Jurnal Penelitian Inovatif (JUPIN)
Vol. 6, No. 2, Mei 2026, Hal. 1623-1632
https://jurnal-id.com/index.php/jupin

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

Analisis Sentimen Publik terhadap Sinetron Asmara Gen Z di Platform X Menggunakan Metode Naive Bayes dan Support Vector Machine

Ida Hidayati[1], Elis Nurwasih[2] , Muhammad Ilham Mudya Riyadi[3], Cindy Gabriel Boro[4]

[1,2,3,4] Teknik Informatika, Fakultas Teknik, Universitas Papua, Indonesia
Email: [1]202365059@student.unipa.ac.id, [2]202365063@student.unipa.ac.id,
[3]202365081@student.unipa.ac.id, [4]202365069@student.unipa.ac.id

Abstrak

Perkembangan media sosial telah menghasilkan data opini publik dalam jumlah besar yang berpotensi dianalisis untuk memahami persepsi masyarakat terhadap konten hiburan. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap sinetron Asmara Gen Z di platform X menggunakan metode Naive Bayes dan membandingkan kinerjanya dengan Support Vector Machine (SVM). Data dikumpulkan melalui proses crawling menggunakan library snscrape dengan beberapa kata kunci terkait, menghasilkan 1.116 cuitan. Tahapan preprocessing yang diterapkan meliputi cleaning, normalisasi, case folding, tokenizing, stopword removal, dan stemming menggunakan library Sastrawi. Pelabelan sentimen dilakukan secara otomatis menggunakan pendekatan lexicon-based sentiment analysis dengan sistem custom scoring berbasis bobot polaritas kata (skor 2 untuk positif, 1 untuk negatif, 0 untuk netral) yang mengklasifikasikan data ke dalam tiga kategori: positif, negatif, dan netral. Representasi fitur teks dilakukan menggunakan metode TF-IDF. Dataset dibagi menjadi 80% data latih dan 20% data uji. Hasil evaluasi menunjukkan bahwa metode SVM menghasilkan akurasi 85,27% dengan F1-score 84,93%, sedangkan Naive Bayes menghasilkan akurasi 79,37% dengan F1-score 78,53%. SVM terbukti lebih unggul dibandingkan Naive Bayes dalam klasifikasi sentimen pada data media sosial, khususnya pada kelas sentimen negatif, karena kemampuannya menangani variasi bahasa informal yang lebih tinggi.

Kata kunci: *analisis sentimen, asmara gen z, naive bayes, support vector machine, twitter.*

Abstract

The rapid growth of social media has generated large volumes of public opinion data with significant potential for analyzing societal perceptions of entertainment content. This study aims to analyze public sentiment toward the Indonesian soap opera Asmara Gen Z on the X platform using the Naive Bayes method and to compare its performance with Support Vector Machine (SVM). Data were collected through a crawling process using the snscrape library based on relevant keywords, yielding 1,116 tweets. Preprocessing steps included cleaning, normalization, case folding, tokenizing, stopword removal, and stemming using the Sastrawi library. Sentiment labeling was performed automatically using a lexicon-based sentiment analysis approach with a custom scoring system based on lexical polarity weights (score 2 for positive, 1 for negative, 0 for neutral), classifying data into three categories: positive, negative, and neutral. Text feature representation was carried out using the TF-IDF method. The dataset was split into 80% training data and 20% testing data. Evaluation results showed that SVM achieved an accuracy of 85.27% with an F1-score of 84.93%, while Naive Bayes achieved an accuracy of 79.37% with an F1-score of 78.53%. SVM demonstrated superior performance compared to Naive Bayes in sentiment classification on social media data, particularly for negative sentiment, due to its stronger ability to handle informal language variation classification.

Keywords: *analisis sentimen, asmara gen z, naive bayes, support vector machine, twitter*

1. PENDAHULUAN

Akses masyarakat terhadap berbagai bentuk hiburan berubah drastis seiring percepatan teknologi informasi dan komunikasi. Sinetron tetap menjadi salah satu konten hiburan paling diminati di Indonesia, terus bertransformasi mengikuti selera pemirsa. Berdasarkan data Nielsen Television Audience Measurement (2023), genre sinetron dan drama lokal secara konsisten menempati rating tertinggi di televisi nasional dengan rata-rata pangsa penonton lebih dari 35% pada slot prime time.

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

Wibowo & Harmonis (2024) mencatat bahwa sinetron tidak semata-mata menghibur, tetapi juga menyampaikan nilai sosial dan budaya yang membentuk cara berpikir publik.

Media sosial kini menjelma menjadi wadah utama ekspresi dan berbagi pengalaman. Platform X (Twitter) khususnya, berfungsi sebagai ruang diskusi publik yang bergerak cepat dengan beragam topik yang mengalir secara langsung. Pemirsa televisi, termasuk penggemar sinetron, dengan mudah melontarkan tanggapan melalui platform ini (Luttrell, 2025).

Data tidak terstruktur dalam jumlah besar lahir dari interaksi media sosial. Opini, emosi, dan persepsi pengguna terhadap suatu fenomena tersimpan di dalamnya. Pengolahan data tersebut menjadi informasi berharga membutuhkan teknik analisis yang tepat (Wankhade et al., 2022).

Salah satu sinetron yang menyedot perhatian publik belakangan ini ialah Asmara Gen Z. Kisah hidup generasi muda dengan konflik-konflik relevan diangkat dalam tayangan ini. Pujian dan kritik mengalir dari penonton sebagai respons. Cuitan-cuitan di platform X menyimpan beragam opini tersebut, menjadikannya sumber data potensial untuk kajian lebih dalam.

Bidang Natural Language Processing (NLP) mengenal analisis sentimen sebagai teknik pengidentifikasi dan pengklasifikasi opini dalam teks ke dalam tiga kategori: positif, negatif, atau netral (Lu et al., 2024). Penerapan teknik ini menuntut metode klasifikasi yang mampu mengolah data teks secara efektif dan efisien.

Naive Bayes termasuk metode yang populer. Kesederhanaan, kecepatan proses, dan kemampuannya menangani data besar menjadi keunggulan utama. Pendekatan probabilistik dengan asumsi independensi antar fitur menjadi dasar kerja metode ini, cocok untuk data teks media sosial.

Perbandingan dengan metode lain diperlukan untuk hasil yang lebih komprehensif. Support Vector Machine (SVM) hadir sebagai pembanding dengan performa tinggi. Kemampuan klasifikasi SVM pada data berdimensi tinggi lebih baik, plus sanggup menemukan hyperplane pemisah antar kelas yang optimal. Sejumlah penelitian membuktikan akurasi SVM lebih tinggi dari Naive Bayes, terutama pada dataset kompleks.

Karakteristik kedua metode berbeda secara fundamental. Naive Bayes unggul di kecepatan dan efisiensi komputasi. SVM unggul di akurasi dan kemampuan generalisasi model. Perbandingan ini penting karena penelitian ini bertujuan mendeteksi metode mana yang lebih optimal untuk klasifikasi sentimen pada data media sosial seputar sinetron Asmara Gen Z.

Pelabelan data dalam penelitian ini tidak hanya mengandalkan machine learning. Pendekatan lexicon-based sentiment analysis turut diterapkan pada tahap pelabelan. Daftar kata positif dan negatif dimanfaatkan untuk memberi label sentimen secara otomatis pada data cuitan dari platform X. Taboada et al. menjelaskan bahwa pendekatan ini menentukan polaritas sentimen berdasarkan kemunculan kata-kata bernilai sentimen positif atau negatif dalam suatu teks. Proses pelabelan otomatis pada data media sosial terbantu secara efektif oleh metode ini, mengurangi beban pelabelan manual yang memakan waktu.

Uraian di atas mengarah pada satu tujuan utama: menganalisis sentimen publik terhadap sinetron Asmara Gen Z di platform X menggunakan Naive Bayes sekaligus membandingkan kinerjanya dengan SVM. Gambaran jelas tentang persepsi masyarakat serta penentuan metode klasifikasi paling efektif untuk analisis sentimen berbasis data media sosial diharapkan menjadi hasil akhir. Kontribusi bagi pengembangan riset analisis sentimen dan referensi bagi industri hiburan dalam memahami respons audiens juga menjadi luaran yang diharapkan.

Penelitian sebelumnya telah banyak mengeksplorasi Naive Bayes dan SVM dalam analisis sentimen berbahasa Indonesia. Dang et al. (2020) membuktikan Naive Bayes memberikan performa memadai pada data media sosial dengan waktu pelatihan singkat. Rai et al. (2021) menunjukkan SVM secara konsisten unggul akurasi dibanding Naive Bayes pada berbagai dataset teks. Wankhade et al. (2022) dalam kajian komprehensifnya menyimpulkan satu hal: pemilihan metode klasifikasi sangat bergantung pada karakteristik data dan kebutuhan komputasi sistem. Berangkat dari temuan-temuan tersebut, penelitian ini mengisi celah pengetahuan dengan mengaplikasikan kedua metode pada konteks spesifik sinetron Indonesia di platform X.

Praktik standar pengolahan teks berbahasa Indonesia menjadi acuan tahapan preprocessing dalam penelitian ini. Koto et al. (2020) menekankan normalisasi teks dan stemming bahasa Indonesia sebagai

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

langkah kritis penyiapan data media sosial sebelum klasifikasi. Ekstraksi fitur menggunakan TF-IDF dipilih berdasarkan rekomendasi Trstenjak et al. (2020) karena kemampuannya merepresentasikan kepentingan suatu kata dalam konteks dokumen secara keseluruhan. Evaluasi model mengandalkan metrik precision, recall, dan F1-score seperti direkomendasikan Sokolova & Lapalme (2009) agar penilaian performa bersifat komprehensif dan tidak bias terhadap distribusi kelas yang timpang.

2. METODE PENELITIAN

Satu fokus utama penelitian ini menganalisis sentimen publik terhadap sinetron Asmara Gen Z menggunakan data yang diambil dari platform X. Permasalahan yang hendak dipecahkan berupa proses klasifikasi teks untuk menentukan label sentimen dari setiap data tweet, dengan tiga kemungkinan kategori yaitu positif, negatif, dan netral, berdasarkan representasi fitur teks yang telah diekstraksi. Proses klasifikasi secara garis besar bekerja dengan memetakan setiap dokumen teks ke dalam kelas sentimen tertentu menggunakan fitur-fitur yang diambil dari dokumen tersebut (Wankhade et al., 2022; Lu et al., 2024).

Sumber data penelitian adalah cuitan-cuitan yang membahas sinetron Asmara Gen Z. Pengumpulan data dilakukan melalui proses crawling memanfaatkan Google Colab, ditunjang oleh tools berbasis web dan library Python bernama snscrape. Beberapa kata kunci relevan dipakai sebagai panduan pengambilan data, yaitu AGZ, Fatqeel, Asmara Gen Z, Harqeel, Mora, dan Aqeela, dalam rentang waktu Januari 2024 hingga Maret 2024. Hasil dari proses crawling tersebut berjumlah 1.116 data yang kemudian digabungkan dan dijadikan dataset utama dalam penelitian ini.

Seluruh data yang berhasil dikumpulkan selanjutnya menjalani serangkaian tahap pra-pemrosesan atau preprocessing. Tahapan tersebut meliputi cleaning untuk membersihkan data dari elemen tidak perlu, normalisasi untuk menyeragamkan bentuk kata, case folding untuk mengubah semua huruf menjadi huruf kecil, tokenization untuk memecah teks menjadi kata-kata terpisah, stopword removal untuk menghilangkan kata-kata umum yang tidak memiliki makna signifikan, dan stemming untuk mengembalikan kata ke bentuk dasarnya. Proses normalisasi data juga dilakukan secara khusus untuk memperbaiki kata tidak baku, berbagai bentuk singkatan, serta beragam kesalahan penulisan yang kerap muncul dalam data media sosial. Semua tahapan ini bertujuan untuk meningkatkan kualitas data agar terstruktur dengan baik dan siap memasuki fase analisis (Koto et al., 2020; Wankhade et al., 2022).

Pada fase pelabelan data, penelitian ini mengaplikasikan pendekatan lexicon-based sentiment analysis. Pelabelan berlangsung secara otomatis dengan cara menghitung jumlah kemunculan kata positif dan kata negatif dalam setiap cuitan, menggunakan kamus sentimen berbahasa Indonesia yang diadaptasi dari InSet (Indonesian Sentiment Lexicon) yang dikembangkan oleh Tala (2003) dan diperkaya dengan daftar kata gaul media sosial yang dikurasi secara manual. Taboada et al. memaparkan bahwa pendekatan lexicon-based sentiment analysis bekerja dengan menentukan polaritas sentimen berdasarkan daftar kata yang mengandung nilai positif atau negatif dalam suatu teks. Sebuah data akan diberi label positif jika kata-kata positif lebih mendominasi, label negatif jika kata-kata negatif lebih mendominasi, dan label netral jika tidak ada kecenderungan dominasi sentimen yang jelas. Untuk memastikan kualitas pelabelan, dilakukan validasi terhadap sampel acak sebanyak 100 cuitan secara manual oleh dua anotator, menghasilkan tingkat kesepakatan (inter-annotator agreement) sebesar 82%, yang menunjukkan bahwa proses pelabelan otomatis cukup andal. Penelitian ini juga menerapkan sistem custom scoring dengan pemberian bobot berbeda pada setiap kategori: skor 2 untuk sentimen positif, skor 1 untuk sentimen negatif, dan skor 0 untuk sentimen netral. Pemberian bobot ini bertujuan memberikan penekanan lebih pada sentimen positif dalam proses perhitungan agregat, sejalan dengan fokus analisis pada persepsi publik yang bersifat afirmatif terhadap sinetron.

Setelah seluruh tahap pra-pemrosesan selesai dijalankan, analisis korelasi data dilakukan untuk melihat hubungan antar fitur atau antar kata yang muncul dalam dataset. Analisis ini berguna untuk memahami keterkaitan antara berbagai variabel teks sekaligus mengidentifikasi fitur-fitur mana yang memiliki pengaruh terhadap keberhasilan klasifikasi sentimen.

Langkah berikutnya adalah mengubah data ke dalam format numerik menggunakan metode Term Frequency-Inverse Document Frequency atau yang lebih dikenal dengan TF-IDF. Metode ini bekerja dengan memberikan bobot pada setiap kata berdasarkan tingkat kepentingan relatifnya dalam suatu

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

dokumen. Dengan demikian, kata-kata yang lebih relevan terhadap isi dokumen akan memberikan kontribusi lebih besar dalam proses klasifikasi (Trstenjak et al., 2020).

Metode utama yang dipakai dalam penelitian ini adalah Naive Bayes Classifier, yaitu algoritma klasifikasi yang berbasis pada teori probabilitas. Algoritma ini dikenal luas karena kesederhanaan kerjanya dan efisiensinya dalam mengolah data teks dalam jumlah besar. Naive Bayes sering menjadi pilihan dalam berbagai studi analisis sentimen karena mampu memberikan hasil yang memuaskan dengan waktu komputasi yang tergolong cepat (Fatra & Manguma, 2025; Dang et al., 2020).

Sebagai pembanding atau yang sering disebut state of the art, penelitian ini juga mengimplementasikan metode Support Vector Machine atau SVM. Metode ini terkenal memiliki performa tinggi dalam tugas klasifikasi teks, terutama karena kemampuannya menangani data dengan dimensi yang tinggi serta menghasilkan model yang stabil dan tingkat akurasinya baik (Rai et al., 2021; Wankhade et al., 2022).

Pertimbangan memilih Naive Bayes dan SVM sebagai metode yang dibandingkan berangkat dari karakteristik keduanya yang saling melengkapi satu sama lain. Naive Bayes memiliki keunggulan di sisi kecepatan pemrosesan dan efisiensi komputasi. SVM sebaliknya, lebih unggul dalam hal akurasi hasil klasifikasi dan kemampuan generalisasi model terhadap data baru. Perbandingan antara kedua metode ini menjadi penting untuk menentukan mana yang paling optimal ketika diterapkan pada data analisis sentimen dari media sosial.

Untuk mengukur seberapa baik kinerja masing-masing model, digunakan metrik evaluasi berupa akurasi, precision, recall, dan F1-score. Keempat metrik ini dihitung berdasarkan confusion matrix yang merekam jumlah prediksi benar dan salah dari setiap kelas. Penggunaan gabungan metrik ini bertujuan menyajikan gambaran performa model klasifikasi secara menyeluruh dan tidak bias (Sokolova & Lapalme, 2009).

Keseluruhan proses analisis dalam penelitian ini dijalankan dengan bahasa pemrograman Python. Beberapa pustaka atau library yang dimanfaatkan antara lain scikit-learn untuk keperluan machine learning, pandas untuk manipulasi data, dan numpy untuk perhitungan numerik. Rangkaian tahapan penelitian secara berurutan mencakup pengumpulan data, pelaksanaan pra-pemrosesan, analisis korelasi antar fitur, ekstraksi fitur menggunakan TF-IDF, pelatihan model dengan data latih, pengujian model dengan data uji, serta evaluasi akhir terhadap hasil yang diperoleh.

3. HASIL DAN PEMBAHASAN

Penelitian ini melakukan crawling untuk memperoleh data cuitan dari platform X yang berkaitan dengan sinetron Asmara Gen Z. Hasil dan pembahasan dari analisis sentimen yang telah dilaksanakan adalah sebagai berikut.

3.1. Scraping Data

Scraping data merupakan proses pengambilan data cuitan dari platform X menggunakan library snscrape berdasarkan kata kunci tertentu yang berkaitan dengan sinetron Asmara Gen Z.

	K	L	M	N
1	tweet_url	user_id_str	username	full_text_content
2	https://x.c	1,73E+18		yakan anjir lah parah si qeel qeel agz nih
3	https://x.c	1,56E+18		selama ini ternyata kita gak nonton agz guys beda judul
4	https://x.c	1,71E+18		yg pny hati ga cuma team crew dan cast agz ajaa fevers juga punya hati dan ga layak di bully secara psikologis spt itu
5	https://x.c	1,40E+18		bener bener ya moqeel agz udh dibikin semusnah itu sama penulisnya dibuat sejelek mungkin entah itu dari pandangan penonton atau pandangan karakternya jahat banget yaallah diginiin sebegininya jangan sampe bakalan jadi boomerang atau bahkan karma buat penulisnya
6	https://x.c	4,677E+09		iyaaa kannn emoji piyik juga gemes dan mayan gitu loh ada tiga variasi mana warnanya kuning cerah ceria meskipun kadang tidak merepresentasikan dark romance si mora agz itu hahaha
7	https://x.c	2,157E+09		baru sadar aqeela nelfon om leonkayanya sabtu agz beneran ada gong krn keluarga aqeela ngumpul semua
8	https://x.c	1,83E+18		day kesekian nonton agz ya jet
9	https://x.c	1,10E+18		terserah aku giveupppp dengan agz ni sakittttt
10	https://x.c	149459291		ini kalau karam fz sama hq udahan ya yapping agz yaudah walaikumsalam see you in another project yg idola kita sama bes
11	https://x.c	1,87E+18		takutnya karena fq di hype terus penulis malah belokin lagi alurnya bikin f itu unknown bukan h karena kalo baca teorinya kedengaran masuk akal penulis agz kan suka bikin plot twist yg tibatiba
12	https://x.c	1,72E+18		jujur tp sakit bnget anjirr gue jd penonton kalo boleh gue mending masuk agz universe aja join gaspol atau slay queen aj lg happy mereka soalnya biar gatau apa dr pd jd penonton njirr sakit bgt penulis tambahin saya dong biar sekalian gue sadarin tuh satu semuanya

Gambar 1. Data Mentah

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

3.2. Cleaning Data

Langkah awal dalam penelitian ini menjalankan proses cleaning data. Tujuan dari proses ini membersihkan data hasil crawling dari berbagai karakter yang tidak diperlukan, contohnya URL, mention, hashtag, emoji, tanda baca, angka, serta simbol lain yang berpotensi mengganggu jalannya analisis sentimen. Penghapusan data duplikat dan data kosong juga dilakukan pada tahap ini untuk meningkatkan kualitas dataset.

Data mentah dari proses crawling berjumlah 1.116 cuitan seputar sinetron Asmara Gen Z. Data ini masih kaya akan noise karena bersumber dari media sosial yang sifatnya tidak terstruktur. Dengan demikian, proses cleaning menjadi tahap krusial sebelum memasuki preprocessing lanjutan.

N
full_text_content
yakan anjir lah parah si qeel qeel agz nih
selama ini ternyata kita gak nonton agz guys beda judul
yg pny hati ga cuma team crew dan cast agz ajaa fevers juga punya hati dan ga layak di bully secara psikologis spt itu
bener bener ya moqeel agz udh dibikin semusnah itu sama penulisnya dibuat sejelek mungkin entah itu dari pandangan penonton atau pandangan karakternya jahat banget yaallah diginiin sebegininya jangan sampe bakalan jadi boomerang atau bahkan karma buat penulisnya
iyaaa kannn emoji piyik juga gemes dan mayan gitu loh ada tiga variasi mana warnanya kuning cerah ceria meskipun kadang tidak merepresentasikan dark romance si mora agz itu hahaha
baru sadar aqeela nelfon om leonkayanya sabtu agz beneran ada gong krn keluarga aqeela ngumpul semua
day kesekian nonton agz ya jet
terserah aku giveupppp dengan agz ni sakittttt
ini kalau karam fz sama hq udahan ya yapping agz yaudah walaikumsalam see you in another project yg idola kita sama bes
takutnya karena fq di hype terus penulis malah belokin lagi alurnya bikin f itu unknown bukan h karena kalo baca teorinya kedengaran masuk akal penulis agz kan suka bikin plot twist yg tibatiba
jujur tp sakit bnget anjirr gue jd penonton kalo boleh gue mending masuk agz universe aja join gaspol atau slay queen aj lg happy mereka soalnya biar gatau apa dr pd jd penonton njirr sakit bgt penulis tambahin saya dong biar sekalian gue sadarin tuh satu semuanya

Gambar 2. Cleaning Data

3.3. Normalisasi Data

Normalisasi menjadi langkah lanjutan setelah cleaning data selesai dikerjakan. Tujuan dari normalisasi memperbaiki kata-kata tidak baku, singkatan, bahasa gaul, dan kesalahan penulisan yang kerap ditemukan di media sosial. Proses ini menyeragamkan kata-kata dengan makna yang sama, yang pada akhirnya mempermudah jalannya analisis.

Perubahan kata gk menjadi tidak, yg menjadi yang, dan bgt menjadi banget merupakan contoh normalisasi yang dilakukan. Kata-kata dengan penulisan berlebihan juga ikut diperbaiki ke bentuk standar.

	A
1	full_text_normalized
2	ya ampun parah sekali si qeel ini
3	selama ini ternyata kita tidak menonton asmara gen z teman-teman judulnya berbeda
4	yang memiliki hati bukan hanya tim kru dan pemeran asmara gen z saja penggemar juga memiliki hati dan tidak layak dibully secara psikologis
5	benar-benar ya karakter moqeel di asmara gen z sudah dibuat sehancur itu oleh penulisnya dibuat seburuk mungkin baik dari sudut pandang penonton maupun karakter lain jahat sekali jangan sampai ini menjadi bumerang atau karma bagi penulisnya
6	iya kan emoji anak ayam juga lucu dan lumayan ada tiga variasi dengan warna kuning cerah meskipun kadang tidak merepresentasikan sisi gelap romansa mora di asmara gen z
7	baru sadar aqeela menelepon om leon sepertinya hari sabtu di asmara gen z benar-benar ada momen penting karena keluarga aqeela berkumpul
8	hari ke sekian menonton asmara gen z
9	terserah aku sudah menyerah dengan asmara gen z ini menyakitkan
10	jika karam fatahzarah dan harryqeela berakhir ya sudah sampai jumpa di proyek lain yang kita sukai
11	takutnya karena fatqeel terus diangkat penulis malah mengubah alur dan membuat fatah menjadi tidak jelas bukan harry jika membaca teorinya terdengar masuk akal karena penulis asmara gen z suka membuat plot twist mendadak
12	jujur ini sangat menyakitkan sebagai penonton kalau boleh aku lebih memilih masuk ke semesta asmara gen z saja agar tidak merasakan sakit

Gambar 3. Normalisasi Data

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

3.4. Pelabelan dan Custom Scoring

Setelah data melewati tahap normalisasi, proses pelabelan sentimen berjalan secara otomatis dengan pendekatan lexicon-based sentiment analysis. Pada fase ini, setiap cuitan dikelompokkan ke dalam tiga kategori sentimen yaitu positif, negatif, dan netral, berdasarkan kemunculan kata-kata sentimen di dalam teks.

Penelitian ini menerapkan metode custom scoring untuk menentukan label sentimen. Sistem scoring yang dipakai memberi skor 2 untuk sentimen positif, skor 1 untuk sentimen negatif, dan skor 0 untuk sentimen netral. Penentuan skor bertumpu pada ada tidaknya kata positif dan kata negatif yang muncul dalam setiap cuitan.

Sebuah cuitan masuk kategori positif bila mengandung kata seperti bagus, hebat, lucu, senang, atau keren. Sementara itu, cuitan dikategorikan negatif jika memuat kata seperti parah, sedih, kecewa, jahat, atau menyakitkan. Cuitan yang tidak mengandung kata sentimen sama sekali dimasukkan ke dalam kategori netral.

	A	B	C
1	full_text_normalized	score	Label
2	ya ampun parah sekali si qeel ini	1	Negatif
3	selama ini ternyata kita tidak menonton asmara gen z teman-teman judulnya berbeda	0	Netral
4	yang memiliki hati bukan hanya tim kru dan pemeran asmara gen z saja penggemar juga memiliki hati dan tidak layak dibully secara psikologis seperti itu	1	Negatif
5	benar-benar ya karakter moqeel di asmara gen z sudah dibuat sehancur itu oleh penulisnya dibuat seburuk mungkin baik dari sudut pandang penonton maupun karakter lain jahat sekali jangan sampai ini menjadi bumerang atau karma bagi penulisnya	1	Negatif
6	iya kan emoji anak ayam juga lucu dan lumayan ada tiga variasi dengan warna kuning cerah meskipun kadang tidak merepresentasikan sisi gelap romansa mora di asmara gen z	2	Positif
7	baru sadar aqeela menelepon om leon sepertinya hari sabtu di asmara gen z benar-benar ada momen penting karena keluarga aqeela berkumpul semua	0	Netral
8	hari ke sekian menonton asmara gen z	0	Netral
9	terserah aku sudah menyerah dengan asmara gen z ini menyakitkan	1	Negatif
10	jika karam fatahzarah dan harryqeela berakhir ya sudah sampai jumpa di proyek lain yang kita sukai	0	Netral
11	takutnya karena fatqeel terus diangkat penulis malah mengubah alur dan membuat fatah menjadi tidak jelas bukan harry jika membaca teorinya terdengar masuk akal karena penulis asmara gen z suka membuat plot twist mendadak	1	Negatif
12	jujur ini sangat menyakitkan sebagai penonton kalau boleh aku lebih memilih masuk ke semesta asmara gen z saja agar tidak merasakan sakit seperti ini	1	Negatif

Gambar 4. Hasil Pelabelan dan Custom Scoring

3.5. Preprocessing

Penyiapan data teks agar siap diolah oleh algoritma klasifikasi dilakukan melalui tahapan preprocessing. Penelitian ini menjalankan preprocessing yang mencakup case folding, tokenizing, stopword removal, dan stemming. Berikut disajikan data mentah sebelum proses preprocessing berlangsung.

	A	B	C
1	full_text_content	score	Label
2	ya ampun parah sekali si qeel ini	1	Negatif
3	selama ini ternyata kita tidak menonton asmara gen z teman-teman judulnya berbeda	0	Netral
4	yang memiliki hati bukan hanya tim kru dan pemeran asmara gen z saja penggemar juga memiliki hati dan tidak layak dibully secara psikologis seperti itu	1	Negatif
5	benar-benar ya karakter moqeel di asmara gen z sudah dibuat sehancur itu oleh penulisnya dibuat seburuk mungkin baik dari sudut pandang penonton maupun karakter lain jahat sekali jangan sampai ini menjadi bumerang atau karma bagi penulisnya	1	Negatif
6	iya kan emoji anak ayam juga lucu dan lumayan ada tiga variasi dengan warna kuning cerah meskipun kadang tidak merepresentasikan sisi gelap romansa mora di asmara gen z	2	Positif
7	baru sadar aqeela menelepon om leon sepertinya hari sabtu di asmara gen z benar-benar ada momen penting karena keluarga aqeela berkumpul semua	0	Netral
8	hari ke sekian menonton asmara gen z	0	Netral

Gambar 5. Data Mentah

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

3.5.1. Case Folding

Proses mengubah seluruh huruf pada teks ke bentuk huruf kecil dinamakan case folding. Tujuan dari langkah ini menstandarkan penulisan kata, agar kata yang serupa tetapi berbeda penggunaan huruf kapital tidak lagi dianggap sebagai entitas yang terpisah.

Sebagai ilustrasi, tiga variasi tulisan Asmara Gen Z, asmara gen z, dan ASMARA GEN Z bertransformasi menjadi satu bentuk yang sama, yaitu asmara gen z. Dengan begitu, analisis menjadi lebih konsisten sekaligus akurat.

	A
1	full_text_content
2	ya ampun parah sekali si qeel ini
3	selama ini ternyata kita tidak menonton asmara gen z teman-teman judulnya berbeda
4	yang memiliki hati bukan hanya tim kru dan pemeran asmara gen z saja penggemar juga memiliki hati dan tidak layak dibully secara psikologis seperti itu
5	benar-benar ya karakter moqeel di asmara gen z sudah dibuat sehancur itu oleh penulisnya dibuat seburuk mungkin baik dari sudut pandang penonton maupun karakter lain jahat sekali jangan sampai ini menjadi bumerang atau karma bagi penulisnya
6	iya kan emoji anak ayam juga lucu dan lumayan ada tiga variasi dengan warna kuning cerah meskipun kadang tidak merepresentasikan sisi gelap romansa mora di asmara gen z
7	baru sadar aqeela menelepon om leon sepertinya hari sabtu di asmara gen z benar-benar ada momen penting karena keluarga aqeela berkumpul semua
8	hari ke sekian menonton asmara gen z

Gambar 6. Hasil Case Folding

3.5.2. Tokenizing

Pemecahan kalimat menjadi kata-kata tunggal atau token disebut tokenizing. Pemisahan dilakukan berdasarkan spasi, sehingga keluaran dari tahap ini berupa daftar kata individual.

Sebagai ilustrasi, kalimat hari ke sekian menonton asmara gen z bertransformasi menjadi token seperti hari, sekian, menonton, asmara, gen, dan z. Tujuan tokenizing memudahkan proses analisis kata pada tahap berikutnya karena algoritma klasifikasi mengandalkan representasi kata dalam dokumen.

	D
1	Tokenizing
2	[ya, ampun, parah, sekali, si, qeel, ini]
3	[selama, ini ternyata, kita, tidak, menonton, asmara, gen, z, teman-teman, judulnya, berbeda]
4	[yang, memiliki, hati, bukan, hanya, tim, kru,dan, pemeran, asmara, gen, z, saja, penggemar, juga, memiliki, hati, dan, tidak, layak, dibully, secara, psikologis, seperti,itu]
5	[benar-benar, ya, karakter, moqeel, di, asmara, gen, z,sudah, dibuat, sehancur, itu, oleh, penulisnya, dibuat, seburuk, mungkin, baik, dari, sudut, pandang, penonton, maupun, karakter, lain, jahat, sekali, jangan, sampai, ini, menjadi, bumerang, atau, karma, bagi, penulisnya]
6	[iya, kan, emoji, anak, ayam, juga, lucu, dan, lumayan, ada,tiga,variasi, dengan, warna, kuning, cerah, meskipun, kadang, tidak, merepresentasikan, sisi, gelap, romansa, morakl, di, asmara, gen, z]
7	[baru, sadar, aqeela, menelepon, om, leon, sepertinya, hari, sabtu,di,asmara, gen, z, benar-benar, ada, momen, penting, karena, keluarga, aqeela, berkumpul, semua]
8	[hari, ke, sekian, menonton, asmara, gen, z]

Gambar 7. Hasil Tokenizing

3.5.3. Stopword Removal

Stopword removal menghapus kata-kata yang tidak memiliki makna penting dalam analisis sentimen. Kata-kata seperti dan, yang, di, ke, ini, dan itu ikut dihapus karena frekuensi kemunculannya terlalu tinggi namun tidak memberi informasi sentimen yang berarti.

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

Setelah proses penghapusan stopword selesai, data menjadi lebih terfokus pada kata-kata penting yang berkaitan dengan opini pengguna. Tahapan ini berkontribusi pada peningkatan kualitas fitur yang dipakai dalam proses klasifikasi.

	E
1	Stopword Removal
2	['ampun', 'parah', 'qeel']
3	['ternyata', 'menonton', 'asmara', 'gen', 'z', 'teman-teman', 'judulnya', 'berbeda']
4	['memiliki', 'hati', 'tim', 'kru', 'pemeran', 'asmara', 'gen', 'z', 'penggemar', 'memiliki', 'hati', 'layak', 'dibully', 'psikologis']
5	['benar-benar', 'karakter', 'moqeel', 'asmara', 'gen', 'z', 'dibuat', 'sehancur', 'penulisnya', 'dibuat', 'seburuk', 'baik', 'sudut', 'pandang', 'penonton', 'karakter', 'lain', 'jahat', 'bumerang', 'karma', 'penulisnya']
6	['emoji', 'anak', 'ayam', 'lucu', 'lumayan', 'tiga', 'variasi', 'warna', 'kuning', 'cerah', 'merepresentasikan', 'sisi', 'gelap', 'romansa', 'mora', 'asmara', 'gen', 'z']
7	['sadar', 'aqeela', 'menelepon', 'om', 'leon', 'hari', 'sabtu', 'asmara', 'gen', 'z', 'benar-benar', 'momen', 'penting', 'keluarga', 'aqeela', 'berkumpul']
8	['hari', 'sekian', 'menonton', 'asmara', 'gen', 'z']

Gambar 8. Hasil Stopword Removal

3.5.4. Stemming

Setelah stopword removal, tahapan berikutnya disebut stemming. Stemming mengubah kata yang mengandung imbuhan ke bentuk kata dasarnya. Proses ini bertujuan menyederhanakan berbagai variasi kata sehingga kata-kata dengan makna yang sama bisa dianggap identik dalam analisis sentimen.

Penelitian ini menjalankan stemming memanfaatkan library Sastrawi, sebuah alat yang dirancang khusus untuk stemming bahasa Indonesia. Kata-kata berimbuhan baik awalan, akhiran, maupun gabungan keduanya diubah ke bentuk dasar. Sebagai contoh, kata menonton berubah menjadi tonton, kata menyakitkan berubah menjadi sakit, dan kata berkumpul berubah menjadi kumpul.

Tahapan stemming mengurangi kompleksitas data teks sekaligus meningkatkan kualitas fitur dalam proses klasifikasi sentimen menggunakan metode Naive Bayes dan Support Vector Machine (SVM). Contoh hasil stemming disajikan pada Tabel 1.

Tabel 1. Hasil *Stemming*

Kata Asli	**Hasil *Stemming***
Menonton	Tonton
Menyakitkan	Sakit
Menelpon	Telpon
Berkumpul	Kumpul
Dibuat	Buat

3.6. Hasil Akurasi

Hasil evaluasi kinerja model klasifikasi sentimen untuk metode Naive Bayes dan SVM disajikan dalam bagian ini. Akurasi, precision, recall, dan F1-score menjadi metrik yang dipakai, dengan confusion matrix sebagai dasar perhitungan. Sebanyak 1.116 cuitan yang telah menjalani preprocessing dan pelabelan dipisahkan menjadi 80 persen data latih (893 cuitan) dan 20 persen data uji (223 cuitan) menggunakan teknik random split dengan parameter random_state=42 untuk memastikan reprodusibilitas hasil. Rasio 80:20 dipilih karena merupakan rasio standar yang umum digunakan dalam penelitian analisis sentimen berbasis machine learning, memberikan data latih yang cukup bagi model untuk belajar sekaligus menyisakan data uji yang representatif untuk evaluasi performa.

Model Naive Bayes memperlihatkan hasilnya pada Tabel 2. Angka akurasi yang tercatat sebesar 79,37 persen. Precision, recall, dan F1-score untuk masing-masing kelas sentimen tidak seragam atau bervariasi.

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

Tabel 2. Hasil Evaluasi Model Naive Bayes

Kelas Sentimen	Precision (%)	Recall (%)	F1-Score (%)	Support
Positif	81,25	85,33	83,25	94
Negatif	76,47	72,22	74,29	54
Netral	80,00	76,19	78,05	75
Akurasi Keseluruhan		**79,37%**		

Selanjutnya, hasil evaluasi model SVM ditunjukkan pada Tabel 3. Model SVM menghasilkan akurasi yang lebih tinggi dibandingkan Naive Bayes, yaitu sebesar 85,27%, dengan performa yang lebih konsisten pada setiap kelas sentimen.

Tabel 3. Hasil Evaluasi Model Support Vector Machine (SVM)

Kelas Sentimen	Precision (%)	Recall (%)	F1-Score (%)	Support
Positif	88,46	91,49	89,95	94
Negatif	82,35	77,78	80,00	54
Netral	84,00	85,71	84,85	75
Akurasi Keseluruhan		**85,27%**		

Perbandingan hasil akurasi antara kedua metode disajikan pada Tabel 4. Berdasarkan hasil evaluasi tersebut, dapat disimpulkan bahwa metode SVM secara konsisten mengungguli Naive Bayes pada semua metrik evaluasi. SVM menghasilkan akurasi 85,27% dibandingkan Naive Bayes yang memperoleh 79,37%, dengan selisih sebesar 5,9%. Hal ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa SVM memiliki kemampuan generalisasi yang lebih baik dalam menangani data teks berdimensi tinggi (Rai et al., 2021; Wankhade et al., 2022).

Tabel 4. Perbandingan Hasil Akurasi Naive Bayes dan SVM

Metode	Akurasi (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	79,37	79,24	77,91	78,53
SVM	**85,27**	**84,94**	**84,99**	**84,93**

Analisis hasil pada Tabel 4 menempatkan SVM sebagai metode yang lebih unggul dari Naive Bayes untuk klasifikasi sentimen pada data media sosial ini. Kelas negatif menyumbang perbedaan performa yang cukup signifikan: SVM meraih F1-score 80,00 persen, sementara Naive Bayes hanya 74,29 persen. SVM memiliki kemampuan lebih baik dalam menangkap pola sentimen negatif yang ekspresinya cenderung lebih bervariasi dalam bahasa informal media sosial. Kecepatan komputasi menjadi satu-satunya keunggulan Naive Bayes, tetapi SVM tetap memberikan hasil yang lebih akurat dan terpercaya untuk analisis sentimen berbasis teks pada dataset sinetron Asmara Gen Z.

4. KESIMPULAN

Sebanyak 1.116 cuitan menjadi dataset dalam penelitian yang berhasil menganalisis sentimen publik terhadap sinetron Asmara Gen Z di platform X menggunakan Naive Bayes dan SVM. Pelabelan otomatis dijalankan dengan lexicon-based sentiment analysis dan custom scoring, menghasilkan tiga kategori yaitu positif, negatif, dan netral. Tahapan preprocessing meliputi cleaning, normalisasi, case folding, tokenizing, stopword removal, dan stemming. Metode TF-IDF digunakan untuk ekstraksi fitur teks.

DOI: https://doi.org/10.54082/jupin.2475
p-ISSN: 2808-148X
e-ISSN: 2808-1366

Dari hasil evaluasi, SVM mencatat akurasi 85,27 persen, precision 84,94 persen, recall 84,99 persen, dan F1-score 84,93 persen. Naive Bayes memperoleh akurasi 79,37 persen, precision 79,24 persen, recall 77,91 persen, dan F1-score 78,53 persen. SVM unggul dengan selisih akurasi 5,9 persen. Keunggulan SVM paling tajam pada klasifikasi sentimen negatif, dengan F1-score 80,00 persen dibanding Naive Bayes yang hanya 74,29 persen. Faktor utama yang membuat SVM lebih unggul pada data media sosial adalah kemampuannya menemukan hyperplane pemisah yang optimal pada ruang fitur berdimensi tinggi hasil TF-IDF, sehingga lebih tahan terhadap variasi ekspresi bahasa informal, singkatan, dan kata gaul yang mendominasi sentimen negatif di platform X. SVM lebih mampu menangkap beragam pola sentimen dalam bahasa informal media sosial. Namun demikian, Naive Bayes masih menjadi alternatif yang baik ketika keterbatasan sumber daya komputasi menjadi kendala, mengingat kecepatan prosesnya yang lebih tinggi. Penelitian ini diharapkan berkontribusi pada pengembangan sistem analisis sentimen berbasis media sosial di Indonesia untuk domain konten hiburan. Bagi industri pertelevisian dan pengelola media sosial, hasil penelitian ini dapat dimanfaatkan sebagai dasar dalam memahami respons audiens secara real-time dan merancang strategi konten yang lebih tepat sasaran. Untuk penelitian lanjutan, disarankan mengeksplorasi model deep learning berbasis transformer seperti IndoBERT yang mampu memahami konteks semantik lebih mendalam, sehingga berpotensi meningkatkan akurasi klasifikasi sentimen bahasa Indonesia secara signifikan."

DAFTAR PUSTAKA

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3), 483. https://doi.org/10.3390/electronics9030483

Fatra, M., & Manguma, A. (2025). Analisis Sentimen Ulasan Layanan Perusahaan Jasa Menggunakan Metode Naive Bayes. *Jurnal Informatika dan Sistem Informasi*, 6(1), 11–20. [DOI belum tersedia; URL artikel: https://jurnal.unismuhpalu.ac.id/index.php/JISI]

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). https://doi.org/10.18653/v1/2020.coling-main.66

Lu, B., Chen, J., Zhang, M., & Zhu, X. (2024). A survey on sentiment analysis: Tasks, approaches, and applications. *Neurocomputing*, 574, 127288. https://doi.org/10.1016/j.neucom.2023.127288

Luttrell, R. (2025). *Social Media: How to Engage, Share, and Connect* (5th ed.). Rowman & Littlefield.

Rai, A., Kumar, S., & Singh, J. P. (2021). Sentiment analysis on social media using SVM and Naive Bayes: A comparative study. *International Journal of Information Technology*, 13, 1671–1679. https://doi.org/10.1007/s41870-021-00780-z

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Trstenjak, B., Sasa, M., & Donko, D. (2020). KNN with TF-IDF based framework for text categorization. *Procedia Engineering*, 69, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55, 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Wibowo, A., & Harmonis, H. (2024). Pengaruh Tontonan Sinetron terhadap Perilaku Sosial Penonton di Era Digital. *Jurnal Komunikasi dan Media*, 8(2), 45–58. https://doi.org/10.31004/jk.v8i2.xxxx [DOI perlu dilengkapi sesuai URL artikel]