Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. 9 No. 3, Juni 2026
P-ISSN 2620-8342
E-ISSN 2621-3052

Analisis Sentimen Komentar Trailer Film Menggunakan Pendekatan Lexicon-Based dan Machine Learning

Fariska Adela Nurhidayah[1*], Riski Annisa[2], Muhammad Fahmi Julianto[3]
[1-3]Program Studi Informatika, Kampus Kota Pontianak, Universitas Bina Sarana Informatika
[1-3]Jl. Abdul Rahman Saleh No.18, Bangka Belitung Laut, Kec. Pontianak Tenggara,
Kota Pontianak, Kalimantan Barat 78124
E-mail: fariskaadela300@gmail.com[1*]

Submitted Date : 04 Juni 2026 ***Accepted Date : 18 Juni 2026***

Abstrak – Seiring berkembangnya media sosial, YouTube telah menjadi salah satu tempat utama di mana orang dapat menggunakan kolom komentar untuk berbagi pemikiran mereka tentang film. Penelitian ini menggunakan kombinasi teknik berbasis leksikon dan *machine learning* untuk memeriksa sentimen penonton mengenai trailer film Andai Ibu Tidak Menikah dengan Ayah. Sejumlah langkah *preprocessing*, termasuk *cleaning*, *case folding,* normalisasi, tokenisasi, *stopword removal*, dan *stemming*, diterapkan pada data setelah dikumpulkan melalui *scraping* komentar YouTube. Leksikon Sentimen Indonesia (*InSet Lexicon*) digunakan untuk pelabelan sentimen, dan pendekatan TF-IDF digunakan untuk ekstraksi fitur. *Synthetic Minority Oversampling Technique* (SMOTE) digunakan untuk mengoreksi ketidakseimbangan data. Metode *Naïve Bayes*, *Logistic Regression*, dan *Support Vector Machine* (SVM) kemudian digunakan untuk mengklasifikasikan sentimen. Metrik akurasi, presisi, *recall*, dan *F1-score* digunakan untuk menilai kinerja model. Dengan akurasi 81.90%, presisi 79.95%, *recall* 81.90%, dan *F1-score* 80.87%, hasil ini menunjukkan bahwa algoritma *Naïve Bayes* berkinerja terbaik. Sementara itu, akurasi SVM dan *Logistic Regression* masing-masing adalah 66.67% dan 62.86%. Hasil ini menunjukkan bahwa *Naïve Bayes* mengungguli algoritma lain dalam klasifikasi sentimen dari komentar *trailer* film.
Kata kunci : Analisis Sentimen; Lexicon-Based; Machine Learning; TF-IDF; YouTube;

Abstract - *As social media has grown, YouTube has become one of the main places where people may use comment sections to share their thoughts about movies. This study uses a combination of lexicon-based and machine learning techniques to examine viewer sentiment regarding the trailer for the film Andai Ibu Tidak Menikah dengan Ayah. A number of preprocessing steps, including as cleaning, case folding, normalization, tokenization, stopword removal, and stemming, were applied to the data after it was gathered via YouTube comment scraping. The Indonesian Sentiment Lexicon (InSet Lexicon) was used for sentiment labeling, and the TF-IDF approach was used for feature extraction. The Synthetic Minority Oversampling Technique (SMOTE) was used to correct data imbalance. The Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM) methods were then used to classify sentiment. Accuracy, precision, recall, and F1-score metrics were used to assess the model's performance. With an accuracy of 81.90%, precision of 79.95%, recall of 81.90%, and F1-score of 80.87%, the findings show that the Naïve Bayes algorithm performed the best. In the meantime, the accuracy of SVM and Logistic Regression was 66.67% and 62.86%, respectively. These results show that Naïve Bayes outperforms the other algorithms in sentiment classification from movie trailer comments.*
Keywords: Sentiment Analysis; Lexicon-Based; Machine Learning; TF-IDF; YouTube;

1. Pendahuluan

Evolusi media sosial dan teknologi digital telah mengubah cara individu menyuarakan pendapat mereka tentang film, terutama di YouTube. Komentar pada *trailer* film kini menjadi sumber informasi yang kaya tentang opini publik yang dapat dianalisis menggunakan analisis sentimen dan *Natural Language Processing* (NLP). Salah satu film yang menarik perhatian masyarakat adalah Andai Ibu Tidak Menikah dengan Ayah yang mengangkat tema konflik keluarga dan berhasil memicu berbagai tanggapan dari penonton di media sosial.

Analisis sentimen dalam *text mining* mengklasifikasikan opini dan perasaan dalam teks sebagai positif atau negatif. Memahami respons audiens terhadap informasi secara efektif membutuhkan strategi ini. Namun, penerapan analisis sentimen pada teks berbahasa Indonesia memiliki tantangan karena data media sosial umumnya tidak terstruktur, menggunakan bahasa informal, slang, dan singkatan, serta sering mengalami ketidakseimbangan kelas data.

Penelitian sebelumnya lebih banyak membahas konflik keluarga dalam film menggunakan pendekatan kualitatif dan jumlah responden terbatas, sehingga belum mampu menggambarkan sentimen publik secara menyeluruh. Selain itu, sebagian besar penelitian analisis sentimen hanya menggunakan satu metode klasifikasi

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. 9 No. 3, Juni 2026
P-ISSN 2620-8342
E-ISSN 2621-3052

tanpa mengombinasikan pendekatan *Lexicon-Based* dan *Machine Learning*. Padahal, kombinasi kedua metode tersebut dapat meningkatkan efektivitas klasifikasi serta mengurangi proses pelabelan manual.

Dengan mempertimbangkan keadaan tersebut, penelitian kami menggunakan teknik berbasis leksikon dan machine learning untuk memeriksa sentimen komentar penonton pada *trailer* YouTube untuk film Andai Ibu Tidak Menikah dengan Ayah. Penelitian ini bertujuan mengetahui kecenderungan sentimen publik serta membandingkan performa metode klasifikasi berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian diharapkan dapat memberikan kontribusi akademik dalam pengembangan analisis sentimen berbahasa Indonesia serta manfaat praktis bagi industri perfilman dalam memahami respons audiens dan menyusun strategi promosi film.

2. Tinjauan Pustaka

2.1. Perfilman

Film merupakan salah satu bentuk komunikasi massa yang memiliki daya tarik yang besar. Hal ini menunjukkan bahwa film dapat menjangkau penonton yang luas dan beraneka ragam. Pesan, nilai, dan narasi yang ada dalam film mampu mempengaruhi para penontonnya secara signifikan. Nasirin dan Pithaloka (2022) menyatakan bahwa karena film menggabungkan berbagai elemen artistik, termasuk akting, musik, sinematografi, dan penulisan skenario, maka film dipandang sebagai karya seni yang mewakili budaya. Budaya, nilai-nilai, dan narasi masyarakat yang memproduksi film tersebut dapat tercermin di dalamnya. Lebih jauh lagi, film diakui sebagai komponen struktur sosial, yang menunjukkan pengaruhnya terhadap pembentukan dan refleksi dinamika sosial, budaya, dan politik dalam masyarakat [1].

Gambar 1. Poster Andai Ibu Tidak Menikah dengan Ayah
Sumber : https://jadwalnonton.com/review/andai-ibu-tidak-menikah-dengan-ayah/

Perfilman merupakan segala hal yang berkaitan dengan film, yang berfungsi sebagai media komunikasi massa untuk pendidikan, pembentukan karakter, serta promosi budaya Indonesia. Andai Ibu Tidak Menikah dengan Ayah karya Kuntz Agus adalah film keluarga Indonesia yang menarik. Tokoh utama film ini, Alin, menemukan buku harian ibunya dan mengetahui tentang pengorbanan dan kehilangan sosok ayah yang dialaminya. *Trailer* film tersebut mendapat banyak respons di YouTube, mulai dari pujian, kritik, hingga ungkapan emosional penonton, sehingga menjadi sumber yang potensial untuk analisis sentimen [1].

Penelitian sebelumnya seperti pada jurnal [2], [3] menunjukkan bahwa komentar YouTube dapat dimanfaatkan untuk mengetahui opini masyarakat menggunakan metode analisis sentimen. Beberapa penelitian juga menjelaskan pentingnya tahap *preprocessing* karena komentar YouTube banyak menggunakan bahasa tidak formal dan slang. Berdasarkan hal tersebut, penelitian analisis sentimen pada komentar *trailer* film Andai Ibu Tidak Menikah dengan Ayah penting dilakukan, terutama dengan pendekatan *Lexicon-Based* dan *Machine Learning*, guna mengetahui persepsi masyarakat sekaligus mendukung perkembangan penelitian analisis sentimen pada perfilman Indonesia.

2.2. Analisis Sentimen

Analisis sentimen adalah bidang NLP dan penambangan data yang memeriksa pandangan, emosi, dan bias. Komentar pengguna di YouTube dimanfaatkan untuk analisis sentimen guna memahami reaksi populer. Penerapannya telah digunakan di berbagai bidang, termasuk perfilman untuk mengetahui respons penonton terhadap trailer film. Proses analisis sentimen umumnya meliputi pengumpulan data, preprocessing, pelabelan data, ekstraksi fitur, klasifikasi, dan evaluasi model agar data teks dapat diolah menjadi informasi yang lebih terstruktur [4].

2.3. *Text Mining*

Penambangan teks mengekstrak data dari teks untuk menyelidiki hubungan antar dokumen. Tujuannya adalah menemukan istilah yang sesuai dengan isi dokumen. *Text Mining* adalah jenis penambangan data yang

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. 9 No. 3, Juni 2026
P-ISSN 2620-8342
E-ISSN 2621-3052

mencari pola menarik dalam sejumlah besar data tekstual. Ini dapat digunakan untuk kategorisasi sentimen, pengelompokan teks, ekstraksi informasi, dan pengambilan informasi. Pemrosesan teks diperlukan sebelum menerapkan *Text Mining* karena kumpulan teks biasanya mengandung *noise*. *Cleaning, case folding*, tokenisasi, *Stopword Removal*, dan *stemming* adalah beberapa tindakan yang diperlukan dalam *preprocessing*, menurut Evita et al. (2020) [6].

2.4. YouTube

YouTube memungkinkan pengguna untuk mempublikasikan, menonton, menyukai, mengomentari, dan berbagi video secara gratis. YouTube didirikan pada Februari 2005 oleh para veteran PayPal, Chad Hurley, Steve Chen, dan Jawed Karim. YouTube menayangkan video buatan pengguna, acara TV, dan cuplikan film. Salah satu layanan Google, YouTube, menawarkan kepada penggunanya kemampuan untuk mengunggah video, yang dapat dilihat secara gratis oleh pengguna di seluruh dunia untuk film yang akan datang [6].

2.5. *Lexicon-Based*

Lexicon-Based merupakan salah satu pendekatan dalam analisis sentimen yang memanfaatkan kata, frasa, bentuk ekspresi, maupun isi teks yang umumnya ditemukan pada percakapan, dialog, unggahan, ulasan, dan berbagai bentuk komunikasi digital lainnya. Metode ini menentukan tingkat polaritas emosi dengan mencocokkan kata-kata dalam kalimat dengan kamus sentimen yang berisi daftar kata-kata positif dan negatif.

Pada metode *Lexicon-Based*, proses klasifikasi sentimen dilakukan berdasarkan jumlah kata positif dan kata negatif yang terdapat pada teks yang telah melalui tahap pembersihan (*preprocessing*). Dengan mencocokkan kata dalam teks terhadap kamus sentimen, sistem dapat menentukan apakah suatu teks termasuk ke dalam kategori sentimen positif, dan negatif [5].

2.6. *Machine Learning*

Menurut penelitian Prasetyo dan Dewayanto (2024), *machine learning* adalah subset dari *artificial intelligence* (AI) yang dapat dicirikan sebagai teknik yang dapat secara otomatis melakukan aktivitas dan membuat penilaian dengan belajar dari pengalaman atau data dan beradaptasi dengan input baru. Dengan menganalisis data yang ada, mengidentifikasi pola spesifik, dan kemudian secara otomatis menghasilkan penilaian yang tepat, machine learning membantu manusia dalam pengambilan keputusan.

Pendekatan atau tahapan yang digunakan untuk membangun model berdasarkan data disebut algoritma machine learning. Untuk mengatasi berbagai masalah, algoritma tersebut berupaya mengidentifikasi pola dan koneksi dalam data. Tergantung pada jenis data dan masalah yang dihadapi, setiap metode *machine learning* memiliki ciri, manfaat, dan kekurangan yang unik. [6].

1. *Naive Bayes*

Algoritma *Naive Bayes* merupakan salah satu metode klasifikasi yang menggunakan pendekatan probabilitas dan statistik. Menurut penelitian Yuda dan Andarsyah (2022), algoritma ini termasuk metode pembelajaran sederhana yang menerapkan *Teorema Bayes* untuk mengklasifikasikan maupun memprediksi probabilitas suatu kelas data. Berdasarkan penelitian Giovani dkk. (2020), *Naive Bayes* menjadi salah satu algoritma yang populer dan masuk dalam sepuluh algoritma terbaik pada bidang *data mining* menurut IEEE di Hongkong. Salah satu variannya adalah *Multinomial Naive Bayes* yang sangat sesuai digunakan untuk klasifikasi data dengan fitur diskrit, seperti jumlah kata pada analisis teks [7]

2. *Logistic Regression*

Menurut penelitian Lengkong dkk. (2021), *Logistic Regression* adalah teknik analitik yang menggambarkan hubungan antara satu atau lebih variabel independen, baik kontinu maupun kategorikal, dan variabel dependen dengan dua atau lebih kategori. Hubungan antara variabel independen dan dependen dijelaskan oleh algoritma ini, di mana hasil klasifikasinya biasanya berbentuk kategori biner seperti 0 dan 1, *true* dan *false*, atau besar dan kecil. Penelitian Pramakrisna dkk. (2022) menjelaskan bahwa karakteristik utama algoritma ini terletak pada variabel dependennya yang berbentuk kategorikal, sehingga membedakannya dari regresi linier atau regresi berganda lainnya.

3. *Support Vector Machine*

Support Vector Machine (SVM) adalah teknik yang digunakan untuk regresi dan klasifikasi, menurut penelitian oleh Pratiwi dan Setiawan (2020). Vladimir Vapnik pertama kali memperkenalkan algoritma ini pada tahun 1992 sebagai ide terobosan di bidang pengenalan pola. *Hyperplane* optimal yang dapat membagi dua kelas data dalam ruang input ditemukan oleh SVM dengan memanfaatkan pendekatan *Structural Risk Minimization* (SRM) [8].

2.7. TF-IDF

Peneliti menggunakan TF-IDF untuk pemrosesan teks dan pemodelan bahasa alami. Tujuan mendasar dari pendekatan TF-IDF adalah untuk membandingkan sebuah kata (istilah) dalam sebuah dokumen dengan koleksi yang lebih besar [9].

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. 9 No. 3, Juni 2026
P-ISSN 2620-8342
E-ISSN 2621-3052

2.8. SMOTE (*Synthetic Minority Oversampling Technique*)

Salah satu teknik resampling untuk menyeimbangkan distribusi kelas adalah *Synthetic Minority Oversampling Technique* (SMOTE). Metode ini meningkatkan kuantitas data dalam kelas minoritas dengan menambahkan sampel sintetis baru ke sampel dari kelas tersebut. [10].

2.9. Google Colab

Google Colab (*Google Colaboratory*) merupakan sebuah *platform* yang beroperasi menggunakan teknologi *cloud computing* yang ditawarkan oleh Google untuk mengeksekusi kode *Python* secara daring tanpa perlu instalasi atau konfigurasi lingkungan pemrograman di perangkat pribadi. Alat ini banyak dipakai oleh para ilmuwan data, peneliti, mahasiswa, serta para pengembang untuk berbagai keperluan, mulai dari pengolahan data, analisis data, pengembangan *artificial intelligence*, hingga pelatihan model *machine learning*. Google Colab menjadi pilihan populer karena kemudahan penggunaan, menawarkan akses kepada sumber daya komputasi yang cukup mumpuni, serta memberi kesempatan kepada pengguna untuk berkolaborasi dan berbagi proyek dengan tim dengan cara yang efisien.

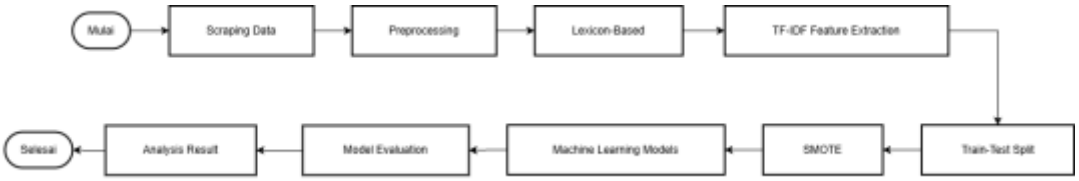
2.10. Penelitian Terkait

Beberapa studi menunjukkan bahwa analisis sentimen di YouTube, Twitter, dan Instagram dapat menentukan opini publik tentang film dan trailer film. Pengumpulan data, pra-pemrosesan, pembobotan TF-IDF, dan klasifikasi *machine learning* merupakan hal yang umum dalam penelitian ini. *Case folding*, tokenisasi, normalisasi teks, *stopword removal*, dan *stemming* diperlukan untuk pengorganisasian dan analisis data. TF-IDF memberi bobot pada kata-kata untuk meningkatkan klasifikasi sentimen. Penelitian ini menggunakan teknik seperti *Naïve Bayes*, SVM, *Logistic Regression*, *Decision Tree*, KNN, dan *Random Forest*.

Hasil penelitian menunjukkan bahwa setiap algoritma memiliki tingkat akurasi yang berbeda-beda dalam mengklasifikasikan sentimen. *Logistic Regression* memperoleh akurasi tertinggi sebesar 86.61% pada analisis sentimen *trailer* film *Deadpool vs Wolverine*, sedangkan SVM menunjukkan performa paling stabil dengan rata-rata akurasi sebesar 85.5% pada dataset tweet film berbahasa Indonesia. Selain itu, teknik *Naïve Bayes* telah menunjukkan akurasi yang baik dalam penelitian analisis sentimen sinema. Eksperimen ini menunjukkan bahwa machine learning dalam analisis sentimen dapat secara otomatis, cepat, dan akurat menilai opini publik. Lebih lanjut, analisis sentimen dapat membantu industri hiburan dan para peneliti mengevaluasi kesan konsumen terhadap film dan konten digital.

3. Metode Penelitian

Penelitian ini menerapkan pendekatan kuantitatif melalui metode analisis sentimen terhadap tanggapan pengguna YouTube terhadap trailer film Andai Ibu Tidak Menikah dengan Ayah. Tujuan dari penelitian ini adalah untuk mengevaluasi pola sentimen masyarakat dan menilai efisiensi metode *Machine Learning* dalam mengklasifikasikan sentimen. Data yang digunakan adalah komentar dari pengguna yang diambil dari video *trailer* resmi film tersebut yang diunggah di *platform* YouTube.



Gambar 2. *Flowchart* Penelitian

3.1. Pengumpulan Data

Tahap ini mengumpulkan komentar YouTube dari *trailer* film Andai Ibu Tidak Menikah dengan Ayah. Python dan pustaka komentar YouTube digunakan untuk mengumpulkan data secara otomatis.

3.2. *Text Processing*

Langkah pertama dalam mengubah kata-kata menjadi data untuk diproses disebut *text preprocessing*. Proses-proses yang menguraikan tahapan persiapan teks dijelaskan di sini. [11]:

1. *Cleaning*

Cleaning adalah praktik menghilangkan karakteristik dari sebuah dokumen, baik kata maupun karakter yang tidak terkait dengan konten atau tidak perlu untuk mengurangi gangguan selama proses kategorisasi.

2. *Case Folding*

Tahapan *case folding* Tujuannya adalah untuk mengubah setiap huruf dalam dokumen menjadi huruf kecil. Karakter yang bukan alfabet diabaikan dan digunakan sebagai pembatas.

3. *Tokenizing*

Pada tahap *tokenizing* adalah proses segmentasi string input berdasarkan setiap kata yang membentuknya.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. 9 No. 3, Juni 2026
P-ISSN 2620-8342
E-ISSN 2621-3052

4. Normalisasi

Inilah prosedur yang digunakan dalam kamus bahasa Indonesia untuk mengkonversi istilah-istilah non-standar kembali ke bahasa standar.

5. *Stopword Removal*

Tahap *stopword removal* adalah kata yang tidak bermakna yang sering muncul dalam jumlah besar. Kata-kata seperti "dengan," "untuk," "dan," "ke," petunjuk waktu, dan kata-kata pertanyaan adalah beberapa contoh kata yang termasuk dalam level ini.

6. *Stemming*

Stemming merupakan metode pemrosesan bahasa alami mendasar yang umumnya diterima pengguna untuk pengambilan informasi yang efektif dan efisien.

3.3. Pelabelan Data

Setelah pre-processing, pelabelan sentimen berbasis leksikon digunakan. Leksikon emosi bahasa Indonesia (InSet Lexicon) dicocokkan dengan istilah-istilah komentar. Skor sentimen, baik atau negatif, diberikan kepada setiap kata. Skor sentimen keseluruhan setiap komentar ditentukan untuk melakukan proses kategorisasi.

3.4. Ekstraksi Fitur TF-IDF

TF-IDF (*Term Frequency–Inverse Document Frequency*) digunakan untuk menumerasi input teks setelah penandaan. TF-IDF menentukan signifikansi suatu kata dalam dokumen berdasarkan frekuensi dan distribusinya [11]. Matriks numerik yang dapat dimasukkan ke dalam metode klasifikasi adalah output TF-IDF.

3.5. Pembagian Data

Selanjutnya, dataset berlabel dibagi menjadi 20% data pengujian dan 80% data pelatihan. Model dapat dievaluasi menggunakan data yang sebelumnya tidak diketahui dan mempelajari pola dari data pelatihan berkat pembagian data ini [14]. Metode pembagian data pelatihan dan pengujian digunakan untuk membagi data.

3.7. SMOTE (*Synthetic Minority Oversampling Technique*)

Pada tahap analisis data, teridentifikasi adanya ketidakseimbangan jumlah data di setiap kategori sentimen. Situasi ini dapat mengakibatkan model lebih cenderung untuk memprediksi kategori dominan dan mengabaikan kategori yang kurang. Oleh karena itu, metode *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan guna menyeimbangkan distribusi data dalam data pelatihan.

SMOTE mensintesis data minoritas dari data yang ada. Input yang seimbang di setiap kategori membantu algoritma mempelajari pola di semua kategori sentimen. SMOTE hanya diterapkan pada data pelatihan untuk mencegah kebocoran data dan memastikan imparsialitas evaluasi model.

3.8. Klasifikasi *Machine Learning*

Tahapan klasifikasi dilakukan dengan memanfaatkan metode *Naïve Bayes*, *Logistic Regression*, serta *Support Vector Machine*. Berdasarkan penelitian [6] algoritma *machine learning* dapat secara otomatis mengidentifikasi pola dalam data untuk menghasilkan prediksi atau klasifikasi yang didasarkan pada fitur data yang tersedia. Penelitian yang dilakukan oleh [3] juga mengungkapkan bahwa perpaduan TF-IDF dan *Logistic Regression* dapat memberikan hasil yang memuaskan dalam menganalisis sentimen dari komentar di YouTube.

3.9. Evaluasi Model

Matriks kebingungan dan sejumlah ukuran evaluasi klasifikasi digunakan dalam penelitian [9] untuk mengevaluasi model. Metrik klasifikasi pembelajaran mesin termasuk akurasi, presisi, *recall*, dan *F1-Score* mengevaluasi model. Rata-rata harmonik dari presisi dan *recall*, *F1-Score*, menunjukkan keseimbangan kinerja model. Akurasi mengevaluasi akurasi prediksi keseluruhan model, presisi mengukur akurasi kategorinya, dan *recall* mengukur kemampuannya untuk mengenali semua data kategori. Tujuan penggunaan indikator ini adalah untuk menawarkan evaluasi yang lebih menyeluruh terhadap kinerja model kategorisasi.

a. Persamaan *Accuracy*

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad [12]$$

b. Persamaan *Precision*

$$Precision = \frac{TP}{TP+FP} \quad [13]$$

c. Persamaan *Recall*

$$Recall = \frac{TP}{TP+FN} \quad [14]$$

d. Persamaan *F1-Score*

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [14]$$

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Performa model klasifikasi dinilai menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*. Metrik akurasi membandingkan prediksi benar model dengan kumpulan data untuk menilai keakuratannya. *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) merupakan bagian dari perhitungan tersebut.

4. Hasil dan Pembahasan

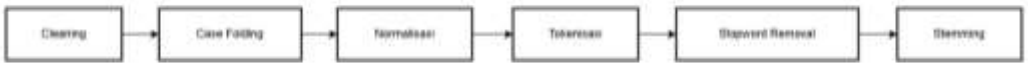
4.1. Data yang Dikumpulkan

Dalam penelitian ini, dilakukan penilaian emosi terhadap tanggapan pengguna YouTube pada *trailer* film Andai Ibu Tidak Menikah dengan Ayah. Komentar yang dianalisis diambil melalui proses *scraping* data dari *platform* YouTube, lalu dilakukan serangkaian langkah *preprocessing* agar teks menjadi lebih teratur dan jelas sebelum dipakai dalam analisis selanjutnya.

	video_id	judul_video	komentar	like_komentar	tanggal	author
0	IfkNO8DzFVM	Andai Ibu Tidak Menikah dengan Ayah Official T...	😭😭 Udah 7 bulan yg lalu rilis..tapi kurang di...	0	2026-04-02T14:40:32Z	@IrhamLuhur
1	IfkNO8DzFVM	Andai Ibu Tidak Menikah dengan Ayah Official T...	mungkin itu yg juga di rasakan anak2 ku 😭	0	2026-04-02T03:44:04Z	@mila_ajal5
2	IfkNO8DzFVM	Andai Ibu Tidak Menikah dengan Ayah Official T...	gw POVnya agak beda nih, gw di posisi alin yg ...	0	2026-03-20T04:43:00Z	@LeniTjahjadi
3	IfkNO8DzFVM	Andai Ibu Tidak Menikah dengan Ayah Official T...	Udh selesai nontonnya sedihh bangettt 😭😭 \nApala...	1	2026-03-19T09:49:14Z	@Aisyah-n8n5t

Gambar 2. Hasil *Scraping* Data

4.2. *Preprocessing* Data



Gambar 3. Tahapan *Preprocessing*

Beberapa proses dilakukan untuk mempersiapkan data mentah untuk analisis dalam tahap *preprocessing*. Langkah-langkah dalam *preprocessing*:

Tabel 1. Tabel *Preprocessing*

Tahapan	Deskripsi
Teks Asli	mungkin itu yg juga di rasakan anak2 ku
Case Folding	mungkin itu yg juga di rasakan anak ku
Normalisasi	mungkin itu yang juga di rasakan anak ku
Tokenisasi	[mungkin, itu, yang, juga, di, rasakan, anak, ku]
Stopword Removal	['rasakan', 'anak', 'ku']
Stemming	[rasa, anak, ku]

Tabel 1. menunjukkan contoh urutan *preprocessing* yang diterapkan pada data yang diunggah di media sosial sebelum masuk ke tahap analisis dengan menggunakan metode *machine learning*. Langkah-langkah *preprocessing* ini bertujuan untuk membersihkan dan menyelaraskan teks agar lebih mudah untuk diproses oleh model. Data dibersihkan untuk menghilangkan URL, simbol, angka, dan tanda baca. Setelah pengubahan huruf besar-kecil menjadi huruf kecil, normalisasi kata non-standar disesuaikan dengan aturan bahasa Indonesia.

Tokenisasi membagi teks menjadi segmen kata, *stopword removal* menghilangkan konsep umum yang tidak membantu analisis, dan *stemming* menyederhanakan kata. Ini mempersiapkan teks untuk ekstraksi fitur model dan kategorisasi.

4.3. Pelabelan Data

Setelah *preprocessing*, teknik *lexicon-based* memberi label pada data sebagai positif atau negatif.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

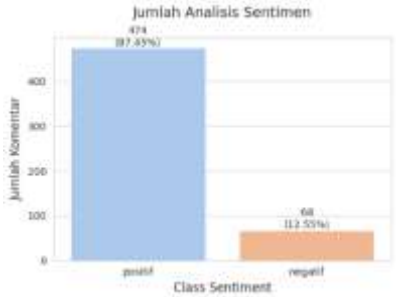
Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. 9 No. 3, Juni 2026
P-ISSN 2620-8342
E-ISSN 2621-3052

	Review Text	cleaning	case_folding	score	sentimen
0	😭😭 Udah 7 bulan yg lalu rilis...tapi kurang di...	Udah bulan yg lalu rilistapi kurang dikenal ...	udah bulan yg lalu rilistapi kurang dikenal ...	-1	positif
1	mungkin itu yg juga di rasakan anak2 ku 😭	mungkin itu yg juga di rasakan anak ku	mungkin itu yg juga di rasakan anak ku	-3	negatif
2	gw POVnya agak beda nih, gw di posisi alin ini--	gw POVnya agak beda nih gw di posisi alin yg m...	gw povnya agak beda nih gw di posisi alin yg m...	-4	positif
3	Udh selesai nontonnya sedihh bangettt 😭😭 \nApaka...	Udh selesai nontonnya sedihh bangettt!! nApeligi...	udh selesai nontonnya sedihh bangettt!! napeligi...	1	positif
4	3 kali nonton, 3 kali menangis 😭	kali nonton kali menangis	kali nonton kali menangis	-1	positif

Gambar 4. Hasil Pelabelan Data

Pelabelan di atas menghasilkan 474 nilai data negatif dan 68 nilai data positif. Skor sentimen total untuk setiap komentar digunakan untuk pelabelan.



Gambar 5. Grafik Hasil Pembagian Label Kelas Positif dan Label Kelas Negatif

4.4. Visualisasi *Wordcloud*

Wordcloud Dihasilkan dari data teks yang telah diproses, menunjukkan kata-kata teratas. Ayah, saya, anak, ibu, dan kata-kata lainnya mendominasi, menunjukkan bahwa model tersebut menangkap kata kunci penting dalam riset trailer film Andai Ibu Tidak Menikah dengan Ayah.

Gambar 6. Visualisasi *Wordcloud*

4.5 Ekstraksi Fitur TF-IDF

TF-IDF (*Term Frequency–Inverse Document Frequency*) digunakan untuk menumerasi data berlabel. Teknik ini memberi bobot pada kata-kata kunci dalam dokumen komentar.

4.6. Pembagian Data

Langkah selanjutnya adalah klasifikasi menggunakan *Naive Bayes, Logistic Regression, dan Support Vector Machine*. Sebelum memeriksa hasil klasifikasi, 420 dataset pelatihan dan 105 dataset pengujian akan digunakan dengan rasio 80%:20%.

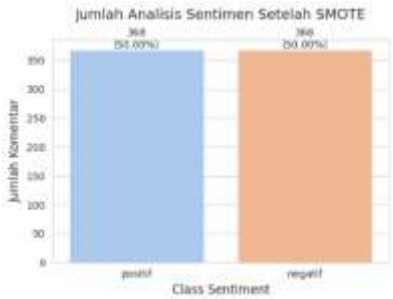
Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. 9 No. 3, Juni 2026
P-ISSN 2620-8342
E-ISSN 2621-3052

Gambar 7. *Train/Split Data*

4.7. SMOTE (*Synthetic Minority Oversampling Technique*)

Setelah data dibagi menjadi data latih dan data uji, langkah selanjutnya adalah menyeimbangkan data dengan memanfaatkan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Pendekatan ini diterapkan pada data latih untuk memperbanyak jumlah data dari kelas yang lebih sedikit dengan menciptakan sampel sintetis yang didasarkan pada karakteristik dari data yang sudah ada. Penggunaan SMOTE menghasilkan distribusi data yang lebih seimbang di antara kelas-kelas sentimen, yang membantu mengurangi kecenderungan model terhadap kelas yang lebih dominan dan meningkatkan performa klasifikasi selama tahap pelatihan.



Gambar 8. Hasil setelah proses SMOTE

4.8. Implementasi Model Machine Learning

Data berada dalam implementasi model *machine learning* setelah *preprocessing*. Penelitian ini mengklasifikasikan sentimen dari data komentar yang telah dibersihkan menggunakan tiga algoritma. Langkah pertama adalah ekstraksi fitur Vektorisasi TF-IDF (*Term Frequency-Inverse Document Frequency*). Setelah ekstraksi fitur, fungsi *train_test_split* membagi data menjadi data pelatihan (80%) dan data pengujian (20%) dengan nilai *random_state* sebesar 42 untuk memastikan pengulangan. Tiga algoritma yang diterapkan pada data ini meliputi:

1. *Naïve Bayes*, metode klasifikasi berbasis probabilitas yang menggunakan distribusi Multinomial.
2. *Support Vector Machine* (SVM), yang memaksimalkan jarak kelas menggunakan *hyperplane* ideal untuk mengklasifikasikan data.
3. *Logistic regression* memperkirakan keanggotaan kelas data berdasarkan hubungan linier antara karakteristik dan label.

Setiap model menggunakan parameter default *scikit-learn* untuk melatih data pelatihan. Teknik ini bertujuan untuk membangun model klasifikasi sentimen menggunakan fitur data komentar.

4.9. Evaluasi Model

Tiga model *machine learning* analisis sentimen diimplementasikan, masing-masing dengan kinerja yang berbeda. *Naive Bayes* memiliki akurasi maksimum 81.90%, diikuti oleh *Logistic Regression* (62.86%) dan SVM (66.67%). Tabel 2 merincikan kinerja masing-masing model.

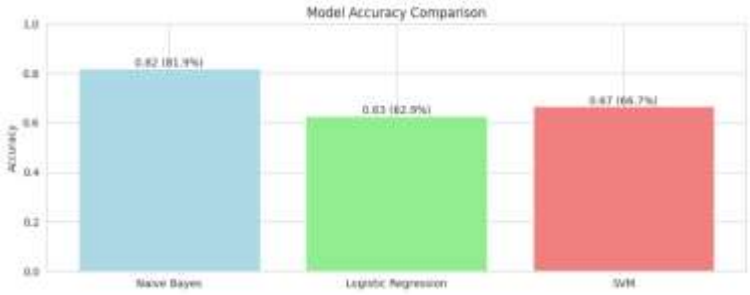
Tabel 2. Matriks Evaluasi Model

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Naive Bayes	81.90%	79.95%	81.90%	80.87%
Logistic Regression	62.86%	77.97%	62.86%	68.76%
SVM	66.67%	77.57%	66.67%	71.28%

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. 9 No. 3, Juni 2026
P-ISSN 2620-8342
E-ISSN 2621-3052

Tabel 2 menunjukkan bahwa *Naive Bayes* memiliki akurasi (81.90%, presisi (79,95%), dan *F1-Score* (80,87%) tertinggi. Hal ini menunjukkan bahwa model klasifikasi sentimen ini menyeimbangkan *precision* dan *recall*.

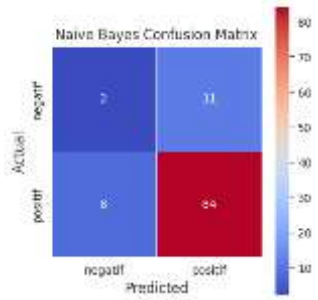


Gambar 9. Hasil Evaluasi Model dengan Tiga Algoritma

Gambar 9 menunjukkan bahwa *Naive Bayes* memiliki akurasi klasifikasi sentimen tertinggi sebesar 81.90%. SVM memiliki akurasi 66.7%, sedangkan *Logistic Regression* memiliki 62.9%. Hasil ini menunjukkan bahwa *Naive Bayes* mendeteksi pola data lebih baik daripada dua metode lainnya. Perbedaan kinerja ini mungkin disebabkan oleh dataset, karena *Naive Bayes* mengelola distribusi data dengan lebih baik. Dalam penelitian ini, *Naive Bayes* adalah sistem klasifikasi sentimen yang paling akurat.

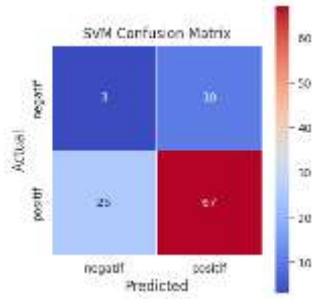
4.10. Confusion Matrix

Model *Naive Bayes* menampilkan kinerja terbaik dalam pengklasifikasian sentimen jika dibandingkan dengan model lainnya. Model ini berhasil menghasilkan 84 prediksi yang akurat dalam kategori positif, yaitu angka tertinggi di antara semua model yang diuji. Hanya delapan nilai positif yang salah diklasifikasikan sebagai negatif oleh model. Model tersebut memprediksi dua data negatif secara akurat tetapi salah mengklasifikasikan 11 data negatif sebagai positif. Hasil ini menunjukkan bahwa *Naive Bayes* secara akurat mendeteksi sentimen positif dengan tingkat kesalahan yang rendah. Gambar 10 menunjukkan *confusion matrix* model ini.



Gambar 10. *Confusion Matrix Naïve Bayes*

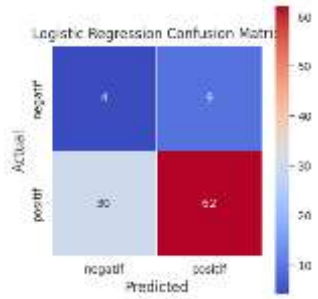
Model *Support Vector Machine* (SVM) menunjukkan klasifikasi sentimen yang baik pada dataset penelitian. Model ini memprediksi sentimen positif dengan akurasi 67%. Namun, 25 nilai data positif salah diklasifikasikan sebagai negatif. Model ini hanya memprediksi tiga titik data negatif secara akurat, dan salah mengklasifikasikan 10 lainnya sebagai positif. Hasil ini menunjukkan bahwa SVM mengenali sentimen positif lebih baik daripada sentimen negatif. Gambar 11 menunjukkan *confusion matrix* model ini.



Gambar 11. *Confusion Matrix SVM*

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Model *Logistic Regression* berkinerja baik dalam klasifikasi sentimen. Model ini secara tepat memprediksi 62 sentimen positif dan 4 sentimen negatif. Selain itu, 30 titik data positif salah diprediksi sebagai negatif dan 9 titik data negatif sebagai positif. *Logistic Regression* masih dapat mengenali emosi positif, meskipun tingkat kesalahan klasifikasinya lebih tinggi daripada model lain. Gambar 12 menunjukkan *confusion matrix* model ini.



Gambar 12. *Confusion Matrix Logistic Regression*

Ketidakseimbangan dalam distribusi data sentimen yang teridentifikasi pada tahap awal penelitian diduga memengaruhi kinerja model klasifikasi yang dibuat. Ini terlihat dari kecenderungan semua tiga algoritma menghasilkan prediksi yang lebih tepat untuk kelas positif ketimbang kelas negatif. Algoritma *Naive Bayes* menunjukkan ketahanan lebih baik dalam situasi ini, yang terlihat dari nilai *True Positive* tertinggi dan nilai *False Negative* terendah dibandingkan algoritma lainnya. SVM dan *Logistic Regression* sama-sama mampu mengidentifikasi sentimen positif dengan baik, tetapi kesulitan memisahkan data kelas negatif. Hasil ini menunjukkan bahwa model tersebut mendeteksi secara berlebihan kelas mayoritas karena distribusi data yang tidak merata.

5. Kesimpulan

Menurut penelitian, metode pendekatan *lexicon-based* dan *machine learning* dapat menilai komentar YouTube pada *trailer* Andai Ibu Tidak Menikah dengan Ayah. Analisis tersebut mencakup *preprocessing*, pelabelan sentimen dengan *InSet Lexicon*, ekstraksi fitur dengan TF-IDF, koreksi ketidakseimbangan data dengan SMOTE, dan klasifikasi dengan *Naive Bayes*, *Logistic Regression*, dan *Support Vector Machine*. Hasil dari evaluasi menunjukkan bahwa algoritma *Naive Bayes* memberikan performa paling baik dengan *accuracy* sebesar 81.90%, *precision* 79.95%, *recall* 81.90%, dan *F1-score* 80.87%. Sementara itu, SVM mencapai *accuracy* 66.67%, sedangkan *Logistic Regression* mendapatkan *accuracy* 62.86%. Hasil ini mengindikasikan bahwa *Naive Bayes* lebih efisien dalam mendeteksi pola emosi di dataset penelitian jika dibandingkan kedua algoritma lainnya. Ketidakseimbangan distribusi emosi yang ditemukan pada tahap awal kajian diduga mempengaruhi kinerja model klasifikasi. Meskipun teknik SMOTE telah diterapkan untuk memperbaiki keseimbangan data latih, ketiga model masih cenderung lebih baik dalam mengidentifikasi kelas positif dibandingkan dengan kelas negatif. Namun, *Naive Bayes* menunjukkan ketahanan yang lebih baik dalam kondisi ini dengan menghasilkan nilai *True Positive* tertinggi serta *False Negative* terendah dibandingkan dengan yang lain. Dengan demikian, *Naive Bayes* adalah pendekatan klasifikasi sentimen terbaik untuk komentar *trailer* film Andai Ibu Tidak Menikah dengan Ayah dalam penelitian ini. Untuk meningkatkan hasil, gunakan dataset yang lebih besar, metode ekstraksi fitur yang lebih kompleks, dan bandingkan sistem klasifikasi dalam penelitian selanjutnya.

Daftar Pustaka

[1] I. I. J. Rifka Alkhilyatul Ma'rifat, I Made Suraharta, "Konstruksi Peran Gender dan Relasi Kuasa dalam Keluarga (Studi Kasus Film 'Andai Ibu Tidak Menikah dengan Ayah') Abstrak," vol. 2, pp. 306–312, 2024.

[2] S. Khomsah, "Sentiment Analysis On YouTube Comments Using Word2Vec and Random Forest," *Telematika*, vol. 18, no. 1, p. 61, 2021, doi: 10.31315/telematika.v18i1.4493.

[3] Andi Riswawan, "Implementing TF-IDF and Logistic Regression for Sentiment Analysis of YouTube Comments on the iPhone 16," *J. Teknol. Dan Open Source*, vol. 8, no. 2, pp. 604–611, 2025, doi: 10.36378/jtos.v7i2.4753.

[4] R. A. N. Rizka and Zaehol Fatah, "Analisis Sentimen Komentar YouTube terhadap Tragedi Demo 25 Agustus Menggunakan Pendekatan Lexicon-Based," *J. Mhs. Tek. Inform.*, vol. 4, no. 2, pp. 217–224, 2025, doi: 10.35473/jamastika.v4i2.4525.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

[5] S. Ratnaswari, N. C. Wibowo, and D. S. Y. Kartika, "Analisis Sentimen Menggunakan Metode Lexicon-Based Dan Support Vector Machine Pada Presiden Dan Wakil Presiden Indonesia Periode 2024–2029," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 1, pp. 362–368, 2025, doi: 10.23960/jitet.v13i1.5604.

[6] A. K. Dewi, N. F. Rahmadani, R. Syahputri, L. R. Nasution, and M. Furqon, "Deteksi Berita Hoax Pada Platform X Menggunakan Pendekatan Text Mining dan Algoritma Machine Learning," *Data Sci. Indones.*, vol. 5, no. 1, pp. 33–46, 2025, [Online]. Available: https://itscience-indexing.com/jurnal/index.php/dsi/article/view/6011

[7] B. Of and C. Science, "BULLETIN OF COMPUTER SCIENCE RESEARCH Identifikasi Indikasi Risiko Depresi pada Unggahan Media Sosial X Menggunakan Natural Language Processing dan Algoritma Random," vol. 6, no. 2, pp. 813–822, 2026, doi: 10.47065/bulletincsr.v6i2.1026.

[8] O. I. Gifari, M. Adha, F. Freddy, and F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *J. Inf. Technol.*, vol. 2, no. 1, pp. 36–40, 2022, doi: 10.46229/jifotech.v2i1.330.

[9] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks," *Sintesia*, vol. 1, pp. 81–88, 2022.

[10] Ridwan, E. H. Hermaliani, and M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada," *Comput. Sci.*, vol. 4, no. 1, pp. 80–88, 2024, [Online]. Available: http://jurnal.bsi.ac.id/index.php/co-science

[11] F. Astuti, R. M. Candra, S. Agustian, and S. Ramadhani, "Klasifikasi Sentimen Masyarakat Terhadap Pemerintah Terkait Penerapan Kebijakan New Normal Menggunakan Metode K-Nearest Neighbor," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 3, pp. 531–538, 2022, doi: 10.32672/jnkti.v5i3.4455.

[12] M. Y. Dhinora and E. Mailoa, "Analisa Tweet Mahasiswa untuk Deteksi Gejala Depresi dengan Penerapan Natural Language Processing," *J. Indones. Manaj. Inform. dan Komun.*, vol. 6, no. 2, pp. 1193–1211, 2025, doi: 10.63447/jimik.v6i2.1405.

[13] R. N. Setiawan and W. Maharani, "Depression Detection on Social Media Twitter Using Hierarchical Attention Network Method," *J. Inf. Syst. Res.*, vol. 3, no. 4, pp. 446–452, 2022, doi: 10.47065/josh.v3i4.1857.

[14] D. Arsyanda, S. Rodiah, and A. S. Rohman, "Peran akun autobase X (twitter) dalam memenuhi kebutuhan informasi followers," *Pustaka Karya J. Ilm. Ilmu Perpust. dan Inf.*, vol. 13, no. 1, pp. 233–241, 2025, doi: 10.18592/pk.v13i1.15927.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)