I N F O R M A T I K A
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

ANALISIS SENTIMEN MASYARAKAT PADA PLATFORM MEDIA SOSIAL X TERHADAP ISU PERSELINGKUHAN MENGGUNAKAN K-NEAREST NEIGHBORS

Havni Virul[1], Abdul Halim Hasugian[2]
[1,2]Program Studi Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sumatera Utara, Jl. Lap. Golf No.120, Kp. Tengah, Kec. Pancur Batu, Kabupaten Deli Serdang, Sumatera Utara 20353
E-mail : havnivirul12@gmail.com[1] , abdulhalimhasugian@uinsu.ac.id[2]

ABSTRAK

Perkembangan media sosial, khususnya platform X (Twitter), menjadikannya ruang utama bagi masyarakat untuk menyampaikan opini terhadap berbagai isu sosial, termasuk isu perselingkuhan yang sering memicu perdebatan moral dan emosional. Banyaknya opini yang muncul mendorong perlunya analisis sentimen untuk memahami kecenderungan persepsi publik secara objektif dan sistematis. Penelitian ini bertujuan menganalisis sentimen masyarakat terhadap isu perselingkuhan di media sosial X menggunakan algoritma K-Nearest Neighbor (KNN). Metode yang digunakan adalah pendekatan kuantitatif dengan data tweet hasil web scraping berdasarkan kata kunci terkait perselingkuhan. Data kemudian diproses melalui tahap preprocessing, meliputi cleaning, case folding, tokenisasi, normalisasi, stopword removal, dan stemming. Selanjutnya, data direpresentasikan menggunakan TF-IDF dan diklasifikasikan dengan algoritma KNN berbasis jarak Euclidean. Hasil penelitian menunjukkan bahwa KNN mampu mengklasifikasikan sentimen positif dan negatif dengan akurasi 91,49%. Model menunjukkan performa sangat baik pada sentimen negatif, meskipun recall sentimen positif masih terbatas. Secara keseluruhan, algoritma KNN terbukti efektif dan andal untuk analisis sentimen isu sosial di media sosial.

Kata kunci *: Analisis Sentimen, Media Sosial X, Perselingkuhan, K-Nearest Neighbor, TF-IDF.*

ABSTRACT

The development of social media, particularly platform X (Twitter), has made it a primary platform for people to express their opinions on various social issues, including the issue of infidelity, which often triggers moral and emotional debates. The numerous emerging opinions have prompted the need for sentiment analysis to objectively and systematically understand public perception trends. This study aims to analyze public sentiment towards the issue of infidelity on social media X using the K-Nearest Neighbor (KNN) algorithm. The method used is a quantitative approach with tweet data from web scraping based on keywords related to infidelity. The data is then processed through a preprocessing stage, including cleaning, case folding, tokenization, normalization, stopword removal, and stemming. Next, the data is represented using TF-IDF and classified with the Euclidean distance-based KNN algorithm. The results show that KNN is able to classify positive and negative sentiments with an accuracy of 91.49%. The model performs very well on negative sentiments, although recall of positive sentiments is still limited. Overall, the KNN algorithm has proven effective and reliable for sentiment analysis of social issues on social media.

Keywords : *Sentiment Analysis, Social Media X, Infidelity, K-Nearest Neighbor, TF-IDF.*

1. PENDAHULUAN

Perkembangan teknologi informasi telah membawa perubahan signifikan dalam cara masyarakat berkomunikasi dan berbagi informasi (Saidah & Mayary, 2020). Media sosial, khususnya Twitter atau kini disebut X, menjadi platform yang populer untuk menyampaikan opini dan pandangan terhadap berbagai isu sosial (Dharmawan et al., 2020). Kemampuan *platform* X dalam menyebarkan informasi secara cepat dan luas menjadikannya sumber data yang berharga untuk menganalisis sentimen publik terhadap isu-isu tertentu. Menurut

I N F O R M A T I K A
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

Lestari dan Mahdiana (2021), X digunakan secara luas oleh masyarakat untuk menyampaikan opini terkait topik yang sedang hangat dibahas, seperti kebijakan pemerintah atau isu sosial lainnya (Lestari & Mahdiana, 2021).

Analisis sentimen merupakan metode yang digunakan untuk mengidentifikasi dan mengklasifikasikan opini atau perasaan pengguna terhadap suatu topik menjadi kategori positif, negatif, atau netral (Fachrudin et al., 2024). Metode ini telah banyak diterapkan dalam berbagai penelitian untuk memahami persepsi publik terhadap produk, layanan, maupun isu sosial. Dharmawan et al. (2020) menerapkan analisis sentimen pada X untuk mengevaluasi layanan Sistem Informasi Akademik Mahasiswa Universitas Brawijaya, dengan menggunakan metode K-Nearest Neighbor (KNN) dan mencapai akurasi sebesar 86% (Dharmawan et al., 2020).

Isu perselingkuhan merupakan salah satu topik yang sering menjadi perbincangan di media sosial, termasuk X. X adalah platform mikroblogging yang memungkinkan pengguna untuk membagikan pesan singkat, yang dikenal sebagai "tweet", dengan batasan karakter tertentu. Karakteristik utama X meliputi kecepatan dalam penyebaran informasi, penggunaan tagar (#) untuk mengelompokkan topik, serta kemampuan untuk menyampaikan opini secara langsung dan real-time. Platform ini telah menjadi sumber data yang kaya untuk analisis sentimen karena volume dan keragaman kontennya (Gibran et al., 2024). Perselingkuhan tidak hanya dipandang sebagai persoalan pribadi, tetapi juga sebagai isu sosial yang menimbulkan perdebatan moral, stigma, serta dapat memengaruhi persepsi masyarakat terhadap nilai keluarga, pernikahan, dan kepercayaan dalam hubungan. Perbincangan mengenai kasus perselingkuhan publik figur misalnya, kerap memicu gelombang opini yang beragam, mulai dari kecaman, dukungan, hingga sikap netral. Opini yang terbentuk di media sosial mencerminkan nilai-nilai sosial dan moral yang berlaku serta dapat memengaruhi cara masyarakat memandang fenomena ini. Namun, hingga saat ini belum banyak penelitian yang secara khusus menganalisis sentimen masyarakat terhadap isu perselingkuhan di media sosial dengan pendekatan analisis sentimen. Padahal, studi semacam ini penting untuk memahami bagaimana masyarakat menilai perselingkuhan, apakah didominasi oleh sentimen negatif, netral, atau bahkan positif.(Halimi et al., 2021).

Metode K-Nearest Neighbor (KNN) merupakan salah satu algoritma yang efektif dalam klasifikasi data, termasuk dalam analisis sentimen. KNN bekerja dengan mengklasifikasikan data berdasarkan kedekatan dengan data lain yang telah diklasifikasikan sebelumnya. Dalam penelitian oleh Khoirunnisa et al. (2025), KNN digunakan untuk menganalisis sentimen masyarakat terhadap Pemilu 2024 melalui media sosial X, dengan hasil akurasi mencapai 97,50%, menunjukkan keunggulan KNN dibandingkan algoritma lainnya seperti *Naïve Bayes* dan *Decision Tree* (Khoirunnisa et al., 2024). Selain itu, Dewo (2020) dalam penelitiannya mengenai opini publik terhadap pemindahan ibu kota, menunjukkan bahwa KNN mencapai akurasi sebesar 95,02%, lebih tinggi dibandingkan dengan algoritma *Decision Tree* dan *Support Vector Machine* (Dewo, 2022). Temuan-temuan ini mengindikasikan bahwa KNN memiliki kinerja yang unggul dalam mengklasifikasikan sentimen dari data teks yang kompleks dan beragam.

Penggunaan KNN dalam analisis sentimen terhadap isu sosial juga telah dilakukan oleh Pamungkas dan Kharisudin (2021), yang menganalisis tanggapan masyarakat Indonesia terhadap pandemi COVID-19 di X (Pamungkas & Kharisudin, 2021). Penelitian ini menunjukkan bahwa KNN dapat digunakan secara efektif untuk mengklasifikasikan sentimen dalam data teks yang besar dan beragam.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap isu perselingkuhan yang dibagikan melalui media sosial X dengan menggunakan algoritma K-Nearest Neighbor (KNN). Berdasarkan uraian tersebut, penulis akan mengangkat penelitian dengan judul **"Analisis Sentimen Masyarakat pada Platform Media Sosial X (Twitter) Terhadap Isu Perselingkuhan Menggunakan K-Nearest Neighbor (KNN)"**. Diharapkan penelitian ini dapat memberikan gambaran yang lebih jelas mengenai persepsi publik terhadap isu yang diangkat serta memberikan kontribusi terhadap pengembangan kajian analisis sentimen di era digital.

2. METODE PENELITIAN

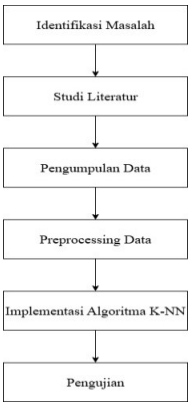
Dalam penelitian ini, pendekatan kuantitatif digunakan untuk mengeksplorasi dan memetakan sentimen masyarakat terhadap isu perselingkuhan di platform media sosial X. Penelitian ini mengandalkan data digital berupa cuitan publik yang diperoleh melalui proses pengumpulan data secara otomatis menggunakan teknik *web scraping* dengan Python. Tweet yang dikumpulkan dipilih berdasarkan kata kunci tertentu yang berkaitan dengan topik perselingkuhan.

Setelah data diperoleh, tahap selanjutnya adalah preprocessing teks, yang mencakup beberapa tahapan seperti *case folding*, tokenisasi, *stopword removal*, *stemming*, dan konversi ke representasi

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

numerik menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*). Proses ini bertujuan untuk membersihkan dan menstandarkan data agar siap digunakan dalam proses klasifikasi.

Untuk klasifikasi sentimen, digunakan algoritma K-Nearest Neighbor (KNN), yang bekerja berdasarkan prinsip kedekatan antar data. Algoritma ini akan mengelompokkan data ke dalam tiga kategori utama, yaitu positif, dan negatif berdasarkan jarak terdekat dalam ruang fitur. Dengan metode ini, penelitian ini diharapkan mampu memberikan gambaran yang objektif dan terukur mengenai opini publik terhadap isu perselingkuhan yang ramai dibicarakan di media sosial.



Gambar 1 Kerangka Penelitian Pengumpulan Data

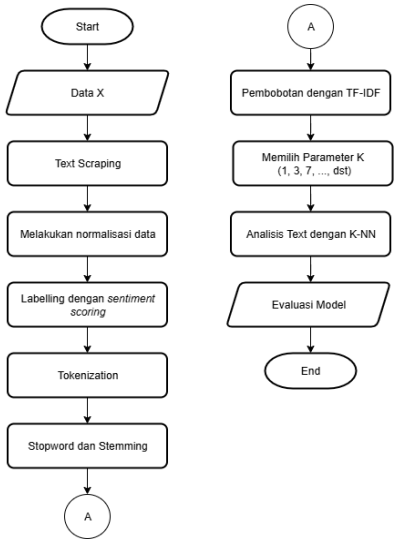
Penelitian ini menggunakan 500 komentar dari platform media sosial X yang memuat kata kunci terkait isu perselingkuhan dengan hashtag #perselingkuhan. Pengambilan data dilakukan melalui metode web scraping menggunakan API atau tools seperti Python Tweepy dan Snscrape. Data yang dikumpulkan diharapkan representatif dan relevan untuk dijadikan bahan analisis sentimen masyarakat terhadap isu perselingkuhan.

A. Pra-Pemrosesan Data (Preprocessing)

Melakukan tahapan preprocessing pada data teks, yang mencakup case folding (mengubah semua huruf menjadi huruf kecil), tokenisasi (memisahkan teks menjadi kata-kata), stopword removal (menghapus kata-kata yang tidak memiliki makna penting), dan stemming (mengembalikan kata ke bentuk dasarnya). Setelah itu, data teks diubah menjadi representasi numerik menggunakan metode TF-IDF (Term Frequency–Inverse Document Frequency) agar dapat digunakan sebagai input dalam proses klasifikasi K-Nearest Neighbor (KNN).

B. Implementasi Metode K-Nearest Neighbor (KNN)

Menerapkan algoritma K-Nearest Neighbor (KNN) untuk melakukan klasifikasi sentimen komentar di platform X menjadi dua kategori, yaitu positif dan negatif. Proses klasifikasi dilakukan dengan mengukur jarak kemiripan antar data dalam ruang fitur yang telah direpresentasikan secara numerik menggunakan TF-IDF. Data uji kemudian diklasifikasikan berdasarkan kelas mayoritas dari k tetangga terdekat, sehingga setiap komentar dapat diprediksi sentimennya secara akurat. Adapun perancangan diagram alir (*flowchart*) dari sistem analisis sentimen masyarakat terhadap isu perselingkuhan menggunakan algoritma K-Nearest Neighbor (KNN) dapat dilihat pada Gambar berikut.



Gambar 2 Flowchart Algoritma K-NN

Berdasarkan Gambar di atas, tahapan analisis sentimen masyarakat pada *platform* media sosial X terhadap isu perselingkuhan menggunakan metode K-Nearest Neighbor (K-NN) adalah sebagai berikut:

1. Mulai (*Start*)

   Proses dimulai dengan tahap awal yaitu pengumpulan data *tweet* dari *platform* media sosial X.
2. Data *Tweet*

   Pada tahap *input* awal, data *tweet* dikumpulkan menggunakan *Twitter Streaming* API. Pencarian akan dilakukan menggunakan *query* berdasarkan kata kunci. Sebagai contoh kata yang berkaitan dengan tweet perselingkuhan yakni “selingkuh”, “perselingkuhan”, “diselingkuhin”, “pelakor”.
3. *Text Scraping*

   Pada tahap ini dilakukan pengambilan data mentah berupa teks *tweet* berdasarkan kata kunci (*keyword*) yang mengandung opini publik

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

mengenai isu perselingkuhan menggunakan teknik *scraping*.

4. Melakukan Normalisasi Data

   Proses normalisasi dilakukan untuk menyamakan format teks, seperti mengubah seluruh huruf menjadi huruf kecil, menghapus tanda baca, angka, dan karakter khusus lainnya.

5. *Labelling* dengan *Sentiment Scoring*

   Setelah dinormalisasi, data diberi label sentimen (positif dan negatif) berdasarkan nilai *sentiment scoring*.

6. *Tokenization*

   Teks kemudian dipisahkan menjadi token (kata-kata individu).

7. *Stopwords* dan *Stemming*

   Kata-kata umum yang tidak memiliki makna penting (seperti "yang", "dan", "di") dihapus untuk meningkatkan kualitas analisis dan dilakukan *stemming* untuk mengembalikan kata ke bentuk dasarnya.

8. Pembobotan dengan TF-IDF

   Setelah *preprocessing* selesai, dilakukan pembobotan kata dengan metode TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk mengetahui seberapa penting sebuah kata dalam dokumen.

9. Memilih Parameter K

   Nilai parameter K ditentukan (misalnya 1, 3, 7, dan seterusnya) sebagai jumlah tetangga terdekat yang akan digunakan dalam proses klasifikasi K-NN.

10. Analisis Teks dengan K-NN

    Data yang telah dibobot kemudian dianalisis menggunakan algoritma K-Nearest Neighbor untuk mengklasifikasikan sentimen *tweet*.

11. Evaluasi Model

    Model yang telah dibangun dievaluasi menggunakan metrik evaluasi seperti akurasi, *presisi*, *recall*, dan *f1-score* untuk mengetahui performa dari klasifikasi yang dilakukan.

12. Selesai (*End*)

    Proses analisis selesai dan sistem siap untuk digunakan kembali untuk dataset lain jika diperlukan.

C. Evaluasi dan Pengujian

Melakukan evaluasi kinerja model klasifikasi K-Nearest Neighbor (KNN) dengan menggunakan metrik seperti akurasi, precision, recall, dan F1-score. Evaluasi ini bertujuan untuk menilai seberapa efektif model dalam mengklasifikasikan sentimen komentar masyarakat di platform X terhadap isu perselingkuhan, sehingga dapat memastikan bahwa metode yang digunakan mampu menghasilkan prediksi yang andal dan representatif.

D. Alat dan Bahan Penelitian

1. Perangkat Keras

Adapun perangkat keras (hardware) yang digunakan dalam penelitian ini berupa laptop dengan spesifikasi yang memadai untuk pengolahan data dan analisis sentimen, meliputi prosesor, kapasitas RAM, ruang penyimpanan, serta kemampuan grafis yang mendukung eksekusi program Python, library untuk web scraping, dan algoritma K-Nearest Neighbor (KNN). Dapat dilihat pada tabel 1.

Tabel 1. Analisis Kebutuhan Perangkat Keras

Perangkat Keras	Spesifikasi
CPU	Intel(R) Core i3 1.8 GHz
RAM	8 GB
Memori	500 GB
Kartu Grafis	Intel HD Graphics

2. Perangkat Lunak

Perangkat keras (*software*) yang digunakan pada penelitian ini dapat dilihat lebih detail pada Tabel 2.

Tabel 2. Analisis Kebutuhan Perangkat Lunak

Perangkat Lunak	Spesifikasi
Sistem Operasi	Windows 10 Pro 64-bit
Code Editor	Google Colaboratory
Bahasa Pemrograman	Python

3. HASIL DAN PEMBAHASAN

A. Analisis Data

Pada penelitian ini digunakan sebanyak 512 data tweet yang diambil dari platform Twitter (X) menggunakan proses scraping melalui API Twitter dengan menggunakan hashtag #perselingkuhan sebagai kata kunci pencarian. Data yang diperoleh kemudian melalui tahap preprocessing untuk memastikan kualitas dan konsistensi teks sebelum dilakukan analisis. Proses preprocessing tersebut meliputi beberapa langkah, yaitu cleaning (menghapus tanda baca, link, angka, dan karakter

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

tidak relevan), casefolding (mengubah seluruh teks menjadi huruf kecil), tokenizing (memecah kalimat menjadi kata-kata), normalisasi (menyeragamkan kata tidak baku menjadi bentuk baku), stopword removal (menghapus kata umum yang tidak memiliki makna penting), serta stemming (mengembalikan kata ke bentuk dasarnya). Tahapan ini dilakukan untuk menghasilkan data teks yang siap dianalisis lebih lanjut menggunakan metode klasifikasi sentimen. Setelah tahap preprocessing selesai, data teks diubah menjadi representasi numerik menggunakan metode TF-IDF (Term Frequency–Inverse Document Frequency) agar dapat diproses oleh algoritma klasifikasi.

B. Representasi Data

Berdasarkan data yang diperoleh, sebanyak 500 data tweet berhasil dikumpulkan dari platform Twitter. Dari keseluruhan data tersebut, sebagian diambil sebagai data representatif untuk menggambarkan proses analisis yang dilakukan. Dalam hal ini, digunakan 5 data sebagai data latih (training data) dan 2 data sebagai data uji (testing data) untuk menunjukkan tahapan pengolahan dan penerapan algoritma K-Nearest Neighbor (KNN). Pemilihan data ini dilakukan secara representatif agar dapat memberikan gambaran yang jelas mengenai bagaimana model bekerja dalam mengklasifikasikan sentimen positif dan negatif pada isu perselingkuhan. Dapat dilihat pada tabel berikut.

Tabel 3. Representasi Data

No	komentar
1	gapapa kalo masi nangisin hubungan yg rusak karena perselingkuhan kann
2	jangan harap bahagia menjalin hubungan dri hasil perselingkuhan ya anjg
3	@dialogsenja__ Pernah memaafkan perselingkuhan karna yakin dia akan berubah ternyata itu hal gila tak masuk logika.
4	@WAPamungkas Jgn gitu budaya kita menunjukan kasih sayang itu dengan figur bapak yang gak ada kasus KDRT yang selalu meningkat dan juga perselingkuhan
5	gak lancar tuh hidup orang yang dukung perselingkuhan karmanya belum datang sekarang emang tapi liat aja nanti
6	Puqi banyak banget kasus pembunuhan ama perselingkuhan makin ga nafsu gua hidup
7	DUH GW GABISA BGT MERASA MENDUKUNG PERSELINGKUHAN ANJRIT

C. Preprocessing Data *Cleaning*

Cleaning data merupakan tahap awal dalam proses preprocessing yang bertujuan untuk membersihkan data teks dari elemen-elemen yang tidak relevan agar hasil analisis menjadi lebih akurat. Pada tahap ini, dilakukan penghapusan tanda baca, angka, *URL, mention* (*@username*), hashtag, emoji, serta spasi berlebih yang tidak memiliki makna penting dalam analisis sentimen. Dengan melakukan cleaning, data teks menjadi lebih bersih, konsisten, dan siap untuk diproses ke tahap selanjutnya.

D. Casefolding

Casefolding merupakan tahap dalam preprocessing data yang bertujuan untuk menyeragamkan bentuk huruf pada teks, yaitu dengan mengubah seluruh karakter menjadi huruf kecil (lowercase). Langkah ini dilakukan agar tidak terjadi perbedaan makna antara kata yang sama tetapi memiliki bentuk huruf berbeda, seperti “Selingkuh” dan “selingkuh” yang dianggap identik. Dengan demikian, casefolding membantu meningkatkan konsistensi data teks sebelum masuk ke tahap berikutnya.

E. Tokenizing

Tokenizing merupakan tahap dalam preprocessing data yang berfungsi untuk memecah teks atau kalimat menjadi satuan kata-kata (token). Tahapan ini penting karena memungkinkan setiap kata dianalisis secara terpisah dalam proses selanjutnya. Misalnya, kalimat “selingkuh merusak kepercayaan” akan diubah menjadi token [“selingkuh”, “merusak”, “kepercayaan”].

F. Normalisasi

Normalisasi merupakan tahap dalam preprocessing data yang bertujuan untuk menyeragamkan kata tidak baku, singkatan, atau ejaan tidak konsisten menjadi bentuk yang baku agar maknanya seragam. Tahapan ini penting karena teks di media sosial sering mengandung kata-kata yang ditulis tidak sesuai kaidah bahasa, seperti penggunaan huruf ganda, singkatan, atau penyingkatan informal. Contohnya, kata “slngkuh” atau “selingkuhh” akan dinormalisasi menjadi “selingkuh”. Dengan adanya proses normalisasi, data teks menjadi lebih bersih, konsisten, dan siap untuk diproses ke tahap berikutnya.

G. Stopword

Stopword merupakan tahap dalam preprocessing data yang berfungsi untuk menghapus kata-kata umum yang tidak memiliki makna penting dalam analisis. Kata-kata seperti “dan”, “yang”, “di”, atau “itu” sering muncul dalam teks tetapi tidak

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

memberikan kontribusi signifikan terhadap konteks atau sentimen. Dengan menghapus stopword, fokus analisis dapat diarahkan pada kata-kata yang benar-benar memiliki makna atau pengaruh terhadap hasil klasifikasi. Proses ini membantu meningkatkan efisiensi dan akurasi dalam analisis teks berikutnya.

H. Stemming

Stemming merupakan tahap dalam preprocessing data yang bertujuan untuk mengembalikan kata ke bentuk dasarnya (*root word*) dengan cara menghapus imbuhan seperti awalan, akhiran, sisipan, maupun gabungan keduanya. Proses ini penting agar kata-kata yang memiliki makna sama dapat dianggap sebagai satu representasi yang seragam dalam analisis. Misalnya, kata "berselingkuh", "selingkuhannya", dan "menyelingkuhi" semuanya akan diubah menjadi bentuk dasar "selingkuh". Dengan melakukan stemming, jumlah variasi kata dapat dikurangi sehingga memudahkan proses analisis dan meningkatkan akurasi pada tahap klasifikasi sentimen.

I. Pelabelan Data

Pada tahap ini, data yang telah melewati proses preprocessing selanjutnya akan dilakukan pelabelan sentimen dengan menggunakan pendekatan lexicon-based. Metode ini bekerja dengan mencocokkan setiap kata dalam teks dengan daftar kata yang telah memiliki nilai atau skor sentimen, baik positif maupun negatif. Kata yang mengandung makna positif, seperti "bahagia" atau "lega", akan diberi label positif, sedangkan kata dengan makna negatif, seperti "marah" atau "selingkuh", akan diberi label negatif. Proses pelabelan ini bertujuan untuk mengelompokkan teks berdasarkan polaritas emosinya sehingga data siap digunakan dalam tahap analisis dan klasifikasi menggunakan algoritma KNN. Hasil pelabelan data dapat dilihat pada tabel berikut.

Tabel 4. Pelabelan data

Label	Hasil Preprocessing Data
Data Latih	
Negatif	['nangis', 'hubung', 'rusak', 'selingkuh']
Negatif	['harap', 'bahagia', 'jalin', 'hubung', 'hasil', 'selingkuh', 'anjing']
Negatif	['maaf', 'selingkuh', 'ubah', 'gila', 'masuk', 'logika']
Positif	['budaya', 'tunjuk', 'kasih', 'sayang', 'figur', 'keras', 'tingkat', 'selingkuh']
Positif	['lancar', 'hidup', 'orang', 'dukung', 'selingkuh', 'karma', 'lihat']
Data Uji	
?	['puqi', 'bunuh', 'selingkuh', 'nafsu', 'hidup']
?	['aduh', 'dukung', 'selingkuh', 'anjing']

J. Pembobotan TF - IDF

Pada tahap ini dilakukan proses pembobotan menggunakan metode TF-IDF (Term Frequency Inverse Document Frequency) untuk menentukan tingkat kepentingan setiap kata dalam kumpulan dokumen. Perhitungan dimulai dengan menentukan nilai TF (Term Frequency) dan DF (Document Frequency), di mana TF menunjukkan seberapa sering suatu kata muncul dalam dokumen, sedangkan DF menunjukkan jumlah dokumen yang memuat kata tersebut. Selanjutnya dihitung nilai IDF (Inverse Document Frequency) dengan memperhatikan total jumlah dokumen dan jumlah dokumen yang mengandung kata tertentu, sehingga kata yang jarang muncul akan memiliki bobot lebih tinggi dibandingkan kata yang sering muncul. Hasil akhir berupa bobot TF-IDF diperoleh dari perkalian antara nilai TF dan IDF, yang merepresentasikan tingkat kepentingan suatu kata dalam dokumen terhadap keseluruhan korpus data.

K. Menentukan Nilai TF dan DF

Pada tahap ini, dilakukan perhitungan nilai Term Frequency (TF) dan Document Frequency (DF) sebagai langkah awal dalam proses pembobotan TF-IDF. Nilai TF menggambarkan seberapa sering suatu kata muncul dalam satu dokumen (tweet), sedangkan nilai DF menunjukkan jumlah dokumen yang mengandung kata tersebut di seluruh kumpulan data. Perhitungan TF dan DF ini bertujuan untuk mengetahui tingkat kemunculan dan penyebaran setiap kata sehingga dapat memberikan gambaran awal mengenai kata-kata yang paling dominan dan relevan dalam data tweet yang dianalisis. Tahapan ini menjadi dasar penting sebelum menghitung nilai IDF dan bobot akhir TF-IDF yang akan digunakan dalam proses klasifikasi sentimen.

L. Menghitung Nilai TF - IDF

Setelah nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) diperoleh, langkah selanjutnya adalah menghitung nilai *TF-IDF* untuk setiap kata pada masing-masing dokumen. Perhitungan *TF-IDF* dilakukan dengan mengalikan nilai TF dengan nilai IDF dari setiap kata. Nilai ini digunakan untuk menentukan seberapa penting suatu kata dalam sebuah dokumen dibandingkan dengan keseluruhan dokumen dalam korpus. Semakin tinggi nilai *TF-IDF*, maka semakin besar tingkat kepentingan kata tersebut dalam merepresentasikan isi dokumen. Tahap ini bertujuan untuk menghasilkan representasi numerik dari teks yang siap digunakan sebagai masukan pada proses klasifikasi menggunakan algoritma K-Nearest Neighbor (KNN). Rumus yang digunakan dalam proses perhitungan *TF-IDF* adalah sebagai berikut.

$$W = TF \times IDF$$

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

Keterangan:
W : bobot dokumen ke-d terhadap kata ke-t
TF : jumlah kata dalam dokumen yang dicari
IDF : *Inversed Document Frequency*
Berikut adalah sampel dalam menerapkan rumus tersebut pada data pertama :
$W = TF \times IDF = 1 \times 1{,}602 = 1{,}602$

M. Normalisasi Data

Setelah diperoleh nilai bobot TF-IDF untuk setiap kata dalam dokumen, tahap selanjutnya adalah melakukan normalisasi data. Proses ini bertujuan untuk menyamakan skala nilai antar fitur agar tidak ada atribut yang mendominasi dalam perhitungan jarak pada algoritma klasifikasi. Normalisasi dilakukan dengan mengubah nilai bobot TF-IDF ke dalam rentang 0–1, sehingga seluruh data berada pada skala yang seragam. Dengan demikian, proses analisis dan klasifikasi dapat berjalan lebih optimal dan menghasilkan hasil yang lebih akurat.
Keterangan:
d : dokumen ke-d
t : kata ke-t dari kata kunci
TF : jumlah kata dalam dokumen yang dicari
Sampel dalam menerapkan rumus tersebut, diantaranya :

$$TF_{norm}(1,1) = \frac{TF(t,d)}{\sqrt{\sum_i (TF(t,d))^2}} = \frac{1{,}602}{\sqrt{(1{,}602)^2 + (1{,}426)^2 \ldots + (1{,}602)^2}} = \frac{1{,}602}{\sqrt{89{,}9986}} = 0{,}168873$$

N. Klasifikasi Algoritma K-Nearest Neighbor (KNN)

Proses klasifikasi dengan algoritma K-Nearest Neighbor (KNN) dimulai dengan menghitung jarak antar data menggunakan Euclidean Distance untuk mengetahui kemiripan data uji dengan data latih. Selanjutnya, jarak tersebut diurutkan, dan dipilih k tetangga terdekat sebagai acuan. Terakhir, dilakukan voting kelas mayoritas, di mana kelas yang paling sering muncul di antara tetangga terdekat ditetapkan sebagai hasil klasifikasi sentimen, baik positif maupun negatif.

O. Menghitung Jarak dengan *Euclidean Distance*

Setelah teks diubah menjadi vektor numerik, jarak antara teks uji dan teks latih dihitung menggunakan *Euclidean Distance* untuk menilai tingkat kemiripan. Jarak yang lebih kecil menunjukkan kesamaan yang lebih tinggi, yang akan digunakan dalam menentukan tetangga terdekat untuk klasifikasi sentimen. Adapun rumus yang digunakan dalam menghitung jarak menggunakan *Euclidean Distance* adalah sebagai berikut.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Di mana:
x_i dan y_i = elemen ke -i dari vektor teks latih dan teks uji.
n = jumlah elemen (dimensi) dalam vektor.

1. Data Uji 1

$$d(1,1) = \sqrt{\begin{matrix}(0{,}1689-0)^2 + (0{,}1503-0)^2 + (0{,}1689-0)^2 + \\ (0{,}1054-0{,}1054)^2 + \cdots + (0-0{,}1689)^2 + (0-0{,}1689)^2 + (0-0)^2\end{matrix}}$$

$$d(2,1) = \sqrt{\begin{matrix}(0-0)^2 + (0{,}1503-0)^2 + (0-0)^2 + (0{,}1054-0{,}1054)^2 + \cdots \\ +(0-0{,}1689)^2 + (0-0{,}1689)^2 + (0-0)^2\end{matrix}}$$

$$d(3,1) = \sqrt{\begin{matrix}(0-0)^2 + (0-0)^2 + (0-0)^2 + (0{,}1054-0{,}1054)^2 \\ + \cdots + (0-0{,}1689)^2 + (0-0{,}1689)^2 + (0-0)^2\end{matrix}}$$

$$d(4,1) = \sqrt{\begin{matrix}(0-0)^2 + (0-0)^2 + (0-0)^2 + (0{,}1054-0{,}1054)^2 \\ + \cdots + (0-0{,}1689)^2 + (0-0{,}1689)^2 + (0-0)^2\end{matrix}}$$

$$d(5,1) = \sqrt{\begin{matrix}(0-0)^2 + (0-0)^2 + (0-0)^2 + (0{,}1054-0{,}1054)^2 \\ + \cdots + (0-0{,}1689)^2 + (0-0{,}1689)^2 + (0-0)^2\end{matrix}}$$

$d(1,1) = \sqrt{0{,}187778} = 0{,}433334$
$d(2,1) = \sqrt{0{,}267408} = 0{,}517115$
$d(3,1) = \sqrt{0{,}250739} = 0{,}500738$
$d(4,1) = \sqrt{0{,}307775} = 0{,}554775$
$d(5,1) = \sqrt{0{,}222221} = 0{,}471403$

2. Pengurutan Jarak

Setelah semua jarak antara data uji dan data latih dihitung, nilai-nilai jarak tersebut kemudian diurutkan dari yang paling kecil hingga terbesar. Jarak terkecil menunjukkan data latih yang paling mirip dengan data uji, sehingga memiliki pengaruh terbesar dalam penentuan kelas. Proses pengurutan ini bertujuan untuk memudahkan pemilihan k tetangga terdekat, yaitu sejumlah data latih dengan kemiripan tertinggi yang akan digunakan sebagai dasar dalam klasifikasi sentimen data uji.

Tabel berikut memperlihatkan hasil pengurutan jarak antara data uji 1 dengan seluruh data latih, di mana jarak terkecil menunjukkan data latih yang paling mirip dan menjadi acuan utama dalam proses klasifikasi sentimen.

Tabel 5. Pengurutan Jarak Data Uji 1

Dokumen	Jarak	Label
D1	0,433334	Negatif
D5	0,471403	Positif
D3	0,500738	Negatif
D2	0,517115	Negatif
D4	0,554775	Positif

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

Tabel berikut memperlihatkan hasil pengurutan jarak antara data uji 2 dengan seluruh data latih, di mana jarak terkecil menunjukkan data latih yang paling mirip dan menjadi acuan utama dalam proses klasifikasi sentimen.

Tabel 6. Pengurutan Jarak Data Uji 2

Dokumen	Jarak	Label
D1	0,39158	Negatif
D2	0,433334	Negatif
D5	0,433334	Positif
D3	0,465077	Negatif
D4	0,522812	Positif

3. Pemilihan *k* Tetangga Terdekat

Dalam penelitian ini, parameter k ditetapkan sebesar 3, yang berarti hanya tiga data latih terdekat yang akan dipertimbangkan dalam proses klasifikasi. Ketiga tetangga ini mewakili data latih yang paling mirip dengan data uji berdasarkan jarak Euclidean, sehingga memiliki pengaruh terbesar dalam menentukan kelas sentimen. Pemilihan k yang tepat sangat penting karena memengaruhi akurasi model.

Tabel berikut menunjukkan hasil pemilihan tiga data latih terdekat (k = 3) untuk data uji 1, yang akan digunakan sebagai acuan dalam menentukan kelas sentimen berdasarkan kemiripan jarak.

Tabel 7. Tetangga terdekat (k=3) data uji 1

Dokumen	Jarak	Label
D1	0,433334	Negatif
D5	0,471403	Positif
D3	0,500738	Negatif
D2	0,517115	Negatif
D4	0,554775	Positif

Tabel berikut menunjukkan hasil pemilihan tiga data latih terdekat (k = 3) untuk data uji 2, yang akan digunakan sebagai acuan dalam menentukan kelas sentimen berdasarkan kemiripan jarak.

Tabel 8. Tetangga terdekat (k=3) data uji 2

Dokumen	Jarak	Label
D1	0,39158	Negatif
D2	0,433334	Negatif
D5	0,433334	Positif
D3	0,465077	Negatif
D4	0,522812	Positif

4. Voting Kelas Mayoritas

Setelah menentukan tiga tetangga terdekat (k = 3), langkah berikutnya adalah menetapkan kelas data uji berdasarkan mayoritas kelas dari tetangga tersebut. Proses ini dilakukan dengan melakukan voting, di mana kelas yang paling sering muncul di antara ketiga tetangga terpilih akan menjadi kelas akhir data uji. Tabel berikut menampilkan hasil voting kelas mayoritas untuk kedua data uji, yang menunjukkan bagaimana model KNN menentukan sentimen berdasarkan kemiripan dengan data latih.

Tabel 9. Hasil Klasifikasi Algoritma KNN

Dokumen	Hasil Voting Label
Data Uji 1	Negatif
Data Uji 2	Negatif

5. Penerapan Algoritma KNN Pada Python

Pada tahap ini, akan disajikan kode program Python yang digunakan untuk mendukung proses analisis data. Kode-kode tersebut dijalankan melalui platform Google Colab, yang memudahkan eksekusi skrip, visualisasi hasil, dan pengujian model secara interaktif. Penyajian potongan kode ini bertujuan agar pembaca dapat memahami alur pemrosesan data dan implementasi algoritma secara lebih jelas.

Gambar di bawah menampilkan hasil akurasi, presisi, recall, dan F1-score dari pengujian algoritma K-Nearest Neighbor (KNN). Metode evaluasi ini memberikan gambaran menyeluruh mengenai kinerja model, di mana akurasi menunjukkan tingkat ketepatan keseluruhan, presisi menilai ketepatan prediksi untuk setiap kelas, recall mengukur kemampuan model menangkap semua data relevan, dan F1-score menyatukan kedua metrik tersebut untuk memberikan nilai evaluasi yang seimbang.

```
print(mt.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

     Negatif       0.91      1.00      0.96        85
     Positif       1.00      0.11      0.20         9

    accuracy                           0.91        94
   macro avg       0.96      0.56      0.58        94
weighted avg       0.92      0.91      0.88        94
```

Gambar 3. Evaluasi Model

```
accuracy = accuracy_score(y_test, y_pred)
print("Akurasi: {:.2f}%".format(accuracy * 100))
```

```
Akurasi: 91.49%
```

Gambar 4. Akurasi Model

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17 No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

Tahap berikutnya yaitu menghitung secara manual nilai akurasi, presisi, recall, dan F1-score sebagai langkah evaluasi terhadap kinerja model. Proses ini bertujuan untuk menilai tingkat ketepatan, konsistensi, serta kemampuan model dalam mengklasifikasikan sentimen tweet secara menyeluruh. Melalui perhitungan ini, dapat diketahui seberapa baik model mengenali pola data dan membedakan antar kelas sentimen dengan benar, sehingga hasil analisis yang diperoleh menjadi lebih valid dan reliabel.

1. Akurasi

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Total\ Data} \times 100\% = \frac{1+85}{94} \times 100\% = \frac{86}{94} \times 100\% = 91{,}49\%$$

2. Presisi

$$Presisi = \frac{TP}{(TP + FP)}$$

a. Negatif

$$Presisi = \frac{TP}{(TP + FP)} = \frac{85}{(85 + 8)} = 91{,}4\%$$

b. Positif

$$Presisi = \frac{TP}{(TP + FP)} = \frac{1}{(1 + 0)} = 100\%$$

3. Recall

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

a. Negatif

$$Recall = \frac{TP}{(TP + FN)} = \frac{85}{(85 + 0)} = 100\%$$

b. Positif

$$Presisi = \frac{TP}{(TP + FP)} = \frac{1}{(1 + 8)} = 11{,}11\%$$

4. F1-Score

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

a. Negatif

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} = 2 \times \frac{(0{,}91 \times 1)}{(0{,}91 + 1)} = 95{,}5\%$$

b. Positif

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} = 2 \times \frac{(1 \times 0{,}11)}{(1 + 0{,}11)} = 20\%$$

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma K-Nearest Neighbors (KNN) terbukti efektif dalam menganalisis sentimen masyarakat terhadap isu perselingkuhan pada platform media sosial X. Penerapan KNN mencakup beberapa tahap, mulai dari preprocessing data teks yang meliputi cleaning, casefolding, tokenizing, normalisasi, penghapusan stopword, dan stemming, konversi teks menjadi representasi numerik menggunakan TF-IDF, perhitungan jarak Euclidean antara data uji dan data latih, hingga penentuan kelas sentimen berdasarkan mayoritas dari k tetangga terdekat. Hasil evaluasi menunjukkan bahwa model mampu membedakan sentimen positif dan negatif dengan akurasi yang tinggi, memberikan gambaran yang jelas mengenai persepsi publik terhadap isu yang sensitif ini. Dengan demikian, algoritma KNN dapat dikatakan sebagai metode yang andal, efisien, dan praktis untuk memahami opini masyarakat dari data teks dalam jumlah besar di media sosial, sekaligus memiliki potensi untuk diterapkan pada analisis sentimen isu-isu sosial lainnya.

Berdasarkan hasil evaluasi model K-Nearest Neighbors (KNN) dengan k = 3, dapat disimpulkan bahwa algoritma ini menunjukkan kinerja yang baik dalam mengklasifikasikan sentimen masyarakat terhadap isu perselingkuhan pada platform media sosial X, dengan akurasi keseluruhan sebesar 91,49% yang menandakan sebagian besar prediksi sesuai dengan label sesungguhnya. Analisis lebih lanjut menunjukkan bahwa model mampu mengidentifikasi sentimen negatif dengan sangat akurat (precision : 0,91, recall : 1,00, F1-score : 0,96), sementara performa pada sentimen positif relatif rendah (precision : 1,00, recall : 0,11, F1-score : 0,20) Dengan demikian, KNN terbukti menjadi metode yang andal, efisien, dan mampu memberikan gambaran yang jelas mengenai persepsi masyarakat, serta potensial untuk digunakan dalam analisis opini publik pada isu sensitif lainnya di media sosial.

5. REFERENSI

Dewo, B. T. (2022). *Analisis Sentimen Twitter Terhadap Ibu Kota Pindah Dengan Perbandingan Metode Klasifikasi K-Nearest Neighbor, Decision Tree, Dan Support Vector Machine*.

Dharmawan, L. R., Arwani, I., & Ratnawati, D. E. (2020). Analisis Sentimen Pada Sosial Media Twitter Terhadap Layanan Sistem Informasi Akademik Mahasiswa Universitas Brawijaya Dengan Metode K- Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *4*(3), 959–965.

Fachrudin, M. F., Angkoso, C. V., & Fatah, D. A. (2024). *Analisis Sentimen Pada Sosial Media Twitter Terhadap Kualitas Jaringan Internet Telkomsel Menggunakan Ensemble K-Nearest Neighbour -Support Vector Machine Sentiment Analysis On Twitter Social Media On Telkomsel ' S Internet Network Quality Using Ensemble . 11*(6), 1253–1264. Https://Doi.Org/10.25126/Jtiik.2024118713

Gibran, M. K., Rifki, M. I., Hasugian, A. H., Siahaan, A. T. A. A., Sahputra, A., & Ong, R. (2024). Sentiment Analysis of Platform X Users on Starlink Using Naive Bayes. Instal: Jurnal Komputer, 16(3), 210-220. https://scholar.archive.org/work/te4arwrevjhhdg47urjqwoha2y/access/wayback/https://journalinstal.cattleyadf.org/index.php/Instal/article/ download/8/8

Halimi, A., Kusrini, K., & Arief, M. R. (2021). Analisis Sentimen Masyarakat Indonesia Terhadap Pembelajaran Online Dari Di Media Sosial Twitter Menggunakan Lexicon Dan K-Nearest Neighbor. *Coreai: Jurnal Kecerdasan Buatan, Komputasi Dan Teknologi Informasi*, *2*(1), 18–28. Https://Doi.Org/10.33650/Coreai.V2i1.2283

Khoirunnisa, C. S., Tukiyat, & Anggai, S. (2024). Analisis Sentimen Opini Masyarakat Terhadap Pemilu 2024 Melalui Media Sosial X Dengan Menggunakan Naive Bayes, K-Nearest Neighbor Dan Decision Tree. *Jurnal Ilmu Komputer*, *2*.

Lestari, D. A., & Mahdiana, D. (2021). Penerapan Algoritma K-Nearest Neighbor Pada Twitter Untuk Analisis Sentimen Masyarakat Terhadap Larangan Mudik 2021. *Informatik : Jurnal Ilmu Komputer*, *17*(2), 123. Https://Doi.Org/10.52958/Iftk.V17i2.3629

Pamungkas, F. S., & Kharisudin, I. (2021). Analisis Sentimen Dengan Svm, Naive Bayes Dan Knnuntuk Studi Tanggapan Masyarakat Indonesia Terhadappandemi Covid-19 Pada Media Sosial Twitter. *Prosiding Seminar Nasional Matematika*, *4*, 1–7.

Saidah, S., & Mayary, J. (2020). Analisis Sentimen Pengguna Twitter Terhadap Dompet Elektronik Dengan Metode Lexicon Based Dan K – Nearest Neighbor. *Jurnal Ilmiah Informatika Komputer*, *25*(1), 1–17. Https://Doi.Org/10.35760/Ik.2020.V25i1.2411