 574 data, diikuti oleh kelas positif 352 data, dan jumlah yang sangat kecil pada kelas netral 74 data. Ketidakseimbangan ini merupakan salah satu keterbatasan dalam penelitian, karena dapat memengaruhi performa model klasifikasi terutama dalam mengenali kelas yang jumlah datanya lebih sedikit seperti netral dan positif. Kondisi ini mengindikasikan adanya kecenderungan bias dari model terhadap kelas mayoritas. permasalahan ini dapat diatasi dengan menerapkan teknik penyeimbangan data seperti *oversampling* (misalnya dengan metode *SMOTE*) atau *undersampling*, serta pendekatan lain seperti algoritma yang bersifat *cost-sensitive*. Namun, karena pendekatan penyeimbangan ini tidak diterapkan dalam penelitian ini, maka ketidakseimbangan data secara eksplisit menjadi keterbatasan yang perlu dicatat. Peneliti selanjutnya disarankan untuk mengatasi isu ketidakseimbangan ini agar performa klasifikasi dapat merata di semua kelas sentimen.

3.4 *Tranformation* Data

Setelah tahap pelabelan data diselesaikan, data kemudian diproses menggunakan metode TF-IDF untuk menghitung bobot masing-masing kata. Teknik ini menggunakan fungsi *TfidfVectorizer* dari pustaka *scikit-learn* guna menghasilkan daftar kosakata yang terurut secara alfabetis. Proses pembobotan ini didasarkan pada seberapa sering sebuah kata muncul dalam dokumen, yang merepresentasikan tingkat kepentingan atau relevansi kata tersebut terhadap konten dokumen.

3.5 Klasifikasi Data

Ada dua algoritma yang digunakan yaitu *Random Forest* dan *K-Nearest Neighbor*. Dataset sebanyak 1.000 entri dibagi dalam rasio 80:20 dan 70:30 :

Tabel 5. *Split* Data

Persentase Data		Jumlah Data	
Latih	Uji	Latih	Uji
80 %	20 %	800	200
70 %	30 %	700	300

Tabel 5 menggambarkan proporsi data dan total entri yang dimanfaatkan dalam proses pelatihan model. Data dipisahkan ke dalam dua skema distribusi, yaitu 80:20 dan 70:30, yang dimaksudkan untuk mengamati kestabilan dan konsistensi performa model terhadap variasi proporsi data pelatihan dan pengujian. Selain itu, pendekatan ini berguna dalam mengevaluasi perbandingan kinerja antara *Random Forest* dan *K-Nearest Neighbor* ketika dihadapkan pada kondisi data yang tidak seragam.

3.6 Evaluasi Data

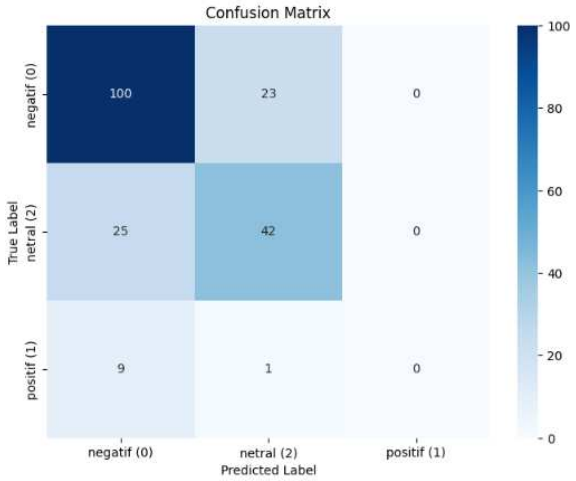
Setelah model berhasil dibangun, hasil evaluasi dari ketiga algoritma yang digunakan dalam skenario pembagian data 80:20 dan 70:30 sebagai berikut :

Tabel 6. Hasil Evaluasi Model Rasio 80:20

Model	*Accuracy*	*Precision*	*Recall*	F1-Score
K-Nearest Neighbor	0.71	0.67	0.71	0.69
Random Forest	0.73	0.69	0.73	0.70

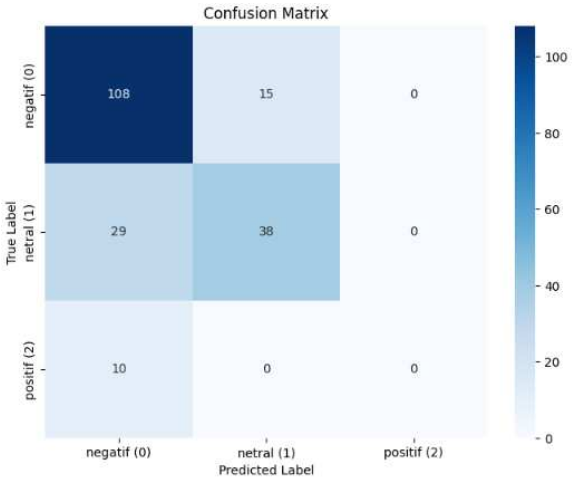
Pada Tabel 6 evaluasi kinerja model yang dilakukan pada skema pembagian data latih dan uji sebesar 80:20 menunjukkan bahwa Random Forest secara konsisten memberikan hasil klasifikasi yang lebih baik dibandingkan K-Nearest Neighbor (K-NN). Hal ini terlihat dari capaian nilai akurasi Random Forest sebesar 0,73, yang melampaui akurasi K-NN sebesar 0,71. Selain itu, Random Forest juga memperoleh nilai precision 0,69, recall 0,73, serta F1-score 0,70. Sebaliknya, K-NN menghasilkan precision sebesar 0,67, F1-score sebesar 0,69 dan recall sebesar 0,71. Dengan demikian, Random Forest dinilai lebih handal dalam mendeteksi pola-pola sentimen dan menghasilkan klasifikasi yang lebih stabil dibandingkan K-NN.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547



Gambar 5. *K-Nearest Neighbor*

Berdasarkan Gambar 5 merujuk pada visualisasi confusion matrix, terlihat bahwa model memiliki kinerja paling optimal dalam mengidentifikasi kelas negatif (label 0), dengan 100 data berhasil diprediksi dengan tepat dan hanya 23 yang salah dikategorikan sebagai netral. Sebaliknya, pada kelas netral (label 2), model menunjukkan tingkat kesulitan yang cukup tinggi, di mana hanya 42 dari total 67 data berhasil dikenali dengan benar, sementara 25 lainnya keliru terklasifikasi sebagai negatif. Adapun untuk kelas positif (label 1), model sepenuhnya gagal melakukan prediksi yang tepat karena seluruh 9 data dialokasikan ke dalam kelas negatif dan 1 data kedalam kelas netral. Temuan ini mengindikasikan adanya kecenderungan bias model terhadap kelas negatif serta ketidakmampuannya dalam mengenali kategori netral dan positif secara akurat.



Gambar 6. *Random Forest*

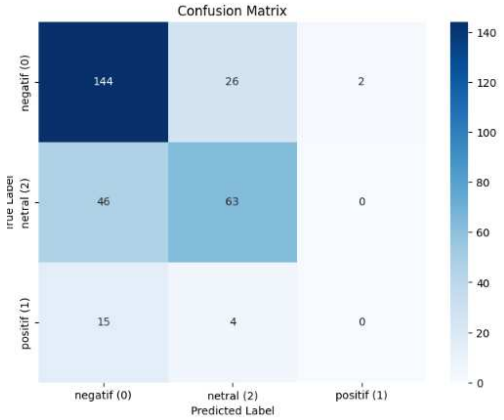
Pada Gambar 6 hasil confusion matrix mengindikasikan bahwa performa tertinggi dicapai pada kategori negatif (label 0), dengan jumlah prediksi tepat mencapai 108 data dan hanya 15 entri yang salah terbaca sebagai netral. Untuk kategori netral (label 1), model mampu mengenali 38 data secara tepat, walau 29 data masih salah teridentifikasi sebagai negatif, dan tidak ada yang tergolong dalam kelas positif. Sementara itu, model sepenuhnya gagal mengidentifikasi kelas positif (label 2), karena seluruh 10 data pada kategori ini justru terbaca sebagai negatif.

Tabel 7. Hasil Evaluasi Model Rasio 70:30

Model	*Accuracy*	*Precision*	*Recall*	F1-Score
K-Nearest Neighbor	0.69	0.65	0.69	0.66
Random Forest	0.69	0.65	0.69	0.66

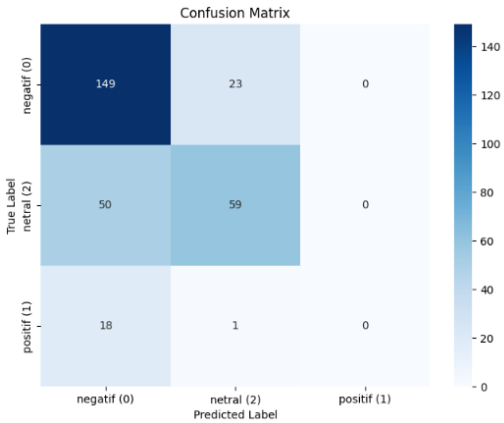
Berdasarkan Tabel 7 hasil evaluasi performa model dengan rasio 70:30, terlihat bahwa algoritma *Random Forest* memiliki nilai yang sama dengan *K-Nearest Neighbor* dalam seluruh metrik evaluasi. *Random Fores*t dan *K-Nearest Neighbor* mencatatkan nilai *accuracy* sebesar 0,69. Jika dilihat dari indikator evaluasi seperti tingkat F1-score, *precision*, maupun *recall*, algoritma Random Forest dan *K-Nearest Neighbor* memperoleh nilai masing-masing sebesar 0,66; 0,65 dan 0,69..

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547



Gambar 7. *K-Nearest Neighbor*

Berdasarkan Gambar 7 merujuk pada visualisasi confusion matrix, model menunjukkan tingkat akurasi tertinggi dalam mendeteksi sentimen negatif (label 0), dengan 144 data berhasil diklasifikasikan secara benar dan hanya 28 prediksi yang meleset (26 kasus diprediksi sebagai netral dan 2 sebagai positif). Sementara itu, pada kategori netral (label 2), model berhasil mengenali 63 entri dengan tepat, meskipun 46 lainnya keliru dipetakan ke kelas negatif. Adapun kelas positif (label 1) menunjukkan hasil terendah, karena tidak ada satu pun dari total 19 data yang berhasil diklasifikasikan secara benar (15 dialokasikan ke kategori negatif dan 4 sisanya ke netral).



Gambar 8. *Random Forest*

Berdasarkan Gambar 8 *confusion matrix* di atas, model menunjukkan performa terbaik dalam mengklasifikasikan kelas negatif (label 0), dengan 149 prediksi yang benar dan 23 prediksi yang keliru diklasifikasikan sebagai netral. Untuk kelas netral (label 2), model mampu mengklasifikasikan dengan benar sebanyak 59 data, namun 50 data lainnya salah diprediksi sebagai negatif. Sementara itu, kelas positif (label 1) tidak dapat dikenali sama sekali oleh model, karena seluruh 18 data positif diprediksi sebagai negatif dan 1 data sebagai netral.

Hasil evaluasi model menunjukkan performa yang tidak merata terhadap ketiga kelas sentimen. Salah satu temuan paling signifikan adalah bahwa model sepenuhnya gagal mengidentifikasi kelas sentimen positif dalam semua skenario pengujian. Baik algoritma Random Forest maupun K-Nearest Neighbor tidak berhasil mengklasifikasikan satu pun data positif dengan benar. Semua entri positif justru terklasifikasi sebagai negatif atau netral. Kegagalan ini kemungkinan besar disebabkan oleh ketidakseimbangan distribusi data latih. Kondisi ini membuat model lebih banyak belajar dari pola sentimen negatif sehingga memiliki kecenderungan bias terhadap kelas tersebut. Selain itu, representasi fitur hasil TF-IDF yang digunakan dalam penelitian ini kemungkinan belum cukup mampu membedakan karakteristik sentimen positif secara efektif. Kelas netral juga mengalami kesulitan serupa. Banyak data netral salah diklasifikasikan sebagai negatif, menghasilkan nilai *False Positive* (FP) dan *False Negative* (FN) yang tinggi. Kesalahan ini berdampak langsung terhadap nilai *precision* dan *recall*, serta mengurangi keandalan model dalam mendeteksi opini netral masyarakat. Berikut adalah perbandingan performa model berdasarkan confusion matrix dari kedua algoritma dan dua rasio pembagian data:

a. Rasio 80:20

Tabel 8. Perbandingan Nilai Kelas Sentimen Rasio 80:20

Metode	Kelas	TP	TN	FP	FN
KNN	Negatif	100	43	34	23
	Netral	42	109	24	25
	Positif	0	190	0	10

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547

	Negatif	108	38	39	15
Random Forest	Netral	38	118	15	29
	Positif	0	190	0	10

b. Rasio 70:30

Tabel 9. Perbandingan Nilai Kelas Sentimen Rasio 70:30

Metode	Kelas	TP	TN	FP	FN
	Negatif	144	63	61	28
KNN	Netral	63	161	30	46
	Positif	0	279	2	19
	Negatif	149	60	68	23
Random Forest	Netral	59	167	24	50
	Positif	0	281	0	19

Dari Tabel 8 dan Tabel 9 secara umum *Random Forest* dengan rasio data 80:20 merupakan kombinasi terbaik dalam penelitian ini. Kombinasi ini memberikan keseimbangan antara *akurasi*, *presisi*, dan *recall* yang lebih konsisten dibandingkan metode lainnya. KNN memiliki potensi dalam mengenali kelas netral namun rentan terhadap fluktuasi performa. Rasio 70:30 memberikan keuntungan dari sisi jumlah data latih tetapi tidak memberikan dampak signifikan pada peningkatan kualitas klasifikasi terutama pada kelas minoritas. Temuan ini menggaris bawahi pentingnya tidak hanya memilih algoritma yang tepat, tetapi juga memperhatikan karakteristik distribusi data dan proporsi pembagian data yang sesuai, agar model dapat mempelajari pola sentimen secara efektif dan adil.

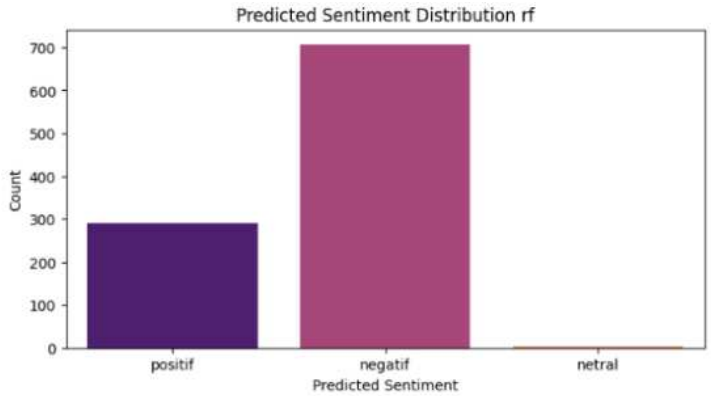
3.7 Pengujian

Setelah selesai membuat model, Pada tahap ini, pengujian dilakukan menggunakan data yang belum dilengkapi dengan label, dataset akan diberikan label menggunakan model model yang sudah di buat.

	full_text	predicted_sentiment_rf	predicted_sentiment_knn
0	nyaman pakek shell vpower curang	positif	positif
1	pickup larang isi minyak murah gmn gitu yaa	negatif	negatif
2	lcgc masukin kn bikin murah biar beli bbm okta...	negatif	negatif
3	tuju rata mobil motor kelas premium kudu perta...	positif	negatif
4	yg mana niih suara kmren ngebully anies dukung...	negatif	negatif

Gambar 9. Hasil pelabelan menggunakan model rasio 80:20

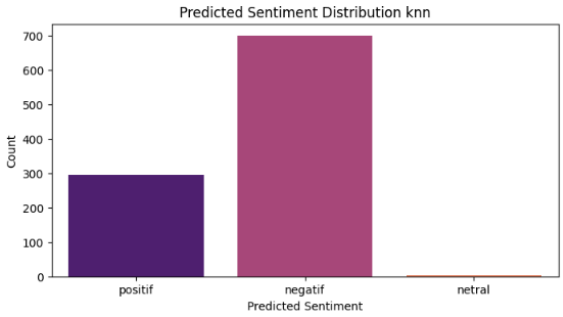
Gambar 9 model yang dikembangkan dengan rasio 80:20 menunjukkan performa yang solid dalam mengklasifikasikan sentimen pada data baru yang memiliki kesamaan topik dengan data pelatihan. Konsistensi dalam proses pelabelan menjadi indikator bahwa model mampu mengidentifikasi pola sentimen pada data yang tidak dikenal dalam pelatihan sebelumnya. Kondisi ini sekaligus memperlihatkan bahwa model mampu beradaptasi dan menerapkan pembelajaran pada data baru dengan karakteristik yang sejenis.



Gambar 10. Hasil pelabelan menggunakan model *random forest*

Gambar 10 menampilkan hasil pelabelan data menggunakan algoritma *random forest*. Grafik menunjukan bahwa sebagian besar termasuk ke dalam sentimen negatif, jumlah yang sangat dominan dibadingkan dengan kategori netral, dan positif.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 340-352
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.547



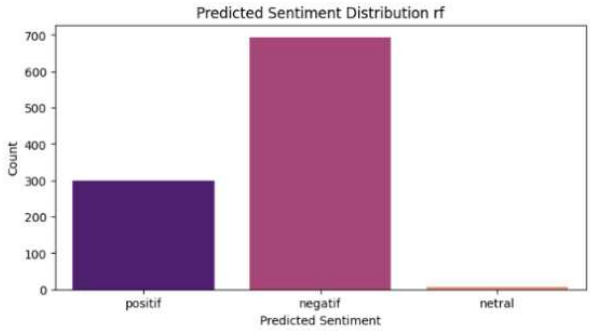
Gambar 11. Hasil pelabelan menggunakan model knn

Gambar 11 menampilkan hasil pelabelan data menggunakan algoritma *k-nearest neighbor*. Grafik menunjukan bahwa sebagian besar termasuk ke dalam sentimen negatif, jumlah yang sangat dominan dibadingkan dengan kategori netral, dan positif.

	full_text	predicted_sentiment_rf	predicted_sentiment_knn
0	nyaman pakek shell vpower curang	positif	positif
1	pickup larang isi minyak murah gmn gitu yaa	negatif	negatif
2	lcgc masukin kn bikin murah biar beli bbm okta...	negatif	negatif
3	tuju rata mobil motor kelas premium kudu perta...	positif	positif
4	yg mana niih suara kmren ngebully anies dukung...	negatif	negatif

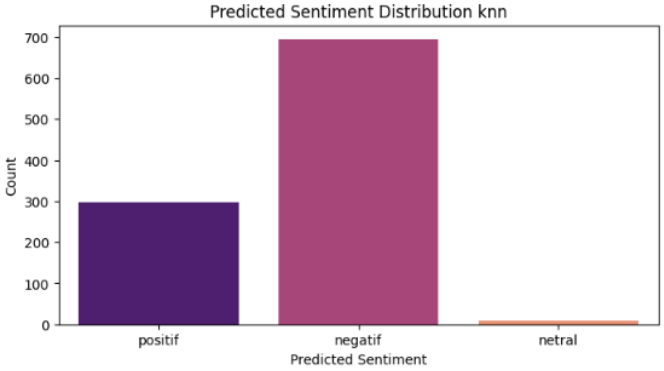
Gambar 12. Hasil pelabelan menggunakan model rasio 70:30

Gambar 12 model yang dikembangkan dengan rasio 70:30 menunjukkan performa yang solid dalam mengklasifikasikan sentimen pada data baru yang memiliki kesamaan topik dengan data pelatihan. Konsistensi dalam proses pelabelan menjadi indikator bahwa model mampu mengidentifikasi pola sentimen pada data yang tidak dikenal dalam pelatihan sebelumnya. Kondisi ini sekaligus memperlihatkan bahwa model mampu beradaptasi dan menerapkan pembelajaran pada data baru dengan karakteristik yang sejenis.



Gambar 13. Hasil pelabelan menggunakan model *random forest*

Gambar 13 menampilkan hasil pelabelan data menggunakan algoritma *random forest*. Grafik menunjukan bahwa sebagian besar termasuk ke dalam sentimen negatif, jumlah yang sangat dominan dibadingkan dengan kategori netral, dan positif.



Gambar 14. Hasil pelabelan menggunakan model knn

Gambar 14 menampilkan hasil pelabelan data menggunakan algoritma *k-nearest neighbor*. Grafik menunjukan bahwa sebagian besar termasuk ke dalam sentimen negatif, jumlah yang sangat dominan dibadingkan dengan kategori netral, dan positif.

3.8 Pembahasan

Proses crawling data menghasilkan sebanyak 3.081 komentar dari media sosial Instagram yang berkaitan dengan isu pembatasan BBM Pertalite. Setelah melalui sejumlah proses prapengolahan data, seperti pembersihan teks, penyeragaman huruf, normalisasi kata, pemecahan kalimat menjadi token, penghapusan kata-kata umum, serta reduksi kata ke bentuk dasar (stemming), total data yang dapat digunakan dianalisis tersisa sebanyak 3.062 data. Dari total data tersebut, sebanyak 1.000 data diberi label sentimen secara manual oleh pakar linguistik, sementara 1.000 digunakan untuk pengujian model.

Pada tahap klasifikasi, studi ini melakukan perbandingan antara dua metode machine learning, yakni K-Nearest Neighbor (KNN) dan Random Forest. Evaluasi terhadap kinerja dan kestabilan masing-masing model dilakukan dengan menerapkan dua skema pembagian data yang berbeda, yaitu 80:20 dan 70:30. Skema 80:20 menunjukkan bahwa sebanyak 800 data digunakan untuk proses pelatihan dan 200 sisanya untuk pengujian, sementara skema 70:30 menggunakan 700 data untuk pelatihan dan 300 data untuk pengujian. Tujuan dari penggunaan dua skenario ini adalah untuk menilai kestabilan dan generalisasi model dalam menghadapi variasi proporsi data latih dan uji. Evaluasi hasil dapat ditemukan pada tabel.

Tabel 10. Evaluasi Model

Rasio	Model	Akurasi
80:20	*K-Nearest Neighbor*	0.71
	Random Forest	0.73
70:30	*K-Nearest Neighbor*	0.69
	Random Forest	0.69

Pada pengujian dengan dua rasio data, Random Forest secara konsisten menunjukkan performa terbaik dibandingkan *K-Nearest Neighbor* dalam seluruh metrik evaluasi. Pada rasio 80:20, Random Forest unggul dengan akurasi 0.73, sedangkan KNN berada di posisi kedua dengan akurasi 0.71. Sementara itu, pada rasio 70:30, performa antara keduanya sama, dengan *Random Forest* mencatatkan akurasi 0.69 dan KNN 0.69. Temuan dari studi ini mengungkap bahwa algotitma *Random Forest* menunjukkan tingkat kestabilan dan ketepatan yang lebih optimal ketimbang algoritma *K-Nearest Neighbor* dalam proses pengelompokan sentimen. Walau demikian, kedua metode masih menunjukkan performa yang layak dalam mengklasifikasi.

4. KESIMPULAN

Studi ini berhasil merancang serta menguji model analisis sentimen terkait kebijakan pembatasan BBM jenis Pertalite, dengan memanfaatkan data komentar pengguna Instagram dan menerapkan algoritma *Random Forest* serta *K-Nearest Neighbor* (KNN). Hasil pengujian menunjukkan bahwa Random Forest secara konsisten memberikan performa lebih baik dibandingkan KNN pada seluruh indikator penilaian performa model dengan dua proporsi data, yaitu 80:20 dan 70:30, mencatat akurasi tertinggi hingga 73%. Temuan ini mengindikasikan bahwa algoritma Random Forest menunjukkan performa yang lebih konsisten dan presisi dalam mendeteksi pola sentimen publik terkait isu kebijakan. Selain itu, hasil klasifikasi mengungkap bahwa sebagian besar opini yang disampaikan masyarakat bernada negatif, yang mengisyaratkan adanya ketidakpuasan terhadap kebijakan pembatasan penggunaan BBM jenis Pertalite. Namun, penelitian ini memiliki beberapa keterbatasan, antara lain ketidakseimbangan jumlah data antar kategori sentimen, dominasi data dari satu platform media sosial, serta belum optimalnya identifikasi terhadap sentimen positif dan netral. Untuk riset selanjutnya disarankan agar ruang lingkup data diperluas dengan melibatkan berbagai platform media sosial serta rentang waktu yang lebih luas, sehingga dapat memberikan gambaran pandangan masyarakat yang lebih menyeluruh. Selain itu, penerapan algoritma yang lebih kompleks juga layak dipertimbangkan demi memperoleh hasil analisis yang lebih optimal.

REFERENCES

[1] E. Setiawati and A. L. Suryanli, "DAMPAK EKONOMIS DAN PSIKOLOGIS KENAIKAN HARGA BBM," *j. ekon. manaj. akunt. perbank. syariah*, vol. 12, no. 1, pp. 298–316, Mar. 2023, [Online]. Available: https://journal.uwgm.ac.id/ekonomika/article/view/1949

[2] A. Shinta and K. Y. S. Putri, "Efektivitas Media Sosial Instagram Terhadap Personal Branding Bintang Emon Pada Pengguna Instagram," *Communicology: Jurnal Ilmu Komunikasi*, vol. 9, no. 1, pp. 98–122, 2021, doi: 10.21009/COMMUNICOLOGY.021.08.

[3] F. A. Indriyani, A. Fauzi, and S. Faisal, "Analisis sentimen aplikasi tiktok menggunakan algoritma naïve bayes dan support vector machine," *TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika*, vol. 10, no. 2, pp. 176–184, 2023, doi: https://doi.org/10.57152/malcom.v4i3.1482.

[4] M. Samantri, "Perbandingan Algoritma Support Vector Machine dan Random Forest untuk Analisis Sentimen Terhadap Kebijakan Pemerintah Indonesia Terkait Kenaikan Harga BBM Tahun 2022," *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, pp. 1–9, 2024, doi: https://doi.org/10.35870/jtik.v8i1.1202.

[5] G. F. I. M. D. A. Chrisley Heltroyce, "Analisis Sentimen Terhadap Kenaikan Harga Bahan Bakar Minyak Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Kesehatan, Sains, dan Teknologi (JAKASAKTI)*, vol. 3, no. 1, pp. 227–232, Apr. 2024, doi: https://doi.org/10.36002/js.v3i1.

[6] N. Alvionika, S. Faisal, R. Rahmat, and A. F. N. Masruriyah, "Analisis Sentimen Pada Komentar Instagram Provider By. U Menggunakan Metode K-Nearest Neighbors (KNN)," *Jurnal Algoritma*, vol. 21, no. 2, pp. 50–63, 2024, doi: https://doi.org/10.33364/algoritma/v.21-2.1672.

[7] S. Rahayu, Y. Mz, J. E. Bororing, and R. Hadiyat, "Implementasi Metode K-Nearest Neighbor (K-NN) untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial FLIP," *Edumatic J. Pendidik. Inform*, vol. 6, no. 1, pp. 98–106, 2022, doi: 10.29408/edumatic.v6i1.5433.

[8] A. Amelia, L. N. Hayati, and H. Darwis, "Analisis Sentimen Masyarakat Terhadap Sistem Pembayaran Mypertamina dengan Metode Random Forest, SVM, dan Naïve Bayes," *LINIER: Literatur Informatika dan Komputer*, vol. 1, no. 1, pp. 28–44, 2024, doi: https://doi.org/10.33096/linier.v1i1.2269.

[9] F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, "Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 6, no. 9, pp. 4305–4313, 2022, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/11562

[10] H. Taufiqqurrahman, F. T. Anggraeny, and M. M. Al Haromainy, "Perbandingan Algoritma Naïve Bayes Dan K-Nearest Neighbor Pada Analisis Sentimen Ulasan Aplikasi Mypertamina," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3934–3939, 2023, doi: https://doi.org/10.36040/jati.v7i6.7801.

[11] D. A. Agustina, S. Subanti, and E. Zukhronah, "Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma Support Vector Machine," *Indonesian Journal of Applied Statistics*, vol. 3, no. 2, pp. 109–122, 2021, doi: 10.26418/jp.v8i3.56478.

[12] T. F. Basar, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen Pengguna Twitter terhadap Pembayaran Cashless menggunakan Shopeepay dengan Algoritma Random Forest," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 3, pp. 1426–1433, 2022, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10830

[13] S. , & W. S. Yahya, "Analisis Sentimen Analisis Sentimen Publik Terhadap Pariwisata Aceh di Media Sosial X Menggunakan Algoritma Naive Bayes Classifier.," *Bulletin of Information Technology (BIT*, vol. 5, no. 4, pp. 269–278, 2024, doi: https://doi.org/10.47065/bit.v5i4.1700.

[14] N. A. R. Putri, "Analisis Jaringan pada Media Sosial X dengan# Boikot Menggunakan Social Network Analysis," *Analisis Jaringan pada Media Sosial X dengan# Boikot Menggunakan Social Network Analysis*, vol. 2, no. 1, pp. 11–5, 2024, [Online]. Available: https://ojisnu.isnuponorogo.org/index.php/ijitech/article/view/79

[15] M. Afdal and L. R. Elita, "Penerapan Text Mining Pada Aplikasi Tokopedia Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, vol. 8, no. 1, pp. 78–87, Feb. 2022, [Online]. Available: https://scispace.com/pdf/penerapan-text-mining-pada-aplikasi-tokopedia-menggunakan-2hggn3of.pdf

[16] Y. Khoiruddin, A. Fauzi, and A. M. Siregar, "Analisis Sentimen Gojek Indonesia Pada Twitter Menggunakan Algoritme Naïve Bayes Dan Support Vector Machine," *Progresif: Jurnal Ilmiah Komputer*, vol. 19, no. 1, pp. 391–400, 2023, doi: 10.35889/progresif.v19i1.1173.

[17] Yeni Kustiyahningsih, Ikromul Islam, Bain Khusnul Khotimah, and Jaka Purnama, "Sentiment Analysis for Indonesian Salt Policy uses Naïve Bayes and Information Gain Methods," *Technium: Romanian Journal of Applied Sciences and Technology*, vol. 17, no. 1, pp. 440–445, Nov. 2023, doi: 10.47577/technium.v17i.10121.

[18] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *INSYST: Journal of Intelligent System and Computation*, vol. 1, no. 1, pp. 43–49, 2019, doi: https://doi.org/10.52985/insyst.v1i1.36.

[19] Yerik Afrianto Singgalen, "Analisis Sentimen Wisatawan terhadap Taman Nasional Bunaken dan Top 10 Hotel Rekomendasi Tripadvisor Menggunakan Algoritma SVM dan DT berbasis CRISP-DM," *Journal of Computer System and Informatics (JoSYC)*, vol. 4, no. 2, pp. 367–379, Feb. 2023, doi: https://doi.org/10.47065/josyc.v4i2.3092.

[20] R. Rahmadini, E. E. LorencisLubis, A. Priansyah, R. W. N. Yolanda, and T. Meutia, "Penerapan Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Mahasiswa Akuntansi Samudra*, vol. 4, no. 4, pp. 223–235, 2023, doi: https://doi.org/10.33059/jmas.v4i4.7074.

[21] D. E. R. B. R. Cahyo Gusti Indrayanto, "Analisis Sentimen Data Ulasan Pengguna Aplikasi MyPertamina di Indonesia pada Google Play Store menggunakan Metode Random Forest," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 3, pp. 1131–1139, Mar. 2023, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12390