Deteksi Sentimen Ujaran Kebencian dalam *Tweet* Media Sosial X dalam Konteks RKUHP Menggunakan Algoritma KNN

Yuda Septiawan[1]*, Juliansyah Adi Putra[2], Adimas Aglasia[3], Danang Ade Muktiawan[4], Meiliza[5]

[1,2,3]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya, Bandar Lampung, Indonesia

[4]Program Studi Sistem Komputer, Fakultas Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya, Bandar Lampung, Indonesia

[5]Program Studi Pendidikan Teknologi Informasi, Fakultas Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya, Bandar Lampung, Indonesia

Email: yuda.septiawan@darmajaya.ac.id[1], juliansyahadiputra.1811010121@.darmajaya.ac.id[2], dimas@darmajaya.ac.id[3], danang@darmajaya.ac.id[4], meiliza@darmajaya.ac.id[5]

Abstrak

Penyebaran ujaran kebencian di platform media sosial, terutama dalam konteks undang-undang kontroversial seperti Rancangan Kitab Undang-Undang Hukum Pidana (RKUHP), menimbulkan tantangan kritis dalam menjaga diskursus publik yang konstruktif. Studi ini bertujuan untuk mengidentifikasi dan mengklasifikasikan sentimen dalam *tweet* yang berkaitan dengan RKUHP menggunakan algoritma *K-Nearest Neighbor* (KNN). Sebuah dataset berisi 703 *tweet* dikumpulkan melalui API X dengan kata kunci "RKUHP", mencerminkan puncak percakapan online pada Juni 2023. Setelah disaring untuk konten Bahasa Indonesia, *tweet* diproses melalui pembersihan (penghapusan tanda baca, tag, URL), normalisasi, dan terjemahan ke dalam Bahasa Inggris. Penandaan sentimen dilakukan menggunakan *TextBlob*, dan kemudian diverifikasi oleh ahli linguistik untuk meningkatkan akurasi label. *Term Frequency–Inverse Document Frequency* (TF–IDF) diterapkan untuk ekstraksi fitur, dan *cosine similarity* digunakan sebagai metrik jarak. Beberapa nilai K (3, 5, 7, 9) diuji, dengan K = 3 menghasilkan akurasi tertinggi sebesar 56,74%. Evaluasi menunjukkan bahwa KNN dapat mendeteksi sentimen terkait ujaran kebencian dalam *tweet* Indonesia secara moderat, meskipun masih ada keterbatasan dalam menangani sarkasme, netralitas, dan ambiguitas terjemahan. Hasil ini mendukung upaya untuk meningkatkan deteksi ujaran kebencian otomatis dengan mengintegrasikan *embeddings* kontekstual, mengatasi ketidakseimbangan kelas, dan memanfaatkan pembelajaran *ensembel*.

Kata Kunci: *K-Nearest Neighbor; Hate Speech; Sentiment Analysis*; RKUHP; X

ABSTRACT

The spread of hate speech on social media platforms, particularly in the context of controversial legislation such as the Indonesian Criminal Code Draft (RKUHP), presents critical challenges to preserving constructive public discourse. This study aims to identify and classify sentiment in tweets related to RKUHP using the K-Nearest Neighbor (KNN) algorithm. A dataset of 703 tweets was collected via the X API using "RKUHP" as the keyword, reflecting the peak of online conversation during June 2023. After filtering for Bahasa Indonesia content, tweets were processed through cleansing (removal of punctuation, tags, URLs), normalization, and translation into English. Sentiment labeling was performed using TextBlob, and subsequently verified by a linguistic expert to

enhance label accuracy. Term Frequency–Inverse Document Frequency (TF–IDF) was applied for feature extraction, and cosine similarity was used as the distance metric. Several K values (3, 5, 7, 9) were tested, with K = 3 yielding the highest accuracy at 56.74%. The evaluation shows that KNN can moderately detect sentiment related to hate speech in Indonesian tweets, though limitations persist in handling sarcasm, neutrality, and translation ambiguity. The results support efforts to improve automated hate speech detection by integrating contextual embeddings, addressing class imbalance, and leveraging ensemble learning.

Keywords: *K-Nearest Neighbor; Hate Speech; Sentiment Analysis; RKUHP; X*

1. PENDAHULUAN

Platform media sosial, terutama X, telah muncul sebagai arena dinamis untuk diskusi publik dan partisipasi warga, terutama selama peristiwa sosial-politik penting seperti debat seputar Rancangan Kitab Undang-Undang Hukum Pidana (RKUHP) (Taradhita & Putra, 2021). Meskipun platform-platform ini mendemokratisasi komunikasi dan memperkuat suara-suara yang beragam, mereka juga secara tidak sengaja menjadi saluran untuk penyebaran cepat ujaran kebencian, yang didefinisikan sebagai ungkapan yang merendahkan, mengancam, atau menghasut permusuhan terhadap individu atau kelompok berdasarkan atribut seperti ras, agama, etnis, atau kewarganegaraan (Hadi & Utami, 2024). Kehadiran konten yang dihasilkan pengguna di platform-platform ini menimbulkan tantangan bagi moderasi manual akibat volume dan kecepatan posting yang sangat besar (Saputra et al., 2023). Akibatnya, pendekatan otomatis yang memanfaatkan *machine learning* (ML) telah menjadi alat penting untuk mendeteksi dan mengurangi penyebaran ujaran kebencian secara massal. Pendekatan ini menawarkan solusi berbasis data, skalabel, dan adaptif yang melengkapi upaya moderasi manusia.

Di antara berbagai algoritma ML, klasifikasi *K-Nearest Neighbor* (KNN) menarik perhatian karena kesederhanaan konseptualnya dan kinerja yang tangguh, terutama ketika dipadukan dengan representasi fitur klasik seperti *Term Frequency–Inverse Document Frequency* (TF–IDF) (Kusuma & Chowanda, 2023). Studi empiris menunjukkan bahwa KNN, bersama dengan TF–IDF, mencapai akurasi kompetitif yang sering melebihi 85% dalam tugas klasifikasi sentimen pada data X Indonesia terkait layanan publik, debat politik, dan gerakan sosial (Ibrohim & Budi, 2023). Efektivitas algoritma *K-Nearest Neighbor* (KNN) dalam mendeteksi ujaran kebencian sangat dipengaruhi oleh beberapa faktor yang saling terkait. Representasi fitur, seperti TF–IDF, dapat menangkap distribusi frekuensi kata tetapi seringkali kurang mampu mewakili nuansa semantik, ketergantungan konteks, dan petunjuk pragmatik yang umum ditemukan dalam diskursus media sosial Indonesia (Arifin & Mahdiana, 2024). Selain itu, pemilihan parameter, terutama nilai K, memainkan peran kritis karena nilai yang tidak tepat dapat meningkatkan sensitivitas terhadap noise atau menyebabkan bias terhadap

kelas mayoritas, sehingga membatasi kemampuan generalisasi model (Saputra et al., 2023). Tantangan lain terletak pada kualitas dataset, yang semakin rumit akibat karakteristik linguistik media sosial Indonesia, termasuk bahasa informal, slang, pergantian kode bahasa, dan sarkasme, yang semuanya menambah kompleksitas dalam tugas deteksi ujaran kebencian (Malik et al., 2022).

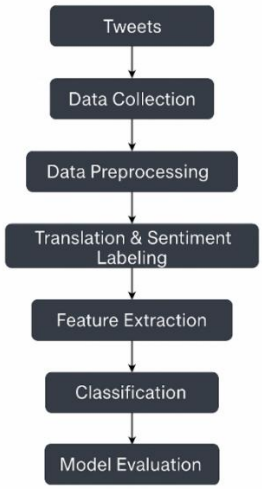
Studi sebelumnya tentang deteksi ujaran kebencian berbasis ML dalam konteks Indonesia mencakup teknik dari *Naive Bayes* dan *Support Vector Machines* (SVM) hingga model *deep learning* seperti *Convolutional Neural Networks* (CNN) dan *Bidirectional* LSTMs (Saleh et al., 2021). Penelitian ini secara konsisten menekankan peran kritis prapemrosesan data, rekayasa fitur, dan optimasi hiperparameter dalam mencapai kinerja klasifikasi optimal. Namun, penelitian yang secara khusus berfokus pada diskursus legislatif Indonesia, seperti debat RKUHP, masih terbatas. Kesenjangan ini patut diperhatikan karena diskursus legislatif sering kali mengintegrasikan strategi retorika kompleks, termasuk seruan emosional, ironi, dan bahasa yang sarat muatan politik. Alat retorika semacam ini dapat mengaburkan batas antara kritik yang sah dan ujaran kebencian, menantang klasifikasi generik yang kurang sensitif terhadap konteks (Kusuma & Chowanda, 2023).

Mengingat kompleksitas tersebut, penelitian ini bertujuan untuk menganalisis sentimen pada *tweet* terkait debat RKUHP dengan penekanan khusus pada identifikasi konten ujaran kebencian yang muncul dalam diskusi publik. Selain itu, studi ini menilai kinerja algoritma *K-Nearest Neighbor* (KNN) dengan menggunakan ekstraksi fitur TF–IDF dan kesamaan kosinus untuk memberikan gambaran komprehensif mengenai tingkat kesesuaiannya dalam tugas klasifikasi ini. Lebih lanjut, penelitian ini juga menyelidiki pengaruh variasi nilai K (3, 5, 7, dan 9) terhadap akurasi klasifikasi, sehingga dapat menentukan konfigurasi KNN yang paling optimal dalam konteks diskursus media sosial Indonesia. Dengan fokus pada konteks sosio-legislatif ini, studi ini berkontribusi pada penelitian tentang deteksi ujaran kebencian yang disesuaikan secara lokal, menawarkan wawasan empiris yang langsung dapat diterapkan pada ekosistem digital Indonesia. Selain itu, temuan ini akan memberikan informasi kepada praktisi dan pembuat kebijakan tentang kemampuan dan batasan pendekatan ML klasik seperti KNN saat dihadapkan pada analisis diskursus yang kompleks dan spesifik budaya dalam dataset Indonesia berskala sedang.

2. METODE

2.1. Kerangka Kerja Penelitian

Penelitian ini dirancang mengikuti alur metodologis terstruktur yang terdiri dari enam tahap utama yaitu pengumpulan data, prapemrosesan data, terjemahan dan penandaan sentimen, ekstraksi fitur, klasifikasi menggunakan algoritma *K-Nearest Neighbor* (KNN), dan evaluasi model. Alur kerja multi-tahap ini dari teks media sosial mentah hingga output model telah banyak digunakan dalam penelitian analisis sentimen X terkini, menggambarkan pendekatan terstruktur dalam tugas penambangan teks (Padhy et al., 2024). Tujuan utama adalah mengevaluasi efektivitas algoritma KNN dalam mengklasifikasikan sentimen yang terkait dengan ujaran kebencian dalam *tweet* yang membahas Rancangan Kitab Undang-Undang Hukum Pidana (RKUHP). Kerangka penelitian secara keseluruhan ditampilkan pada Gambar 1, menggambarkan pendekatan sistematis yang digunakan untuk mengubah konten media sosial mentah menjadi output klasifikasi sentimen yang terstruktur.



Gambar 1. Kerangka Kerja Penelitian

2.2. Data Preprocessing

Untuk memastikan dataset kami siap digunakan untuk analisis sentimen dan pembelajaran mesin (TF-IDF + KNN), kami menerapkan pipeline terstruktur berikut:

1. Pengimporan Data dan Pemeriksaan Kualitas Awal

   Mengimpor *tweet* dan menghapus duplikat, entri yang hilang, serta konten non-Indonesia untuk menjaga integritas data.

2. Pembersihan Data

   Menggunakan *Python* (re, NLTK) untuk menghilangkan karakter *non-alfanumerik*, tanda baca, emoji, hashtag, mention pengguna, dan URL praktik yang terbukti meningkatkan akurasi klasifikasi sentimen pada teks X (Taradhita & Putra, 2021).

3. Konversi Huruf Kecil & Normalisasi *Lexis*

   Mengonversi ke huruf kecil dan menstandarkan variasi leksikal informal (misalnya, "makasiiiii" → "makasi") untuk mengurangi kelangkaan kosakata dan meningkatkan konsistensi fitur (Samad et al., 2020).

4. Terjemahan ke Bahasa Inggris

   Tweet diterjemahkan secara otomatis ke Bahasa Inggris untuk mengakomodasi analis sentimen *TextBlob*; terjemahan berbasis transformer telah terbukti efektif untuk tugas klasifikasi sentimen *tweet* non-Inggris (Barriere & Balahur, 2020).

5. Penghapusan Kata Pengisi & Lemmatisasi

   Kata pengisi Bahasa Inggris dihapus dan lemmatisasi diterapkan menggunakan NLTK/*Spacy*, dengan hati-hati mempertahankan token yang mengandung konteks sesuai saran Samad dkk. (Samad et al., 2020).

6. Vektorisasi TF-IDF

   Mengubah teks menjadi vektor fitur numerik menggunakan TF-IDF, metode standar dan efektif yang dikombinasikan dengan klasifikasi KNN/SVM (Saputra et al., 2023).

7. Penandaan Sentimen Dua Tahap
   a. Penandaan otomatis melalui TextBlob (polaritas >0: positif; =0: netral; <0: negatif).
   b. Peninjauan manual oleh penutur asli untuk memperbaiki terjemahan yang ambigu, mengikuti praktik terbaik dalam anotasi hibrida (Wenando et al., 2025).

8. Pemilihan Fitur

   Jika diperlukan, pilih N fitur TF-IDF teratas untuk mengurangi dimensi dan mencegah *overfitting*.

9. Persiapan Masukan Model

   Vektor TF-IDF yang telah dilabeli siap untuk dimasukkan ke dalam klasifikasi *K-Nearest Neighbors*.

2.3. Ekstraksi Fitur

Untuk mengubah teks yang telah diproses menjadi representasi numerik yang sesuai untuk pembelajaran mesin, metode *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan. TF-IDF menangkap pentingnya relatif setiap istilah dalam dokumen, mengabaikan istilah yang sering muncul di seluruh korpus sambil menonjolkan istilah yang membedakan (Ahmed et al., 2023). Bobot TF-IDF w_t untuk istilah t dalam dokumen d dihitung seperti pada

persamaan 1. Representasi ini terbukti sangat efektif dalam menangkap fitur teks yang relevan untuk pembedaan sentimen.

$$\boldsymbol{w_t = tf(t,d) \cdot \log(\frac{N}{df(t)})} \quad \text{.....................................}(1)$$

Di mana:

a. tf(t,d) adalah frekuensi kemunculan kata t dalam dokumen d.
b. df(t) adalah jumlah dokumen yang mengandung kata t.
c. N adalah jumlah total dokumen.

2.4. Klasifikasi Menggunakan K-Nearest Neighbor (KNN)

Algoritma *K-Nearest Neighbor* (KNN) dipilih sebagai klasifikasi dasar karena kesederhanaannya, kemudahan interpretasi, dan efektivitasnya dalam lingkungan dengan sumber daya terbatas. KNN adalah algoritma pembelajaran non-parametrik dan malas yang mengklasifikasikan titik data baru berdasarkan label mayoritas di antara K sampel pelatihan terdekat, diukur dalam ruang fitur. Dalam studi ini, kesamaan kosinus digunakan sebagai metrik jarak, yang sangat cocok untuk data berdimensi tinggi dan jarang seperti vektor TF-IDF. Kesamaan kosinus antara vektor A dan B ditunjukkan pada persamaan 2.

$$\cos(\theta) = \frac{A \cdot B}{||A||||B||} \quad \text{.....................................}(2)$$

Di mana A dan B adalah vektor TF-IDF dari dua *tweet*, $A \cdot B$ menandakan hasil kali dot, dan $\|A\|$, $\|B\|$ adalah norma *Euclidean.* Algoritma ini diuji dengan empat nilai K (3, 5, 7, dan 9) untuk menentukan pengaturan parameter optimal dalam klasifikasi sentimen. Pendekatan serupa dalam literatur terbaru telah menunjukkan kinerja yang baik, (Olajuwon et al., 2024) mencapai akurasi klasifikasi tinggi pada data X terkait Piala Dunia Qatar 2023 menggunakan TF-IDF + KNN, dan (Romli et al., 2021) melaporkan akurasi hingga 82% dengan KNN-*cosine* pada *tweet* Indonesia tentang pembatasan sosial.

2.5. Evaluasi Model

Kinerja model dievaluasi menggunakan akurasi sebagai metrik utama, dihitung sesuai Persamaan (3). Dataset dibagi menjadi 80% dan 20% untuk pelatihan dan pengujian melalui sampling acak terstratifikasi, memastikan distribusi kelas sentimen yang proporsional. Pendekatan ini sejalan dengan studi-studi terbaru (Munawaroh & Alamsyah, 2023) yang juga menggunakan metrik akurasi berbasis matriks kebingungan setelah pembagian pelatihan-pengujian dan mencakup evaluasi kinerja KNN secara komparatif. Metrik ini memberikan penilaian yang sederhana namun informatif mengenai keakuratan model di berbagai kategori sentimen. Penggunaan nilai K yang beragam memungkinkan

analisis perbandingan untuk menentukan konfigurasi terbaik berdasarkan dataset dan batasan prapemrosesan yang ada. Akurasi dihitung dengan persamaan 3.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(3)}$$

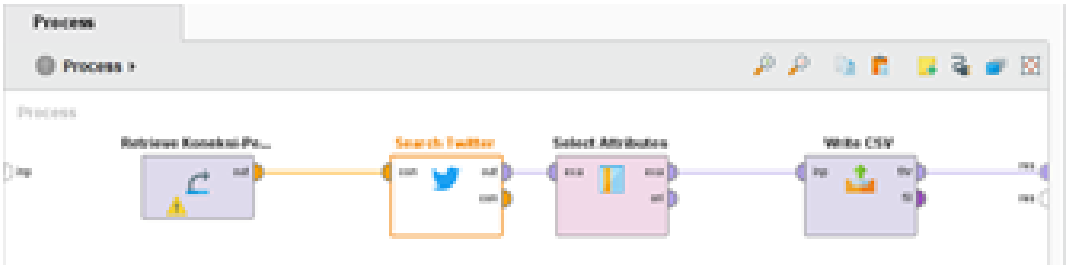
Di mana:

a. TP = Positif Benar
b. TN = Negatif Benar
c. FP = Positif Palsu
d. FN = Negatif Palsu

3. HASIL DAN PEMBAHASAN

3.1. Data Collection

Data set untuk penelitian ini dikumpulkan dari X melalui API resminya menggunakan bantuan software RapidMiner dengan mengambil data pada periode 1–30 Juni 2023. Periode ini dipilih untuk bertepatan dengan puncak perbincangan online setelah RKUHP disahkan oleh Dewan Perwakilan Rakyat Indonesia pada 6 Desember 2022 (Irawan et al., 2023). Sesuai dengan analisis tren dan studi sentimen RKUHP sebelumnya (Michel et al., 2023), kata kunci "RKUHP" digunakan sebagai kueri pencarian. Pengumpulan data dilakukan dengan cara menangkap tweet dengan cara mendapatkan data secara langsung (*Crawl*) menggunakan API (*Application Interface*) pada X. Gambar 2 menampilkan pengumpulan data menggunakan RapidMiner.



Gambar 2. Pengumpulan Data Menggunakan RapidMiner

3.2. Preprocessing Data

Setelah mendapatkan data, akan dilakukan preprocessing data. Preprocessing data dilakukan dengan menghilangkan *noise,* menghapus data duplikat, memeriksa data untuk ketidakkonsistenan, dan memperbaiki kesalahan dalam data, seperti kesalahan ketik. Pada tahap ini data akan dibagi menjadi data testing dan data training dengan rasio 80% untuk data testing dan 20% data training dan didapatkan hasil pada Tabel 1.

Tabel 1. Pembagian Data Testing dan Data Training

Jumlah Data (N)	Data Testing (N x 20%)	Data Training (N x 80%)
704	141	563

3.3. Cleaning Data

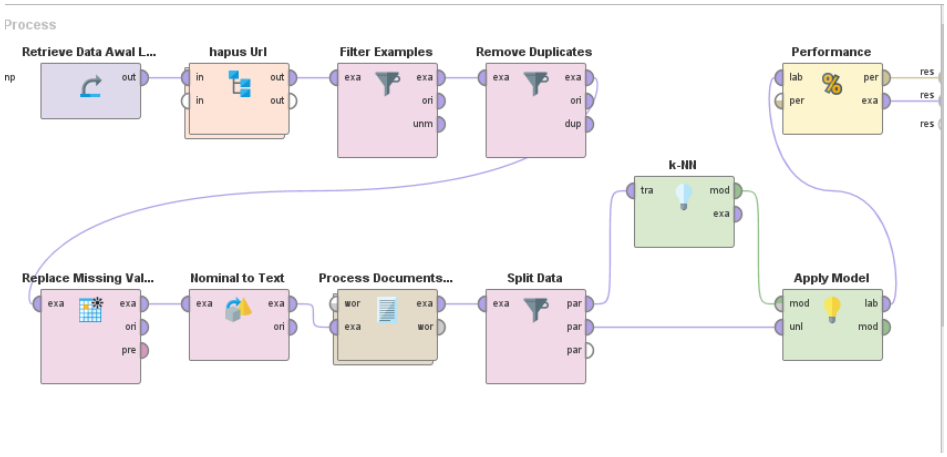
Tahap cleaning data dilakukan sebagai tahap awal yang sangat penting dalam penelitian ini. Hal ini dikarenakan data yang didapatkan dari X masih dalam bentuk teks yang tidak sesuai kaidah dan tidak lengkap. Normalisasi adalah proses penskalaan nilai atribut dari data sehingga terletak pada rentang tertentu. Selain itu pada tahap normalisasi juga dilakukan cleaning data menggunakan pemrograman Python, hal ini bertujuan untuk memastikan isi dari data tersebut. Cleaning data yang akan menghapus atribut-atribut yang tidak diperlukan di dalam data seperti, tanda baca, emot ikon, angka dan lain-lain. Atribut-atribut yang akan dihapus contohnya "@[A-Za-z0-9]+, [0-9]+, #, [^\w], '_', [\n]+, :, 'RT[\s]+, ^https?:\/\/.[\r\n], ^http?:\/\/.[\r\n]". Proses cleaning data ditunjukkan pada Gambar 3.

```
def cleanTweet(text):
    text = re.sub(r'@[A-Za-z0-9]+', '', text)
    text = re.sub(r'[0-9]+', '', text)
    text = re.sub(r'#','',text)
    text = re.sub(r'[^\w]', ' ', text)
    text = re.sub(r'_', ' ', text)
    text = re.sub(r'[\n]+', '', text)
    text = re.sub(r':', '', text)
    text = re.sub(r'RT[\s]+', '', text)


    return text
df['clean_tweets'] = df['Text'].apply(cleanTweet)

df.head()
```

Gambar 3. Cleaning Data dengan Python

3.4. Implementasi Algoritma K-Nearest Neighbor Menggunakan Klasifikasi

Gambar 4 menampilkan implementasi Algoritma Klasifikasi K-Nearest Neighbor menggunakan RapidMiner dengan alat K-Nearest Neighbor. Pada tahap ini, fitur yang digunakan dalam pengimplementasian metode Klasifikasi K-Nearest Neighbor ada beberapa yaitu diantaranya, Fitur *Subprocess,* Fitur *Filter Example* dan *Remove Duplicate*, Fitur *Replace Missing Value* dan *Nominal to Text,* Fitur *Process Documents from Data* dan Fitur *Split Data.* Fitur-fitur ini digunakan agar data yang digunakan dan dapat digunakan secara optimal pada proses pengimpelentasiannya pada algoritma K-Nearest Neighbor. Adapun pada Tabel 2 merupakan model yang akan digunakan dalam menetukan prediksi pada label data yang sudah melewati tahapan training untuk melihat seberapa akurat data yang akan digunakan pada model Klasifikasi K-Nearest Neighbor. Untuk melihat hasil dari akurasi maka "K" yang akan digunakan adalah K=3, K=5, K=7 dan K=9.



Gambar 4. Implementasi Algoritma *K-Nearest Neighbor* RapidMiner

Tabel 2. Parameter data yang digunakan

Parameter "K"	Measure Types	Mixed Measures
3, 5, 7, 9	MixedMeasures	MixedEcludianDistance

Setelah ditentukan parameter "K" yang akan digunakan, maka dapat dilakukan percobaan pada tiap-tipa parameter guna mendapatkan hasil dari tiap parameter. Dan berikut hasil dari yang diperoleh dari percobaan pada tiap parameter "K". Pada percobaan dengan parameter K=3 didapatkan hasil sebagai berikut yang ditunjukan pada Gambar 5.

accuracy: 56.74%

	true Positive	true Negative	true Neutral	class precision
pred. Positive	14	8	3	56.00%
pred. Negative	7	20	7	58.82%
pred. Neutral	22	14	46	56.10%
class recall	32.56%	47.62%	82.14%	

Gambar 5. Akurasi Prediksi Parameter K=3

Gambar 5 menunjukkan hasil evaluasi model klasifikasi dengan tingkat akurasi sebesar 56,74%. Berdasarkan tabel *confusion matrix*, terlihat bahwa model memiliki performa terbaik dalam mengidentifikasi kelas *neutral*, dengan nilai *recall* sebesar 82,14% dan *precision* sebesar 56,10%. Sementara itu, kelas *positive* dan *negative* memiliki *recall* yang lebih rendah, yaitu masing-masing 32,56% dan 47,62%, menandakan bahwa model masih sering keliru dalam membedakan kedua kelas tersebut. Secara keseluruhan, model ini perlu dilakukan penyempurnaan, terutama dalam meningkatkan kemampuan klasifikasi untuk kelas *positive* dan *negative* agar akurasi keseluruhan dapat meningkat.

accuracy: 52.48%

	true Positive	true Negative	true Neutral	class precision
pred. Positive	9	6	5	45.00%
pred. Negative	8	21	7	58.33%
pred. Neutral	26	15	44	51.76%
class recall	20.93%	50.00%	78.57%	

Gambar 6. Akurasi Prediksi Parameter K=5

Pada percobaan dengan parameter K=5 didapatkan hasil yang ditunjukkan pada Gambar 6. Gambar 6 tersebut memperlihatkan hasil evaluasi performa model klasifikasi dengan akurasi sebesar 52,48%. Berdasarkan *confusion matrix*, model menunjukkan performa terbaik pada kelas *neutral*, dengan *recall* sebesar 78,57% dan *precision* sebesar 51,76%, menandakan kemampuan yang cukup baik dalam mengenali data netral. Namun, performa model pada kelas *positive* dan *negative* masih rendah, dengan *recall* masing-masing 20,93% dan 50,00%, menunjukkan bahwa banyak data positif dan negatif yang belum terdeteksi secara akurat. Secara keseluruhan, model ini masih memerlukan peningkatan, khususnya dalam penyeimbangan data antar kelas dan optimasi parameter agar hasil prediksi menjadi lebih stabil dan akurat.

accuracy: 56.03%

	true Positive	true Negative	true Neutral	class precision
pred. Positive	11	4	5	55.00%
pred. Negative	9	25	8	59.52%
pred. Neutral	23	13	43	54.43%
class recall	25.58%	59.52%	76.79%	

Gambar 7. Akurasi Prediksi Parameter K=7

Pada percobaan dengan parameter K=7 didapatkan hasil yang ditunjukkan pada Gambar Gambar 7. Gambar 7 menampilkan hasil evaluasi model klasifikasi dengan akurasi sebesar 56,03%. Berdasarkan tabel *confusion matrix*, model memiliki performa tertinggi pada kelas *neutral*, dengan *recall* sebesar 76,79% dan *precision* sebesar 54,43%, yang menunjukkan kemampuan cukup baik dalam mengenali data netral. Sementara itu, kelas *negative* memiliki *recall* sebesar 59,52% dan *precision* yang relatif baik yaitu 59,52%, menandakan prediksi yang cukup seimbang. Namun, untuk kelas *positive*, performa masih rendah dengan *recall* hanya 25,58%, yang mengindikasikan bahwa banyak data positif belum teridentifikasi dengan benar. Secara keseluruhan, model menunjukkan performa sedang dan masih perlu ditingkatkan, khususnya dalam mendeteksi kelas positif agar hasil klasifikasi lebih akurat dan merata di semua kategori.

Pada percobaan dengan parameter K=9 didapatkan hasil yang ditunjukkan pada Gambar

8. Gambar 8 menunjukkan hasil evaluasi performa model klasifikasi dengan tingkat akurasi sebesar 53,19%. Berdasarkan tabel *confusion matrix*, model memiliki performa terbaik pada kelas *neutral*, dengan *recall* sebesar 71,43% dan *precision* sebesar 54,05%, menunjukkan kemampuan yang cukup baik dalam mengenali data netral. Kelas *negative* menempati posisi kedua dengan *recall* 57,14% dan *precision* 53,33%, menandakan keseimbangan yang moderat antara ketepatan dan kelengkapan prediksi. Namun, performa pada kelas *positive* masih rendah dengan *recall* sebesar 25,58%, menandakan banyak data positif yang tidak terdeteksi secara akurat. Secara keseluruhan, model ini masih perlu ditingkatkan, terutama dalam mengoptimalkan deteksi kelas positif agar hasil klasifikasi menjadi lebih merata dan akurasi keseluruhan meningkat**.**

accuracy: 53.19%

	true Positive	true Negative	true Neutral	class precision
pred. Positive	11	4	7	50.00%
pred. Negative	12	24	9	53.33%
pred. Neutral	20	14	40	54.05%
class recall	25.58%	57.14%	71.43%	

Gambar 8. Akurasi Prediksi Parameter K=9

3.5. Evaluasi Performance

Evaluasi *performance* menunjukkan bahwa kinerja K-Nearest Neighbor (KNN) sangat sensitif terhadap pilihan parameter K. Tabel 3 menampilkan akurasi klasifikasi pada berbagai nilai. Akurasi optimal sebesar 56,74% dicapai pada K = 3, menunjukkan bahwa ukuran tetangga yang lebih kecil lebih baik dalam menangkap kluster sentimen lokal di ruang fitur TF-IDF berdimensi tinggi kami. Saat K meningkat (5, 7, 9), akurasi menurun kemungkinan karena penambahan tetangga yang lebih heterogen, yang melemahkan sinyal spesifik kelas. Dari hasil seluruh percobaan terdapat data sebelum dan sesudah diimplementasikan pada model KNN dengan data training yang sudah dibagi menggunakan RapidMiner pada Tabel 4.

Tabel 3. Kinerja Klasifikasi KNN Untuk Berbagai Nilai K

K Value	*Accuracy (%)*
3	**56.74**
5	52.48
7	56.03
9	53.19

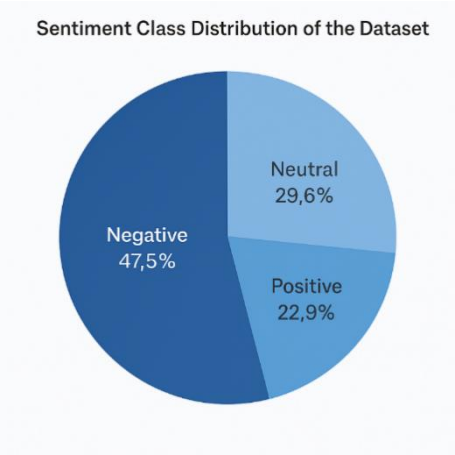
Tabel 4. Data Awal dan Data Prediksi dengan "K"

Label	Data Awal	K=3	K=5	K=7	K=9
Positif	18	25	20	20	22

Label	Data Awal	K=3	K=5	K=7	K=9
Netral	62	82	85	79	74
Negatif	61	34	36	42	45
Total	141	141	141	141	141

3.6. Distribusi Sentimen dan Wawasan Dataset

Berdasarkan data yang diperoleh pada percobaan sebelumnya, Gambar 9 menyoroti ketidakseimbangan kelas yang signifikan dalam dataset kami yang terdiri dari 703 *tweet*, dengan 47,5% sentimen negatif, 29,6% netral, dan 22,9% positif. Kecenderungan ke arah sentimen negatif ini menunjukkan penolakan publik yang dominan, kemungkinan terkait dengan draf RKUHP yang kontroversial. Ukuran kelas positif yang relatif kecil dapat menghambat kemampuan model untuk secara akurat mengidentifikasi *tweet* yang mendukung, sementara kategori netral yang menempati hampir sepertiga dari data dapat mengaburkan batas-batas sentimen yang jelas.



Gambar 9. Distribusi Kelas Sentimen pada Set Data

Ketidakseimbangan kelas semacam ini menimbulkan risiko klasifikasi yang sudah dikenal bias terhadap kelas mayoritas. Penelitian secara konsisten menunjukkan bahwa model yang dilatih pada dataset yang tidak seimbang cenderung terlalu sering memprediksi label mayoritas, yang mengakibatkan *recall* yang buruk untuk kelas minoritas. Misalnya, teknik resampling sintetis seperti SMOTE telah terbukti dapat meningkatkan kinerja kelas minoritas sebesar 10–12% dalam tugas sentimen, dibandingkan dengan model yang tidak di-*resampling*. Sebaliknya, tanpa langkah-langkah tersebut, bahkan klasifikasi canggih pun mungkin kesulitan untuk generalisasi secara merata di semua kelas.

Studi *deep learning* seperti yang menggunakan BiLSTM dengan *embeddings FastText* menunjukkan bahwa ketidakseimbangan kelas menyebabkan kinerja lebih rendah pada kategori yang

kurang terwakili, dengan selisih F1-score hingga ~15 poin persentase antara sentimen minoritas dan mayoritas. Untuk mengatasi masalah ini, literatur menyarankan beberapa mitigasi:

a. Strategi Resampling: Penerapan SMOTE (atau RUS) secara konsisten meningkatkan akurasi seimbang di berbagai algoritma pembelajaran terawasi.
b. Protokol Pelatihan Seimbang: Teknik seperti sampling terstratifikasi, fungsi kerugian yang *diweighted* berdasarkan kelas, dan penyesuaian ambang batas membantu memperbaiki bias model tanpa meningkatkan kinerja secara artifisial.

3.7. Implikasi Teoritis dan Temuan Perbandingan

Meskipun model K-Nearest Neighbor (KNN) dalam studi ini mencapai kinerja puncak moderat sebesar 56,74%, hasil tersebut tetap relatif lebih rendah dibandingkan dengan yang dilaporkan dalam studi serupa. Misalnya, (Su et al., 2023) memperoleh akurasi 64,2% menggunakan KNN pada dataset bahasa Inggris yang seimbang. Perbedaan ini dapat dikaitkan dengan beberapa faktor, termasuk kompleksitas linguistik bahasa Indonesia yang kaya akan pergantian kode, ungkapan idiomatik, dan variasi dialek yang seringkali menantang pipeline NLP standar. Selain itu, pergeseran semantik yang dihasilkan oleh terjemahan mesin dari bahasa Indonesia ke bahasa Inggris dapat menyamarkan nuansa emosional dan memengaruhi skor polaritas, sementara ketidakseimbangan kelas dalam dataset kemungkinan mengurangi kinerja model pada kelas minoritas masalah umum dalam pembelajaran data yang tidak seimbang.

Secara teoritis, penelitian ini memperkuat posisi KNN sebagai dasar yang berharga untuk klasifikasi sentimen dalam lingkungan linguistik dengan sumber daya terbatas, memberikan landasan untuk eksplorasi lebih lanjut model *deep learning* yang lebih sadar konteks atau hibrida. Secara praktis, temuan ini memberikan wawasan berguna bagi lembaga pemerintah dalam memantau dan mengidentifikasi potensi ujaran kebencian atau polarisasi opini publik di media sosial, terutama terkait isu-isu kebijakan seperti RKUHP. Bagi masyarakat umum, penelitian ini meningkatkan kesadaran tentang pola komunikasi digital dan dampak sosial potensial dari ujaran kebencian. Oleh karena itu, studi ini tidak hanya memperkaya diskursus akademik tentang analisis sentimen bahasa Indonesia tetapi juga berkontribusi dalam membangun ekosistem digital yang lebih sehat dan inklusif.

4. PENUTUP

Simpulan dan Saran

Studi ini membuktikan bahwa algoritma KNN dapat digunakan sebagai acuan dasar untuk analisis sentimen tweet Indonesia terkait debat RKUHP, dengan kinerja terbaik dicapai pada K = 3, menghasilkan akurasi 56,74%. Meskipun hasil ini memberikan acuan yang berarti, KNN masih kesulitan menangkap nuansa semantik yang lebih dalam dan ketergantungan

kontekstual. Diskursus sosio-politik di media sosial Indonesia sering kali ditandai dengan ironi, sarkasme, dan pergantian kode bahasa, yang sulit diklasifikasikan secara efektif oleh algoritma sederhana seperti KNN. Dibandingkan dengan model berbasis *transformer* seperti IndoBERT, yang mencapai akurasi di atas 80%, selisih kinerja lebih dari 20% menyoroti keterbatasan KNN dalam tugas analisis sentimen yang kompleks. Untuk penelitian masa depan, perbaikan harus berfokus pada adopsi model *deep learning* canggih seperti IndoBERTweet, yang lebih efektif dalam menangkap makna kontekstual. Meningkatkan prapemrosesan dengan memasukkan leksikon khusus sentimen, *embeddings slang*, dan sumber daya linguistik yang disesuaikan dapat meningkatkan representasi. Mengatasi ketidakseimbangan kelas melalui metode seperti SMOTE, pelatihan sensitif, atau strategi *ensembel* juga akan memperkuat kinerja klasifikasi. Selain itu, menerapkan teknik optimasi *hiperparameter* sistematis, termasuk pencarian *grid* atau pencarian acak, dapat membantu memaksimalkan akurasi dan generalisasi model.

DAFTAR PUSTAKA

Ahmed, W., Semary, N., Amin, K., & Adel Hammad, M. (2023). Sentiment Analysis on Twitter Using Machine Learning Techniques and TF-IDF Feature Extraction: A Comparative Study. *IJCI. International Journal of Computers and Information*, *10*(3), 52–57. https://doi.org/10.21608/ijci.2023.236052.1128

Arifin, M., & Mahdiana, D. (2024). Implementation of Deep learning models in hate speech detection on twitter using an natural language processing approach. *Jurnal Teknik Informatika (Jutif)*, *5*(5), 1257–1266. https://doi.org/10.52436/1.jutif.2024.5.5.2043

Barriere, V., & Balahur, A. (2020). Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation. *Proceedings of the 28th International Conference on Computational Linguistics*, 266–271. https://doi.org/10.18653/v1/2020.coling-main.23

Hadi, K., & Utami, E. (2024). Analysis of K-NN with the Integration of Bag of Words, TF-IDF, and N-Grams for Hate Speech Classification on Twitter. *JUITA: Jurnal Informatika*, *12*(2), 289. https://doi.org/10.30595/juita.v12i2.23829

Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon*, *9*(8), e18647. https://doi.org/10.1016/j.heliyon.2023.e18647

Kusuma, J. F., & Chowanda, A. (2023). Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter. *JOIV : International Journal on Informatics Visualization*, *7*(3), 773–780. https://doi.org/10.30630/joiv.7.3.1035

Malik, J. S., Qiao, H., Pang, G., & Hengel, A. van den. (2022). *Deep Learning for Hate Speech Detection: A Comparative Study* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2202.09517

Munawaroh, K., & Alamsyah, A. (2023). Performance Comparison of SVM, Naïve Bayes, and KNN Algorithms for Analysis of Public Opinion Sentiment Against COVID-19 Vaccination on Twitter. *Journal of Advances in Information Systems and Technology*, *4*(2), 113–125. https://doi.org/10.15294/jaist.v4i2.59493

Olajuwon, S. Muh. R., Kusrini, K., & Kusnawi, K. (2024). Analyzing Public Sentiment Regarding the Qatar 2023 World Cup Debate Using TF-IDF and K-Nearest Neighbor Weighting. *Sinkron*, *8*(2), 679–688. https://doi.org/10.33395/sinkron.v8i2.13275

Padhy, M., Modibbo, U. M., Rautray, R., Tripathy, S. S., & Bebortta, S. (2024). Application of Machine Learning Techniques to Classify Twitter Sentiments Using Vectorization Techniques. *Algorithms*, *17*(11), 486. https://doi.org/10.3390/a17110486

Romli, I., Prameswari R, S., & Kamalia, A. Z. (2021). Sentiment Analysis about Large-Scale Social Restrictions in Social Media Twitter Using Algoritm K-Nearest Neighbor. *Jurnal Online Informatika*, *6*(1), 96–102. https://doi.org/10.15575/join.v6i1.670

Saleh, H., Alhothali, A., & Moria, K. (2021). *Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2111.01515

Samad, M. D., Khounviengxay, N. D., & Witherow, M. A. (2020). *Effect of Text Processing Steps on Twitter Sentiment Classification using Word Embedding* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2007.13027

Saputra, N. A., Aeni, K., & Saraswati, N. M. (2023). Indonesian Hate Speech Text Classification Using Improved K-Nearest Neighbor with TF-IDF-ICSρF. *Scientific Journal of Informatics*, *11*(1), Article 1. https://doi.org/10.15294/sji.v11i1.48085

Su, J., Chen, Q., Wang, Y., Zhang, L., Pan, W., & Li, Z. (2023). Sentence-level sentiment analysis based on supervised gradual machine learning. *Scientific Reports*, *13*(1), 14500. https://doi.org/10.1038/s41598-023-41485-8

Taradhita, D. A. N., & Putra, I. K. G. D. (2021). Hate Speech Classification in Indonesian Language Tweets by Using Convolutional Neural Network. *Journal of ICT Research and Applications*, *14*(3), 225–239. https://doi.org/10.5614/itbj.ict.res.appl.2021.14.3.2

Wenando, F. A., Yusoff, N., Izrin, N., Ahmad, S. R. N., Salim, M., Puspa, M. A., & Novaliendry, D. (2025). Optimizing Hate Speech Detection in Indonesian Social Media: An ADASYN and LSTM-Based Approach. *Journal Européen Des Systèmes Automatisés*, *58*(1), 13–20. https://doi.org/10.18280/jesa.580102