Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3168~3176
e-ISSN: 2548-6861

Opinion Classification on IMDb Reviews Using Naïve Bayes Algorithm

Amiliya Putri [1]*, Khothibul Umam [2]*, Hery Mustofa [3]*, Adzhal Arwani Mahfudh [4]*
* Teknologi Informasi, Fakultas Sains dan Teknologi, UIN Walisongo Semarang
2208096082@student.walisongo.ac.id [1], khothibul_umam@walisongo.ac.id [2], herymustofa@walisongo.ac.id [3], adzhal@walisongo.ac.id [4]

Article Info

Article history:

Received 2025-06-16
Revised 2025-10-14
Accepted 2025-11-05

Keyword:

IMDb,
Naïve Bayes,
Opinion Classification,
Natural Language Processing,
Sentiment Analysis.

ABSTRACT

This study aims to classify user opinions on IMDb movie reviews using the *Multinomial Naïve Bayes* algorithm. The dataset consists of 50,000 reviews, evenly distributed between 25,000 positive and 25,000 negative reviews. The preprocessing stage includes cleaning, case folding, stopword removal, tokenization, and lemmatization using the NLTK library. Text features are represented through the TF-IDF method to capture the significance of each word in the documents. The *Multinomial Naïve Bayes* model was trained using the hold-out validation technique with an 80:20 split for training and testing data. Hyperparameter tuning of α (*Laplace smoothing*) was conducted to enhance model stability and accuracy. The model's performance was evaluated using accuracy, precision, recall, and F1-score metrics, supported by a confusion matrix visualization. The results show that the model achieved an accuracy of 87%, with precision of 87.9%, recall of 85.4%, and an F1-score of 86.6%. In comparison, *Logistic Regression* as a baseline algorithm achieved an accuracy of 91%. Nevertheless, the *Naïve Bayes* algorithm remains competitive and computationally efficient for large-scale text data, making it highly relevant for sentiment analysis of movie reviews.

I. PENDAHULUAN

Seiring berkembangnya zaman digital moderen, kemajuan internet yang signifikan telah memudahkan akses serta distribusi informasi secara luas. Setiap orang kini memiliki kesempatan untuk mengakses, membagikan, dan memberikan pendapat secara terbuka melalui berbagai platform digital. Kegiatan tersebut kini telah menjadi rutinitas dalam kehidupan masyarakat modern dan membawa pengaruh yang signifikan di berbagai bidang kehidupan [1]. Salah satunya adalah opini atau ulasan pengguna yang kini memegang peran penting dalam memengaruhi keputusan orang lain, baik dalam memilih produk, makanan, pakaian, hingga tontonan seperti film. Ulasan dari pengguna dapat memberikan gambaran awal dan membantu seseorang dalam menentukan pilihan film yang ingin mereka tonton. Hal ini semakin jelas terlihat pada platform besar seperti IMDb, di mana ulasan dan penilaian pengguna menjadi referensi utama bagi para pecinta film dalam menentukan pilihannya.

Situs web *Internet Movie Database* (IMDb) dikenal sebagai tempat yang menyediakan berbagai data terkait perfilman dan proses produksinya dari berbagai belahan dunia. Informasi tersebut mencakup detail mengenai para pemeran, sutradara, penulis skenario, hingga kru lainnya seperti penata rias dan penyusun musik latar. IMDb pertama kali muncul sebagai sebuah basis data film yang dikelola oleh komunitas penggemar pada tahun 1990. Kemudian, layanan ini mulai tersedia secara daring sejak tahun 1993. Sejak tahun 1998, IMDb dimiliki dan dijalankan oleh IMDb.com, Inc., yang merupakan anak perusahaan dari Amazon.com [2]. IMDb menyediakan ruang bagi komunitas pengguna untuk secara aktif memberikan ulasan dan penilaian terhadap film. Tidak hanya dari kalangan umum, para ahli juga memiliki platform tersendiri untuk menyampaikan ulasan dan memberikan opini secara profesional.

Film merupakan salah satu objek yang menarik untuk dianalisis, karena dalam memberikan pendapat tentang sebuah film, pengguna cenderung membahas berbagai elemen yang membentuk film tersebut [3]. Banyak pengguna IMDb yang membagikan tangapan terhadap film yang mereka tonton, baik berupa opini positif maupun negatif. Ulasan-ulasan ini menjadikan film sebagai objek yang kaya akan informasi yang dapat digali kembali [3].

Namun, seiring bertambahnya volume ulasan, penilaian opini dengan cara tradisional tidak lagi memberikan hasil yang efisien. Untuk menjawab tantangan ini, dibutuhkan teknologi yang mampu secara otomatis mengidentifikasi dan mengklasifikasikan opini berbasis teks. Dalam konteks pengolahan data teks, *Natural Language Processing* (NLP) menjadi salah satu teknik penting yang memampukan sistem komputer untuk menafsirkan dan mengklasifikasikan bahasa manusia secara otomatis.

Bidang NLP termasuk dalam ranah *machine learning* yang berfokus pada pengembangan kemampuan komputer untuk memahami serta memproses bahasa manusia dalam bentuk teks sehari-hari [4]. Teknologi ini dirancang agar komputer bisa mengenali aturan tata bahasa dan makna kalimat, lalu mengubahnya menjadi format yang bisa dipahami oleh mesin. Dengan bantuan perangkat lunak NLP, proses analisis data teks dapat dilakukan secara otomatis, termasuk dalam mengidentifikasi maksud maupun sentimen dari pesan, sehingga memungkinkan sistem untuk merespons interaksi manusia secara *real-time*.

Beberapa penelitian terdahulu telah meneliti penerapan algoritma Naïve Bayes di berbagai bidang. Salah satunya dilakukan oleh Awangga dan Khonsa (2022), yang melakukan perbandingan performa antara algoritma Random Forest dan Multinomial Naïve Bayes menggunakan dua kumpulan data berbeda, yaitu ulasan terkait produk farmasi serta film. Hasilnya menunjukkan bahwa kedua metode memiliki tingkat akurasi yang sama, yaitu 0,56%. Namun, dari sisi efisiensi, Naïve Bayes lebih unggul karena hanya membutuhkan waktu pelatihan 7 detik, sedangkan Random Forest memerlukan hingga 120 detik [1]. Penelitian lain yang dilakukan oleh Apriliyani dan rekan-rekan (2024) menggunakan algoritma Naïve Bayes untuk melakukan analisis sentimen terhadap ulasan pengguna aplikasi Duolingo yang diunduh dari Google Play Store. Data yang digunakan berjumlah sekitar 1.000 ulasan, dengan distribusi yang seimbang antara sentimen positif dan negatif. Hasil eksperimen menunjukkan bahwa model tersebut memperoleh tingkat akurasi sebesar 86%, *precision* 89%, *recall* 83%, dan *F1-score* 86%. Hasil tersebut menegaskan bahwa Naïve Bayes tetap kompetitif meskipun diterapkan pada domain berbeda, sehingga relevan untuk digunakan dalam klasifikasi opini pada dataset IMDb yang lebih besar [5]. Selain itu, Widyaningtias dkk. (2025) memanfaatkan Naïve Bayes untuk mengklasifikasikan genre film terpopuler bulanan berdasarkan data penayangan dari platform streaming. Model yang dikembangkan menghasilkan akurasi sebesar 57,74%. Walaupun tingkat akurasi tidak setinggi pada kasus klasifikasi sentimen, penelitian ini menunjukkan fleksibilitas Naïve Bayes dalam mengolah data berskala besar, baik teks maupun numerik [6]. Nurtikasari dkk. (2022 memanfaatkan algoritma Naïve Bayes untuk menganalisis sentimen penonton terhadap film *Ngeri-Ngeri Sedap* dengan sumber data berasal dari platform Twitter. Dataset yang digunakan mencakup sekitar 404 cuitan, yang dikategorikan menjadi tiga jenis sentimen, yaitu positif, negatif, dan netral. Berdasarkan hasil pengujian, model mencapai akurasi sebesar 75%, *precision* 80%, dan *recall* 79%, menunjukkan bahwa Naïve Bayes mampu memberikan hasil yang cukup baik sehingga mendukung penerapannya pada dataset IMDb yang lebih besar [7]. Sementara itu Pratiwi dan Nugroho (2017) meneliti prediksi rating film menggunakan Naïve Bayes dengan dataset dari Kaggle. Berdasarkan hasil eksperimen, model ini mampu mencapai tingkat ketepatan sekitar 65,56% untuk akurasi, 81,20% untuk *precision*, dan 66,77% untuk *recall*. Temuan ini menunjukkan bahwa Naïve Bayes merupakan metode yang andal dalam pengelompokan berbasis teks serta memiliki kinerja yang cukup kompetitif di berbagai bidang penerapan [8].

Berdasarkan tinjauan literatur, penelitian ini menggunakan Multinomial Naïve Bayes (MNB) untuk mengklasifikasikan opini pada ulasan film di IMDb. Algoritma ini dipilih karena sesuai untuk teks, mempertimbangkan frekuensi kata, serta unggul dalam memproses data berskala besar secara efisien. Model akan memprediksi apakah sebuah ulasan bernuansa positif atau negatif berdasarkan istilah dalam teks. Hasil klasifikasi diharapkan dapat merepresentasikan persepsi publik terhadap film dan memberikan masukan berharga bagi pembuat maupun penonton [9] [10].

Penelitian ini menggunakan dataset IMDb berjumlah 50.000 ulasan yang seimbang, terdiri atas 25.000 ulasan bernada positif dan 25.000 ulasan lainya negatif. Tahapan *preprocessing* dilakukan secara menyeluruh, meliputi *cleaning*, *case folding*, *stopword removal*, *tokenization*, dan *lemmatization* dengan pembobotan TF-IDF untuk menghasilkan representasi fitur yang lebih akurat serta mengatasi tantangan data *sparsity*. Selain itu, word cloud digunakan untuk menggambarkan kata dominan pada tiap sentimen.

Dataset IMDb 50K yang dimanfaatkan pada penelitian ini berasal dari sumber terbuka dan telah banyak digunakan dalam berbagai kajian terkait analisis sentimen. Namun, penggunaannya tetap relevan karena dataset ini bersifat seimbang dan memungkinkan perbandingan langsung dengan penelitian terdahulu. Dengan karakteristik tersebut, dataset IMDb menjadi dasar yang tepat untuk menguji efektivitas algoritma *Multinomial Naïve Bayes* dalam skala besar.

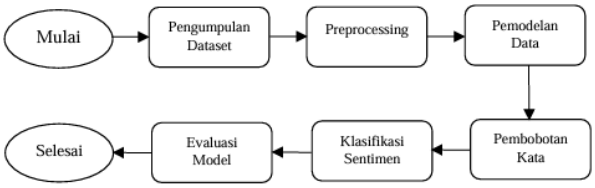
Hasil penelitian ini diharapkan dapat memberikan nilai tambah, baik dalam pengembangan metode analisis sentimen maupun penerapannya pada data ulasan film. Pertama, mengembangkan pipeline analisis sentimen yang lebih komprehensif dan efisien melalui penerapan tahapan *preprocessing* lengkap, meliputi *cleaning*, *case folding*, *stopword removal*, *tokenization*, dan *lemmatization* yang dikombinasikan dengan pembobotan TF-IDF. Kombinasi ini dirancang untuk menghasilkan representasi teks yang lebih akurat dan bermakna, sekaligus mengatasi permasalahan *data sparsity* yang umum terjadi pada data ulasan film.

Kedua, menghadirkan evaluasi model yang lebih mendalam dengan menambahkan analisis kesalahan (*error analysis*), serta perbandingan kinerja dengan algoritma *baseline* Logistic Regression. Pendekatan ini memungkinkan penilaian performa model secara lebih menyeluruh, tidak hanya berdasarkan metrik umum seperti akurasi, tetapi juga dari segi stabilitas dan kemampuan generalisasi terhadap data baru. Ketiga, menyediakan potensi penerapan hasil model secara praktis, baik sebagai komponen dalam sistem rekomendasi film maupun sebagai *dashboard* analisis opini penonton yang menampilkan tren sentimen secara visual. Hasil integrasi model ini dapat dimanfaatkan oleh pengembang platform dan sektor perfilman untuk menganalisis perilaku penonton serta mengadaptasi strategi peningkatan pengalaman pengguna secara lebih terarah.

Dengan demikian, kebaruan penelitian ini tidak terletak pada pengembangan algoritma baru, melainkan pada optimalisasi dan penerapan pipeline analisis sentimen berbasis Naïve Bayes secara lebih menyeluruh dan aplikatif untuk mendukung pengambilan keputusan berbasis opini publik.

II. METODE

Perancangan sistem dalam penelitian ini dilakukan secara bertahap dan dirangkum secara keseluruhan dalam bentuk bagan alur yang ditampilkan pada diagram alur berikut.



Gambar 1. Flowchart Alur Penelitian

A. Pengambilan Dataset

Dataset yang digunakan diambil dari situs Kaggle. Dataset ini mencakup sebanyak 50.000 data ulasan berbahasa inggris, masing-masing memiliki dua atribut utama: kolom *review* yang memuat isi ulasan film, dan kolom *sentiment* yang berfungsi sebagai penanda kategori sentimen dari setiap ulasan, dengan label positif atau negatif [11]. Dari total data, terdapat 25.000 ulasan positif dan 25.000 ulasan negatif, sehingga dataset ini bersifat seimbang (*balanced dataset*). Dataset ini dimanfaatkan sebagai sumber data utama dalam proses pelatihan dan pengujian model analisis sentimen.

B. Preprocessing

Setelah data dikumpulkan, langkah selanjutnya adalah pemrosesan awal guna menyiapkan data mentah sehingga bisa diolah lebih lanjut oleh system [12]. Dalam penelitian ini, tahapan *preprocessing* dilakukan melewati lima proses pengolahan sebagai berikut:

1) Cleaning: Tahap pembersihan dilakukan untuk menghapus unsur-unsur yang tidak dibutuhkan dari teks, seperti tag HTML, tautan (URL), angka, dan tanda baca. Tahapan ini bertujuan untuk melakukan pembersihan terhadap teks ulasan agar kontennya lebih terstruktur dan siap diolah secara otomatis oleh sistem [12].

2) Case folding: Proses normalisasi teks yang berfungsi menyamakan semua karakter dengan mengubahnya menjadi huruf kecil [13]. Dalam pemrosesan teks, komputer memperlakukan huruf besar dan kecil sebagai simbol yang berbeda. Oleh karena itu, penyamaan format huruf menjadi penting untuk memastikan konsistensi data dan mempermudah tahap analisis berikutnya.

3) Stopwords: Tahapan penghapusan kata-kata umum yang sering muncul pada dokumen, namun tidak memberikan makna signifikan terhadap proses analisis teks [13]. Contohnya termasuk kata *"the"*, *"and"*, *"is"*, serta kata-kata sejenisnya. Mengingat dataset dalam penelitian ini berbahasa Inggris, maka daftar *stopwords* yang diterapkan juga berasal dari bahasa Inggris. Pada penelitian ini digunakan daftar *stopword*s standar dari pustaka NLTK (*Natural Language Toolkit*) tanpa penyesuaian tambahan. Dengan membersihkan kata-kata tersebut, analisis menjadi lebih terfokus pada istilah yang benar-benar mencerminkan opini pengguna serta membantu mengurangi gangguan (*noise*) dalam data ulasan.

4) Tokenization: Merupakan teknik untuk membagi kalimat menjadi bagian-bagian kecil berupa kata, yang dikenal sebagai token atau potongan teks, guna mempermudah proses analisis [12].

5) Lemmatization: Proses mereduksi kata ke bentuk dasarnya atau *lemma*, agar kata-kata yang memiliki arti serupa dapat direpresentasikan secara seragam [14]. Contohnya, kata *"better"* akan dikenali sebagai bentuk turunan dari *"good"*, dan dikembalikan ke lema *"good"*. Proses ini membantu meningkatkan kualitas fitur yang diekstrak dari dokumen karena membuat data lebih bersih dan terstruktur [15]. Berdasarkan penelitian, *lemmatization* memiliki kecenderungan menghasilkan performa yang lebih baik dalam klasifikasi teks dibandingkan metode lain seperti *stemming* [14]. Dalam penelitian ini, proses *lemmatization* dilakukan menggunakan WordNetLemmatizer dari pustaka NLTK, yang memanfaatkan basis data WordNet untuk mengembalikan kata ke bentuk dasarnya.

Proses *preprocessing* dalam studi ini diimplementasikan menggunakan Python, di mana pustaka NLTK digunakan untuk mendukung langkah *tokenisas*i, *stopword removal*, dan *lemmatization*. Sementara itu, proses pembobotan fitur menggunakan TF-IDF, pemodelan, serta evaluasi dilakukan dengan pustaka Scikit-learn.

C. Pemodelan Data

Setelah tahapan *preprocessing* selesai dilakukan, Langkah selanjutnya adalah pemodelan data. Langkah ini bertujuan untuk menciptakan sebuah sistem klasifikasi yang dapat mengelompokkan data sesuai dengan pola-pola yang teridentifikasi dalam data latih [3]. Pada penelitian ini, dataset IMDb dibagi menjadi dua subset menggunakan teknik *hold-out validation*, yaitu 80% untuk data latih (*training set*) dan 20% untuk data uji (*testing set*) [16]. Pembagian dilakukan secara acak (*random split*), namun dengan menetapkan parameter *random_state*, sehingga hasil pembagian data tetap konsisten setiap kali proses dijalankan. Hal ini penting agar evaluasi model dapat direplikasi dengan adil tanpa dipengaruhi variasi pembagian data. Data latih digunakan untuk membangun model, sedangkan data uji berfungsi untuk mengukur kemampuan model dalam memprediksi ulasan baru yang belum pernah dilihat sebelumnya.

Selain pembagian data menggunakan *hold-out validation*, penelitian ini juga menerapkan proses *k-fold cross-validation* secara internal selama tahap pencarian parameter terbaik. Proses ini dilakukan melalui *RandomizedSearchCV* dengan nilai *cv = 4*, yang berarti data latih dibagi menjadi empat lipatan. Setiap lipatan secara bergantian berperan sebagai data validasi, sementara tiga lipatan lainnya digunakan untuk pelatihan. Pendekatan ini membantu memperoleh hasil tuning parameter yang lebih stabil dan menghindari bias akibat pembagian data tunggal. Nilai *best_score_* yang dihasilkan merepresentasikan rata-rata performa model terbaik dari proses *cross-validation* tersebut.

Selanjutnya, dilakukan penyetelan *hyperparameter* terhadap nilai α (*alpha*), yaitu parameter *Laplace smoothing* pada algoritma Multinomial Naïve Bayes. Proses tuning ini juga dilakukan menggunakan *RandomizedSearchCV* dengan rentang pencarian nilai α dari 10^{-3} hingga 10^{3}. Penerapan *Laplace smoothing* penting untuk mengatasi permasalahan probabilitas nol pada kata-kata yang jarang muncul, sehingga membantu meningkatkan stabilitas serta akurasi model dalam proses klasifikasi sentimen.

D. Pembobotan Kata

Sebelum data dimasukkan ke dalam model klasifikasi, istilah-istilah dalam teks perlu diganti menjadi format numerik agar nantinya dapat diproses oleh komputer. Salah satu pendekatan umum dalam pemberian bobot terhadap istilah ialah TF-IDF yang berperan untuk menilai tingkat relevansi sebuah kata terhadap sebuah teks dibandingkan dengan keseluruhan korpus lainnya [17]. Bobot TF-IDF meningkat ketika suatu kata memiliki kemunculan yang dominan pada satu dokumen tetapi tidak umum di dokumen lain. Mekanisme ini membantu model mengenali tingkat kepentingan kata terhadap konteks dokumen [5].

Dalam representasi data teks, kondisi *sparsity* (kepadatan data yang rendah) merupakan hal yang umum terjadi karena setiap dokumen hanya memuat sebagian kecil dari keseluruhan kosakata. Permasalahan ini ditangani melalui penerapan *TF-IDF Vectorizer*, yang secara implisit mengatasi *sparsity* dengan memberikan bobot lebih tinggi pada kata yang relevan dan unik, sekaligus menurunkan pengaruh kata-kata umum.

E. Klasifikasi Sentimen

Setelah pembobotan kata selesai, proses selanjutnya adalah pengkategorian sentimen. Langkah ini dimaksudkan untuk memisahkan ulasan film dari IMDb ke dalam dua jenis respons, yakni sentimen positif dan negatif. Dalam studi ini, pendekatan yang diterapkan untuk pengelompokan adalah Naïve Bayes, sebuah teknik dalam penambangan data yang mampu membangun model klasifikasi berdasarkan data yang sudah diproses sebelumnya [18]. Tahapan pengelompokan ini dilakukan dengan menghitung kemungkinan setiap kelas berdasarkan frekuensi kemunculan label dalam data pelatihan (*probabilitas prior*) serta kontribusi masing-masing fitur kata dalam teks [3]. Naïve Bayes akan menentukan kelas sentimen dari teks baru berdasarkan pola yang dipelajari dari data *training*. Naïve Bayes unggul dalam segi efisiensi dan performa yang tinggi, terutama dalam menangani dataset berukuran besar.

F. Evaluasi Model

Tahap terakhir dalam penelitian ini adalah penilaian model, yang berfungsi untuk menilai efektivitas algoritma klasifikasi dalam mengelompokkan ulasan film berdasarkan sentimennya. Penilian dilakukan dengan memanfaatkan *confusion matrix*, yang adalah metode umum untuk menilai kinerja model klasifikasi. *Confusion matrix* terdiri dari empat bagian, yakni *True Positive* (TP), yang menunjukan jumlah ulasan dengan label positif yang berhasil teridentifikasi sebagai positif; *False Positive* (FP), yang merujuk pada ulasan dengan sentimen negatif yang salah dikategorikan sebagai positif; *True Negative* (TN), yakni ulasan bernuansa negatif yang berhasil dikenali sebagai negatif secara akurat; sedangkan *False Negative* (FN), yang menggambarkan ulasan positif yang keliru diklasifikasikan sebagai negatif. Berdasarkan keempat komponen ini, dihitunglah metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-score* untuk memberikan gambaran menyeluruh terhadap tingkat ketepatan dan efektivitas model, sebagaimana dijelaskan dalam persamaan (1) hingga (4):

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

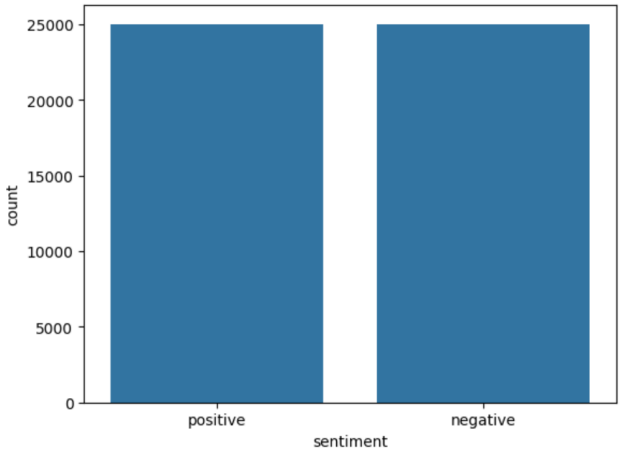
$$\text{F1-score} = 2\ x\ \frac{recall\ x\ presisi}{recall+presisi} \quad (4)$$

Selain confusion matrix dan metrik klasifikasi, penelitian ini juga menggunakan analisis kesalahan, uji robustness, serta membandingkan hasil model dengan algoritma Logistic Regression sebagai *baseline* untuk menilai keunggulan performa Naïve Bayes. Selain itu, word cloud turut digunakan untuk menampilkan kata-kata dominan pada tiap kategori sentimen, sehingga hasil evaluasi menjadi lebih komprehensif baik dari sisi kuantitatif maupun visual.

III. HASIL DAN PEMBAHASAN

A. Pengambilan Dataset

Tahapan pengujian diawali dengan pemanfaatan data IMDb yang diunduh dari situs Kaggle.



Gambar 2. Visualisasi Hasil Data Imbang

Berdasarkan hasil visualisasi, diketahui bahwa dataset bersifat seimbang, dengan total 50.000 ulasan film berbahasa Inggris yang terdiri atas 25.000 ulasan positif dan sisanya negatif. Dataset ini digunakan untuk menguji performa algoritma Multinominal Naïve Bayes dalam mengidentifikasi sentimen ulasan.

B. Preprocessing

Sesudah tahap pengumpulan data, dilakukan proses *preprocessing* yang berfungsi untuk menormalkan dan membersihkan data sebelum masuk ke tahap pemodelan**.** Tahapan ini memegang peranan penting karena berfungsi untuk menyaring, merapikan, dan menyiapkan data mentah sebelum dimasukkan ke dalam proses pelatihan model klasifikasi. Melalui tahap ini, data yang semula belum terstruktur akan diubah menjadi bentuk yang lebih bersih dan konsisten sehingga dapat diolah secara optimal oleh sistem. Ringkasan hasil prapemrosesan disajikan pada tabel berikut:

TABEL I
HASIL PREPROCESSING ULASAN FILM

Review Awal	Hasil Cleaning
Cleaning	Checking Out is an extraordinary film that towers above most film production Its refreshing witty hu...
Case folding	checking out is an extraordinary film that towers above most film production its refreshing witty hu...
Tokenizing	[checking, out, is, an, extraordinary, film, that, towers, above, most, film, production, its, refreshing, witty ...]
Stopwords	[checking, extraordinary, film, towers, film, production, refreshing, witty, humor, never ...]
Lemmatization	checking extraordinary film tower film production refreshing witty humor never excuse remain superfi...

C. Pemodelan Data

Proses pemodelan dilakukan setelah tahap *preprocessing* data selesai. Dataset IMDb dibagi menggunakan teknik *hold-out validation* dengan proporsi 80% untuk data pelatihan dan 20% untuk data pengujian. Proses pembagian ini dilakukan secara acak dengan parameter *random_state=42* untuk menjaga konsistensi hasil setiap kali kode dijalankan. Dari total 50.000 data ulasan, sebanyak 40.000 digunakan untuk proses pelatihan model, sementara 10.000 sisanya dimanfaatkan sebagai data uji.

Selama proses pelatihan, dilakukan pula pencarian parameter terbaik menggunakan RandomizedSearchCV dengan nilai *cv = 4* sebagai bagian dari *k-fold cross-validation*. Proses ini bertujuan untuk memastikan performa model tetap stabil dan tidak bergantung pada satu pembagian data tunggal.

Hasil tuning menunjukkan bahwa nilai α (Laplace smoothing) optimal berada dalam rentang yang menghasilkan *best cross-validation score* sebesar 0.8611. Hasil tersebut mengindikasikan bahwa model mampu mempertahankan kinerja yang stabil dan konsisten selama proses pelatihan. Selanjutnya, model yang telah terbentuk diuji menggunakan data pengujian untuk mengukur sejauh mana kemampuannya dalam melakukan generalisasi terhadap data yang belum pernah dipelajari sebelumnya.

D. Pembobotan Kata

Setelah pemodelan disiapkan, tahap berikutnya adalah pembobotan kata melalui pendekatan TF-IDF. Pendekatan ini digunakan untuk menetapkan nilai numerik pada tiap kata

dalam teks, tujuannya untuk menggambarkan tingkat relevansi kata tersebut dalam suatu dokumen. TF-IDF sendiri merupakan hasil penggabungan dari dua prinsip utama: TF, yang jumlah kemunculan kata tertentu di dalam dokumen dan IDF, yang mengukur tingkat keunikan kata tersebut dengan membandingkan kemunculannya di seluruh dokumen dalam kumpulan data.

	bad	character	even	film	get	good	great	like
0	0.0	0.000000	0.000000	0.000000	0.707709	0.000000	0.000000	0.000000
1	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.284245	0.000000
2	0.0	0.383109	0.380946	0.000000	0.000000	0.000000	0.431491	0.000000
3	0.0	0.000000	0.000000	0.458828	0.000000	0.000000	0.000000	0.262054
4	0.0	0.289177	0.000000	0.211458	0.144400	0.271022	0.000000	0.000000

Gambar 3. Hasil TF-IDF pada lima dokumen pertama

Pada gambar 3, setiap baris mewakili satu ulasan, sedangkan setiap kolom menunjukkan kata penting (fitur). Nilai dalam tabel menunjukkan bobot TF-IDF dari kata-kata seperti "*film*", "*get*", "*great*", dan "*character*". Misalnya, pada dokumen ke-0, kata "*get*" memiliki bobot tertinggi yaitu 0.707709, yang menunjukkan bahwa kata tersebut sangat relevan dan dominan dalam menyampaikan opini pada ulasan tersebut. Sebaliknya, nilai 0.0 mengindikasikan bahwa kata tersebut tidak muncul di dokumen atau terlalu umum sehingga kontribusinya terhadap klasifikasi menjadi rendah.

Hasil ini menunjukkan bahwa metode TF-IDF mampu menekankan kata-kata yang memiliki pengaruh paling besar terhadap pembentukan sentimen, sambil meminimalkan peran kata-kata umum yang kurang memberikan informasi penting. Bobot yang dihasilkan dari proses ini selanjutnya dijadikan sebagai representasi fitur pada tahap pelatihan model klasifikasi.

E. Klasifikasi Sentimen

Setelah dokumen berhasil direpresentasikan secara numerik, proses klasifikasi dapat dilakukan. Teknik yang diterapkan adalah algoritma Naïve Bayes. Model menganalisis data pelatihan untuk mengenali karakteristik ulasan positif dan negatif, kemudian menerapkan pengetahuan tersebut untuk memprediksi sentimen dari data uji. Gambar 4 menyajikan hasil evaluasi terhadap data pelatihan (*train report*) yang telah dilakukan.

Train report

	precision	recall	f1-score	support
negative	0.90	0.93	0.92	20039
positive	0.93	0.90	0.91	19961
accuracy			0.92	40000
macro avg	0.92	0.92	0.92	40000
weighted avg	0.92	0.92	0.92	40000

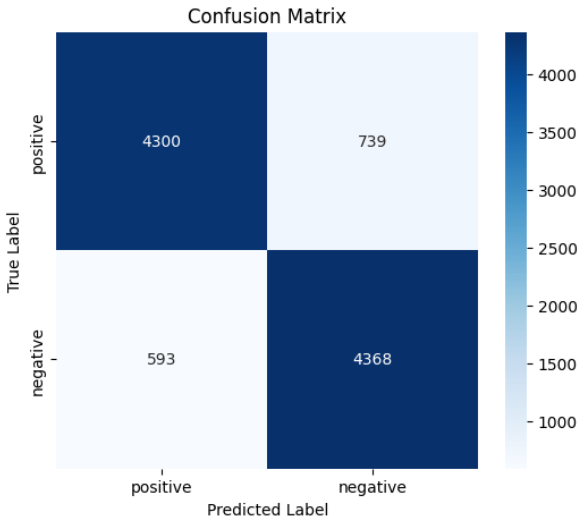
Gambar 4. Laporan hasil klasifikasi data pelatihan

Berdasarkan evaluasi, model berhasil memberikan performa yang optimal dengan tingkat ketepatan mencapai 92%. Untuk kelas sentimen negatif precision 90% dan recall 93%. Disisi lain, untuk sentimen positif, precision 93% dan recall 90%. Kedua kelas memiliki nilai *f1-score* sekitar 91%, yang menunjukan proporsi seimbangan antara tingktan ketepatan dan kemampuan deteksi dalam proses klasifikas. Nilai *macro average* dan *weighted average* juga berada pada angka yang sama, yaitu 92%, menunjukkan bahwa model mampu mendeteksi pola dari data pelatihan secara optimal dan konsisten.

F. Evaluasi Model

Untuk mengetahui seberapa baik model bekerja dalam praktiknya, dilakukan tahap evaluasi terhadap hasil klasifikasi. Evaluasi ini mencakup *confusion matrix* dan *word cloud*. Kedua metode ini digunakan untuk menilai performa model dalam mengklasifikasikan opini pada ulasan film di IMDb.

1) Confusion Matrix



Gambar 5. Visualisasi Output Klasifikasi Confusion Matrix

Berdasarkan hasil visualisasi, dapat diketahui bahwa model berhasil mengklasifikasikan 4368 ulasan negatif (*True Negative*). Namun, ada 593 ulasan negatif yang salah teridentifikasi sebagai positif (*False Positive*). Selain itu, model juga berhasil mengklasifikasikan 4300 ulasan positif (*True Positive*), sementara 739 ulasan positif teridentifikasi secara tidak tepat sebagai ulasan negatif (*False Negative*). Bedasarkan data tersebut, metrik evaluasi dihitung sebagai berikut :

$$Accuracy = = \frac{4300 + 4368}{4300 + 739 + 593 + 4368} = \frac{8668}{10000} = 0.87$$

$$Precision = \frac{4300}{4300 + 593} = \frac{4300}{4893} = 0.879$$

$$Recall = = \frac{4300}{4300 + 739} = \frac{4300}{5039} = 0.854$$

$$F1 = \frac{2\ x\ (0.879\ x\ 0.854)}{0.879 + 0.854} = \frac{1.501}{1.733} = 0.866$$

Tingkat akurasi menggambarkan seberapa besar proporsi prediksi model yang sesuai dengan label sebenarnya pada data uji. Ukuran *precision* merepresentasikan ketepatan model dalam mengklasifikasikan data positif secara benar, sedangkan *recall* mengukur sejauh mana model dapat mengenali seluruh data positif yang ada. Nilai *F1-score* berfungsi untuk menyeimbangkan kedua metrik tersebut melalui rata-rata harmonik. Berdasarkan pengujian, model menunjukkan akurasi 87%, *precision* 87,9%, *recall* 85,4%, dan *F1-score* 86,6%, sehingga dapat disimpulkan bahwa model memiliki kemampuan klasifikasi yang stabil dan seimbang antara dua kelas sentimen.

Selain memberikan gambaran kuantitatif, *confusion matrix* juga digunakan untuk mengidentifikasi pola kesalahan klasifikasi. Hasil menunjukkan bahwa kesalahan terbanyak terjadi pada ulasan positif yang diprediksi sebagai negatif (FN = 739), dibandingkan ulasan negatif yang diklasifikasikan sebagai positif (FP = 593). Hal ini menandakan bahwa model cenderung kurang sensitif terhadap opini positif yang mengandung kritik halus atau struktur kalimat kompleks. Untuk memperjelas pola tersebut, dilakukan *error analysis* dengan meninjau dua kategori utama, yaitu *False Positive* dan *False Negative*. Ringkasan hasil analisis ditampilkan pada Tabel 2 berikut:

TABEL 2
HASIL KESALAHAN KLASIFIKASI

Jenis	Asli	Pred.	Ulasan	Analisis Singkat
FP	Neg	Pos	"Although this series and the mini film in particular were very important at the time of release…"	Model keliru karena adanya kata awal bernuansa positif (*important*, *significant*) sehingga gagal menangkap kritik utama.
FN	Pos	Neg	"Okay, I didn't get the Purgatory thing ... really caught my attention."	Kata negatif (*didn't get*) membuat konteks positif keliru diprediksi.

Keterangan:
FP = *False Positive*
FN = *False Negative*
Asli = Label sentiment sebenarnya dari data ulasan
Pred. = Label sentiment hasil prediksi model

Selain itu, penelitian ini turut mengevaluasi tingkat *robustness* dari model yang dikembangkan. Ketahanan model terhadap data baru diuji menggunakan *testing set* (X_test) yang telah dipisahkan sejak awal proses dan tidak disertakan dalam tahap pelatihan. Berdasarkan hasil pengujian, model menunjukkan kinerja yang stabil dengan nilai akurasi sebesar 87%, sedikit lebih rendah dibandingkan hasil pelatihan sebesar 92%. Selisih ini masih dalam batas yang wajar dan mengindikasikan bahwa model tidak mengalami *overfitting*. Dengan demikian, model dapat dikatakan memiliki kemampuan generalisasi yang baik terhadap data baru yang belum pernah ditemui sebelumnya.

Uji khusus terhadap *noisy data* (seperti ulasan dengan typo, singkatan, atau tata bahasa tidak baku) belum dilakukan. Meski demikian, tahapan *preprocessing* yang meliputi *cleaning*, *case folding*, *stopword removal*, dan *lemmatization* telah membantu meningkatkan ketahanan model terhadap sebagian noise yang terdapat dalam dataset IMDb.

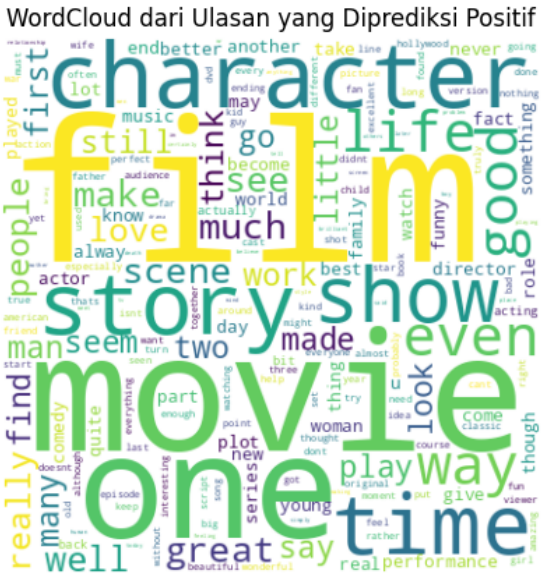
Untuk memperkuat evaluasi, penelitian ini juga membandingkan performa Multinomial Naïve Bayes dengan Logistic Regression sebagai *baseline*. Logistic Regression dipilih karena merupakan algoritma populer dalam analisis sentimen dan dikenal memiliki kinerja kompetitif.

TABEL 2
PERBANDINGAN NAÏVE BAYES DENGAN LOGISTIC REGRESSION

Metrik	Naive Bayes (Test)	Logistic Regression (Test)
Accuracy	0.87	0.91
Precision (Neg)	0.86	0.92
Recall (Neg)	0.88	0.89
F1-score (Neg)	0.87	0.91
Precision (Pos)	0.88	0.90
Recall (Pos)	0.85	0.92
F1-score (Pos)	0.87	0.91
Macro Avg F1	0.87	0.91
Weighted Avg F1	0.87	0.91

Berdasarkan hasil tersebut, Logistic Regression menunjukkan performa lebih baik hampir di semua metrik dengan akurasi 91% dibandingkan Naïve Bayes 87%. Logistic Regression juga lebih konsisten dalam mendeteksi ulasan positif, yang sebelumnya menjadi kelemahan utama Naïve Bayes. Meski begitu, Naïve Bayes tetap kompetitif dengan keunggulan efisiensi komputasi sehingga cocok diterapkan pada skala besar.

2) Word Cloud



Gambar 6. Hasil Word Cloud Positif

Output pada ulasan positif menampilkan bahwa kata-kata seperti "*film*", "*movie*", "*character*", "*story*", dan "*great*" paling sering muncul. Ini mencerminkan apresiasi pengguna terhadap alur cerita, karakter, dan pengalaman menonton secara keseluruhan. Hal ini menunjukkan bahwa ulasan yang diprediksi positif memang cenderung berisi opini yang baik terhadap film.

Gambar 7. Hasil World Cloud Negatif

Pada ulasan negatif, kata-kata seperti "*movie*", "*film*", "*character*", "*one*", dan "*really*" mendominasi. Muncul juga kata seperti "*bad*", "*nothing*", dan "*make*", yang menunjukkan ketidakpuasan terhadap aspek film. Ini mendukung bahwa model berhasil mengenali opini negatif dari ulasan.

Hasil klasifikasi opini menggunakan Multinomial Naïve Bayes pada dataset IMDb ini dapat diintegrasikan ke dalam aplikasi praktis. Misalnya, model dapat dijadikan komponen sistem rekomendasi film, di mana ulasan dengan sentimen positif lebih dominan akan meningkatkan skor rekomendasi suatu film kepada calon penonton. Selain itu, model juga dapat digunakan dalam dashboard analisis ulasan, yang menampilkan tren opini penonton secara visual (positif vs. negatif) dalam bentuk grafik atau ringkasan sentimen. Dengan integrasi ini, hasil penelitian tidak hanya bersifat akademis, tetapi juga bermanfaat dalam mendukung pengambilan keputusan, baik oleh pengguna individu maupun pihak industri perfilman.

IV. KESIMPULAN

Berdasarkan hasil penelitian dan analisis yang telah dilakukan, dapat disimpulkan bahwa algoritma Multinomial Naïve Bayes berhasil diterapkan dalam klasifikasi opini pengguna pada ulasan film IMDb, dengan akurasi sebesar 87% pada data uji setelah melalui tahapan *preprocessing* teks dan representasi data menggunakan TF-IDF. Model juga menghasilkan nilai precision sebesar 87,9%, recall sebesar 85,4%, dan F1-score sebesar 86,6%, yang menunjukkan performa klasifikasi yang baik dan seimbang. Hasil perbandingan dengan algoritma Logistic Regression menunjukkan bahwa Logistic Regression memperoleh akurasi lebih tinggi sebesar 91%, namun Naïve Bayes tetap kompetitif dengan keunggulan pada efisiensi komputasi dan kemampuan generalisasi terhadap data baru. Analisis kesalahan mengungkap bahwa sebagian besar kesalahan klasifikasi terjadi pada ulasan positif yang teridentifikasi sebagai negatif, terutama karena ambiguitas makna dan variasi ekspresi bahasa. Kinerja model klasifikasi opini ini sangat dipengaruhi oleh kualitas tahap *preprocessing* dan representasi fitur teks. Untuk penelitian mendatang, pengembangan dapat difokuskan pada perluasan proses *hyperparameter tuning*, peningkatan jumlah data pelatihan, serta penerapan metode tambahan seperti Random Forest, SVM, LSTM, atau BERT guna memperoleh hasil klasifikasi yang lebih optimal dan adaptif terhadap variasi bahasa serta konteks sentimen yang beragam industri dan pengembang platform digital.

DAFTAR PUSTAKA

[1] R. M. Awangga and N. H. Khonsa', "Analisis Performa Algoritma Random Forest dan Naive Bayes Multinomial pada Dataset Ulasan Obat dan Ulasan Film," *InComTech J. Telekomun. dan Komput.*, vol. 12, no. 1, p. 60, Apr. 2022, doi: 10.22441/incomtech.v12i1.14770.

[2] I. A. . Dityawan, "Pengaruh Rating dalam Situs IMDb terhadap Keputusan Menonton di Kota Bandung. (Studi Pada film Halfworlds)," Universitas Telkom, 2016. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/pustaka/121802/pengaruh-rating-dalam-situs-imdb-terhadap-keputusan-menonton-di-kota-bandung-studi-pada-film-halfworlds-.html#:~:text=Internet Movie Database (IMDb) adalah,sampai penata rias dan soundtrack.

[3] F. Ratnawati, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter," *J. INOVTEK POLBENG - SERI Inform.*, vol. 3, no. 1, pp. 50–59, 2018, [Online]. Available: https://ejournal.polbeng.ac.id/index.php/ISI/article/view/335

[4] "Apa itu Pemrosesan Bahasa Alami (NLP)?," Amazon Web Server. Accessed: Apr. 17, 2025. [Online]. Available: https://aws.amazon.com/id/what-is/nlp/

[5] M. Apriliyani, M. I. Musyaffaq, S. Nur'Aini, and K. Umam, "Implementasi analisis sentimen pada ulasan aplikasi Duolingo di Google Playstore menggunakan algoritma Naïve Bayes," *AITI J. Teknol. Inf.*, vol. 21, no. 2, pp. 302–303, 2024, [Online]. Available: https://ejournal.uksw.edu/aiti/article/view/12100

[6] G. R. Widyaningtias, M. Adam, and E. Daniati, "Klasifikasi Genre Film Terpopuler Bulanan Menggunakan Algoritma Naive Bayes Berbasis Data Penayangan," *SEMNAS INOTEK*, vol. 9, pp. 2183–2188, 2025, [Online]. Available: https://proceeding.unpkediri.ac.id/index.php/inotek/

[7] Y. Nurtikasari, Syariful Alam, and Teguh Iman Hermanto, "Analisis Sentimen Opini Masyarakat Terhadap Film Pada Platform Twitter Menggunakan Algoritma Naive Bayes," *INSOLOGI J. Sains dan Teknol.*, vol. 1, no. 4, pp. 411–423, Aug. 2022, doi: 10.55123/insologi.v1i4.770.

[8] R. W. Pratiwi and Y. S. Nugroho, "Prediksi Rating Film Menggunakan Metode Naïve Bayes," *Duta.com*, vol. 12, no. 1, pp. 91–108, 2017, [Online]. Available: file:///C:/Users/ADVAN/Downloads/reverensi artikel nlp/admin,+211-212-1-SM.pdf

[9] C. Rizal, D. A. Kifta, R. H. Nasution, A. Rengganis, and R. Watrianthos, "Opinion classification for IMDb review based using naive bayes method," 2023, p. 030025. doi: 10.1063/5.0171628.

[10] K. Pradeep, C. R. TintuRosmin, S. S. Durom, and G. S. Anisha, "Decision Tree Algorithms for Accurate Prediction of Movie Rating," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Mar. 2020, pp. 853–858. doi: 10.1109/ICCMC48092.2020.ICCMC-000158.

[11] C. Prianto, N. H. Harani, and I. Firmansyah, "Analisis Sentimen Terhadap Kandidat Presiden Republik Indonesia Pada Pemilu 2019 di Media Sosial Twitter," *J. MEDIA Inform. BUDIDARMA*, vol. 3, no. 4, p. 405, Oct. 2019, doi: 10.30865/mib.v3i4.1549.

[12] Arif Widiasan Subagio, Anggraini Puspita Sari, and Andreas Nugroho Sihananto, "Klasifikasi Lexicon-Based Sentiment Analysis Tragedi Kanjuruhan pada Twitter Menggunakan Algoritma Convolutional Neural Network," *J. Ilm. Sist. Inf. dan Ilmu Komput.*, vol. 4, no. 1, pp. 166–177, Jan. 2024, doi: 10.55606/juisik.v4i1.759.

[13] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma SVM untuk Analisis Sentiment pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," *J. Ilm. Edutic*, vol. 7, no. 1, pp. 1–11, 2020, [Online]. Available: https://journal.trunojoyo.ac.id/edutic/article/viewFile/8779/5125

[14] I. M. Yulietha, S. Al Faraby, and Adiwijaya, "Klasifikasi Sentiment Review Film Menggunakan Algoritma Support Vector Machine," *e-Proceeding Eng.*, vol. 4, no. 3, pp. 4748–4749, 2017, [Online]. Available: https://core.ac.uk/download/pdf/299917715.pdf

[15] S. Srinidhi, "Lemmatisasi dalam Pemrosesan Bahasa Alami (NLP) dan Pembelajaran Mesin," bulitin. Accessed: Apr. 19, 2025. [Online]. Available: https://builtin.com/machine-learning/lemmatization

[16] B. V. Haekal, L. Ernawati, and N. Chamida, "Klasifikasi Kepuasan Pengguna Layanan Aplikasi Shopee Menggunakan Metode Decision Tree C4.5," *IFTK*, vol. 17, no. 3, p. 193, 2021, [Online]. Available: file:///C:/Users/ADVAN/Downloads/theresiawati,+188-196.pdf

[17] G. Sanjaya and K. M. Lhaksmana, "No TitleaAnalisis Sentimen Komentar YouTube tentang Terpilihnya Menteri Kabinet Indonesia Maju Menggunakan Lexicon Based," *e-Proceeding Eng.*, vol. 7, no. 3, pp. 9698–9710, 2020, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/14205

[18] A. Tangkelayuk, "The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 1109–1119, Jun. 2022, doi: 10.35957/jatisi.v9i2.2048.